| Statistica Sinica Preprint No: SS-2024-0261 | |
|---|---|
| **Title** | Semi-supervised Regression Analysis with Model Misspecification and High-dimensional Data |
| **Manuscript ID** | SS-2024-0261 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202024.0261 |
| **Complete List of Authors** | Ye Tian,<br>Peng Wu and<br>Zhiqiang Tan |
| **Corresponding Authors** | Zhiqiang Tan |
| **E-mails** | ztan@stat.rutgers.edu |

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



# Semi-supervised Regression Analysis with Model Misspecification and High-dimensional Data

Ye Tian[a], Peng Wu[b] and Zhiqiang Tan[a*]

[a]Rutgers University,

[b]Beijing Technology and Business University

*Abstract:* *The accessibility of vast volumes of unlabeled data has sparked growing interest in semi-supervised learning (SSL) and covariate shift transfer learning (CSTL). In this paper, we present an inference framework for estimating regression coefficients in conditional mean models within both SSL and CSTL settings, while allowing for the misspecification of conditional mean models. We develop an augmented inverse probability weighted (AIPW) method, employing regularized calibrated estimators for both propensity score (PS) and outcome regression (OR) nuisance models, with PS and OR models being sequentially dependent. We show that when the PS model is correctly specified, the proposed estimator achieves consistency, asymptotic normality, and valid confidence intervals, even with possible OR model misspecification and high-dimensional data. Moreover, by suppressing detailed technical choices, we demonstrate that previous methods can be unified within our AIPW framework. Our theoretical findings are verified through extensive simulation studies and a real-world data application.*

*Key words and phrases:* Augmented inverse probability weighted estimator, Covariate shift transfer learning, High-dimensional data, Semi-supervised learning.

*Correspondence to: ztan@stat.rutgers.edu.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

## 1. Introduction

In recent years, vast volumes of unlabeled data have become increasingly accessible, sparking growing interest in how to leverage these data in both academic research and industrial applications. One of the active areas of research is semi-supervised learning (SSL). In addition, covariate shift transfer learning (CSTL) also exploits information from unlabeled data. Both have various application scenarios like computer vision (Sohn et al., 2020; Zhou and Levine, 2021; Zheng et al., 2022), natural language process (Chen and Huang, 2016; Ruder et al., 2019; Zhao et al., 2022), causal inference (Alvari et al., 2019; Aloui et al., 2023; Zhang et al., 2023), healthcare data analysis (Castro et al., 2020; Liu et al., 2023; Tang et al., 2024), etc.

In both SSL and CSTL settings, we have access to a labeled dataset $\mathcal{L}$ and an unlabeled dataset $\mathcal{U}$, where the labeled dataset $\mathcal{L}$ contains observations with both the covariates $\boldsymbol{X}$ and the outcome $Y$, while the unlabeled dataset $\mathcal{U}$ consists solely of observations with the covariates $\boldsymbol{X}$. The training set $\mathcal{T}$ is the union of $\mathcal{L}$ and $\mathcal{U}$. Nevertheless, there is a key distinction between the classic SSL and CSTL setups (Chapelle et al., 2006; Liu et al., 2023). In CSTL, the conditional distributions of $Y$ given $\boldsymbol{X}$ in the labeled and unlabeled datasets are assumed to be the same, whereas the marginal distributions of $\boldsymbol{X}$ are different (hence the term covariate shift), and the estimator is ultimately evaluated on unlabeled data. However, under the classic SSL setup, it is assumed that the distributions of labeled data, unlabeled

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

data, and population are the same, making no difference in evaluating the estimator on which distribution. To accommodate both SSL and CSTL, we consider a more general setting. We only assume the conditional distribution of $Y$ given $\boldsymbol{X}$ is the same in labeled and unlabeled data, while marginal distributions of $\boldsymbol{X}$ are permitted to be different. Estimators evaluated on the population and unlabeled data are both considered.

While there is a long history of SSL (Chapelle et al., 2006; Zhu, 2008) and CSTL (Quiñonero-Candela et al.,2009), a growing literature has considered inference procedures only recently. Notable advancements have been made in estimating the population mean $\mathbb{E}(Y)$ (Zhang et al., 2019; Zhang and Bradic, 2021) and regression coefficients or fixed linear combinations in (generalized) linear models regressing $Y$ against a sub-vector of $\boldsymbol{X}$ (Chakrabortty, 2016; Liu et al., 2023) or in linear models regressing $Y$ against full $\boldsymbol{X}$ (Chakrabortty and Cai, 2018; Chakrabortty et al., 2019; Deng et al., 2024; Zhang et al., 2023; Chen and Zhang, 2023). See Section 5.1 and the Supplementary Material Section S2 for further information. Inferences of quantile regression (Chakrabortty et al., 2022), explained variance (Cai and Guo, 2020), and model performance metrics such as true and false positive rates (Gronsbell and Cai, 2017) have also attracted interest.

In this article, we focus on the inference of regression coefficients in (conditional) mean models for $Y$ against a sub-vector of $\boldsymbol{X}$ in SSL settings, hence called semi-supervised regression analysis. We demonstrate a unified framework for estimating

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

and inferring these coefficients, particularly in cases where the (conditional) mean model and outcome regression (OR) model $\mathbb{E}(Y|\boldsymbol{X})$ may be misspecified. Previous SSL and CSTL methods that considered this type of problems, such as Chakrabortty (2016), Chakrabortty and Cai (2018), Zhang et al. (2019), Zhang and Bradic (2021) and Liu et al. (2023), can largely be accommodated in the augmented inverse probability weighting (AIPW) framework (Robins et al., 1994; Tan, 2020a; Wu et al., 2024). See Section 5 and Section 8 for further details.

Despite significant advancements made, there remain limitations in the previous AIPW methods. Methods developed in SSL settings usually treat the problem as the one where data are missing completely at random (MCAR) (Chakrabortty, 2016; Chakrabortty and Cai, 2018; Zhang et al., 2019; Zhang and Bradic, 2021). This restricts their application scenarios and overlooks the significance of constructing the propensity score (PS) model. In contrast, our setting is in general a missing-at-random (MAR) problem (Little and Rubin, 2019), and the estimation of the PS model is no longer negligible. In the setting of MCAR, the PS remains a constant, whereas in the setting of MAR, the PS varies with the covariates $\boldsymbol{X}$. In MAR problems, similarly as in Tan (2020a), the estimation of PS and OR models needs to be carefully handled in a way different from regularized least squares or maximum likelihood as in previous papers, so that $\sqrt{N}$-consistent estimation can be achieved with possible misspecification of the OR model, where $N$ is the sample size of $\mathcal{T}$.

In summary, we mainly make two contributions. First, we present an inference

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

framework that accommodates several previous settings, including the estimation of population mean and regression coefficients in conditional mean models of $Y$ given any sub-vector of $\boldsymbol{X}$ under both SSL and CSTL setups. Second, we propose a novel AIPW method that enables $\sqrt{N}$-consistent and asymptotically normal estimation and achieves valid confidence intervals under suitable sparsity conditions, when the PS model is correctly specified but the OR model may be misspecified. This robustness to model misspecification is achieved by carefully exploiting the connection between PS and OR models and designing estimating equations for nuisance parameters, differently from regularized least-squares or maximum-likelihood estimation. Previous related methods (Chakrabortty et al., 2019) achieve $\sqrt{N}$-consistency when both PS and OR models are correctly specified and the estimated PS and OR functions converge to the *true values* at fast enough rates (specifically, the product of estimation errors is smaller than $N^{-1/2}$). In contrast, our proposed estimator is shown to be $\sqrt{N}$-consistent even when the estimated OR function converges to a *target value* different from the true value and the estimated PS function converges to the true value, both slower than $N^{-1/2}$ (but faster than $N^{-1/4}$), with misspecified OR model and correctly specified PS model. Hence the aforementioned product of estimation errors may be greater than $N^{-1/2}$.

This work is also related to the causal inference problem under the strong ignorability assumption (Rosenbaum and Rubin, 1983). Specifically, the SSL problem is similar to estimating the average treatment effect (ATE) and the conditional ATE

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

(CATE) (Zimmert and Lechner, 2019; Fan et al., 2022; Wu et al., 2024). The CSTL problem can be viewed as an analog to the estimation of the average treatment effect on the treated (ATT).

The rest of this paper is organized as follows. In section 2, we present our setup and define the target parameters of interest. In Section 3, we construct a novel AIPW estimator for the target parameter. We show theoretical properties of the proposed estimator in Section 4, and compare them with the previous literature in Section 5. Numerical implementation is introduced in Section 6. An application to a crime study is presented in Section 7. Extension of proposed methods to the CSTL setting is given in Section 8 followed by concluding discussions in Section 9.

## 2. Setup and preliminaries

### 2.1 Data and target parameters

Let $Y \in \mathbb{R}$ be a response variable and $\boldsymbol{X} = (1, X_1, \ldots, X_d)^{\mathrm{T}} \in \mathbb{R}^{d+1}$ be a covariate vector with the first element being the constant 1. In addition, let $R \in \{0, 1\}$ be the indicator of whether $Y$ is observed: $R = 1$ if observed and $R = 0$ if missing. Assume that $\{(\boldsymbol{X}_i, Y_i, R_i) : i = 1, \ldots, N\}$ is an independent and identically distributed (i.i.d.) sample from a joint distribution of $(\boldsymbol{X}, Y, R)$, denoted as $\mathbb{P}$. The observed dataset, $\{(\boldsymbol{X}_i, R_i Y_i, R_i) : i = 1, \ldots, N\}$, can be split into a labeled dataset $\mathcal{L} = \{(\boldsymbol{X}_i, Y_i, R_i = 1),\ i = 1, \ldots, n\}$ and an unlabeled dataset $\mathcal{U} = \{(\boldsymbol{X}_i, R_i = 0), i =$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



$n+1,\ldots,N\}$. For $\boldsymbol{Z} \in \mathbb{R}^m$, a sub-vector of $\boldsymbol{X}$, it is of interest to fit a regression model for the conditional mean $\mathbb{E}(Y|\boldsymbol{Z})$:

$$\mathbb{E}(Y|\boldsymbol{Z}) = \psi(\beta^{*\mathrm{T}}\boldsymbol{Z}), \tag{2.1}$$

where $\mathbb{E}(\cdot)$ denotes the expectation under $\mathbb{P}$, $\beta^*$ is a parameter vector, and $\psi(\cdot)$ is an (increasing) inverse link function, such as the identity function $\psi(u) = u$ and logit function $\psi(u) = 1/\{1+\exp(-u)\}$. When $\boldsymbol{Z}$ is a strict sub-vector of $\boldsymbol{X}$, the parameter $\beta^*$ can be seen to capture the marginal (or full) effect of $\boldsymbol{Z}$ on $Y$, as any indirect effect of $\boldsymbol{Z}$ on $Y$ through other covariates in $\boldsymbol{X}$ is marginalized out conceptually. This is similar to marginal structural models (Robins, 1999; Wu et al., 2024). In comparison, the coefficient sub-vector associated with $\boldsymbol{Z}$ in the full regression of $Y$ on $\boldsymbol{X}$ can be seen to represent the direct effect of $\boldsymbol{Z}$ on $Y$ after accounting for the influence of other covariates in $\boldsymbol{X}$ on $Y$.

Model (2.1) is allowed to be misspecified, that is, $\mathbb{E}(Y|\boldsymbol{Z})$ may not be in the form $\psi(\beta^{\mathrm{T}}\boldsymbol{Z})$. With possible model misspecification, $\beta^*$ is defined as the solution to

$$\mathbb{E}\left[\{Y - \psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\,\boldsymbol{Z}\right] = 0. \tag{2.2}$$

For a generalized linear model with $\psi(\cdot)$ as the canonical inverse link, the estimating equation (2.2) leads to maximum likelihood estimation, so that $\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})$ can be

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



interpreted as the best likelihood-based approximation to $\mathbb{E}(Y|\boldsymbol{Z})$ using model (2.1).

The regression model (2.1) is flexible. The target parameter $\beta^*$ accommodates a variety of estimands in the previous literature.

(a) If we set $\boldsymbol{Z} = 1$ and $\psi(u) = u$, then $\beta^* = \mathbb{E}(Y)$. The problem corresponds to the semi-supervised estimation of the population mean (Zhang et al., 2019; Zhang and Bradic, 2021).

(b) If we set $\boldsymbol{Z}$ to be a univariate covariate, for example, $\boldsymbol{Z} = X_1$, then $\beta^*$ corresponds to the regression coefficient in the regression model of $Y$ given the particular covariate $X_1$. This problem was studied by Liu et al. (2023) and Wu et al. (2024).

(c) If we set $\boldsymbol{Z} = \boldsymbol{X}$, then $\beta^*$ corresponds to the coefficient vector in the regression model of $Y$ given the full covariate vector $\boldsymbol{X}$ (Chakrabortty, 2016; Chakrabortty and Cai, 2018; Chakrabortty et al., 2019; Deng et al., 2024; Zhang et al., 2023).

The description above is general, without specifying dependency of the dimensions of $\boldsymbol{X}$ and $\boldsymbol{Z}$ on $N$. Nevertheless, for our statistical theory (Section 4), we allow the dimension of $\boldsymbol{X}$ to increase as $N$ increases, but the dimension of $\boldsymbol{Z}$ is fixed, and we study inference about the entire vector $\beta^*$. Hence our work is distinct from Chakrabortty et al. (2019) in case (c) where $\boldsymbol{Z} = \boldsymbol{X}$ is high-dimensional and then the estimation objective is inferences about individual elements of $\beta^*$. See the Supplementary Material Section S2 for a detailed comparison of our work and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



several papers mentioned in case (c). On the other hand, for Chakrabortty (2016) and Chakrabortty and Cai (2018) in case (c), $\boldsymbol{Z} = \boldsymbol{X}$ is fixed-dimensional and inference about $\beta^*$ is studied by incorporating kernel smoothing in fitting the OR function. Hence such settings can be properly accommodated by our theory with fixed-dimensional $\boldsymbol{Z}$ but high-dimensional $\boldsymbol{X}$ as basis functions (denoted as $\boldsymbol{F}$ or $\boldsymbol{G}$ later) for OR fitting.

In addition to the parameter vector $\beta^*$, it may also be of interest to consider target parameters defined within unlabeled data corresponding to the CSTL setting, as studied in several recent papers (Liu et al., 2023; He et al., 2024). To incorporate this setting, we consider a regression model for the conditional mean in the unlabeled data: $\mathbb{E}(Y|R = 0, \boldsymbol{Z}) = \psi(\beta^{0*\mathrm{T}}\boldsymbol{Z})$. With possible model misspecification, $\beta^{0*}$ is defined as the solution to

$$\mathbb{E}\left[(1-R)\left\{Y - \psi(\beta^{\mathrm{T}}\boldsymbol{Z})\right\}\boldsymbol{Z}\right] = 0. \tag{2.3}$$

To illustrate the main ideas, we focus on the estimation of $\beta^*$, and defer the associated results for the estimation of $\beta^{0*}$ to Section 8.

## 2.2 General assumptions

Without imposing any assumption, we cannot obtain a consistent estimator of $\beta^*$ due to the missingness of $Y$ in the unlabeled data. Below, we introduce the identifiability

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



assumption.

**Assumption 1.** $Y \perp\!\!\!\perp R \mid \boldsymbol{X}$.

Assumption 1 is crucial for the identifiability of $\beta^*$ and $\beta^{0*}$. It ensures that $\mathbb{E}(Y|\boldsymbol{X}, R=1) = \mathbb{E}(Y|\boldsymbol{X}, R=0)$, indicating that the conditional mean of $Y$ is the same for both unlabeled and labeled data after accounting for the full covariates $\boldsymbol{X}$. This establishes the connection between the labeled and unlabeled data. Moreover, Assumption 1 implies that $\mathbb{P}(R=1|\boldsymbol{X}, Y) = \mathbb{P}(R=1|\boldsymbol{X})$, meaning that the label indicator $R$ depends solely on the covariates $\boldsymbol{X}$, i.e., $R$ is missing at random (Molenberghs et al., 2015; Imbens and Rubin, 2015). It should be noted that Assumption 1 does not imply $Y \perp\!\!\!\perp R \mid \boldsymbol{Z}$ when $\boldsymbol{Z} \neq \boldsymbol{X}$. If $Y \perp\!\!\!\perp R \mid \boldsymbol{Z}$ and the regression models are correctly specified, then $\beta^* = \beta^{0*}$. Otherwise, the equality may not hold.

Despite bearing many similarities, Assumption 1 differs from the classic SSL setup in the missing mechanism represented by the probabilistic behavior of $R$. SSL assumes the missing-completely-at-random mechanism (MCAR) (Chakrabortty, 2016; Chakrabortty and Cai, 2018; Zhang et al., 2019; Zhang and Bradic, 2021), that is, $R \perp\!\!\!\perp (\boldsymbol{X}, Y)$, thus $\mathbb{P}(R=1|\boldsymbol{X})$ is a constant, independent of $\boldsymbol{X}$. In contrast, we allow $R$ to probabilistically depend on $\boldsymbol{X}$. In addition, we make the following technical assumption.

**Assumption 2.** $\pi^*(\boldsymbol{X}) = \mathbb{P}(R=1|\boldsymbol{X}) > c$ almost surely, for some constant $0 < c < 1$ independent of $N$ and $d$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



This condition is introduced to ensure that each unit has a positive probability of belonging to $\mathcal{L}$. Then the labeled dataset $\mathcal{L}$ is of a non-negligible size compared with $N$. By Assumption 2, the ratio $n/N$ may randomly fluctuate but, as $N \to \infty$, converges to a value in the interval $(0, 1]$, equal to $\mathbb{E}(n/N) = \mathbb{P}(R = 1)$. This distinguishes our sampling process from the stratified sampling process widely used in the previous literature (Chakrabortty, 2016; Chakrabortty and Cai, 2018; Zhang et al., 2019; Zhang and Bradic, 2021), where the sizes of labeled and unlabeled datasets, $n$ and $N - n$, are deterministic. For the asymptotic analysis, they assume that both $n$ and $N$ tend to $\infty$ such that $n/N$ converges to a value in $[0, 1]$, including zero. See Section 5 for a detailed discussion.

### 2.3 AIPW estimating equations

For estimating $\beta^*$, we introduce the augmented inverse probability weighting (AIPW) estimating equations. Essentially the same AIPW estimating equation has been used in the previous literature, albeit under somewhat different settings than ours. See Section 5 for a connection and comparison of our method with the previous methods.

With the true PS model $\pi^*(\boldsymbol{X}) = \mathbb{P}(R = 1|\boldsymbol{X})$, under Assumption 1, we have $\mathbb{E}\left[\{R/\pi^*(\boldsymbol{X})\}\{Y - \psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\right] = \mathbb{E}\left[\{Y - \psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\right]$. Then a sample estimating equation for $\beta^*$ is $\tilde{\mathbb{E}}\left[\{R/\hat{\pi}(\boldsymbol{X})\}\{Y - \psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\right] = 0$, where $\hat{\pi}(\boldsymbol{X})$ is an estimator of $\pi^*(\boldsymbol{X})$ and $\tilde{\mathbb{E}}$ denotes the sample mean, defined as $\tilde{\mathbb{E}}(U) = N^{-1}\sum_{i=1}^{N} U_i$ for a variable $U$. Let $\hat{\beta}_{\mathrm{IPW}}$ be a solution to the previous estimating equation. If the PS model

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



is correctly specified, then under certain regularity conditions, $\hat{\pi}(\boldsymbol{X}) \xrightarrow{\mathbb{P}} \pi^*(\boldsymbol{X})$ and $\hat{\beta}_{\text{IPW}} \xrightarrow{\mathbb{P}} \beta^*$; If the PS model is misspecified, $\hat{\pi}(\boldsymbol{X}) \not\xrightarrow{\mathbb{P}} \pi^*(\boldsymbol{X})$ and $\hat{\beta}_{\text{IPW}} \not\xrightarrow{\mathbb{P}} \beta^*$. To mitigate the possible inconsistency of $\hat{\beta}_{\text{IPW}}$, the AIPW method introduces an augmented term. Specifically, let $m^*(X) = \mathbb{E}(Y|\boldsymbol{X})$ be the true OR function and $\hat{m}(\boldsymbol{X})$ be a corresponding estimator, the AIPW estimating equation is

$$\tilde{\mathbb{E}}\left[\frac{R}{\hat{\pi}(\boldsymbol{X})}\left\{Y-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\right\}\boldsymbol{Z}+\left\{1-\frac{R}{\hat{\pi}(\boldsymbol{X})}\right\}\left\{\hat{m}(\boldsymbol{X})-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\right\}\boldsymbol{Z}\right]=0. \tag{2.4}$$

Let $\hat{\beta}_{\text{AIPW}}$ be the solution to equation (2.4). If the PS model is misspecified, the augmented term $\tilde{\mathbb{E}}\left[\{1-R/\hat{\pi}(\boldsymbol{X})\}\left\{\hat{m}(\boldsymbol{X})-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\right\}\boldsymbol{Z}\right]$ corrects the bias of $\tilde{\mathbb{E}}\left[\{R/\hat{\pi}(\boldsymbol{X})\}\{Y-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\right]$ by introducing the estimator $\hat{m}(\boldsymbol{X})$. In addition, if the PS model is correctly specified, the augmented term improves the estimation efficiency of $\beta^*$ by leveraging the association between $\boldsymbol{X}$ and $Y$. It can be shown that the left side of equation (2.4) converges in probability to that of equation (2.2), if either $\hat{\pi}(\boldsymbol{X}) \xrightarrow{\mathbb{P}} \pi^*(\boldsymbol{X})$ or $\hat{m}(\boldsymbol{X}) \xrightarrow{\mathbb{P}} m^*(\boldsymbol{X})$, which is the property of double robustness.

In the classic SSL setup, estimating $\beta^*$ is considered to be an MCAR problem, where $\hat{\pi}(\boldsymbol{X})$ is a constant, independent of $\boldsymbol{X}$, and the estimator $\hat{m}(\boldsymbol{X})$ is usually defined using an OR model by (unweighted) least squares, maximum likelihood, or variations (Chakrabortty, 2016; Chakrabortty and Cai, 2018; Zhang et al., 2019). However, our semi-supervised regression is formulated as a MAR problem, where

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

$\hat{\pi}(\boldsymbol{X})$ depends on $\boldsymbol{X}$. In such a scenario, as shown in the next section, the estimators $\hat{\pi}(\boldsymbol{X})$ and $\hat{m}(\boldsymbol{X})$ for the PS and OR functions can be defined in a sequential manner, different from least squares or maximum likelihood, in order to obtain desirable properties with possible model misspecification.

## 3. Method

We develop a novel AIPW method that achieves $\sqrt{N}$-consistency in the setting of sparse high-dimensional PS and OR models, even if the estimation of the PS model exhibits convergence rates slower than $N^{-1/2}$ and the OR model is misspecified.

### 3.1 Model specification for nuisance parameters

AIPW estimation based on the estimating equation (2.4) requires constructing the estimators $\hat{\pi}(\boldsymbol{X})$ and $\hat{m}(\boldsymbol{X})$ for $\pi^*(\boldsymbol{X})$ and $m^*(\boldsymbol{X})$, using some PS and OR models. In contrast with the previous literature, we introduce a dependency between $\hat{\pi}(\boldsymbol{X})$ and $\hat{m}(\boldsymbol{X})$ by carefully specifying basis functions and incorporating weighted estimation.

Specifically, let $\boldsymbol{F}(\boldsymbol{X}) = \{1, f_1(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X})\}^{\mathrm{T}}$ be a vector of known functions of $\boldsymbol{X}$. We allow $p$ to be high-dimensional, tending to infinity as $N$ increases. As in Tan (2020a), we propose using logistic regression as a working model for the PS

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



function $\pi^*(\boldsymbol{X})$,

$$\mathbb{P}(R=1|\boldsymbol{X}) = \pi(\boldsymbol{X};\gamma) = [1+\exp\{-\gamma^{\mathrm{T}}\boldsymbol{F}(\boldsymbol{X})\}]^{-1}, \tag{3.1}$$

where $\gamma$ is an unknown coefficient parameter.

**Remark 1.** In several related works on classic SSL and stratified sampling setups, making efforts to estimate the PS may not be necessary. Firstly, in the classic SSL setup, the true PS is a constant, leading to a constant PS model. Secondly, in the stratified sampling setup, the proportion $n/N$ is fixed and known, which corresponds to a known PS function. Thus, researchers concentrate on specifying OR models.

Next, we turn to modeling the OR function $m^*(\boldsymbol{X})$. The working model for $m^*(\boldsymbol{X})$ is specified as

$$\mathbb{E}(Y|\boldsymbol{X}) = m(\boldsymbol{X};\alpha) = \psi\{\alpha^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{X})\}, \tag{3.2}$$

where $\boldsymbol{G}(\boldsymbol{X}) = \{1, g_1(\boldsymbol{X}), \ldots, g_q(\boldsymbol{X})\}^{\mathrm{T}}$ is a vector of known functions of $\boldsymbol{X}$ and $q$ can be high-dimensional. In contrast to the previous literature, to ensure valid inference even when the OR model is misspecified, we carefully specify a choice of $\boldsymbol{G}(\boldsymbol{X})$ as follows:

$$\boldsymbol{G}(\boldsymbol{X}) = [\boldsymbol{F}(\boldsymbol{X})^{\mathrm{T}}, \{\boldsymbol{Z}\otimes\boldsymbol{F}(\boldsymbol{X})\}^{\mathrm{T}}]^{\mathrm{T}}, \tag{3.3}$$

where $\boldsymbol{Z}\otimes\boldsymbol{F}(\boldsymbol{X})$ consists of all interactions between $\boldsymbol{Z}$ and $\boldsymbol{F}(\boldsymbol{X})$ (i.e., all prod-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



ucts of individual components from $\boldsymbol{Z}$ and $\boldsymbol{F}(\boldsymbol{X})$). Equation (3.3) represents the minimal choice for $\boldsymbol{G}(\boldsymbol{X})$, and additional covariates can also be incorporated, such as nonlinear terms of $\boldsymbol{Z}$ and $\boldsymbol{F}(\boldsymbol{X})$. Under sparsity conditions, these additional terms can be readily accommodated.

### 3.2 Estimation procedures

The proposed method consists of the following three steps: (a) estimating the parameter $\gamma$ in the PS model (3.1); (b) estimating the parameter $\alpha$ in the OR model (3.2); (c) estimating the target parameter $\beta$.

For estimating $\gamma$, we utilize a regularized calibrated estimator (Tan, 2020b), defined as

$$\hat{\gamma} = \underset{\gamma \in \mathbb{R}^{p+1}}{\operatorname{argmin}} L_{\mathrm{RCAL}}(\gamma) = \underset{\gamma \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{\ell_{\mathrm{CAL}}(\gamma) + \lambda_\gamma \|\gamma_{1:p}\|_1\}, \tag{3.4}$$

where $\ell_{\mathrm{CAL}}(\gamma) = \tilde{\mathbb{E}}[R \exp\{-\gamma^{\mathrm{T}} \boldsymbol{F}(\boldsymbol{X})\} + (1-R)\gamma^{\mathrm{T}} \boldsymbol{F}(\boldsymbol{X})]$, $\lambda_\gamma$ is a pre-specified tuning parameter, $||\cdot||_1$ denotes the $L_1$-norm, and for any vector $\nu$, $\nu_{i:j}$ is the sub-vector of $\nu$ consisting of its $i$–th to $j$–th elements (both ends included). For a possibly misspecified model $\pi(\boldsymbol{X}; \gamma)$, under suitable regularity conditions, $\hat{\gamma}$ converges in probability to its target value $\bar{\gamma}$ defined by $\bar{\gamma} = \operatorname{argmin}_{\gamma \in \mathbb{R}^{p+1}} \mathbb{E}[R \exp\{-\gamma^{\mathrm{T}} \boldsymbol{F}(\boldsymbol{X})\} + (1-R)\gamma^{\mathrm{T}} \boldsymbol{F}(\boldsymbol{X})]$.

For estimating $\alpha$, we adopt a regularized weighted maximum likelihood estima-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



tor (Tan, 2020a), defined as

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^{q+1}}{\operatorname{argmin}} L_{\mathrm{RWL}}(\alpha; \hat{\gamma}) = \underset{\alpha \in \mathbb{R}^{q+1}}{\operatorname{argmin}} \{\ell_{\mathrm{WL}}(\alpha; \hat{\gamma}) + \lambda_\alpha \|\alpha_{1:q}\|_1\}, \tag{3.5}$$

where $\ell_{\mathrm{WL}}(\alpha; \hat{\gamma}) = \tilde{\mathbb{E}}(Rw(\boldsymbol{X}; \hat{\gamma})[-Y\alpha^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{X}) + \Psi\{\alpha^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{X})\}])$, $w(X, \hat{\gamma}) = \{1 - \pi(\boldsymbol{X}, \hat{\gamma})\}/\pi(\boldsymbol{X}, \hat{\gamma})$, $\Psi(u) = \int_0^u \psi(t)dt$ is the antiderivative of $\psi$ and $\lambda_\alpha$ is a tuning parameter. Similar to the target value $\bar{\gamma}$, we define the target value of $\alpha$ as $\bar{\alpha} = \operatorname{argmin}_{\alpha \in \mathbb{R}^{q+1}} \mathbb{E}\left(Rw(\boldsymbol{X}; \bar{\gamma})\left[-Y\alpha^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{X}) + \Psi\{\alpha^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{X})\}\right]\right)$.

After obtaining the estimators of $\gamma$ and $\alpha$, the proposed calibrated AIPW estimator of $\beta$, denoted as $\hat{\beta}$, is the solution to

$$\tilde{\mathbb{E}}\{\tau(\boldsymbol{O}, \hat{\alpha}, \beta, \hat{\gamma})\} = 0, \tag{3.6}$$

where $\boldsymbol{O} = (\boldsymbol{X}, \boldsymbol{Z}, Y, R)$ and $\tau(\boldsymbol{O}, \alpha, \beta, \gamma) = \{R/\pi(\boldsymbol{X}; \gamma)\}\{Y - \psi(\alpha^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{Z} + \{\psi(\alpha^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}$. Our estimating equations (3.6) share the same form as the estimating equations (2.4). However, there is a crucial distinction between our method and previous related methods based on (2.4). In previous methods, the PS and OR models are specified and fitted independently of each other, typically both by least squares, maximum likelihood, or variations. In our method, the PS and OR models are specified and fitted in a sequentially dependent manner. This design allows our estimator $\hat{\beta}$ to achieve $\sqrt{N}$-consistency in the presence of misspecified OR models.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

## 4. Theoretical analysis

In this section, we present the theoretical analysis of the proposed estimator $\hat{\beta}$. In Section 4.1, we examine theoretical properties of estimators $\hat{\gamma}$ and $\hat{\alpha}$ in PS and OR models. Then, we study the asymptotic properties of the proposed estimator $\hat{\beta}$ in Section 4.2. Finally, in Section 4.3, we extend our analysis to the classic SSL setting (stratified sampling with constant PS).

### 4.1 Properties of the estimators for nuisance parameters

For simplicity, we denote $\boldsymbol{F}(\boldsymbol{X})$ and $\boldsymbol{G}(\boldsymbol{X})$ as $\boldsymbol{F}$ and $\boldsymbol{G}$, respectively. All the regularity Assumptions are given in the Supplementary Material. In addition, let $|\cdot|$ denote the cardinality of a set. We first present the properties of $\hat{\gamma}$ based on Tan (2020a) [Theorems 1 and 3].

**Proposition 1.** *Suppose that Assumption S1 in the Supplementary Material is satisfied, and $\lambda_\gamma$ in (3.4) is specified by $\lambda_\gamma = A_0\lambda_0$, where $A_0 > 1$ is a constant defined in Assumption S1. Then, with probability at least $1 - 8\epsilon$,*

$$D^{\dagger}_{\mathrm{CAL}}(\hat{\gamma}^{\mathrm{T}}\boldsymbol{F}, \bar{\gamma}^{\mathrm{T}}\boldsymbol{F}) + (A_0 - 1)\lambda_0\|\hat{\gamma} - \bar{\gamma}\|_1 \leq M_0|S_{\bar{\gamma}}|\lambda_0^2, \tag{4.1}$$

*where $M_0 > 0$ is a constant, and $D^{\dagger}_{\mathrm{CAL}}(\hat{\gamma}^{\mathrm{T}}F, \bar{\gamma}^{\mathrm{T}}F)$ is the symmetrized Bregman Divergence w.r.t $\ell_{\mathrm{CAL}}(\gamma)$, i.e., $D^{\dagger}_{\mathrm{CAL}}(\hat{\gamma}^{\mathrm{T}}\boldsymbol{F}, \bar{\gamma}^{\mathrm{T}}\boldsymbol{F}) = -\tilde{\mathbb{E}}[R\{\exp(-\hat{\gamma}^{\mathrm{T}}\boldsymbol{F}) - \exp(-\bar{\gamma}^{\mathrm{T}}\boldsymbol{F})\}$ $(\hat{\gamma}^{\mathrm{T}}\boldsymbol{F} - \bar{\gamma}^{\mathrm{T}}\boldsymbol{F})]$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



Note that $D^{\dagger}_{\mathrm{CAL}}(\hat{\gamma}^{\mathrm{T}}\boldsymbol{F}, \bar{\gamma}^{\mathrm{T}}\boldsymbol{F}) \geq 0$, then equation (4.1) implies that $\|\hat{\gamma} - \bar{\gamma}\|_1 \leq \{M_0/(A_0 - 1)\}|S_{\bar{\gamma}}|\lambda_0$, which indicates that the $L_1$-convergence rate of the proposed regularized calibrated estimator $\hat{\gamma}$ is $|S_{\bar{\gamma}}|\lambda_0$, where $|S_{\bar{\gamma}}|$ is the nonzero size of $\bar{\gamma}$ and $\lambda_0 = c_\gamma\sqrt{\ln\{(1+p)/\epsilon\}/N}$ for some constant $c_\gamma$. For example, taking $\epsilon = 1/(1+p)$ gives $\lambda_0 = c_\gamma\sqrt{2\ln(1+p)/N}$, which leads to $\|\hat{\gamma} - \bar{\gamma}\|_1 = O(|S_{\bar{\gamma}}|\sqrt{\ln(1+p)/N})$.

**Proposition 2.** *Suppose Assumptions S1 and S2 in the Supplementary Material are satisfied. If* $\ln\{(1+p)/\epsilon\}/N < 1$ *and* $\lambda_\alpha$ *in* (3.5) *is specified as* $A_1\lambda_1$*, where* $A_1 > 1$ *is a constant defined in Assumption S2. Then with probability at least* $1 - 10\epsilon$*,*

$$D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\gamma}) + \exp(\eta_{01})(A_1 - 1)\lambda_1\|\hat{\alpha} - \bar{\alpha}\|_1 \leq M_{11}|S_{\bar{\gamma}}|\lambda_0^2 + M_{12}|S_{\bar{\alpha}}|\lambda_1^2, \quad (4.2)$$

*where* $\eta_{01}$ *is a constant defined in Lemma S9,* $M_{11}$ *and* $M_{12}$ *are constants defined in Section S6.2 of the Supplementary Material;* $D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\gamma})$ *is the symmetrized Bregman divergence given by*

$$D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\gamma}) = \tilde{\mathbb{E}}\left[Rw(\boldsymbol{X}; \bar{\gamma})\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G} - \bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\right]. \quad (4.3)$$

Proposition 2 gives the convergence rate of $\hat{\alpha}$. Since $D^{\dagger}_{\mathrm{WL}}(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}, \bar{\gamma}) \geq 0$ and $\lambda_1 \geq \lambda_0$ (Assumption S2(vi)), $\|\hat{\alpha} - \bar{\alpha}\|_1 \leq c_m(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}}|)\lambda_1$ for some constant $c_m > 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



## 4.2 Large sample properties of the proposed estimator

In this subsection, we present the properties of the proposed estimator $\hat{\beta}$.

**Theorem 1.** Suppose Assumptions 1, 2, and S1 – S3 in the Supplementary Material are satisfied, and the PS model (3.1) is correctly specified with $\pi(\cdot;\bar{\gamma}) = \pi^*(\cdot)$. If $\ln\{(1+p\}/\epsilon)/N < 1$, then the following results hold.

(i) The estimator $\hat{\beta}$ is consistent and asymptotically normal, and $\sqrt{N}(\hat{\beta}-\beta^*) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma})$, where $\xrightarrow{d}$ denotes convergence in distribution, $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{-1}$ with $\boldsymbol{\Gamma} = \mathbb{E}\{\psi_1(\beta^{*\mathrm{T}}\boldsymbol{Z})\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\}$, and $\boldsymbol{\Lambda} = \mathbb{E}\{\tau(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})\tau(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})^{\mathrm{T}}\}$.

(ii) A consistent estimator of $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Gamma}}^{-1}$, where $\hat{\boldsymbol{\Gamma}} = \tilde{\mathbb{E}}\{\psi_1(\hat{\beta}^{\mathrm{T}}\boldsymbol{Z})\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\}$ and $\hat{\boldsymbol{\Lambda}} = \tilde{\mathbb{E}}\left\{\tau(\boldsymbol{O},\hat{\alpha},\hat{\beta},\hat{\gamma})\tau(\boldsymbol{O},\hat{\alpha},\hat{\beta},\hat{\gamma})^{\mathrm{T}}\right\}$. Thus, for a constant vector $\boldsymbol{c}$ with the same dimension of $\beta$, an asymptotic $(1-\eta)$ confidence interval for $\boldsymbol{c}^{\mathrm{T}}\beta^*$ is $\boldsymbol{c}^{\mathrm{T}}\hat{\beta} \pm z_{\eta/2}\sqrt{\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{c}/N}$, where $z_{\eta/2}$ is the $(1-\eta/2)$ quantile of the standard normal distribution.

Theorem 1 shows that if the PS model is correct, and $\alpha \to \bar{\alpha}$ at certain rate (faster than $N^{-1/4}$), regardless of whether the OR working model is correct, the proposed estimator $\hat{\beta}$ is $\sqrt{N}$–consistent and asymptotically normal, and the proposed CIs based on $\hat{\boldsymbol{\Sigma}}$ are valid. In Section S1 of the Supplementary Material, we provide a detailed discussion to explain how these properties are achieved and why, for a general choice of Z, the correct specification of the PS model is assumed, although either PS model or OR model is assumed to be correct in related work (Tan, 2020a).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



In contrast, the estimator $\hat{\beta}_{\mathrm{IPW}}$ is not $\sqrt{N}$-consistent in general, even with a correctly specified PS model, because in high-dimensional settings, the convergence rate of $\hat{\pi}(\boldsymbol{X})$ is typically slower than $N^{-1/2}$, leading to a slower convergence rate of $\hat{\beta}_{\mathrm{IPW}}$. Similarly, when the PS model is correctly specified but the OR model is misspecified, the convergence rate of the AIPW estimator with the PS and OR functions being estimated using conventional regularized maximum likelihood as in double machine learning (Chernozhukov et al., 2018) may also be slower than $N^{-1/2}$. Such double-machine learning estimators are only shown to achieve $\sqrt{N}$-consistency when both PS and OR models are correctly specified and the estimated PS and OR functions converge to the *true values* at fast enough rates (specifically, the product of estimation errors is smaller than $N^{-1/2}$).

### 4.3 Extension to stratified sampling with constant PS

To facilitate comparison with existing methods described in Section 5, we extend the theoretical analysis of the proposed estimator to the classic SSL setting (stratified sampling with constant PS), where the sizes of labeled and unlabeled datasets, $n$ and $N-n$, are deterministic. For fixed $n$ and $N$, the observed data are generated as follows:

- The labeled dataset $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \overset{\text{i.i.d.}}{\sim} \mathbb{P}(\boldsymbol{X}, Y | R = 1)$.
- The unlabeled dataset $(\boldsymbol{X}_{n+1}, \ldots, \boldsymbol{X}_N) \overset{\text{i.i.d.}}{\sim} \mathbb{P}(\boldsymbol{X}, Y | R = 0)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



Moreover, by letting $\boldsymbol{F} = 1$, $\hat{\pi}(\boldsymbol{X}) = \pi(\boldsymbol{X}; \hat{\gamma}) = n/N$ (constant PS) and allowing a general choice of $\boldsymbol{G}$ instead of (3.3), our AIPW estimator, denoted as $\hat{\beta}^s$, can be rewritten as a solution to the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)\}\,\boldsymbol{Z}_i + \frac{1}{N}\sum_{i=1}^{N}\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i) - \psi(\beta^{\mathrm{T}}\boldsymbol{Z}_i)\}\boldsymbol{Z}_i = 0. \tag{4.4}$$

Due to the constant $\hat{\pi}(\boldsymbol{X})$, our estimator $\hat{\alpha}$ reduces to the regularized unweighted maximum likelihood or least squares estimator. The following proposition for $\hat{\beta}^s$ can be readily derived.

**Proposition 3.** *Suppose that the conditions of Theorem 1 are satisfied with* $\boldsymbol{F} = 1$, $\pi^*(\boldsymbol{X}) \equiv n/N$, *and a general choice of* $\boldsymbol{G}$, *where Assumption S3(v) reduces to* $|S_{\bar{\alpha}}|\ln(q+1) = o_p(\sqrt{n})$. *Then we have* $\sqrt{n}(\hat{\beta}^s - \beta^*) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma}^{\boldsymbol{s}})$, *where* $\boldsymbol{\Sigma}^{\boldsymbol{s}} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}^s\boldsymbol{\Gamma}^{-1}$, *with* $\boldsymbol{\Gamma}$ *defined as in Theorem 1 and* $\boldsymbol{\Lambda}^s = \mathbb{E}([Y - \{(N-n)/N\}\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - (n/N)\psi(\beta^{*\mathrm{T}}\boldsymbol{Z})]^2\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}) + \{(N-n)n/N^2\}\mathbb{E}\left[\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}^2\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right]$.

For comparison, by letting $\boldsymbol{F} = 1$ and $\pi^*(\boldsymbol{X}) \equiv n/N$ in Theorem 1, the variance matrix $\boldsymbol{\Sigma}$ for $\hat{\beta}$ such that $\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma})$ is $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{-1}$, where $\boldsymbol{\Lambda} = \mathbb{E}\left([(N/n)\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}^2 + \{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}^2]\,\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right) + 2\mathbb{E}([\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}]\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})$. By direct calculation (see Section S8 of the Supplementary Material), we have $\boldsymbol{\Sigma}/N = \boldsymbol{\Sigma}^s/n$, which means the asymptotic variances of $\hat{\beta}^s$ and $\hat{\beta}$ are the same. Hence, in the classic SSL with constant PS, the asymptotic variances of our estimators, $\hat{\beta}$ under random sampling or $\hat{\beta}^s$ under stratified

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

sampling, are equivalent to each other.

## 5. Comparison with previous methods

We first summarize various previous methods, all of which can be integrated into the AIPW estimation framework. We also compare the asymptotic variances of our methods with previous ones. See the Supplementary Material Section S2 for a detailed comparison of our paper with several related papers with regression of $Y$ on high-dimensional $\boldsymbol{Z} = \boldsymbol{X}$ as mentioned in Section 2.1.

### 5.1 Unified framework

Various methods have been proposed in the classic SSL setting, i.e., stratified sampling with constant PS ($\boldsymbol{F} = 1$) as in Section 4.3 (Chakrabortty, 2016; Chakrabortty and Cai, 2018; Zhang et al., 2019; Zhang and Bradic, 2021). From an AIPW point of view, the major difference among previous methods lies in the choices of OR working models. For example, Zhang et al. (2019) and Zhang and Bradic (2021) proposed linear OR working models for the estimation of $\mathbb{E}(Y)$, i.e., with $\boldsymbol{Z} = 1$. Chakrabortty (2016) and Chakrabortty and Cai (2018) proposed using non-parametric or semi-parametric OR working models, such as kernel smoothing or partially linear model, for regression analysis with $\boldsymbol{Z}$ a sub-vector of $\boldsymbol{X}$.

If we disregard the specific choice of the OR working model, the previous methods can be incorporated into the AIPW estimating framework. In our notation, the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



previous estimators can be reformulated as solutions of

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)\}\boldsymbol{Z}_i+\frac{1}{N}\sum_{i=1}^{N}\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\beta^{\mathrm{T}}\boldsymbol{Z}_i)\}\boldsymbol{Z}_i=0, \tag{5.1}$$

for different choices of $\boldsymbol{Z}_i$ and $\psi(\cdot)$. Specifically, Zhang et al. (2019) and Zhang and Bradic (2021) correspond to the case of $\boldsymbol{Z}=1$ and $\psi(\cdot)$ is the identity function in (5.1), while Chakrabortty (2016) corresponds to the case where $\boldsymbol{Z}$ is any sub-vector of $\boldsymbol{X}$ and $\psi(\cdot)$ is an arbitrary inverse link function. Suppose a constant PS model is used with $\pi^*(\boldsymbol{X})\equiv n/N$, then the AIPW estimating equation (3.6) or, in the simplified form, (4.4) in Section 4.3, coincides with (5.1).

In addition, Chakrabortty and Cai (2018) adopted a variation of AIPW estimating equations. By the assumption $\lim_{n,N\to\infty} n/N\to 0$ and controlling kernel smoothing in fitting OR working models, they made it possible to drop the labeled part and only retain the augmented term of unlabeled data in (5.1). Their estimating equations can be reformulated in our notation as $\{1/(N-n)\}\sum_{i=n+1}^{N}\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\beta^{\mathrm{T}}\boldsymbol{Z}_i)\}\boldsymbol{Z}_i=0$, with $\psi(\cdot)$ to be identity function and $\boldsymbol{Z}=\boldsymbol{X}$, corresponding to full linear regression.

## 5.2 Variance comparison

Under stratified sampling with constant PS, both estimators of Zhang et al. (2019) and Zhang and Bradic (2021) of $\mathbb{E}(Y)$ achieve asymptotic normality and their

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



asymptotic variance is $\mathrm{Var}(Y - \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) + (n/N)\mathrm{Var}(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})$. Under this setting, by Proposition 3 with $\boldsymbol{Z} = 1$, our AIPW estimator has the asymptotic variance $\mathbb{E}\left\{(Y - \bar{\alpha}^{\mathrm{T}}\boldsymbol{G})^2 + (n/N)(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G} - \beta^*)^2 + 2(n/N)(Y - \bar{\alpha}^{\mathrm{T}}\boldsymbol{G})(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G} - \beta^*)\right\}$, where $\beta^* = \mathbb{E}(Y) = \mathbb{E}(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})$ and $\mathbb{E}\{(Y - \bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\boldsymbol{G}\} = 0$ by definition of $\bar{\alpha}$ and the fact that $\boldsymbol{G}$ includes 1. Then $\mathbb{E}\left\{(Y - \bar{\alpha}^{\mathrm{T}}\boldsymbol{G})(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G} - \beta^*)\right\} = \mathbb{E}\{(Y - \bar{\alpha}^{\mathrm{T}}\boldsymbol{G})(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\} = 0$, and our asymptotic variance reduces to $\mathrm{Var}(Y - \bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) + (n/N)\mathrm{Var}(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})$ matching results in Zhang et al. (2019) and Zhang and Bradic (2021).

Under stratified sampling with constant PS, the estimators of regression coefficients in conditional mean models proposed by Chakrabortty (2016) and Chakrabortty and Cai (2018) achieve asymptotic normality under the assumption that $\lim_{n,N\to\infty} n/N \to 0$. Their asymptotic variances satisfy

$$\boldsymbol{\Gamma}^{-1}\mathrm{Var}[\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{Z}]\boldsymbol{\Gamma}^{-1}. \tag{5.2}$$

In this setup, by Proposition 3, our estimator has the asymptotic variance $\boldsymbol{\Gamma}^{-1}\mathrm{Var}[\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\boldsymbol{Z}]\boldsymbol{\Gamma}^{-1} + (n/N)\boldsymbol{\Gamma}^{-1}\mathbb{E}[\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}^2\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}]\boldsymbol{\Gamma}^{-1} + 2(n/N)\boldsymbol{\Gamma}^{-1}\mathbb{E}[\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}]\boldsymbol{\Gamma}^{-1}$, which, compared with (5.2), in general has additional term $(n/N)\boldsymbol{\Gamma}^{-1}\Big\{\mathbb{E}([\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}^2 + 2\{Y - \psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G})\}\{\psi(\bar{\alpha}^{\mathrm{T}}\boldsymbol{G}) - \psi(\beta^{*\mathrm{T}}\boldsymbol{Z})\}]\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})\Big\}\boldsymbol{\Gamma}^{-1}$.

The additional term reduces to 0 under the condition $\lim_{n,N\to\infty} n/N \to 0$, implying that our result aligns with those of Chakrabortty (2016) and Chakrabortty

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

and Cai (2018) with the same condition.

## 6. Numerical implementation and simulation

We design experiments to evaluate the finite-sample performance of the proposed method and compare it with several alternative methods: the IPW method with Lasso-regularized maximum likelihood estimation for the PS model, AIPW methods with Lasso-regularized maximum likelihood estimation for both the PS and OR models without cross-fitting as in Tan (2020a), and AIPW methods with cross-fitting and Lasso-regularized maximum likelihood estimation for both the PS and OR models (Chernozhukov et al., 2018; Zhang and Bradic, 2021). These competing estimators are denoted as IPW, $\text{AIPW}_{\text{RML}}$, and $\text{AIPW}_{\text{CF}}$, respectively. For the PS and OR models, the basis functions $\boldsymbol{F}$ and $\boldsymbol{G}$ are specified as follows.

- $\text{AIPW}_{\text{RCAL}}$: Given $\boldsymbol{X} = (1, X_1, \ldots, X_d)^{\mathrm{T}}$, let $\{\xi_i\}_{i=1}^{k}$ be $k$ points equally spaced within $(-a, a)$, where $d = 3$, $k = 49$, and $a = 3$. Let $f_{ij}(\boldsymbol{X}) = (X_i - \xi_j)_+$, $i = 1, \ldots, d, j = 1, \ldots, k$. Let $\boldsymbol{F} = \{1, f_{11}(\boldsymbol{X}), \ldots, f_{1n_k}(\boldsymbol{X}), \ldots, f_{d1}(\boldsymbol{X}), \ldots, f_{dn_k}(\boldsymbol{X})\}^{\mathrm{T}}$ be basis functions in the PS model, and $\boldsymbol{G} = \{\boldsymbol{F}^{\mathrm{T}}, (\boldsymbol{Z} \bigotimes \boldsymbol{F})^{\mathrm{T}}\}^{\mathrm{T}}$ be basis functions in the OR model. The dimension of $\boldsymbol{F}$ is 148. For $\boldsymbol{Z} = 1,\ X_1\ , \boldsymbol{X}$, the dimensions of $\boldsymbol{G}$ are 148, 285 and 589, respectively.
- IPW: Let $\boldsymbol{F}$ be the basis functions for the PS model.
- $\text{AIPW}_{\text{RML}}$: Let $\boldsymbol{F}$ and $\boldsymbol{G} = \boldsymbol{F}$ be the basis functions for both PS and OR models, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

- AIPW$_{\text{CF}}$: Let $\boldsymbol{F}$ and $\boldsymbol{G} = \boldsymbol{F}$ be the basis functions for both PS and OR models, respectively.

We consider the estimators of population mean for $\boldsymbol{Z} = 1$, regression coefficients in the mean model for $\boldsymbol{Z} = X_1$ and $\boldsymbol{Z} = \boldsymbol{X}$, respectively. The data generating mechanisms and the associated numerical results are presented in Section S9 of the Supplementary Material. We evaluate methods with five metrics: Bias (Monte Carlo bias), $\sqrt{\text{Var}}$ (Monte Carlo standard deviation), $\sqrt{\text{EVar}}$ (square root of the mean of variance estimates), CP90 (coverage proportions of the 90% CIs), and CP95 (coverage proportions of the 95% CIs). The simulation results demonstrate that the proposed method AIPW$_{\text{RCAL}}$ has the smallest $\sqrt{\text{Var}}$ and $\sqrt{\text{EVar}}$, and Bias. Moreover, CP90 and CP95 of the proposed method are more aligned with their nominal values of 0.90 and 0.95, respectively. This indicates the effectiveness of the proposed method in terms of estimating both the population mean and regression coefficients. See Section S9.3 of the Supplementary Material for more details.

## 7. Application

### 7.1 Data description

The Communities and Crime dataset comprises 1994 records of crime-related information from communities in the USA, which combine socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



data from the 1995 FBI UCR. Each record includes a response `ViolentCrimesPerPop` (total number of violent crimes per 100,000 population) and 127 covariates, encompassing both location information, such as `state` as well as `county`, and socio-economic factors, such as `PctTeen2Par` (percent of kids age 12-17 in two parent households), `HousVacant` (number of vacant households), `PctHousNoPhone` (percent of occupied housing units without phone) and `PopDens` (population density in persons per square mile). In this study, we are interested in examining the influence of univariate covariates on the response. We consider the case where $\boldsymbol{Z} = (1, X_i)^{\mathrm{T}}$ for a particular univariate covariate $X_i$ as discussed in Section 2.1 and denote $\beta^* = (\beta_0, \beta_1)$.

Due to the presence of numerous missing values in high-dimensional covariates, we eliminate covariates with high missing ratios. See details of the pre-rocessing procedure in Section S10.1 of the Supplementary Material. After pre-processing, the analytical dataset consists of 1993 observations and 26 covariates (i.e., $d = 26$). The shift in covariates $\boldsymbol{X}$ is naturally introduced by the different states where the communities are located. We set label indicator $R$ for communities in New Jersey (Code 34) to be 1 and that for communities in other states to be 0 and remove the associated response data if $R = 0$, resulting in 211 labeled observations and 1782 unlabeled observations. The covariate shift of the joint distribution of $\boldsymbol{X}$ was confirmed to exist using a Gaussian kernel two-sample test with maximum mean discrepancy (You, 2023). Additionally, we assess the shift of each individual

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



covariate by a bootstrap version of the Kolmogorov–Smirnov test (Sekhon, 2011). For results of those tests, please see Section S10.2 of the Supplementary Material.

We randomly take 90% of labeled data and 90% of unlabeled data to form the training set with the remaining data used for the testing set. From the remaining 26 covariates, we select four representative ones: `PctTeen2Par`, `HousVacant`, `PctHousNoPhone` and `PopDens`, which illustrate different aspects of the socio-economic characteristics of communities. Notice that the covariate shift exists in all four covariates.

We compare the proposed method with IPW, $\text{AIPW}_{\text{RML}}$ and $\text{AIPW}_{\text{CF}}$ methods with piecewise linear basis functions introduced in Section 6. The PS and OR working models are estimated the same way as described in Section S9.2 of the Supplementary Material. For details of the procedures for designing basis functions, please see Section S10.3 of the Supplementary Material.

## 7.2 Results

Table 1 presents the estimates of the regression coefficients $\hat{\beta}_1$ along with the prediction mean squared error (MSE), which are calculated using the test data. It reveals that the point estimates of the regression coefficient are similar across the different methods. Notably, our estimators achieve the lowest prediction MSE except `PctTeen2Par`, highlighting the superior performance of our methods in minimizing predictive errors.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



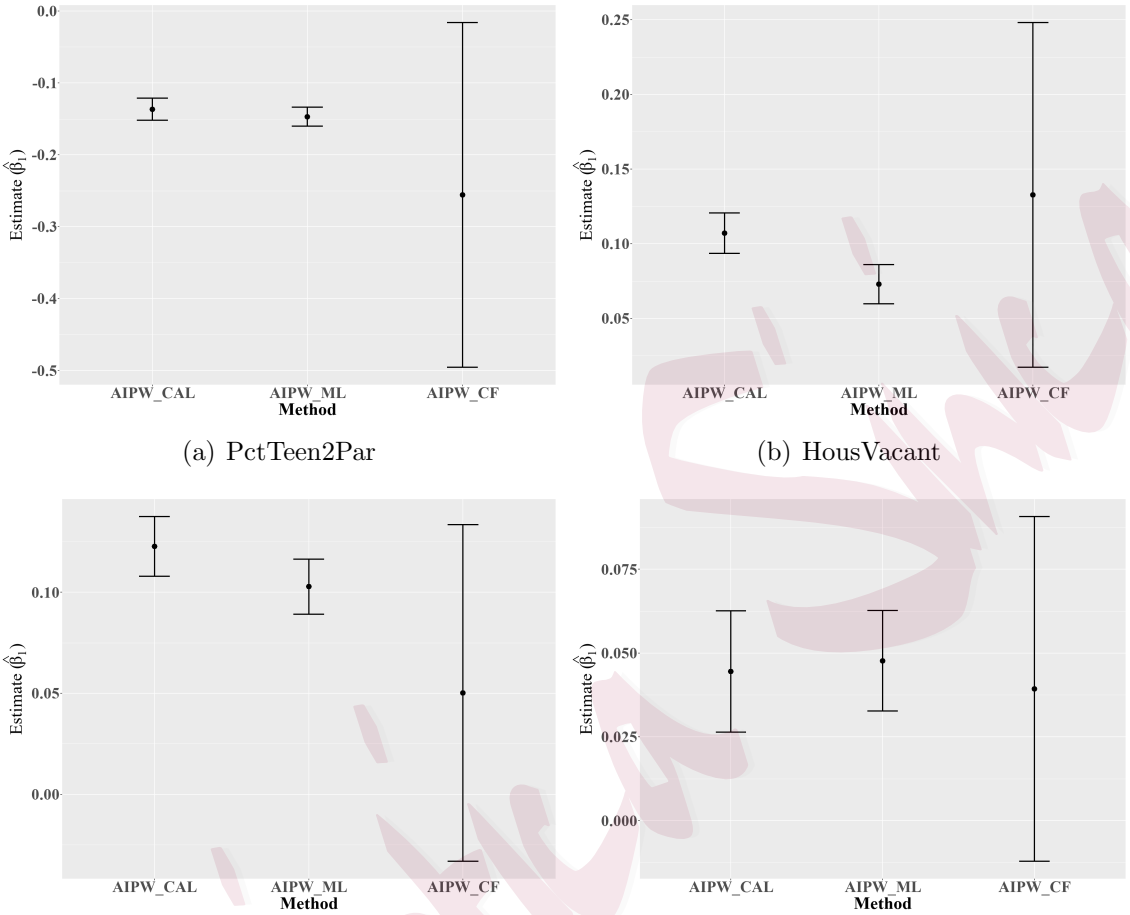
Table 1: Summary of $\hat{\beta}_1$ and prediction MSE

| | $\hat{\beta}_1$ | | | | prediction MSE | | | |
|---|---|---|---|---|---|---|---|---|
| | $\text{AIPW}_{\text{RCAL}}$ | IPW | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ | $\text{AIPW}_{\text{RCAL}}$ | IPW | $\text{AIPW}_{\text{RML}}$ | $\text{AIPW}_{\text{CF}}$ |
| `PctTeen2Par` | -0.137 | -0.172 | -0.147 | -0.256 | 0.034 | 0.032 | 0.034 | 0.044 |
| `HousVacant` | 0.107 | 0.261 | 0.073 | 0.133 | 0.046 | 0.085 | 0.046 | 0.047 |
| `PctHousNoPhone` | 0.123 | 0.245 | 0.103 | 0.050 | 0.036 | 0.058 | 0.039 | 0.047 |
| `PopDens` | 0.045 | 0.069 | 0.048 | 0.039 | 0.050 | 0.052 | 0.052 | 0.055 |

Moreover, signs of estimates of coefficients are the same among different methods for each covariate $Z$ of interest. and they coincide with common sense and previous studies. For example, the coefficients of `PctTeen2Par` are negative, since it is believed to have protective effects in assaults (Luo and Qi, 2017); the coefficients of `HousVacan` is positive, and criminological theories predict a positive association between vacancy and crime since empty structures of houses could provide locations for some crimes (e.g., prostitution, drug dealing), and the absence of residents may prevent social organization and reduce guardianship (Roth, 2019). Moreover, $\text{AIPW}_{\text{RML}}$ and ours are close in all cases, while IPW estimators and $\text{AIPW}_{\text{CF}}$ estimators are far from others in some cases.

In Figure 1, we compare the 95% CIs of $\text{AIPW}_{\text{RCAL}}$, $\text{AIPW}_{\text{RML}}$ and $\text{AIPW}_{\text{CF}}$. From the CIs, we see that for $\text{AIPW}_{\text{RML}}$ and our estimators all four single effects are significant. CIs of our estimators and of $\text{AIPW}_{\text{RML}}$'s have similar lengths and are overlapped, except `HousVacant`. The reason of the small difference is that the estimates of $\beta_0$ are a bit different. CIs of $\text{AIPW}_{\text{CF}}$ are much longer; for `PctHousNoPhone` and `PopDens`, the estimates are not significant. Both phenomenons show that $\text{AIPW}_{\text{CF}}$ is not as efficient as other two methods.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261




(a) PctTeen2Par (b) HousVacant

(c) PctHousNoPhone (d) PopDens

Figure 1: Comparison of 95% CIs of $\text{AIPW}_{\text{RCAL}}$, $\text{AIPW}_{\text{RML}}$ and $\text{AIPW}_{\text{CF}}$

## 8. Extension to estimation of $\beta^{0*}$

Consider the estimation of $\beta^{0*}$, defined as a solution to estimating equations (2.3). Under Assumption 1, $\mathbb{E}\left[R\{1-\pi^*(\boldsymbol{X})\}/\pi^*(\boldsymbol{X})\left\{Y-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\right\}\boldsymbol{Z}\right] = \mathbb{E}\left[\{1-R\}\right.$ $\left.\{Y-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\,\boldsymbol{Z}\right]$. Then a natural sample estimating equation for $\beta^{0*}$ is $\tilde{\mathbb{E}}[R\{1-\hat{\pi}(\boldsymbol{X})\}/\hat{\pi}(\boldsymbol{X})\{Y-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}] = 0$. We augment the estimating equations similarly

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

as described in Section 2.3 and obtain the sample AIPW estimating equations:

$$\tilde{\mathbb{E}}\left[\frac{R\{1-\hat{\pi}(\boldsymbol{X})\}}{\hat{\pi}(\boldsymbol{X})}\{Y-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}+\left\{1-\frac{R}{\hat{\pi}(\boldsymbol{X})}\right\}\{\hat{m}(\boldsymbol{X})-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}\boldsymbol{Z}\right]=0. \tag{8.1}$$

For the PS and OR models, we adopt a similar construction as in Section 3. Our AIPW estimator for $\beta^{0*}$, $\hat{\beta}^0$, is defined as the solution to the estimating equations $\tilde{\mathbb{E}}\{\tau^0(\boldsymbol{O},\hat{\alpha},\beta,\hat{\gamma})\}=0$, where $\tau^0(\boldsymbol{O},\alpha,\beta,\gamma)=([R\{1-\pi(\boldsymbol{X};\gamma)\}/\pi(\boldsymbol{X};\gamma)]\{Y-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\}+[\{\pi(\boldsymbol{X};\gamma)-R\}/\pi(\boldsymbol{X};\gamma)]\{\psi(\alpha^{\mathrm{T}}\boldsymbol{G})-\psi(\beta^{\mathrm{T}}\boldsymbol{Z})\})\boldsymbol{Z}$. The asymptotic properties of $\beta^{0*}$ is given in Theorem 2.

**Theorem 2.** Under Assumptions 1–2 and S1 – S3 in the Supplementary Material, if the PS model (3.1) is correctly specified with $\pi(\cdot;\bar{\gamma})=\pi^*(\cdot)$, and $\ln\{(1+p)/\epsilon\}/N<1$, we have that

(i) the estimator $\hat{\beta}^0$ is consistent and asymptotically normal, and $\sqrt{N}(\hat{\beta}^0-\beta^{0*})\xrightarrow{d}\mathrm{N}(0,\boldsymbol{\Sigma^0})$, where $\boldsymbol{\Sigma^0}=\boldsymbol{\Gamma}^{0-1}\boldsymbol{\Lambda}^0\boldsymbol{\Gamma}^{0-1}$ with $\boldsymbol{\Gamma}^0=\mathbb{E}[\{1-\pi(\boldsymbol{X};\bar{\gamma})\}\psi_1(\beta^{0*\mathrm{T}}\boldsymbol{Z})\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}]$ and $\boldsymbol{\Lambda}^0=\mathbb{E}\{\tau^0(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})\tau^0(\boldsymbol{O},\bar{\alpha},\beta^*,\bar{\gamma})^{\mathrm{T}}\}$.

(ii) a consistent estimator of $\boldsymbol{\Sigma^0}$ is $\hat{\boldsymbol{\Sigma}}^0=\hat{\boldsymbol{\Gamma}}^{0-1}\hat{\boldsymbol{\Lambda}}^0\hat{\boldsymbol{\Gamma}}^{0-1}$, where $\hat{\boldsymbol{\Gamma}}^0=\tilde{\mathbb{E}}\{\psi_1(\hat{\beta}^{0\mathrm{T}}\boldsymbol{Z})\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\}$ and $\hat{\boldsymbol{\Lambda}}^0=\tilde{\mathbb{E}}\left\{\tau^0(\boldsymbol{O},\hat{\alpha},\hat{\beta}^{0\mathrm{T}},\hat{\gamma})\tau^0(\boldsymbol{O},\hat{\alpha},\hat{\beta}^{0\mathrm{T}},\hat{\gamma})^{\mathrm{T}}\right\}$. Thus, for a constant vector $\boldsymbol{c}$ with the same dimension of $\beta$, an asymptotic $(1-\eta)$ confidence interval for $\boldsymbol{c}^{\mathrm{T}}\beta^{0*}$ is $\boldsymbol{c}^{\mathrm{T}}\hat{\beta}^{0\mathrm{T}}\pm z_{\eta/2}\sqrt{\boldsymbol{c}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{\boldsymbol{0}}\boldsymbol{c}/N}$.

Theorem 2 shows that if the PS model is correct, regardless of the correctness of

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

OR working models, the proposed estimator $\hat{\beta}^0$ is consistent, asymptotically normal, and the proposed CIs based on $\hat{\mathbf{\Sigma}}^0$ are valid. Similarly to the estimation of $\beta^*$, conclusions in Theorem 2 also hold in low-dimensional settings with a reduced form of Assumptions S1 to S3.

We point out that the method of Liu et al. (2023) for CSTL can be viewed as an AIPW estimator of $\beta^{0*}$ under the stratified sampling setting, where the labeled and unlabeled datasets $\mathcal{L}$ and $\mathcal{U}$ are treated as two independent samples of fixed sizes $n$ and $N-n$. They employed partial linear models for both PS and OR working models. By replacing their choices of semi-parametric nuisance models with our parametric models, the estimator of $\beta^{0*}$ in Liu et al. (2023) can be reformulated as the solution to the following estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n} w(\boldsymbol{X}_i;\hat{\gamma}^s)\left[\{Y_i-\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)\}\boldsymbol{Z}_i\right]+\frac{1}{N-n}\sum_{i=n+1}^{N}\left[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)-\psi(\beta^{\mathrm{T}}\boldsymbol{Z}_i)\}\,\boldsymbol{Z}_i\right]=0. \tag{8.2}$$

where $w(\boldsymbol{X}_i;\hat{\gamma}^s)=\exp(-\hat{\gamma}^{s\mathrm{T}}\boldsymbol{F}_i)$ and $\boldsymbol{F}_i$ is the abbreviation of $\boldsymbol{F}(\boldsymbol{X}_i)$; $\hat{\gamma}^s=(\hat{\gamma}^s_0,\hat{\gamma}^{s\mathrm{T}}_{1:p})^{\mathrm{T}}$ is an estimator of the parameter $\gamma^s$ in an exponential tilt model, defined as

$$\mathrm{d}G_1=\exp(\gamma^s_0+\gamma^{s\mathrm{T}}_{1:p}\boldsymbol{F}_{1:p})\mathrm{d}G_0, \tag{8.3}$$

where $G_0$ and $G_1$ are two probability distributions for the unlabeled and labeled data in $\boldsymbol{F}_{1:p}$ and $\gamma^s_0=-\ln\left\{\int\exp(\gamma^{s\mathrm{T}}_{1:p}\boldsymbol{F}_{1:p})\mathrm{d}G_0\right\}$ to ensure that $\int\mathrm{d}G_1=1$. The

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

exponential tilt model (8.3) can be shown to be equivalent to the logistic PS model (3.1), where the coefficients are related as follows (Prentice and Pyke, 1979; Qin, 1998; Tian et al., 2026):

$$\gamma_0 = \gamma_0^s + \ln\left(\frac{\rho_m}{1-\rho_m}\right), \quad \gamma_{1:p} = \gamma_{1:p}^s, \tag{8.4}$$

where $\rho_m = \mathbb{P}(R=1)$, the true value of the proportion of missing data. When analyzing the asymptotic property in stratified sampling settings, we assume $n/N$ to be constant and, consequently, assume that $\rho_m = n/N$. On the other hand, our estimating equations (8.1) can be rewritten as

$$\frac{1}{N}\sum_{i=1}^{n} w(\boldsymbol{X}_i;\hat{\gamma})\left[\{Y_i - \psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i)\}\boldsymbol{Z}_i\right] + \frac{1}{N}\sum_{i=n+1}^{N}\left[\{\psi(\hat{\alpha}^{\mathrm{T}}\boldsymbol{G}_i) - \psi(\beta^{\mathrm{T}}\boldsymbol{Z}_i)\}\,\boldsymbol{Z}_i\right] = 0, \tag{8.5}$$

where $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_{1:p})^{\mathrm{T}}$ is an estimator of the parameter $\gamma$ in logistic PS model (3.1). If the estimators $\hat{\gamma}^s$ and $\hat{\gamma}$ satisfy the same relationship as (8.4), i.e., $\hat{\gamma}_0 = \hat{\gamma}_0^s + \ln\{n/(N-n)\}$ and $\hat{\gamma}_{1:p} = \hat{\gamma}_{1:p}^s$, then one can see that $\{(N-n)/n\}w(\boldsymbol{X}_i;\hat{\gamma}^s) = \exp(-\hat{\gamma}^{\mathrm{T}}\boldsymbol{F}_i) = w(\boldsymbol{X}_i;\hat{\gamma})$, and the two equations (8.2) and (8.5) match each other. Thus, the different forms of (8.2) and (8.5) can be explained by the relationship of the coefficient estimates between the exponential tilt model (8.3) and the logistic regression model (3.1).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261

## 9. Discussion

We present a new AIPW method for the inference of regression coefficients in (conditional) mean models in SSL and CSTL settings. We demonstrate that various previous methods can be unified in our AIPW framework by suppressing detailed technical choices. Our AIPW estimator achieves asymptotic normality, and valid CIs can be obtained, whether or not the OR working model is correctly specified, with high-dimensional data. Finite sample performances of the proposed method are confirmed by a simulation study and an application to a real-world dataset.

Currently, the proposed CIs can only achieve single robustness to the misspecification of the OR model. Doubly robust CIs can be developed using the approach of Ghosh and Tan (2022), albeit at the cost of increasing technical and numerical complexities. In addition, how to handle the case where $\lim_{n,N\to\infty} n/N \to 0$ under the random sampling process is also technically challenging, since the "positivity assumption" (Assumption 2) typical in missing data theory is violated. New analysis needs to be developed to address the problem.

## Supplementary Material

The online Supplementary Material contains a heuristic discussion on conditions for the proposed estimator to be $\sqrt{N}$-consistent and asymptotic normal, a comparison of our paper with several related papers with regression of $Y$ on high-dimensional

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



$\boldsymbol{Z} = \boldsymbol{X}$ and papers under stratified sampling settings, detailed proofs of theorems as well as propositions, and details of the numerical implementation and application.

## Acknowledgments

The authors thank the assistant editor and the anonymous reviewers for their helpful comments and valuable suggestions. Ye Tian conducted this research while at Rutgers University and is now affiliated with Northeast Normal University. Peng Wu was supported by the National Natural Science Foundation of China (No. 12301370), the funding from the Beijing Municipal Education Commission for the Emerging Interdisciplinary Platform for Digital Business at Beijing Technology and Business University, and the Beijing Key Laboratory of Applied Statistics and Digital Regulation.

## References

Aloui, A., J. Dong, C. P. Le, and V. Tarokh (2023). Transfer learning for individual treatment effect estimation. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 56–66.

Alvari, H., E. Shaabani, S. Sarkar, G. Beigi, and P. Shakarian (2019). Less is more: Semi-supervised causal inference for detecting pathogenic users in social media. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 154–161.

Cai, T. T. and Z. Guo (2020). Semisupervised inference for explained variance in high dimensional linear

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



regression and its applications. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*, 391–419.

Castro, D. C., I. Walker, and B. Glocker (2020). Causality matters in medical imaging. *Nature Communications 11*, 3673.

Chakrabortty, A. (2016). *Robust Semi-Parametric Inference in Semi-Supervised Settings*. Ph. D. thesis, Harvard University.

Chakrabortty, A. and T. Cai (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics 46*, 1541–1572.

Chakrabortty, A., G. Dai, and R. J. Carroll (2022). Semi-supervised quantile estimation: Robust and efficient inference in high dimensional settings. *arXiv preprint arXiv:2201.10208*.

Chakrabortty, A., J. Lu, T. T. Cai, and H. Li (2019). High dimensional m-estimation with missing outcomes: A semi-parametric framework. *arXiv preprint arXiv:1911.11345*.

Chapelle, O., B. Schölkopf, and A. Zien (2006). *Semi-Supervised Learning*. The MIT Press.

Chen, B. and F. Huang (2016). Semi-supervised convolutional networks for translation adaptation with tiny amount of in-domain data. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 314–323.

Chen, K. and Y. Zhang (2023). Enhancing efficiency and robustness in high-dimensional linear regression with additional unlabeled data. *arXiv preprint arXiv:2311.17685*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Jour-*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



*nal 21*, C1–C68.

Deng, S., Y. Ning, J. Zhao, and H. Zhang (2024). Optimal and safe estimation for high-dimensional semi-supervised learning. *Journal of the American Statistical Association 119*, 2748–2759.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics 40*, 313–327.

Ghosh, S. and Z. Tan (2022). Doubly robust semiparametric inference using regularized calibrated estimation with high-dimensional data. *Bernoulli 28*, 1675–1703.

Gronsbell, J. L. and T. Cai (2017). Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*, 579–594.

He, Z., Y. Sun, and R. Li (2024). Transfusion: Covariate-shift robust transfer learning for high-dimensional regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 703–711.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics Social and Biomedical Science*. Cambridge University Press.

Little, R. J. and D. B. Rubin (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Liu, M., Y. Zhang, K. P. Liao, and T. Cai (2023). Augmented transfer regression learning with semi-nonparametric nuisance models. *Journal of Machine Learning Research 24*, 1–50.

Luo, R. and X. Qi (2017). Signal extraction approach for sparse multivariate response regression. *Journal of Multivariate Analysis 153*, 83–97.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke (2015). *Handbook of Missing Data Methodology*. Chapman & Hall/CRC.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



Prentice, R. L. and R. Pyke (1979). Logistic disease incidence models and case-control studies. *Biometrika 66*, 403–411.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika 85*, 619–630.

Quiñonero-Candela, J., M. Sugiyama, A. Schwaighofer, and N. D. Lawrence (2009). *Dataset Shift in Machine Learning*. The MIT Press.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 95–133. Springer New York.

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

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

Roth, J. J. (2019). Empty homes and acquisitive crime: Does vacancy type matter? *American Journal of Criminal Justice 44*, 770–787.

Ruder, S., M. E. Peters, S. Swayamdipta, and T. Wolf (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15–18.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *Journal of Statistical Software 42*, 1–52.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



Sohn, K., D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, pp. 596–608.

Tan, Z. (2020a). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics 48*, 811–837.

Tan, Z. (2020b). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika 107*, 137–158.

Tang, C., X. Zeng, L. Zhou, Q. Zhou, P. Wang, X. Wu, H. Ren, J. Zhou, and Y. Wang (2024). Semi-supervised medical image segmentation via hard positives oriented contrastive learning. *Pattern Recognition 146*, 110020.

Tian, Y., X. Zhang, and Z. Tan (2026). On semi-supervised estimation using exponential tilt mixture models. *Journal of Statistical Planning and Inference 241*, 106314.

Wu, P., Z. Tan, W. Hu, and X. Zhou (2024). Model-assisted inference for covariate-specific treatment effects with high-dimensional data. *Statistica Sinica 34*, 459–479.

You, K. (2023). *maotai: Tools for Matrix Algebra, Optimization and Inference*. R package version 0.2.5.

Zhang, A., L. D. Brown, and T. T. Cai (2019). Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics 47*, 2538 – 2566.

Zhang, Y. and J. Bradic (2021). High-dimensional semi-supervised learning: In search of optimal inference of the mean. *Biometrika 109*, 387–403.

Zhang, Y., A. Chakrabortty, and J. Bradic (2023). Semi-supervised causal inference: Generalizable and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



double robust inference for average treatment effects under selection bias with decaying overlap. *arXiv preprint arXiv:2305.12789*.

Zhang, Y., A. Giessing, and Y.-C. Chen (2023). Efficient inference on high-dimensional linear models with missing outcomes. *arXiv preprint arXiv:2309.06429*.

Zhao, Y., Y. Zheng, B. Yu, Z. Tian, D. Lee, J. Sun, Y. Li, and N. L. Zhang (2022). Semi-supervised lifelong language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3937–3951.

Zheng, M., S. You, L. Huang, F. Wang, C. Qian, and C. Xu (2022). Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14481.

Zhou, A. and S. Levine (2021). Bayesian adaptation for covariate shift. In *Advances in Neural Information Processing Systems*, pp. 914–927.

Zhu, X. (2008). Semi-supervised learning literature survey. Technical Report No. 1530, Department of Computer Sciences, University of Wisconsin-Madison, USA.

Zimmert, M. and M. Lechner (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*.

Ye Tian

Address: School of Mathematics and Statistics, Northeast Normal University, Jilin 130024, China

E-mail: tianye@nenu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0261



Peng Wu

Address: School of Mathematics and Statistics, Beijing Technology and Business University, Beijing 100048, China

E-mail: pengwu@btbu.edu.cn

Zhiqiang Tan

Address: Department of Statistics, Rutgers University New-Brunswick, NJ 08854, USA

E-mail: ztan@stat.rutgers.edu