Doubly Robust Inference when Combining Probability and Non-probability Samples with High-dimensional Data

Shu Yang[1], Jae Kwang Kim[2], and Rui Song[1]
[1]Department of Statistics, North Carolina State University
[2]Department of Statistics, Iowa State University

Abstract

Non-probability samples become increasingly popular in survey statistics but may suffer from selection biases that limit the generalizability of results to the target population. We consider integrating a non-probability sample with a probability sample which provides high-dimensional representative covariate information of the target population. We propose a two-step approach for variable selection and finite population inference. In the first step, we use penalized estimating equations with folded-concave penalties to select important variables for the sampling score of selection into the non-probability sample and the outcome model. We show that the penalized estimating equation approach enjoys the selection consistency property for general probability samples. The major technical hurdle is due to the possible dependence of the sample under the finite population framework. To overcome this challenge, we construct martingales which enable us to apply Bernstein concentration inequality for martingales. In the second step, we focus on a doubly robust estimator of the finite population mean and re-estimate the nuisance model parameters by minimizing the asymptotic squared bias of the doubly robust estimator. This estimating strategy mitigates the possible first-step selection error and renders the doubly robust estimator root-n consistent if either the sampling probability or the outcome model is correctly specified.

Keywords: Data integration; Double robustness; Generalizability; Penalized estimating equation; Variable selection

1 INTRODUCTION

Probability sampling is regarded as the gold-standard in survey statistics for finite population inference. Fundamentally, probability samples are selected under known sampling designs and therefore are representative of the target population. However, many practical challenges arise in collecting and analyzing probability sample data such as cost, time duration, and increasing non-response rates (Keiding and Louis, 2016). As the advancement of technology, big non-probability samples become increasingly available for research purposes, such as remote sensing data, Internet samples, etc. Although non-probability samples do not contain information on the sampling mechanism, they provide rich information about the target population and can be potentially helpful for finite population inference. These complementary features of probability samples and non-probability samples raise the question of whether it is possible to develop data integration methods that leverage the advantages of both data sources.

Existing methods for data integration can be categorized into three types. The first type is the so-called propensity score adjustment (Rosenbaum and Rubin, 1983). In this approach, the probability of a unit being selected into the non-probability sample, which is referred to as the propensity or sampling score, is modeled and estimated for all units in the non-probability sample. The subsequent adjustments, such as propensity score weighting or stratification, can then be used to adjust for selection biases; see, e.g., Lee and Valliant (2009), Valliant and Dever (2011), Elliott et al. (2017) and Chen et al. (2018). Stuart et al. (2011; 2015) and Buchanan et al. (2018) use propensity score weighting to generalize results from randomized trials to a target population. O'Muircheartaigh and Hedges (2014) propose propensity score stratification for analyzing a non-randomized social experiment. One notable disadvantage of the propensity score methods is that they rely on an explicit

propensity score model and are biased and highly variable if the model is misspecified (Kang and Schafer, 2007). The second type uses calibration weighting (Deville and Särndal, 1992, Kott, 2006). This technique calibrates auxiliary information in the non-probability sample with that in the probability sample, so that after calibration the weighted distribution of the non-probability sample is similar to that of the target population (DiSogra et al., 2011). The third type is mass imputation, which imputes the missing values for all units in the probability sample. In the usual imputation for missing data analysis, the respondents in the sample constitute a training dataset for developing an imputation model. In the mass imputation, an independent non-probability sample is used as a training dataset, and imputation is applied to all units in the probability sample; see, e.g., Breidt et al. (1996), Rivers (2007), Kim and Rao (2012), Chipperfield et al. (2012), Bethlehem (2016), and Yang and Kim (2018).

Let $X \in \mathbb{R}^p$ be a vector of auxiliary variables (including an intercept) that are available from two data sources, and let $Y \in \mathbb{R}$ be the study variable of interest. We consider combining a probability sample with X, referred to as Sample A, and a non-probability sample with (X, Y), referred to as Sample B, to estimate μ the population mean of Y. Because the sampling mechanism of a non-probability sample is unknown, the target population quantity is not identifiable in general. Researchers rely on an identification strategy that requires a non-informative sampling assumption imposed on the non-probability sample. To ensure this assumption holds, researchers should control for all covariates that are predictors of both sampling and the outcome variable. In practice, subject matter experts will recommend a rich set of potential useful variables but will not identify the exact variables to adjust for. In the presence of many auxiliary variables, variable selection becomes important, because existing methods may become unstable or even infeasible, and irrelevant

auxiliary variables can introduce a large variability in estimation. There is a large literature on variable selection methods for prediction, but little work on variable selection for data integration that can successfully recognize the strengths and the limitations of each data source and utilize all information captured for finite population inference. Gao and Carroll (2017) propose a pseudo-likelihood approach to combining multiple non-survey data with high dimensionality; this approach requires all likelihoods be correctly specified and therefore is sensitive to model misspecification. Robust inference has not been addressed in the context of data integration with high-dimensional data.

We propose a doubly robust variable selection and estimation strategy that harnesses the representativeness of the probability sample and the outcome and covariate information in the non-probability sample. The double robustness entails that the final estimator is consistent for the true value if either the probability of selection into the non-probability sample, referred to as the sampling score, or the outcome model is correctly specified, not necessarily both (a double robustness condition); see, e.g., Bang and Robins (2005), Tsiatis (2006), Cao et al. (2009), and Han and Wang (2013). To handle potentially high-dimensional covariates, our strategy separates the variable selection step and the estimation step for finite population mean to achieve two different goals.

In the first step, we select a set of variables that are important predictors of either the sampling score or the outcome model by penalized estimating equations. Following most of the empirical literature, we assume the sampling score follows a logistic regression model with the unknown parameter $\alpha \in \mathbb{R}^p$ and the outcome follows a generalized linear model (accommodating different types of the outcome) with the unknown parameter $\beta \in \mathbb{R}^p$. Importantly, we separate the estimating equations for α and β in order to achieve stability in variable selection under the double robustness condition. Specifically, we construct the

estimating equation for α by calibrating the weighted average of X from Sample B, weighted by the inverse of the sampling score, to the design weighted average of X from Sample A (i.e., a design estimate of population mean of X). We construct the estimating equation for β by minimizing the standard least squared error loss under the outcome model. To establish the selection properties, we consider the "large n, diverging p" framework. To the best of our knowledge, the asymptotic properties of penalized estimating estimation based on survey data have not been studied in the literature. Our major technical challenge is that under the finite population framework, the sampling indicator of Sample A may not be independent even under simple random sampling. To overcome this challenge, we construct martingale random variables with a weak dependence that allows applying Bernstein inequality. This construction is innovative and crucial in establishing our new selection consistency result.

In the second step, we consider a doubly robust estimator of μ, $\widehat{\mu}_{\text{dr}}(\widehat{\alpha}, \widehat{\beta})$ and re-estimate (α, β) based on the joint set of variables selected from the first step. We propose using different estimating equations to estimate (α, β), derived by minimizing the asymptotic squared bias of $\widehat{\mu}_{\text{dr}}(\widehat{\alpha}, \widehat{\beta})$. This estimation strategy is not new; see, e.g., Kim and Haziza (2014) for missing data analyses and Vermeulen and Vansteelandt (2015; 2016) for causal inference of treatment effects in low-dimensional data; however, we demonstrate its new role in high-dimensional data to mitigate the possible selection error in the first step. In essence, our strategy for estimating (α, β) renders the first order term in the Taylor expansion of $\widehat{\mu}_{\text{dr}}(\widehat{\alpha}, \widehat{\beta})$ with respect to (α, β) to be exactly zero, and the remaining terms are negligible under regularity conditions. Therefore, the proposed estimator allows model misspecification of either the sampling score or the outcome model. Moreover, we propose a consistent variance estimator allowing for doubly robust inferences.

The paper proceeds as follows. Section 2 provides the basic setup of the paper. Section 3 presents the proposed two-step procedure for variable selection and doubly robust estimation of the finite population mean. Section 4 describes the computation algorithm for solving penalized estimating equations. Section 5 presents the theoretical properties for variable selection and doubly robust estimation. Section 6 reports simulation results that illustrate the finite-sample performance of the proposed method. In Section 7, we present an application to analyze a non-probability sample collected by the Pew Research Centre. Section 8 concludes with a discussion. We relegate all proofs to the supplementary material.

2 BASIC SETUP

2.1 Notation: Two Samples

Let $\mathcal{U} = \{1, \ldots, N\}$ be the index set of N units for the finite population, with N being the known population size. The finite population consists of $\mathcal{F}_N = \{(X_i, Y_i) : i \in \mathcal{U}\}$. Let the parameter of interest be the finite population mean $\mu = N^{-1} \sum_{i=1}^{N} Y_i$. We consider two data sources: one from a probability sample, referred to as Sample A, and the other one from a non-probability sample, referred to as Sample B. Table 1 illustrates the observed data structure. Sample A consists of observations $\mathcal{O}_A = \{(d_{A,i} = \pi_{A,i}^{-1}, X_i) : i \in \mathcal{A}\}$ with sample size n_A, where $\pi_{A,i} = P(i \in \mathcal{A})$ is known throughout Sample A, and Sample B consists of observations $\mathcal{O}_B = \{(X_i, Y_i) : i \in \mathcal{B}\}$ with sample size n_B. We define $I_{A,i}$ and $I_{B,i}$ to be the indicators of selection to Sample A and Sample B, respectively. Although the non-probability sample contains rich information on (X, Y), the sampling mechanism is unknown, and therefore we cannot compute the first-order inclusion probability for Horvitz–

Table 1: Two data sources. "$\surd$" and "?" indicate observed and unobserved data, respectively.

		Sample weight	Covariate	Study Variable
		π^{-1}	X	Y
Probability	1	$\surd$	$\surd$	?
Sample	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{O}_A$	n_A	$\surd$	$\surd$	?
Non-probability	$n_A + 1$	?	$\surd$	$\surd$
Sample	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{O}_B$	$n_A + n_B$	?	$\surd$	$\surd$

Sample A is a probability sample, and Sample B is a non-probability sample.

Thompson estimation. The naive estimators without adjusting for the sampling process are subject to selection biases (Meng, 2018). On the other hand, although the probability sample with sampling weights represents the finite population, it does not observe the study variable of interest.

2.2 An Identification Assumption

Before presenting the proposed methodology for integrating the two data sources, we first discuss the identification assumption. Let $f(Y \mid X)$ be the conditional distribution of Y given X in the superpopulation model ζ that generates the finite population. We make the following primary assumption.

Assumption 1 *(i) The sampling indicator I_B of Sample B and the response variable Y is*

independent given X; i.e. $P(I_B = 1 \mid X, Y) = P(I_B = 1 \mid X)$, referred to as the sampling score $\pi_B(X)$, and (ii) $\pi_B(X) > N^{\gamma-1}\delta_B > 0$ for all X, where $\gamma \in (2/3, 1]$.

Assumption 1 (i) implies that $E(Y \mid X) = E(Y \mid X, I_B = 1)$, denoted by $m(X)$, can be estimated based solely on Sample B. Assumption 1 (ii) specifies a lower bound of $\pi_B(X)$ for the technicality in Section 5. A standard condition in the literature imposes a strict positivity in the sense that $\pi_B(X) > \delta_B > 0$; however, it implies that $n_B = O(N)$, which may be restrictive in survey sampling. Here, we relax this condition and allow $n_B = O(N^\gamma)$, where γ can be strictly less than 1.

Assumption 1 is a key assumption for identification. Under Assumption 1, $E(\mu)$ is identifiable based on Sample A by $E\{I_A m(X)\}$ or Sample B by $E\{I_B Y/\pi_B(X)\}$. However, this assumption is not verifiable from the observed data. To ensure this assumption holds, researchers often consider many potentially predictors for the sampling indicator I_B or the outcome Y, resulting in a rich set of variables in X.

2.3 Existing Estimators

In practice, the sampling score function $\pi_B(X)$ and the outcome mean function $m(X)$ are unknown and need to be estimated from the data. Let $\pi_B(X^{\mathrm{T}}\alpha)$ and $m(X^{\mathrm{T}}\beta)$ be the posited models for $\pi_B(X)$ and $m(X)$, respectively, where α and β are unknown parameters. First, under Assumption 1, we can obtain $\widehat{\beta}$ by fitting the outcome model based solely on $\mathcal{O}_B = \{(X_i, Y_i) : i \in \mathcal{B}\}$. Second, following Valliant and Dever (2011), we can obtain $\widehat{\alpha}$ by fitting the sampling score model based on the blended data $\mathcal{O}_A \cup \mathcal{O}_B = \{(d_{A,i}, X_i, I_i = 0) : i \in \mathcal{A}\} \cup \{(X_i, I_i = 1) : i \in \mathcal{B}\}$, weighted by the design weights from Sample A. The resulting estimator $\widehat{\alpha}$ is valid if the size of Sample B is relatively small (Valliant and Dever, 2011).

Given $m(X;\widehat{\beta})$ and $\pi_B(X;\widehat{\alpha})$, researchers have proposed different estimators for μ. We provide examples below and discuss their properties and limitations.

Example 1 (Inverse probability of sampling score weighting) *The inverse probability of sampling score weighting estimator is*

$$\widehat{\mu}_{\text{IPW}} = \widehat{\mu}_{\text{IPW}}(\widehat{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} \frac{I_{B,i}}{\pi_B(X_i^{\text{T}}\widehat{\alpha})} Y_i. \tag{1}$$

The justification for $\widehat{\mu}_{\text{IPW}}$ relies on the correct specification of $\pi_B(X)$ and the consistency of $\widehat{\alpha}$. If $\pi_B(X_i^{\text{T}}\alpha)$ is misspecified or $\widehat{\alpha}$ is inconsistent, $\widehat{\mu}_{\text{IPW}}$ is biased.

Example 2 (Calibration weighting) *The calibration weighting estimator is*

$$\widehat{\mu}_{\text{cal}} = \widehat{\mu}_{\text{cal}} = \frac{1}{N}\sum_{i=1}^{N} \omega_i I_{B,i} Y_i, \tag{2}$$

where $\{\omega_i : i \in \mathcal{S}_B\}$ *satisfies* $\sum_{i\in\mathcal{S}_B} \omega_i X_i = \sum_{i\in\mathcal{S}_A} d_{A,i} X_i$.

The justification for $\widehat{\mu}_{\text{cal}}$ relies on the linearity of the outcome model, i.e., $m(X) = X^{\text{T}}\beta^*$ for some β^*, or the linearity of the inverse probability of sampling weight, i.e., $\pi_B(X)^{-1} = X^{\text{T}}\alpha^*$ for some α^* (Fuller, 2009; Theorem 5.1). The linearity conditions are unlikely to hold for non-continuous variables. In these cases, $\widehat{\mu}_{\text{cal}}$ is biased.

Example 3 (Outcome regression based on Sample A) *The outcome regression estimator is*

$$\widehat{\mu}_{\text{reg}} = \widehat{\mu}_{\text{reg}}(\widehat{\beta}) = \frac{1}{N}\sum_{i=1}^{N} I_{A,i} d_{A,i} m(X_i;\widehat{\beta}). \tag{3}$$

The justification for $\widehat{\mu}_{\text{reg}}$ relies on the correct specification of $m(X)$ and the consistency of $\widehat{\beta}$. If $m(X^{\text{T}}\beta)$ is misspecified or $\widehat{\beta}$ is inconsistent, $\widehat{\mu}_{\text{reg}}$ is biased.

Example 4 (Doubly robust estimator) *The doubly robust estimator is*

$$\widehat{\mu}_{\mathrm{dr}} = \widehat{\mu}_{\mathrm{dr}}(\widehat{\alpha}, \widehat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{I_{B,i}}{\widehat{\pi}_B(X_i^{\mathrm{T}}\widehat{\alpha})}\{Y_i - m(X_i;\widehat{\beta})\} + I_{A,i}d_{A,i}m(X_i;\widehat{\beta})\right]. \tag{4}$$

The estimator $\widehat{\mu}_{\mathrm{dr}}$ is doubly robust with fixed-dimensional X (Chen et al., 2018), in the sense that it achieves the consistency if either $\pi_B(X_i^{\mathrm{T}}\alpha)$ or $m(X^{\mathrm{T}}\beta)$ is correctly specified, but not necessarily both. The double robustness is attractive; therefore, we shall investigate the potential of $\widehat{\mu}_{\mathrm{dr}}$ in high-dimensional data.

3 METHODOLOGY IN HIGH-DIMENSIONAL DATA

A major challenge arises in the presence of a large number of covariates, not all of them are necessary for making inference of the population mean of the outcome. This necessitates variable selection. For simplicity of exposition, we introduce the following notation. For any vector $\alpha \in \mathbb{R}^p$, denote the number of nonzero elements in α as $||\alpha||_0 = \sum_{j=1}^p I(\alpha_j \neq 0)$, the L_1-norm as $||\alpha||_1 = \sum_{j=1}^p |\alpha_j|$, the L_2-norm as $||\alpha||_2 = \sqrt{\sum_{j=1}^p \alpha_j^2}$, and the L_∞-norm as $||\alpha||_\infty = \max_{j=1}^p |\alpha_j|$. For any $\mathcal{J} \subseteq \{1,\ldots,p\}$, let $\alpha_{\mathcal{J}}$ be the sub-vector of α formed by elements of α whose indexes are in $\mathcal{J}$. Let $\mathcal{J}^c$ be the complement of $\mathcal{J}$. For any $\mathcal{J}_1, \mathcal{J}_2 \subseteq \{1,\ldots,p\}$ and matrix $\Sigma \in \mathbb{R}^{p\times p}$, let $\Sigma_{\mathcal{J}_1,\mathcal{J}_2}$ be the sub-matrix of Σ formed by rows in $\mathcal{J}_1$ and columns in $\mathcal{J}_2$. Following the literature on variable selection, we can first standardize the covariates so that approximately they have variances equal to one to stabilize the variable selection procedure. We make the following modeling assumptions.

Assumption 2 (Sampling score model) *We assume a logistic regression model for $\pi_B(X)$; i.e., logit$\{\pi_B(X^{\mathrm{T}}\alpha)\} = X^{\mathrm{T}}\alpha$ for $\alpha \in \mathbb{R}^p$. Define α^* to be the p-dimensional parameter that*

minimizes the Kullback-Leibler divergence

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^p} E\left[\pi_B(X)\log\frac{\pi_B(X^{\mathrm{T}}\alpha)}{\pi_B(X)} + \{1-\pi_B(X)\}\log\frac{1-\pi_B(X^{\mathrm{T}}\alpha)}{1-\pi_B(X)}\right].$$

Assumption 3 (Outcome model) *We assume a generalized linear regression model for $m(X)$; i.e. $m(X^{\mathrm{T}}\beta)$ for $\beta \in \mathbb{R}^p$, where $m(\cdot)$ is a link function by an abuse the notation. Define $\beta^* = \arg\min_\beta E\left[I_B\{Y - m(X^{\mathrm{T}}\beta)\}^2\right]$.*

The models $\pi_B(X^{\mathrm{T}}\alpha)$ and $m(X^{\mathrm{T}}\beta)$ are working models and they may be misspecified. If the sampling score model is correctly specified, $\pi_B(X) = \pi_B(X^{\mathrm{T}}\alpha^*)$. If the outcome model is correctly specified, $m(X) = m(X^{\mathrm{T}}\beta^*)$.

The proposed procedure consists of two steps: the first step selects important variables in the sampling score model and the outcome model, and the second step focuses on doubly robust estimation of the population mean.

In the first step, we propose solving penalized estimating equations for variable selection. To select important variables in $\pi_B(X^{\mathrm{T}}\alpha)$, the traditional loss function under the logistic regression model,

$$\frac{1}{N}\sum_{i=1}^{N}\left[\log\{1+\pi_B(X_i^{\mathrm{T}}\alpha)\} - I_{B,i}X_i^{\mathrm{T}}\alpha\right],$$

is not feasible, because it requires the availability of the population information on X. To overcome this difficulty, the key insight is that under Assumption 2,

$$E\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\alpha)}X_i\right\} = E\left(\frac{I_{A,i}}{\pi_{A,i}}X_i\right) = E(X_i).$$

Therefore, we define the estimating function for α as

$$U_1(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\alpha)} - \frac{I_{A,i}}{\pi_{A,i}}\right\}X_i.$$

To select important variables in $m(X^{\mathrm{T}}\beta)$, under Assumption 1, we have $E(Y \mid X) = E(Y \mid X, I_B = 1)$. Therefore, we define the estimating function for β as

$$U_2(\beta) = \frac{1}{N}\sum_{i=1}^{N} I_{B,i}\left\{Y_i - m(X_i^{\mathrm{T}}\beta)\right\} X_i.$$

Let $U(\theta) = (U_1(\alpha)^{\mathrm{T}}, U_2(\beta)^{\mathrm{T}})^{\mathrm{T}}$ be the joint estimating function for $\theta = (\alpha^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$. When p is large, following Johnson et al. (2008), we consider solving the penalized estimating function

$$U^{\mathrm{p}}(\alpha, \beta) = U(\alpha, \beta) - \begin{pmatrix} q_{\lambda_\alpha}(|\alpha|)\mathrm{sign}(\alpha) \\ q_{\lambda_\beta}(|\beta|)\mathrm{sign}(\beta) \end{pmatrix}, \tag{5}$$

for (α, β), where $q_{\lambda_\alpha}(\alpha) = \{q_{\lambda_\alpha}(|\alpha_0|), \ldots, q_{\lambda_\alpha}(|\alpha_p|)\}^{\mathrm{T}}$ and $q_{\lambda_\beta}(\beta) = \{q_{\lambda_\beta}(|\beta_0|), \ldots, q_{\lambda_\beta}(|\beta_p|)\}^{\mathrm{T}}$ are some continuous functions, $q_{\lambda_\alpha}(|\alpha|)\mathrm{sign}(\alpha)$ is the element-wise product of $q_{\lambda_\alpha}(\alpha)$ and $\mathrm{sign}(\alpha)$, and $q_{\lambda_\beta}(|\beta|)\mathrm{sign}(\beta)$ is the element-wise product of $q_{\lambda_\beta}(\beta)$ and $\mathrm{sign}(\beta)$. We let $q_\lambda(x) = \mathrm{d}p_\lambda(x)/\mathrm{d}x$, where $p_\lambda(x)$ is some penalization function. Although the same discussion applies to different non-concave penalty functions, we specify $p_\lambda(x)$ to be a folded-concave smoothly clipped absolute deviation (SCAD) penalty function (Fan and Lv, 2011). Accordingly, we have

$$q_\lambda(|\theta|) = \lambda\left\{I(|\theta| < \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| \geq \lambda)\right\}, \tag{6}$$

for $a > 0$, where $(\cdot)_+$ is the truncated linear function; i.e., if $x \geq 0$, $(x)_+ = x$, and if $x < 0$, $(x)_+ = 0$. Following the suggestion of Fan and Lv (2011), we use $a = 3.7$. We select the variables if the corresponding estimates of (5) are nonzero in either the sampling score or the outcome model, indexed by $\mathcal{C}$.

Remark 1 *To help understand the penalized estimating equation, we discuss two scenarios. If $|\alpha_j|$ is large, then $q_{\lambda_\alpha}(|\alpha_j|)$ is zero, and therefore $U_{1,j}(\alpha)$ is not penalized. Whereas, if*

$|\alpha_j|$ is small but nonzero, then $q_{\lambda_\alpha}(|\alpha_j|)$ is large, and therefore $U_{1,j}(\alpha)$ is penalized with a penalty term. The penalty term then forces $\widehat{\alpha}_j$ to be zero and excludes the jth element in X from the final selected set of variables. The same discussion applies to $U_2(\beta)$ and $q_{\lambda_\beta}(|\beta|)$.

In the second step, we consider the estimator of the population mean $\widehat{\mu}_{\rm dr}(\widehat{\alpha},\widehat{\beta})$ in (4) with $(\widehat{\alpha},\widehat{\beta})$ re-estimated based on $X_{\mathcal{C}}$. As we will show in Section 5, $\mathcal{C}$ contains the true important variables in either the sampling score model or the outcome model with probability approaching one (the oracle property). Therefore, if either the sampling score model or the outcome model is correctly specified, the asymptotic bias of $\widehat{\mu}_{\rm dr}(\alpha^*,\beta^*)$ is zero; however, if both models are misspecified, the asymptotic bias of $\widehat{\mu}_{\rm dr}(\alpha^*,\beta^*)$ is

$$\begin{aligned}
\text{a.bias}(\alpha^*,\beta^*) &= E\left\{\widehat{\mu}_{\rm dr}(\alpha^*,\beta^*)-\mu\right\} \\
&= E\left(\frac{1}{N}\sum_{i=1}^{N}\left[\frac{I_{B,i}}{\pi_B(X_i^{\rm T}\alpha^*)}\{Y_i-m(X_i^{\rm T}\beta^*)\}+I_{A,i}d_{A,i}m(X_i^{\rm T}\beta^*)\right]\right)-\mu \\
&= E\left[\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{I_{B,i}}{\pi_B(X_i^{\rm T}\alpha^*)}-1\right\}\{Y_i-m(X_i^{\rm T}\beta^*)\}\right] \\
&\quad +E\left\{\frac{1}{N}\sum_{i=1}^{N}\left(I_{A,i}d_{A,i}-1\right)m(X_i^{\rm T}\beta^*)\right\}.
\end{aligned}$$

In order to minimize $\{\text{a.bias}(\alpha^*,\beta^*)\}^2$, we consider the joint estimating function

$$J(\alpha,\beta)=\begin{pmatrix} J_1(\alpha,\beta)\\ J_2(\alpha,\beta)\end{pmatrix}=\begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} I_{B,i}\left\{\frac{1}{\pi_B(X_i^{\rm T}\alpha)}-1\right\}\{Y_i-m(X_i^{\rm T}\beta)\}X_{i\mathcal{C}}\\ \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{I_{B,i}}{\pi_B(X_i^{\rm T}\alpha)}-d_{A,i}I_{A,i}\right\}\partial m(X_i^{\rm T}\beta)/\partial\beta_{\mathcal{C}}\end{pmatrix} \tag{7}$$

for estimating (α,β), constrained on $\{(\alpha^{\rm T},\beta^{\rm T})^{\rm T}\in\mathbb{R}^{2p}:\alpha_{\mathcal{C}^c}=0,\beta_{\mathcal{C}^c}=0\}$.

Remark 2 *The two steps use different estimating functions (5) and (7), respectively, for selection and estimation with the following advantages. First, (5) separates the selection*

for α and β in $U_1(\alpha)$ and $U_2(\beta)$, so it stabilizes the selection procedure if either the sampling score model or the outcome model is misspecified. Second, using the joint estimating function (7) for estimation leads to an attractive feature in the estimation of μ: this estimating strategy mitigates the possible first-step selection error and renders $\widehat{\mu}_{\mathrm{dr}}(\widehat{\alpha}, \widehat{\beta})$ root-n consistent if either the sampling probability or the outcome model is correctly specified in high-dimensional data. We relegate the details to Section 5.

In summary, our two-step procedure for variable selection and estimation is as follows. spacing

Step 1. To facilitate joint selection of variables for the sampling score and outcome, solve the penalized joint estimating equations $U^{\mathrm{p}}(\alpha, \beta) = 0$ in (5), denoted by $(\widetilde{\alpha}, \widetilde{\beta})$. Let $\widehat{\mathcal{M}}_\alpha = \{j : \widetilde{\alpha}_j \neq 0\}$ and $\widehat{\mathcal{M}}_\beta = \{j : \widetilde{\beta}_j \neq 0\}$.

Step 2. Let the set of variables for estimation be $\mathcal{C} = \widehat{\mathcal{M}}_\alpha \cup \widehat{\mathcal{M}}_\beta$. Obtain the proposed estimator as

$$\widehat{\mu}_{\text{p-dr}} = \widehat{\mu}_{\text{p-dr}}(\widehat{\alpha}, \widehat{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left\{ I_{B,i} \frac{Y_i - m(X_i^{\mathrm{T}} \widehat{\beta})}{\pi_B(X_i^{\mathrm{T}} \widehat{\alpha})} + I_{A,i} d_{A,i} m(X_i^{\mathrm{T}} \widehat{\beta}) \right\}, \tag{8}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are obtained by solving the joint estimating equations (7) for α and β with $\alpha_{\mathcal{C}^c} = 0$ and $\beta_{\mathcal{C}^c} = 0$.

Remark 3 *Variable selection circumvents the instability or infeasibility of direct estimation of (α, β) with high-dimensional X. Moreover, in Step 2 for estimation, we consider a union of covariates $X_{\mathcal{C}}$, where $\mathcal{C} = \widehat{\mathcal{M}}_\alpha \cup \widehat{\mathcal{M}}_\beta$. Brookhart et al. (2006) and Shortreed and Ertefaie (2017) show that including variables that are related to the outcome in the propensity score model will increase the precision of the estimated average treatment effect*

without increasing bias. This implies that an efficient variable selection and estimation method should take into account both sampling-covariate and outcome-covariate relationships. As a result, $\widehat{\mu}_{\mathrm{dr}}(\widehat{\alpha}, \widehat{\beta})$ *may have a better performance than the oracle estimator which uses the true important variables in the sampling score and the outcome model. This is particularly true when one of the models is misspecified. Our simulation study in Section 6 demonstrates that* $\widehat{\mu}_{\mathrm{dr}}$ *with variable selection has a similar performance as the orcale estimator for the continuous outcome and outperforms the oracle estimator for the binary outcome.*

4 COMPUTATION

In this section, we discuss the computation for solving the penalized estimating function (5). Following Johnson et al. (2008), we use an iterative algorithm that combines the Newton–Raphson algorithm for solving estimating equation and the minorization-maximization algorithm for non-convex penalty of Hunter and Li (2005).

First, by the minorization-maximization algorithm, the penalized estimator $\widetilde{\theta} = (\widetilde{\alpha}, \widetilde{\beta})$ solving (5) satisfies

$$U^{\mathrm{p}}(\widetilde{\theta}) = U(\widetilde{\theta}) - \begin{pmatrix} q_{\lambda_{\widetilde{\alpha}}}(|\widetilde{\alpha}|)\mathrm{sign}(\widetilde{\alpha})\frac{|\widetilde{\alpha}|}{\epsilon+|\widetilde{\alpha}|} \\ q_{\lambda_{\widetilde{\beta}}}(|\widetilde{\beta}|)\mathrm{sign}(\widetilde{\beta})\frac{|\widetilde{\beta}|}{\epsilon+|\widetilde{\beta}|} \end{pmatrix} = 0, \tag{9}$$

for ϵ is a predefined small number. In our implementation, we choose ϵ to be 10^{-6}.

Second, we solve (9) by the Newton-Raphson algorithm. It may be challenging to implement the Newton-Raphson algorithm directly, because it involves inverting a large matrix. For ease and stability in those cases, we can use a coordinate decent algorithm (Friedman et al., 2007) by cycling through and updating each of the coordinates.

Following most of the empirical literature, we assume that $\pi_B(X^{\mathrm{T}}\alpha)$ follows a logistic regression model. Define $m^{(k)}(t) = \mathrm{d}^k m(t)/\mathrm{d}^k t$ for $k \geq 1$. We denote

$$\begin{aligned} \nabla(\theta) = \partial U(\theta)/\partial\theta^{\mathrm{T}} &= \mathrm{diag}\{\partial U_1(\alpha)/\partial\alpha^{\mathrm{T}}, \partial U_2(\beta)/\partial\beta^{\mathrm{T}}\}, \\ \frac{\partial U_1(\alpha)}{\partial\alpha^{\mathrm{T}}} &= -\frac{1}{N}\sum_{i=1}^{N} I_{B,i}\frac{1-\pi_B(X_i^{\mathrm{T}}\alpha)}{\pi_B(X_i^{\mathrm{T}}\alpha)}X_i X_i^{\mathrm{T}}, \\ \frac{\partial U_2(\beta)}{\partial\beta^{\mathrm{T}}} &= -\frac{1}{N}\sum_{i=1}^{N} I_{B,i} m^{(1)}(X_i^{\mathrm{T}}\beta)^2 X_i X_i^{\mathrm{T}}, \end{aligned} \tag{10}$$

and

$$E(\theta) = \begin{pmatrix} q_{\lambda_1}(|\theta_1|) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & q_{\lambda_{2p}}(|\theta_{2p}|) \end{pmatrix}.$$

Let θ start at an initial value $\widetilde{\theta}^{[0]}$. With the other coordinates fixed, the kth Newton-Raphson update for θ_j $(j = 1, \ldots, 2p)$, the jth element of θ, is

$$\widetilde{\theta}_j^{[k]} = \widetilde{\theta}_j^{[k-1]} + \left\{\nabla_{jj}(\widetilde{\theta}^{[k-1]}) + N \cdot E_{jj}(\widetilde{\theta}^{[k-1]})\right\}^{-1}\left\{U_j(\widetilde{\theta}^{[k-1]}) - N \cdot E_{jj}(\widetilde{\theta}^{[k-1]})\widetilde{\theta}_j^{[k-1]}\right\}, \tag{11}$$

where $\nabla_{jj}(\theta)$ and $E_{jj}(\theta)$ are the jth diagonal elements in $\nabla(\theta)$ and $E(\theta)$, respectively. The procedure cycles through all the $2p$ elements of θ and is repeated until convergence.

We use K-fold cross-validation to select the tuning parameter $(\lambda_\alpha, \lambda_\beta)$. To be specific, we partition both samples into approximately K equal sized subsets and pair subsets of Sample A and subsets of Sample B randomly. Of the K pairs, we retain one single pair as the validation data and the remaining $K-1$ pairs as the training data. We fit the models based on the training data and estimate the loss function based on the validation data. We repeat the process K times, with each of the K pairs used exactly once as the validation

data. Finally, we aggregate the K estimated loss function. We select the tuning parameter as the one that minimizes the aggregated loss function over a pre-specified grid.

Because the weighting estimator uses the sampling score $\pi_B(X)$ to calibrate the distribution of $X_{\mathcal{C}}$ between Sample B and the target population, we use the following loss function for selecting λ_α:

$$\text{Loss}(\lambda_\alpha) = \sum_{j=1}^{p} \left[\sum_{i=1}^{N} \left\{ \frac{I_{B,i}}{\pi_B\{X_i^{\mathrm{T}} \widehat{\alpha}(\lambda_\alpha)\}} - \frac{I_{A,i}}{\pi_{A,i}} \right\} X_{i,j} \right]^2,$$

where $\widetilde{\alpha}(\lambda_\alpha)$ is the penalized estimator $\widetilde{\alpha}$ with the tuning parameter λ_α. We use the prediction error loss function for selecting λ_β:

$$\text{Loss}(\lambda_\beta) = \sum_{i=1}^{N} I_{B,i} \left[Y_i - m\{X_i^{\mathrm{T}} \widehat{\beta}(\lambda_\beta)\} \right]^2,$$

where $\widetilde{\beta}(\lambda_\beta)$ is the penalized estimator $\widetilde{\beta}$ with the tuning parameter λ_β.

5 ASYMPTOTIC RESULTS FOR VARIABLE SELECTION AND ESTIMATION

We establish the asymptotic properties for the proposed double variable selection and doubly robust estimation. We can establish theoretical results for general sampling mechanisms for Sample A requiring specific regularity conditions. In this section, for technical convenience, we assume that Sample A is collected by simple random sampling or Poisson sampling with the following regularity conditions.

Assumption 4 *For all* $1 \leq i \leq N$, $\pi_{A,i} \geq N^{\gamma-1}\delta_A > 0$, *where* $\gamma \in (2/3, 1]$.

Similar to Assumption 1 (ii), we relax the strict positivity on $\pi_{A,i}$ and render $n_A = O(N^\gamma)$ for γ possibly strictly less than 1. Let $n = \min(n_A, n_B)$, which is $O(N^\gamma)$ under Assumptions 1 and 4.

Let the support of model parameters be

$$\mathcal{M}_\alpha = \{1 \le j \le p : \alpha_j^* \neq 0\}, \quad \mathcal{M}_\beta = \{1 \le j \le p : \beta_j^* \neq 0\}, \quad \mathcal{M}_\theta = \mathcal{M}_\alpha \cup \{p + \mathcal{M}_\beta\}.$$

Define $s_\alpha = ||\alpha^*||_0$, $s_\beta = ||\beta^*||_0$, $s_\theta = s_\alpha + s_\beta$, and $\lambda_\theta = \min(\lambda_\alpha, \lambda_\beta)$.

Assumption 5 *The following regularity conditions hold.*

(A1) *The parameter θ belongs to a compact subset in $\mathbb{R}^{2p}$, and θ^* lies in the interior of the compact subset.*

(A2) *$\{X_i : i \in \mathcal{U}\}$ are fixed and uniformly bounded.*

(A3) *There exist constants c_1 and c_2 such that*

$$0 < c_1 \le \lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^N X_i^{\mathrm{T}} X_i\right) \le \lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^N X_i^{\mathrm{T}} X_i\right) \le c_2 < \infty,$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are the minimum and the maximum eigenvalue of a matrix, respectively.

(A4) *Let $\epsilon_i(\beta) = Y_i - m(X_i^{\mathrm{T}}\beta)$ be the ith residual. There exists a constant c_3 such that $E\{|\epsilon_i(\beta^*)|^{2+\delta}\} \le c_3$ for all $1 \le i \le N$ and some $\delta > 0$. There exist constants c_4 and c_5 such that $E[\exp\{c_4|\epsilon_i(\beta^*)|\} \mid X_i] \le c_5$ for all $1 \le i \le N$.*

(A5) *$m^{(1)}(X_i^{\mathrm{T}}\beta)$, $m^{(2)}(X_i^{\mathrm{T}}\beta)$, and $m^{(3)}(X_i^{\mathrm{T}}\beta)$ are uniformly bounded away from ∞ on $\mathcal{N}_{\theta,\tau} = \{\theta \in \mathbb{R}^{2p} : ||\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}|| \le \tau\sqrt{s_\theta/n}, \theta_{\mathcal{M}_\theta^c} = 0\}$ for some $\tau > 0$.*

(A6) $\min_{j\in\mathcal{M}_\alpha}|\alpha_j^*|/\lambda_\alpha \to \infty$ *and* $\min_{k\in\mathcal{M}_\beta}|\beta_k^*|/\lambda_\beta \to \infty$, *as* $n\to\infty$.

(A7) $s_\theta = o(n^{1/3})$, $\lambda_\alpha, \lambda_\beta \to 0$, $(\log n)^2 = o(n\lambda_\theta^2)$, $\log(p) = o\{n\lambda_\theta^2/(\log n)^2\}$, $ps_\theta^4(\log n)^6 = o(n^3\lambda_\theta^2)$, $ps_\theta^4(\log n)^8 = o(n^4\lambda_\theta^4)$, *as* $n\to\infty$.

These assumptions are typical in the penalization literature. (A2) specifies a fixed design which is well suited under the finite population inference framework. (A4) holds for Gaussian distribution, sub-Gaussian distribution, and so on. (A5) holds for common models. For example, for the linear regression model with $m(X_i^{\mathrm{T}}\beta) = X_i^{\mathrm{T}}\beta$, then

$$m^{(1)}(X_i^{\mathrm{T}}\beta) = \beta,\ m^{(2)}(X_i^{\mathrm{T}}\beta) = m^{(3)}(X_i^{\mathrm{T}}\beta) = 0.$$

For the logistic regression model with $m(X_i^{\mathrm{T}}\beta) = \exp(X_i^{\mathrm{T}}\beta)/\{1+\exp(X_i^{\mathrm{T}}\beta)\}$, then

$$\begin{aligned}
m^{(1)}(X_i^{\mathrm{T}}\beta) &= -m(X_i^{\mathrm{T}}\beta)\{1-m(X_i^{\mathrm{T}}\beta)\},\\
m^{(2)}(X_i^{\mathrm{T}}\beta) &= -m(X_i^{\mathrm{T}}\beta)\{1-m(X_i^{\mathrm{T}}\beta)\}\{2m(X_i^{\mathrm{T}}\beta)-1\},\\
m^{(3)}(X_i^{\mathrm{T}}\beta) &= -m(X_i^{\mathrm{T}}\beta)\{1-m(X_i^{\mathrm{T}}\beta)\}\{6m(X_i^{\mathrm{T}}\beta)^2-6m(X_i^{\mathrm{T}}\beta)+1\}.
\end{aligned}$$

Under these models, (A1) and (A2) imply (A5). (A7) specifies the restrictions on the dimension of covariates p and the dimension of the true nonzero coefficients s_θ. To gain insight, when the true model size s_θ is fixed, (A7) holds for $p = O(n)$, i.e., p can be the same size as n.

We establish the asymptotic properties of the penalized estimating equation procedure.

Theorem 1 *Under Assumptions 1–5, there exists an approximate penalized solution* $\widetilde{\theta}$,

which satisfies the selection consistency properties:

$$P(|U_j^{\mathrm{p}}(\widetilde{\theta})| = 0, j \in \mathcal{M}_\theta) \;\; \rightarrow \;\; 1, \tag{12}$$

$$P\left(|U_j^{\mathrm{p}}(\widetilde{\theta})| \leq \frac{\lambda_\theta}{\log n}, j \in \mathcal{M}_\theta^c\right) \;\; \rightarrow \;\; 1, \tag{13}$$

$$P\left(\widetilde{\theta}_{\mathcal{M}_\theta^c} = 0\right) \;\; \rightarrow \;\; 1, \tag{14}$$

and

$$\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta} = O_P(\sqrt{s_\theta/n}), \tag{15}$$

as $n \to \infty$.

Results (12) and (13) imply that $U(\widetilde{\theta}) = o_P\left(\lambda_\theta / \log n\right)$. Results (14) and (15) imply that with probability approaching to one, the penalized estimating equation procedure would not over-select irrelevant variables and estimate the true nonzero coefficients at the $\sqrt{s_\theta/n}$ convergence rate, which is the so-called oracle property of variable selection.

Proof (Theorem 1). We provide a proof for a key step below and defer the complete proof till the supplementary material. The major technical hurdle for the proof is induced by the finite population inference framework, which does not exist in the conventional high-dimensional data setting. To help understand the problem, assume that Sample A is selected under simple random sampling. A major step in the proof is to show that Bernstein's inequality holds for $N^{-1}\sum_{i=1}^N Z_{i,j}$, where $Z_{i,j} = N n_A^{-1} I_{A,i} X_{i,j}$ and $j \in \mathcal{M}_{\alpha^c}$. It is important to note that the $Z_{i,j}$'s $(1 \leq i \leq N)$ are not independent random variables, because $I_{A,i}$ and $I_{A,i'}$ are dependent for any $i \neq i'$. To overcome the technical challenge, we decompose

$$N^{-1}\sum_{i=1}^N Z_{i,j} = N^{-1}\sum_{i=1}^N (W_{i,j} + V_{i,j}),$$

where

$$
\begin{aligned}
W_{1,j} &= \frac{N}{n_A}\left(I_{A,1} - \frac{n_A}{N}\right) X_{1,j}, && V_{1,j} = 0, \\
W_{2,j} &= \frac{N}{n_A}\left(I_{A,2} - \frac{n_A - I_{A,1}}{N - I_{A,1}}\right) X_{2,j}, && V_{2,j} = \frac{N}{n_A}\left(\frac{n_A - I_{A,1}}{N - I_{A,1}} - \frac{n_A}{N}\right) X_{2,j}, \\
&\vdots && \vdots \\
W_{i,j} &= \frac{N}{n_A}\left(I_{A,i} - \frac{n_A - k_i}{N - k_i}\right) X_{i,j}, && V_{i,j} = \frac{N}{n_A}\left(\frac{n_A - k_i}{N - k_i} - \frac{n_A}{N}\right) X_{i,j}, \left(k_i = \sum_{l=1}^{i-1} I_{A,l}\right) \\
&\vdots && \vdots
\end{aligned} \tag{16}
$$

Under Assumptions 4 and 5, $N^{-1}\sum_{i=1}^{N} V_{i,j} \to 0$ as $n_A \to \infty$. Moreover, the key insight is that $\{W_{1,j}, W_{2,j}, \ldots\}$ are martingales, in the sense that $E(W_{i,j} \mid W_{1,j}, \ldots, W_{i-1,j}) = 0$ for all $1 \leq i \leq N$. This enables us to apply the Bernstein's inequality for martingales (Fan et al., 2015) and establish our final results. ■

We now establish the asymptotic properties of $\widehat{\mu}_{\text{p-dr}}(\widehat{\alpha}, \widehat{\beta})$. Define a sequence of events $\mathcal{D}_n = \{\mathcal{M}_\theta \subset \mathcal{C}\}$, where we emphasize that $\mathcal{D}_n$ depends on n although we suppress the dependence of $\mathcal{M}_\theta$ and $\mathcal{C}$ on n. Following the same argument for (15), given the event $\mathcal{D}_n$, we have $\{(\widehat{\alpha} - \alpha^*)^{\mathrm{T}}, (\widehat{\beta} - \beta^*)^{\mathrm{T}}\} = O_p(\sqrt{s_\theta/n})$. Combining with $P(\mathcal{D}_n) \to 1$, we have

$$
\{(\widehat{\alpha} - \alpha^*)^{\mathrm{T}}, (\widehat{\beta} - \beta^*)^{\mathrm{T}}\} = O_p(\sqrt{s_\theta/n}). \tag{17}
$$

By Taylor expansion,

$$
\begin{aligned}
n^{1/2}\left\{\widehat{\mu}_{\text{p-dr}}(\widehat{\alpha},\widehat{\beta})-\mu\right\} &= n^{1/2}\left\{\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)-\mu\right\}+n^{1/2}\left\{\frac{\widehat{\mu}_{\text{p-dr}}(\widehat{\alpha},\widehat{\beta})}{\partial(\alpha^{\mathrm{T}},\beta^{\mathrm{T}})}\right\}\begin{pmatrix}\widehat{\alpha}-\alpha^*\\ \widehat{\beta}-\beta^*\end{pmatrix} \\
&\quad +O_P\left\{n^{1/2}\left\|\begin{pmatrix}\widehat{\alpha}-\alpha^*\\ \widehat{\beta}-\beta^*\end{pmatrix}\right\|_2^2\right\} \\
&= n^{1/2}\left\{\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)-\mu\right\}+O_P\left\{n^{1/2}\left\|\begin{pmatrix}\widehat{\alpha}-\alpha^*\\ \widehat{\beta}-\beta^*\end{pmatrix}\right\|_2^2\right\} \quad (18)\\
&= n^{1/2}\left\{\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)-\mu\right\}+o_p(1), \quad (19)
\end{aligned}
$$

where $\widehat{\mu}_{\text{p-dr}}(\alpha,\beta)$ is defined in (8). Equation (18) follows because we solve (7) for (α,β). Equation (19) follows because of (17) and Assumption 5 (A7). As a result, the way for estimating (α^*,β^*) leads to the asymptotic equivalence between $\widehat{\mu}_{\text{p-dr}}(\widehat{\alpha},\widehat{\beta})$ and $\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)$.

Moreover, we show that $\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)$ is asymptotically unbiased of μ under the double robustness condition. We note that

$$
\begin{aligned}
& n^{1/2}E[\{\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)-\mu\}] \\
= & \frac{n^{1/2}}{N}\sum_{i=1}^{N}E\left[\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\alpha^*)}-1\right\}\{Y_i-m(X_i^{\mathrm{T}}\beta^*)\}+\left(I_{A,i}d_{A,i}-1\right)m(X_i^{\mathrm{T}}\beta^*)\right] \\
= & \frac{n^{1/2}}{N}\sum_{i=1}^{N}\left[E\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\alpha^*)}-1\mid X_i\right\}E\left\{Y_i-m(X_i^{\mathrm{T}}\beta^*)\mid X_i\right\}+E\left\{\left(I_{A,i}d_{A,i}-1\right)m(X_i^{\mathrm{T}}\beta^*\right\}\right] \\
= & \frac{n^{1/2}}{N}\sum_{i=1}^{N}E\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\alpha^*)}-1\mid X_i\right\}E\left\{Y_i-m(X_i^{\mathrm{T}}\beta^*)\mid X_i\right\}. \quad (20)
\end{aligned}
$$

If $\pi_B(X^{\mathrm{T}}\alpha)$ is correctly specified, then $\pi_B(X^{\mathrm{T}}\alpha^*)=\pi_B(X)$ and therefore (20) is zero; if $m(X_i^{\mathrm{T}}\beta)$ is correctly specified, then $m(X_i^{\mathrm{T}}\beta^*)=m(X_i)$ and therefore (20) is zero.

The asymptotic variance of the linearized term is

$$V\left[n^{1/2}\left\{\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)-\mu\right\}\right] = n^{1/2}E\left[V\left\{\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)-\mu \mid I_B,X,Y\right\}\right] + n^{1/2}V\left[E\left\{\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)-\mu \mid I_B,X,Y\right\}\right] := V_1+V_2,$$

where the conditional distribution in $E(\cdot \mid I_B,X,Y)$ and $V(\cdot \mid I_B,X,Y)$ is the sampling distribution for Sample A. The first term V_1 is the sampling variance of the Horvitz–Thompson estimator. Thus,

$$V_1 = E\left\{\frac{n}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{A,ij}-\pi_{A,i}\pi_{A,j})\frac{m(X_i^{\mathrm{T}}\beta^*)}{\pi_{A,i}}\frac{m(X_j^{\mathrm{T}}\beta^*)}{\pi_{A,j}}\right\}. \tag{21}$$

For the second term V_2, note that

$$E\left\{\widehat{\mu}_{\text{p-dr}}(\alpha^*,\beta^*)-\mu \mid I_B,X,Y\right\} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{I_{B,i}}{\pi_{B,i}(X_i^{\mathrm{T}}\alpha^*)}-1\right\}\left\{Y_i-m(X_i^{\mathrm{T}}\beta^*)\right\}.$$

Thus,

$$V_2 = \frac{n}{N^2}\sum_{i=1}^{N}E\left[\left\{\frac{I_{B,i}}{\pi_{B,i}(X_i^{\mathrm{T}}\alpha^*)}-1\right\}^2\left\{Y_i-m(X_i^{\mathrm{T}}\beta^*)\right\}^2\right]. \tag{22}$$

Theorem 2 below summarizes the asymptotic properties of $\widehat{\mu}_{\text{p-dr}}$.

Theorem 2 *Under Assumptions 1–5, if either $\pi_B(X^{\mathrm{T}}\alpha)$ or $m(X^{\mathrm{T}}\beta)$ is correctly specified,*

$$n^{1/2}\left\{\widehat{\mu}_{p\text{-}dr}(\widehat{\alpha},\widehat{\beta})-\mu\right\} \to \mathcal{N}\left(0,V\right),$$

as $n\to\infty$, where $V=\lim_{n\to\infty}(V_1+V_2)$, V_1 and V_2 are defined in (21) and (22), respectively.

To estimate V_1, we can use the design-based variance estimator applied to $m(X_i^{\mathrm{T}}\widehat{\beta})$ as

$$\widehat{V}_1 = \frac{n}{N^2}\sum_{i\in\mathcal{S}_A}\sum_{j\in\mathcal{S}_A}\frac{(\pi_{A,ij}-\pi_{A,i}\pi_{A,j})}{\pi_{A,ij}}\frac{m(X_i^{\mathrm{T}}\widehat{\beta})}{\pi_{A,i}}\frac{m(X_j^{\mathrm{T}}\widehat{\beta})}{\pi_{A,j}}. \tag{23}$$

To estimate V_2, we further express V_2 as

$$V_2 = \frac{n}{N^2}\sum_{i=1}^{N} E\left[\left\{\frac{I_{B,i}}{\pi_{B,i}(X_i^{\mathrm{T}}\alpha^*)^2} - \frac{2I_{B,i}}{\pi_{B,i}(X_i^{\mathrm{T}}\alpha^*)}\right\}\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\}^2 + \{Y_i - m(X_i^{\mathrm{T}}\beta^*)\}^2\right]. \tag{24}$$

Let $\sigma^2(X_i^{\mathrm{T}}\beta^*) = E\left[\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\}^2\right]$, and let $\widehat{\sigma}^2(X_i)$ be a consistent estimator of $\sigma^2(X_i^{\mathrm{T}}\beta^*)$. We can then estimate V_2 by

$$\widehat{V}_2 = \frac{n}{N^2}\sum_{i=1}^{N}\left[\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\widehat{\alpha})^2} - \frac{2I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\widehat{\alpha})}\right\}\left\{Y_i - m(X_i^{\mathrm{T}}\widehat{\beta})\right\}^2 + I_{A,i}d_{A,i}\widehat{\sigma}^2(X_i)\right].$$

By the law of large numbers, $\widehat{V}_2$ is consistent for V_2 regardless whether one of $\pi_{B,i}(X_i^{\mathrm{T}}\alpha)$ or $\pi_{B,i}(X_i^{\mathrm{T}}\beta)$ is misspecificed, and therefore it is doubly robust.

Theorem 3 (Double robustness of $\widehat{V}$) *Under Assumptions 1–5, if either $\pi_B(X^{\mathrm{T}}\alpha)$ or $m(X^{\mathrm{T}}\beta)$ is correctly specified, $\widehat{V} = \widehat{V}_1 + \widehat{V}_2$ is consistent for V.*

6 SIMULATION STUDY

6.1 Setup

In this section, we evaluate the finite-sample performance of the proposed procedure. We first generate a finite population $\mathcal{F}_N = \{(X_i, Y_i) : i = 1, \ldots N\}$ with $N = 10,000$, where X_i is a p-dimensional vector of covariates, and Y_i is a continuous or binary outcome variable. We set $p = 50$. We generate X_i independently from standard normal with mean 0 and variance 1 and set the first component to be 1. From the finite population, we select a non-probability sample $\mathcal{B}$, where we generate the inclusion indicator by $I_{B,i} \sim \mathrm{Ber}(\pi_{B,i})$, and we obtain a probability sample $\mathcal{A}$ of average size $n_A = 500$ under Poisson sampling

with $\pi_{A,i} \propto (0.25 + |X_{1i}| + 0.03|Y_i|)$. The parameter of interest is the population mean $\mu = N^{-1}\sum_{i=1}^{N} Y_i$.

For the non-probability sampling probability, we consider both linear and nonlinear sampling score models

- PSM I: $\text{logit}(\pi_{B,i}) = \alpha_0^{\mathrm{T}} X$, where $\alpha_0 = (-2, 1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$,
- PSM II: $\text{logit}(\pi_{B,i}) = \alpha_0^{\mathrm{T}} \text{sign}(X) \log(X)$, where $\alpha_0 = (-3.5, 1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$.

The sample size n_B is about $2,000$. In both PSM I and PSM II, the first four covariates are important variables.

For generating a continuous outcome variable Y_i, we consider both linear and nonlinear outcome models

- OM I: $Y_i = \beta_0^{\mathrm{T}} X_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$.
- OM II: $Y_i = \beta_0^{\mathrm{T}} \log(X_i^2) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 2.5)$.

For each model, we consider two parameter settings: (a) the first four covariates are important variables with $\beta_0 = (1, 1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$, and (b) the third to sixth variables are important variables with $\beta_0 = (1, 0, 0, 1, 1, 1, 1, 0 \ldots, 0)^{\mathrm{T}}$. In Setting (a), the sampling score model and the outcome model do not have non-overlap covariates; while in Setting (b), the sampling score model and the outcome model have non-overlap covariates.

For generating a binary outcome variable Y_i, we consider both linear and nonlinear outcome models

- OM I: $Y \sim$Ber$\{\pi_Y(X)\}$ with $\text{logit}\{\pi_Y(X)\} = \beta_0^{\mathrm{T}} X$,
- OM II: $Y \sim$Ber$\{\pi_Y(X)\}$ with $\text{logit}\{\pi_Y(X)\} = \beta_0^{\mathrm{T}} \log(X^2)$.

For each model, we consider two parameter settings: (a) the first four covariates are important variables with $\beta_0 = (1/3, -1, -1, 1, 1, 0, \ldots, 0)^{\mathrm{T}} \times 3$, and (b) the third to sixth variables are important variables with $\beta_0 = (1/3, -1, -1, 0, 0, 1, 1, 0 \ldots, 0)^{\mathrm{T}} \times 3$.

We consider the following scenarios for both continuous and binary outcomes: (i) OM I and PSM I, (ii) OM I and PSM II, (iii) OM II and PSM I, (iv) OM II and PSM II.

We consider the following estimators:

1. Naive, $\hat{\mu}_{\text{naive}}$, the naive estimator using the simple average of Y_i from Sample B, which provides the degree of the selection bias;

2. Oracle, $\hat{\mu}_{\text{ora}}$, the doubly robust estimator $\widehat{\mu}_{\text{dr}}(\widehat{\alpha}_{\text{ora}}, \widehat{\beta}_{\text{ora}})$, where $\widehat{\alpha}_{\text{ora}}$ and $\widehat{\beta}_{\text{ora}}$ are based on the joint estimation restricting to the known important covariates for comparison purpose;

3. p-ipw, $\hat{\mu}_{\text{p-ipw}}$, the penalized inverse probability of sampling weighting estimator $\hat{\mu}_{\text{IPW}} = N^{-1} \sum_{i \in \mathcal{B}} \widehat{\pi}_{B,i}^{-1} Y_i$, where $\widehat{\pi}_{B,i} = P(I_{B,i} = 1 \mid X_i^{\mathrm{T}} \widehat{\alpha})$, and $\widehat{\alpha}$ is obtained by a weighted penalized regression of $I_{B,i}$ on X_i using the combined sample of A and B, weighted by the design weights;

4. p-reg, $\hat{\mu}_{\text{p-reg}}$, the penalized regression estimator $\hat{\mu}_{\text{p-reg}} = N^{-1} \sum_{i \in \mathcal{A}} d_{A,i} m(X; \widehat{\beta})$, where $\widehat{\beta}$ is obtained by a penalized regression of Y_i on X_i based on Sample B;

5. p-dr, $\hat{\mu}_{\text{p-dr}}$, the proposed penalized double estimating equation estimator.

We also note that $\widehat{\mu}_{\text{dr}}$ without variable selection is severely biased and unstable and therefore is excluded for comparison.

6.2 Simulation Results

All simulation results are based on 500 Monte Carlo runs. Table 2 reports the selection performance of the proposed penalized estimating equation approach in terms of the proportion of the proposed procedure under-selecting (Under), over-selecting (Over), the average false negatives (FN: the average number of selected covariates that have the true zero coefficients), and the average false positives (FP: the average number of selected covariates that have the true zero coefficients). The proposed procedure selects all covariates with nonzero coefficients in both outcome model and the sampling score model under the true model specification. Moreover, the number of false positives is small under the true model specification.

Figures 1 and 2 display the estimation simulation results for the continuous come with $\beta_0 = (1, 1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$ and $\beta_0 = (1, 0, 0, 1, 1, 1, 1, 0 \ldots, 0)^{\mathrm{T}}$, respectively. The naive estimator $\hat{\mu}_{\text{naive}}$ shows large biases across scenarios. The oracle estimator $\hat{\mu}_{\text{ora}}$ is doubly robust, in the sense that either the outcome or the sampling score is correctly specified, it is unbiased. The penalized inverse probability of sampling weighting estimator $\hat{\mu}_{\text{p-ipw}}$ shows larges biases except for Scenario (ii). The weighted estimator $\widehat{\alpha}$ is based on the blended sample combining Sample A and Sample B, where the units in Sample A are weighted by the known sampling weights and the units in Sample B are weighted by 1. This approach is justifiable only if the sampling rate of Sample B is relatively small compared the the population size. The penalized regression estimator $\hat{\mu}_{\text{p-reg}}$ is only singly robust. When the outcome model is misspecified as in Scenarios (ii) and (iv), it shows large biases. The proposed penalized double estimating equation estimator $\hat{\mu}_{\text{p-dr}}$ is doubly robust, and its performance is comparable to the oracle estimator that requires knowing the true important variables. Moreover, in the case with $\beta_0 = (1, 0, 0, 1, 1, 1, 1, 0 \ldots, 0)^{\mathrm{T}}$, $\hat{\mu}_{\text{p-dr}}$ is slightly more

efficient than $\hat{\mu}_{\text{ora}}$. This efficiency gain is due to using the union of covariates selected for the sampling score model and the outcome model. This phenomenon is consistent with the findings in Brookhart et al. (2006) and Shortreed and Ertefaie (2017).

Figures 3 and 4 display the estimation results for the binary outcome with $\beta_0 = (1/3, -1, -1, 1, 1, 0, \ldots, 0)^{\mathrm{T}} \times 3$ and $\beta_0 = (1/3, -1, -1, 0, 0, 1, 1, 0 \ldots, 0)^{\mathrm{T}} \times 3$, respectively. The same discussion above for the continuous outcome applies here. Moreover, when the outcome model is incorrectly specified, the oracle estimator has a large variability. In this case, the proposed estimator outperforms the oracle estimator, because the variable selection step helps to stabilize the estimation performance.

Table 3 reports the simulation results for the coverage properties for the continuous outcome and binary outcome. Under the double robustness condition (i.e., if either the outcome model or the sampling score model is correctly specified), the coverage rates are close to the nominal coverage; while if both models are misspecified, the coverage rates are off the nominal coverage.

7 AN APPLICATION

We analyze two datasets from the 2005 Pew Research Centre (PRC, http://www.pewresearch.org/) and the 2005 Behavioral Risk Factor Surveillance System (BRFSS). The goal of the PRC study was to evaluate the relationship between individuals and community (Chen et al., 2018, Kim et al., 2018). The 2005 PRC dataset is from a non-probability sample provided by eight different vendors, which consists of $n_B = 9,301$ subjects. We focus on two study variables, a continuous Y_1 (days had at least one drink last month) and a binary Y_2 (an indicator of voted local elections). The 2005 BRFSS sample is a probability sample, which

Table 2: Simulation results for selection performance for the proposed double penalized estimating equation procedure under four scenarios: under OM I (II), the outcome model is correctly specified (misspecified), and under PSM I (II), the probability of sampling score model is correctly specified (misspecified)

	β^*				α^*			
	Under ($\times 10^2$)	Over ($\times 10^2$)	FN	FP	Under ($\times 10^2$)	Over ($\times 10^2$)	FN	FP
Simulation 1: Continuous outcome								
in setting (a) with overlap important variables								
(i) OM I and PSM I	0.0	28.2	0.0	1.2	0.0	0.0	0.0	0.0
(ii) OM II and PSM I	0.0	75.6	0.0	2.6	0.0	0.0	0.0	0.0
(iii) OM I and PSM II	0.0	32.6	0.0	1.3	100.0	100.0	4.0	1.0
(iv) OM II and PSM II	1.8	98.8	0.0	8.0	100.0	100.0	4.0	1.0
in setting (b) with non-overlap important variables								
(i) OM I and PSM I	0.0	31.8	0.0	1.4	0.0	0.0	0.0	0.0
(ii) OM II and PSM I	99.8	68.4	1.9	2.2	0.0	0.0	0.0	0.0
(iii) OM I and PSM II	0.0	32.8	0.0	1.4	100.0	100.0	4.0	1.0
(iv) OM II and PSM II	99.43	93.1	1.8	5.6	100.0	100.0	4.0	1.0
Simulation 2: Binary outcome								
in setting (a) with overlap important variables								
(i) OM I and PSM I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(ii) OM II and PSM I	89.4	0.0	1.5	0.0	0.0	0.0	0.0	0.0
(iii) OM I and PSM II	0.0	0.0	0.0	0.0	100.0	100.0	4.0	1.0
(iv) OM II and PSM II	100.0	0.0	4.0	0.0	100.0	100.0	4.0	1.0
in setting (b) with non-overlap important variables								
(i) OM I and PSM I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(ii) OM II and PSM I	100.0	0.0	3.9	0.0	0.0	0.0	0.0	0.0
(iii) OM I and PSM II	0.0	0.0	0.0	0.0	100.0	100.0	4.0	1.0
(iv) OM II and PSM II	100.0	0.0	4.0	0.0	100.0	100.0	4.0	1.0

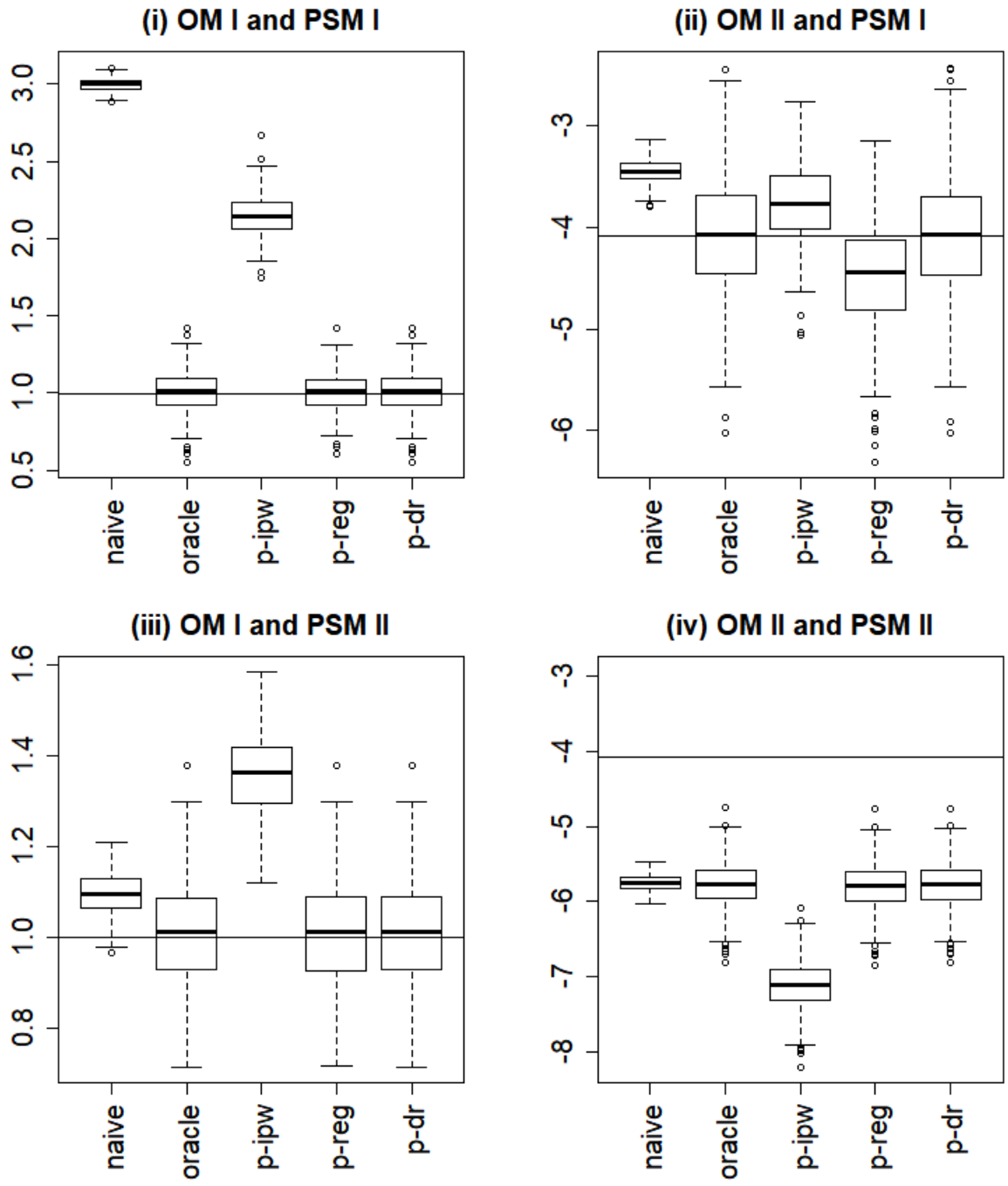

Figure 1: Estimation results for the *continuous outcome* with $\beta_0 = (1, 1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$ under four scenarios: under OM I (II), the outcome model is correctly specified (misspecified), and under PSM I (II), the probability of sampling score model is correctly specified (misspecified) with five estimators: "naive" is the naive estimator using the simple average of outcome from Sample B; "oracle" is the doubly robust estimator with known important covariates; "p-ipw" is the penalized inverse probability of sampling weighting estimator; "p-reg" is the penalized regression estimator; and "p-dr" is the proposed penalized double estimating equation estimator

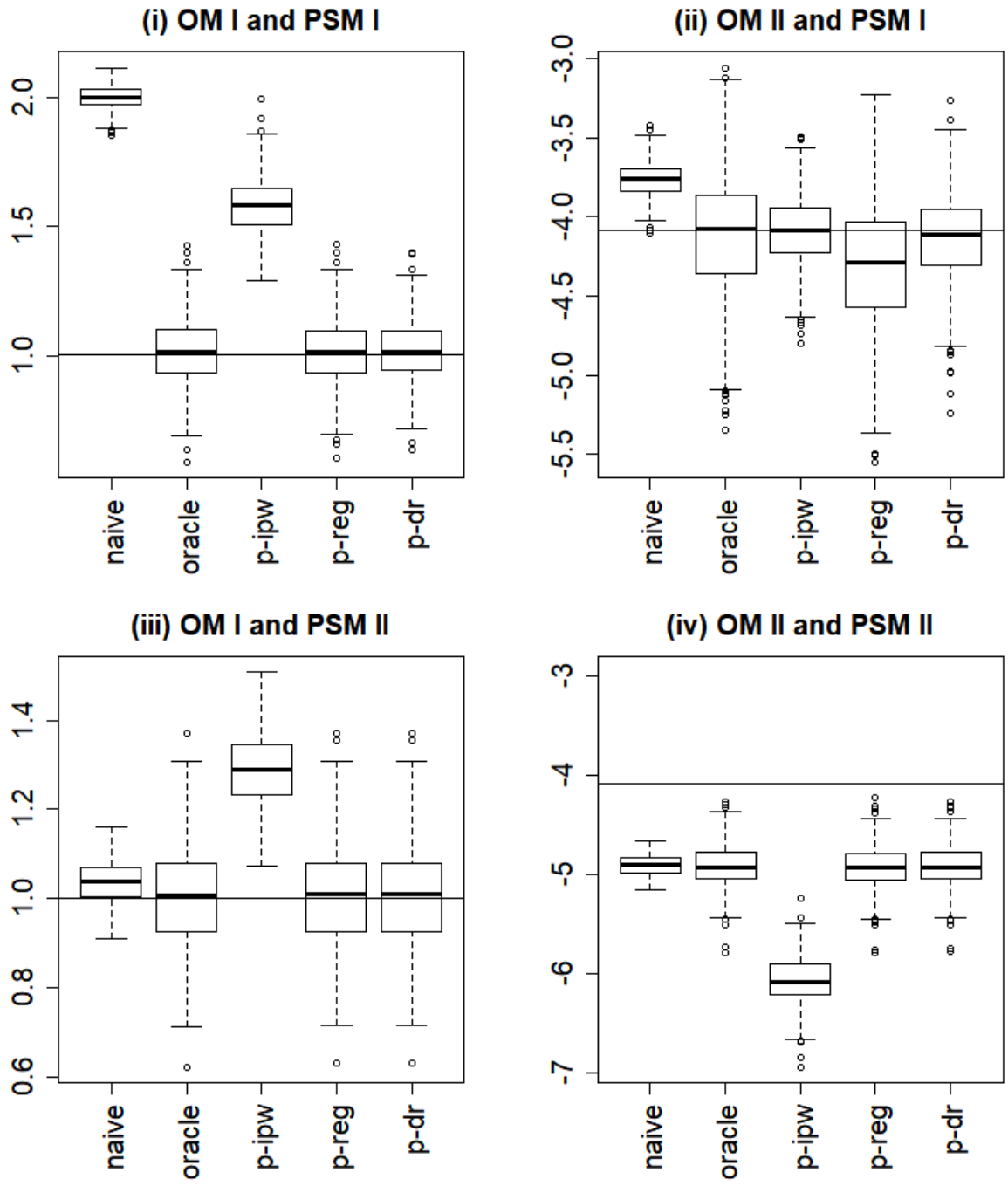

Figure 2: Estimation results for the *continuous outcome* with $\beta_0 = (1, 0, 0, 1, 1, 1, 1, 0 \ldots, 0)^{\mathrm{T}}$ under four scenarios: under OM I (II), the outcome model is correctly specified (misspecified), and under PSM I (II), the probability of sampling score model is correctly specified (misspecified) with five estimators: "naive" is the naive estimator using the simple average of outcome from Sample B; "oracle" is the doubly robust estimator with known important covariates; "p-ipw" is the penalized inverse probability of sampling weighting estimator; "p-reg" is the penalized regression estimator; and "p-dr" is the proposed penalized double estimating equation estimator

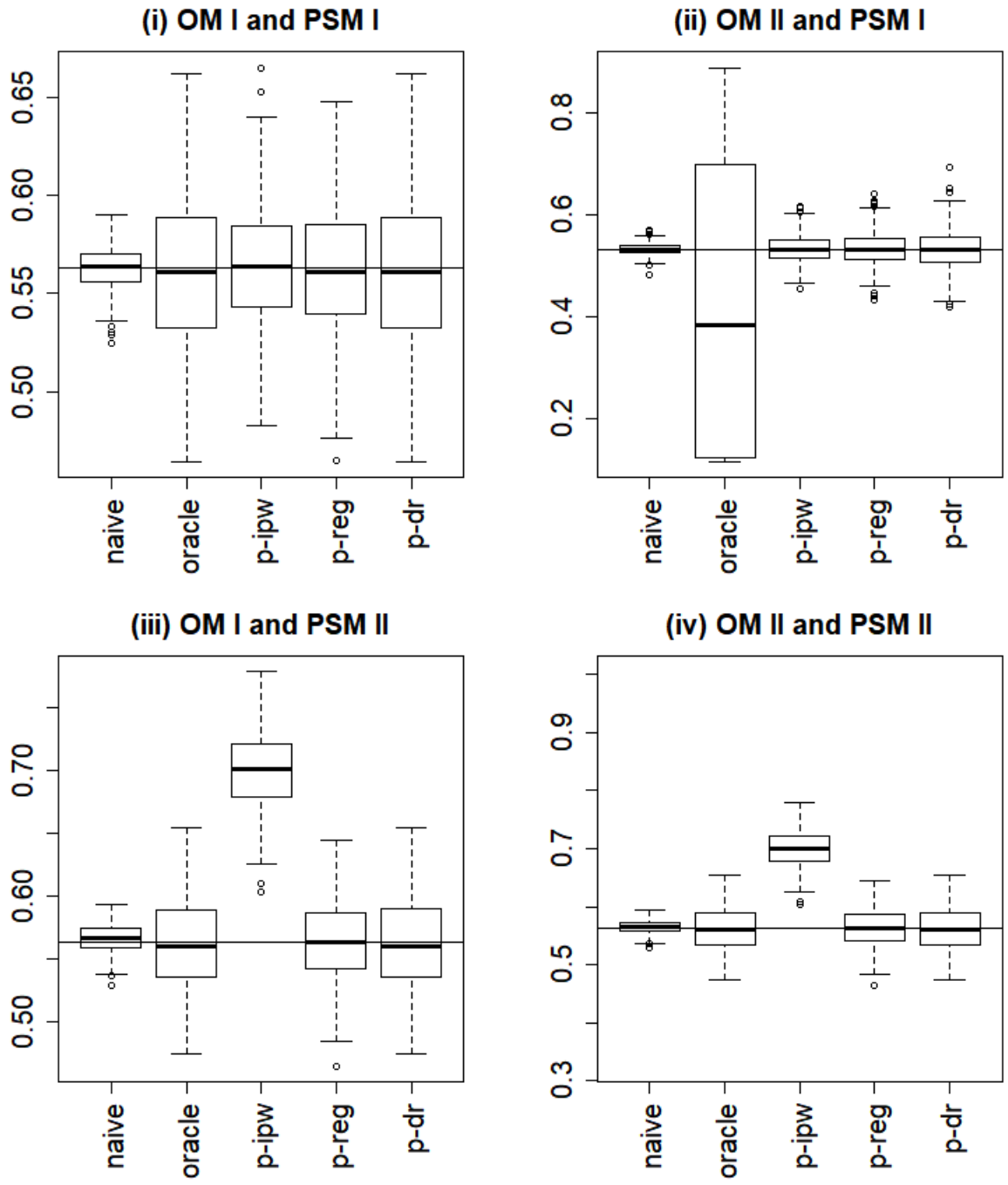

Figure 3: Estimation results for the *binary outcome* with$\beta_0 = (1/3, -1, -1, 1, 1, 0, \ldots, 0)^{\mathrm{T}} \times 3$ under four scenarios: under OM I (II), the outcome model is correctly specified (misspecified), and under PSM I (II), the probability of sampling score model is correctly specified (misspecified) with five estimators: "naive" is the naive estimator using the simple average of outcome from Sample B; "oracle" is the doubly robust estimator with known important covariates; "p-ipw" is the penalized inverse probability of sampling weighting estimator; "p-reg" is the penalized regression estimator; and "p-dr" is the proposed penalized double estimating equation estimator

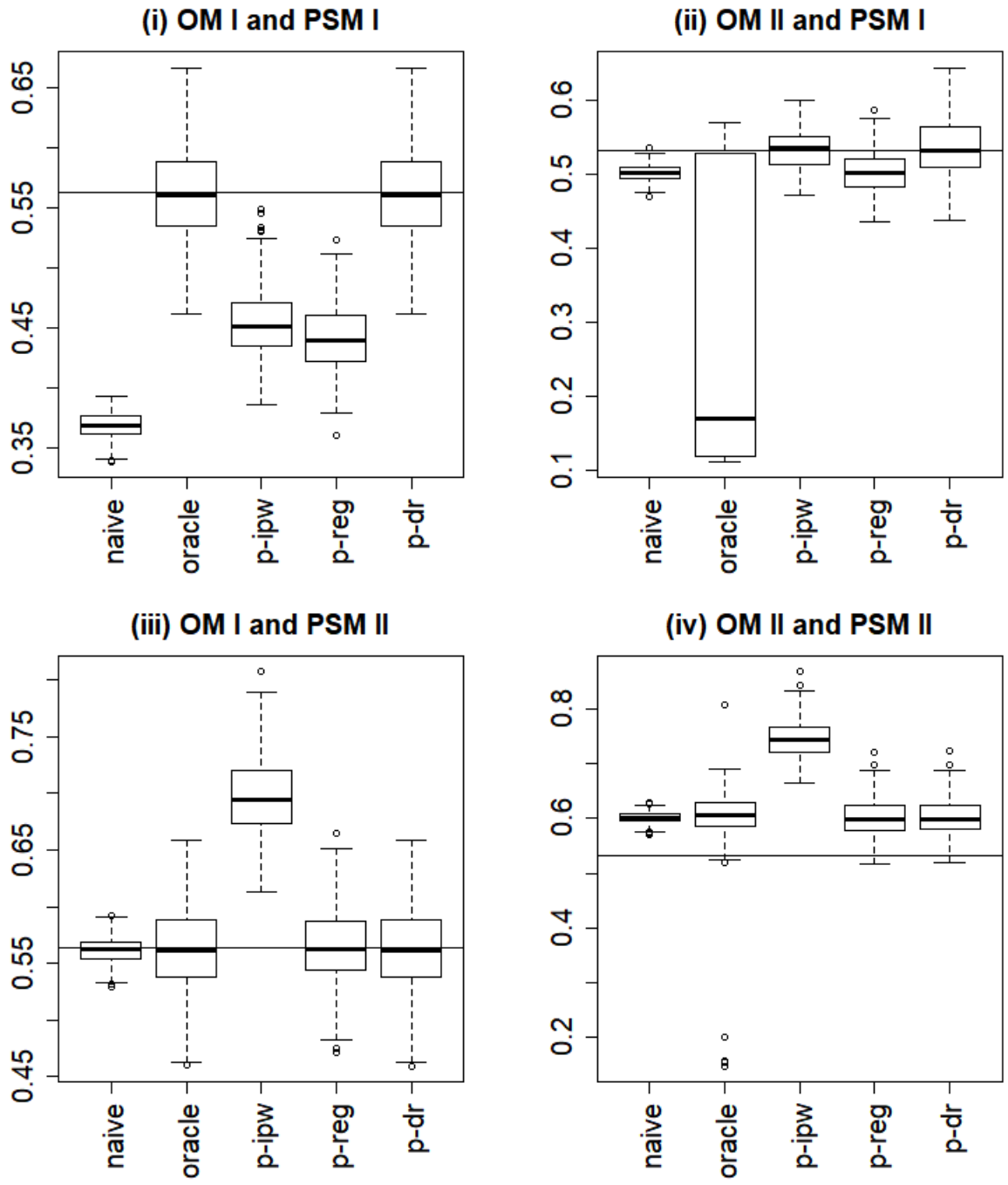

Figure 4: Estimation results for the *binary outcome* with $\beta_0 = (1/3, -1, -1, 0, 0, 1, 1, 0 \ldots, 0)^{\mathrm{T}} \times 3$ under four scenarios: under OM I (II), the outcome model is correctly specified (misspecified), and under PSM I (II), the probability of sampling score model is correctly specified (misspecified) with five estimators: "naive" is the naive estimator using the simple average of outcome from Sample B; "oracle" is the doubly robust estimator with known important covariates; "p-ipw" is the penalized inverse probability of sampling weighting estimator; "p-reg" is the penalized regression estimator; and "p-dr" is the proposed penalized double estimating equation estimator

Table 3: Simulation results for the coverage properties for the continuous and binary outcomes: empirical coverage rate and (empirical coverage rate±2×Monte Carlo standard error)

	The continuous outcome		The binary outcome	
Setting	(a)	(b)	(a)	(b)
(i) OM I and PSM I	95.8 (94.0,97.5)	96.2 (94.5,97.9)	95.8 (94.0,97.6)	95.6 (93.7,97.4)
(ii) OM II and PSM I	95.8 (94.2.97.4)	95.0 (93.0,96.9)	94.2 (92.1,96.2)	93.3 (91.1,95.6)
(iii) OM I and PSM II	93.4 (91.2,95.6)	94.6 (92.6,96.6)	96.6 (95.0,98.5)	96.0 (94.2,97.7)
(iv) OM II and PSM II	0.0 (0.0,0.0)	6.3 (4.1,8.5)	34.5 (30.0,38.9)	38.0 (43.3,52.3)

consists of $n_A = 441,456$ subjects with survey weights. This dataset does not have measurements on the study variables of interest; however, it contains a rich set of common covariates with the PRC dataset listed in Figure 5. To illustrate the heterogeneity in the study populations, Figure 5 contrasts the covariate means from the PRC data and the design-weighted covariate means (i.e., the estimated population covariate means) from the BRFSS dataset. The covariate distributions from the PRC sample and the BRFSS sample are considerably different, e.g., age, education (high school or less), financial status (no money to see doctors, own house), retirement rate, and health (smoking). Therefore, the naive analyses of the study variables based on the PRC dataset are subject to selection biases.

We compute the naive and proposed estimators. To apply the proposed method, we assume the sampling score to be a logistic regression model, the continuous outcome to be a linear regression model, and the binary outcome model to be a logistic regression model adjusting for all available covariates. Using cross validation, the double selection procedure identifies 18 important covariates (all available covariates except for the northeast region) in

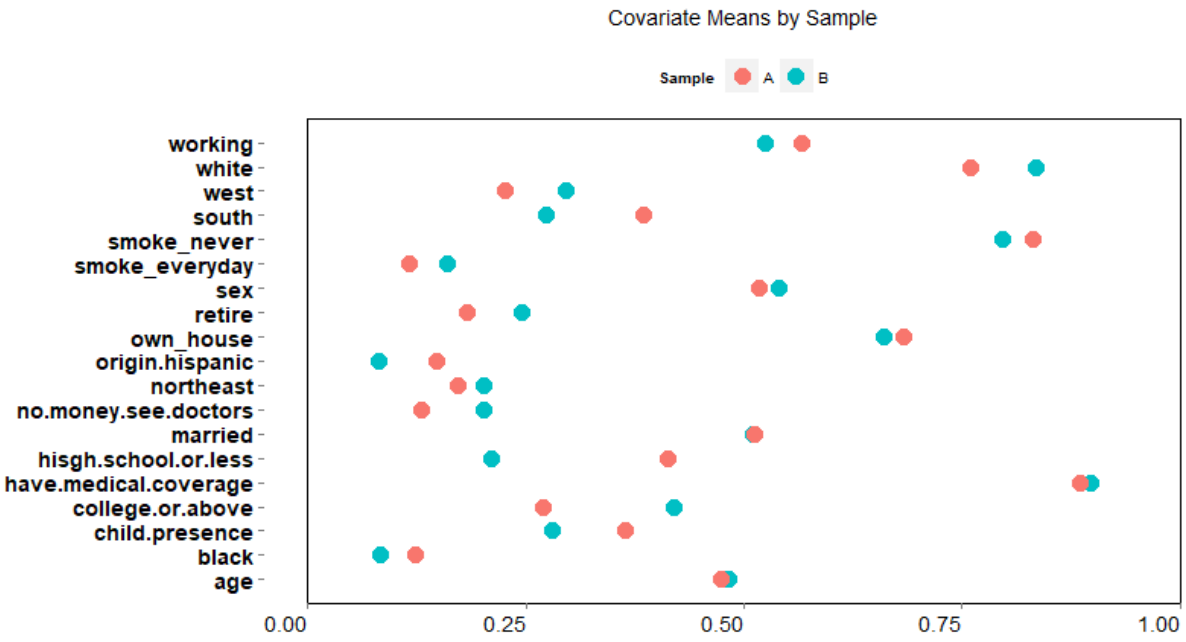

Figure 5: The covariate means by two samples: age is divided by 100

the sampling score and the binary outcome model, and it identifies 15 important covariates (all available covariates except for black, indicator of smoking everyday, the northeast region and the south region).

Table 4 presents the point estimate and the standard error. For estimating the standard error, because the second-order inclusion probabilities are unknown, following the survey literature, we approximate the variance estimator in (23) by assuming the survey design is single-stage Poisson sampling. We find significant differences in the results between our proposed estimator and the corresponding naive estimator. As demonstrated by the simulation in Section 6, the naive estimator may be biased due to selection biases, and the proposed estimator utilizes a probability sample to correct for such biases. From the

results, on average, the target population had at least one drink for 4.84 days over the last month, and 71.8% of the target population voted local elections.

Table 4: Point estimate, standard error and 95% Wald confidence interval

	Y_1 (days had at least one drink last month)			Y_2 (weather voted local elections)		
	Est	SE	CI	Est$\times 10^2$	SE$\times 10^2$	CI$\times 10^2$
Naive	5.36	0.90	(5.17,5.54)	75.3	0.5	(74.4,76.3)
Proposed method	4.84	0.15	(4.81,4.87)	71.8	0.2	(71.3,72.2)

8 CONCLUDING REMARK

Doubly robust estimation is widely used in the literature of missing data and causal inference. We have developed a systematic approach for doubly robust estimation with high dimensional covariates in the context of data integration. The proposed method is based on a two-step approach, where the first step selects important covariates using penalized joint estimating equations, and the second step uses a doubly robust estimator for the population mean. The variable selection procedure enjoys the oracle property under mild regularity conditions. Moreover, by minimizing the asymptotic squared bias term, the effect of first-step selection error is negligible in the doubly robust estimation.

The proposed method is based on Assumption 1 entailing that the sampling mechanism of the non-probability sample is ignorable. If this assumption is in question, we can consider a non-ignorable model for the sampling mechanism and apply a similar two-step method for model selection and parameter estimation. Such extension will be a topic for future research.

SUPPLEMENTARY MATERIAL

Supplementary material provides technical details and proofs. (.pdf)

References

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

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching?, *Social Science Computer Review* **34**: 59–77.

Breidt, F. J., McVey, A. and Fuller, W. A. (1996). Two-phase estimation by imputation, *J. Indian Soc. Agri. Statist.* **49**: 79–90.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Stürmer, T. (2006). Variable selection for propensity score models, *American Journal of Epidemiology* **163**: 1149–1156.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J. and Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights, *J. R. Statist. Soc. A* p. doi: 10.1111/rssa.12357.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data, *Biometrika* **96**: 723–734.

Chen, Y., Li, P. and Wu, C. (2018). Doubly robust inference with non-probability survey samples, *arXiv preprint arXiv:1805.06432* .

Chipperfield, J., Chessman, J. and Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals, *Aust. New Zeal. J. Statist.* **54**: 223–238.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association* **87**: 376–382.

DiSogra, C., Cobb, C., Chan, E. and Dennis, J. M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics, *Joint Statistical Meetings (JSM), Survey Research Methods*, pp. 4501–4515.

Elliott, M. R., Valliant, R. et al. (2017). Inference for nonprobability samples, *Statistical Science* **32**: 249–264.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality, *IEEE Transactions on Information Theory* **57**: 5467–5484.

Fan, X., Grama, I. and Liu, Q. (2015). Exponential inequalities for martingales with applications, *Electronic Journal of Probability* **20**: 1–22.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. et al. (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics* **1**: 302–332.

Fuller, W. A. (2009). *Sampling Statistics*, Wiley, Hoboken, NJ.

Gao, X. and Carroll, R. J. (2017). Data integration with high dimensionality, *Biometrika* **104**: 251–272.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness, *Biometrika* **100**: 417–430.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms, *Annals of Statistics* **33**: 1617–1642.

Johnson, B. A., Lin, D. and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models, *Journal of the American Statistical Association* **103**: 672–680.

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

Keiding, N. and Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys, *J. R. Statist. Soc. A* **179**: 319–376.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling, *Statistica Sinica* **24**: 375–394.

Kim, J. K., Park, S., Chen, Y. and Wu, C. (2018). Combining non-probability and probability survey samples through mass imputation, *arxiv.org/abs/1812.10694* .

Kim, J. K. and Rao, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach, *Biometrika* **99**: 85–100.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors, *Survey Methodology* **32**: 133–142.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment, *Sociological Methods & Research* **37**: 319–343.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election, *The Annals of Applied Statistics* **12**: 685–726.

O'Muircheartaigh, C. and Hedges, L. V. (2014). Generalizing from unrepresentative experiments: a stratified propensity score approach, *J. R. Statist. Soc. C* **63**: 195–210.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

Rivers, D. (2007). Sampling for web surveys, *Proc. Survey Res. Meth. Sect.*, American Statistical Association.

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

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

Stuart, E. A., Bradshaw, C. P. and Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations, *Prevention Science* **16**: 475–485.

Stuart, E. A., Cole, S. R., Bradshaw, C. P. and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials, *J. R. Statist. Soc. A* **174**: 369–386.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*, Springer, New York.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys, *Sociological Methods & Research* **40**: 105–137.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation, *Journal of the American Statistical Association* **110**: 1024–1036.

Vermeulen, K. and Vansteelandt, S. (2016). Data-adaptive bias-reduced doubly robust estimation, *The International Journal of Biostatistics* **12**: 253–282.

Yang, S. and Kim, J. K. (2018). Integration of survey data and big observational data for finite population inference using mass imputation, *arXiv preprint arXiv:1807.02817* .

Supplementary Material for "Doubly Robust Inference when Combining Probability and Non-probability Samples with High-dimensional Data"

Shu Yang, Jae Kwang Kim, and Rui Song

This supplementary material provides technical details and proofs.

S1 BERNSTEIN INEQUALITIES

We first state some useful results.

Lemma S1 (Bernstein inequalities) *1. Let $Z_1, \ldots, Z_N$ be independent zero-mean random variables. Suppose that $|Z_i| \leq M$ almost surely, for all $1 \leq i \leq N$ and some positive constant M. Then, for all $t > 0$,*

$$P\left(\left|\sum_{i=1}^{N} Z_i\right| > t\right) \leq 2\exp\left\{-\frac{2^{-1}t^2}{\sum_{i=1}^{N} E(Z_i^2) + 3^{-1}Mt}\right\}.$$

2. Let $Z_1, \ldots, Z_N$ be independent zero-mean random variables. Suppose that $E(|Z_i|^k) \leq 2^{-1}k!M^{k-2}E(Z_i{}^2)$ for all $k \geq 2$, $1 \leq i \leq N$, and some positive constant M. Then,

$$P\left(\left|\sum_{i=1}^{N} Z_i\right| > t\right) \leq 2\exp\left\{-\frac{2^{-1}t^2}{\sum_{i=1}^{N} E(Z_i^2) + Mt}\right\}.$$

3. *Let $Z_1, \ldots, Z_N$ be possibly non-independent random variables. Suppose that $E(Z_i \mid Z_1, \ldots, Z_{i-1}) = 0$, $E(Z_i^2 \mid Z_1, \ldots, Z_{i-1}) \leq R_i E(Z_i^2)$, $E(Z_i^k \mid Z_1, \ldots, Z_{i-1}) \leq k! M^{k-2} \times R_i E(Z_i^2 \mid Z_1, \ldots, Z_{i-1})/2$ for all $k \geq 2$, $1 \leq i \leq N$, and some positive constant M. Then,*

$$P\left(\left|\sum_{i=1}^{N} Z_i\right| > t\right) \leq 2\exp\left\{-\frac{4^{-1}t^2}{\sum_{i=1}^{N} R_i E(Z_i^2)}\right\},$$

for $0 < t \leq (2M)^{-1}\sqrt{\sum_{i=1}^{N} R_i E(Z_i^2)}$.

S2 PROOF OF THEOREM 1

To simplify the exposition, we introduce more notation. Let $\theta^* = (\alpha^{*\mathrm{T}}, \beta^{*\mathrm{T}})^{\mathrm{T}}$ be the combined parameter values, and let $\mathcal{M}_\theta = \mathcal{M}_\alpha \cup \{p + \mathcal{M}_\beta\}$ be the index set where $\theta_j \neq 0$ for $j \in \mathcal{M}_\theta$. Let $s_\theta = s_\alpha + s_\beta$ and $\lambda_\theta = \min(\lambda_\alpha, \lambda_\beta)$. Define the sets

$$\begin{aligned}
\mathcal{N}_{\theta,\tau} &= \left\{\theta \in \mathbb{R}^{2p} : ||\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}|| \leq \tau\sqrt{s_\theta/n}, \theta_{\mathcal{M}_\theta^c} = 0\right\}, \\
\partial\mathcal{N}_{\theta,\tau} &= \left\{\theta \in \mathbb{R}^{2p} : ||\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}|| = \tau\sqrt{s_\theta/n}, \theta_{\mathcal{M}_\theta^c} = 0\right\},
\end{aligned}$$

for $\tau > 0$.

Step 1: Proof of (15). We show the existence of $\widetilde{\theta}$ by construction. We construct $\widetilde{\theta}$ in a way that $\widetilde{\theta}_{\mathcal{M}_\theta}$ is the oracle solution to $U_{\mathcal{M}_\theta}(\theta)$ and $\widetilde{\theta}_{\mathcal{M}_\theta^c} = 0$.

We show that $\widetilde{\theta}$ satisfies $\widetilde{\theta} - \theta^* = O_P(\sqrt{s_\theta/n})$. Toward this end, we follow Ortega and Rheinboldt (1970) and show that for any $\epsilon > 0$, there exists a $\tau > 0$ such that for all sufficiently large n,

$$P\left\{\sup_{\theta \in \partial\mathcal{N}_{\theta,\tau}} (\theta - \theta^*)^{\mathrm{T}} U(\theta) < 0\right\} \geq 1 - \epsilon. \tag{S1}$$

Because we constrain on $\partial\mathcal{N}_{\theta,\tau}$, we have $(\theta-\theta^*)^{\mathrm{T}}U(\theta) = (\theta_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})^{\mathrm{T}}U_{\mathcal{M}_\theta}(\theta)$. By Taylor expansion,

$$\begin{aligned}(\theta_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})^{\mathrm{T}}U_{\mathcal{M}_\theta}(\theta) &= (\theta_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})^{\mathrm{T}}U_{\mathcal{M}_\theta}(\theta^*) + (\theta_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})^{\mathrm{T}}\nabla_{\mathcal{M}_\theta\mathcal{M}_\theta}(\widetilde{\theta}^*)(\theta_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta}) \\ &:= T_1+T_2,\end{aligned}$$

where $\widetilde{\theta}^*$ satisfies that $\widetilde{\theta}^*_{\mathcal{M}_\theta^c}=0$ and $\widetilde{\theta}^*_{\mathcal{M}_\theta}$ is between $\theta_{\mathcal{M}_\theta}$ and $\theta^*_{\mathcal{M}_\theta}$, and $\nabla(\theta)$ is defined in (10).

Considering T_1, for any $\theta_{\mathcal{M}_\theta}$ such that $||\theta_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta}|| = \tau\sqrt{s_\theta/n}$, by Cauchy-Schwarz inequality, we have

$$|T_1| \le \tau\sqrt{s_\theta/n}||U_{\mathcal{M}_\theta}(\theta^*)||. \tag{S2}$$

Moreover, we have

$$\begin{aligned}E\left\{||U_{\mathcal{M}_\theta}(\theta^*)||^2\right\} &= E\left\{\left|\left|\frac{1}{N}\sum_{i=1}^N I_{B,i}\frac{1-\pi_B(X_i^{\mathrm{T}}\alpha^*)}{\pi_B(X_i^{\mathrm{T}}\alpha^*)}X_{i,\mathcal{M}_\alpha}\right|\right|^2\right\} \\ &\quad +E\left[\left|\left|\frac{1}{N}\sum_{i=1}^N I_{B,i}\left\{Y_i-m(X_i^{\mathrm{T}}\beta^*)\right\}X_{i,\mathcal{M}_\beta}\right|\right|^2\right] \\ &= \mathrm{trace}\left[\frac{1}{N^2}\sum_{i=1}^N\frac{\left\{1-\pi_B(X_i^{\mathrm{T}}\alpha^*)\right\}^2}{\pi_B(X_i^{\mathrm{T}}\alpha^*)}X_{i,\mathcal{M}_\alpha}X_{i,\mathcal{M}_\alpha}^{\mathrm{T}}\right] \\ &\quad +\mathrm{trace}\left[\frac{1}{N^2}\sum_{i=1}^N\pi_B(X_i^{\mathrm{T}}\alpha^*)E\left\{\epsilon_i(\beta^*)^2\mid X_i\right\}X_{i,\mathcal{M}_\beta}X_{j,\mathcal{M}_\beta}^{\mathrm{T}}\right] \\ &\le \frac{1}{N^2}\sum_{i=1}^N C\left\{N^{1-\gamma}s_\alpha\lambda_{\max}\left(X_{i,\mathcal{M}_\alpha}X_{i,\mathcal{M}_\alpha}^{\mathrm{T}}\right)+s_\beta\lambda_{\max}\left(X_{i,\mathcal{M}_\beta}X_{i,\mathcal{M}_\beta}^{\mathrm{T}}\right)\right\} \quad \text{(S3)} \\ &= O(s_\theta/n), \quad \text{(S4)}\end{aligned}$$

where (S3) follows by Assumption 1, Assumption 5 (A3) and (A4), and (S4) follows by Assumption 1 (i) and Assumption 5 (A5). Combining (S2) and (S4), $|T_1| < \tau O_P(s_\theta/n)$.

Considering T_2, we have

$$\begin{aligned} T_2 &= (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} \nabla_{\mathcal{M}_\theta \mathcal{M}_\theta}(\widetilde{\theta}^*)(\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}) \\ &= (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} \nabla_{\mathcal{M}_\theta \mathcal{M}_\theta}(\theta^*)(\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}) \\ &\quad + (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} \left\{ \nabla_{\mathcal{M}_\theta \mathcal{M}_\theta}(\widetilde{\theta}^*) - \nabla_{\mathcal{M}_\theta \mathcal{M}_\theta}(\theta^*) \right\} (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}) \\ &:= T_{21} + T_{22}. \end{aligned}$$

For T_{21}, we have

$$\begin{aligned} T_{21} &= (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} \nabla_{\mathcal{M}_\theta \mathcal{M}_\theta}(\theta^*)(\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}) \\ &\leq -N^{-1} \sum_{i=1}^{N} C \lambda_{\max} \left(X_{i,\mathcal{M}_\theta} X_{i,\mathcal{M}_\theta}^{\mathrm{T}} \right) \left|\left| (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}) \right|\right|^2 \\ &\leq -C\tau^2 (s_\theta / n). \end{aligned}$$

For T_{22}, we have

$$\nabla_{\mathcal{M}_\theta \mathcal{M}_\theta}(\widetilde{\theta}^*) - \nabla_{\mathcal{M}_\theta \mathcal{M}_\theta}(\theta^*) = \begin{pmatrix} \frac{\partial U_{1,\mathcal{M}_\alpha}(\widetilde{\alpha}^*)}{\partial \alpha^{\mathrm{T}}_{\mathcal{M}_\alpha}} - \frac{\partial U_{1,\mathcal{M}_\alpha}(\alpha^*)}{\partial \alpha^{\mathrm{T}}_{\mathcal{M}_\alpha}} & 0 \\ 0 & \frac{\partial U_{2,\mathcal{M}_\beta}(\widetilde{\theta}^*)}{\partial \beta^{\mathrm{T}}_{\mathcal{M}_\beta}} - \frac{\partial U_{2,\mathcal{M}_\beta}(\theta^*)}{\partial \beta^{\mathrm{T}}_{\mathcal{M}_\beta}} \end{pmatrix},$$

where

$$\begin{aligned} \frac{\partial U_{1,\mathcal{M}_\alpha}(\widetilde{\alpha}^*)}{\partial \alpha^{\mathrm{T}}_{\mathcal{M}_\alpha}} - \frac{\partial U_{1,\mathcal{M}_\alpha}(\alpha^*)}{\partial \alpha^{\mathrm{T}}_{\mathcal{M}_\alpha}} &= -\frac{1}{N} \sum_{i=1}^{N} I_{B,i} \left\{ \frac{1 - \pi_B(X_i; \widetilde{\alpha}^*)}{\pi_B(X_i; \widetilde{\alpha}^*)} - \frac{1 - \pi_B(X_i^{\mathrm{T}} \alpha^*)}{\pi_B(X_i^{\mathrm{T}} \alpha^*)} \right\} X_{i,\mathcal{M}_\alpha} X_{i,\mathcal{M}_\alpha}^{\mathrm{T}} \\ &= \frac{1}{N} \sum_{i=1}^{N} I_{B,i} \frac{1 - \pi_B(X_i; \widetilde{\alpha}^{**})}{\pi_B(X_i; \widetilde{\alpha}^{**})} X_{i,\mathcal{M}_\alpha}^{\mathrm{T}} (\widetilde{\alpha}^*_{\mathcal{M}_\alpha} - \alpha^*_{\mathcal{M}_\alpha}) X_{i,\mathcal{M}_\alpha} X_{i,\mathcal{M}_\alpha}^{\mathrm{T}}, \end{aligned}$$

$$\begin{aligned} \frac{\partial U_{2,\mathcal{M}_\beta}(\widetilde{\theta}^*)}{\partial \beta^{\mathrm{T}}_{\mathcal{M}_\beta}} - \frac{\partial U_{2,\mathcal{M}_\beta}(\theta^*)}{\partial \beta^{\mathrm{T}}_{\mathcal{M}_\beta}} &= -\frac{1}{N} \sum_{i=1}^{N} I_{B,i} \left\{ m^{(1)}(X_i^{\mathrm{T}} \widetilde{\beta}^*)^2 - m^{(1)}(X_i^{\mathrm{T}} \beta^*)^2 \right\} X_{i,\mathcal{M}_\beta} X_{i,\mathcal{M}_\beta}^{\mathrm{T}} \\ &= \frac{1}{N} \sum_{i=1}^{N} I_{B,i} 2 m^{(1)}(X_i^{\mathrm{T}} \widetilde{\beta}^{**}) m^{(2)}(X_i^{\mathrm{T}} \widetilde{\beta}^{**}) X_{i,\mathcal{M}_\alpha}^{\mathrm{T}} (\widetilde{\beta}^*_{\mathcal{M}_\theta} - \beta^*_{\mathcal{M}_\theta}) X_{i,\mathcal{M}_\beta} X_{i,\mathcal{M}_\beta}^{\mathrm{T}}, \end{aligned}$$

$\widetilde{\alpha}^{**}$ is between $\widetilde{\alpha}^{*}$ and α^{*}, and $\widetilde{\beta}^{**}$ is between $\widetilde{\beta}^{*}$ and β^{*}. Let

$$B = \sup_{1\le i\le N, k=1,2,3, \theta\in\mathcal{N}_{\theta,\tau}} \left\{ N^{\gamma-1} \left| \frac{1-\pi_B(X_i^{\mathrm{T}}\alpha)}{\pi_B(X_i^{\mathrm{T}}\alpha)} \right|, \left| 2m^{(1)}(X_i^{\mathrm{T}}\beta) m^{(2)}(X_i^{\mathrm{T}}\beta) \right| \right\} \cdot ||X_{i,\mathcal{M}_\theta}||_\infty.$$

Then, we have $B < \infty$ by Assumption 1 and Assumption 5 (A2) and (A4). Therefore, we have

$$\begin{aligned}
|T_{22}| &\le (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} \left| \nabla_{\mathcal{M}_\theta\mathcal{M}_\theta}(\widetilde{\theta}^*) - \nabla_{\mathcal{M}_\theta\mathcal{M}_\theta}(\theta^*) \right| (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}) \\
&\le B \cdot N^{1-\gamma} ||\widetilde{\theta}^*_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}|| \cdot ||\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}||^2 \cdot \lambda_{\max} \left(N^{-1} \sum_{i=1}^{N} X_{i,\mathcal{M}_\theta} X_{i,\mathcal{M}_\theta}^{\mathrm{T}} \right) \\
&\le C \cdot N^{1-\gamma} \sqrt{s_\theta} \left(\tau \sqrt{s_\theta/n} \right)^3 \\
&= \tau^3 o\left(s_\theta/n\right),
\end{aligned}$$

where the last line follows because $n = O(N^\gamma)$ and $N^{1-3\gamma/2} = o(1)$ by Assumption 1.

Then, for a sufficiently large τ, T_{21} dominates $(\theta - \theta^*)^{\mathrm{T}} U(\theta)$ and T_{21} is negative for all sufficiently large n. Therefore, (S1) holds, and as a result, $\widetilde{\theta} - \theta^* = O_P(\sqrt{s_\theta/n})$.

Step 2. Proof of (12). By our construction of $\widetilde{\theta}$, for $j \in \mathcal{M}_\theta$, we have $U_j(\widetilde{\theta}) = 0$. Therefore, to show (12), it suffices to show that $P\left\{ q_{\lambda_\theta}(|\widetilde{\theta}_j|) = 0 : j \in \mathcal{M}_\theta \right\} \to 1$. By (6), it is equivalent to show that $P\left(|\widetilde{\theta}_j| \ge a\lambda_\theta : j \in \mathcal{M}_\theta \right) \to 1$. Note that

$$\begin{aligned}
\min_{j\in\mathcal{M}_\theta} |\widetilde{\theta}_j| &= \min_{j\in\mathcal{M}_\theta} |\theta^*_j + \widetilde{\theta}_j - \theta^*_j| \\
&\ge \min_{j\in\mathcal{M}_\theta} |\theta^*_j| - \max_{j\in\mathcal{M}_\theta} |\widetilde{\theta}_j - \theta^*_j| \\
&\ge \min_{j\in\mathcal{M}_\theta} |\theta^*_j| - ||\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}||.
\end{aligned}$$

Therefore, we have

$$P\left\{ \left(\min_{j\in\mathcal{M}_\theta} |\theta^*_j| - ||\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}|| \right) \ge a\lambda_\theta \right\} = P\left\{ ||\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}|| \le \left(\min_{j\in\mathcal{M}_\theta} |\theta^*_j| - a\lambda_\theta \right) \right\} \to 1,$$

as $\min_{j\in\mathcal{M}_\theta}|\theta^*_j|/\lambda_\theta \to \infty$ and $||\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}|| = o(\lambda_\theta)$. Therefore, $P\left(\min_{j\in\mathcal{M}_\theta}|\widetilde{\theta}_j| \geq a\lambda_\theta\right) \to 1$, as $n \to \infty$.

Proof of (13). By construction of $\widetilde{\theta}$, for $j \in \mathcal{M}^c_\theta$, we have $\widetilde{\theta}_j = 0$ and therefore $q_{\lambda_\theta}(\widetilde{\theta}_j)\text{sign}(\widetilde{\theta}_j) = 0$. To show (13), it suffices to show that

$$P\left\{\max_{j\in\mathcal{M}^c_\theta}|U_j(\widetilde{\theta})| \leq \frac{\lambda_\theta}{\log n}\right\} \to 1. \tag{S5}$$

To show (S5), we define $D_j(\theta) = \partial^2 U_j(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$ and consider the Taylor expansion:

$$U_j(\widetilde{\theta}) = U_j(\theta^*) + \nabla_j(\theta^*)(\widetilde{\theta} - \theta^*) + (\widetilde{\theta} - \theta^*)^{\mathrm{T}} D_j(\widetilde{\theta}^*)(\widetilde{\theta} - \theta^*),$$

where $\widetilde{\theta}^*$ is between $\widetilde{\theta}$ and θ^*. By the definition of $\widetilde{\theta}$, we have $\widetilde{\theta}_{\mathcal{M}^c_\theta} = 0$ and therefore

$$U_j(\widetilde{\theta}) = U_j(\theta^*) + \nabla_{j,\mathcal{M}_\theta}(\theta^*)(\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}) + (\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} D_{j,\mathcal{M}_\theta\mathcal{M}_\theta}(\widetilde{\theta}^*)(\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})$$

We then have

$$\begin{aligned}
P\left\{\max_{j\in\mathcal{M}^c_\theta}|U_j(\widetilde{\theta})| > \frac{\lambda_\theta}{\log n}\right\} \;\leq\; & P\left\{\max_{j\in\mathcal{M}^c_\theta}|U_j(\theta^*)| > \frac{\lambda_\theta}{3\log n}\right\} \\
& + P\left\{\max_{j\in\mathcal{M}^c_\alpha}|\nabla_{j,\mathcal{M}_\theta}(\theta^*)(\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})| > \frac{\lambda_\theta}{3\log n}\right\} \\
& + P\left\{\max_{k\in\mathcal{M}^c_\theta}\left|(\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} D_{j,\mathcal{M}_\theta\mathcal{M}_\theta}(\widetilde{\theta}^*)(\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})\right| > \frac{\lambda_\theta}{3\log n}\right\} \\
=\; & T_3 + T_4 + T_5.
\end{aligned}$$

Therefore, to show (S5), it suffices to show that $T_k = o(1)$ for $k = 3, 4, 5$.

First, we show that $T_3 = o(1)$. We first expand the expression for $U_j(\theta^*)$. For $1 \leq j \leq p$,

$$U_j(\theta^*) = \frac{1}{N}\sum_{i=1}^N\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\alpha^*)} - 1\right\}X_{i,j} - \frac{1}{N}\sum_{i=1}^N\left(\frac{I_{A,i}}{\pi_{A,i}} - 1\right)X_{i,j},$$

and for $p+1 \leq j \leq 2p$,

$$U_j(\theta^*) = \frac{1}{N}\sum_{i=1}^{N} I_{B,i}\left\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\right\} X_{i,j}.$$

Therefore, we have

$$\begin{aligned}
T_3 &= P\left\{\max_{j\in\mathcal{M}_\theta^c} |U_j(\theta^*)| > \frac{\lambda_\theta}{3\log n}\right\} \\
&\leq P\left\{\max_{j\in\mathcal{M}_\alpha^c}\left|\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\alpha^*)} - 1\right\} X_{i,j}\right| > \frac{\lambda_\theta}{9\log n}\right\} \\
&\quad + P\left\{\max_{j\in\mathcal{M}_\alpha^c}\left|\frac{1}{N}\sum_{i=1}^{N}\left(\frac{I_{A,i}}{\pi_{A,i}} - 1\right) X_{i,j}\right| > \frac{\lambda_\theta}{9\log n}\right\} \\
&\quad + P\left\{\max_{j\in\mathcal{M}_\alpha^c}\left|\frac{1}{N}\sum_{i=1}^{N} I_{B,i}\left\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\right\} X_{i,j}\right| > \frac{\lambda_\theta}{9\log n}\right\} \\
&\leq \sum_{j\in\mathcal{M}_\alpha^c} P\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\alpha^*)} - 1\right\} X_{i,j}\right| > \frac{\lambda_\theta}{9\log n}\right\} \\
&\quad + \sum_{j\in\mathcal{M}_\alpha^c} P\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\left(\frac{I_{A,i}}{\pi_{A,i}} - 1\right) X_{i,j}\right| > \frac{\lambda_\theta}{9\log n}\right\} \\
&\quad + \sum_{j\in\mathcal{M}_\alpha^c} P\left\{\left|\frac{1}{N}\sum_{i=1}^{N} I_{B,i}\left\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\right\} X_{i,j}\right| > \frac{\lambda_\theta}{9\log n}\right\} \\
&= T_{31} + T_{32} + T_{33}.
\end{aligned}$$

To evaluate T_{31}, we consider $N^{-1}\sum_{i=1}^{N} Z_{i,j}$, where $Z_{i,j} = \{I_{B,i}/\pi_B(X_i^{\mathrm{T}}\alpha^*) - 1\} X_{i,j}$. Note that the $Z_{i,j}$'s $(1 \leq i \leq N)$ are independent mean zero random variables. By Assumption 1 and Assumption 5 (A2) and (A4), the $Z_{i,j}$'s satisfy the conditions in Lemma S1 (i). By

Bernstein inequality, we have

$$
\begin{aligned}
P\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\frac{I_{B,i}-\pi_B(X_i^{\mathrm{T}}\alpha^*)}{\pi_B(X_i^{\mathrm{T}}\alpha^*)}X_{i,j}\right|>\frac{\lambda_\theta}{9\log n}\right\} &\leq 2\exp\left\{-\frac{\frac{1}{2}\left(\frac{N\lambda_\theta}{9\log n}\right)^2}{\sum_{i=1}^{N}\frac{1-\pi_B(X_i^{\mathrm{T}}\alpha^*)}{\pi_B(X_i^{\mathrm{T}}\alpha^*)}X_{i,j}^2+\frac{1}{3}M\left(\frac{N\lambda_\theta}{9\log n}\right)}\right\} \\
&\leq 2\exp\left\{-C\frac{\left(\frac{N\lambda_\theta}{\log n}\right)^2}{N^{2-\gamma}}\right\} \\
&\leq 2\exp\left\{-Cn\left(\frac{\lambda_\theta}{\log n}\right)\right\}, \qquad \text{(S6)}
\end{aligned}
$$

where the last inequality follows by Assumption 1. To evaluate T_{32}, we consider $N^{-1}\sum_{i=1}^{N}Z_{i,j}$, where $Z_{i,j}=(I_{A,i}/\pi_{A,i}-1)\,X_{i,j}$. We consider two scenarios for the sampling mechanism of Sample A: i) simple random sampling and ii) Poisson sampling. Under Scenario i), the $Z_{i,j}$'s $(1\leq i\leq N)$ are not independent random variables, because $I_{A,i}$ and $I_{A,i'}$ are dependent for any $i\neq i'$. Under simple random sampling ($\pi_{A,i}=n_A/N$ for $1\leq i\leq N$), we construct random variables $\{(W_{i,j},V_{i,j}):1\leq i\leq N,1\leq j\leq p\}$ as in (16). Then, under Assumptions 4 and 5, $N^{-1}\sum_{i=1}^{N}V_{i,j}\to 0$ as $n_A\to\infty$, and $\{W_{1,j},W_{2,j},\ldots\}$ are martingales, in the sense that $E(W_{i,j}\mid W_{1,j},\ldots,W_{i-1,j})=0$ for all $1\leq i\leq N$. Because

$$
\frac{1}{N}\sum_{i=1}^{N}Z_{i,j}=\frac{1}{N}\sum_{i=1}^{N}\left(W_{i,j}+V_{i,j}\right),
$$

we have

$$
\begin{aligned}
P\left(\left|\frac{1}{N}\sum_{i=1}^{N}Z_{i,j}\right|>\frac{\lambda_\theta}{9\log n}\right) &\leq P\left(\left|\frac{1}{N}\sum_{i=1}^{N}W_{i,j}\right|>\frac{\lambda_\theta}{18\log n}\right)+P\left(\left|\frac{1}{N}\sum_{i=1}^{N}V_{i,j}\right|>\frac{\lambda_\theta}{18\log n}\right) \\
&\leq P\left(\left|\frac{1}{N}\sum_{i=1}^{N}W_{i,j}\right|>\frac{\lambda_\theta}{18\log n}\right)+o(1).
\end{aligned}
$$

We consider $P\left\{\left|N^{-1}\sum_{i=1}^{N} W_{i,j}\right| > \lambda_\theta/(18\log n)\right\}$. We verify conditions in Lemma S1 (iii):

$$E(W_{i,j}^2) = \left(\frac{N}{n_A}\right)^2 X_{i,j}^2 \left\{\left(\frac{N-n_A}{N-k_i}\right)^2 \frac{n_A}{N} + \left(\frac{n_A-k_i}{N-k_i}\right)^2 \frac{N-n_A}{N}\right\}$$

$$E(W_{i,j}^2 \mid W_{1,j},\ldots,W_{i-1,j}) = \left(\frac{N}{n_A}\right)^2 X_{i,j}^2 \left(\frac{N-n_A}{N-k_i} \times \frac{n_A-k_i}{N-k_i}\right) \le R_i E(W_{i,j}^2),$$

where

$$\begin{aligned} R_i &= \max_{1\le k\le n_A} \frac{\frac{N-n_A}{N-k} \times \frac{n_A-k}{N-k}}{\left(\frac{N-n_A}{N-k}\right)^2 \frac{n_A}{N} + \left(\frac{n_A-k}{N-k}\right)^2 \frac{N-n_A}{N}} \\ &= \max_{1\le k\le n_A} \left\{\frac{N(n_A-k)}{(N-n_A)n_A + (n_A-k)^2}\right\} \le C. \end{aligned}$$

Moreover, for $k \ge 2$,

$$\begin{aligned} E(|W_{i,j}|^k \mid W_{1,j},\ldots,W_{i-1,j}) &= \left|\left(\frac{N}{n_A}\right) X_{i,j}\right|^k \left\{\left(\frac{N-n_A}{N-k_i}\right)^k \frac{n_A-k_i}{N-k_i} + \left(\frac{n_A-k_i}{N-k_i}\right)^k \frac{N-n_A}{N-k_i}\right\} \\ &= \left|\left(\frac{N}{n_A}\right) X_{i,j}\right|^{k-2} \left\{\left(\frac{N-n_A}{N-k_i}\right)^{k-1} + \left(\frac{n_A-k_i}{N-k_i}\right)^{k-1}\right\} \\ &\quad \times \left(\frac{N}{n_A}\right)^2 X_{i,j}^2 \left(\frac{N-n_A}{N-k_i} \times \frac{n_A-k_i}{N-k_i}\right) \\ &\le 2^{-1} k! M^{k-2} R_i E(W_{i,j}^2 \mid W_{1,j},\ldots,W_{i-1,j}) \end{aligned}$$

for some positive constant M. By Bernstein inequality,

$$\begin{aligned} P\left(\left|N^{-1}\sum_{i=1}^{N} W_{i,j}\right| > \frac{\lambda_\theta}{18\log n}\right) &\le 2\exp\left\{-\frac{\frac{1}{4}\left(\frac{N\lambda_\theta}{18\log n}\right)^2}{\sum_{i=1}^{N} R_i E(W_{i,j}^2)}\right\} \\ &\le 2\exp\left\{-Cn\left(\frac{\lambda_\theta}{\log n}\right)^2\right\}. \end{aligned}$$

Therefore,

$$P\left(\left|N^{-1}\sum_{i=1}^{N} Z_{i,j}\right| > \frac{\lambda_\theta}{9\log n}\right) \leq \exp\left\{-Cn\left(\frac{\lambda_\theta}{\log n}\right)^2\right\}.$$

Under Scenario ii), the $Z_{i,j}$'s $(1 \leq i \leq N)$ are independent mean zero random variables. Similar to (S6), we have

$$\begin{aligned} P\left\{\left|N^{-1}\sum_{i=1}^{N}\left(\frac{I_{A,i}-\pi_{A,i}}{\pi_{A,i}}\right)X_{i,k}\right| > \frac{\lambda_\theta}{9\log n}\right\} &\leq 2\exp\left\{-\frac{1}{2}\frac{\left(\frac{N\lambda_\theta}{9\log n}\right)^2}{\sum_{j=1}^{N}\frac{1-\pi_{A,i}}{\pi_{A,i}}X_{i,k}^2 + \frac{1}{3}M\left(\frac{N\lambda_\theta}{9\log n}\right)}\right\} \\ &\leq 2\exp\left\{-Cn\left(\frac{\lambda_\theta}{\log n}\right)^2\right\}. \end{aligned} \tag{S7}$$

To evaluate T_{33}, we consider $N^{-1}\sum_{i=1}^{N} Z_{i,j}$, where $Z_{i,j} = I_{B,i}\left\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\right\}X_{i,j} = I_{B,i}\epsilon_i(\beta^*)X_{i,j}$. By Assumption 1 and Assumption 5 (A2) and (A4), we have

$$\begin{aligned} E\left(|Z_{i,j}|^k\right) &= E\left[|I_{B,i}\epsilon_i(\beta^*)X_{i,j}|^k\right] \\ &\leq CE\left(|\epsilon_i(\beta^*)|^k\right) \\ &\leq Ck!c_4^{-k}E\left[\exp\left\{c_4|\epsilon_i(\beta^*)|\right\}\right] \qquad \text{(S8)} \\ &\leq Ck!c_4^{-k}c_5 \\ &\leq 2^{-1}k!M^{k-2}\delta, \end{aligned}$$

for some positive constants M and δ, where (S8) follows by Taylor expansion of the exponential function. Therefore, the $Z_{i,j}$'s satisfy the conditions in Lemma S1 (iii). By Bernstein's inequality,

$$P\left\{\left|\frac{1}{N}\sum_{i=1}^{N} I_{B,i}\left\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\right\}X_{i,j}\right| > \frac{\lambda_\theta}{9\log n}\right\} \leq 2\exp\left\{-Cn\left(\frac{\lambda_\theta}{\log n}\right)^2\right\}.$$

Therefore, by Assumption 5 (A7), $\log p = o\{n\lambda_\theta^2/(\log n)^2\}$ and $n\lambda_\theta^2/(\log n)^2 \to \infty$, we have

$$T_3 \le 2\exp\left[\log p - Cn\left(\frac{\lambda_\theta}{\log n}\right)^2\right] = o(1).$$

Second, we show that $T_4 = o(1)$. We have

$$\begin{aligned} T_4 &= P\left\{\max_{j\in\mathcal{M}_\theta^c} |\nabla_{j,\mathcal{M}_\theta}(\theta^*)(\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})| > \frac{\lambda_\theta}{9\log n}\right\} \\ &= P\left\{\max_{j\in\mathcal{M}_\theta^c} |\nabla_{j,\mathcal{M}_\theta}(\theta^*)(\widetilde{\theta}_{\mathcal{M}_\alpha} - \theta^*_{\mathcal{M}_\alpha})| > \frac{\lambda_\theta}{9\log n}, ||\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}|| \le \sqrt{s_\theta/n}\log n\right\} \\ &\quad + P\left\{||\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}|| > \sqrt{s_\theta/n}\log n\right\} \\ &\le P\left\{\max_{j\in\mathcal{M}_\theta^c} ||\nabla_{j,\mathcal{M}_\theta}(\theta^*)|| > \frac{\lambda_\theta\sqrt{n}}{9\sqrt{s_\theta}(\log n)^2}\right\} + o(1), \end{aligned}$$

where $o(1)$ in the last line is because $\widetilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta} = O_P(\sqrt{s_\theta/n})$. To evaluate T_4 further, we note that for $1 \le j \le p$,

$$\nabla_{j,\mathcal{M}_\theta}(\theta^*) = \begin{pmatrix} -\frac{1}{N}\sum_{i=1}^N I_{B,i}\frac{1-\pi_B(X_i^{\mathrm{T}}\alpha^*)}{\pi_B(X_i^{\mathrm{T}}\alpha^*)}X_{i,j}X_{i,\mathcal{M}_\alpha} \\ 0 \end{pmatrix}^{\mathrm{T}},$$

for $p+1 \le j \le 2p$,

$$\nabla_{j,\mathcal{M}_\theta}(\theta^*) = \begin{pmatrix} 0 \\ -\frac{1}{N}\sum_{i=1}^N I_{B,i}m^{(1)}(X_i^{\mathrm{T}}\beta^*)^2 X_{i,j}X_{i,\mathcal{M}_\beta}^{\mathrm{T}} \end{pmatrix}.$$

We then have

$$\begin{aligned} P\left\{\max_{j\in\mathcal{M}_\theta^c} ||\nabla_{j,\mathcal{M}_\theta}(\theta^*)|| > \frac{\lambda_\theta\sqrt{n}}{9\sqrt{s_\theta}(\log n)^2}\right\} &\le P\left\{\max_{j\in\mathcal{M}_\theta^c} ||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2 > \frac{C\lambda_\theta^2 n}{s_\theta(\log n)^4}\right\} \\ &\le P\left\{\max_{j\in\mathcal{M}_\theta^c} \left|||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2 - E||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2\right| > \frac{C\lambda_\theta^2 n}{2s_\theta(\log n)^4}\right\} \\ &\quad + P\left\{\max_{j\in\mathcal{M}_\theta^c} E||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2 > \frac{C\lambda_\theta^2 n}{2s_\theta(\log n)^4}\right\}. \end{aligned}$$

Moreover, by Assumption 1 and Assumption 5 (A1),

$$\max_{j\in\mathcal{M}_\theta^c} E||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2 = \max_{j\in\mathcal{M}_\theta^c} E\left[\sum_{j'\in\mathcal{M}_\theta}\{\nabla_{j,j'}(\theta^*)\}^2\right] \le Cs_\theta \max\left(N^{-\gamma}, N^{\gamma-2}\right) \le Cs_\theta/n.$$

By Assumption 5 (A7), for a sufficiently large n , we have

$$\begin{aligned}
T_4 &\le P\left\{\max_{j\in\mathcal{M}_\theta^c}\left|||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2 - E||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2\right| > \frac{Cn\lambda_\theta^2}{2s_\theta(\log n)^4}\right\} + o(1) \\
&\le \sum_{j\in\mathcal{M}_\theta^c} P\left\{\left|||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2 - E||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2\right| > \frac{Cn\lambda_\theta^2}{2s_\theta(\log n)^4}\right\} + o(1) \\
&\le C\cdot\sum_{j\in\mathcal{M}_\theta^c} \frac{E\left(\sum_{j'\in\mathcal{M}_\theta}\left[\nabla_{j,j'}(\theta^*)^2 - E\left\{\nabla_{j,j'}(\theta^*)^2\right\}\right]\right)^2 s_\theta^2(\log n)^8}{n^2\lambda_\theta^4} + o(1) \qquad \text{(S9)} \\
&= O\left\{p\left(\frac{s_\theta}{N^\gamma}\right)^2 \frac{s_\theta^2(\log n)^8}{n^2\lambda_\theta^4}\right\} \\
&= O\left\{\frac{ps_\theta^4(\log n)^8}{n^4\lambda_\theta^4}\right\} = o(1),
\end{aligned}$$

where (S9) follows by Markov inequality, and $o(1)$ in the last line follows by Assumption 5 (A7).

Third, we show that $T_5 = o(1)$. We have

$$\begin{aligned}
T_5 &\leq P\left\{\max_{j\in\mathcal{M}_\theta^c}\left|(\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})^{\mathrm{T}}D_{j,\mathcal{M}_\theta}(\widetilde{\theta}^*)(\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})\right|>\frac{\lambda_\theta}{3\log n}\right\}\\
&\leq P\left\{\max_{j\in\mathcal{M}_\theta^c}\left|(\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})^{\mathrm{T}}D_{j,\mathcal{M}_\theta}(\widetilde{\theta}^*)(\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})\right|>\frac{\lambda_\theta}{3\log n}, ||\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta}||\leq\sqrt{s_\theta/n}\log n\right\}\\
&\quad +P\left\{||\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta}||>\sqrt{s_\theta/n}\log n\right\}\\
&\leq \sum_{j\in\mathcal{M}_\theta^c}P\left[\mathrm{trace}\left\{D_{j,\mathcal{M}_\theta}(\widetilde{\theta}^*)\right\}>\frac{n\lambda_\theta}{3s_\theta(\log n)^3}\right]+o(1)\\
&\leq c\sum_{j\in\mathcal{M}_\theta^c}\left[\frac{E\left(\left[\mathrm{trace}\left\{D_{j,\mathcal{M}_\theta}(\widetilde{\theta}^*)\right\}\right]^2\right)s_\theta^2(\log n)^6}{n^2\lambda_\theta^2}\right]+o(1), \qquad \text{(S10)}
\end{aligned}$$

where (S10) follows by Markov inequality. Because for $1\leq j\leq p$,

$$D_{j,\mathcal{M}_\theta\mathcal{M}_\theta}(\theta^*)=\begin{pmatrix} -\frac{1}{N}\sum_{i=1}^N I_{B,i}\frac{1-\pi_B(X_i^{\mathrm{T}}\alpha^*)}{\pi_B(X_i^{\mathrm{T}}\alpha^*)}X_{i,j}X_{i,\mathcal{M}_\alpha}X_{i,\mathcal{M}_\alpha}^{\mathrm{T}} & 0\\ 0 & 0\end{pmatrix},$$

and for $p+1\leq j\leq 2p$,

$$D_{j,\mathcal{M}_\theta\mathcal{M}_\theta}(\theta^*)=\begin{pmatrix} 0 & 0\\ 0 & -\frac{1}{N}\sum_{i=1}^N I_{B,i}2m^{(1)}(X_i^{\mathrm{T}}\beta^*)m^{(2)}(X_i^{\mathrm{T}}\beta^*)X_{i,j}X_{i,\mathcal{M}_\beta}X_{i,\mathcal{M}_\beta}^{\mathrm{T}}\end{pmatrix},$$

by Assumption 5 (A1), (A4), (A5) and (A6), we have

$$E\left(\left[\mathrm{trace}\left\{D_{j,\mathcal{M}_\theta}(\widetilde{\theta}^*)\right\}\right]^2\right)\leq Cs_\theta^2/N^\gamma,$$

for all j. Therefore, $T_5 = O\left\{ps_\theta^4(\log n)^6/(n^3\lambda_\theta^2)\right\}+o(1)=o(1)$.

Combining all results together, we complete the proof for Theorem 1.

S3 PROOF OF (17)

We outline the proof for that on the event $\mathcal{D}_n$, $\{(\widehat{\alpha}-\alpha^*)^{\mathrm{T}}, (\widehat{\beta}-\beta^*)^{\mathrm{T}}\} = O_p(\sqrt{s_\theta/n})$. Without further mentioning, we now constrain the parameters and estimators by $\theta^*_{\mathcal{C}^c} = 0$ and $\widehat{\theta}_{\mathcal{C}^c} = 0$. On the event $\mathcal{D}_n$, $\mathcal{C}$ contains all indexes for the true important covariates. We construct $\widehat{\theta}$ such that $\widehat{\theta}_{\mathcal{M}_\theta}$ is the oracle solution to $J_{\mathcal{M}_\theta}(\theta)$ and $\widehat{\theta}_{\mathcal{M}_\theta^c} = 0$ and show that $\widehat{\theta}$ satisfies $\widehat{\theta} - \theta^* = O_P(\sqrt{s_\theta/n})$.

Toward this end, we follow the proof in Section S2 and show that for any $\epsilon > 0$, there exists a $\tau > 0$ such that for all sufficiently large n,

$$P\left\{\sup_{\theta\in\partial\mathcal{N}_{\theta,\tau}} (\theta-\theta^*)^{\mathrm{T}} J(\theta) < 0\right\} \geq 1-\epsilon. \tag{S11}$$

Because we constrain on $\partial\mathcal{N}_{\theta,\tau}$, we have $(\theta-\theta^*)^{\mathrm{T}} J(\theta) = (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} J_{\mathcal{M}_\theta}(\theta)$. By Taylor expansion,

$$(\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} J_{\mathcal{M}_\theta}(\theta) \;=\; (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} J_{\mathcal{M}_\theta}(\theta^*) + (\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})^{\mathrm{T}} \nabla^J_{\mathcal{M}_\theta\mathcal{M}_\theta}(\widetilde{\theta}^*)(\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta})$$

where $\widetilde{\theta}^*$ satisfies that $\widetilde{\theta}^*_{\mathcal{M}_\theta^c} = 0$ and $\widetilde{\theta}^*_{\mathcal{M}_\theta}$ is between $\theta_{\mathcal{M}_\theta}$ and $\theta^*_{\mathcal{M}_\theta}$, and $\nabla^J(\theta) = \partial J(\theta)/\partial\theta^{\mathrm{T}}$. Following the same argument as in Section S2, (S11) holds, and as a result, on the event $\mathcal{D}_n$, $\widehat{\theta} - \theta^* = O_P(\sqrt{s_\theta/n})$. Combining with $P(\mathcal{D}_n) \to 1$, (17) holds.