ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 2, *pp.* 445–465

Doubly robust inference when combining probability and non-probability samples with high dimensional data

Shu Yang,

North Carolina State University, Raleigh, USA

Jae Kwang Kim,

Iowa State University, Ames, USA

and Rui Song

North Carolina State University, Raleigh, USA

[Received March 2019. Final revision November 2019]

Summary. 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 and show selection consistency for general samples. 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 analysing probability sample data such as data collection costs and increasing non-response rates (Keiding and Louis, 2016). With advances of technology, non-probability samples become increasingly available for research purposes, such as remote sensing data and web-based volunteer samples. Non-probability samples 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 sources of data.

Address for correspondence: Shu Yang, Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.
E-mail: syang24@ncsu.edu

 1369–7412/20/82445

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 modelled 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, for example, Lee and Valliant (2009), Valliant and Dever (2011), Elliott and Valliant (2017) and Chen, Li and Wu (2018). Stuart *et al.* (2011, 2015) and Buchanan *et al.* (2018) used propensity score weighting to generalize results from randomized trials to a target population. O'Muircheartaigh and Hedges (2014) proposed propensity score stratification for analysing a non-randomized social experiment. One notable disadvantage of propensity score methods is that they rely on an explicit propensity score model and may be 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; Chen, Valliant and Elliott, 2018; Chen *et al.*, 2019). This technique forces the moments or the empirical distribution of auxiliary variables to be the same between the probability sample and the non-probability sample, so that after calibration the weighted distribution in the non-probability sample appears similar to that in 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 data set for developing an imputation model. In the mass imputation, the non-probability sample is used as a training data set, and imputation is applied to all units in the probability sample; see, for example, Breidt *et al.* (1996), Rivers (2007), Kim and Rao (2012), Chipperfield *et al.* (2012) 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 sources of data, and let Y be a general-type study variable of interest. We consider combining a probability sample observing X, referred to as sample A, and a non-probability sample observing (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 uses the non-informative sampling assumption imposed on the non-probability sample. To ensure that this assumption holds, researchers try to control for all covariates that are predictors of both sampling and the outcome variable. In practice, subject matter experts recommend a rich set of potentially useful variables but typically will not identify the set of variables to adjust for. In the presence of a large number of auxiliary variables, variable selection is 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 source of data and utilize all information captured for finite population inference. Gao and Carroll (2017) proposed a pseudolikelihood approach to combining multiple non-survey data with high dimensionality; this approach requires that all likelihoods are correctly specified and therefore is sensitive to model misspecification. Chen, Valliant and Elliott (2018) proposed a model-based calibration approach using lasso regression; this approach relies on a correctly specified outcome model. To our knowledge, 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 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, which is referred to as the sampling score, or the outcome model is correctly specified, but not necessarily both (a

double-robustness condition); see, for example, Bang and Robins (2005), Tsiatis (2006), Cao *et al.* (2009) and Han and Wang (2013). To handle high dimensional covariates, our strategy separates the variable selection step and the estimation step for the 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 using penalized estimating equations. We assume that the sampling score follows a logistic regression model with unknown parameter $\alpha \in \mathbb{R}^p$ and the outcome follows a generalized linear model with unknown parameter $\beta \in \mathbb{R}^p$. Importantly, we separate the estimating equations for α and β 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 weighted average of X from sample A (i.e. a design-unbiased 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. The major technical challenge is that, under the finite population framework, the selection indicators of sample A are not independent in general. To overcome this challenge, we construct martingale random variables with a weak dependence that enables the application of the Bernstein inequality. This construction is used in establishing our selection consistency result.

In the second step, we re-estimate (α, β) on the basis of the joint set of covariates selected from the first step and consider a doubly robust estimator of μ, $\hat{\mu}_{\mathrm{dr}}(\hat{\alpha}, \hat{\beta})$. We propose to use different estimating equations for (α, β), derived by minimizing the asymptotic squared bias of $\hat{\mu}_{\mathrm{dr}}(\hat{\alpha}, \hat{\beta})$. This estimation strategy is not new; see, for example, Kim and Haziza (2014) for missing data analyses in low dimensional data; here, 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 series expansion of $\hat{\mu}_{\mathrm{dr}}(\hat{\alpha}, \hat{\beta})$ with respect to (α, β) to be exactly zero, and the remaining terms are negligible under regularity conditions. This estimating strategy makes the doubly robust estimator root n consistent if either the sampling probability or the outcome model is correctly specified. This also enables us to construct a simple and consistent variance estimator allowing for doubly robust inferences. Importantly, the estimator proposed enables model misspecification of either the sampling score or the outcome model. In the existing high dimensional causal inference literature, the doubly robust estimators have been shown to be robust to selection errors by using penalization (Farrell, 2015) or approximation errors by using machine learning (Chernozhukov *et al.*, 2018). However, this double-robustness feature requires both nuisance models to be correctly specified. We relax this requirement by allowing one of the nuisance models to be misspecified. We clarify that, even though the set of variables for estimation may include the variables that are solely related to the sampling score but not the outcome and therefore may harm efficiency of estimating μ (De Luna *et al.*, 2011; Patrick *et al.*, 2011), it is important to include these variables for $\hat{\mu}_{\mathrm{dr}}(\hat{\alpha}, \hat{\beta})$ to achieve consistency in the case when the outcome model is misspecified and the sampling score model is correctly specified; see Section 6.

The paper proceeds as follows. Section 2 provides the basic set-up of the problem. 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 method. In Section 7, we present an application to analyse a non-probability sample collected by the Pew Research Center (PRC). We relegate all proofs to the on-line supplementary material.

2. Basic set-up

2.1. Notation: two samples

Let $\mathcal{U}=\{1,\ldots,N\}$ be the index set of N units for the finite population with N known. The finite population consists of $\mathcal{F}_N=\{(X_i,Y_i): i\in\mathcal{U}\}$. The parameter of interest is the finite population mean $\mu=N^{-1}\Sigma_{i=1}^{N}Y_i$. We consider two sources of data: a probability sample, referred to as sample A, and a non-probability sample, referred to as sample B. Table 1 illustrates the observed data structure. Sample A consists of observations $\mathcal{O}_{\mathrm{A}}=\{(d_{\mathrm{A},i}=\pi_{\mathrm{A},i}^{-1},X_i): i\in\mathcal{A}\}$ with sample size n_{A}, where $\pi_{\mathrm{A},i}=P(i\in\mathcal{A})$ is known in sample A. Sample B consists of observations $\mathcal{O}_{\mathrm{B}}=\{(X_i,Y_i): i\in\mathcal{B}\}$ with sample size n_{B}. We define $I_{\mathrm{A},i}$ and $I_{\mathrm{B},i}$ to be the selection indicators corresponding 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–Thompson estimation. The naive estimators applied to sample B without adjusting for the sampling process are subject to selection biases (Meng, 2018).

2.2. An identification assumption

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

Assumption 1.

(a) The selection indicator I_{B} of sample B and the response variable Y are independent given X, i.e. $P(I_{\mathrm{B}}=1|X,Y)=P(I_{\mathrm{B}}=1|X)$, which is referred to as the sampling score $\pi_{\mathrm{B}}(X)$, and
(b) $\pi_{\mathrm{B}}(X)>N^{\gamma-1}\delta_{\mathrm{B}}>0$ for all X, where $\gamma\in(\frac{2}{3},1]$.

Assumption 1(a) implies that $m(X)=E(Y|X)=E(Y|X,I_{\mathrm{B}}=1)$ can be estimated solely on the basis of sample B. Assumption 1(b) specifies a lower bound of $\pi_{\mathrm{B}}(X)$. A standard condition in the literature imposes a strict positivity in the sense that $\pi_{\mathrm{B}}(X)>\delta_{\mathrm{B}}>0$; however, it implies that $n_{\mathrm{B}}^{-1}=O(N^{-1})$, which may be restrictive in survey practice. Here, we relax this condition and allow $n_{\mathrm{B}}^{-1}=O(N^{-\gamma})$, where γ can be strictly less than 1.

Assumption 1 is a key assumption for identification. Under assumption 1, $E(\mu)$ is identifiable on the basis of sample A by $E\{I_{\mathrm{A}}m(X)\}$ or sample B by $E\{I_{\mathrm{B}}Y/\pi_{\mathrm{B}}(X)\}$. However, this

Table 1. Two sources of data†

Sample		*Sampling weight* π^{-1}	*Covariate* X	*Study variable* Y
Probability sample	1	√	√	?
	⋮	⋮	⋮	⋮
	n_{A}	√	√	?
Non-probability sample	$n_{\mathrm{A}}+1$	?	√	√
	⋮	⋮	⋮	⋮
	$n_{\mathrm{A}}+n_{\mathrm{B}}$	?	√	√

†Sample A is a probability sample, and sample B is a non-probability sample. '√' and '?' indicate observed and unobserved data respectively.

assumption is not verifiable from the observed data. To ensure that this assumption holds, researchers often consider many possible predictors for the selection 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 $\pi_{\mathrm{B}}(X)$ and the outcome mean function $m(X)$ are unknown and need to be estimated from the data. Let $\pi_{\mathrm{B}}(X^{\mathrm{T}}\alpha)$ and $m(X^{\mathrm{T}}\beta)$ be the postulated models for $\pi_{\mathrm{B}}(X)$ and $m(X)$ respectively, where α and β are unknown parameters. Various estimators of μ have been proposed in the literature, each requiring different model assumptions and estimation strategies. We provide examples below and discuss their properties and limitations.

2.3.1. Example 1 (inverse probability of sampling score weighting)

Given an estimator $\hat{\alpha}$, the inverse probability of sampling score weighting estimator is

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

The justification for $\hat{\mu}_{\mathrm{IPW}}$ relies on a correct specification of $\pi_{\mathrm{B}}(X)$ and the consistency of $\hat{\alpha}$. There are different approaches to obtain $\hat{\alpha}$. Following Valliant and Dever (2011), we can obtain $\hat{\alpha}$ by fitting the sampling score model on the basis of the combined data $\mathcal{O}_{\mathrm{A}} \cup \mathcal{O}_{\mathrm{B}} = \{(\omega_i = d_{\mathrm{A},i}, X_i, I_i = 0) : i \in \mathcal{A}\} \cup \{(\omega_i = 1, X_i, I_i = 1) : i \in \mathcal{B}\}$, weighted by ω_i. The resulting estimator $\hat{\alpha}$ is valid if n_{B} is relatively small (Valliant and Dever, 2011). Elliott and Valliant (2017) proposed an alternative strategy based on the Bayes rule: $\pi_{\mathrm{B}}(X) \propto P(I_{\mathrm{A}} = 1|X) O_{\mathrm{B}}(X)$, where $O_{\mathrm{B}}(X) = P(I_{\mathrm{B}} = 1|X, \mathcal{O}_{\mathrm{A}} \cup \mathcal{O}_{\mathrm{B}}) / P(I_{\mathrm{B}} = 0|X, \mathcal{O}_{\mathrm{A}} \cup \mathcal{O}_{\mathrm{B}})$ is the odds of selection into sample B among the combined sample. This approach does not require n_{B} to be small; however, if X does not correspond to the design variables for sample A, it requires postulating an additional model for $P(I_{\mathrm{A}} = 1|X)$. Moreover, variable selection based on this approach is not straightforward with a high dimensional X. To obtain $\hat{\alpha}$, we use the following estimating equation for α:

$$\sum_{i=1}^{N} \left\{ \frac{I_{\mathrm{B},i}}{\pi(X_i^{\mathrm{T}}\alpha)} - \frac{I_{\mathrm{A},i}}{\pi_{\mathrm{A},i}} \right\} h(X_i; \alpha) = 0, \tag{2}$$

for some $h(X_i; \alpha)$ such that equation (2) has a unique solution. Kott and Liao (2017) advocated the use of $h(X; \alpha) = X$ and Chen, Li and Wu (2018) advocated the use of $h(X; \alpha) = \pi(X^{\mathrm{T}}; \alpha)X$.

2.3.2. Example 2 (outcome regression based on sample A)

The outcome regression estimator is

$$\hat{\mu}_{\mathrm{reg}} = \hat{\mu}_{\mathrm{reg}}(\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N} I_{\mathrm{A},i} d_{\mathrm{A},i} m(X_i^{\mathrm{T}}\hat{\beta}), \tag{3}$$

where $\hat{\beta}$ is obtained by fitting the outcome model based solely on $\mathcal{O}_{\mathrm{B}}$ under assumption 1.

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

2.3.3. Example 3 (calibration weighting)

The calibration weighting estimator is

$$\hat{\mu}_{\mathrm{cal}} = \hat{\mu}_{\mathrm{cal}} = \frac{1}{N}\sum_{i=1}^{N} \omega_i I_{\mathrm{B},i} Y_i, \tag{4}$$

where $\{\omega_i : i \in \mathcal{B}\}$ satisfies constraint (i) $\Sigma_{i\in\mathcal{S}_\mathrm{B}}\omega_i X_i = \Sigma_{i\in\mathcal{S}_\mathrm{A}} d_{\mathrm{A},i} X_i$ or constraint (ii) $\Sigma_{i\in\mathcal{S}_\mathrm{B}}\omega_i m(X_i;\hat{\beta}) = \Sigma_{i\in\mathcal{A}} d_{\mathrm{A},i} m(X_i;\hat{\beta})$ (McConville *et al.*, 2017; Chen, Valliant and Elliott, 2018; Chen *et al.*, 2019).

The justification for $\hat{\mu}_{\mathrm{cal}}$ subject to constraint (i) relies on the linearity of the outcome model, i.e. $m(X) = X^{\mathrm{T}}\beta^*$ for some β^*, or the linearity of the inverse probability of sampling weight, i.e. $\pi_\mathrm{B}(X)^{-1} = X^{\mathrm{T}}\alpha^*$ for some α^* (Fuller (2009), theorem 5.1). The linearity conditions are unlikely to hold for non-continuous variables. In these cases, $\hat{\mu}_{\mathrm{cal}}$ may be biased. The justification for $\hat{\mu}_{\mathrm{cal}}$ subject to constraint (ii) relies on a correct specification of $m(X;\beta)$.

2.3.4. Example 4 (doubly robust estimator)
The doubly robust estimator is

$$\hat{\mu}_{\mathrm{dr}} = \hat{\mu}_{\mathrm{dr}}(\hat{\alpha},\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{I_{\mathrm{B},i}}{\hat{\pi}_\mathrm{B}(X_i^{\mathrm{T}}\hat{\alpha})}\{Y_i - m(X_i^{\mathrm{T}}\hat{\beta})\} + I_{\mathrm{A},i} d_{\mathrm{A},i} m(X_i^{\mathrm{T}}\hat{\beta})\right]. \tag{5}$$

The estimator $\hat{\mu}_{\mathrm{dr}}$ is doubly robust with fixed dimensional X (Chen, Li and Wu, 2018), in the sense that it achieves consistency if either $\pi_\mathrm{B}(X^{\mathrm{T}}\alpha)$ 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 $\hat{\mu}_{\mathrm{dr}}$ in a high dimensional set-up.

3. Methodology in high dimensional data

In the presence of a large number of covariates, not all of them are relevant for making inference of the population mean of the outcome. Including unnecessary covariates in the model makes the computation unstable and increases estimation errors. Variable selection is required to handle high dimensional covariates. For any vector $\alpha \in \mathbb{R}^p$, denote the number of non-zero elements in α as $\|\alpha\|_0 = \Sigma_{j=1}^p I(\alpha_j \neq 0)$, the L_1-norm as $\|\alpha\|_1 = \Sigma_{j=1}^p |\alpha_j|$, the L_2-norm as $\|\alpha\|_2 = \sqrt{\Sigma_{j=1}^p \alpha_j^2}$ and the L_∞-norm as $\|\alpha\|_\infty = \max_{1\leqslant j\leqslant p}|\alpha_j|$. For any $\mathcal{J} \subseteq \{1,\ldots,p\}$, let $\alpha_\mathcal{J}$ be the subvector of α formed by elements of α whose indices are in $\mathcal{J}$. Let $\mathcal{J}^{\mathrm{c}}$ 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 submatrix of Σ formed by rows in $\mathcal{J}_1$ and columns in $\mathcal{J}_2$. Following the literature on variable selection, we first standardize the covariates so that they have variances approximately equal to 1, which makes the variable-selection procedure more stable. We make the following modelling assumptions.

Assumption 2 (sampling score model). The sampling mechanism of sample B, $\pi_\mathrm{B}(X)$, follows a logistic regression model $\pi_\mathrm{B}(X^{\mathrm{T}}\alpha)$, i.e. $\mathrm{logit}\{\pi_\mathrm{B}(X^{\mathrm{T}}\alpha)\} = X^{\mathrm{T}}\alpha$ for $\alpha \in \mathbb{R}^p$.

Assumption 3 (outcome model). The outcome mean function $m(X)$ follows a generalized linear regression model, i.e. $m(X) = m(X^{\mathrm{T}}\beta)$ for $\beta \in \mathbb{R}^p$, where $m(\cdot)$ denotes the link function.

Define α^* to be the p-dimensional parameter that minimizes the Kullback–Leibler divergence,

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

and $\beta^* = \arg\min_{\beta\in\mathbb{R}^p} E[\{Y - m(X^{\mathrm{T}}\beta)\}^2]$.

In assumption 2, we adopt the logistic regression model for the sampling score following most of the empirical literature; but our framework can be extended to the case of other models

such as the probit model. The models $\pi_B(X^T\alpha)$ and $m(X^T\beta)$ are working models, which may be misspecified. If the sampling score model is correctly specified, we have $\pi_B(X) = \pi_B(X^T\alpha^*)$. If the outcome model is correctly specified, we have $m(X) = m(X^T\beta^*)$.

The procedure proposed 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 to solve penalized estimating equations for variable selection. Using equation (2) with $h(X;\alpha) = X$, we define the estimating function for α as

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

To select important variables in $m(X^T\beta)$, we define the estimating function for β as

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

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

$$U^p(\alpha,\beta) = U(\alpha,\beta) - \begin{pmatrix} q_{\lambda_\alpha}(|\alpha|)\mathrm{sgn}(\alpha) \\ q_{\lambda_\beta}(|\beta|)\mathrm{sgn}(\beta) \end{pmatrix}, \tag{6}$$

where $q_{\lambda_\alpha}(\alpha) = (q_{\lambda_\alpha}(|\alpha_0|), \ldots, q_{\lambda_\alpha}(|\alpha_p|))^T$ and $q_{\lambda_\beta}(\beta) = (q_{\lambda_\beta}(|\beta_0|), \ldots, q_{\lambda_\beta}(|\beta_p|))^T$ are some continuous functions, $q_{\lambda_\alpha}(|\alpha|)\mathrm{sgn}(\alpha)$ is the elementwise product of $q_{\lambda_\alpha}(\alpha)$ and $\mathrm{sgn}(\alpha)$, and $q_{\lambda_\beta}(|\beta|)\mathrm{sgn}(\beta)$ is the elementwise product of $q_{\lambda_\beta}(\beta)$ and $\mathrm{sgn}(\beta)$. 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 penalty function (Fan and Lv, 2011). Accordingly, we have

$$q_\lambda(|\theta|) = \lambda\left\{I(|\theta| < \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| \geqslant \lambda)\right\}, \tag{7}$$

for $a > 0$, where $(\cdot)_+$ is the truncated linear function, i.e., if $x \geqslant 0$, $(x)_+ = x$ and, if $x < 0$, $(x)_+ = 0$. We use $a = 3.7$ following the suggestion of Fan and Li (2001). We select the variables if the corresponding estimates of coefficients are non-zero in either the sampling score or the outcome model, indexed by $\mathcal{C}$.

Remark 1. To help to understand function (6) we discuss two scenarios. If $|\alpha_j|$ is large, then $q_{\lambda_\alpha}(|\alpha_j|)$ is 0, and therefore $U_{1,j}(\alpha)$ is not penalized. In contrast, if $|\alpha_j|$ is small but non-zero, then $q_{\lambda_\alpha}(|\alpha_j|)$ is large, and $U_{1,j}(\alpha)$ is penalized with a penalty term. The penalty term then forces $\hat{\alpha}_j$ to be 0 and excludes the jth element of 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 doubly robust estimator $\hat{\mu}_{dr}(\hat{\alpha}, \hat{\beta})$ in equation (5) with $(\hat{\alpha}, \hat{\beta})$ re-estimated on the basis of $X_{\mathcal{C}}$. As we shall show in Section 5, the set $\mathcal{C}$ contains the true important variables in either the sampling score model or the outcome model with probability approaching 1 (the oracle property). Therefore, if either the sampling score model or the outcome model is correctly specified, the asymptotic bias of $\hat{\mu}_{dr}(\alpha^*, \beta^*)$ is 0; however, if both models are misspecified, the asymptotic bias of $\hat{\mu}_{dr}(\alpha^*, \beta^*)$ is

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

To minimize $\text{a.bias}(\alpha,\beta)^2$, we consider the estimating function

$$\frac{\partial\,\text{a.bias}(\alpha,\beta)^2}{\partial(\alpha_{\mathcal{C}}^{\text{T}},\beta_{\mathcal{C}}^{\text{T}})^{\text{T}}} = 2\,\text{a.bias}(\alpha,\beta)\begin{pmatrix} I_{\text{B}}\left\{\dfrac{1}{\pi_{\text{B}}(X^{\text{T}}\alpha)}-1\right\}\{Y-m(X^{\text{T}}\beta)\}X_{\mathcal{C}} \\ \left\{\dfrac{I_{\text{B}}}{\pi_{\text{B}}(X^{\text{T}}\alpha)}-d_{\text{A}}I_{\text{A}}\right\}\partial m(X^{\text{T}}\beta)/\partial\beta_{\mathcal{C}}\end{pmatrix} \tag{8}$$

and the corresponding empirical estimating function

$$J(\alpha,\beta)=\begin{pmatrix}J_1(\alpha,\beta)\\ J_2(\alpha,\beta)\end{pmatrix}=\begin{pmatrix}\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}I_{\text{B},i}\left\{\dfrac{1}{\pi_{\text{B}}(X_i^{\text{T}}\alpha)}-1\right\}\{Y_i-m(X_i^{\text{T}}\beta)\}X_{i\mathcal{C}}\\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left\{\dfrac{I_{\text{B},i}}{\pi_{\text{B}}(X_i^{\text{T}}\alpha)}-d_{\text{A},i}I_{\text{A},i}\right\}\dfrac{\partial m(X_i^{\text{T}}\beta)}{\partial\beta_{\mathcal{C}}}\end{pmatrix} \tag{9}$$

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

To summarize, our two-step procedure is as follows.

Step 1: solve the penalized joint estimating equation $U^{\text{p}}(\alpha,\beta)=0$, denoted by $(\tilde{\alpha},\tilde{\beta})$. Let $\hat{\mathcal{M}}_\alpha=\{j:\tilde{\alpha}_j\neq0\}$, $\hat{\mathcal{M}}_\beta=\{j:\tilde{\beta}_j\neq0\}$ and $\mathcal{C}=\hat{\mathcal{M}}_\alpha\cup\hat{\mathcal{M}}_\beta$.

Step 2: obtain the proposed estimator as

$$\hat{\mu}_{\text{p-dr}}=\hat{\mu}_{\text{p-dr}}(\hat{\alpha},\hat{\beta})=\frac{1}{N}\sum_{i=1}^{N}\left\{I_{\text{B},i}\frac{Y_i-m(X_i^{\text{T}}\hat{\beta})}{\pi_{\text{B}}(X_i^{\text{T}}\hat{\alpha})}+I_{\text{A},i}d_{\text{A},i}m(X_i^{\text{T}}\hat{\beta})\right\}, \tag{10}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are obtained by solving $J(\alpha,\beta)=0$ for α and β with $\alpha_{\mathcal{C}^{\text{c}}}=0$ and $\beta_{\mathcal{C}^{\text{c}}}=0$.

Remark 2. The two steps use different estimating functions (6) and (9) for selection and estimation with the following advantages. First, function (6) 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 equation (9) for estimation leads to an attractive feature for inference about μ. We point out that, although the joint estimating function (9) is motivated by minimizing the asymptotic bias of $\hat{\mu}_{\text{dr}}(\alpha^*,\beta^*)$ when both nuisance models are misspecified, we do not expect the proposed estimator for μ to be unbiased in this case. Instead, using function (9) has the advantage in the case when either the sampling score or the outcome model is correctly specified. It is well known that post-selection inference is notoriously difficult even when both models are correctly specified because the estimation step is based on a random set of variables being selected. We show that our estimation strategy based on function (9) mitigates the possible first-step selection error and makes $\hat{\mu}_{\text{dr}}(\hat{\alpha},\hat{\beta})$ root n consistent if either the sampling probability or the outcome model is correctly specified in high dimensional data. Heuristically this is achieved because the first Taylor series expansion term is set to be 0 because of function (8). We relegate the details to Section 5.

Remark 3. Variable selection circumvents the instability or infeasibility of direct estimation of (α,β) with high dimensional X. Moreover, in step 2, we consider the union of covariates

$X_{\mathcal{C}}$, where $\mathcal{C} = \hat{\mathcal{M}}_\alpha \cup \hat{\mathcal{M}}_\beta$. It is worth comparing this choice with two other common choices in the literature. The first considers separate sets of variables for the two models, i.e. the sampling score is fitted on the basis of $\hat{\mathcal{M}}_\alpha$, and the outcome model is fitted on the basis of $\hat{\mathcal{M}}_\beta$. However, we note that in the joint estimating equations $J_1(\alpha, \beta)$ and $J_2(\alpha, \beta)$ should have the same dimension; otherwise, a solution to $J(\alpha, \beta) = 0$ may not exist. Moreover, Brookhart *et al.* (2006) and Shortreed and Ertefaie (2017) have shown that including outcome predictors 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, $\hat{\mu}_{\mathrm{dr}}(\hat{\alpha}, \hat{\beta})$ may have a better performance than the oracle estimator that uses the true important variables separately in the sampling score and the outcome model. Second, many researchers have suggested that including predictors that are solely related to the sampling score but not the outcome may harm estimation efficiency (De Luna *et al.*, 2011; Patrick *et al.*, 2011). However, this strategy is effective when both the sampling score and the outcome models are correctly specified. When the sampling score model is correctly specified but the outcome model is misspecified, restricting the variables to be the outcome predictors may make the sampling score misspecified by using the wrong set of variables. The simulation study suggests that $\hat{\mu}_{\text{p-dr}}$ restricted to the set of variables in $\hat{\mathcal{M}}_\beta$ is not doubly robust.

4. Computation

In this section, we discuss the computation for solving the penalized estimating equation (6). Following Johnson *et al.* (2008), we use an iterative algorithm that combines the Newton–Raphson algorithm for solving estimating equations and the minorization–maximization algorithm for the non-convex penalty of Hunter and Li (2005).

First, by the minorization–maximization algorithm, the penalized estimator $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta})$ solving equation (6) satisfies

$$U^{\mathrm{p}}(\tilde{\theta}) = U(\tilde{\theta}) - \begin{pmatrix} q_{\lambda_{\tilde{\alpha}}}(|\tilde{\alpha}|)\mathrm{sgn}(\tilde{\alpha})\dfrac{|\tilde{\alpha}|}{\epsilon + |\tilde{\alpha}|} \\ q_{\lambda_{\tilde{\beta}}}(|\tilde{\beta}|)\mathrm{sgn}(\tilde{\beta})\dfrac{|\tilde{\beta}|}{\epsilon + |\tilde{\beta}|} \end{pmatrix} = 0, \tag{11}$$

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

Second, we solve equation (11) using the Newton–Raphson algorithm. It may be challenging to implement the Newton–Raphson algorithm directly, because it involves inverting a large matrix. For computational stability, we use a co-ordinate decent algorithm (Friedman *et al.*, 2007) by cycling through and updating each of the co-ordinates. Define $m^{(k)}(t) = \mathrm{d}^k m(t)/\mathrm{d}^k t$ for $k \geqslant 1$:

$$\nabla(\theta) = \frac{\partial U(\theta)}{\partial \theta^{\mathrm{T}}} = \mathrm{diag}\left\{ \frac{\partial U_1(\alpha)}{\partial \alpha^{\mathrm{T}}}, \frac{\partial U_2(\beta)}{\partial \beta^{\mathrm{T}}} \right\}, \tag{12}$$

$$\frac{\partial U_1(\alpha)}{\partial \alpha^{\mathrm{T}}} = -\frac{1}{N}\sum_{i=1}^{N} I_{\mathrm{B},i} \frac{1 - \pi_{\mathrm{B}}(X_i^{\mathrm{T}}\alpha)}{\pi_{\mathrm{B}}(X_i^{\mathrm{T}}\alpha)} X_i X_i^{\mathrm{T}},$$

$$\frac{\partial U_2(\beta)}{\partial \beta^{\mathrm{T}}} = -\frac{1}{N}\sum_{i=1}^{N} I_{\mathrm{B},i} m^{(1)}(X_i^{\mathrm{T}}\beta)^2 X_i X_i^{\mathrm{T}},$$

and

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

Let θ start at an initial value $\tilde{\theta}^{[0]}$. With the other co-ordinates fixed, the kth Newton–Raphson update for θ_j is

$$\tilde{\theta}_j^{[k]} = \tilde{\theta}_j^{[k-1]} + \{\nabla_{jj}(\tilde{\theta}^{[k-1]}) + N\Lambda_{jj}(\tilde{\theta}^{[k-1]})\}^{-1}\{U_j(\tilde{\theta}^{[k-1]}) - N\Lambda_{jj}(\tilde{\theta}^{[k-1]})\tilde{\theta}_j^{[k-1]}\}, \tag{13}$$

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

We use K-fold cross-validation to select tuning parameters $(\lambda_\alpha, \lambda_\beta)$. More specifically, 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 on the basis of the training data and estimate the loss function on the basis of 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 parameter that minimizes the aggregated loss function over a prespecified grid.

Because the weighting estimator uses the sampling score $\pi_{\mathrm{B}}(X)$ 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:

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

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

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

where $\tilde{\beta}(\lambda_\beta)$ is the penalized estimator $\tilde{\beta}$ with 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 method. We assume that sample A is collected by simple random sampling or Poisson sampling with the following regularity conditions. Although it appears restrictive, our results extend to high entropy sampling designs; see remark 4.

Assumption 4. For all $1 \leqslant i \leqslant N$, $\pi_{\mathrm{A},i} \geqslant N^{\gamma-1}\delta_{\mathrm{A}} > 0$, where $\gamma \in (\frac{2}{3}, 1]$.

Similarly to assumption 1(b), we relax the strict positivity on $\pi_{\mathrm{A},i}$ and assume $n_{\mathrm{A}} = O(N^\gamma)$ for γ possibly strictly less than 1. Let $n = \min(n_{\mathrm{A}}, n_{\mathrm{B}})$, which is $O(N^\gamma)$ under assumptions 1 and 4.

Remark 4. We discuss the applicability of our asymptotic framework to sample A with high entropy sampling designs. Examples of high entropy sampling designs include simple random sampling, correlated Poisson sampling designs, normalized conditional Poisson sampling designs, Rao–Sampford sampling, the Chao design and the stratified design; see Berger (1998a, b),

Brewer and Donadio (2003) and Grafström (2010). The asymptotic properties for high entropy sampling designs are determined solely by their first-order inclusion probabilities. Therefore, for two high entropy sampling designs with the same first-order inclusion probabilities, their asymptotic behaviours are the same. In particular, we can consider the conditional Poisson sampling design (Hájek, 1964; Tillé, 2011), which appears when conditioning the Poisson design on a fixed sample size n_{A}. Let $p_{\mathrm{A}} = \{p_{\mathrm{A},i} : i = 1, \ldots, N\}$ with $\Sigma_{i=1}^{N} p_{\mathrm{A},i} = n_{\mathrm{A}}$ be the inclusion probabilities for the conditional Poisson sampling design. Given p_{A}, it is possible to find $\pi_{\mathrm{A}} = \{\pi_{\mathrm{A},i} : i = 1, \ldots, N\}$ with $\Sigma_{i=1}^{N} \pi_{\mathrm{A},i} = n_{\mathrm{A}}$ such that the conditional Poisson sampling design with p_{A} is asymptotically equivalent to the Poisson sampling design with the inclusion probabilities π_{A} (Hájek, 1964; Conti, 2014). Therefore, for a high entropy design, to apply our theoretical results, we check the conditions for the corresponding inclusion probability π_{A} under Poisson sampling.

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

Condition 1. The parameter θ belongs to a compact subset in $\mathbb{R}^{2p}$, and θ^* lies in the interior of the compact subset.

Condition 2. $\{X_i : i \in \mathcal{U}\}$ are fixed and uniformly bounded.

Condition 3. There are constants c_1 and c_2 such that

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

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

Condition 4. Let $\epsilon_i(\beta) = Y_i - m(X_i^{\mathrm{T}}\beta)$ be the ith residual. There is a constant c_3 such that $E\{|\epsilon_i(\beta^*)|^{2+\delta}\} \leqslant c_3$ for all $1 \leqslant i \leqslant N$ and some $\delta > 0$. There are constants c_4 and c_5 such that $E[\exp\{c_4|\epsilon_i(\beta^*)|\}|X_i] \leqslant c_5$ for all $1 \leqslant i \leqslant N$.

Condition 5. $m^{(1)}(X_i^{\mathrm{T}}\beta)$, $m^{(2)}(X_i^{\mathrm{T}}\beta)$ and $m^{(3)}(X_i^{\mathrm{T}}\beta)$ are uniformly bounded away from ∞ on $\mathcal{N}_{\theta,\tau} = \{\theta \in \mathbb{R}^{2p} : \|\theta_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta}\| \leqslant \tau\sqrt{(s_\theta/n)}, \theta_{\mathcal{M}_\theta^{\mathrm{c}}} = 0\}$ for some $\tau > 0$.

Condition 6. $\min_{j \in \mathcal{M}_\alpha} |\alpha_j^*|/\lambda_\alpha \to \infty$ and $\min_{k \in \mathcal{M}_\beta} |\beta_k^*|/\lambda_\beta \to \infty$, as $n \to \infty$.

Condition 7. $s_\theta = o(n^{1/3})$, $\lambda_\alpha, \lambda_\beta \to 0$, $\log(n)^2 = o(n\lambda_\theta^2)$, $\log(p) = o\{n\lambda_\theta^2/\log(n)^2\}$, $ps_\theta^4 \log(n)^6 = o(n^3\lambda_\theta^2)$ and $ps_\theta^4 \log(n)^8 = o(n^4\lambda_\theta^4)$, as $n \to \infty$.

These assumptions are typical in the literature on penalization methods. Condition 2 specifies a fixed design which is well suited under the finite population inference framework. Condition 4 holds for Gaussian distributions, sub-Gaussian distributions, and so on. Condition 5 holds for common models. Condition 7 specifies the restrictions on the dimension of covariates p and the dimension of the true non-zero coefficients s_θ. To gain insight, when the true model size s_θ is fixed, condition 7 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 is an approximate penalized solution $\tilde{\theta}$, which satisfies the selection consistency properties:

$$P\{|U_j^{\mathrm{p}}(\tilde{\theta})| = 0, j \in \mathcal{M}_\theta\} \to 1, \tag{14}$$

$$P\left\{|U_j^{\mathrm{p}}(\tilde{\theta})| \leqslant \frac{\lambda_\theta}{\log(n)}, j \in \mathcal{M}_\theta^{\mathrm{c}}\right\} \to 1, \tag{15}$$

$$P(\tilde{\theta}_{\mathcal{M}_\theta^{\mathrm{c}}} = 0) \to 1 \tag{16}$$

and

$$\tilde{\theta}_{\mathcal{M}_\theta} - \theta^*_{\mathcal{M}_\theta} = O_P\{\sqrt{(s_\theta/n)}\}, \tag{17}$$

as $n \to \infty$.

Results (14) and (15) imply that $U(\tilde{\theta}) = o_P\{\lambda_\theta/\log(n)\}$. Results (16) and (17) imply that, with probability approaching 1, the penalized estimating equation procedure would not overselect irrelevant variables and estimate the true non-zero coefficients at the $\sqrt{(s_\theta/n)}$ convergence rate, which is the so-called oracle property of variable selection.

We now establish the asymptotic properties of $\hat{\mu}_{\text{p-dr}}(\hat{\alpha}, \hat{\beta})$. Define a sequence of events $\mathcal{D}_n = \{\mathcal{M}_\theta \subset \mathcal{C}\}$, where we emphasize that $\mathcal{D}_n$ depends on n but we suppress the dependence of $\mathcal{M}_\theta$ and $\mathcal{C}$ on n. Following the same argument as for equation (17), given the event $\mathcal{D}_n$, we have $\{(\hat{\alpha} - \alpha^*)^{\mathrm{T}}, (\hat{\beta} - \beta^*)^{\mathrm{T}}\} = O_p\{\sqrt{(s_\theta/n)}\}$. Combining with $P(\mathcal{D}_n) \to 1$, we have

$$\{(\hat{\alpha} - \alpha^*)^{\mathrm{T}}, (\hat{\beta} - \beta^*)^{\mathrm{T}}\} = O_p\{\sqrt{(s_\theta/n)}\}. \tag{18}$$

By Taylor series expansion,

$$\begin{aligned} n^{1/2}\{\hat{\mu}_{\text{p-dr}}(\hat{\alpha}, \hat{\beta}) - \mu\} &= n^{1/2}\{\hat{\mu}_{\text{p-dr}}(\alpha^*, \beta^*) - \mu\} + n^{1/2}\frac{\hat{\mu}_{\text{p-dr}}(\hat{\alpha}, \hat{\beta})}{\partial(\alpha^{\mathrm{T}}, \beta^{\mathrm{T}})}\begin{pmatrix} \hat{\alpha} - \alpha^* \\ \hat{\beta} - \beta^* \end{pmatrix} \\ &\quad + O_P\left\{n^{1/2}\left\|\begin{pmatrix} \hat{\alpha} - \alpha^* \\ \hat{\beta} - \beta^* \end{pmatrix}\right\|_2^2\right\} \\ &= n^{1/2}\{\hat{\mu}_{\text{p-dr}}(\alpha^*, \beta^*) - \mu\} + O_P\left\{n^{1/2}\left\|\begin{pmatrix} \hat{\alpha} - \alpha^* \\ \hat{\beta} - \beta^* \end{pmatrix}\right\|_2^2\right\} \qquad (19) \\ &= n^{1/2}\{\hat{\mu}_{\text{p-dr}}(\alpha^*, \beta^*) - \mu\} + o_p(1), \qquad (20) \end{aligned}$$

where $\hat{\mu}_{\text{p-dr}}(\alpha, \beta)$ is defined in equation (10). Equation (19) follows because we solve equation (9) for (α, β). Equation (20) follows because of equation (18) and assumption 5. As a result, the way for estimating (α, β) leads to asymptotic equivalence between $\hat{\mu}_{\text{p-dr}}(\hat{\alpha}, \hat{\beta})$ and $\hat{\mu}_{\text{p-dr}}(\alpha^*, \beta^*)$.

We now show that $\hat{\mu}_{\text{p-dr}}(\alpha^*, \beta^*)$ is asymptotically unbiased for μ if either $\pi_{\mathrm{B}}(X^{\mathrm{T}}\alpha)$ or $m(X^{\mathrm{T}}\beta)$ is correctly specified. We note that

$$n^{1/2}E[\{\hat{\mu}_{\text{p-dr}}(\alpha^*, \beta^*) - \mu\}] = \frac{n^{1/2}}{N}\sum_{i=1}^{N} E\left[E\left\{\frac{I_{\mathrm{B},i}}{\pi_{\mathrm{B}}(X_i^{\mathrm{T}}\alpha^*)} - 1|X_i\right\}E\{Y_i - m(X_i^{\mathrm{T}}\beta^*)|X_i\}\right]. \tag{21}$$

If $\pi_{\mathrm{B}}(X^{\mathrm{T}}\alpha)$ is correctly specified, then $\pi_{\mathrm{B}}(X^{\mathrm{T}}\alpha^*) = \pi_{\mathrm{B}}(X)$ and therefore equation (21) is 0; if $m(X_i^{\mathrm{T}}\beta)$ is correctly specified, then $m(X_i^{\mathrm{T}}\beta^*) = m(X_i)$ and therefore equation (21) is 0.

Following the variance decomposition of Shao and Steel (1999), the asymptotic variance of the linearized term is

$$\begin{aligned} V[n^{1/2}\{\hat{\mu}_{\text{p-dr}}(\alpha^*, \beta^*) - \mu\}] &= n^{1/2}E[V\{\hat{\mu}_{\text{p-dr}}(\alpha^*, \beta^*) - \mu|I_{\mathrm{B}}, X, Y\}] \\ &\quad + n^{1/2}V[E\{\hat{\mu}_{\text{p-dr}}(\alpha^*, \beta^*) - \mu|I_{\mathrm{B}}, X, Y\}] := V_1 + V_2, \end{aligned}$$

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

$$V_1 = E\left\{ \frac{n}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{A,ij} - \pi_{A,i}\pi_{A,j}) \frac{m(X_i^T\beta^*)}{\pi_{A,i}} \frac{m(X_j^T\beta^*)}{\pi_{A,j}} \right\}. \tag{22}$$

For the second term V_2, note that

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

Thus,

$$V_2 = \frac{n}{N^2} \sum_{i=1}^{N} E\left[\left\{ \frac{I_{B,i}}{\pi_{B,i}(X_i^T\alpha^*)} - 1 \right\}^2 \{Y_i - m(X_i^T\beta^*)\}^2 \right]. \tag{23}$$

Theorem 2 summarizes the asymptotic properties of $\hat{\mu}_{\text{p-dr}}(\hat{\alpha}, \hat{\beta})$.

Theorem 2. Under assumptions 1–5, if either $\pi_B(X^T\alpha)$ or $m(X^T\beta)$ is correctly specified,

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

as $n \to \infty$, where $V = \lim_{n\to\infty}(V_1 + V_2)$, and V_1 and V_2 are defined in equations (22) and (23) respectively.

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

$$\hat{V}_1 = \frac{n}{N^2} \sum_{i\in\mathcal{S}_A} \sum_{j\in\mathcal{S}_A} \frac{\pi_{A,ij} - \pi_{A,i}\pi_{A,j}}{\pi_{A,ij}} \frac{m(X_i^T\hat{\beta})}{\pi_{A,i}} \frac{m(X_j^T\hat{\beta})}{\pi_{A,j}}. \tag{24}$$

To estimate V_2, we further express V_2 as

$$V_2 = \frac{n}{N^2} \sum_{i=1}^{N} E\left[\left\{ \frac{I_{B,i}}{\pi_{B,i}(X_i^T\alpha^*)^2} - \frac{2I_{B,i}}{\pi_{B,i}(X_i^T\alpha^*)} \right\} \{Y_i - m(X_i^T\beta^*)\}^2 + \{Y_i - m(X_i^T\beta^*)\}^2 \right]. \tag{25}$$

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

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

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

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

Remark 5. It is worth discussing the relationship of our proposed method to existing variable-selection methods in the survey literature. On the basis of a single probability sample source, McConville *et al.* (2017) proposed a model-assisted survey regression estimator of finite population totals using the lasso (the least absolute shrinkage and selection operator) to improve the efficiency. Chen, Valliant and Elliott (2018) and Chen *et al.* (2019) proposed model-based calibration estimators using the lasso based on non-probability samples integrating with

auxiliary known totals or probability samples respectively. However, their methods require that the working outcome model includes sufficient population information and therefore are not doubly robust. To the best of our knowledge, our paper is the first to propose doubly robust inference of finite population means after variable selection.

6. Simulation study

6.1. Set-up

In this section, we evaluate the finite sample performance of the procedure proposed. We first generate a finite population $\mathcal{F}_N = \{(X_i, Y_i) : i = 1, \ldots, N\}$ with $N = 10000$, where Y_i is a continuous or binary outcome variable, and $X_i = (1, X_{1,i}, \ldots, X_{p-1,i})^{\mathrm{T}}$ is a p-dimensional vector of covariates with the first component being 1 and other components independently generated from the standard normal distribution. We set $p = 50$. From the finite population, we select a non-probability sample $\mathcal{B}$ of size $n_{\mathrm{B}} \approx 2000$, according to the selection indicator $I_{\mathrm{B},i} \sim \mathrm{Ber}(\pi_{\mathrm{B},i})$. We select a probability sample $\mathcal{A}$ of the average size $n_{\mathrm{A}} = 500$ under Poisson sampling with $\pi_{\mathrm{A},i} \propto (0.25 + |X_{1i}| + 0.03|Y_i|)$. The parameter of interest is the population mean $\mu = N^{-1}\Sigma_{i=1}^{N} Y_i$.

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

(a) $\mathrm{logit}(\pi_{\mathrm{B},i}) = \alpha_0^{\mathrm{T}} X_i$, where $\alpha_0 = (-2, 1, 1, 1, 1, 0, 0, 0, \ldots, 0)^{\mathrm{T}}$ (model PSM I);
(b) $\mathrm{logit}(\pi_{\mathrm{B},i}) = 3.5 + \alpha_0^{\mathrm{T}} \log(X_i^2) - \sin(X_{3,i} + X_{4,i}) - X_{5,i} - X_{6,i}$, where $\alpha_0 = (0, 0, 0, 3, 3, 3, 3, 0, \ldots, 0)^{\mathrm{T}}$ (model PSM II).

For generating a continuous outcome variable Y_i, we consider both linear and non-linear outcome models with $\beta_0 = (1, 0, 0, 1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$:

(a) $Y_i = \beta_0^{\mathrm{T}} X_i + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 1)$ (model OM I);
(b) $Y_i = 1 + \exp\{3 \sin(\beta_0^{\mathrm{T}} X_i)\} + X_{5,i} + X_{6,i} + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 1)$ (model OM II).

For generating a binary outcome variable Y_i, we consider both linear and non-linear outcome models with $\beta_0 = (1, 0, 0, 3, 3, 3, 3, 0, \ldots, 0)^{\mathrm{T}}$,

(a) $Y \sim \mathrm{Ber}\{\pi_Y(X)\}$ with $\mathrm{logit}\{\pi_Y(X)\} = \beta_0^{\mathrm{T}} X$ (model OM III);
(b) $Y \sim \mathrm{Ber}\{\pi_Y(X)\}$ with $\mathrm{logit}\{\pi_Y(X)\} = 2 - \log\{(\beta_0^{\mathrm{T}} X)^2\} + 2X_{5,i} + 2X_{6,i}$ (model OM IV).

We consider the following estimators:

(a) *naive*, $\hat{\mu}_{\mathrm{naive}}$, the naive estimator using the simple average of Y_i from sample B, which provides the degree of the selection bias of sample B;
(b) *oracle*, $\hat{\mu}_{\mathrm{ora}}$, the doubly robust estimator $\hat{\mu}_{\mathrm{dr}}(\hat{\alpha}_{\mathrm{ora}}, \hat{\beta}_{\mathrm{ora}})$, where $\hat{\alpha}_{\mathrm{ora}}$ and $\hat{\beta}_{\mathrm{ora}}$ are based on the joint estimator restricting to the known important covariates for comparison;
(c) *p-ipw*, $\hat{\mu}_{\text{p-ipw}}$, the penalized inverse probability of sampling weighting estimator $\hat{\mu}_{\mathrm{IPW}} = N^{-1}\Sigma_{i \in \mathcal{B}} \hat{\pi}_{\mathrm{B},i}^{-1} Y_i$, where $\mathrm{logit}(\hat{\pi}_{\mathrm{B},i}) = X_i^{\mathrm{T}} \hat{\alpha}$ using a logistic regression model, and $\hat{\alpha}$ is obtained by a weighted penalized regression of $I_{\mathrm{B},i}$ on X_i based on the combined data from sample A and sample B, with the units in sample A weighted by the known sampling weights and the units in sample B weighted by 1.
(d) *p-reg*, $\hat{\mu}_{\text{p-reg}}$, the penalized regression estimator $\hat{\mu}_{\text{p-reg}} = N^{-1}\Sigma_{i \in \mathcal{A}} d_{\mathrm{A},i} m(X; \hat{\beta})$, where $\hat{\beta}$ is obtained by a penalized regression of Y_i on X_i based on sample B;
(e) *p-dr0*, $\hat{\mu}_{\text{p-dr0}}$, the penalized double estimating equation estimator based on the set of outcome predictors $\hat{\mathcal{M}}_{\beta}$;

(f) *p-dr*, $\hat{\mu}_{\text{p-dr}}$, the proposed penalized double estimating equation estimator based on the union of sampling and outcome predictors $\hat{\mathcal{M}}_\alpha \cup \hat{\mathcal{M}}_\beta$.

We also note that $\hat{\mu}_{\text{dr}}$ 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 penalization procedure in terms of the proportion of underselecting ('Under') or overselecting ('Over'), the average false negative results, FN (the average number of selected covariates that have the true 0 coefficients), and the average false positive results, FP (the average number of selected covariates that have the true 0 coefficients). The procedure proposed selects all covariates with non-zero coefficients in both the outcome model and the sampling score model under the true model specification. Moreover, the number of false positive results is small under the true model specification.

Fig. 1 displays the simulation results for the continuous outcome. The naive estimator $\hat{\mu}_{\text{naive}}$ shows large biases across all scenarios. The oracle estimator $\hat{\mu}_{\text{ora}}$ is doubly robust, in the sense that, if 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 largest biases except for scenario (ii). This approach is justifiable only if the sampling rate of sample B is relatively small compared with 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 estimator $\hat{\mu}_{\text{p-dr}}$ based on $\hat{\mathcal{M}}_\alpha \cup \hat{\mathcal{M}}_\beta$ is doubly robust, and its performance is comparable with the oracle estimator that requires knowing the true important variables. Moreover, $\hat{\mu}_{\text{p-dr}}$ 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 is consistent

Table 2. Simulation results for selection performance for the proposed double-penalized estimating equation procedure under four scenarios†

Scenario	β^*				α^*			
	Under $(\times 10^2)$	*Over* $(\times 10^2)$	*FN*	*FP*	*Under* $(\times 10^2)$	*Over* $(\times 10^2)$	*FN*	*FP*
Continuous outcome								
(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	70.6	15.0	0.9	0.2	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	0.0	0.4	0.0	0.4	100.0	100.0	3.5	4.3
Binary outcome								
(i) OM III and PSM I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(ii) OM IV and PSM I	100.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0
(iii) OM III and PSM II	0.0	0.0	0.0	0.0	100.0	100.0	4.0	1.0
(iv) OM IV and PSM II	100.0	0.0	4.0	0.0	100.0	96.0	4.0	1.0

†Under OM I and OM II, or OM III and OM IV, the outcome model is respectively correctly specified and misspecified, and, under PSM I and PSM II, the probability of sampling score model is respectively correctly specified or misspecified.

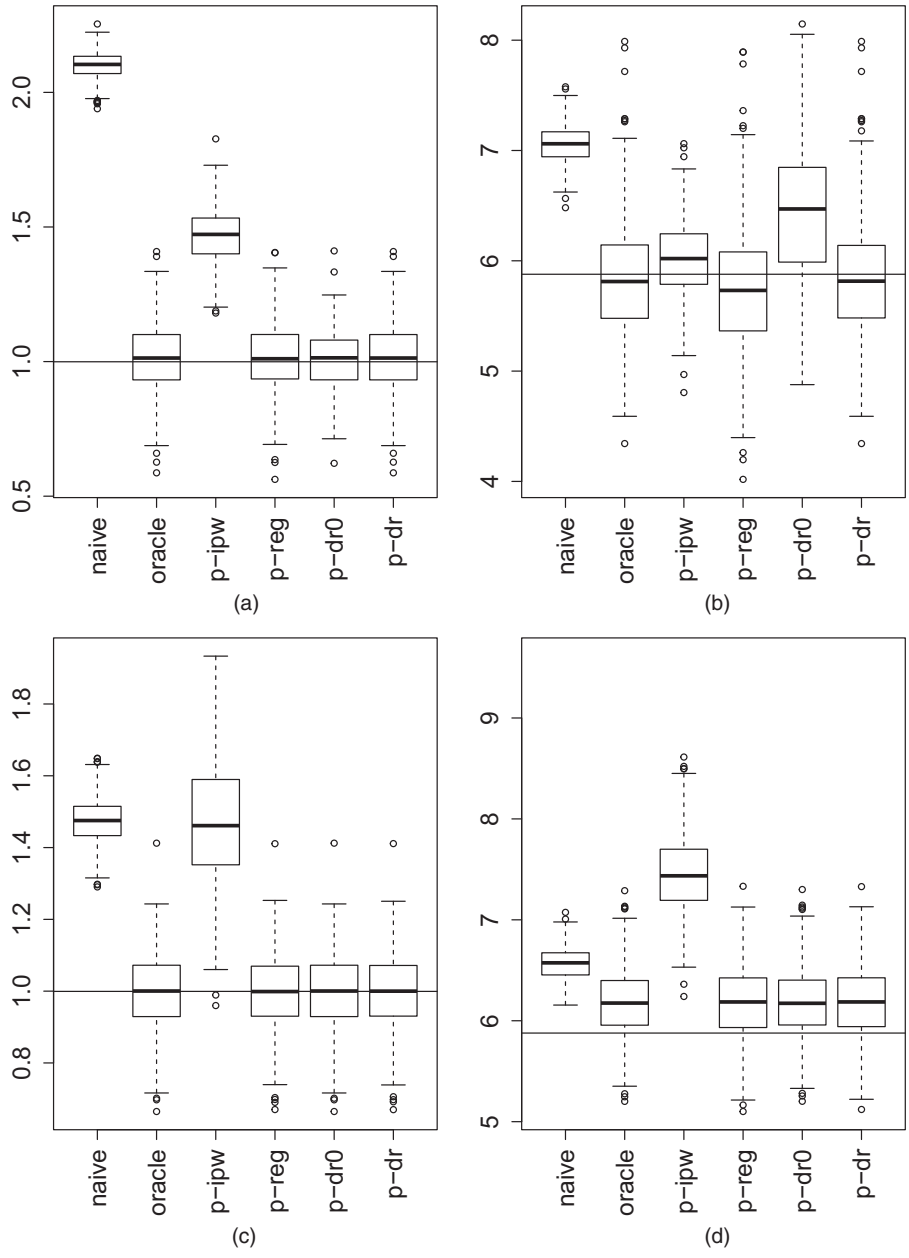

Fig. 1. Estimation results for the *continuous outcome* under four scenarios (under OM I and OM II, the outcome model is respectively correctly specified and misspecified, and, under PSM I and PSM II, the probability of sampling score model is respectively correctly specified and misspecified): (a) scenario (i), OM I and PSM I; (b) scenario (ii), OM II and PSM I; (c) scenario (iii), OM I and PSM II; (d) scenario (iv), OM II and PSM II

with the findings in Brookhart *et al.* (2006) and Shortreed and Ertefaie (2017). The proposed penalized double estimating equation estimator $\hat{\mu}_{\text{p-dr0}}$ based on $\hat{\mathcal{M}}_\beta$ is slightly more efficient than $\hat{\mu}_{\text{p-dr}}$ based on $\hat{\mathcal{M}}_\alpha \cup \hat{\mathcal{M}}_\beta$ in scenario (i) when both the outcome and the sampling score models are correctly specified; however, $\hat{\mu}_{\text{p-dr0}}$ has a large bias in scenario (ii) when the outcome model is misspecified and therefore is not doubly robust anymore; see remark 3.

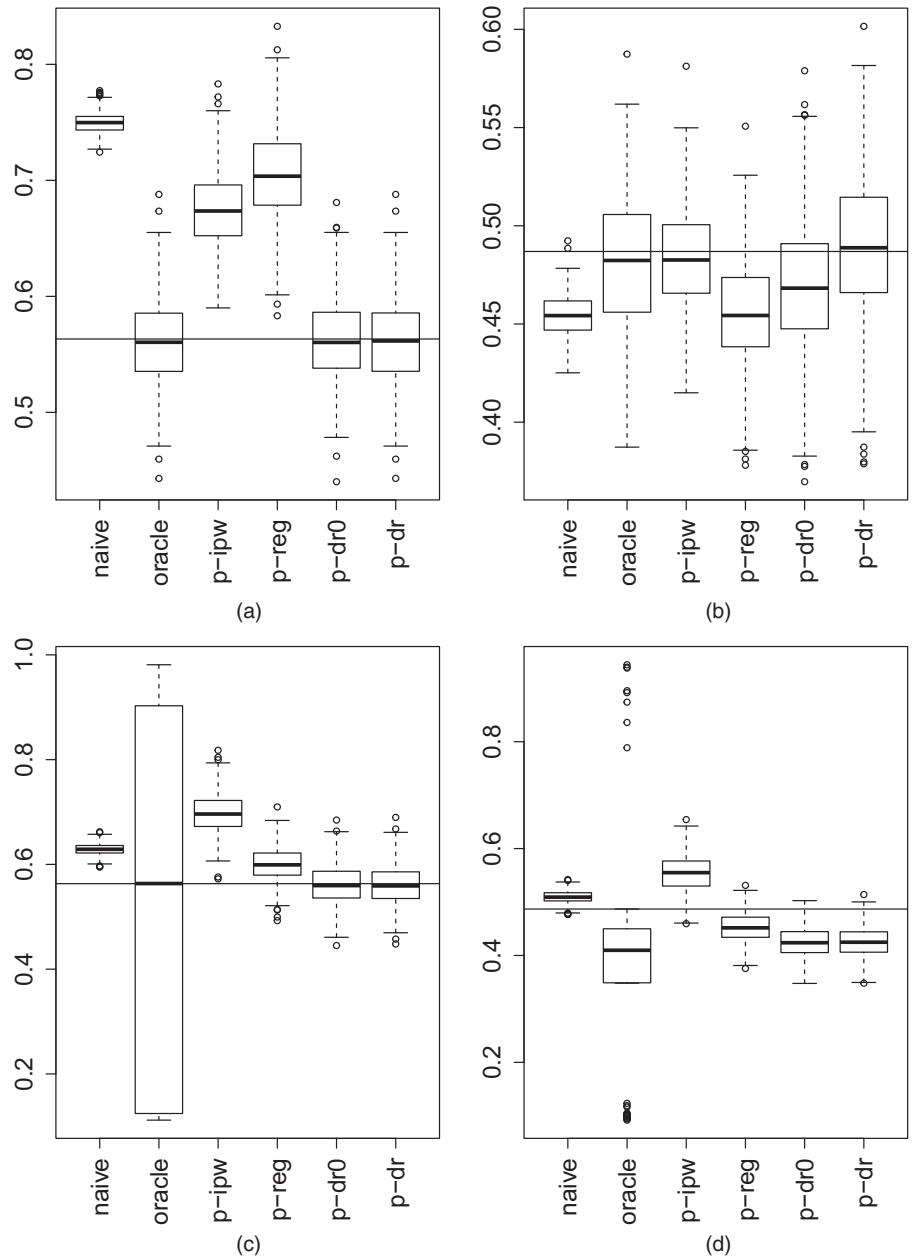

Fig. 2. Estimation results for the *binary outcome* under four scenarios (under OM III and OM IV, the outcome model is respectively correctly specified and misspecified, and, under PSM I and PSM II, the probability of sampling score model is respectively correctly specified and misspecified): (a) scenario (i), OM III and PSM I; (b) scenario (ii), OM IV and PSM I; (c) scenario (iii), OM III and PSM II; (d) scenario (iv), OM IV and PSM II

Fig. 2 displays the estimation results for the binary outcome. The same discussion above applies here. Moreover, when the outcome model is incorrectly specified, the oracle estimator has a large variability. In this case, the estimator proposed outperforms the oracle estimator, because the variable selection step helps to stabilize the estimation performance.

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

Scenario	*Results for continuous outcome*	*Results for binary outcome*
(i) OM I or OM III and PSM I	95.2 (93.3, 97.1)	95.7 (93.9, 97.6)
(ii) OM II or OM IV and PSM I	94.6 (92.6, 96.6)	95.5 (93.6, 97.4)
(iii) OM I or OM III and PSM II	96.2 (94.2, 97.8)	95.6 (93.8, 97.5)
(iv) OM II or OM IV and PSM II	88.2 (85.3, 91.1)	42.9 (38.3, 47.6)

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, whereas, if both models are misspecified, the coverage rates are off the nominal coverage.

7. An application

We analyse two data sets from the 2005 PRC (`http://www.pewresearch.org/`) and the 2005 behavioural risk factor surveillance system (BRFSS). The goal of the PRC study was to evaluate the relationship between individuals and community (Chen, Li and Wu, 2018; Kim *et al.*, 2018). The 2005 PRC data set is from a non-probability sample provided by eight vendors, which consists of $n_B = 9301$ 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 in local elections). In contrast, the 2005 BRFSS sample is a probability sample, which consists of $n_A = 441456$ subjects with survey weights. This data set does not have measurements on the study variables of interest; however, it contains a rich set of common covariates with the PRC data set listed in Fig. 3. To illustrate the heterogeneity in the study populations, Fig. 3 contrasts the covariate means from the PRC data and the design-weighted covariate means (i.e. the estimated population covariate means) from the BRFSS data set. 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 data set are subject to selection biases.

We compute the naive and proposed estimators. To apply the method proposed, we assume that the sampling score is a logistic regression model, the continuous outcome follows a linear regression model and that the binary outcome follows a logistic regression model. Using fivefold cross-validation, the double-selection procedure identifies 18 important covariates (all available covariates except for the north-east region) in the sampling score and the binary outcome model, and it identifies 15 important covariates (all available covariates except for black, an indicator of smoking every day, the north-east region and the south region) in the continuous outcome model.

Table 4 presents the point estimates, the standard errors and the 95% Wald confidence intervals. For estimating the standard error, because the second-order inclusion probabilities are unknown, following the survey literature, we compute the variance estimator in equation (24) by assuming that the survey design is single-stage Poisson sampling. We find significant differences in the results between the naive estimator and the proposed estimator. As demonstrated by the simulation study in Section 6, the naive estimator may be biased because of selection biases, and

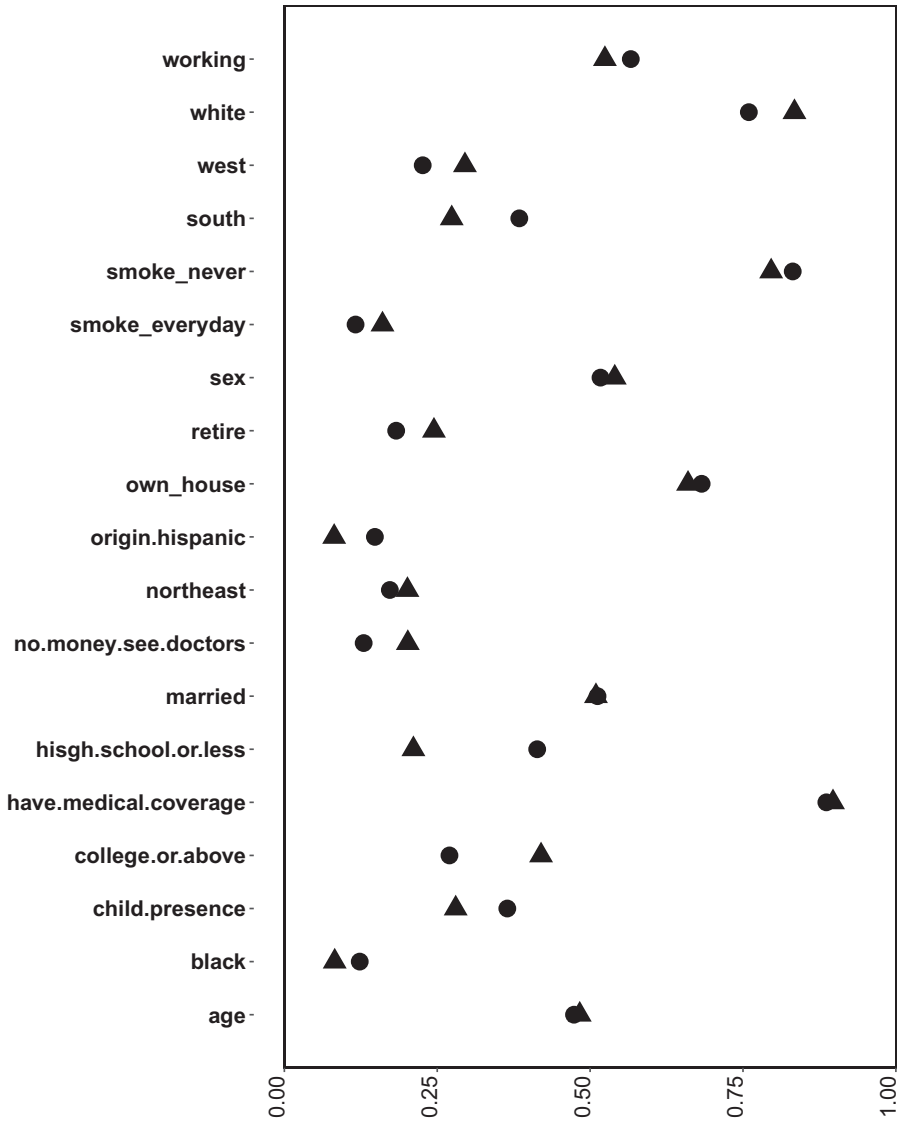

Fig. 3. Covariate means by two samples (age is divided by 100): ●, sample A; ▲, sample B

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

Method	Y_1 *(days had at least 1 drink last month)*			Y_2 *(whether voted in local elections)*		
	Estimate	*Standard error*	*CI*	*Estimate* $\times 10^2$	*Standard error* $\times 10^2$	*CI* $\times 10^2$
Naive	5.36	0.90	(5.17, 5.54)	75.3	0.5	(74.4, 76.3)
Proposed	4.84	0.15	(4.81, 4.87)	71.8	0.2	(71.3, 72.2)

the estimator proposed 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 in local elections.

8. Supplementary material

The supplementary material provides technical details and proofs.

Acknowledgements

Dr Yang is partially supported by National Science Foundation grant DMS 1811245, National Cancer Institute grant P01 CA142538 and Oak Ridge Associated Universities. Dr Kim is partially supported by National Science Foundation grant MMS 1733572. Dr Song is partially supported by National Science Foundation grant DMS 1555244 and National Cancer Institute grant P01 CA142538. An R package `IntegrativeFPM` that implements the method proposed is available from `https://github.com/shuyang1987/IntegrativeFPM`.

References

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

Berger, Y. G. (1998a) Rate of convergence for asymptotic variance of the Horvitz–Thompson estimator. *J. Statist. Planng Inf.*, **74**, 149–168.

Berger, Y. G. (1998b) Rate of convergence to normal distribution for the Horvitz–Thompson estimator. *J. Statist. Planng Inf.*, **67**, 209–226.

Bethlehem, J. (2016) Solving the nonresponse problem with sample matching? *Socl Sci. Comput. Rev.*, **34**, 59–77.

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

Brewer, K. and Donadio, M. E. (2003) The high entropy variance of the Horvitz–Thompson estimator. *Surv. Methodol.*, **29**, 189–196.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Stürmer, T. (2006) Variable selection for propensity score models. *Am. J. Epidem.*, **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, **181**, 1193–1209.

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. *J. Am. Statist. Ass.*, to be published, doi 10.1080/01621459.2019.1677241.

Chen, J. K. T., Valliant, R. and Elliott, M. R. (2018) Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Surv. Methodol.*, **44**, 117–144.

Chen, J. K. T., Valliant, R. L. and Elliott, M. R. (2019) Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Appl. Statist.*, **68**, 657–681.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018) Double/debiased machine learning for treatment and structural parameters. *Econmetr. J.*, **21**, C1–C68.

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.

Conti, P. L. (2014) On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya* B, **76**, 234–259.

De Luna, X., Waernbaum, I. and Richardson, T. S. (2011) Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, **98**, 861–875.

Deville, J.-C. and Särndal, C.-E. (1992) Calibration estimators in survey sampling. *J. Am. Statist. Ass.*, **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. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 4501–4515.

Elliott, M. R. and Valliant, R. (2017) Inference for nonprobability samples. *Statist. Sci.*, **32**, 249–264.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J. and Lv, J. (2011) Nonconcave penalized likelihood with np-dimensionality. *IEEE Trans. Inform. Theory*, **57**, 5467–5484.
Farrell, M. H. (2015) Robust inference on average treatment effects with possibly more covariates than observations. *J. Econmetr.*, **189**, 1–23.
Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**, 302–332.
Fuller, W. A. (2009) *Sampling Statistics*. Hoboken: Wiley.
Gao, X. and Carroll, R. J. (2017) Data integration with high dimensionality. *Biometrika*, **104**, 251–272.
Grafström, A. (2010) Entropy of unequal probability sampling designs. *Statist. Methodol.*, **7**, 84–97.
Hájek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.*, **35**, 1491–1523.
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. *Ann. Statist.*, **33**, 1617–1642.
Johnson, B. A., Lin, D. and Zeng, D. (2008) Penalized estimating functions and variable selection in semiparametric regression models. *J. Am. Statist. Ass.*, **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 (with discussion). *J. R. Statist. Soc.* A, **179**, 319–376.
Kim, J. K. and Haziza, D. (2014) Doubly robust inference with missing data in survey sampling. *Statist. Sin.*, **24**, 375–394.
Kim, J. K., Park, S., Chen, Y. and Wu, C. (2018) Combining non-probability and probability survey samples through mass imputation. *Preprint*. (Available from `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. *Surv. Methodol.*, **32**, 133–142.
Kott, P. S. and Liao, D. (2017) Calibration weighting for nonresponse that is not missing at random: allowing more calibration than response-model variables. *J. Surv. Statist. Methodol.*, **5**, 159–174.
Lee, S. and Valliant, R. (2009) Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociol. Meth. Res.*, **37**, 319–343.
McConville, K. S., Breidt, F. J., Lee, T. C. and Moisen, G. G. (2017) Model-assisted survey regression estimation with the LASSO. *J. Surv. Statist. Methodol.*, **5**, 131–158.
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. *Ann. Appl. Statist.*, **12**, 685–726.
O'Muircheartaigh, C. and Hedges, L. V. (2014) Generalizing from unrepresentative experiments: a stratified propensity score approach. *Appl. Statist.*, **63**, 195–210.
Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J. and Stürmer, T. (2011) The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmepidem. Drug Safty*, **20**, 551–559.
Rivers, D. (2007) Sampling for web surveys. *Jt Statist. Meet., Salt Lake City*.
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.
Shao, J. and Steel, P. (1999) Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *J. Am. Statist. Ass.*, **94**, 254–265.
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. *Prev. Sci.*, **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.
Tillé, Y. (2011) *Sampling Algorithms*. Berlin: Springer.
Tsiatis, A. (2006) *Semiparametric Theory and Missing Data*. New York: Springer.
Valliant, R. and Dever, J. A. (2011) Estimating propensity adjustments for volunteer web surveys. *Sociol. Meth. Res.*, **40**, 105–137.
Yang, S. and Kim, J. K. (2018) Integration of survey data and big observational data for finite population inference using mass imputation. *Preprint arXiv:1807.02817*.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for "Doubly robust inference when combining probability and non-probability samples with high-dimensional data"'.

Supplementary Material for "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

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,

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

*syang24@ncsu.edu, Department of Statistics, North Carolina 27695, U.S.A.

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 (17). 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 (12).

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

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

Moreover, we have

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

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

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

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

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

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

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

have

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

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 (14). By our construction of $\widetilde{\theta}$, for $j \in \mathcal{M}_\theta$, we have $U_j(\widetilde{\theta}) = 0$. Therefore, to show (14), it suffices to show that $P\left\{ q_{\lambda_\theta}(|\widetilde{\theta}_j|) = 0 : j \in \mathcal{M}_\theta \right\} \to 1$. By (7), it is equivalent to show that $P\left(|\widetilde{\theta}_j| \geq a\lambda_\theta : j \in \mathcal{M}_\theta \right) \to 1$. Note that

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

Therefore, we have

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

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 (15). By construction of $\widetilde{\theta}$, for $j \in \mathcal{M}_\theta^c$, we have $\widetilde{\theta}_j = 0$ and therefore $q_{\lambda_\theta}(\widetilde{\theta}_j)\mathrm{sign}(\widetilde{\theta}_j) = 0$. To show (15), it suffices to show that

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

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 $\widetilde{\theta}_{\mathcal{M}_\theta^c} = 0$ 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

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

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

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

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

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 $i \neq i'$. 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

$$
\begin{aligned}
W_{1,j} &= \frac{N}{n_A}\left(I_{A,1} - \frac{n_A}{N}\right) X_{1,j}, && V_{1,j} = 0, \\
W_{2,j} &= \frac{N}{n_A}\left(I_{A,2} - \frac{n_A - I_{A,1}}{N - I_{A,1}}\right) X_{2,j}, && V_{2,j} = \frac{N}{n_A}\left(\frac{n_A - I_{A,1}}{N - I_{A,1}} - \frac{n_A}{N}\right) X_{2,j}, \\
&\vdots && \vdots \\
W_{i,j} &= \frac{N}{n_A}\left(I_{A,i} - \frac{n_A - k_i}{N - k_i}\right) X_{i,j}, && V_{i,j} = \frac{N}{n_A}\left(\frac{n_A - k_i}{N - k_i} - \frac{n_A}{N}\right) X_{i,j}, \left(k_i = \sum_{l=1}^{i-1} I_{A,l}\right)
\end{aligned}
\tag{S7}
$$

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 $1 \le i \le N$. 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

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

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 $k \geq 2$,

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

for some positive constant M. By Bernstein inequality,

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

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

$$\begin{aligned} P\left\{ \left| N^{-1} \sum_{i=1}^N \left(\frac{I_{A,i} - \pi_{A,i}}{\pi_{A,i}} \right) X_{i,k} \right| > \frac{\lambda_\theta}{9 \log n} \right\} &\leq 2 \exp \left\{ - \frac{1}{2} \frac{\left(\frac{N \lambda_\theta}{9 \log n} \right)^2}{\sum_{j=1}^N \frac{1 - \pi_{A,i}}{\pi_{A,i}} X_{i,k}^2 + \frac{1}{3} M \left(\frac{N \lambda_\theta}{9 \log n} \right)} \right\} \\ &\leq 2 \exp \left\{ -Cn \left(\frac{\lambda_\theta}{\log n} \right)^2 \right\}. \end{aligned} \tag{S8}$$

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

$$\begin{aligned} E\left(|Z_{i,j}|^k\right) &= E\left[|I_{B,i}\epsilon_i(\beta^*)X_{i,j}|^k\right] \\ &\leq CE\left(|\epsilon_i(\beta^*)|^k\right) \\ &\leq Ck!c_4^{-k}E\left[\exp\left\{c_4|\epsilon_i(\beta^*)|\right\}\right] \qquad \text{(S9)} \\ &\leq Ck!c_4^{-k}c_5 \\ &\leq 2^{-1}k!M^{k-2}\delta, \end{aligned}$$

for some positive constants M and δ, where (S9) 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,

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

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

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

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

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

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 $1 \leq j \leq p$,

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

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

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

We then have

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

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

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

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

$$\begin{aligned} T_4 &\leq P\left\{\max_{j\in\mathcal{M}_\theta^c} \left|||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2 - E||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2\right| > \frac{Cn\lambda_\theta^2}{2s_\theta(\log n)^4}\right\} + o(1) \\ &\leq \sum_{j\in\mathcal{M}_\theta^c} P\left\{\left|||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2 - E||\nabla_{j,\mathcal{M}_\theta}(\theta^*)||^2\right| > \frac{Cn\lambda_\theta^2}{2s_\theta(\log n)^4}\right\} + o(1) \\ &\leq C\cdot\sum_{j\in\mathcal{M}_\theta^c} \frac{E\left(\sum_{j'\in\mathcal{M}_\theta}\left[\nabla_{j,j'}(\theta^*)^2 - E\left\{\nabla_{j,j'}(\theta^*)^2\right\}\right]\right)^2 s_\theta^2(\log n)^8}{n^2\lambda_\theta^4} + o(1) \quad \text{(S10)} \\ &= O\left\{p\left(\frac{s_\theta}{N^\gamma}\right)^2 \frac{s_\theta^2(\log n)^8}{n^2\lambda_\theta^4}\right\} \\ &= O\left\{\frac{ps_\theta^4(\log n)^8}{n^4\lambda_\theta^4}\right\} = o(1), \end{aligned}$$

where (S10) 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

$$\begin{aligned}
T_5 &\leq P\left\{\max_{j\in\mathcal{M}_\theta^c}\left|(\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})^{\mathrm{T}}D_{j,\mathcal{M}_\theta}(\widetilde{\theta}^*)(\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})\right|>\frac{\lambda_\theta}{3\log n}\right\}\\
&\leq P\left\{\max_{j\in\mathcal{M}_\theta^c}\left|(\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})^{\mathrm{T}}D_{j,\mathcal{M}_\theta}(\widetilde{\theta}^*)(\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta})\right|>\frac{\lambda_\theta}{3\log n}, ||\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta}||\leq\sqrt{s_\theta/n}\log n\right\}\\
&\quad +P\left\{||\widetilde{\theta}_{\mathcal{M}_\theta}-\theta^*_{\mathcal{M}_\theta}||>\sqrt{s_\theta/n}\log n\right\}\\
&\leq \sum_{j\in\mathcal{M}_\theta^c}P\left[\mathrm{trace}\left\{D_{j,\mathcal{M}_\theta}(\widetilde{\theta}^*)\right\}>\frac{n\lambda_\theta}{3s_\theta(\log n)^3}\right]+o(1)\\
&\leq c\sum_{j\in\mathcal{M}_\theta^c}\left[\frac{E\left(\left[\mathrm{trace}\left\{D_{j,\mathcal{M}_\theta}(\widetilde{\theta}^*)\right\}\right]^2\right)s_\theta^2(\log n)^6}{n^2\lambda_\theta^2}\right]+o(1), \qquad \text{(S11)}
\end{aligned}$$

where (S11) 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 (18)

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,

$$P\left\{\sup_{\theta \in \partial\mathcal{N}_{\theta,\tau}} (\theta - \theta^*)^{\mathrm{T}} J(\theta) < 0\right\} \geq 1 - \epsilon. \tag{S12}$$

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, (S12) holds, and as a result, on the event $\mathcal{D}_n$, $\widehat{\theta} - \theta^* = O_P(\sqrt{s_\theta/n})$. Combining with $P(\mathcal{D}_n) \to 1$, (18) holds.

References

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