Information projection approach to propensity score estimation for handling selection bias under missing at random

Hengfang Wang\(^1\) and Jae Kwang Kim\(^2\) *

\(^1\) School of Mathematics and Statistics, Fujian Normal University
\(^2\) Department of Statistics, Iowa State University

Summary. Propensity score weighting is widely used to improve the representativeness and correct the selection bias in the voluntary sample. The propensity score is often developed using a model for the sampling probability, which can be subject to model misspecification. In this paper, we consider an alternative approach of estimating the inverse of the propensity scores using the density ratio function satisfying the self-efficiency condition. The smoothed density ratio function is obtained by the solution to the information projection onto the space satisfying the moment conditions on the balancing scores. By including the covariates for the outcome regression models only in the density ratio model, we can achieve efficient propensity score estimation. Penalized regression is used to identify important covariates. We further extend the proposed approach to the multivariate missing case. Some limited simulation studies are presented to compare with the existing methods. 

Keywords: Calibration estimation; Density ratio model; Self-efficiency; Missing data

\(^*\) jkim@iastate.edu, Department of Statistics, Iowa State University, Ames, Iowa, U.S.A.
1 Introduction

In statistical analysis using sample data, the concept of representativeness is crucial. The fundamental issue is that the final sample might be subject to selection bias and might not accurately reflect the target population. A naive analysis that ignores selection bias can produce incorrect results. Another way to frame the selection bias problem is as a missing data problem. Making valid statistical inference with missing data is a fundamental problem in statistics (Little and Rubin 2019, Kim and Shao 2021).

The propensity score (PS) weighting approach is widely used for adjusting the selection bias in the final sample. Frequently, the propensity score is calculated using the selection probability model. In principle, regression models for binary responses, such as logistic regression, can be used to estimate the selection probability in the presence of observed auxiliary data. The target parameter can then be estimated using an inverse probability weighting estimator to provide an unbiased estimate. However, correctly specifying the propensity score model can be tricky, and we frequently lack sufficient understanding of the selection mechanism. Furthermore, the final estimation can be unstable when some propensity scores are close to zero (Ma and Wang 2020). The bias and variance of the resulting estimator are likely to be amplified by the nature of such an ‘inverse’ fashion.

The existing methods for propensity score estimation are either based on the maximum likelihood method (Rosenbaum and Rubin 1983, Robins et al. 1994) or calibration method (Folsom 1991, Tan 2010, Kott and Chang 2010, Graham et al. 2012, Hainmueller 2012, Imai and Ratkovic 2014, Chan et al. 2016, Zhao 2019). The calibration method can also be justified under an outcome model, which gives a doubly robust flavor. However, the objective function for calibration estimation is not fully agreed.

In this paper, we introduce the self-efficiency condition as a new paradigm for PS estimation. The self-efficiency condition means that the resulting PS estimator is equivalent to the prediction-based estimator using the outcome response model. Due to the equivalence, the self-efficient propensity score estimator is more efficient than existing approaches. To build a self-efficient PS estimation, we first model density ratios with information projection (Csiszár and Shields 2004), which minimizes the Kullback-Leibler divergence measure for models with moment constraints on the balance scores. Density ratio models are widely employed to address classification and two-sample problems (Qin 1998, Nguyen et al. 2010, Chen and Liu 2013). To our best knowledge, however, it has not been studied in the context of correcting selection bias. Once the density ratio model has been constructed using the information projection approach, the model parameters can be estimated using the empirical likelihood technique (Qin and Lawless 1994).
The self-efficient PS estimation is also closely related to weight smoothing. Information projection of the density ratio function combined with calibration estimation is used to accomplish weight smoothing. We give some asymptotic theories for the smoothed PS estimator utilizing the balance score function. The proposed estimator is doubly robust and locally efficient in the sense of [Robins et al. (1994)]. By applying the weight smoothing based on the covariate in the outcome model, we can achieve the efficiency gain. In addition, a two-step estimation approach is designed to facilitate parameter estimation computation. Variance estimation of the PS estimator can be implemented using either the linearization method or bootstrap.

Furthermore, the proposed paradigm includes discussing propensity score estimation in high-dimensional covariate scenarios. When there are more auxiliary variables than we require, the PS estimator can be inefficient. Only significant covariates for the outcome model should be included in the log density ratio model for effective PS estimation. We consider penalized regression to identify important covariates. Given the observed study variable and the corresponding auxiliary variables, we can implement penalized regression methods to select important covariates and obtain an efficient PS estimator.

Finally, using the proposed framework, the PS estimation can be easily extended to handle multivariate missing data. We can partition the sample into multiple groups based on missing patterns and apply the density ratio estimation method to obtain the inverse propensity scores. By constructing a self-efficient PS estimation, we can combine information from multiple sources with different missing patterns.

The paper is organized as follows. In Section 2, the basic setup is introduced. In Section 3, the proposed method is developed using the information projection technique. In Section 4, its asymptotic properties are investigated. In Section 5, the proposed method is extended to handle high dimensional covariate cases. In Section 6, an extension to multivariate missing data setup is developed. Two limited simulation studies are presented in Section 7 to compare with other existing methods and to understand the effect of choice of calibration variables. Concluding remarks are made in Section 8.

## 2 Basic Setup

Suppose that the parameter of interest $\theta \in \mathbb{R}^p$ can be written as the unique solution to $\mathbb{E}\{U(\theta; X, Y)\} = 0$, where $Y$ is the study variable that is subject to missingness, $X$ is the auxiliary variable that is always observed, and $U$ is a smooth function of $\theta$ with nonsingular gradient matrix $\mathbb{E}\{\nabla \theta U(\theta)\}$. Thus, the joint density of $(X, Y)$ is completely
unspecified except for the moment condition $\mathbb{E}\{U(\theta; X, Y)\} = 0$. We further assume that the second moments of $U(\theta; X, Y)$ exist.

Suppose we have an $N$ independently and identically distributed realization of $(X, Y)$, denoted as $\{(x_i, y_i) : i = 1, \ldots, N\}$. Since $Y$ is subject to missingness, the actual dataset is $\{(x_i, \delta_i y_i, \delta_i) : i = 1, \ldots, N\}$, where $\delta_i$ is the sampling indicator variable defined as

$$\delta_i = \begin{cases} 1 & \text{if } y_i \text{ is observed}, \\
0 & \text{otherwise}. \end{cases}$$

We assume that the sampling mechanism is missing at random (MAR) in the sense of Rubin (1976). That is,

$$Y \perp \delta \mid X.$$  \hspace{1cm} (1)

Under MAR, we are interested in finding the propensity weights $\omega(x_i)$ such that the solution to

$$\sum_{i=1}^{N} \delta_i \omega(x_i) U(\theta; x_i, y_i) = 0$$  \hspace{1cm} (2)

leads to consistent estimation of $\theta$. Equation (2) is called the propensity score equation. One popular choice for $\omega(x)$ is

$$\omega(x) = \{\mathbb{P}(\delta = 1 \mid x)\}^{-1}. \hspace{1cm} (3)$$

There are two problems with the choice in (3). First, we do not know the true propensity score function $\mathbb{P}(\delta = 1 \mid x)$. Second, the resulting estimator is not necessarily efficient. To resolve these problems, we may consider the following conditions:

$$\sum_{i=1}^{N} \delta_i \omega(x_i) U(\theta; x_i, y_i) = \sum_{i=1}^{N} \left[ \delta_i U(\theta; x_i, y_i) + (1 - \delta_i) \hat{\mathbb{E}}\{U(\theta; x_i, Y) \mid x_i\} \right], \hspace{1cm} (4)$$

where $\hat{\mathbb{E}}\{U(\theta; x_i, Y) \mid x_i\}$ is the best predictor of $U(\theta; x_i, y_i)$ for $\delta_i = 0$. Condition (4) is attractive as the solution to the PS estimating equation in (2) is equivalent to the expected estimating equation, which leverages the best predictors of the unobserved parts of the population estimating equation. We call condition (4) self-efficiency condition, as we can obtain an efficient estimator of $\theta$ by solving the expected estimating equation (Wang and Pepe, 2000). The term “self” is employed to emphasize that efficiency is achieved without including additional terms to the propensity score equation.
To compute the conditional expectation in (4), one could make an extra assumption on the outcome model \( y \mid x \) and estimate the parameters in the model. To avoid the difficulty associated with the modeling for \( y \mid x \), we assume that

\[
\mathbb{E}\{U(\theta; x, Y) \mid x\} \in \text{span}\{1, b_1(x), \ldots, b_L(x)\} := \mathcal{H}
\]

for some \( b_1(x), \ldots, b_L(x) \). Once condition (5) is met, the additional modeling for \( y \mid x \) can be omitted. Under (5), by taking the conditional expectations on both terms in (4), we can check that the self-efficiency condition implies

\[
\sum_{i=1}^{N} \delta_i \omega(x_i) b(x_i) = \sum_{i=1}^{N} b(x_i),
\]

where \( b(x) = (1, b_1(x), \ldots, b_L(x)) \) is a vector of basis functions in \( \mathcal{H} \). Thus, the self-efficient propensity score method achieves the balancing property for \( b(x) \). The balancing property is also called calibration property in survey sampling ([Deville and Särndal] 1992, [Fuller] 2009, [Wu and Thompson] 2020).

Another implication of the self-efficiency is related to weight smoothing. Note that, under assumption (5), the self-efficiency condition in (4) implies that \( \omega(x_i) \) is a function of \( b_1(x), \ldots, b_L(x) \), which means that the final PS weights are functions of the covariates in the outcome regression model. By including the covariates in the outcome model only, we can achieve weight smoothing and efficient estimation.

Thus, under assumption (5), it makes sense to impose the self-efficiency for obtaining the propensity weights. How to uniquely determine \( \omega(x) \) to satisfy the self-efficiency condition is our main research problem. We address this problem by introducing the density ratio function and information projection in the next section.

### 3 Proposed method

#### 3.1 Density ratio function

To introduce the density ratio function, we use \( f(\cdot) \) to denote generic density functions. Further, let \( f_j(x) \) denote the conditional density of \( X \) given \( \delta = j \) for \( j = 0, 1 \). Using this notation, we define the following density ratio function:

\[
r(x) = \frac{f_0(x)}{f_1(x)}.
\]
By Bayes theorem, we obtain
\[
\frac{P(\delta = 0 \mid x)}{P(\delta = 1 \mid x)} = \frac{P(\delta = 0)}{P(\delta = 1)} \times r(x).
\]
Assuming \( c = P(\delta = 0)/P(\delta = 1) \) is known, we can express
\[
\frac{1}{P(\delta = 1 \mid X)} = 1 + c \cdot r(X)
\]
and use
\[
\omega_i = \frac{1}{P(\delta_i = 1 \mid x_i)} = 1 + c \cdot r(x_i)
\]
as the propensity score weight for unit \( i \) with \( \delta_i = 1 \). So, our propensity score weighting problem reduces to density ratio function estimation problem. Note that we can easily estimate \( c \) by
\[
\hat{c} = \frac{N_0}{N_1}, \quad N_1 = \sum_{i=1}^{N} \delta_i \quad \text{and} \quad N_0 = N - N_1.
\]
Let \( b(x) = (b_1(x), \cdots, b_L(x)) \) be a vector of basis functions of \( \mathcal{H} \). Condition [5] implies that, as far as estimation of \( \theta \) is concerned, the MAR condition holds conditional on \( b(X) \), that is,
\[
Y \perp \delta \mid b(X).
\] (8)
According to [Rosenbaum and Rubin (1983)], \( b(x) \) in (8) is called the balancing score function.

Using \( b(x) \) satisfying [8], we can construct
\[
r^*(x) = \mathbb{E} \{ r(x) \mid b(x), \delta = 1 \},
\] (9)
where \( r(x) = f_0(x)/f_1(x) \) is the true density ratio function. To distinguish \( r^*(x) \) from the original density ratio function \( r(x) \), we may call \( r^*(x) \) the smoothed density ratio function in the sense that it satisfies
\[
\frac{1}{N_1} \sum_{i=1}^{N} \delta_i r^*(x_i)[1, b(x_i)] = \frac{1}{N_0} \sum_{i=1}^{N} (1 - \delta_i) [1, b(x_i)].
\] (10)
Thus, the smoothed density ratio function \( r^*(x) \) is regarded as a low-dimensional projection of \( r(x) \) onto the space satisfying [10].

We can use \( r^*(x) \) in [9] to construct a smoothed PS estimating function of \( \theta \) as follows:
\[
\hat{U}_{SPS}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta_i \{ 1 + (N_0/N_1)r^*(x_i) \} U(\theta; x_i, y_i).
\] (11)
Furthermore, since \( r(\mathbf{x}) \) is a function of \( \mathbf{x} \) only, we obtain
\[
\mathbb{E}\{ r(\mathbf{x}) \mid \mathbf{b}(\mathbf{x}), y, \delta = 1 \} = \mathbb{E}\{ r(\mathbf{x}) \mid \mathbf{b}(\mathbf{x}), \delta = 1 \},
\]
(12)
where \( r(\mathbf{x}) = f_0(\mathbf{x})/f_1(\mathbf{x}) \) is the true density ratio function. Note that result (12) implies that
\[
\mathbb{E}\{ \hat{U}_{\text{PS}}(\theta) \mid \mathbf{b}(\mathbf{x}), y, \delta \} = \hat{U}_{\text{SPS}}(\theta),
\]
where
\[
\hat{U}_{\text{PS}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta_i \{1 + (N_0/N_1) r(\mathbf{x}_i)\} U(\theta; \mathbf{x}_i, y_i).
\]
Therefore, we can establish the following result.

**Proposition 1** Under (8), we obtain
\[
\mathbb{E}\{ \hat{U}_{\text{PS}}(\theta) \} = \mathbb{E}\{ \hat{U}_{\text{SPS}}(\theta) \},
\]
(13)
and
\[
\nabla\{ \hat{U}_{\text{PS}}(\theta) \} \geq \nabla\{ \hat{U}_{\text{SPS}}(\theta) \}.
\]
(14)

Proposition 1 implies that the solution to \( \hat{U}_{\text{SPS}}(\theta) = 0 \) is more efficient than the solution to \( \hat{U}_{\text{PS}}(\theta) = 0 \). Such phenomenon has been recognized by Little and Vartivarian (2005), Beaumont (2008), Kim and Skinner (2013) and Park et al. (2019). To find a model of the smoothed density ratio function \( r^*(\mathbf{x}) \) in (9), we introduce information projection in the following subsection.

### 3.2 Information projection

To motivate the proposed method, suppose that we have two probability distributions \( \mathbb{P}_0, \mathbb{P}_1 \), with \( \mathbb{P}_0 \) absolutely continuous with respect to \( \mathbb{P}_1 \). For simplicity, we also assume that \( \mathbb{P}_k \) are absolutely continuous with respect to Lebesgue measure \( \mu \), with density \( f_k \) with support \( \mathcal{X} \subset \mathbb{R}^p \), for \( k = 0, 1 \). The Kullback-Leibler divergence between \( \mathbb{P}_0 \) and \( \mathbb{P}_1 \) is defined by
\[
D(\mathbb{P}_0 \parallel \mathbb{P}_1) = \int \log \left( \frac{d\mathbb{P}_0}{d\mathbb{P}_1} \right) d\mathbb{P}_0 = \int \log \left( \frac{f_0}{f_1} \right) f_0 d\mu.
\]
Finding the density ratio function can be formulated as an optimization problem using the KL divergence. Specifically, we wish to minimize

$$Q(f_0) = \int \log \left( \frac{f_0}{f_1} \right) f_0 d\mu$$  \hspace{1cm} (15)

with respect to $f_0$ satisfying some moment conditions of the variables that are available from the whole sample. Let $b(x)$ be the given balancing score obtained from assumption (5). Because $x_i$ are observed throughout the sample, we can estimate $\mathbb{E}\{b(X)\}$ from the sample. Given $\mathbb{E}\{b(X)\}$, to utilize the auxiliary information, we have

$$p \int b(x) f_1(x) d\mu + (1-p) \int b(x) f_0(x) d\mu = \mathbb{E}\{b(X)\},$$  \hspace{1cm} (16)

where $p = P(\delta = 1)$. Thus, we wish to find the minimizer of $Q(f_0)$ in (15) subject to the constraint in (16). The following lemma gives the form of the solution to the optimization problem.

**Lemma 1** The minimizer of $Q(f_0)$ in (15) subject to the constraint in (16) is

$$f^*_0(x) = f_1(x) \times \frac{\exp \{ \lambda_1^T b(x) \}}{\mathbb{E}_1 \left[ \exp \{ \lambda_1^T b(x) \} \right]},$$  \hspace{1cm} (17)

where $\lambda_1$ is chosen to satisfy (16).

The solution (17) is obtained using the information projection method in information geometry. Roughly speaking, constraint in (16) is a moment condition on $f_0(x)$ to satisfy (9). Thus, we can use Lagrange multiplier method to obtain the solution as an exponential tilting form. Note that (17) is equivalent to assuming a model for the density ratio function. That is, our model is

$$\log \{ r^*(x; \lambda) \} = \lambda_0 + \lambda_1^T b(x),$$  \hspace{1cm} (18)

where $\lambda_0$ is the normalizing constant satisfying

$$\int r^*(x; \lambda) f_1(x) d\mu = 1.$$

By (7), the log-linear density ratio model (18) is equivalent to the logistic regression model for the sample selection probability

$$P(\delta = 1 \mid x) = \frac{1}{1 + (1/p - 1) \cdot \exp \left\{ \lambda_0 + \lambda_1^T b(x) \right\}}.$$  \hspace{1cm} (19)
Thus, our framework provides another justification of the logistic regression model for the sample selection probability that has been considered in Folsom (1991), Kott (2006), and Tan (2020), among others. But, the parameter estimation discussed in §3.3 is different from the existing methods. Also, the covariates in the parametric model in (19) is derived from the outcome model assumption in (5). Nonetheless, including other covariates in the calibration weighting can make the resulting PS estimator robust against model misspecification (Han and Wang, 2013, Chan et al., 2014, Chen and Haziza, 2017, Yang et al., 2020).

Using (17), we can express (16) as

\[ p \int b(x) \left[ 1 + \frac{N_0}{N_1} \cdot \exp \{ \lambda_0 + \lambda_1^T b(x) \} \right] f_1(x)d\mu = \mathbb{E}\{b(X)\}, \tag{20} \]

where \( \lambda_0 \) satisfies

\[ \int \exp \{ \lambda_0 + \lambda_1^T b(x) \} f_1(x)d\mu = 1. \tag{21} \]

The function \( r^*(x) = \exp(\lambda_0 + \lambda_1^T b(x)) \) can be viewed as the projection of the true density ratio function onto the space satisfying (16).

### 3.3 Model parameter estimation

We now wish to estimate the parameters in the log-linear density ratio model in (18) and estimate \( \theta \). Note that the parameter is defined through \( \mathbb{E}\{U(\theta; X, Y)\} = 0 \), which can be expressed as

\[ p \int U(\theta; x, y) \left[ 1 + \frac{N_0}{N_1} \cdot \exp \{ \lambda_0 + \lambda_1^T b(x) \} \right] f_1(x, y)d\mu = 0 \tag{22} \]

as an integral equation for \( \theta \), where \( \lambda_0 \) and \( \lambda_1 \) satisfy the constraints (20) and (21).

To estimate the parameters, we use

\[ \hat{P}_1(x, y) = \frac{1}{N_1} \sum_{i=1}^{N} \delta_i \mathbb{I}\{(x, y) = (x_i, y_i)\} \]

to find the minimizer of \( D(\hat{P}_1 \parallel P_1) \) among \( P_1 \) satisfying the integral constraints (20), (21), and (22). The problem can be formulated as an optimization problem using the empirical likelihood (EL) method of Qin and Lawless (1994). In the EL method, a fully
nonparametric density $p_i$ can be assumed to replace $f_1(x, y)$ in the integral equations. That is, we maximize

$$\ell(p) = \sum_{i=1}^{N} \delta_i \log(p_i)$$

subject to

$$\sum_{i=1}^{N} \delta_i p_i = 1,$$  \hspace{1cm} \text{(23)}

$$\frac{N_1}{N} \sum_{i=1}^{N} \delta_i b(x_i) \left\{1 + \frac{N_0}{N_1} \cdot \exp\{\lambda_0 + \lambda_1^T b(x_i)\}\right\} p_i = \frac{1}{N} \sum_{i=1}^{N} b(x_i),$$  \hspace{1cm} \text{(24)}

$$\sum_{i=1}^{N} \delta_i U(\theta; x_i, y_i) \left\{1 + \frac{N_0}{N_1} \cdot \exp\{\lambda_0 + \lambda_1^T b(x_i)\}\right\} p_i = 0,$$  \hspace{1cm} \text{(25)}

and

$$\sum_{i=1}^{N} \delta_i \exp\{\lambda_0 + \lambda_1^T b(x_i)\} p_i = 1.$$  \hspace{1cm} \text{(26)}

The solution to the above optimization problem lead to $\hat{p}_i = \hat{p}_i(\lambda, \theta)$ and we obtain the profile empirical likelihood

$$\ell_e(\theta, \lambda) = \sum_{i=1}^{N} \delta_i \log\{\hat{p}_i(\lambda, \theta)\}.$$  \hspace{1cm} \text{(27)}

The maximizer of $\ell_e(\theta, \lambda)$ can be used as the final estimator of $\theta$. However, such a joint optimization of $\theta$ and $\lambda$ is computationally expensive.

To introduce an alternative computation, note that the form of the smoothed propensity weights

$$\tilde{\omega}_i^* = \left\{1 + \frac{N_0}{N_1} \cdot \exp\left(\hat{\lambda}_0 + \hat{\lambda}_1^T b(x_i)\right)\right\} \hat{p}_i$$

depends on $\theta$ only through $\hat{\theta}$. Thus, we can first exclude (25) in the EL optimization to remove the dependency on $\theta$ in $\tilde{\omega}_i^*$ and then obtain $\hat{\theta}$ by solving $\sum_{i=1}^{N} \delta_i \tilde{\omega}_i^* U(\theta; x_i, y_i) = 0$ for $\theta$. This two-step procedure will greatly simplify the computation.

Note that maximizing $\ell(p) = \sum_{i=1}^{N} \delta_i \log(p_i)$ subject to (23) only gives $\hat{p}_i = 1/N_1$. Imposing (24) and (26) to the unconstrained optimization does not play any role on the
final solution because $\lambda_0$ and $\lambda_1$ are unknown. Thus, the solution to the optimization problem remains the same:

$$\hat{p}_i = 1/N_1$$

with $\hat{\lambda}_0$ and $\hat{\lambda}_1$ satisfying

$$\sum_{i=1}^{N} \delta_i \left\{ 1 + (N_0/N_1) \cdot \exp\{\hat{\lambda}_0 + \hat{\lambda}_1^T b(x_i)\} \right\} [1, b(x_i)] = \sum_{i=1}^{N} [1, b(x_i)],$$

(28)

which is a calibration equation for $[1, b(x)]$.

Once the model parameters in (18) are estimated, we can compute the smoothed propensity score estimating function

$$\hat{U}_{SPS}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta_i \hat{\omega}_i^* U(\theta; x_i, y_i),$$

(29)

where

$$\hat{\omega}_i^* = 1 + (N_0/N_1) \cdot \exp\{\hat{\lambda}_0 + \hat{\lambda}_1^T b(x_i)\}$$

(30)

and $\hat{\lambda}_0$ and $\hat{\lambda}_1$ are computed from (28). The final estimator of $\theta$ is obtained by solving $\hat{U}_{SPS}(\theta) = 0$.

The proposed propensity score estimating function in (29) satisfies the self-efficiency. To see this, without loss of generality, assume that $\theta$ is a scalar. Note that, for a fixed $\theta$, we can write

$$\hat{U}_{SPS}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta_i \hat{\omega}_i^* U_i = \frac{1}{N} \sum_{i=1}^{N} \left\{ m_i(\beta) + \delta_i \hat{\omega}_i^* \{ U_i - m_i(\beta) \} \right\}$$

(31)

where $U_i = U(\theta; x_i, y_i)$ and $m_i(\beta) = \beta_0 + \sum_{j=1}^{L} \beta_j b_j(x_i)$ for any $\beta_0, \beta_1, \ldots, \beta_L$. Now, since $\hat{\omega}_i^* = 1 + (N_0/N_1) \cdot \exp\{\hat{\lambda}_0 + \hat{\lambda}_1^T b(x_i)\}$, the smoothed PS estimator in (31) is algebraically equivalent to

$$\hat{U}_{SPS}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \{ \delta_i U_i + (1 - \delta_i)m_i(\beta) \} + \frac{1}{N} \sum_{i=1}^{N} \delta_i \exp\{\hat{\lambda}_0 + \hat{\lambda}_1^T b(x_i)\} \{ U_i - m_i(\beta) \}.$$  

(32)

The second term in the right hand side of (32) is zero at $\beta = \hat{\beta}$ where $\hat{\beta}$ satisfies

$$\sum_{i=1}^{N} \delta_i \exp\{\hat{\lambda}_0 + \hat{\lambda}_1^T b(x_i)\} \left\{ U_i - m_i(\hat{\beta}) \right\} = 0.$$
Thus, we have

$$\frac{1}{N} \sum_{i=1}^{N} \delta_i \hat{\omega}_i U_i = \frac{1}{N} \sum_{i=1}^{N} \left\{ \delta_i U_i + (1 - \delta_i) m_i(\beta) \right\}.$$  \hspace{1cm} (33)

Therefore, the proposed estimator satisfies the self-efficiency condition in (4).

Equation (33) means that the final inverse propensity weights $\hat{\omega}_i$ do not directly use the regression model for prediction, but it implements regression prediction indirectly. Thus, the smoothed PS estimator incorporates the outcome regression model through calibration equation in (28) and achieves the self-efficiency. On the other hand, model calibration of Wu and Sitter (2001) uses the outcome model directly in the calibration equation. Similar ideas have been considered in the context of nonparametric calibration estimation in survey sampling. For examples, Montanari and Ranalli (2005) used a single-layer Neural Network model and Breidt et al. (2005) used penalized Spline model for nonparametric calibration estimation.

**Remark 1** Assumption (5) states that the basis functions do not depend on $\theta$, which does not always hold. If the basis functions in (5) depends on $\theta$, that is,

$$E\{U(\theta; x, Y) \mid x \} \in \text{span}\{1, b_1(x; \theta), \ldots, b_L(x; \theta)\},$$  \hspace{1cm} (34)

we can use the following iterative procedure.

1. Given the current parameter estimate $\hat{\theta}^{(t)}$, find $\hat{b}_{ki}^{(t)} = b_k(x_i; \hat{\theta}^{(t)})$, for $k = 1, \ldots, L$, such that (34) holds for $\theta = \hat{\theta}^{(t)}$.

2. Estimate the parameter in

$$\hat{\omega}_i^{(t)} = 1 + (N_0/N_1) \exp \left\{ \phi_0 + \phi_1 \hat{b}_{i1}^{(t)} + \cdots + \phi_L \hat{b}_{L1}^{(t)} \right\}$$  \hspace{1cm} (35)

by apply the calibration equation on $[1, \hat{b}_{i1}^{(t)}, \ldots, \hat{b}_{Li}^{(t)}]$.

3. Use

$$\hat{\omega}_i^{(t)} = 1 + (N_0/N_1) \exp \left\{ \hat{\phi}_0 + \hat{\phi}_1 \hat{b}_{i1}^{(t)} + \cdots + \hat{\phi}_L \hat{b}_{Li}^{(t)} \right\}$$

$$\sum_{i=1}^{N} \delta_i \hat{\omega}_i^{(t)} U(\theta; x_i, y_i) = 0.$$
4. Set $t = t + 1$ and goto Step 1 until convergence.

This is essentially an extension of the EM algorithm [Dempster et al., 1977] applied to regression model. Step (a)-(b) corresponds to the E-step and Step (c) corresponds to the M-step. In the E-step, instead of using a parametric model, we use the information projection to compute the conditional expectation under assumption $[\delta]$. 

4 Statistical properties

We now discuss the asymptotic properties of the smoothed PS estimator $\hat{\theta}_{SPS}$ which is the solution to $\hat{U}_{SPS}(\theta) = 0$ using $\hat{U}_{SPS}(\theta)$ in (29). To formally discuss the asymptotic properties of $\hat{\theta}_{SPS}$, let $\hat{U}_2(\lambda)$ be the estimating function for $\lambda = (\lambda_0, \lambda_1)$. By (28), we can express

$$ \hat{U}_2(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \delta_i \omega^*(x_i; \lambda)[1, b(x_i)] - \frac{1}{N} \sum_{i=1}^{N} [1, b(x_i)], $$

(36)

where $\omega^*(x; \lambda) = 1 + (N_0/N_1) \exp\{\lambda_0 + \lambda^T b(x)\}$. Let $\lambda^*$ be the true parameter value in the log-linear density ratio model in (18) so that $\exp\{\lambda_0^* + b^T(x)\lambda_1^*\} = r^*(x)$. By the constraint in (16), we have $E\{\hat{U}_2(\lambda^*)\} = 0$, where the reference distribution is the conditional distribution of $x$ given $\delta$. The unbiasedness of $\hat{U}_2(\lambda^*)$ can also be derived under the selection probability model associated with (18). That is, under the selection probability model

$$ P(\delta = 1 \mid x) = \frac{\exp\{\lambda_0^* + b^T(x)\lambda_1^*\}}{1 + \exp\{\lambda_0 + b^T(x)\lambda_1\}} := \pi^*(x), $$

(37)

we can also obtain $E\{\hat{U}_2(\lambda^*)\} = 0$.

Thus, as long as the sufficient conditions for $E\{\hat{U}_2(\lambda^*)\} = 0$ are satisfied, we can establish the weak consistency of $\hat{\lambda}$ and apply the standard Taylor linearization to obtain the following theorem, where $\lambda^*$ is the true value of parameter in model (18).

The regularity conditions and the proof are presented in the Supplementary Material.

**Theorem 1** Assume that density ratio model (18) holds. Under the regularity conditions described in the Supplementary Material, we have

$$ \hat{U}_{SPS}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left[ z_i^T \beta^* + \delta_i \omega^*(x_i; \lambda^*)\{U(\theta; x_i, y_i) - z_i^T \beta^*\} \right] + o_p(N^{-1/2}), $$

(38)
where \( z_i^T = (1, b^T(x_i)) \), \( \omega^*(x; \lambda) = 1 + (N_0/N_1) \exp\{\lambda_0 + \lambda^T b(x)\} \), and \( \beta^* = \beta^*(\theta) \) is the probability limit of the solution to

\[
\sum_{i=1}^N \delta_i \exp(z_i^T \lambda^*) \{U(\theta; x_i, y_i) - z_i^T \beta\} z_i = 0.
\]

(39)

By Theorem 1, we can obtain, ignoring the smaller order terms,

\[
\hat{U}_{\text{SPS}}(\theta) - \hat{U}_N(\theta) = \frac{1}{N} \sum_{i=1}^N \delta_i \omega^*(x_i; \lambda^*) - 1 \{U(\theta; x_i, y_i) - z_i^T \beta^*\}
\]

(40)

where \( \hat{U}_N(\theta) = N^{-1} \sum_{i=1}^N U(\theta; x_i, y_i) \). Under the selection probability model (37), the first term of (40) has zero expectation. Also, if the regression outcome model satisfies

\[
E_t \left[ U(\theta; x, y) \mid x \right] \in \mathcal{H} \equiv \text{span}\{1, b_1(x), \ldots, b_L(x)\},
\]

(41)

then we obtain \( z_i^T \beta^* = E\{U(\theta; x, Y) \mid x_i\} \) and the second term of (40) has zero expectation. Thus, the expression in (40) gives the doubly robust property (Bang and Robins, 2005, Tsiatis, 2007, Cao et al., 2009, Kim and Haziza, 2014) in that the resulting estimator is justified either the outcome regression model or the selection probability model is correctly specified.

Corollary 1

Suppose that the assumptions for Theorem 1 hold. If \( \hat{U}(\theta; x) = E\{U(\theta; x, Y) \mid x\} \) satisfies (41), we obtain \( \hat{U}(\theta; x) = z_i^T \beta^* \) and

\[
\sqrt{N} \left( \hat{\theta}_{\text{SPS}} - \theta_0 \right) \xrightarrow{\mathcal{L}} N(0, \tau^{-1} V_1 (\tau^{-1})^T),
\]

(42)

where \( \tau = E\{\partial U(\theta_0; X, Y)/\partial \theta^T\} \), \( V_1 = \mathbb{V} \{\hat{U}(\theta_0; X)\} + \mathbb{E} \left[ \delta(\omega^*(X; \lambda^*))^2 \mathbb{V}\{U(\theta_0; X, Y) \mid X\} \right] \),

(43)

and \( \omega^*(X; \lambda) = 1 + (N_0/N_1) \exp\{\lambda_0 + \lambda^T b(x)\} \).

Remark 2 (Remark 2)

The variance term in (43) deserves a further discussion. Let \( \pi(x) \) be the true selection probability. We may use the true selection probability to construct a doubly robust estimator of the form

\[
\hat{U}_{\text{DR}}(\theta) = \frac{1}{N} \sum_{i=1}^N \left[ \hat{U}(\theta; x_i) + \frac{\delta_i}{\pi(x_i)} \{U(\theta; x_i, y_i) - \hat{U}(\theta; x_i)\} \right].
\]

(44)
Now, by (7) and (9), we have
\[
\omega^*(x) = \mathbb{E} \left\{ \frac{1}{\pi(x)} \mid b(x), \delta = 1 \right\}.
\]

By Jensen’s inequality, we can obtain
\[
V \{ \hat{U}_{DR}(\theta) \} = V \{ \hat{U}(\theta; X) \} + \mathbb{E} \left[ \delta \{ \pi(X) \}^{-2} V \{ U(\theta; X, Y) \mid X \} \right] \geq V \{ \hat{U}(\theta; X) \} + \mathbb{E} \left[ \delta \{ \omega^*(x) \}^{-2} V \{ U(\theta; X, Y) \mid X \} \right] = V \{ \hat{U}_{SPS}(\theta) \}.
\]
Thus, the smoothed PS estimator is more efficient than the doubly robust estimator in (44). Furthermore, if the selection probability model satisfies (37), we have equality in (45) and the smoothed PS estimator is locally optimal in the sense that it achieves the variance lower bound of Robins et al. (1994).

**Remark 3 (Remark 3)** Condition (41) is a critical condition for the validity of the proposed estimator. If the space \( \mathcal{H} = \text{span}\{b(x)\} \) is large enough, then (41) is likely to be satisfied and result (42) will hold. However, if \( \mathcal{H} \) is too large, then we can find \( \mathcal{H}_0 \subset \mathcal{H} \) such that \( \mathbb{E}(Y \mid x) \in \mathcal{H}_0 \). In this case, we can construct a smoothed density ratio function using the basis functions in \( \mathcal{H}_0 \) only. By Remark 2, it is more efficient than the smoothed PS estimator using the basis function in \( \mathcal{H} \). Therefore, including unnecessary calibration variables in the calibration equation will increase the variance. This is consistent with the empirical findings of Brookhart et al. (2006) and Shortreed and Ertefaie (2017). We will discuss this result further in Section 5. See also the second simulation study in Section 7.2.

We now discuss variance estimation. Using (38), we obtain
\[
\hat{\theta} - \theta_0 = -\tau^{-1} \frac{1}{N} \sum_{i=1}^{N} d(x_i, y_i, \delta_i; \theta_0, \lambda^*) + o_p(N^{-1/2}),
\]
where \( \tau = \mathbb{E}\{\partial U(\theta_0; X, Y)/\partial \theta^T\} \),
\[
d(x_i, y_i, \delta_i; \theta_0, \lambda^*) = \sum_{k=0}^{L} b_k(x_i)\beta_k^* + \delta_i \omega^*(x_i; \lambda^*) \{ U(\theta_0; x_i, y_i) - \sum_{k=0}^{L} b_k(x_i)\beta_k^* \},
\]
and $\beta_k^*$ are the solution to
\[ \sum_{i=1}^{N} \delta_i \exp\{\lambda_0^* + \lambda_i^T b_i(x_i)\} \times \left\{ \frac{U(\theta_0; x_i, y_i) - \sum_{j=0}^{L} b_j(x_i)\beta_j}{\sum_{k=0}^{L} b_k(x_i)\beta_k} \right\} b_k(x_i) = 0 \]
for all $k = 0, 1, \cdots, L$.

Now, the variance estimation of $\theta_{SPS}$ can be constructed from (46). Specifically, let
\[ \tilde{d}_i = \sum_{k=0}^{L} b_k(x_i)\beta_k + \delta_i \left\{ 1 + \frac{N}{N_{1}} \cdot r^*(x_i; \lambda) \right\} \left\{ \frac{U(\hat{\theta}; x_i, y_i) - \sum_{k=0}^{L} b_k(x_i)\beta_k}{\sum_{k=0}^{L} b_k(x_i)\beta_k} \right\}, \]
where $\beta_k^*$ are the solution to
\[ \sum_{i=1}^{N} \delta_i \exp\{\hat{\lambda}_0 + \lambda_i^T b(x_i)\} \left\{ \frac{U(\hat{\theta}; x_i, y_i) - \sum_{j=0}^{L} b_j(x_i)\beta_j}{\sum_{k=0}^{L} b_k(x_i)\beta_k} \right\} b_k(x_i) = 0 \]
for all $k = 0, 1, \cdots, L$. The linearization variance estimator is then written as
\[ \tilde{V}(\theta_{SPS}) = \frac{1}{N} \tilde{\tau}^{-1} \tilde{\Sigma}_{dd} \tilde{\tau}^{-1'}, \]
where $\tilde{\tau} = N^{-1} \sum_{i=1}^{N} \delta_i \omega_i^* U(\hat{\theta}; x_i, y_i), \tilde{\Sigma}_{dd} = (N-1)^{-1} \sum_{i=1}^{N} (\tilde{d}_i - \tilde{d}_N)^{\otimes 2}$ and $\tilde{d}_N = N^{-1} \sum_{i=1}^{N} \tilde{d}_i$.

The empirical likelihood approach also provides a way to develop a likelihood ratio test for parameters in a completely analogous way to that for parametric likelihoods. Let $\ell_e(\theta, \lambda)$ be the profile empirical log-likelihood in (27) under constraints (23), (24) and (26). Further, define the profile empirical likelihood
\[ \ell_p(\theta) = \max_{\lambda} \ell_e(\theta, \lambda). \quad (47) \]
We have the following result.

**Theorem 2** Under regularity conditions in the Supplementary Material,
\[ 2\{\ell_p(\hat{\theta}) - \ell_p(\theta_0)\} \xrightarrow{p} \chi^2_p, \quad (48) \]
as $N \to \infty$, where $\hat{\theta}$ is obtained from the two-step procedure and $\theta_0$ is the true parameter.

For $\theta = (\theta_1, \theta_2)$, the empirical likelihood ratio test for $H_0 : \theta_1 = \theta_1^0$ can also be developed similarly to [Qin and Lawless (1994)].

16
5 Dimension Reduction

We now consider the case of $\theta = \mathbb{E}(Y)$. In Section 4, we have seen that either $\mathbb{E}(Y \mid x)$ lies in the linear space $\mathcal{H} = \text{span}\{b(x)\}$ or $b(x)$ is the basis for logistic regression (37) can give the root-$n$-consistency of the proposed estimator. However, as pointed out in Remark 3, including other $x$-variables outside the outcome model into $\mathcal{H}$ may lead to efficiency loss.

To explain the idea further, we assume that $b(x) = x_M$, where $M$ is an index set for a subset of $x$. The following lemma presents an interesting result.

**Lemma 2** If MAR condition in (7) holds and the reduced model for $y$ holds such that

$$f(y \mid x) = f(y \mid x_M) \quad (49)$$

for $x = (x_M, x_{M^c})$, then we can obtain the reduced MAR given $x_M$. That is,

$$Y \perp \delta \mid X_M. \quad (50)$$

Note that (50) is a special case of the reduced MAR in (8) using $b(x) = x_M$ as the balancing score function. In the spirit of Remark 3, we can see that the smoothed PS estimator using $b(x) = x_M$ is more efficient than the PS estimator using $b(x) = x$. Therefore, it is better to apply a model selection procedure to select the important variables which satisfies (49).

To find $x_M$ satisfying (49), we utilize two-stage estimation strategy to complete the smoothed propensity score function:

1. Step 1: Use a penalized regression method to select the basis function for the regression of $y$ on $x$.

2. Step 2: Use the basis function in Step 1 to obtain the calibration equation in (28) and construct $\hat{r}(x) = r^*(x; \hat{\lambda}) = \exp\{\hat{\lambda}_0 + b^T(x_i)\hat{\lambda}_1\}$.

In the first stage, we adapt the penalized estimating equations (Johnson et al., 2008) to select important covariates in the outcome model. Generally speaking, we utilize an $Z$-estimator in the outcome model and we denote the corresponding score function as $U(\alpha)$. For example, $U(\alpha)$ can be written as

$$U(\alpha) = \frac{1}{N_1} \sum_{i=1}^{N_1} \delta_i x_i (x_i^T \alpha - y_i), \quad (51)$$
where $\alpha \in \mathbb{R}^{d+1}$. The penalized estimating equations can be written as

$$U^P(\alpha) = U(\alpha) - q_\lambda(|\alpha|) \text{sgn}(\alpha), \quad (52)$$

where $q_\lambda(|\alpha|) = (q_\lambda(|\alpha_0|), \ldots, (q_\lambda(|\alpha_d|))^T$, $q_\lambda(\cdot)$ is a continuous function and $q_\lambda(|\alpha|) \text{sgn}(\alpha)$ is an elementwise product between two vectors. Further, let $p_\lambda(x) = \int q_\lambda(x) dx$. In $M$-estimation framework, $p_\lambda(x)$ usually serves as a penalization function. Various penalization functions can be used but we only consider the smoothly clipped absolute deviation function (SCAD) [Fan and Li, 2001]. In particular,

$$q_\lambda(\alpha) = \lambda \left\{ \mathbb{I}(|\alpha| < \lambda) + \frac{(a\lambda - |\alpha|)}{(a-1)\lambda} \mathbb{I}(|\alpha| \geq \lambda) \right\}, \quad (53)$$

where $(x)_+ = \max\{x, 0\}$, $\mathbb{I}$ is the indicator function and $a$ is constant specified as 3.7 in Fan and Li (2001). Further, we let $\hat{\mathcal{M}}$ to denote the variable index set selected after SCAD procedure. Here we use a working model to select the variables. According to Johnson et al. (2008), under some regularity conditions, we have

$$\mathbb{P}(\hat{\alpha}_j \neq 0) \to 1, \text{ for } j \in \mathcal{M};$$
$$\mathbb{P}(\hat{\alpha}_j = 0) \to 1, \text{ for } j \in \mathcal{M}^c, \quad (54)$$

which constructs the model selection consistency. After the first stage, we now obtain the important variable set $\hat{\mathcal{M}}$.

In the second stage, we use the selected variables to compute the smoothed density ratio function. That is, use

$$\log\{r^*(x, \mathcal{M})\} = \lambda_0 + \lambda_1^T x_\mathcal{M}.$$ 

The resulting PS estimator is then computed by

$$\hat{\theta}_{SPS} = N^{-1} \sum_{i=1}^{N} \delta_i \hat{\omega}^*(x_i) y_i$$

with $\hat{\omega}^*(x) = 1 + (N_0/N_1) \exp(x_M^T \hat{\lambda})$.

**Corollary 2** Suppose that the assumptions for Theorem 2 hold. Also, the additional assumptions listed in the Supplementary Material hold. If $x_\mathcal{M}$ satisfies $\mathbb{E}(Y \mid x) = \mathbb{E}(Y \mid x_\mathcal{M})$, then we obtain, with probability goes to 1,

$$\sqrt{N} \left( \hat{\theta}_{SPS} - \theta_0 \right) \xrightarrow{L} \mathcal{N}(0, V_r), \quad (55)$$

18
where

\[ V_r = \mathbb{V} \{ \mathbb{E}(Y \mid X_M) \} + \mathbb{E} \left[ \delta \{ 1 + (N_0/N_1) r^*(X_M) \}^2 \mathbb{V}(Y \mid X_M) \right]. \]

By Corollary 2, we can safely ignore the uncertainty due to the model selection in the first step. The asymptotic results in (55) is based on the model selection consistency in (54). That is, if we define \( D_n = \{ M = \hat{M} \} \) where \( \hat{M} \) is obtained from the first stage procedure, the linearization in (55) is conditional on \( D_n \). By (54), we obtain \( P(D_n) \rightarrow 1 \) and the limiting distribution of \( T_n \equiv \sqrt{n}(\hat{\theta}_{SPS} - \theta) \) given \( \hat{M} \), denoted by \( L(T_n \mid \hat{M}) \), is asymptotically equivalent to \( L(T_n \mid M) \). See also Theorem 1 of Yang et al. (2020) for a similar argument.

6 Multivariate missing data

We now consider the case of multivariate study variables, denoted by \( Y_1, \ldots, Y_p \), and they are subject to missingness. There are \( 2^p \) possible missing patterns with \( p \) study variables. Let \( T \leq 2^p \) be the realized number of different missing patterns in the sample. Thus, the sample is partitioned into \( T \) disjoint subsets with the same missing patterns. The parameter of interest is defined through \( \mathbb{E}(U(\theta; Y) \mid y_{i,obs(t)}) \) for \( t = 2, \ldots, T \). Condition (56) is the self-efficiency condition in (4).

To construct the density ratio function satisfying (56), we first find \( H_t = \text{span} \{ b_{1}^{(t)}(y_{obs(t)}), \ldots, b_{L(t)}^{(t)}(y_{obs(t)}) \} \) such that \( \mathbb{E}\{U(\theta; y_t) \mid y_{i,obs(t)} \} \in H_t \). Thus, using the I-projection idea in

\[ \sum_{t=1}^{T} \sum_{i \in S_t} \mathbb{E}\{U(\theta; y_i) \mid y_{i,obs(t)} \} \]

where \( y_{i,obs(t)} \) is the observed part of \( y_i \) for \( i \in S_t \). Instead of using a model for each conditional distribution, we can use the density ratio model such that

\[ N_1^{-1} \sum_{i \in S_t} r_t^*(y_{i,obs(t)}) U(\theta; y_t) = N_t^{-1} \sum_{i \in S_t} \mathbb{E}\{U(\theta; y_i) \mid y_{i,obs(t)} \} \] (56)

for \( t = 2, \ldots, T \). Condition (56) is the self-efficiency condition in (4).
Section 3, we may assume
\[
\log\{r_t^*(y_{obs(t)}; \phi(t))\} = \phi_0^{(t)} + \sum_{j=1}^{L(t)} \phi_j^{(t)} b_j^{(t)}(y_{obs(t)})
\] (57)
as the log-linear model for density ratio function. The model parameters in (57) can be estimated by calibration equation derived from (56) and model assumption (57):
\[
N_1^{-1} \sum_{i \in S_1} r_t^*(y_{i,obs(t)}; \hat{\phi}^{(t)}) (1, b_i^{(t)}) = N_1^{-1} \sum_{i \in S_1} (1, b_i^{(t)})
\]
with respect to \(\phi^{(t)}\) for \(t = 2, \ldots, T\), where \(b_i^{(t)}\) is a vector of \(b_j^{(t)}(y_{i,obs(t)})\) for \(j = 1, \ldots, L(t)\).

Now, the smoothed PS estimator of \(\theta\) can be obtained by solving
\[
\hat{\theta}_{SPS}(\theta) = N^{-1} \sum_{i \in S_1} \tilde{\omega}_i^* U(\theta; y_i) = 0,
\] (58)
where
\[
\tilde{\omega}_i^* = \sum_{t=1}^{T} \frac{N_t}{N_1} r_t^*(y_{i,obs(t)}; \hat{\phi}^{(t)})
\]
is the final weights for PS estimation.

To investigate the asymptotic behavior of the solution to (58), the density ratio for missing pattern \(t\) can be simplified as \(r_t^*(z_{i,t}; \lambda_t)\), for \(t = 2, \ldots, T\). Let the true parameter of interest be \(\theta_0\) and the true parameter for density ratio be \(\lambda_0 = (\lambda_2^T, \ldots, \lambda_T^T)^T\). Then we have the following theorem.

**Theorem 3** Under the regularity conditions stated in the Supplementary Material, for multivariate missing case, the solution \(\hat{\theta}_{SPS}\) to (58) satisfies
\[
\sqrt{N} \left( \hat{\theta}_{SPS} - \theta_0 \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} d(x_i, y_i, \delta_i; \lambda_0) + o_p(1),
\]
where
\[
d(x_i, y_i, \delta_i; \lambda_0) = - \left[ \mathbb{E} \left\{ \frac{\partial}{\partial \theta} U(\theta_0; x, y) \right\} \right]^{-1} \times \left[ \delta_{i,1} U_i + \sum_{t=2}^{T} \delta_{i,t} \tilde{\beta}_i z_{i,t} + \delta_{i,1} \sum_{t=2}^{T} \frac{N_t}{N_1} r_t^*(z_{i,t}; \lambda_t) \left\{ U_i - \tilde{\beta}_i z_{i,t} \right\} \right],
\] (59)
\[ U_i = U(\theta; x_i, y_i), \quad z_{i,t}^* = (1, b_{i,t}^{(k)}) \] and \( \tilde{\beta}_t \) is the probability limit to the solution of

\[ \sum_{i=1}^{N} \delta_{i,t} r_t^*(z_{i,t}; \lambda_t) \{ U(\theta_0; x_i, y_i) - \beta_t z_{i,t} \} z_{i,t}^T = 0. \]

The above theorem depicts the asymptotic behavior of our proposed estimators in multivariate missing case. Its proof is presented in the Supplementary Materials.

7 Simulation Study

7.1 Simulation Study One

Two limited simulation studies are performed to check the performance of the proposed method. In the first simulation study, we compare the proposed method with other existing methods under the MAR setup with few covariates. The setup for the first simulation study employed a \( 2 \times 2 \) factorial structure with two factors. The first factor is the outcome regression (OR) model that generates the sample. The second factor is the response mechanism (RM). We generate \( \delta \) and \( x = (x_1, x_2, x_3, x_4)^T \) based on the RM first. We have two different setup for the response mechanism as follows:

- **RM1 (Logistic model):**
  \[
  x_{ik} \sim N(2, 1), \quad \delta_i \sim \text{Ber}(p_i),
  \]
  \[
  \text{logit}(p_i) = 1 - x_{i1} + 0.5 x_{i2} + 0.5 x_{i3} - 0.25 x_{i4}.
  \]

- **RM2 (Gaussian mixture model):**
  \[
  \delta_i \sim \text{Bern}(0.6)
  \]
  \[
  x_{ik} \sim N(2, 1), \quad \text{for } k = 1, 2, 3,
  \]
  \[
  x_{i4} \sim \begin{cases} 
  N(3, 1), & \text{if } \delta_i = 1 \\
  N(1, 1), & \text{otherwise.}
  \end{cases}
  \]

Once \( x \) and \( \delta \) are generated, we generate \( y \) from two different outcome models, OR1 and OR2, respectively. That is, we generate \( y \) from
• OR1: \( y_i = 1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + e_i \).

• OR2: \( y_i = 1 + 0.5x_{i1}x_{i2} + 0.5x_{i3}^2x_{i4}^2 + e_i \).

Here, \( e_i \sim N(0, 1) \).

From each of the sample, we compare four different PS estimators of \( \theta \):

1. The proposed information projection (IP) PS estimator using the maximum entropy method in Section 3 using \( x_i = (1, x_{i1}, x_{i2}, x_{3i}, x_{i4})^T \) as the control variable for calibration. Thus, the proposed PS estimator can be written as \( \hat{\theta}_{\text{SPS}} = N^{-1} \sum_{i=1}^{N} \delta_i \hat{\omega}_i y_i \) where \( \hat{\omega}_i \) is of the form in (30) with

\[
\sum_{i=1}^{N} \delta_i \hat{\omega}_i (1, x_{i1}, x_{i2}, x_{3i}, x_{i4}) = \sum_{i=1}^{N} (1, x_{i1}, x_{i2}, x_{3i}, x_{i4}). \tag{60}
\]

2. The classical PS estimator using maximum likelihood estimation (MLE) of the response probability with Bernoulli distribution with parameter \( \logit(p_i) = x_i^T \lambda \).

3. Covariate balancing propensity score method (CBPS) of Imai and Ratkovic (2014). The CBPS estimator of \( \theta \) is obtained by applying empirical likelihood method that maximizes \( l(\omega) = \sum_{i=1}^{n} \delta_i \log(\omega_i) \) subject to (60).

4. Entropy balancing propensity score method (EBPS) of Hainmueller (2012) using calibration variable \( (1, x_1, x_2, x_3, x_4)^T \). That is, find the weights \( \omega_i \) that maximize \( Q(\omega) = \sum_{i=1}^{n} \delta_i \omega_i \log(\omega_i) \) subject to (60).

To check the performance of the four estimators, we use sample size \( N = 5,000 \) with 5,000 Monte Carlo samples. The results are presented in Figure 1 where IP is our proposed method.

When we use \((1, x_1, x_2, x_3, x_4)\) as the calibration variable, OR1 matches with the working outcome model and RM1 matches with the working response model. Among the four methods considered, our proposed method and EBPS method perform better than the other two methods. Entropy balancing propensity score method also shows good performances when OR1 or RM1 is true as it is doubly robust, but when both models fail (i.e. OR2RM2 setup), the performance is really poor. Our proposed method is also doubly robust and it performs reasonably well even when both models fail. Note that our proposed PS weights can be written as

\[
\hat{\omega}_i = 1 + N_0 \times \frac{\exp(x_i^T \hat{\lambda})}{\sum_{k=1}^{N} \delta_k \exp(x_k^T \hat{\lambda})}.
\]
Figure 1: Boxplots with four estimators for four models under simulation study one: (a) for OR1RM1, (b) OR1RM2, (c) for OR2RM1 and (d) for OR2RM2.
Thus, the PS estimator of $\theta = \mathbb{E}(Y)$ can be expressed as

$$
\hat{\theta}_{\text{SPS}} = \frac{1}{N} \sum_{i=1}^{N} \{ \delta_i y_i + (1 - \delta_i) \hat{y}_i \}
$$

where

$$
\hat{y}_i = \frac{\sum_{j=1}^{N} \delta_j \exp(x_j^T \hat{\lambda}) y_j}{\sum_{k=1}^{N} \delta_k \exp(x_k^T \lambda)}.
$$

Thus, the proposed PS estimator can be expressed as a special case of fractional hot deck imputation of Yang and Kim (2014), which is a robust estimation method. Other PS estimators does not allow for this interpretation.

### 7.2 Simulation Study Two

We performed another simulation study to understand the effect of the calibration variables for smoothed PS weighting. We first generate $x = (x_1, x_2, x_3)$ from multivariate normal distribution with mean $(1, 1, 1)$ and variance $\Sigma$ where

$$
\Sigma = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0.5 \\
0 & 0.5 & 1
\end{pmatrix}
$$

and generate $y$ from $y = 1 + 0.5x_1 - x_2 + e$ where $e \sim N(0, \sigma^2 = 1)$. We also generate $\delta_i \sim \text{Bernoulli}(\pi_i)$ where $\text{logit}(\pi_i) = -x_{1i} + \phi(x_{2i} - 1) + x_{3i}$, with $\phi = 0$ for Scenario 1 and $\phi = 1$ for Scenario 2, respectively. The parameter of interest is $\theta = \mathbb{E}(Y)$. The population size is $N = 1,000$. We considered three choices of the calibration variables in the proposed PS method.

1. The proposed PS estimator using $(1, x_1, x_2)$ as the calibration variable
2. The proposed PS estimator using $(1, x_1, x_3)$ as the calibration variable
3. The proposed PS estimator using $(1, x_1, x_2, x_3)$ as the calibration variable

Monte Carlo samples of size $B = 1,000$ are used to compute the Monte Carlo bias, Monte Carlo standard errors and the root mean squared errors of the estimators considered. The simulation result in Table 1 shows that the PS estimator using $(1, x_1, x_2)$ is the
most efficient among the three calibration estimators considered. That is, using calibration variables in the outcome model only achieves the best efficiency. In Scenario 2, the calibration estimator using the covariates for the outcome model is more efficient than the calibration estimator using the covariates for the selection model. The simulation result is consistent with our theory in Remark 3.

8 Conclusion

In handling the selection bias of the voluntary sample, the problem of estimating the inverse propensity score is recast as a problem of estimating the density ratio function. The density ratio function is defined for the balancing score function, and the propensity scores that arise can be efficient if the balancing score includes the actual mean function. The information projection technique builds a self-efficient propensity score estimating function. The variable selection technique for the outcome model can be utilized to generate efficient propensity score weights. The proposed method can be utilized as a unifying tool for merging information from multiple data sources.

There are several directions for further extensions of the proposed method. The proposed method can be extended to handle data integration problems (Chen et al., 2020) in survey sampling. Also, the proposed method is based on the assumption of missing at random. Extension to nonignorable nonresponse (Kim and Yu, 2011) can be also an interesting research direction. Instead of using Kullback-Leibler divergence, we may use Hellinger divergence (Antoine and Dovonon, 2021; Li et al., 2019) or $\gamma$-power divergence (Eguchi, 2021) to achieve some robustness. In addition, the proposed method can be used for causal inference, including the estimation of the average treatment effect from observational studies (Yang and Ding, 2020). Developing tools for causal inference using the
proposed method will be an important extension of this research.

**Acknowledgement**

The authors would like to thank professors Wayne A. Fuller and Zhiqiang Tan for their constructive comments.

**References**

Antoine, B. and P. Dovonon (2021). Robust estimation with exponentially tilted Hellinger distance. *Journal of Econometrics 224*, 330–344.

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

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika 95*, 539–553.

Breidt, F., G. Claeskens, and J. Opsomer (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika 92*(4), 831–846.

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

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

Chan, K. C. G., S. C. P. Yam, et al. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science 29*, 380–396.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society, Series B 78*, 673–700.

Chen, J. and Y. Liu (2013). Quantile and quantile-function estimations under density ratio model. *Annals of Statistics 41*, 1669–1692.
Chen, S. and D. Haziza (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika* **104**, 439–453.

Chen, Y., P. Li, and C. Wu (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* **115**, 2011–2021.

Csiszár, I. and P. C. Shields (2004). *Information theory and Statistics: A tutorial*. Now Publishers Inc.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–37.

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

Eguchi, S. (2021). Pythagoras theorem in information geometry and applications to generalized linear models. *Handbook of Statistics* **45**, 15–42.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Folsom, R. E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. In *Proceedings of the American Statistical Association, Social Statistics Section*, Volume 197201.

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

Graham, B. S., C. C. de Xavier Pinto, and D. Egel (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies* **79**, 1053–1079.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* **20**, 25–46.

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

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B* **76**, 243–263.
Johnson, B. A., D. Lin, and D. Zeng (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association 103*, 672–680.

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

Kim, J. K. and J. Shao (2021). *Statistical methods for handling incomplete data* (second ed.). CRC press.

Kim, J. K. and C. J. Skinner (2013). Weighting in survey analysis under informative sampling. *Biometrika 100*, 358–398.

Kim, J. K. and C. L. Yu (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association 106*, 157–165.

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

Kott, P. S. and T. Chang (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association 105*, 1265–1275.

Li, L., A. Vidyashankar, G. Diao, and E. Ahmed (2019). Robust inference after random projections via Hellinger distance for location-scale family. *Entropy 21*, 348.

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

Little, R. J. A. and S. Vartivarian (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology 31*, 161–168.

Ma, X. and J. Wang (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association 115*, 1851–1860.

Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association 100*, 1429–1442.

Nguyen, X., M. J. Wainwright, and M. I. Jordan (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory 56*, 5847–5861.
Park, S., J. K. Kim, and K. Kim (2019). A note on propensity score weighting method using paradata in survey sampling. *Survey Methodology 45*, 451–463.

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

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *Annals of Statistics 22*, 300–325.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

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

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika 97*, 661–682.

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

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Wang, C. Y. and M. S. Pepe (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society, Series B 62*, 509–524.

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*, 185–193.

Wu, C. and M. E. Thompson (2020). *Sampling Theory and Practice*. Springer.
Yang, S. and P. Ding (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association* 115, 1540–1554.

Yang, S., J. Kim, and R. Song (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society, Series B* 82, 445–465.

Yang, S. and J. K. Kim (2014). Fractional hot deck imputation for robust inference under item nonresponse in survey sampling. *Survey Methodology* 40, 211–230.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics* 47, 965–993.