Received: 4 April 2024 | Accepted: 22 May 2024

DOI: 10.1002/sta4.707

ORIGINAL ARTICLE

WILEY

Solving the missing at random problem in semi-supervised learning: An inverse probability weighting method

Jin Su | **Shuyi Zhang** | **Yong Zhou**

Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, School of Statistics, Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai, China

Correspondence
Yong Zhou and Shuyi Zhang, Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, School of Statistics, Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai, China.
Email: yzhou@amss.ac.cn and syzhang@fem.ecnu.edu.cn

Funding information
National Natural Science Foundation of China (State Key Program), Grant/Award Number: 71931004; National Key Research and Development Program of China, Grant/Award Numbers: 2021YFA1000100, 2021YFA1000101, 2021YFA1000104; Science and Technology Commission of Shanghai Municipality, Grant/Award Number: 21JS1400501; State Key Program of the National Natural Science Foundation of China, Grant/Award Number: 72331005; Youth Program of the National Natural Science Foundation of China, Grant/Award Number: 72201101; General Program of the National Natural Science Foundation of China, Grant/Award Number: 12271171; Youth Fund of General Project of MoE (Ministry of Education) of Humanities and Social Sciences, Grant/Award Number: 22YJC910013; Natural Science Foundation of Shanghai, Grant/Award Number: 23ZR1419400; Shanghai Pujiang Program, Grant/Award Number: 21PJC034

We propose an estimator for the population mean $\theta_0 = \mathbb{E}(Y)$ under the semi-supervised learning setting with the Missing at Random (MAR) assumption. This setting assumes that the probability of observing Y, denoted by π_M^*, depends on the total sample size M and satisfies $\pi_M^* = o(1)$. To efficiently estimate θ_0, we introduce an adaptive estimator based on inverse probability weighting and cross-fitting. Theoretical analysis reveals that our proposed estimator is consistent and efficient, with a convergence rate of $\sqrt{M\pi_M^*}$, slower than the typical $\sqrt{M}$ rate, due to the diminishing proportion of labelled data as the sample size M increases in the semi-supervised setting. We also prove the consistency of inverse probability weighting (IPW)–Nadaraya–Watson density function estimators. Extensive simulations and an application to the Los Angeles homeless data validate the effectiveness of our approach.

KEYWORDS
dimension reduction, inverse probability weighting, mean estimation, missing at random, semi-supervised learning

1 | INTRODUCTION

Semi-supervised learning (SSL) has gained significant attention in the fields of statistics and machine learning, as it addresses the challenge when there are limited labelled data and the abundance of unlabelled data. SSL aims to leverage both labelled and unlabelled data to improve model performance (Chapelle et al., 2009). Early SSL research focused on classification problems (Ando & Zhang, 2005; Wang & Shen, 2007; Wang et al., 2008), while recent works have extended SSL to regression tasks (Belkin et al., 2006; Johnson & Zhang, 2008). Despite progress, challenges remain existent in providing theoretical guarantees and effectively incorporating domain knowledge. A comprehensive overview of early SSL work can be referred to Zhu and Goldberg (2009).

SSL has been experiencing an explosive growth in recent years. Gronsbell and Cai (2017) considered the efficient SS evaluation of model predictive performance using generalised linear models. The construction of the best linear predictor in SS settings was studied by Azriel et al. (2022). Cai et al. (2022) proposed a SSL approach that combines transfer learning and surrogate-assisted techniques to achieve triple robustness. Song et al. (2023) proposed an estimator based on the projection method to study a class of general M-estimators in the SS setting. Tu et al. (2023) proposed a distributed SS sparse statistical inference method, which utilises additional unlabelled data to estimate the inverse of the Hessian matrix, thereby reducing local bias and improving overall accuracy. Wu et al. (2024) proposed an optimal subsampling method in the SS setting.

The fundamental difference between SSL and traditional missing data lies in the fact that the former allows the probability of observing Y to approach 0. Let $\mathbf{X} \in \mathbb{R}^p$ be a p-dimensional covariate, $Y \in \mathbb{R}$ be the outcome variable, $D \in \{0,1\}$ be a binary indicator variable, which takes the value of 1 if Y is observed and 0 if Y is missing. Let $\pi_M(\mathbf{x}) = \mathbb{P}(D = 1|\mathbf{X} = \mathbf{x})$ be the probability that Y is observed given $\mathbf{X} = \mathbf{x}$, commonly known as the Propensity Score (PS), and $\pi_M^* = \mathbb{E}\{\pi_M(\mathbf{X})\}$. In traditional missing data problems, we require that $\pi_M(\cdot)$ is bounded away from 0 and independent of M, which is the well-known "positive overlap" condition. Semi-supervised Learning and Missing at Random (SSL-MAR) assumption allows $\pi_M(\cdot)$ to depend on both the covariates $\mathbf{X}$ and the total sample size M, while Semi-supervised Learning and Missing Completely at Random (SSL-MCAR) assumption holds when the probability of Y being observed is independent of the covariates (i.e. $\pi_M(\mathbf{X}) = \pi_M^*$). By introducing sample size M into the PS, we allow $\pi_M(\mathbf{X}), \pi_M^* \to 0$ uniformly as $M \to \infty$.

In this paper, our interests focus on statistical inference for the population mean $\theta_0 = \mathbb{E}(Y)$. Under the positive overlap assumption, Hu et al. ((2010), (2012), (2014)) and Huang and Chan (2017) introduced different estimators for the θ_0. Under the SSL-MCAR assumption, Zhang et al. (2019) proposed mean estimators based on least squares, in the ideal SS setting (infinite unlabelled samples) and the ordinary SS setting (finite unlabelled samples), respectively. They proved that when there is a correlation between $\mathbf{X}$ and Y, the proposed estimators outperform the simple sample mean and provided an upper bound for the squared loss. At the same time, they introduced additional covariates under the nonparametric regression model and constructed a sequence of estimators, proving that it can asymptotically achieve the optimal risk. Although their work allows the dimension $p = o(\sqrt{n})$ to diverge, where n is the sample size of labelled data, it is still a strong condition in high-dimensional settings. Zhang and Bradic (2022) proposed a K-fold cross-fitted doubly robust estimator which allows for model misspecification and existence of nuisance parameters by using methods such as penalised estimation and random forests. It is worth noting that, although it may be difficult to recognise at the first glance, some other problems can also be reinterpreted as the estimation of population means, such as the variance of Y or the covariance between Y and $\mathbf{X}$, kernel estimation (Cannings & Fan, 2022). Cai and Guo (2020) studied the semi-supervised inference of explained variance in high-dimensional linear regression, which can also be considered as the estimation of population means. Another example is the Average Treatment Effect (ATE). Cheng et al. (2021) proposed an efficient and robust SS estimator for estimating the ATE. However, the above work has focused on the SSL-MCAR assumption; limited attention has been paid to the SSL-MAR assumption. To the best of our knowledge, Kallus and Mao (2020) and Zhang et al. (2023) are most closely works related to our research. We believe that our work serves as a valuable contribution to this continuously growing and expanding field.

To effectively estimate the population mean $\theta_0 = \mathbb{E}(Y)$, we propose an adaptive estimator based on inverse probability weighting (IPW) and cross-fitting. Specifically, we first introduce a missingness probability model $\pi_M(\mathbf{X})$ given the sample size M to characterise the relationship between the covariates $\mathbf{X}$ and the missingness indicator D, adapting to the SSL-MAR assumption. Meanwhile, to handle high-dimensional covariates, we employ a single index model (SIM) to transform the covariates $\mathbf{X}$ into a one-dimensional index $\mathbf{X}'\boldsymbol{\beta}$, which reduces the dimensionality while retaining the information from the original covariates. Based on this, we construct an IPW estimator by using kernel estimation and cross-fitting.

To summarise, we make the following important contributions to the existing literature. We discuss the problem of mean estimation under the assumptions of SSL-MAR and diverging covariate dimensionality. The SSL-MAR assumption requires that the probability of observing Y is 0 in the ideal semi-supervised setting (i.e. when the total sample size $M = \infty$) and greater than 0 in the finite-sample case (i.e. $M < \infty$). To address this, the PS $\pi_M(\mathbf{X})$ and the observation probability $\pi_M^* = \mathbb{E}\{\pi_M(\mathbf{X})\}$ are introduced, both of which are related to M. It is required that $\sup_{\mathbf{x} \in \mathcal{X}} \pi_M(\mathbf{x})$ and π_M^* converge to 0 as $M \to \infty$. An estimator combining inverse probability weighting and Nadaraya–Watson (IPW-NW) is proposed, and it is proven that the convergence rate of the target parameter is $\sqrt{M\pi_M^*}$ instead of $\sqrt{M}$. Furthermore, we also obtain a series of nuisance parameter estimates with degenerate convergence rates.

The remainder of the paper is organised as follows. In Section 2, we formulate the SS setting and introduce the proposed estimators via a semiparametric imputation. Theoretical properties based on influence function expansions are contained in Section 3. We implement numerical studies including extensive simulation studies and a real data example in Sections 4 and 5, respectively. The paper concludes with brief discussions in Section 6. All useful lemmas, technical proofs and additional numerical results are referred to the supporting information.

2 | METHODOLOGY

2.1 | Notation

Throughout this paper, we adopt the following notation. Denote $Y \in \mathbb{R}$ as the outcome and $\mathbf{X} \in \mathbb{R}^p$ be the covariate vector with dimension p, where p can be fixed or diverging. Let D be an indicator variable, where $D = 1$ if Y is observed and $D = 0$ if Y is missing. Define the complete dataset as $\mathcal{Z} = \{\mathbf{Z}_i = (D_i, D_i Y_i, \mathbf{X}_i')', i = 1, ..., M\}$. The observed dataset can then be partitioned as follows: (i) $\mathcal{L} = \{(D_i, D_i Y_i, \mathbf{X}_i')' : D_i = 1, i = 1, ..., M\}$ of n IID observations and (ii) $\mathcal{U} = \left\{(D_i, D_i Y_i, \mathbf{X}_i')', D_i = 0, i = 1, ..., M\right\}$ of N IID observations, where $\mathcal{L}$ contains the completely observed data and $\mathcal{U}$ contains the data with Y_i missing. Denote the cardinality of a set $\mathcal{A}$ as $|\mathcal{A}|$. We require $|\mathcal{U}| \gg |\mathcal{L}|$ in the SS setting. The propensity score $\pi_M(\mathbf{X})$ is defined as the probability of observing Y given $\mathbf{X}$ and sample size M (i.e. $\pi_M(\mathbf{X}) := \mathbb{P}(D = 1|\mathbf{X})$), and its expectation is denoted by $\pi_M^* = \mathbb{E}\{\pi_M(\mathbf{X})\}$. Additionally, we define $\pi_M^{*\,*-1} = \mathbb{E}\{\pi_M^{-1}(\mathbf{X})\}$, where $\pi_M(\mathbf{X})$, π_M^* and $\pi_M^{*\,*}$ depend on the sample size M. As M approaches infinity, $\pi_M(\mathbf{X})$, π_M^* and $\pi_M^{*\,*}$ approach zero, meaning that the probability of observing Y given $\mathbf{X}$ becomes increasingly small as the sample size becomes larger. For any $\mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{v}\|_r$ denotes the L_r vector norm of $\mathbf{v}$ for any $r \geq 0$, v_i denotes the i^{th} coordinate of $\mathbf{v}$ and $\mathbf{v}_{[i:j]}$ represents the elements of vector $\mathbf{v}$ from the i^{th} element to the j^{th} element, $\forall 1 \leq i, j \leq p+1$. For a matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, $\|\mathbf{A}\|_r := \sup_{\mathbf{v} \neq \mathbf{0}} \|\mathbf{A}\mathbf{v}\|_r / \|\mathbf{v}\|_r$. For sequences a_M and c_M, we write $a_M = O(c_M)$ if $a_M / c_M \leq C_1$ for some constant C_1, $a_M = o(c_M)$ if $a_M / c_M \to 0$, and $a_M \asymp c_M$ if $C_1 \leq a_M / c_M \leq C_2$. The values of constants C_is, which are independent of M and p, vary from one line to another.

Remark 2.1. By allowing $\pi_M(\mathbf{X})$, π_M^* and $\pi_M^{*\,*}$ to depend on the sample size M, we relax the "positive overlap" condition. The rationale is that while labelled data are negligible in the population distribution $\mathbb{P}(D = 1) = 0$, it is crucial for finite-sample estimation. To address this, we introduce a modified data-generating process with M-dependent labelling probabilities, inducing a finite-population distribution $\mathbb{P}_M$ that converges to the target $\mathbb{P}$. It is important to note that $\mathbb{P}_M$ is still a population distribution, not an empirical distribution. It is used to generate the M observations in the finite-sample setting. This setting flexibly describes the SSL-MAR assumption.

Remark 2.2. The distribution of D is different under the finite-population distribution $\mathbb{P}_M$ and the target distribution $\mathbb{P}$ due to the introduction of sample size-dependent labelling probabilities. This ensures sufficient labelled data for finite-sample estimation. However, the distribution of Y is invariant under both $\mathbb{P}_M$ and $\mathbb{P}$, despite potential variations in the labelling process across sample sizes. This invariance is crucial for our method, as it enables the estimation of population characteristics of Y using the finite-sample data. Maintaining consistency is essential for SSL, where the goal is to leverage a large amount of unlabelled data alongside limited labelled data to improve the estimation of population parameters.

2.2 | Estimation via semiparametric inverse propensity weighting

To estimate $\theta_0 = \mathbb{E}(Y)$, the most straightforward is the sample average $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$. However, under the traditional MAR assumption, $\overline{Y}$ is often an inconsistent estimate, since the target parameter for $\overline{Y}$ is $\mathbb{E}(Y|D = 1)$ rather than $\mathbb{E}(Y)$ unless the MCAR assumption holds. On the other hand, a considerable amount of informative unlabelled data remains underutilised under the single imputation scheme. We propose an alternative SS estimator for θ_0 by the semiparametric method. This strategy involves two main steps: dimension reduction and semiparametric calibration. In the dimension reduction step, we aim to reduce the dimensionality of the data by extracting relevant features or components. Instead of $\mathbb{E}(Y|\mathbf{X})$, we target to estimate $\mathbb{E}(Y|S)$. A semiparametric estimator of θ_0 is usually 'motivated' by a working model as illustrated below. For any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\theta_0 = \mathbb{E}(Y) = \mathbb{E}\{\mathbb{E}(Y|\mathbf{X}'\boldsymbol{\beta})\} := \mathbb{E}\{\mathbb{E}(Y|S)\} = \mathbb{E}\{g(S)\} = \mathbb{E}\{g(\mathbf{X}'\boldsymbol{\beta})\}, \tag{2.1}$$

where $S = \mathbf{X}'\boldsymbol{\beta}$ and $g(S) = \mathbb{E}(Y|S)$ is an unknown link model and reduces the dimension of the regressor $\mathbf{X}$ from p to 1. Together with MAR assumption, we assume that the nuisance condition mean function $\mathbb{E}(Y|\mathbf{X}) = \mathbb{E}(Y|\mathbf{X}, D = 1) = g(\mathbf{X}'\boldsymbol{\beta})$ follows a SIM. We use the kernel smoothing method to estimate $g(\cdot)$. Assume that $K(\cdot) : \mathbb{R} \to \mathbb{R}$ is a symmetric (in each dimension) probability density function with finite second-order moments in $\mathbb{R}$. Let $\tilde{K}(\mathbf{u}) = \prod_{d=1}^p K(u_d)$, $K_h(\mathbf{u}) = h^{-p}\prod_{d=1}^p K(u_d/h)$, where $h = h_M = o(1)$ is the bandwidth. We introduce a cross-fitted estimator of $g(\cdot)$: 1. For a fixed $K \geq 2$, we create a random partition $\{\mathcal{I}_k\}_{k=1}^K$ of the index set $\mathcal{I} := \{1, ..., M\}$; 2. For each $k \leq K$, we use the training set $\mathcal{Z}_{-k(i)} := \{\mathbf{Z}_i : i \in \mathcal{I} \setminus \mathcal{I}_k\}$ to obtain the estimator $\widehat{g}_{k,\widehat{w}}(\cdot; \cdot, \mathcal{Z}_{-k(i)})$, as shown in (2.2);

$$\hat{g}_{k,\hat{w}}(s; \hat{\boldsymbol{\beta}}_k, \mathcal{Z}_{-k(i)}) = \frac{\sum_{j \in \mathcal{Z}_{-k(i)}} \frac{D_j}{\hat{\pi}_M^{\hat{w}}(\mathbf{X}_j)} Y_j K_h\left(\mathbf{X}_j'\hat{\boldsymbol{\beta}}_k - \mathbf{x}'\hat{\boldsymbol{\beta}}_k\right)}{\sum_{j \in \mathcal{Z}_{-k(i)}} \frac{D_j}{\hat{\pi}_M^{\hat{w}}(\mathbf{X}_j)} K_h\left(\mathbf{X}_j'\hat{\boldsymbol{\beta}}_k - \mathbf{x}'\hat{\boldsymbol{\beta}}_k\right)}, \tag{2.2}$$

where $\hat{\boldsymbol{\beta}}_k$ is the estimate of $\boldsymbol{\beta}$ on the training set $\mathcal{Z}_{-k(i)}$. 3. For each $i=1,\ldots,M$, if $i\in\mathcal{I}_k$, then we set $\widehat{g}_{k,\widehat{w}}(S_i):=\widehat{g}_{k,\widehat{w}}(S_i;\hat{\boldsymbol{\beta}}_k,\mathcal{Z}_{-k(i)})$. We assume that $\max_k\|\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}\|_1\le b_M$ holds with high probability (w.h.p), $b_M=o(1)$. $\hat{\pi}_M^{\mathrm{w}}(\mathbf{X})$ is a working estimated PS model and $\hat{\pi}_M^{\mathrm{w}}(\mathbf{X})\xrightarrow{p}\pi_M^{\mathrm{w}}(\mathbf{X})$, $\pi_M^{\mathrm{w}}(\mathbf{X})$ can be the same as $\pi_M(\mathbf{X})$ or different.

We focus on estimator (2.3),

$$\hat{\theta}^{\mathrm{sp,w}}=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\mathcal{Z}_{k(i)}|}\sum_{i\in\mathcal{Z}_{k(i)}}\hat{g}_{k,\hat{w}}(S_i). \tag{2.3}$$

Remark 2.3. Ichimura (1993) and Mammen et al. (2016) proposed different estimators for $\boldsymbol{\beta}$ when p is fixed. Alquier and Biau (2013) and Eftekhari and Banerjee (2021) studied kernel smoothing estimation with high-dimensional covariates. Specifically, we can obtain an unbiased estimate of $\boldsymbol{\beta}$ by restricting the analysis on the observed samples $\mathcal{L}_{-k(i)}=\{\mathbf{Z}_j\in\mathcal{Z}_{-k(i)},D_j=1\}$ since $(Y|\mathbf{X})\equiv(Y|\mathbf{X},D=1)$ (Chakrabortty et al., 2019). We can use L_1-penalised regression with canonical link functions under additional assumptions, such as the elliptical symmetry of the covariate distribution given $D=1$(Li & Duan, 1989). If we assume that the marginal covariate distribution is elliptically symmetric, then $\boldsymbol{\beta}$ can be estimated by performing IPW L_1-penalised regression on the complete sample $\mathcal{Z}_{-k(i)}$, with weights constructed as $D/\pi_M(\mathbf{X})$ or $D/\hat{\pi}_M^{\mathrm{w}}(\mathbf{X})$.

Remark 2.4. We assume that $\pi_M(\mathbf{X})=\mathrm{expit}(\log(\pi_M^*)+\boldsymbol{\gamma}'\overline{\Omega}(\mathbf{X}))$, where $\mathrm{expit}(\cdot)$ is a expit link function, $\mathrm{expit}(u)=(1+\exp(-u))^{-1}$, $\overline{\Omega}(\mathbf{X}):=\{1,\Omega(\mathbf{X})\}=\{1,\Omega_1(\mathbf{X}),\ldots,\Omega_L(\mathbf{X})\}$, $\Omega_l(\mathbf{X}):=\left(\mathbf{X}^l_{[1]},\ldots,\mathbf{X}^l_{[p]}\right)'(l=1,\ldots,L)$, and L is a positive integer. Using similar techniques as in Zhang et al. (2023), we have $\pi_M^*\asymp\pi_M^{**}$. We also use it as our working model and denote it with the symbol "w" to represent the working model. $\pi_M^{\mathrm{w}}(\mathbf{X})$ can be the same as $\pi_M(\mathbf{X})$ or different (i.e. we allow for model misspecification). When p is fixed, the unknown parameters $\hat{\boldsymbol{\gamma}}^{\mathrm{w}}$ can be estimated by minimising

$$\mathcal{L}_M(\boldsymbol{\gamma};\hat{\pi}_M^{\mathrm{w}}):=-M^{-1}\sum_{i=1}^{M}\left[D_i\overline{\Omega}(\mathbf{X}_i)'\boldsymbol{\gamma}-\log\left\{1+\hat{\pi}_M^{\mathrm{w}}\exp\left(\overline{\Omega}(\mathbf{X}_i)'\boldsymbol{\gamma}\right)\right\}\right], \tag{2.4}$$

where $\hat{\pi}_M^{\mathrm{w}}=n/M$. When the dimension p diverges, we add an L_1 penalty term to the objective function (2.4) and obtain $\hat{\boldsymbol{\gamma}}$ by minimising the loss function.

$$\mathcal{L}_M^{\lambda}(\boldsymbol{\gamma};\hat{\pi}_M^{\mathrm{w}}):=-M^{-1}\sum_{i=1}^{M}\left[D_i\overline{\Omega}(\mathbf{X}_i)'\boldsymbol{\gamma}-\log\left\{1+\hat{\pi}_M^{\mathrm{w}}\exp\left(\overline{\Omega}(\mathbf{X}_i)'\boldsymbol{\gamma}\right)\right\}\right]+\lambda_M\|\boldsymbol{\gamma}\|_1, \tag{2.5}$$

where $\lambda_M>0$. Note that, although our notation may not explicitly indicate it, $\hat{\pi}_M^{\mathrm{w}}(\mathbf{X}_j)$ is estimated using data from $\mathcal{Z}_{-k(i)}$ excluding the observation $\mathbf{Z}_j$. To maintain notation consistency, denote $\hat{\boldsymbol{\xi}}^{\mathrm{w}}=\hat{\boldsymbol{\gamma}}^{\mathrm{w}}+\log(\hat{\pi}_M^{\mathrm{w}})\mathbf{e}_1$, $\boldsymbol{\xi}^{\mathrm{w}}=\boldsymbol{\gamma}^{\mathrm{w}}+\log(\pi_M^{\mathrm{w}})\mathbf{e}_1$, where $\mathbf{e}_1=(1,\ldots,0)'_{(pL+1)\times1}$. It holds that $\mathbb{P}(\max_k\|\hat{\boldsymbol{\xi}}^{\mathrm{w}}-\boldsymbol{\xi}^{\mathrm{w}}\|_1>r_M)\le p_M$, where $r_M,p_M=o(1)$, $r_M\ge0$ and $p_M\in[0,1]$.

3 | THEORETICAL PROPERTIES

In this section, we establish the asymptotic properties of $\hat{\theta}^{\mathrm{sp,w}}$. Let $f(s;\boldsymbol{\beta})$ be the density function of $S=\mathbf{x}'\boldsymbol{\beta}$, with support $\mathcal{S}$, $\hat{f}_{k,w}(s;\hat{\boldsymbol{\beta}}_k)=\frac{1}{|\mathcal{Z}_{-k(i)}|}\sum_{j\in\mathcal{Z}_{-k(i)}}\frac{D_j}{\pi_M^{\mathrm{w}}(\mathbf{X}_j)}K_h(\mathbf{X}_j'\hat{\boldsymbol{\beta}}_k-\mathbf{x}'\hat{\boldsymbol{\beta}}_k)$, $\hat{f}_{k,\hat{w}}(s;\hat{\boldsymbol{\beta}}_k)=\frac{1}{|\mathcal{Z}_{-k(i)}|}\sum_{j\in\mathcal{Z}_{-k(i)}}\frac{D_j}{\hat{\pi}_M^{\mathrm{w}}(\mathbf{X}_j)}K_h(\mathbf{X}_j'\hat{\boldsymbol{\beta}}_k-\mathbf{x}'\hat{\boldsymbol{\beta}}_k)$, $f_w(s;\boldsymbol{\beta}):=w(s)f(s;\boldsymbol{\beta})$, where $w(s)=\mathbb{E}\{\pi_M(\mathbf{X})/\pi_M^{\mathrm{w}}(\mathbf{X})|S=s\}$. Denote $w_y(s)=\mathbb{E}\{\pi_M(\mathbf{X})/\pi_M^{\mathrm{w}}(\mathbf{X})Y|S=s\}$, $\boldsymbol{\eta}_{\boldsymbol{\beta}}^{(1)}(s):=\mathbb{E}(\mathbf{X}|\mathbf{X}'\boldsymbol{\beta}=s)f(s;\boldsymbol{\beta})$. We assume the following conditions.

Assumption 1. $K(\cdot)$ satisfies $\int K(u)du=1$, $\int uK(u)du=0$, $\int u^2|K(u)|du<\infty$, $\|K(\cdot)\|_\infty\le M_K$, where M_K is a constant. $K(u)\to0$ as $u\to\infty$.

Assumption 2. $K(\cdot)$ has a bounded and integrable derivative $\dot{K}(\cdot)$ (i.e. $\|\dot{K}(\cdot)\|_\infty\le M_{\dot{K}}$ and $\int_{\mathbb{R}}|\dot{K}(u)|du\le C_{\dot{K}}$, where $M_{\dot{K}},C_{\dot{K}}\ge0$ are constants). $\dot{K}(\cdot)$ satisfies a local Lipschitz property: There exists a constant $L>0$ such that for all $u,v\in\mathbb{R}$, if $|u-v|\le L$, then $|\dot{K}(u)-\dot{K}(v)|\le\varphi(u)|u-v|$, where $\varphi(\cdot):\mathbb{R}\to\mathbb{R}^+$ is a bounded and integrable function satisfying $\|\varphi(\cdot)\|_\infty\le M_\varphi$ and $\int\varphi(u)du\le C_\varphi$, where $M_\varphi,C_\varphi\ge0$ are constants.

Assumption 3. There exists a constant $\delta_f>0$ such that $f(s;\boldsymbol{\beta})\ge\delta_f$. $f(s;\boldsymbol{\beta})$ is bounded, with corresponding first- and second-order derivatives also bounded. There exist constants C_1, C_2, δ_g and $\sigma_Y\ge0$ such that $0\le v^2(\mathbf{x})<\infty$. $\|g(\cdot)\|_\infty\le\delta_g$, $\|Y\|_{\psi_2}\le\sigma_Y$. The first-order derivatives of $\boldsymbol{\eta}_{\boldsymbol{\beta}}^{(1)}(s)$ with respect to each component of $\boldsymbol{\beta}$ satisfy $\max_{1\le j\le p}\|\boldsymbol{\eta}_{\boldsymbol{\beta}[j]}^{(1)}(\cdot)\|_\infty\le C_{\eta,1,1}$, $j=1,\ldots,p$.

Assumption 4. $\mathbf{X}$ and $\overline{\Omega}(\mathbf{X})$ are bounded, that is, $\|\mathbf{X}\|_\infty \le M_{\mathbf{X}}$ and $\|\overline{\Omega}(\mathbf{X})\|_\infty \le M_{\mathbf{X},L}$, where $M_{\mathbf{X}}, M_{\mathbf{X},L} \ge 0$ are some constants.

Assumption 5. $\mathbb{P}(\max_k \|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}\|_1 > b_M) \le q_M$, where $b_M, q_M = o(1)$, $b_M \ge 0$, $q_M \in [0,1]$.

Assumption 6. There exist constants $C_{w,1}, C_{w,2} > 0$, such that $C_{w,1} < \inf_{\mathbf{x}\in\mathcal{X}} \pi_M(\mathbf{x})/\pi_M^w(\mathbf{x}) \le \sup_{\mathbf{x}\in\mathcal{X}} \pi_M(\mathbf{x})/\pi_M^w(\mathbf{x}) < C_{w,2}$ holds.

Assumption 7. $M\pi_M^* h^4 \to 0$, $M\pi_M^* h \to \infty$.

Assumptions 1 and 2 are standard conditions for kernel estimation. Similar assumptions are also required in Bravo et al. (2020). Most kernel functions, such as the Gaussian kernel and triangular kernel, satisfy these conditions. Assumptions 3 and 4 are typical conditions that ensure properties such as estimation consistency, convergence rate, and asymptotic normality. Analogous assumptions can be found in Newey and McFadden (1994); Hansen (2008); Chakrabortty et al. (2019). Assumption 5 does not impose any specific constraints on the estimation or properties of $\hat{\boldsymbol{\beta}}_k$, allowing for many popular methods. Assumption 6 requires that the ratio between the true PS model and the working PS model is bounded within a constant range for all $\mathbf{x} \in \mathcal{X}$. Assumption 7 is the undersmoothing condition, which is a standard requirement for achieving consistency in nonparametric estimation, as discussed in Hu et al. (2012) and Lin et al. (2018). Given these conditions, we have the following lemma.

Lemma 3.1. Suppose Assumptions 1–6 hold. Define

$$
\begin{aligned}
\epsilon_{M,1}(t) &\equiv D_1 \frac{t}{\sqrt{M\pi_M^* h}} + D_2 \frac{t^2\sqrt{\log M}}{M\pi_M^* h} + D_3 h^2,\\
\epsilon_{M,2}(t) &\equiv D_4\left(b_M + \frac{b_M^2}{h^2}\right) + D_5\left\{\frac{b_M t}{\sqrt{M\pi_M^* h^3}}\left(1+\frac{b_M}{h}\right) + \frac{b_M\sqrt{\log p}}{\sqrt{M\pi_M^* h^3}}\right\}\\
&\quad + D_6\left\{\frac{b_M t^2\sqrt{\log M}}{M\pi_M^* h^2}\left(1+\frac{b_M}{h}\right) + \frac{b_M \log p\sqrt{\log M}}{M\pi_M^* h^2}\right\},\\
\epsilon_{M,3}(t) &\equiv D_7 r_M\left(\sqrt{\frac{t^2+\log p}{M\pi_M^* h}} + \frac{\sqrt{\log M}(t^2+\log p)}{M\pi_M^* h} + 1\right)\\
&\quad + D_8 r_M b_M\left(\frac{1}{h^2} + \frac{t}{\sqrt{M\pi_M^* h^3}} + \frac{t^2\sqrt{\log M}}{M\pi_M^* h^2}\right)\\
&\quad + D_9 r_M b_M^2\left(\frac{1}{h^3} + \frac{t}{\sqrt{M\pi_M^* h^5}} + \frac{t^2\sqrt{\log M}}{M\pi_M^* h^3}\right),
\end{aligned}
$$

and $\epsilon_M(t) = \epsilon_{M,1}(t) + \epsilon_{M,2}(t) + \epsilon_{M,3}(t)$. Then, for any $t \ge 0$, with at least probability $1 - 18\exp(-t^2) - 3p_M - 5q_M$,

$$
|\hat{f}_{k,\hat{w}}(s;\hat{\boldsymbol{\beta}}_k) - w(s)f(s;\boldsymbol{\beta})| \le \epsilon_M(t). \tag{3.6}
$$

This lemma provides an error bound of $\hat{f}_{k,\hat{w}}(s;\hat{\boldsymbol{\beta}}_k)$ and also ensures that $\hat{f}_{k,\hat{w}}(s;\hat{\boldsymbol{\beta}}_k) \ge \delta_f/2$. Now let us analyse the error bound in (3.6). For $|\hat{f}_{k,w}(s;\hat{\boldsymbol{\beta}}_k) - w(s)f(s;\boldsymbol{\beta})|$,

$$
\begin{aligned}
&\left|\hat{f}_{k,w}(s;\hat{\boldsymbol{\beta}}_k) - w(s)f(s;\boldsymbol{\beta})\right| \le |\hat{f}_{k,w}(s;\hat{\boldsymbol{\beta}}_k) - \hat{f}_{k,w}(s;\boldsymbol{\beta})| + |\hat{f}_{k,w}(s;\boldsymbol{\beta}) - \mathbb{E}\{\hat{f}_{k,w}(s;\boldsymbol{\beta})\}|\\
&+ |\mathbb{E}\{\hat{f}_{k,w}(s;\boldsymbol{\beta})\} - w(s)f(s;\boldsymbol{\beta})| := |\hat{R}_{M,w}(s)| + |\tilde{S}_{M,w}(s)| + |\overline{S}_{M,w}(s)|.
\end{aligned}
$$

These three terms, $|\hat{R}_{M,w}(s)|$, $|\tilde{S}_{M,w}(s)|$ and $|\overline{S}_{M,w}(s)|$, together contribute to the total error bound $\epsilon_{M,1}(t) + \epsilon_{M,2}(t)$ of $|\hat{f}_{k,w}(s;\hat{\boldsymbol{\beta}}_k) - w(s)f(s;\boldsymbol{\beta})|$. Firstly, $|\hat{R}_{M,w}(s)|$ represents the bias in the kernel estimation caused by using the estimated $\hat{\boldsymbol{\beta}}_k$ instead of the true $\boldsymbol{\beta}$, and its contribution to the error bound is represented by $\epsilon_{M,2}(t)$. Secondly, $|\tilde{S}_{M,w}(s)|$ accounts for the variability introduced by using a finite sample to estimate $\mathbb{E}\{\hat{f}_{k,w}(s;\boldsymbol{\beta})\}$. The terms $D_1 t/\sqrt{M\pi_M^* h} + D_2 t^2\sqrt{\log M}/M\pi_M^* h$ in $\epsilon_{M,1}(t)$ represent the error bound contribution from this term. Thirdly, $|\overline{S}_{M,w}(s)|$ captures the difference between $\mathbb{E}\{\hat{f}_{k,w}(s;\boldsymbol{\beta})\}$ and $f(s;\boldsymbol{\beta})$. The term $D_3 h^2$ in $\epsilon_{M,1}(t)$ represents the error bound contributed by this term, which is a bias introduced by the kernel smoothing method itself. Lastly, $\epsilon_{M,3}(t)$ represents the contribution of $|\hat{f}_{k,\hat{w}}(s;\hat{\boldsymbol{\beta}}_k) - \hat{f}_{k,w}(s;\hat{\boldsymbol{\beta}}_k)|$ to the total error bound.

Lemma 3.2. Suppose Assumptions 1–7, $(b_M\sqrt{\log p})/h = o(1)$, $(r_M\sqrt{\log p})/h = o(1)$ and $\log M \log p = O(M\pi_M^* h)$ hold. Then when $M\pi_M^* \to \infty$, we have

$$\begin{aligned}\hat{\theta}^{\mathrm{sp,w}} - \theta_0 &= \frac{1}{M}\sum_{i=1}^{M}\frac{D_i\{Y_i - g(S_i)\}}{\pi_M^{\mathrm{w}}(\mathbf{X}_i)w(S_i)} + \frac{1}{M}\sum_{i=1}^{M}\frac{w_Y(S_i)}{w(S_i)} - \theta_0 \\ &\quad + 1\{m(\mathbf{X}) \neq g(S)\}O_p(b_M + r_M) + O_p\left\{d_M + (b_M + r_M)\sqrt{\frac{\log p}{M\pi_M^*}}\right\},\end{aligned} \tag{3.7}$$

where $d_M = (M^2\pi_M^* h)^{-1/2}$.

Lemma 3.2 shows the influence function of the estimator $\hat{\theta}^{\mathrm{sp,w}}$, which measures the sensitivity of the estimator to small perturbations. The first term in the influence function is proportional to $Y - m(\mathbf{X})$, with weights being the inverse of the propensity score $\pi_M^{\mathrm{w}}(\mathbf{X})w(S) = \pi_M^{\mathrm{w}}(\mathbf{X})\mathbb{E}\{\pi_M(\mathbf{X})/\pi_M^{\mathrm{w}}(\mathbf{X})|S\}$. Observations with lower propensity scores are assigned higher weights to compensate for their reduced representation in the sample. If only one model is correctly specified, consistency holds. Moreover, when the conditional mean function is correctly specified, the estimation errors brought by $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\xi}}^{\mathrm{w}}$ do not affect the asymptotic normality of $\hat{\theta}^{\mathrm{sp,w}}$, highlighting the robustness of the proposed method. It is worth emphasising that in this lemma, we require $\log M \log p = O(M\pi_M^* h)$, which means that the product of the smoothing bandwidth and the effective sample size of labelled samples ($M\pi_M^* h$) should not be smaller than the order of $\log M \log p$. The asymptotic normality of the proposed estimator can be obtained by applying the Lindeberg-Feller central limit theorem and the Slutsky's theorem. Denote $\Psi(\mathbf{Z}) = \frac{D\{Y-g(S)\}}{\pi_M^{\mathrm{w}}(\mathbf{X})w(S)} + g(S) - \theta_0$, $V_M = \mathbb{E}\{\Psi^2(\mathbf{Z})\}$. Then we have the following theorem.

Theorem 3.1. Suppose Assumptions 1–7, $(b_M\sqrt{\log p})/h = o(1)$, $(r_M\sqrt{\log p})/h = o(1)$, $\log M \log p = O(M\pi_M^* h)$ and $m(\mathbf{X}) = g(S)$ hold. If for any $\epsilon > 0$, $\pi_M^*\mathbb{E}\left[\Psi^2(\mathbf{Z})1\left\{|\Psi(\mathbf{Z})| > \epsilon\sqrt{M/\pi_M^*}\right\}\right] \to 0$ as $M \to \infty$. Then as $M\pi_M^* \to \infty$,

$$(M\pi_M^*)^{1/2}\left(\hat{\theta}^{\mathrm{sp,w}} - \theta_0\right) = O_p(1), M^{1/2}V_{M,S}^{-1/2}\left(\hat{\theta}^{\mathrm{sp,w}} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0,1). \tag{3.8}$$

Theorem 3.1 establishes the asymptotic normality of $\hat{\theta}^{\mathrm{sp,w}}$, which relies on the correct specification of $g(S)$. As $M \to \infty$, the tail condition $\pi_M^*\mathbb{E}\left[\Psi^2(\mathbf{Z})1\left\{|\Psi(\mathbf{Z})| > \epsilon\sqrt{M/\pi_M^*}\right\}\right] \to 0$ characterises the behaviour of extreme or rare events associated with the function $\Psi(\mathbf{Z})$. These conditions are crucial for establishing the asymptotic normality of the estimator $\hat{\theta}^{\mathrm{sp,w}}$. Under these conditions, Theorem 3.1 shows that the convergence rate of $\hat{\theta}^{\mathrm{sp,w}}$ is not $\sqrt{M}$, but $\sqrt{M\pi_M^*}$. Similar degenerate convergence rates can be found in the limited overlap literature (Hong et al., 2020; Khan & Tamer, 2010; Rothe, 2017). In classical missing data studies, under the positivity overlap assumption, Crump et al. (2009) directly discarded observations with extremely small or large propensity scores, achieving a convergence rate of $\sqrt{M}$. However, in the semi-supervised setting, the number of unlabelled data far exceeds the number of labelled data, and labelled data can easily have extreme propensity scores. We cannot directly discard these observations since our goal is to fully utilise the valuable information contained in labelled data.

4 | SIMULATION STUDIES

In this section, we compare the performance of the proposed estimator with the supervised benchmark estimator $\overline{Y}$ that relies on fully observed labelled data in terms of bias, standard deviation (SD), standard error (SE), mean squared error (MSE) and efficiency based on MSE and SE. All simulations are repeated 1000 times. We use the plug-in method to obtain the SD and coverage probability (CP) of the 95% Confidence Interval (CI). We use the Gaussian kernel function with bandwidth set as $h_M = \hat{\sigma}|\mathcal{L}_{-k(i)}|^{-1/3}$, where $\hat{\sigma}$ is the standard deviation of $\{\mathbf{X}_i'\hat{\boldsymbol{\beta}}_k\}_{\mathbf{X}_i \in \mathcal{L}_{-k(i)}}$ and $K = 5$. The covariates are generated from $\{\mathbf{X}i\}_{i=1}^{M} \overset{\mathrm{IID}}{\sim} N_{p-1}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = I_{p-1}$, $M = 1000, 2000, 5000$. The dimension $p = 10, 100$, and when $p = 100$, the sparsity $s = 10$. Let $\vec{\mathbf{X}} = (1, \mathbf{X}')'$ and set $\pi_M^* = 0.1$. We consider the following data settings:

(K1) $\pi_M(\mathbf{X}) = \pi_M^*$;
(K2) $\pi_M(\mathbf{X}) = \exp(\log(\pi_M^*) + \vec{\mathbf{X}}'\boldsymbol{\gamma}_0)/\{1 + \exp(\log(\pi_M^*) + \vec{\mathbf{X}}'\boldsymbol{\gamma}_0)\}$.

The parameter $\boldsymbol{\gamma}_0 = (-0.262, 0.4, 0.4, 0.4, -0.4, -0.4, 0.2, 0.2, -0.2, -0.2, -0.2, \mathbf{0}_{1\times(p-s)})'$. Here, $\gamma_{0,1} = -0.262$ is set to ensure that $\mathbb{E}\{\pi_M(\mathbf{X})\} = \pi_M^*$.(Y1) $Y_i = \vec{\mathbf{X}}_i'\boldsymbol{\beta}_0 + \varepsilon_i$;

(Y2) $Y_i = \vec{\mathbf{X}}_i'\boldsymbol{\beta}_0 + (\vec{\mathbf{X}}_i'\boldsymbol{\beta}_0)^2 + \varepsilon_i$;
(Y3) $Y_i = \vec{\mathbf{X}}_i'\boldsymbol{\beta}_0 + \exp(\vec{\mathbf{X}}_i'\boldsymbol{\beta}_0) + \varepsilon_i$.

The parameter $\boldsymbol{\beta}_0 = (0.5, \mathbf{0.1}_{1\times s}, \mathbf{0}_{1\times(p-s)})'$, $\epsilon_i \overset{\mathrm{IID}}{\sim} N(0, 0.1^2)$. For (Y1), (Y2) and (Y3), the values of θ_0 are $0.5, 0.85, 2.23$, respectively. For the above 2×3 data generation processes, we consider the following nine estimators:

TABLE 1 The results of simulation (Y2) on 1000 simulation runs under $p = 10, M = 1000, 2000, 5000$.

	Bias	SE	SD	CP	Eff(M)	Eff(S)	Bias	SE	SD	CP	Eff(M)	Eff(S)
	MCAR(K1)						MAR(K2)					
	$M = 1000$											
$\overline{Y}$	0.001	0.065	0.065	0.941	-	-	0.031	0.067	0.066	0.941	-	-
$\overline{Y}_{\text{IPW}}$	0.001	0.065	0.065	0.941	0.000	0.000	0.000	0.095	0.088	0.920	−0.638	−0.408
$\overline{Y}_{\text{cc}}$	−0.000	0.020	0.021	0.946	0.902	0.687	−0.000	0.020	0.021	0.946	0.923	0.696
$\hat{\theta}_{\mathcal{L}}^{\text{sp,T}}$	−0.004	0.024	0.025	0.949	0.857	0.627	−0.003	0.027	0.020	0.951	0.865	0.598
$\hat{\theta}_{\text{IPW}}^{\text{sp,T}}$	−0.004	0.024	0.025	0.949	0.857	0.627	−0.001	0.028	0.020	0.949	0.852	0.578
$\hat{\theta}_{\mathcal{L}}^{\text{sp,MAR}}$	−0.004	0.024	0.026	0.960	0.862	0.634	−0.003	0.027	0.020	0.963	0.868	0.602
$\hat{\theta}_{\text{IPW}}^{\text{sp,MAR}}$	−0.004	0.024	0.026	0.961	0.862	0.633	−0.001	0.028	0.020	0.958	0.856	0.582
$\hat{\theta}_{\mathcal{L}}^{\text{sp,MCAR}}$	−0.004	0.024	0.025	0.956	0.856	0.627	−0.002	0.028	0.032	0.911	0.850	0.575
$\hat{\theta}_{\text{IPW}}^{\text{sp,MCAR}}$	−0.004	0.024	0.025	0.955	0.856	0.627	−0.002	0.028	0.025	0.911	0.850	0.575
	$M = 2000$											
$\overline{Y}$	0.001	0.048	0.046	0.944	-	-	0.035	0.049	0.047	0.889	-	-
$\overline{Y}_{\text{IPW}}$	0.001	0.048	0.046	0.944	0.000	0.000	0.002	0.068	0.065	0.939	−0.259	−0.382
$\overline{Y}_{\text{cc}}$	−0.000	0.014	0.015	0.951	0.910	0.700	−0.000	0.014	0.015	0.948	0.944	0.708
$\hat{\theta}_{\mathcal{L}}^{\text{sp,T}}$	−0.002	0.016	0.017	0.956	0.886	0.666	−0.002	0.018	0.019	0.954	0.912	0.636
$\hat{\theta}_{\text{IPW}}^{\text{sp,T}}$	−0.002	0.017	0.017	0.956	0.886	0.666	−0.001	0.019	0.020	0.956	0.905	0.620
$\hat{\theta}_{\mathcal{L}}^{\text{sp,MAR}}$	−0.002	0.017	0.017	0.957	0.888	0.670	−0.002	0.018	0.020	0.960	0.913	0.638
$\hat{\theta}_{\text{IPW}}^{\text{sp,MAR}}$	−0.002	0.016	0.017	0.957	0.888	0.669	−0.001	0.019	0.020	0.957	0.907	0.625
$\hat{\theta}_{\mathcal{L}}^{\text{sp,MCAR}}$	−0.002	0.017	0.017	0.953	0.886	0.666	−0.002	0.019	0.017	0.914	0.897	0.605
$\hat{\theta}_{\text{IPW}}^{\text{sp,MCAR}}$	−0.002	0.017	0.017	0.953	0.886	0.666	−0.002	0.019	0.017	0.914	0.897	0.605
	$M = 5000$											
$\overline{Y}$	−0.001	0.030	0.029	0.940	-	-	0.032	0.031	0.030	0.792	-	-
$\overline{Y}_{\text{IPW}}$	−0.001	0.030	0.029	0.940	0.000	0.000	0.001	0.045	0.042	0.927	−0.001	−0.451
$\overline{Y}_{\text{cc}}$	−0.000	0.010	0.009	0.946	0.896	0.677	−0.000	0.010	0.009	0.944	0.955	0.692
$\hat{\theta}_{\mathcal{L}}^{\text{sp,T}}$	−0.001	0.011	0.010	0.938	0.869	0.642	−0.001	0.012	0.012	0.953	0.933	0.627
$\hat{\theta}_{\text{IPW}}^{\text{sp,T}}$	−0.001	0.011	0.010	0.938	0.869	0.642	−0.001	0.012	0.012	0.953	0.932	0.622
$\hat{\theta}_{\mathcal{L}}^{\text{sp,MAR}}$	−0.001	0.011	0.010	0.939	0.870	0.643	−0.001	0.012	0.012	0.952	0.933	0.628
$\hat{\theta}_{\text{IPW}}^{\text{sp,MAR}}$	−0.001	0.011	0.010	0.939	0.870	0.643	−0.001	0.012	0.012	0.951	0.932	0.624
$\hat{\theta}_{\mathcal{L}}^{\text{sp,MCAR}}$	−0.001	0.011	0.010	0.938	0.869	0.641	−0.001	0.010	0.011	0.893	0.922	0.596
$\hat{\theta}_{\text{IPW}}^{\text{sp,MCAR}}$	−0.001	0.011	0.010	0.938	0.869	0.641	−0.001	0.010	0.011	0.893	0.922	0.596

(E1) $\overline{Y}$: The estimator obtained using only labelled data, $\overline{Y} = \sum_{i=1}^{n} Y_i/n$, which is the benchmark estimator.

(E2) $\overline{Y}_{\text{IPW}}$: $\overline{Y}_{\text{IPW}} = \left(\sum_{i=1}^{M} \omega_i Y_i\right) / \left(\sum_{i=1}^{M} \omega_i\right)$, where $\omega_i = D_i/\pi_M(\mathbf{X}_i)$.

(E3) $\overline{Y}_{\text{cc}}$: $\overline{Y}_{cc} = \sum_{i=1}^{M} Y_i/M$: The ideal estimator assuming all data are labelled, which is an ideal estimator that can only be obtained in numerical simulations.

(E4) The estimators $\hat{\theta}_{\mathcal{L}}^{\text{sp,T}}$, $\hat{\theta}_{\text{IPW}}^{\text{sp,T}}$, $\hat{\theta}_{\mathcal{L}}^{\text{sp,MAR}}$, $\hat{\theta}_{\text{IPW}}^{\text{sp,MAR}}$, $\hat{\theta}_{\mathcal{L}}^{\text{sp,MCAR}}$ and $\hat{\theta}_{\text{IPW}}^{\text{sp,MCAR}}$ are obtained by estimating the nuisance parameters of $\hat{\theta}^{\text{sp,w}}$ under different working models, where the superscripts 'T', 'MAR' and 'MCAR' represent using the true PS model, (K2) and (K1) as the working model, respectively, and the subscripts '$\mathcal{L}$' and 'IPW' denote estimating $\hat{\boldsymbol{\beta}}_k$ using least squares ($p = 10$) or Lasso ($p = 100$) with labelled data, or using IPW least squares ($p = 10$) or IPW Lasso ($p = 100$) with all data, respectively.

Tables 1 and 2 present the performance of those nine estimators under different settings. In these tables, SD and CP represent the variance and 95% confidence interval estimates obtained through the plug-in method. Eff(M) and Eff(SE) refer to the efficiency estimates calculated based on MSE and SE, respectively. Additional simulation results using the conditional mean functions Y1 and Y3 are provided in the supporting information.

Consistent with our predictions, $\overline{Y}$ is biased when the missingness type is MAR. When the missingness type is MCAR, $\overline{Y}_{IPW}$ remains unbiased, but its efficiency does not improve as the sample size M increases. It is worth noting that when the data are MAR, $\overline{Y}_{IPW}$ is actually less efficient than $\overline{Y}$, although this situation gradually improves as M increases. As expected, $\overline{Y}_{cc}$ performs better than all other estimators in terms of Bias, SE and efficiency. Estimators $\hat{\theta}_{\mathcal{L}}^{sp,T}$, $\hat{\theta}_{IPW}^{sp,T}$, $\hat{\theta}_{\mathcal{L}}^{sp,MAR}$, $\hat{\theta}_{IPW}^{sp,MAR}$, $\hat{\theta}_{\mathcal{L}}^{sp,MCAR}$ and $\hat{\theta}_{IPW}^{sp,MCAR}$ exhibit similar behaviour—they are all consistent estimators and have higher estimation efficiency than $\overline{Y}$. In fact, their estimation efficiency can even be comparable to that of $\overline{Y}_{cc}$. As M increases, both Eff(M) and Eff(SD) increase.

TABLE 2 The results of simulation (Y2) on 1000 simulation runs under $p = 100, M = 1000, 2000, 5000$.

Bias	Bias	SD	SE(P)	CP	Eff(M)	Eff(S)	Bias	SD	SE(P)	CP	Eff(M)	Eff(S)
	MCAR(K1)						MAR(K2)					
	$M = 1000$											
$\overline{Y}$	−0.001	0.065	0.065	0.938	–	-	0.029	0.066	0.066	0.926	-	-
$\overline{Y}_{IPW}$	−0.001	0.065	0.065	0.938	0.000	0.000	0.000	0.094	0.089	0.942	−0.676	−0.417
$\overline{Y}_{cc}$	0.000	0.021	0.021	0.939	0.895	0.676	0.000	0.021	0.021	0.941	0.915	0.681
$\hat{\theta}_{\mathcal{L}}^{sp,T}$	−0.005	0.027	0.028	0.950	0.820	0.584	−0.001	0.031	0.035	0.948	0.815	0.530
$\hat{\theta}_{IPW}^{sp,T}$	−0.005	0.027	0.028	0.951	0.819	0.582	0.003	0.034	0.037	0.938	0.776	0.484
$\hat{\theta}_{\mathcal{L}}^{sp,MAR}$	−0.005	0.027	0.028	0.946	0.820	0.583	−0.001	0.032	0.026	0.891	0.811	0.524
$\hat{\theta}_{IPW}^{sp,MAR}$	−0.005	0.027	0.028	0.947	0.819	0.583	0.001	0.033	0.027	0.878	0.790	0.498
$\hat{\theta}_{\mathcal{L}}^{sp,MCAR}$	−0.005	0.027	0.028	0.950	0.818	0.582	0.001	0.033	0.028	0.903	0.791	0.500
$\hat{\theta}_{IPW}^{sp,MCAR}$	−0.005	0.027	0.028	0.949	0.820	0.584	−0.001	0.033	0.028	0.907	0.793	0.502
	$M = 2000$											
$\overline{Y}$	−0.001	0.046	0.046	0.955	-	-	0.032	0.046	0.047	0.911	-	-
$\overline{Y}_{IPW}$	−0.001	0.046	0.046	0.955	0.000	0.000	0.000	0.067	0.065	0.941	−0.453	−0.471
$\overline{Y}_{cc}$	0.000	0.015	0.015	0.952	0.898	0.680	0.000	0.015	0.015	0.950	0.930	0.676
$\hat{\theta}_{\mathcal{L}}^{sp,T}$	−0.004	0.017	0.018	0.947	0.857	0.629	−0.002	0.019	0.021	0.962	0.877	0.574
$\hat{\theta}_{IPW}^{sp,T}$	−0.004	0.017	0.018	0.947	0.856	0.628	0.001	0.021	0.022	0.953	0.860	0.543
$\hat{\theta}_{\mathcal{L}}^{sp,MAR}$	−0.004	0.017	0.018	0.948	0.856	0.629	−0.002	0.020	0.017	0.907	0.874	0.569
$\hat{\theta}_{IPW}^{sp,MAR}$	−0.004	0.017	0.018	0.950	0.856	0.629	−0.001	0.020	0.018	0.905	0.865	0.552
$\hat{\theta}_{\mathcal{L}}^{sp,MCAR}$	−0.004	0.017	0.018	0.943	0.855	0.628	−0.001	0.021	0.018	0.904	0.856	0.538
$\hat{\theta}_{IPW}^{sp,MCAR}$	−0.004	0.017	0.018	0.948	0.856	0.628	−0.001	0.021	0.018	0.901	0.855	0.537
	$M = 5000$											
$\overline{Y}$	−0.000	0.031	0.029	0.926	-	-	0.034	0.030	0.030	0.795	-	-
$\overline{Y}_{IPW}$	−0.000	0.031	0.029	0.926	0.000	0.000	−0.001	0.043	0.042	0.943	0.102	−0.428
$\overline{Y}_{cc}$	0.000	0.009	0.009	0.943	0.906	0.694	0.000	0.009	0.009	0.945	0.957	0.686
$\hat{\theta}_{\mathcal{L}}^{sp,T}$	−0.002	0.010	0.010	0.949	0.881	0.659	−0.001	0.012	0.012	0.953	0.932	0.611
$\hat{\theta}_{IPW}^{sp,T}$	−0.002	0.010	0.010	0.949	0.881	0.659	−0.001	0.012	0.012	0.954	0.930	0.602
$\hat{\theta}_{\mathcal{L}}^{sp,MAR}$	−0.002	0.010	0.010	0.952	0.881	0.659	−0.001	0.012	0.011	0.934	0.933	0.613
$\hat{\theta}_{IPW}^{sp,MAR}$	−0.002	0.010	0.010	0.952	0.881	0.659	−0.001	0.012	0.011	0.931	0.930	0.603
$\hat{\theta}_{\mathcal{L}}^{sp,MCAR}$	−0.002	0.010	0.010	0.953	0.881	0.659	−0.001	0.013	0.011	0.879	0.919	0.574
$\hat{\theta}_{IPW}^{sp,MCAR}$	−0.002	0.010	0.011	0.952	0.881	0.659	−0.001	0.013	0.011	0.883	0.919	0.574

Taking Tables 1 and 2 as examples, assume that the data are MCAR or MAR, as long as the conditional mean function is correctly specified, whether the missingness mechanism is correct or not does not affect the estimation efficiency. However, it does affect the consistency of the plug-in variance estimation. Assuming MCAR, using the working PS model K1 or K2 can maintain accurate plug-in variance estimation. But if we persist with K1 when data are truly MAR (model misspecification), the plug-in estimate underestimates the true variance—a pattern seen for both $p = 10$ and $p = 100$ with varying M. Compared with low dimension case, a larger M is needed for well-performing plug-in variance estimation due to high dimensionality. With increasing sample size, we can get a more accurate estimate. We need to analyse the method for variance estimation in detail when the sample sizes is small. We leave the theoretical properties and numerical performance in small samples in future work.

5 | REAL DATA APPLICATION

The Los Angeles Homeless Services Authority (LAHSA) is tasked with conducting research and overseeing homeless services throughout Los Angeles County. To inform strategies and policies aimed at addressing homelessness, LAHSA regularly carries out homeless counts and studies to assess the size and characteristics of Los Angeles' homeless population. However, estimating the number of homeless individuals in a metropolitan area presents significant challenges. One major obstacle is that standard US Census designs involve demographers visiting individuals based on their place of residence, which fails to account for most homeless people (Rossi, 1991). Although visiting shelters and service centers provides some data, many homeless individuals are not included in these counts due to factors such as anonymity, nonutilisation of services or other reasons. These challenges in accurately estimating the homeless population arise from its hidden, transient, and hard-to-reach nature.

LAHSA employed stratified spatial sampling of census tracts to study the homeless population in Los Angeles County by dividing the area into strata and selecting representative samples from each stratum. First, LAHSA visited $n_1 = 244$ 'hot tracts' believed to have large homeless populations—areas where homelessness is likely more prevalent or concentrated. Second, they randomly selected and visited $n_2 = 265$ additional tracts from the remaining county tracts, $N = 1545$ tracts went unvisited.

The predictor vector $\mathbf{X}$ includes seven census-derived predictors: Perc.Industrial, Perc.Residential, Perc.Vacant, Perc.Commercial, Perc.OwnerOcc, Perc.Minority and Median Household Income (Kriegler & Berk, 2010; Zhang et al., 2019). These predictors provide tract information on industrial, residential, vacant, commercial percentages, owner-occupied properties, minority populations and median income. The response Y is the homeless count per tract. We aim to estimate the average number of homeless individuals per tract using seven estimators: The first estimator $\overline{Y}$ solely relies on the estimates obtained from labelled data, with $n_1 = 244$ and $n_2 = 265$. The second and third estimators $\hat{\theta}^{\text{MCAR}}_{\text{SSLS}}$, $\hat{\theta}^{\text{MAR}}_{\text{SSLS}}$ are based on the methodology proposed by Zhang et al. (2019). However, there is a difference between them since Zhang et al. (2019) assumes that Y is

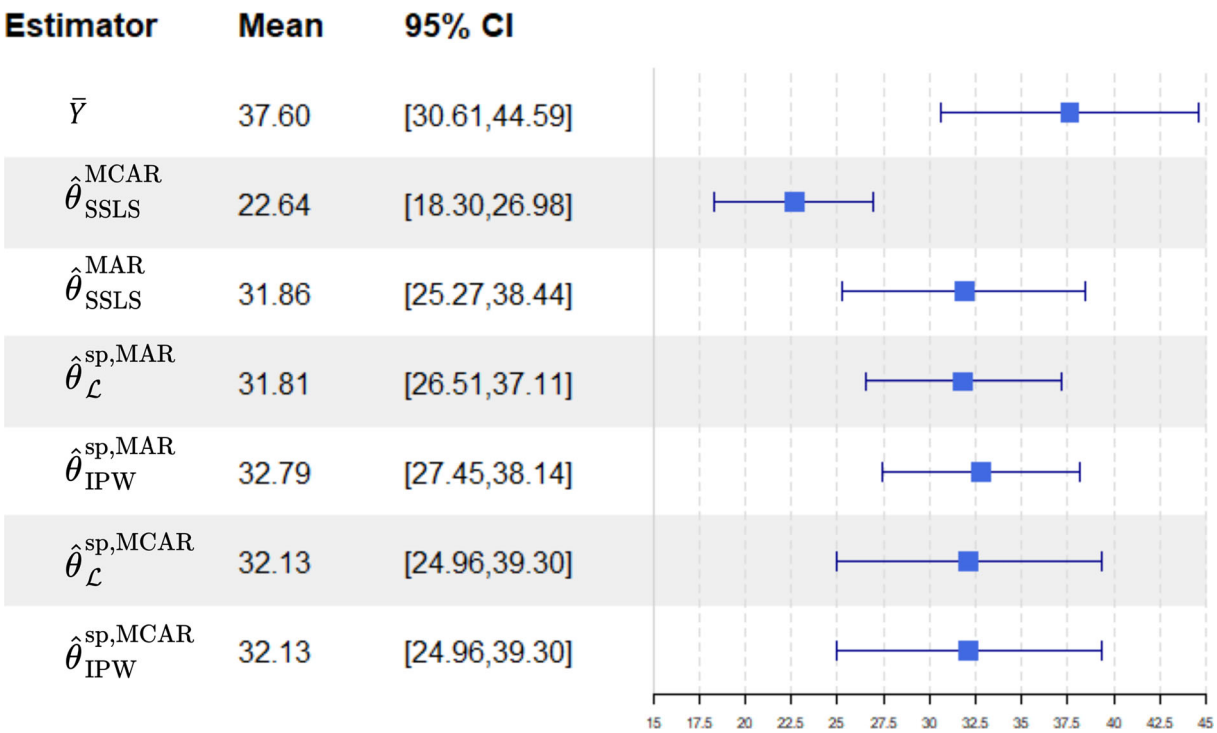

FIGURE 1 Estimates and 95% confidence intervals for the average number of homeless individuals per tract.

missing completely at random. $\hat{\theta}_{\text{SSLS}}^{\text{MCAR}}$ utilises only $n_2 = 265$ labelled data to estimate nuisance parameter, while $\hat{\theta}_{\text{SSLS}}^{\text{MAR}}$ incorporates $n = n_1 + n_2 = 509$ labelled data in the estimation process; variance is estimated following the methodology suggested by Zhang et al. (2019). Besides these, we consider four estimators denoted as $\hat{\theta}_{\mathcal{L}}^{\text{sp,MAR}}$, $\hat{\theta}_{\text{IPW}}^{\text{sp,MCAR}}$, $\hat{\theta}_{\mathcal{L}}^{\text{sp,MCAR}}$ and $\hat{\theta}_{\text{IPW}}^{\text{sp,MCAR}}$. The specific calculation methods for these estimations can be found in the numerical simulation section. We obtain the nuisance parameter estimates using the labelled data with size $n = n_1 + n_2 = 509$.

Figure 1 presents our estimated values for the homeless population in Los Angeles County, along with their corresponding 95% confidence intervals. As expected, $\overline{Y}$, which is based solely on the labelled data, yields the highest estimate. This is likely due to the fact that the labelled data consist of individuals who have been identified as homeless through shelters and service centers, potentially overrepresenting the more visible and accessible homeless population. On the other hand, $\hat{\theta}_{\text{SSLS}}^{\text{MCAR}}$ produces the lowest estimates, possibly because it assumes that the missingness of labels is completely random, which may not adequately capture the complex nature of homelessness. The values of remaining five estimators are relatively close to each other, suggesting that they may be capturing similar aspects of the homeless population. However, $\hat{\theta}_{\mathcal{L}}^{\text{sp,MAR}}$ stands out with the narrowest confidence interval, indicating that it provides a more precise estimate compared to the others. This increased precision could be attributed to the estimator's ability to leverage both labelled and unlabelled data while accounting for the MAR mechanism.

6 | CONCLUSION

We introduce the propensity score model $\pi_M(\mathbf{x})$ to adapt to the SSL-MAR assumption, where missingness of labels depends on both the covariates $\mathbf{X}$ and sample size M. We employ an IPW-NW type estimator to estimate the target parameter $\theta_0 = \mathbb{E}(Y)$. To the best of our knowledge, under the positive overlap assumption in traditional missing data problems, there has been extensive research exploring the performance of IPW-NW type estimators. However, when the positive overlap assumption is violated, its performance has not yet been explored. Moreover, we allow the covariate dimension p to diverge, by introducing a SIM for dimension reduction. We establish the consistency of the IPW-NW type density function estimator in high-dimensional settings. Our proposed method can be easily extended to problems such as M-estimation with fixed-dimensional target parameters. However, when the dimension of the target parameter is diverging, challenges still exist; we need to consider developing more general methods. Furthermore, under the SSL setting, the empirical risk minimisation problem based on U-statistics is also an interesting research direction.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this paper.

ACKNOWLEDGEMENTS

The authors thank the anonymous referees, associate editor and editor for their helpful and constructive comments and suggestions. This research was supported by the National Natural Science Foundation of China (State Key Program: 71931004), the National Key Research and Development Program of China (2021YFA1000100, 2021YFA1000101, 2021YFA1000104), the Science and Technology Commission of Shanghai Municipality (21JS1400501), the National Natural Science Foundation of China (State Key Program: 72331005, Youth Program: 72201101, General Program: 12271171), the General Project of MoE (Ministry of Education) of Humanities and Social Sciences (Youth Fund: 22YJC910013), the Natural Science Foundation of Shanghai (23ZR1419400), and the Shanghai Pujiang Program (21PJC034).

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jin Su https://orcid.org/0009-0005-7255-6010

REFERENCES

Alquier, P., & Biau, G. (2013). Sparse single-index model. *Journal of Machine Learning Research*, *14*(1).

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*(61), 1817–11853.

Azriel, D., Brown, L. D., Sklar, M., Berk, R., Buja, A., & Zhao, L. (2022). Semi-supervised linear regression. *Journal of the American Statistical Association*, *117*(540), 2238–2251.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, *7*(85), 2399–2434.

Bravo, F., Escanciano, J. C., & Keilegom, I. V. (2020). Two-step semiparametric empirical likelihood inference. *The Annals of Statistics*, *48*(1), 1–26.
Cai, T., Li, M., & Liu, M. (2022). Semi-supervised triply robust inductive transfer learning. arXiv preprint arXiv:2209.04977IF: NANANA.
Cai, T. T., & Guo, Z. (2020). Semi-supervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*, 391–419.
Cannings, T. I., & Fan, Y. (2022). The correlation-assisted missing data estimator. *Journal of Machine Learning Research*, *23*(41), 1–49.
Chakrabortty, A., Lu, J., Cai, T. T., & Li, H. (2019). High dimensional M-estimation with missing outcomes: A semi-parametric framework. arXiv preprint arXiv:1911.11345IF: NANANA.
Chapelle, O., Scholkopf, B., & Zien, A. (2009). *Semi-supervised learning*: MIT Press.
Cheng, D., Ananthakrishnan, A. N., & Cai, T. (2021). Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics*, *77*(2), 413–423.
Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199.
Eftekhari, H., & Banerjee, M. (2021). Inference in high-dimensional single-index models under symmetric designs. *Journal of Machine Learning Research*, *22*(1), 1247–1309.
Gronsbell, J., & Cai, T. (2017). Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(3), 579–594.
Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Economic Theory*, *24*(3), 726–748.
Hong, H., Leung, M. P., & Li, J. (2020). Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, *23*(1), 32–47.
Hu, Z., A.Follmann, D., & Qin, J. (2010). Semiparametric dimension reduction estimation for mean response with missing data. *Biometrika*, *97*(2), 305–319.
Hu, Z., A.Follmann, D., & Qin, J. (2012). Semiparametric double balancing score estimation for incomplete data with ignorable missingness. *Journal of the American Statistical Association*, *107*(497), 247–257.
Hu, Z., Follmann, D. A., & Wang, N. (2014). Estimation of mean response via the effective balancing score. *Biometrika*, *101*(3), 613–624.
Huang, M.-Y., & Chan, K. C. G. (2017). Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika*, *104*(3), 583–596.
Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, *58*, 71–120.
Johnson, R., & Zhang, T. (2008). Graph-based semi-supervised learning and spectral kernel design. *IEEE Transactions on Information Theory*, *54*(1), 275–288.
Kallus, N., & Mao, X. (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. arXiv preprint arXiv: 2003.12408IF: NANANA.
Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, *78*(6), 2021–2042.
Kriegler, B., & Berk, R. (2010). Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting. *The Annals of Applied Statistics*, *4*(3), 1234–1255.
Li, K.-C., & Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, *17*(3), 1009–1052.
Lin, H., Zhou, F., Wang, Q., Zhou, L., & Qin, J. (2018). Robust and efficient estimation for the treatment effect in causal inference and missing data problems. *Journal of Econometrics*, *205*(2), 363–380.
Mammen, E., Rothe, C., & Schienle, M. (2016). Semiparametric estimation with generated covariates. *Econometric Theory*, *32*(5), 1140–1177.
Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.
Rossi, P. H. (1991). Strategies for homeless research in the 1990s. *Housing Policy Debate*, *2*(3), 1027–1055.
Rothe, C. (2017). Robust confidence intervals for average treatment effects under limited overlap. *Econometrica*, *85*(2), 645–660.
Song, S., Lin, Y., & Zhou, Y. (2023). A general M-estimation theory in semi-supervised framework. *Journal of the American Statistical Association*, 1–11.
Tu, J., Liu, W., Mao, X., & Xu, M. (2023). Distributed semi-supervised sparse statistical inference. arXiv preprint arXiv:2306.10395IF: NANANA.
Wang, J., & Shen, X. (2007). Large margin semi-supervised learning. *Journal of Machine Learning Research*, *8*(65), 1867–1891.
Wang, J., Shen, X., & Liu, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika*, *95*(1), 149–167.
Wu, X., Huo, Y., Ren, H., & Zou, C. (2024). Optimal subsampling via predictive inference. *Journal of the American Statistical Association*, 1–13.
Zhang, A., Brown, L. D., & Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, *47*(5), 2538–2566.
Zhang, Y., & Bradic, J. (2022). High-dimensional semi-supervised learning: In search of optimal inference of the mean. *Biometrika*, *109*(2), 387–403.
Zhang, Y., Chakrabortty, A., & Bradic, J. (2023). Double robust semi-supervised inference for the mean: Selection bias under MAR labeling with decaying overlap. *Information and Inference: A Journal of the IMA*, *12*(3), 2066–2159.
Zhu, X., & Goldberg, A. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *3*, 1–130.

AUTHOR BIOGRAPHIES

Jin Su is a Ph.D student at the School of Statistics, East China Normal University. Her research focus on semisupervised learning and distributed learning.

Shuyi Zhang is an assistant professor at the Academy of Statistics and Interdisciplinary Sciences, East China Normal University. Her research interests include statistical methods for big data analysis, semisupervised learning, high-dimensional statistics and environmental risk measurement.

Yong Zhou is a professor and dean of School of Statistics, East China Normal University. He has supervised more than 60 Ph.D. students in the areas of Statistics, Management Science and Engineering. He is interested in statistical methods in big data, financial econometrics and risk managements, econometrics, longitudinal and multivariate survival data analysis, among others.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Su, J., Zhang, S., & Zhou, Y. (2024). Solving the missing at random problem in semi-supervised learning: An inverse probability weighting method. *Stat*, *13*(3), e707. https://doi.org/10.1002/sta4.707