An approximate Bayesian inference on propensity score estimation under unit nonresponse

Hejian Sang Jae Kwang Kim [*]

February 14, 2017

Abstract

Nonresponse weighting adjustment using the response propensity score is a popular tool for handling unit nonresponse. Statistical inference after the nonresponse weighting adjustment is complicated because the effect of estimating the propensity model parameter needs to be incorporated. In this paper, we propose an approximate Bayesian approach to handle unit nonresponse with parametric model assumptions on the response probability, but without model assumptions for the outcome variable. The proposed Bayesian method is calibrated to the frequentist inference in that the credible region obtained from the posterior distribution asymptotically matches to the frequentist confidence interval obtained from the Taylor linearization method. Unlike the frequentist approach, however, the proposed method does not involve Taylor linearization. The proposed method can be extended to handle over-identified cases in which there are more estimating equations than the parameters. Besides, the proposed method can also be modified to handle nonignorable nonresponse. Results from two simulation studies confirm the validity of the proposed methods, which are then applied to data from a Korean longitudinal survey.

Key words: Approximate Bayesian computation, Posterior distribution, Missing at random, Nonignorable nonresponse, Nonresponse weighting adjustment.

[*] Department of Statistics, Iowa State University, Ames, IA, 50010, U.S.A

1 Introduction

Missing data is frequently encountered in many areas of statistics. When the response mechanism is missing at random in the sense of Rubin (1976), one of the popular methods of handling missing data is to build a model for the response probability and use the inverse of the estimated response probability to construct weights for estimating parameters. Such weighting method is often called propensity score weighting and the resulting estimator is called propensity score estimator (Rosenbaum and Rubin, 1983). The propensity score method has been well established in the literature. For examples, see Rosenbaum (1987), Flanders and Greenland (1991), Robins et al. (1994), Robins et al. (1995), and Kim and Kim (2007). However, all the above researches were developed via the frequentist approaches. Variance estimates using a Taylor linearization method or bootstrap are used for making frequentist inference.

In this paper, we are interested in developing Bayesian inference for propensity score estimation. One of the main advantages of Bayesian inference is that all the uncertainty in the estimation process can be built into the Bayesian computation automatically. That is, there is no need to conduct variance estimation separately in the Bayesian inference. While the Bayesian method is widely used in many areas of statistics, the literature on the Bayesian approach of propensity score estimation is sparse. An (2010) proposed a Bayesian propensity score estimator jointly modeling the response mechanism and the outcome variable. However, specifying a correct outcome model is difficult under missing data and incorrect specification may lead to biased inference. McCandless et al. (2009) and Kaplan and Chen (2012) also assumed joint models and obtained Bayesian credible regions in the context of casual inference.

In this paper, our interest is in developing a new Bayesian approach without making any model assumptions on the outcome variable. Since no parametric model assumptions on the outcome variable are used, there is no explicit likelihood function corresponding to θ, the main parameter of interest. This makes it difficult to develop a Bayesian method for propensity score estimation. The challenge thus lies in properly incorporating the uncertainty in the propensity score estimation process into the Bayesian framework.

In this paper, we propose a novel approach featuring approximate Bayesian computation based on the summary statistics (Beaumont et al., 2002). The sampling distribution of summary statistics, which is the estimating equation itself, can be used to replace the likelihood part in deriving the posterior distribution. In the proposed Bayesian method, the credible region obtained from the posterior distribution with a flat prior asymptotically matches the frequentist confidence interval obtained from the Taylor linearization method. The computation for the proposed method is relatively simple and easy to understand.

To guarantee the consistency of estimators, the propensity score method requires the correct specification of the response model. To protect against model misspecification, Robins et al. (1994), Scharfstein et al. (1999), and Bang and Robins (2005) proposed the so-called doubly robust estimation, which requires either the propensity score model or the outcome regression model be correctly specified. To achieve efficiency and robustness, we can add into the proposed Bayesian method additional estimating equations obtained from the auxiliary variables observed throughout the full sample. When we incorporate more equations than the parameters, the proposed Bayesian method is modified to solve the over-identifying situation.

The rest of the paper is organized as follows. In Section 2, we introduce the basic setup of the general propensity score estimation problem. The proposed method is presented in Section 3. The main result and asymptotic theory are discussed in Section 4. In Section 5, we developed a related method by extending our proposed method to incorporate the auxiliary information observed throughout the sample. We also presented how to incorporate data augmentation algorithm to handle nonignorable nonresponse in Section 6. The finite sample performance of the proposed methods is examined in an extensive simulation study in Section 7. An application of the proposed methods to a longitudinal survey is presented in Section 8. Some concluding remarks are made in Section 9.

2 Basic Setup

Suppose that we are interested in estimating θ defined through $E\left\{U\left(\theta;\boldsymbol{X},Y\right)\right\}=0$ for some estimating function $U(\theta;\boldsymbol{X},Y)$. Let $\left(\boldsymbol{x}_i,y_i\right),i=1,\cdots,n$, be independently and identically distributed (IID) realizations of random variable $(\boldsymbol{X},Y)$. Under complete data, we can obtain a consistent estimator of θ by solving

$$\frac{1}{n}\sum_{i=1}^{n}U\left(\theta;\boldsymbol{x}_i,y_i\right)=0 \tag{1}$$

for θ. We assume that the solution to (1) is unique almost everywhere.

Now, suppose that $\boldsymbol{X}$ is always observed and Y is subject to missingness. In this case, we can define the response indicator function for unit i as

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

We assume that δ_i are independently generated from a Bernoulli distribution with

$$Pr(\delta_i=1\mid\boldsymbol{x}_i,y_i)=\pi\left(\phi;\boldsymbol{x}_i,y_i\right) \tag{2}$$

for some parameter vector ϕ and $\pi(\cdot)$ is a known function. In the logistic regression model, $\pi(x)=1/\{1+\exp(-x)\}$.

When missing data exist, we cannot apply (1) directly. Instead, using the parametric model for the response probability in (2), we can obtain the propensity score (PS) estimator of θ by the following two steps:

[Step 1] Compute the maximum likelihood (ML) estimator $\hat{\phi}$ of ϕ.

[Step 2] Compute the PS estimator of θ by solving

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\hat{\phi};\boldsymbol{x}_i,y_i)}U\left(\theta;\boldsymbol{x}_i,y_i\right)=0$$

for θ.

The computation for the ML estimator of ϕ can be greatly simplified if the response mechanism is Missing At Random (MAR) in the sense that

$$Pr\left(\delta=1|\boldsymbol{x},y\right)=Pr\left(\delta=1|\boldsymbol{x}\right).$$

In this case, the maximum likelihood estimator of ϕ can be obtained by finding the maximizer of

$$L(\phi) = \prod_{i=1}^{n} \{\pi(\phi; \mathbf{x}_i)\}^{\delta_i} \{1 - \pi(\phi; \mathbf{x}_i)\}^{1-\delta_i}. \tag{3}$$

If MAR does not hold, parameter estimation is more complicated. Assuming a parametric model for $f_1(y \mid \mathbf{x}) = f(y \mid \mathbf{x}, \delta = 1)$, the ML estimator can be obtained by maximizing

$$l_{obs}(\phi) = \sum_{i=1}^{n} \delta_i \log \pi(\phi; \mathbf{x}_i, y_i) + \sum_{i=1}^{n} (1 - \delta_i) \log \int \{1 - \pi(\phi; \mathbf{x}_i, y_i)\} \hat{f}_1(y \mid \mathbf{x}_i) dy,$$

where $\hat{f}_1(y \mid \mathbf{x}_i)$ is an estimator for $f_1(y \mid \mathbf{x}_i)$. Riddles et al. (2016) proposed an alternative computational tool that avoids computing the above integration using an EM algorithm.

We shall first present our proposed method under the MAR assumption. An extension to Not Missing At Random (NMAR) will be discussed in Section 6. Once the PS estimator $\hat{\theta}_{PS}$ of θ is obtained from the above two-step procedure, statistical inference for θ can be made based on the asymptotic normality

$$\sqrt{n}(\hat{\theta}_{PS} - \theta) \xrightarrow{\mathcal{L}} N(0, \sigma^2) \tag{4}$$

for some $\sigma^2 > 0$, where $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution. See Chapter 5 of Kim and Shao (2013) for a justification for (4).

Under the above setup, we shall introduce the proposed Bayesian approach to estimate the parameter and make inference from the posterior distribution. An advantage of the Bayesian approach is that we can incorporate the uncertainty in estimating ϕ into the Bayesian computation automatically.

3 Proposed Method

We now present the proposed Bayesian method in the case of MAR. Under the parametric model assumption (2), the likelihood function for ϕ is given in (3). From the likelihood function, we can derive the score function for ϕ as

$$U_1(\phi) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi(\phi; \boldsymbol{x}_i)} - \frac{1 - \delta_i}{1 - \pi(\phi; \boldsymbol{x}_i)} \right\} \frac{\partial \pi(\phi; \boldsymbol{x}_i)}{\partial \phi} =: \frac{1}{n} \sum_{i=1}^{n} s(\phi; \boldsymbol{x}_i, \delta_i). \tag{5}$$

If we define

$$U_2(\phi, \theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\phi; \boldsymbol{x}_i)} U(\theta; \boldsymbol{x}_i, y_i), \tag{6}$$

the PS estimator $\hat{\theta}_{PS}$ of θ can be viewed as the solution to the joint estimating equations: $U_1(\phi) = 0$ and $U_2(\phi, \theta) = 0$. Taylor linearization can be used to obtain a consistent variance estimator of $\hat{\theta}_{PS}$. See Chapter 5 of Kim and Shao (2013) for more details.

To introduce the proposed Bayesian inference corresponding to $\hat{\theta}_{PS}$, we first define $\boldsymbol{\zeta} = (\theta, \phi)$ and

$$U_n(\boldsymbol{\zeta}) = \begin{pmatrix} U_1(\phi) \\ U_2(\phi, \theta) \end{pmatrix}.$$

Instead of generating the posterior distribution from $p(\boldsymbol{\zeta} \mid \text{sample})$ directly , we use the posterior distribution $p(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}})$ to approximate the posterior distribution $p(\boldsymbol{\zeta} \mid \text{sample})$, where $\hat{\boldsymbol{\zeta}}$ solves $U_n(\boldsymbol{\zeta}) = 0$. Thus, we can consider

$$p(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}}) = \frac{g(\hat{\boldsymbol{\zeta}} \mid \boldsymbol{\zeta})\pi(\boldsymbol{\zeta})}{\int g(\hat{\boldsymbol{\zeta}} \mid \boldsymbol{\zeta})\pi(\boldsymbol{\zeta})d\boldsymbol{\zeta}} \tag{7}$$

as an approximate posterior distribution for $\boldsymbol{\zeta}$, where $g(\hat{\boldsymbol{\zeta}} \mid \boldsymbol{\zeta})$ is the sampling distribution of $\hat{\boldsymbol{\zeta}}$ and $\pi(\boldsymbol{\zeta})$ is the prior distribution for $\boldsymbol{\zeta}$. However, finding the sampling distribution $g(\hat{\boldsymbol{\zeta}} \mid \boldsymbol{\zeta})$ will involve Taylor linearization.

To consider an alternative computation, instead of generating from $p(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}})$ in (7), we use a posterior distribution from

$$p(\boldsymbol{\zeta} \mid U_n) = \frac{g\{U_n(\boldsymbol{\zeta}) \mid \boldsymbol{\zeta}\}\pi(\boldsymbol{\zeta})}{\int g\{U_n(\boldsymbol{\zeta}) \mid \boldsymbol{\zeta}\}\pi(\boldsymbol{\zeta})d\boldsymbol{\zeta}}, \tag{8}$$

where $g\{U_n(\boldsymbol{\zeta}) \mid \boldsymbol{\zeta}\}$ is the sampling distribution of $U_n(\boldsymbol{\zeta})$. To generate samples from (8), we first make a transformation of the parameters, defined as $\boldsymbol{\eta} = E(U_n \mid \boldsymbol{\zeta})$. Thus, $T : \boldsymbol{\zeta} \to \boldsymbol{\eta}$ is an one-to-one transformation of the parameter. We will generate $\boldsymbol{\eta}^*$ from $p(\boldsymbol{\eta} \mid U_n)$ first and then use $\boldsymbol{\zeta}^* = T^{-1}(\boldsymbol{\eta}^*)$ to obtain the posterior distribution values from (8).

Now, to compute $p(\boldsymbol{\eta} \mid U_n)$, first note that, under some regularity conditions,

$$\left[\sqrt{n}U_n|\boldsymbol{\zeta}\right] = \left[\sqrt{n}U_n|\boldsymbol{\eta}\right] \xrightarrow{\mathcal{L}} N(\boldsymbol{\eta}, \Sigma), \tag{9}$$

where notation $[\cdot]$ is used to denote the sampling distribution and $\xrightarrow{\mathcal{L}}$ denotes the convergence in distribution. Writing $\pi(\boldsymbol{\eta})$ as a prior distribution of $\boldsymbol{\eta}$, the posterior distribution of $\boldsymbol{\eta}$ given U_n can be expressed as

$$[\boldsymbol{\eta}|U_n] \propto [U_n|\boldsymbol{\eta}]\,\pi\left(\boldsymbol{\eta}\right).$$

If there is no information for the prior, we can use a flat prior for $\boldsymbol{\eta}$. The sampling distribution $[U_n|\boldsymbol{\eta}]$ serves the role of the likelihood function in the Bayesian inference. Using (9) and a flat prior for $\boldsymbol{\eta}$, we obtain

$$[\boldsymbol{\eta} \mid U_n] \sim N\left(\mathbf{0}, \Sigma/n\right) \tag{10}$$

as the posterior distribution, where a consistent estimator of Σ is

$$\hat{\Sigma} = \begin{pmatrix} n^{-1}\sum_{i=1}^{n} s(\hat{\phi};\boldsymbol{x}_i)^{\otimes 2} & n^{-1}\sum_{i=1}^{n} \delta_i \hat{\pi}_i^{-1} s(\hat{\phi};\boldsymbol{x}_i) U'(\hat{\theta};\boldsymbol{x}_i, y_i) \\ \text{symm.} & n^{-1}\sum_{i=1}^{n} \delta_i \hat{\pi}_i^{-2} U(\hat{\theta};\boldsymbol{x}_i, y_i)^{\otimes 2} \end{pmatrix},$$

where $\hat{\pi}_i = \pi(\hat{\phi};\mathbf{x}_i)$, $\hat{\phi}$ and $\hat{\theta}$ solve $U_n\left(\phi,\theta\right) = \mathbf{0}$, $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}\boldsymbol{A}'$ and $\boldsymbol{A}'$ represents the transpose of $\boldsymbol{A}$. The details of the derivation are presented in Appendix A. After we obtain the posterior distribution of $\boldsymbol{\eta}$, we can use the inverse transformation of T to obtain the posterior distribution of the original parameters. The following algorithm describes how to generate parameters from the posterior distribution of $\boldsymbol{\zeta} = (\phi, \theta)$:

[Step 1] Generate $\boldsymbol{\eta}^*$ from the posterior distribution

$$p(\boldsymbol{\eta} \mid U_n = 0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \hat{\Sigma}/n), \tag{11}$$

where $\hat{\Sigma}$ is a consistent estimator of $Var(\sqrt{n}U_n) = \Sigma$ in (9).

[Step 2] Solve $U_n\left(\boldsymbol{\zeta}\right) = \boldsymbol{\eta}^*$ for $\boldsymbol{\zeta}$ to obtain $\boldsymbol{\zeta}^*$.

Steps 1–2 can be repeated independently to generate independent samples from the posterior distribution. The samples can be used to obtain the posterior distribution of the induced parameters.

As we have illustrated before, the basic idea is that we use the posterior distribution of $p(\boldsymbol{\zeta}|\hat{\boldsymbol{\zeta}})$ to approximate the posterior distribution of $p\left(\boldsymbol{\zeta}|\text{sample}\right)$. This idea

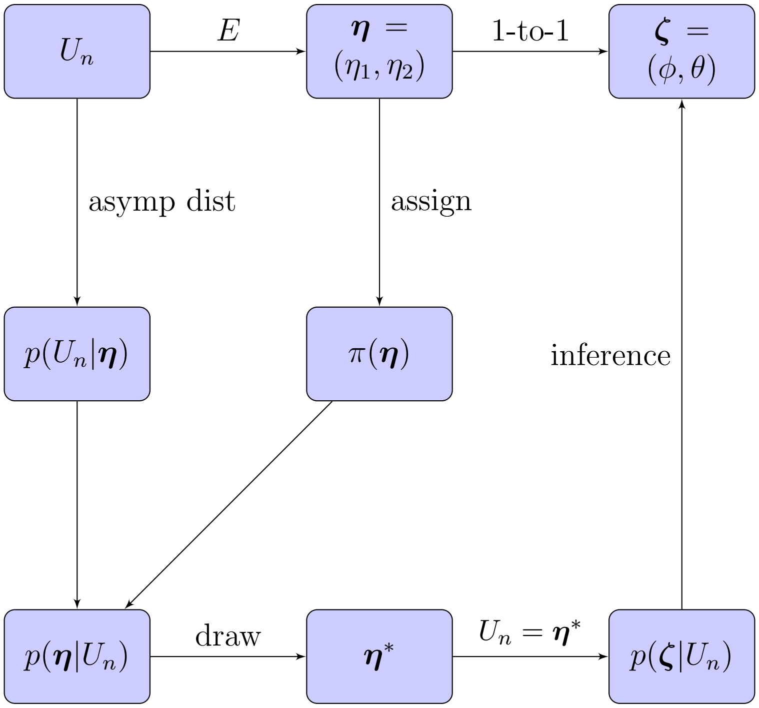
U_n
E
$\boldsymbol{\eta} = (\eta_1, \eta_2)$
1-to-1
$\boldsymbol{\zeta} = (\phi, \theta)$
asymp dist
assign
inference
$p(U_n|\boldsymbol{\eta})$
$\pi(\boldsymbol{\eta})$
$p(\boldsymbol{\eta}|U_n)$
draw
$\boldsymbol{\eta}^*$
$U_n = \boldsymbol{\eta}^*$
$p(\boldsymbol{\zeta}|U_n)$

Figure 1: Proposed Bayesian propensity score method

is similar in spirit to the Approximate Bayesian Computation of Soubeyrand and Haon-Lasportes (2015). Note that we do not use Taylor linearization to obtain the posterior distribution of θ. Instead, we use a transformation technique and generate the posterior distribution of $p(\boldsymbol{\eta}|U_n)$ first. After we obtain the posterior distribution of $\boldsymbol{\eta}$, we use the inverse transformation $T^{-1} : \boldsymbol{\eta} \to \boldsymbol{\zeta}$ to obtain the posterior distribution of the original parameters. The back-transformation plays the role of Taylor linearization in the frequentist approach. See Figure 1 for the illustration of the basic idea. Some asymptotic properties are established in the next section.

4 Asymptotic Properties

To establish the consistency of the parameter estimate and the interval estimate, we assume the following regularity conditions:

[C1] As $n \to \infty$, $U_n(\boldsymbol{\zeta}) \to \boldsymbol{\eta}(\boldsymbol{\zeta})$ in probability uniformly. That is $\sup_{\boldsymbol{\zeta} \in \boldsymbol{Z}} \| U_n(\boldsymbol{\zeta}) - \boldsymbol{\eta}(\boldsymbol{\zeta}) \| \xrightarrow{P} 0$, where $\boldsymbol{Z}$ is the parameter space.

[C2] The mapping $\boldsymbol{\zeta} \mapsto U_n(\boldsymbol{\zeta})$ is continuous and has exactly one zero $\hat{\boldsymbol{\zeta}}$ with probability one as $n \to \infty$.

[C3] Equation $\boldsymbol{\eta}(\boldsymbol{\zeta}) = 0$ has exactly one root at $\boldsymbol{\zeta} = \boldsymbol{\zeta}_0$.

[C4] There exits a neighbor of $\boldsymbol{\zeta}_0$, denoted by $N_n(\boldsymbol{\zeta}_0)$, on which with probability one all $U_n(\boldsymbol{\zeta})$ are continuously differentiable and the Jacobian $\partial U_n(\boldsymbol{\zeta}) / \partial \boldsymbol{\zeta}$ converge uniformly to a non-stochastic limit which is non-singular. Here, $N_n(\boldsymbol{\zeta}_0)$ is a ball with center $\boldsymbol{\zeta}_0$ and radius r_n, where r_n satisfies $r_n \to 0$ and $r_n\sqrt{n} \to \infty$. Also, we assume that $\partial^2 U_{n,j}(\boldsymbol{\zeta}) / \partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}'$ is finite for each entry for $j = 1, 2, \cdots, p$ and with probability one as $n \to \infty$.

[C5] For any $\boldsymbol{\zeta} \in N_n(\boldsymbol{\zeta}_0)$,

$$\sqrt{n}\left(U_n(\boldsymbol{\zeta}) - \boldsymbol{\eta}(\boldsymbol{\zeta})\right) \xrightarrow{\mathcal{L}} N\left(0, \Sigma(\boldsymbol{\zeta})\right) \tag{12}$$

holds for some $\Sigma(\boldsymbol{\zeta}) = Var\left\{\sqrt{n}U_n(\boldsymbol{\zeta}) | \boldsymbol{\zeta}\right\} > 0$ that is independent of n.

As long as the samples satisfy some moment conditions, condition [C1] holds. Condition [C2] and [C3] are used to make sure that the solutions of estimating equation U_n and estimating function $\boldsymbol{\eta}$ exist and are unique to avoid the model non-identifiability problem. The condition [C4] regulates the derivatives of the estimating equation to make sure that the variance converges. Condition [C5] provides the asymptotic distribution for the estimating equation. Under the above conditions, we can obtain

$$\sqrt{n}\left(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\right) \xrightarrow{\mathcal{L}} N\left(0, A(\boldsymbol{\zeta}_0)^{-1}\Sigma(\boldsymbol{\zeta}_0) A'(\boldsymbol{\zeta}_0)^{-1}\right) \tag{13}$$

where $A(\boldsymbol{\zeta}) = \partial\boldsymbol{\eta}(\boldsymbol{\zeta}) / \partial\boldsymbol{\zeta}$.

We now make additional assumptions to establish the posterior consistency and convergence in distribution:

[C6] The prior distribution $\boldsymbol{\eta} \mapsto \pi(\boldsymbol{\eta})$ is positive and Lipschitz continuous over the parameter space.

[C7] For any $\boldsymbol{\zeta} \in N_n(\boldsymbol{\zeta}_0)$, the variance estimator $\hat{\Sigma}(\boldsymbol{\zeta})$ in (11) satisfies $\hat{\Sigma}(\boldsymbol{\zeta}) = \Sigma(\boldsymbol{\zeta})\{1 + o_p(1)\}$.

[C8] For any $\boldsymbol{\zeta} \in N_n(\boldsymbol{\zeta}_0)$, the mapping $\boldsymbol{\zeta} \mapsto |\Sigma(\boldsymbol{\zeta})|^{-1}$ is Lipschitz continuous. Also, the mapping $\boldsymbol{\zeta} \mapsto x' \{\Sigma(\boldsymbol{\zeta})\}^{-1} x$ is Lipschitz continuous in the sense that there exists a constant $C(x)$ satisfying $\left\|x' \{\Sigma(\boldsymbol{\zeta}_1)\}^{-1} x - x' \{\Sigma(\boldsymbol{\zeta}_2)\}^{-1} x\right\| \leq C(x) \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|$, for any $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2 \in N_n(\boldsymbol{\zeta}_0)$, for all $x \in \mathbb{R}^p$, where $p = \dim(\boldsymbol{Z})$. And $C(x)$ is also Lipschitz continuous.

[C9] $\boldsymbol{\zeta} \mapsto U_n(\boldsymbol{\zeta})$ and $\boldsymbol{\zeta} \mapsto \boldsymbol{\eta}(\boldsymbol{\zeta})$ are one to one functions for any $\boldsymbol{\zeta} \in N_n(\boldsymbol{\zeta}_0)$. Also $\boldsymbol{\zeta} \mapsto \boldsymbol{\eta}(\boldsymbol{\zeta})$ is Lipschitz continuous.

Condition [C6] is a common assumption for the prior and the flat prior satisfies this condition. Condition [C7] requires the variance estimator to be consistent. Conditions [C8] to [C9] are the sufficient conditions for the posterior distribution to be approximated by the proposed method. All the conditions can be easily satisfied if we assume variance estimate is continuous and has bounded eigenvalues.

Theorem 4.1 *Let $\hat{\boldsymbol{\zeta}}$ be the solution to $U_n(\boldsymbol{\zeta}) = 0$. Under (C1)–(C9), the posterior distribution $p(\boldsymbol{\zeta} \mid U_n = 0) = p(\boldsymbol{\zeta}|\hat{\boldsymbol{\zeta}})$, generated by the two-step method in Section 3, satisfies*

$$p(\boldsymbol{\zeta}|\hat{\boldsymbol{\zeta}}) \rightarrow \phi_{\hat{\boldsymbol{\zeta}}, Var(\hat{\boldsymbol{\zeta}})}(\boldsymbol{\zeta}) \tag{14}$$

$$p \lim_{n \longrightarrow \infty} \int_{N_n(\boldsymbol{\zeta}_0)} \phi_{\hat{\boldsymbol{\zeta}}, Var(\hat{\boldsymbol{\zeta}})}(\boldsymbol{\zeta})\, d\boldsymbol{\zeta} = 1, \tag{15}$$

where $\phi_{\hat{\boldsymbol{\zeta}}, Var(\hat{\boldsymbol{\zeta}})}(\cdot)$ is the density of the normal distribution with mean $\hat{\boldsymbol{\zeta}}$ and variance $Var(\hat{\boldsymbol{\zeta}})$.

The proof is shown in Appendix B. Result (14) is a convergence of the posterior distribution to normality and result (15) is the posterior consistency. By (14), the confidence region using the proposed Bayesian method is asymptotically equivalent to the frequentist confidence region based on asymptotic normality of $\hat{\boldsymbol{\zeta}}$. Thus, our proposed Bayesian method is calibrated to frequentist inference.

To construct a level α confidence region, let $k^*(\alpha)$ be the largest value of k such that

$$Pr\{\boldsymbol{\zeta} : p(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}}) \geq k\} = 1 - \alpha.$$

The level-α Bayesian High Posterior Density (HPD) confidence region (Chen and Shao, 1999) using k^* is

$$C^*(\alpha) = \left\{ \boldsymbol{\zeta} : p(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}}) \geq k^*(\alpha) \right\}.$$

We can show that $\int_{\hat{C}^*(\alpha)} p(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}}) d\boldsymbol{\zeta} \to 1 - \alpha$ in probability, where $\hat{C}^*(\alpha)$ is the confidence region from Monte Carlo samples, which are generated from the approximate target posterior distribution. See Hyndman (1996).

5 Optimal Estimation

We now extend the proposed method to incorporate additional information from the full sample. Note that the PS estimator applied to $\boldsymbol{\mu}_x = E(X)$ can be computed as the solution to

$$\sum_{i=1}^{n} \frac{\delta_i}{\pi(\hat{\phi}; \boldsymbol{x}_i)} (\boldsymbol{x}_i - \boldsymbol{\mu}_x) = 0$$

which is not necessarily equal to $\hat{\mu}_{x,n} = n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i$. Including this extra information in the propensity score estimation, if done properly, will improve the efficiency of the resulting PS estimator. In the frequentist propensity score method, incorporating such extra information can be implemented by Generalized Method of Moments and it is sometimes called optimal PS estimation. See Cao et al. (2009), Zhou and Kim (2012) and Imai and Ratkovic (2014).

To include such extra information, we may add

$$\begin{aligned} U_3(\phi, \boldsymbol{\mu}_x) &= \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\phi; \boldsymbol{x}_i)} (\boldsymbol{x}_i - \boldsymbol{\mu}_x) \\ U_4(\boldsymbol{\mu}_x) &= \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}_x) \end{aligned}$$

in addition to the original estimating equations based on $U_1(\phi)$ and $U_2(\phi, \theta)$ in (5) and (6), respectively. Note that we cannot directly apply the proposed two-step method in Section 3 in this case because there are more estimating equations than the parameters and the transformation

$$(\boldsymbol{\mu}_x, \phi, \theta) \to (\boldsymbol{\eta}_1, \eta_2, \boldsymbol{\eta}_3, \boldsymbol{\eta}_4)$$

is not one-to-one, where

$$\begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \\ \boldsymbol{\eta}_4 \end{pmatrix} = E\left\{ \left. \begin{pmatrix} U_1(\phi) \\ U_2(\theta,\phi) \\ U_3\left(\phi, \boldsymbol{\mu}_x\right) \\ U_4\left(\boldsymbol{\mu}_x\right) \end{pmatrix} \right| \boldsymbol{\mu}_x, \phi, \theta \right\}.$$

To solve this problem, instead of using the two-step method involving generation of $\boldsymbol{\eta}^*$ first from (11), we consider a direct sampling method that generates $\boldsymbol{\psi}^* = (\boldsymbol{\mu}_x^*, \phi^*, \theta^*)$ from the posterior distribution of $\boldsymbol{\psi} = (\boldsymbol{\mu}_x, \phi, \theta)$ given the observed data directly. To formally describe the procedure, first define

$$U_n(\boldsymbol{\psi}) = (U_1'(\phi), U_2(\phi,\theta), U_3'(\phi,\boldsymbol{\mu}_x), U_4'(\boldsymbol{\mu}_x))'.$$

Under some regularity conditions, we can obtain

$$[U_n|\boldsymbol{\psi}] \sim N(\mathbf{0}, \Sigma(\boldsymbol{\psi})/n) \tag{16}$$

for sufficiently large n, where $\Sigma(\boldsymbol{\psi}) = Var\left\{\sqrt{n}U_n(\boldsymbol{\psi}) \mid \boldsymbol{\psi}\right\}$. Using (16) as the sampling distribution $g(U_n|\boldsymbol{\psi})$ of U_n and using a prior $\pi(\boldsymbol{\psi})$ for $\boldsymbol{\psi}$, the posterior distribution of $\boldsymbol{\psi}$ can be written as

$$p\left(\boldsymbol{\psi}|U_n\right) = \frac{g(U_n|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{\int g(U_n|\boldsymbol{\psi})\pi(\boldsymbol{\psi})d\boldsymbol{\psi}}. \tag{17}$$

Note that we can still use the approximate normality of U_n to play the role of the likelihood function in the approximate Bayesian analysis. Note that even if the prior distribution is normal, the posterior distribution in (17) is no longer normal.

To obtain the posterior draws from (17), we can use a Monte Carlo method based on a version of Metropolis-Hastings algorithm (e.g. Chib and Greenberg (1995)). The computation details of the Monte Carlo method for generating samples from (17) are presented in Appendix C.

Note that, in generating samples from (17), the number of estimating equations is allowed to be greater than the number of parameters. Therefore, the proposed method is quite flexible in the sense that it can be applied to over-identified situations. Since the point estimator is asymptotically equivalent to the optimal PS estimator, the proposed method can thus be called optimal Bayesian PS (OBPS) method.

6 Nonignorable nonresponse

We now consider an application of the proposed Bayesian method to nonignorable nonresponse. Under the setup of Section 2, we first assume a parametric model for the response mechanism

$$Pr(\delta_i = 1|\boldsymbol{x}_i, y_i) = \pi(\phi; \boldsymbol{x}_{i1}, y_i), \tag{18}$$

where $\pi(\cdot)$ is known up to ϕ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$. The auxiliary variable $\boldsymbol{x}_{i2}$ is often called the response instrumental variable to avoid the non-identifiable problem in Wang et al. (2014). In addition, we assume a parametric model for the respondents' outcome model

$$f(y_i \mid \mathbf{x}_i, \delta_i = 1) = f_1(y_i \mid \mathbf{x}_i; \gamma) \tag{19}$$

for some γ. Using (18) and (19), we can obtain the following prediction model for the nonrespondents:

$$f\left(y|\boldsymbol{x}, \delta = 0; \gamma, \phi\right) = f\left(y|\boldsymbol{x}, \delta = 1; \gamma\right) \frac{O\left(\boldsymbol{x}_1, y; \phi\right)}{E\left\{O\left(\boldsymbol{x}_1, y; \phi\right)|\boldsymbol{x}, \delta = 1\right\}}, \tag{20}$$

where $O\left(\boldsymbol{x}_1, y; \phi\right) = Pr\left(\delta = 0|\boldsymbol{x}_1, y\right)/Pr\left(\delta = 1|\boldsymbol{x}_1, y\right)$, $f\left(y|\boldsymbol{x}, \delta = 1\right)$. If $\pi(\phi; \mathbf{x}_{i1}, y_i)$ follows a logistic regression model such as $\pi(\phi; \boldsymbol{x}_{i1}, y_i) = \{1 + \exp(x_{i1}\phi_1 + y_i\phi_2)\}^{-1}$ then $O\left(\boldsymbol{x}_1, y; \phi\right) = \exp(-\phi_2 y)$. See Kim and Yu (2011) for more discussion of the prediction model (20).

If y_i were available throughout the sample, we could use

$$\begin{aligned} S_1\left(\gamma\right) &:= \frac{1}{n}\sum_{i=1}^{n} \delta_i s_1\left(\gamma; \boldsymbol{x}_i, y_i\right) \\ S_2\left(\phi\right) &:= \frac{1}{n}\sum_{i=1}^{n} s_2\left(\phi; \delta_i, \boldsymbol{x}_{1i}, y_i\right), \\ U(\theta) &= \frac{1}{n}\sum_{i=1}^{n} U(\theta; \boldsymbol{x}_i, y_i), \end{aligned}$$

as the estimating functions for $\boldsymbol{\zeta} = (\gamma, \phi, \theta)$, where $s_1(\gamma)$ is the score function of γ with $s_1(\gamma; x_i, y_i) = \partial \log f\left(y_i|\boldsymbol{x}_i, \delta_i = 1; \gamma\right)/\partial\gamma$ and $S_2\left(\phi\right)$ is the score function of ϕ. Writing the joint estimating equations as $U_n(\boldsymbol{\zeta}) = (S_1'(\gamma), S_2'(\phi), U(\theta))'$ and $\boldsymbol{\eta} = E\left\{U_n(\boldsymbol{\zeta}) \mid \boldsymbol{\zeta}\right\}$, the following two-step method can be used to generate the posterior samples of $\boldsymbol{\zeta}$.

[Step 1] Generate $\boldsymbol{\eta}^*$ from the approximate posterior distribution using $p(\boldsymbol{\eta} \mid U_n(\boldsymbol{\zeta}))$. Under a flat prior for $\boldsymbol{\eta}$, the posterior distribution of $\boldsymbol{\eta}$ can be obtained as a multivariate normal distribution with mean $\mathbf{0}$ and variance Σ/n. A consistent estimator of Σ is

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \delta_i s_1(\hat{\gamma}; \boldsymbol{x}_i, y_i) \\ s_2(\hat{\phi}; \delta_i, \boldsymbol{x}_{1i}, y_i) \\ U(\hat{\theta}; \boldsymbol{x}_i, y_i) \end{pmatrix}^{\otimes 2},$$

where $\hat{\boldsymbol{\zeta}} = (\hat{\gamma}, \hat{\phi}, \hat{\theta})$ is the solution to $U_n(\boldsymbol{\zeta}) = \mathbf{0}$ under complete response.

[Step 2] The posterior values of $\boldsymbol{\zeta}$ can be obtained by solving $U_n(\boldsymbol{\zeta}) = \boldsymbol{\eta}^*$ for $\boldsymbol{\zeta}$.

Now, to implement the proposed Bayesian method under missing data, we can use Data Augmentation (DA) method of Tanner and Wong (1987). The DA algorithm consists of I-step and P-step. In I-step, the imputed values of y_i are generated from the prediction model using the current parameter values. In P-step, the posterior values of the parameters are generated from the above two-step method using the current imputed data. To formally describe the proposed method, define $X_n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, $\boldsymbol{\delta}_n = \{\delta_1, \cdots, \delta_n\}$ and $Y_n = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{obs} and Y_{mis} are the observed and missing part of $Y_n = (y_1, \cdots, y_n)$, respectively. The proposed DA algorithm can be described as follows:

I-step: Given current parameter values $\boldsymbol{\zeta}^*$, generate imputed values Y^*_{mis} from the prediction model (20) evaluated at the current parameter values.

P-step: Using the current imputed data, apply the above two-step method of generating the parameter values $\boldsymbol{\zeta}^*$ from $p(\boldsymbol{\zeta} \mid U^*_n(\boldsymbol{\zeta}))$, where $U^*_n(\boldsymbol{\zeta}) = U_n(\boldsymbol{\zeta}; Y_{\text{obs}}, Y^*_{\text{mis}})$.

The two steps are iteratively computed until some convergence criterion is satisfied. Once the posterior values of $\boldsymbol{\zeta}^*$ are obtained, the posterior values of θ^* can be used to perform Bayesian inference for θ. To explain the proposed method further, denote $p_U(\boldsymbol{\zeta} \mid X_n, Y_n, \boldsymbol{\delta}_n) = p(\boldsymbol{\zeta} \mid U_n)$ to emphasize that $U_n(\boldsymbol{\zeta})$ is a function of Y_n. The **I-step** of the proposed method is to generate Y_{mis} from the posterior predictive distribution of Y_{mis} by

$$f(Y_{\text{mis}}|X_n, Y_{\text{obs}}, \boldsymbol{\delta}_n) = \int f(Y_{\text{mis}}|X_n, \boldsymbol{\zeta}) p_U(\boldsymbol{\zeta} \mid X_n, Y_{obs}, \boldsymbol{\delta}_n) d\boldsymbol{\zeta},$$

where

$$p_U(\boldsymbol{\zeta}|X_n, Y_{\text{obs}}, \boldsymbol{\delta}_n) = \int p_U(\boldsymbol{\zeta}|X_n, Y_n, \boldsymbol{\delta}_n) f(Y_{\text{mis}}|X_n, Y_{\text{obs}}, \boldsymbol{\delta}_n) dY_{\text{mis}}$$

is generated from **P-step**. After convergence, the DA algorithm generates ζ from the posterior density

$$p_U(\boldsymbol{\zeta}|X_n, Y_{\text{obs}}, \boldsymbol{\delta}_n) = \frac{\int g(U_n \mid \boldsymbol{\zeta})\pi(\boldsymbol{\zeta}) dY_{\text{mis}}}{\int\int g(U_n \mid \boldsymbol{\zeta})\pi(\boldsymbol{\zeta}) dY_{\text{mis}} d\boldsymbol{\zeta}}.$$

7 Simulation Study

We perform two limited simulation studies to validate our theory and to check the robustness of our proposed methods. In the first simulation, the proposed method is evaluated under ignorable response mechanism. In the second simulation, the proposed method is applied to some nonignorable nonresponse mechasnism.

7.1 Simulation Study One

The first simulation study can be described as a 3 × 4 factorial design, where the factors are outcome regression model for $E(y \mid \mathbf{x})$ and the response mechanism.

For the outcome regression models, we use $y = m(x_1, x_2) + e$ with three different mean functions given by

$$\begin{array}{ll} \text{Function 1:} & m_1(\mathbf{x}) = 2x_1 + 3x_2 - 20 \\ \text{Function 2:} & m_2(\mathbf{x}) = 0.5(x_1 - 2)^2 + x_2 - 2 \\ \text{Function 3:} & m_3(\mathbf{x}) = 0.1\exp(0.1x_1 - 0.2) + 3x_2 + c_3 \end{array},$$

where c_3 is chosen to give the same values for $E(y)$ in different mean functions. The explanatory variables $(x_1, x_2)^T$ are generated from $N(\boldsymbol{\mu}, \Sigma_x)$, with $\boldsymbol{\mu}_x = (2, 8)^T$ and $\Sigma = \text{diag}\{4, 8\}$. The error distribution is $e \sim N(0, \sqrt{|x_1| + 1})$.

For the response mechanism, we use four different response mechanisms. In the first response mechanism (R1), the response indicator function δ_i are independently generated from a Bernoulli distribution with probability

$$p_i(\phi_0, \phi_1) = \frac{\exp(\phi_0 + \phi_1 x_{i1})}{1 + \exp(\phi_0 + \phi_1 x_{i1})} \tag{21}$$

with $(\phi_0, \phi_1) = (0.1, 0.4)$, which makes the overall response rate approximately equal to 70%. In the second response mechanism (R2), we use the sample logistic regression model with $(\phi_0, \phi_1) = (-1.2, 0.15)$, which leads to about 30% response rate. In the third response mechanism (R3), the response indicator function δ_i are independently generated from a Bernoulli distribution with probability

$$p_i(\phi_0, \phi_1) = \Phi(\phi_0 + \phi_1 x_{i1}) \tag{22}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $(\phi_0, \phi_1) = (0, 0.28)$, which leads to about 70% response rate. In the fourth response mechanism (R4), we use the same probit model with $(\phi_0, \phi_1) = (-0.7, 0.1)$ to make the response rate near to 30%.

For each of the $12 = 3 \times 4$ simulation setup, we generate random samples of size $n = 500$ independently $B = 2,000$ times. From each realized sample, we specify a logistic regression model

$$Pr(\delta_i = 1|\mathbf{x}_i, y_i) = \frac{\exp(\phi_0 + \phi_1 x_{i1} + \phi_2 x_{i2})}{1 + \exp(\phi_0 + \phi_1 x_{i1} + \phi_2 x_{i2})} =: \pi(\phi; \mathbf{x}_i)$$

as the response model. Thus, in R3 and R4, the response model is incorrectly specified.

For each Monte Carlo sample, we use the following four methods of inference for $\theta = E(y)$:

1. PS: Frequentist approach based on Taylor linearization. The point estimator $(\hat{\theta}_{PS}, \hat{\phi})$ is computed from

$$U_{PS}(\theta, \phi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi(\phi; \mathbf{x}_i)}(y_i - \theta) = 0$$
$$S(\phi) = \frac{1}{n}\sum_{i=1}^{n} \{\delta_i - \pi(\phi; \mathbf{x}_i)\}(1, \mathbf{x}_i')' = \mathbf{0}.$$

The confidence intervals are constructed by $\hat{\theta}_{PS} \pm 1.96\sqrt{\hat{V}_{PS}}$, where $\hat{V}_{PS}$ is obtained by the Taylor linearization method.

2. Bayesian PS (BPS): Apply the proposed Bayesian method based on the joint estimating functions

$$U_1(\phi) = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i - \pi(\phi;\mathbf{x}_i)\right\}(1,\mathbf{x}_i')' \tag{23}$$

$$U_2(\phi,\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\phi;\mathbf{x}_i)}(y_i - \theta) \tag{24}$$

   The estimators for ϕ, θ are obtained by the median of the draws from the approximate posterior distribution. The confidence interval can be constructed by HPD region introduced in Section 4.

3. Optimal PS (OPS): Generalized method of moments using

$$U_3(\phi,\mu_x) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\phi;\mathbf{x}_i)}(\mathbf{x}_i - \boldsymbol{\mu}_x)$$

$$U_4(\mu_x) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_x)$$

   in addition to (23) and (24). If we denote $U_n(\mu_x,\phi,\theta) = (U_1', U_2, U_3', U_4')'$, then the OPS estimator is obtained by minimizing $U_n^T W^{-1} U_n$, where $W = Var(U_n)$. See Section 5.4 of Kim and Shao (2013).

4. OBPS: Optimal Bayesian PS method discussed in Section 5 using the same estimating functions $U_1(\phi)$, $U_2(\phi,\theta)$, $U_3(\phi,\mu_x)$, and $U_4(\mu_x)$. The point estimators for $\boldsymbol{\mu}_x, \phi, \theta$ are obtained by the median of the draws from the approximate posterior distribution. The confidence intervals can be constructed by the HPD region, introduced in Section 4.

For each of the four methods, 95% confidence intervals for θ are computed from Monte Carlo samples.

Table 1 presents the simulation results, coverage probabilities and average lengths of confidence intervals (CI), for the four methods. Overall, all the coverage probabilities are approximately 95%, which validates our proposed methods BPS and OBPS. For R1 and R2, we have a correctly specified model for the response mechanism. For R1, which has high response rate 70%, both BPS and OBPS methods provide

Table 1: Simulation results: "m" denotes mean function, "c_p" is the coverage probability for the corresponding confidence interval, "CI length" is the average length of the confidence intervals.

Response mechanism	m	method	c_p	CI length	Response mechanism	m	method	c_p	CI length
R1	m_1	PS	0.95	1.83	R3	m_1	PS	0.95	1.86
		BPS	0.95	1.84			BPS	0.95	1.87
		OPS	0.95	1.78			OPS	0.95	1.78
		OBPS	0.95	1.78			OBPS	0.94	1.78
	m_2	PS	0.94	0.88		m_2	PS	0.94	0.89
		BPS	0.94	0.88			BPS	0.94	0.89
		OPS	0.94	0.79			OPS	0.93	0.79
		OBPS	0.94	0.80			OBPS	0.94	0.80
	m_3	PS	0.95	1.56		m_3	PS	0.94	1.58
		BPS	0.94	1.56			BPS	0.94	1.58
		OPS	0.95	1.53			OPS	0.95	1.53
		OBPS	0.94	1.52			OBPS	0.94	1.52
R2	m_1	PS	0.95	1.96	R4	m_1	PS	0.95	1.95
		BPS	0.96	2.00			BPS	0.95	1.99
		OPS	0.95	1.83			OPS	0.94	1.82
		OBPS	0.95	1.83			OBPS	0.95	1.82
	m_2	PS	0.95	1.16		m_2	PS	0.94	1.13
		BPS	0.95	1.16			BPS	0.95	1.13
		OPS	0.94	0.91			OPS	0.93	0.90
		OBPS	0.95	0.97			OBPS	0.95	0.95
	m_3	PS	0.95	1.68		m_3	PS	0.95	1.67
		BPS	0.95	1.72			BPS	0.95	1.70
		OPS	0.95	1.59			OPS	0.95	1.58
		OBPS	0.95	1.58			OBPS	0.95	1.57

valid confidence intervals with correct coverage rates. Comparing the average length of confidence intervals, we can see that PS and BPS methods have approximately equal average CI lengths and OPS and OBPS have approximately equal average CI lengths, which confirms the asymptotic equivalence of the two methods. That is, our proposed Bayesian methods are calibrated to the frequentist inference. The same conclusion can be obtained for R2, which has much lower response rates. For different regression mean functions, we find that both OPS and OBPS methods achieve more efficiency gains when the regression model is not linear and the response rate is low. For the probit response mechanism (R3 and R4), BPS and OBPS still provide valid confidence intervals with correct coverages. Thus, the proposed method seems to be robust against model misspecification of the response model.

7.2 Simulation Study Two

In the second simulation study, we consider an extension of the proposed method to nonignorable nonresponse. In the simulation, we generate the covariate variable $x \sim N(0, 0.5)$ and use the outcome regression model $y = m(x)+e$ to generate y, where $e \sim N(0, 1)$. We consider three different mean functions $m(x)$, which are specified as $m_1(x) = -1 + 2x$, $m_2(x) = -1.25 + 2x + 0.5x^2$ and $m_3(x) = -1 + 8\sin(x)$.

We use two different mechanisms to generate the response indicators. The response indicator function δ_i are independently generated from Bernoulli distribution with the probability for $\delta_i = 1$ equal to

$$p_i(\phi_0, \phi_1) = \begin{cases} \{1 + \exp(-\phi_{10} - \phi_{11} y_i)\}^{-1} & \text{for } \mathcal{R}_1 \\ \Phi(\phi_{20} + \phi_{21} y_i) & \text{for } \mathcal{R}_2, \end{cases} \tag{25}$$

where $(\phi_{10}, \phi_{11}) = (0.8, -0.2), (\phi_{20}, \phi_{21}) = (0.5, -0.1)$ and $\Phi(\cdot)$ is cumulative distribution function of the standard normal distribution. The overall response rates are approximately around 70%. Thus, we have 3 × 2 setup for the simulation study.

For each simulation setup, $n = 500$ samples are generated independently for 2,000 times. For each Monte Carlo sample, we apply the following methods to estimate $\theta = E(y)$:

1. Full sample method: Use $\hat{\theta} = \sum_{i=1}^{n} y_i/n$, which is computed as a benchmark for the comparison.

2. Complete-Case (CC) method: Estimate θ by removing nonresponse. That is, $\hat{\theta}_{CC}$ is obtained by solving $\sum_{i=1}^{n} \delta_i(y_i - \theta) = 0$ for θ.

3. Kott and Chang (2010) (KC) method: Assume the response model is

$$Pr(\delta_i = 1 \mid x_i, y_i) = \pi(\phi; y_i) = \frac{\exp(\phi_0 + \phi_1 y_i)}{1 + \exp(\phi_0 + \phi_1 y_i)}. \tag{26}$$

The KC estimates are obtained by solving

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i}{\pi(\phi;y_i)}-1\right\}(1,x_i)'=\mathbf{0},$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\phi;y_i)}(y_i-\theta)=0.$$

4. Fractional imputation (FI) method: Use $y|(x,\delta=1)\sim N(\beta_0+\beta_1x_i,\sigma^2)$ and the response mechanism in (26) to obtain the predictive model. The maximum likelihood estimator of θ is computed by using Fractional Imputation (FI) method in Kim (2011). Set the size of FI is 20. A description of the FI algorithm is described in Appendix D.

5. Bayesian Data Augmentation (BDA) method: Apply the proposed method in Section 6 using the same model for FI method. In the data augmentation algorithm, we choose the burn-in size as 2,000 and after burn-in, iteration size is 2,000.

Thus, in the last two methods, the outcome model is misspecified under m_2 and m_3. Under $\mathcal{R}_2$, the response mechanism is slightly misspecified.

Table 2: Simulation results: "m" denotes mean function, "bias" is the estimator subtracting true value, "R_std" is the relative standard error which is relative to the standard error of full sample estimator.

Response	m	method	bias	R_std	Response	m	method	bias	R_std
$\mathcal{R}_1$	m_1	CC	-0.16	1.16	$\mathcal{R}_2$	m_1	CC	-0.14	1.17
		KC	-0.00	1.10			KC	-0.00	1.09
		FI	-0.00	1.09			FI	-0.00	1.08
		BDA	0.00	1.10			BDA	-0.00	1.09
	m_2	CC	-0.18	1.14		m_2	CC	-0.14	1.12
		KC	0.00	1.11			KC	-0.00	1.09
		FI	-0.01	1.10			FI	-0.01	1.08
		BDA	-0.00	1.11			BDA	-0.00	1.09
	m_3	CC	-1.13	1.15		m_3	CC	-0.95	1.13
		KC	-0.00	1.03			KC	0.01	1.02
		FI	-0.01	1.04			FI	0.01	1.03
		BDA	-0.00	1.04			BDA	0.01	1.03

Table 3: The coverage probabilities for the proposed method

method	m	res	cp
BDA	m1	$\mathcal{R}_1$	0.95
BDA	m2	$\mathcal{R}_1$	0.94
BDA	m3	$\mathcal{R}_1$	0.95
BDA	m1	$\mathcal{R}_2$	0.95
BDA	m2	$\mathcal{R}_2$	0.94
BDA	m3	$\mathcal{R}_2$	0.95

The simulation results are presented in Table 2 and 3. From Table 2, we can see that the performance of the proposed BDA method is similar to the KC and FI methods. Furthermore, the proposed BDA method can simultaneously construct correct confidence intervals and does not involve Taylor linearization. From Table 3, we can see that the coverage probabilities of the proposed method are around 0.95, which confirms the validity of the proposed BDA method.

8 Application

In this section, we apply the proposed Bayesian propensity score methods to Korea Labor and Income Panel Survey (KLIPS) data. A brief description of the panel survey can be found at http:// www.kli.re.kr/klips/en/about/introduce.jsp. The study variable (y) is the average monthly income for the current year and the auxiliary variable (x) can be demographic variables, such as the age groups and sex. Let (X_i, Y_{it}) be the observations for household i in panel year t. The KLIPS has $n = 5,013$ households and $T = 8$ panel years. We treat the first panel observations as the baseline measurements, and there are no missing data in the first year. In the panel survey, X_i are completely observed and Y_{it} are subject to missingness, for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, T$. Let δ_{it} be the response indicator function of Y_{it}. Define

$$\delta_{it} = \begin{cases} 1 & \text{if we observe } Y_{it} \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in estimating the probability of full response

$$\pi_i = Pr(\delta_{i1} = 1, \cdots, \delta_{iT} = 1 | X_i, Y_{i,obs}), \tag{27}$$

where $Y_{i,obs} = (Y_{i1}, \cdots, Y_{iT})'$ represents the observed responses for household i. The inverse of the π_i in (27) can be used as the propensity weight for the penal survey. For monotone missing data, in the sense of $\delta_{it} = 1$ implying $\delta_{i,t-1} = 1, \cdots, \delta_{i1} = 1$, the probability reduces to

$$\pi_i = \pi_{i1}\pi_{i2} \cdots \pi_{iT},$$

where $\pi_{it} = Pr(\delta_{it} = 1 | \delta_{i,t-1} = 1, X_i, Y_{i1}, \cdots, Y_{i,t-1})$ under MAR assumption.

For arbitrary missing patterns as in KLIPS, we first define $\delta^*_{it} = \prod_{k=1}^{t} \delta_{ik}$. Note that $\delta^*_{it} = 1$ implies that $\delta^*_{i,t-1} = 1$. Furthermore,

$$\begin{aligned} Pr(\delta_{i1} = 1, \cdots, \delta_{iT} & = 1 | X_i, Y_{i,obs}) = Pr(\delta^*_{i1} = 1, \cdots, \delta^*_{iT} = 1 | X_i, Y_{i,obs}) \\ & = \prod_{k=2}^{T} Pr(\delta^*_{ik} = 1 | \delta^*_{i,k-1} = 1, X_i, Y_{i,k-1}) \\ & = \prod_{k=2}^{T} Pr(\delta_{ik} = 1 | \delta^*_{i,k-1} = 1, X_i, Y_{i,k-1}) \\ & = \pi_{i2}\pi_{i3} \cdots \pi_{iT} = \pi_i, \end{aligned}$$

where $\pi_{i1} = 1$ for all samples.

Thus, we can build a parametric model for $\pi_{it} = Pr(\delta_{it} = 1 | \delta^*_{i,t-1} = 1, X_i, Y_{i,t-1})$ and estimate the parameters sequentially. Instead of using the frequentist approach of Zhou and Kim (2012), we apply the BPS method in Section 3 and OBPS method in Section 5 to incorporate the extra information in X.

We are interested in estimating the average income for the final year and constructing confidence intervals for the parameters. Assume the response mechanism follows

$$\pi(\phi_t; X_i, Y_{i,t-1}) =: Pr(\delta_{it} = 1 | \delta^*_{i,t-1} = 1, X_i, Y_{i,t-1}) = \frac{1}{1 + \exp\{-(X'_i, Y_{i,t-1})\phi_t\}}, \quad (28)$$

which is known up to parameter ϕ_t. Thus, we allow that the response probability at year t depends on the last year income y_{t-1}, but not on the current year income. Assume δ_{it}, given $\delta^*_{i,t-1} = 1$, X_i, and $Y_{i,t-1}$, independently follow Bernoulli distribution with probability $\pi(\phi_t; X_i, Y_{i,t-1})$ in (28). Therefore, the score function of ϕ_t is

$$S(\phi_t) = \frac{1}{n} \sum_{i=1}^{n} \{\delta_{it} - \pi(\phi_t; X_i, Y_{i,t-1})\} (X'_i, Y_{i,t-1})' \delta^*_{i,t-1}.$$

Then the joint estimating equations are $U_n(\phi_2, \phi_3, \cdots, \phi_T, \theta) = 0$, where

$$U_n(\phi_2, \phi_3, \cdots, \phi_T, \theta) = n^{-1} \sum_{i=1}^{n} \left[\begin{array}{l} \{\delta_{i2} - \pi(\phi_2; X_i, Y_{i,1})\} (X_i', Y_{i,1})'\delta_{i,1}^* \\ \vdots \\ \{\delta_{iT} - \pi(\phi_T; X_i, Y_{i,T-1})\} (X_i', Y_{i,T-1})'\delta_{i,T-1}^* \\ \pi_i^{-1}\delta_{iT}^* y_{i,T} - \theta, \end{array} \right] \tag{29}$$

and $\theta = E(Y_T)$.

The Bayesian propensity score (BPS) method can be described as

1. Solve $U_n(\phi_2, \phi_3, \cdots, \phi_T, \theta) = \mathbf{0}$ to obtain $\hat{\phi}_2, \cdots, \hat{\phi}_T$, and $\hat{\theta}$.

2. Generate $\eta^* = (\eta_1^{*'}, \eta_2^{*'})'$ from $N(\mathbf{0}, \hat{\Sigma}/n)$, where $\hat{\Sigma}$ is a consistent variance estimator of $\sqrt{n}U_n(\phi_2, \phi_3, \cdots, \phi_T, \theta)$.

3. Solve $(S'(\phi_2), \cdots, S'(\phi_T))' = \eta_1^*$ to obtain $\phi_2^*, \cdots, \phi_T^*$.

4. Compute $\pi_i^* = \pi(\phi_2^*; X_i, Y_{i,1}) \times \cdots \times \pi(\phi_T^*; X_i, Y_{i,T-1})$. Solve

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{iT}^*}{\pi_i^*}(y_{i,T} - \theta) = \eta_2^*$$

to obtain θ^*.

Repeat the above steps independently to generate samples from the posterior distribution of parameters. The variance-covariance matrix $\hat{\Sigma}$ can be derived by

$$\frac{1}{n} \sum_{i=1}^{n} \left[\begin{array}{c} \left\{\delta_{i2} - \pi(\hat{\phi}_2; X_i, Y_{i,1})\right\} (X_i', Y_{i,1})'\delta_{i,1}^* \\ \vdots \\ \left\{\delta_{iT} - \pi(\hat{\phi}_T; X_i, Y_{i,T-1})\right\} (X_i', Y_{i,T-1})'\delta_{i,T-1}^* \\ \hat{\pi}_i^{-1}\delta_{iT}^* y_{i,T} - \hat{\theta} \end{array} \right]^{\otimes 2}.$$

To improve the efficiency of the point estimator, we also apply OBPS method to the same data. In addition to equations in (29), we add

$$\sum_{i=1}^{n} \frac{\delta_{iT}^*}{\pi_i} (X_i - \mu_x) = 0$$

$$\sum_{i=1}^{n} (X_i - \mu_x) = 0,$$

where μ_x is the marginal proportion vector for demographical covariates. Therefore, the posterior distribution of θ can be obtained by applying the proposed algorithm in Section 5.

For a comparison, we also considered a naive method which does not use the propensity model and apply the Bayesian method in the complete cases (CC) only. We apply BPS, OBPS and CC method to $T = 2, 3, 4$. The numerical results are presented below.

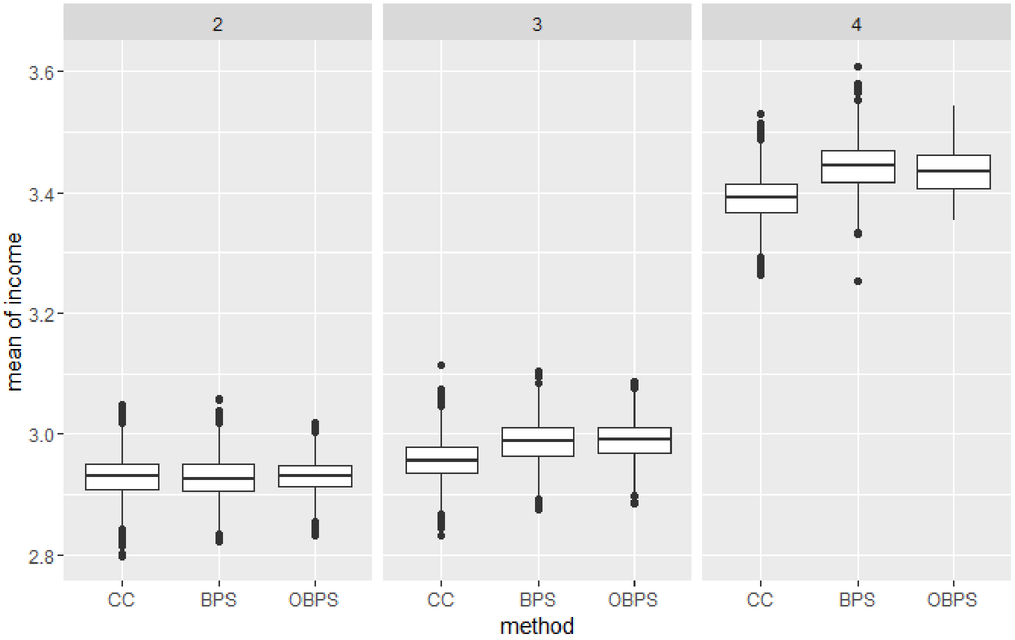

Figure 2: Boxplots for posterior distribution of θ by different methods and different panels. (Magnitude 1,000,000 Won)

From Figure 2, all three methods provide similar estimators for the average income θ.The trend of average income goes up as year T increases. For year $T = 2$, all three methods provide similar mean estimates. But the OBPS method is the most efficient. For year $T = 3$, we see that the CC method provides lower mean estimate than BPS or OBPS, which is due to the nonresponse bias in the CC method. This phenomenon becomes more obvious for year $T = 4$. Also, the lengths of confidence intervals increase as T increases, since the fully observed sample size is decreasing due to panel attrition. The CC method presents smaller values of θ for $T = 4$, which suggests more panel attrition for higher income households. Both BPS and OBPS provide similar mean estimates. But the OBPS method has narrower confidence intervals, which confirms the efficiency of the OBPS method.

9 Concluding Remarks

A new Bayesian inference using PS method is developed using the idea of Approximate Bayesian computation. The proposed method can be widely applicable due to popularity of PS method. The proposed Bayesian approach is calibrated to frequentist inference in the sense that the proposed method provides the same inferential results with its frequentist version asymptotically (Little, 2012). The calibration property holds if the prior distribution for the model parameters is flat. If the prior is informative then the resulting Bayesian inference will be more efficient than frequentist inference thanks to its natural incorporation of the prior information. Thus, the proposed method is applicable when the need of combining information from different sources.

Causal inference, including estimation of average treatment effect from observational studies, can be one promising application area of the PS method (Morgan and Winship, 2014 and Hudgens and Halloran, 2008). Developing tools for causal inference using the Bayesian PS method will be an important extension of this research. Also, Bayesian model selection method (Ishwaran and Rao, 2005) can be naturally applied to this setup. Such extensions will be topics for future research.

Appendix

A. Consistent variance estimator

From the asymptotic distribution in (9), we can write

$$[U_n|\boldsymbol{\eta}] \sim N\left(\boldsymbol{\eta}, \Sigma/n\right).$$

To emphasize that Σ is a function of ϕ, θ, we use $\Sigma =: \Sigma\left(\phi, \theta\right)$. Since the transformation

$$\begin{pmatrix}\phi\\ \theta\end{pmatrix} \rightarrow \begin{pmatrix}\boldsymbol{\eta}_1\\ \eta_2\end{pmatrix}$$

is one-to-one, $\Sigma\left(\phi, \theta\right)$ is equivalent to $\Sigma\left(\boldsymbol{\eta}_1, \eta_2\right)$, where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \eta_2)'$. The corresponding density function is

$$p\left(U_n|\boldsymbol{\eta}\right) \propto \left|\Sigma\left(\boldsymbol{\eta}\right)/n\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(U_n - \boldsymbol{\eta}\right)'\left(\Sigma\left(\boldsymbol{\eta}\right)/n\right)^{-1}\left(U_n - \boldsymbol{\eta}\right)\right\}.$$

Since we have assigned a flat prior, in the sense of $\pi\left(\boldsymbol{\eta}\right) \propto 1$, we can derive the posterior distribution as

$$p\left(\boldsymbol{\eta}|U_n\right) \propto \left|\Sigma\left(\boldsymbol{\eta}\right)/n\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(U_n - \boldsymbol{\eta}\right)'\left(\Sigma\left(\boldsymbol{\eta}\right)/n\right)^{-1}\left(U_n - \boldsymbol{\eta}\right)\right\}.$$

By the definition of $\boldsymbol{\eta}$, U_n is the unbiased estimator of $\boldsymbol{\eta}$. Thus, we write $\hat{\boldsymbol{\eta}} = U_n$. To show that

$$p\left(\boldsymbol{\eta}|U_n\right) \propto \left|\Sigma\left(\hat{\boldsymbol{\eta}}\right)/n\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(U_n - \boldsymbol{\eta}\right)'\left(\Sigma\left(\hat{\boldsymbol{\eta}}\right)/n\right)^{-1}\left(U_n - \boldsymbol{\eta}\right)\right\},$$

we first show that $\Sigma(\cdot)$ is continuous, which can be proved by the dominated convergence theorem applied to $\boldsymbol{\eta}_1(\cdot)$ and $\eta_2(\cdot)$. Now, noting that, by asymptotic distribution (9) and Chebyshev's inequality, we can show that $U_n \xrightarrow{P} \boldsymbol{\eta}$. Thus, we can obtain $\Sigma(\hat{\boldsymbol{\eta}}) \xrightarrow{P} U_n(\boldsymbol{\eta})$. Since Σ is positive definite and $x^{-1/2}$ is continuous if $x > 0$, $|\Sigma(\hat{\boldsymbol{\eta}})|^{-1/2} \xrightarrow{P} |\Sigma(\boldsymbol{\eta})|^{-1/2}$. By the continuous mapping theorem,

$$\sqrt{n}\Sigma(\hat{\boldsymbol{\eta}})^{-1/2}\left(U_n - \boldsymbol{\eta}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}).$$

Therefore, we can derive the posterior distribution as

$$p\left(\boldsymbol{\eta}|U_n\right) \propto \left|\Sigma\left(\hat{\boldsymbol{\eta}}\right)/n\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(U_n - \boldsymbol{\eta}\right)'\left(\Sigma\left(\hat{\boldsymbol{\eta}}\right)/n\right)^{-1}\left(U_n - \boldsymbol{\eta}\right)\right\}.$$

That is $[\boldsymbol{\eta}|U_n = \mathbf{0}] \sim N(\mathbf{0}, \Sigma(U_n = \mathbf{0})/n)$, which is equivalent to

$$p\left(\boldsymbol{\eta}|U_n = \mathbf{0}\right) \propto \left|\Sigma(\hat{\phi}, \hat{\theta})/n\right|^{-1/2} \exp\left[-\frac{1}{2}\left(U_n - \boldsymbol{\eta}\right)' \left\{\Sigma\left(\hat{\phi}, \hat{\theta}\right)/n\right\}^{-1} \left(U_n - \boldsymbol{\eta}\right)\right],$$

where $(\hat{\phi}, \hat{\theta})$ is the solution to $U_n = \mathbf{0}$. Furthermore, the consistency of $\hat{\Sigma} = \hat{\Sigma}(\hat{\phi}, \hat{\theta})$ in (11) can be proved using the law of large numbers.

B. Proof of Theorem 4.1

Step I

From Condition [C9], we assume that $\boldsymbol{\zeta} \mapsto U_n(\boldsymbol{\zeta})$ and $\boldsymbol{\zeta} \mapsto \boldsymbol{\eta}(\boldsymbol{\zeta})$ are one-to-one functions, for any $\boldsymbol{\zeta} \in N(\boldsymbol{\zeta}_0)$. Denote these two mappings as T_n and T respectively. Because of their one-to-one property, their inverse mappings exist for $\boldsymbol{\zeta} \in N_n(\boldsymbol{\zeta}_0)$. Therefore, we can write (12) as

$$\sqrt{n}\left(U_n - \boldsymbol{\eta}\right) \xrightarrow{d} N[0, \Sigma\left\{T^{-1}\left(\boldsymbol{\eta}\right)\right\}],$$

which leads to

$$p(U_n|\boldsymbol{\eta}) \to \phi_{\boldsymbol{\eta}, n^{-1}\Sigma(T^{-1}(\boldsymbol{\eta}))}\left(U_n\right).$$

Thus, by the convergence of U_n to $\boldsymbol{\eta}$ and using the argument similar to the proof for Lemma 1 in Soubeyrand and Haon-Lasportes (2015), we can show that

$$p(\boldsymbol{\eta}|U_n) = \phi_{U_n, n^{-1}\Sigma\left(T_n^{-1}(U_n)\right)}(\boldsymbol{\eta})\left\{1 + o_p(1)\right\}. \tag{B.1}$$

Step II

Note that $U_n(\hat{\boldsymbol{\zeta}}) = 0$, thus $T_n^{-1}(0) = \hat{\boldsymbol{\zeta}}$. From (B.1), we can therefore get the posterior distribution

$$p(\boldsymbol{\eta}|U_n = 0) = p(\boldsymbol{\eta}|\hat{\boldsymbol{\zeta}}) = \phi_{0, n^{-1}\Sigma(\hat{\boldsymbol{\zeta}})}(\boldsymbol{\eta})\left\{1 + o_p(1)\right\}. \tag{B.2}$$

Thus, we can write the density $p(\boldsymbol{\eta}|U_n = 0)$ as

$$\phi_{0, n^{-1}\Sigma(\hat{\boldsymbol{\zeta}})}(\boldsymbol{\eta}) \propto \exp\left\{-\frac{n}{2}\boldsymbol{\eta}'\Sigma^{-1}(\hat{\boldsymbol{\zeta}})\boldsymbol{\eta}\right\}.$$

Furthermore, by the consistency of the variance estimator provided in condition [C7], we can obtain $\hat{\Sigma} := \hat{\Sigma}(\hat{\boldsymbol{\zeta}}) = \Sigma(\hat{\boldsymbol{\zeta}})\left\{1 + o_p(1)\right\}$. Thus,

$$\boldsymbol{\eta}'\Sigma^{-1}(\hat{\boldsymbol{\zeta}})\boldsymbol{\eta} = \boldsymbol{\eta}'\left\{\hat{\Sigma}^{-1}(1 + o_p(1))\right\}\boldsymbol{\eta} = \boldsymbol{\eta}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}\left\{1 + o_p(1)\right\},$$

which leads to

$$\phi_{0,n^{-1}\Sigma(\hat{\zeta})}(\boldsymbol{\eta}) = \phi_{0,n^{-1}\hat{\Sigma}}(\boldsymbol{\eta})\left\{1 + o_p(1)\right\}\exp\left\{-\frac{n}{2}o_p\left(\boldsymbol{\eta}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}\right)\right\}.$$

From [C1] and [C5], we have $U_n(\boldsymbol{\zeta}) \to \boldsymbol{\eta}$ in probability and $U_n = O_p(1/\sqrt{n})$ for $\boldsymbol{\zeta} \in N(\boldsymbol{\zeta}_0)$, which leads to $\boldsymbol{\eta} = O(1/\sqrt{n})$. Thus,

$$\exp\left\{-\frac{n}{2}o_p\left(\boldsymbol{\eta}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}\right)\right\} = \exp\left\{o_p(1)\right\} \to 1,$$

in probability and the following follows

$$p(\boldsymbol{\eta}|U_n = 0) = p(\boldsymbol{\eta}|\hat{\boldsymbol{\zeta}}) = \phi_{0,n^{-1}\hat{\Sigma}}(\boldsymbol{\eta})\left\{1 + o_p(1)\right\}. \tag{B.3}$$

Step III

Let $\boldsymbol{\eta}^*$ be generated from the asymptotic posterior distribution (B.3) which is a normal distribution with mean 0 and variance $\hat{\Sigma}/n$. Therefore, the j-th component $\boldsymbol{\zeta}_j^*$ of $\boldsymbol{\zeta}^*$ satisfies

$$\begin{aligned}
E\left\{\boldsymbol{\zeta}_j^*|\hat{\boldsymbol{\zeta}}_n\right\} &= E\left\{T_{n,j}^{-1}(\boldsymbol{\eta}^*)|\hat{\boldsymbol{\zeta}}_n\right\} \\
&= E\left\{T_{n,j}^{-1}(0) + \left.\frac{\partial T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\eta^* + \frac{1}{2}\boldsymbol{\eta}^{*\prime}\left.\frac{\partial^2 T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\boldsymbol{\eta}^* + o_p(\boldsymbol{\eta}^{*\prime}\boldsymbol{\eta}^*)\right|\hat{\boldsymbol{\zeta}}_n\right\} \\
&= \hat{\boldsymbol{\zeta}}_{n,j} + \frac{1}{2}E\left\{\boldsymbol{\eta}^{*\prime}\left.\frac{\partial^2 T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\boldsymbol{\eta}^*\right|\hat{\boldsymbol{\zeta}}_n\right\} + o\left(\frac{1}{n}\right).
\end{aligned}$$

By $E(\boldsymbol{Z}'\Lambda\boldsymbol{Z}) = tr(\Lambda\Sigma) + \mu'\Lambda\mu$, we derive

$$E\left\{\boldsymbol{\eta}^{*\prime}\left.\frac{\partial^2 T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\boldsymbol{\eta}^*\right|\hat{\boldsymbol{\zeta}}_n\right\} = tr\left[\left.\frac{\partial^2 T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\frac{\hat{\Sigma}}{n}\right] = O\left(\frac{1}{n}\right),$$

under [C4]. Therefore, we have

$$E\left\{\boldsymbol{\zeta}_j^*|\hat{\boldsymbol{\zeta}}\right\} = \hat{\boldsymbol{\zeta}}_{n,j} + O\left(\frac{1}{n}\right),$$

for $j = 1, 2, \cdots, p$, which establishes

$$E\left\{\boldsymbol{\zeta}^*|\hat{\boldsymbol{\zeta}}\right\} = \hat{\boldsymbol{\zeta}}_n + O\left(\frac{1}{n}\right). \tag{B.4}$$

Step IV

Now, the posterior variance of $\boldsymbol{\zeta}_j^*$:

$$\begin{aligned}
Var\left\{\boldsymbol{\zeta}_j^*|\hat{\boldsymbol{\zeta}}\right\} &= Var\left\{T_{n,j}^{-1}(0) + \left.\frac{\partial T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\boldsymbol{\eta}^* + \frac{1}{2}\boldsymbol{\eta}^{*\prime}\left.\frac{\partial^2 T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\boldsymbol{\eta}^* + o_p(\boldsymbol{\eta}^{*\prime}\boldsymbol{\eta}^*)\middle|\hat{\boldsymbol{\zeta}}\right\} \\
&= Var\left\{\left.\frac{\partial T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\boldsymbol{\eta}^* + \frac{1}{2}\boldsymbol{\eta}^{*\prime}\left.\frac{\partial^2 T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\boldsymbol{\eta}^* + o_p(\boldsymbol{\eta}^{*\prime}\boldsymbol{\eta}^*)\middle|\hat{\boldsymbol{\zeta}}\right\}.
\end{aligned}$$

The first term is

$$\begin{aligned}
Var\left\{\left.\tfrac{\partial T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\boldsymbol{\eta}^*\middle|\hat{\boldsymbol{\zeta}}\right\} &= \left.\frac{\partial T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0} Var\left\{\boldsymbol{\eta}^*|\hat{\boldsymbol{\zeta}}\right\}\left\{\left.\frac{\partial T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\right\}' \\
&= O\left(\frac{1}{n}\right). \qquad \text{(B.5)}
\end{aligned}$$

For the second term, using

$$\begin{aligned}
Var(\boldsymbol{Z}'\Lambda\boldsymbol{Z}) &= 2tr(\Lambda\Sigma\Lambda\Sigma) + 4\mu'\Lambda\Sigma\Lambda, \\
Cov(\boldsymbol{Z}'\Lambda_1\boldsymbol{Z}, \boldsymbol{Z}'\Lambda_2\boldsymbol{Z}) &= 2tr(\Lambda_1\Sigma\Lambda_2\Sigma) + 4\mu'\Lambda_1\Sigma\Lambda_2,
\end{aligned}$$

for $\boldsymbol{Z} \sim N(\mu, \Sigma)$, we have

$$Var\left\{\boldsymbol{\eta}^{*\prime}\left.\frac{\partial^2 T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\boldsymbol{\eta}^*\middle|\hat{\boldsymbol{\zeta}}\right\} = 2tr\left\{\left.\frac{\partial^2 T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\frac{\hat{\Sigma}}{n}\left.\frac{\partial^2 T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\frac{\hat{\Sigma}}{n}\right\} = O\left(\frac{1}{n^2}\right).$$

The covariance of two terms is less than the square root of their variances. We have shown that the variance of the first term is in the order of $O(1/n)$ and the variance of the second term is in the order of $O(1/n^2)$. So the covariance has the order of $O(n^{-3/2})$.

Similarly, we can derive

$$Cov(\boldsymbol{\zeta}_j^*, \boldsymbol{\zeta}_k^*|\hat{\boldsymbol{\zeta}}) = \left.\frac{\partial T_{n,j}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0} Var\left\{\boldsymbol{\eta}^*|\hat{\boldsymbol{\zeta}}\right\}\left\{\left.\frac{\partial T_{n,k}^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\right\}' + o\left(\frac{1}{n}\right). \qquad \text{(B.6)}$$

Combining (B.5) and (B.6), we have

$$Var(\zeta^*|\hat{\boldsymbol{\zeta}}) = \left.\frac{\partial T_n^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0} Var\left\{\boldsymbol{\eta}^*|\hat{\boldsymbol{\zeta}}\right\}\left\{\left.\frac{\partial T_n^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0}\right\}' + o\left(\frac{1}{n}\right). \qquad \text{(B.7)}$$

Step V

By Conditions [C1]-[C5], we have

$$\sqrt{n}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \xrightarrow{d} N(0, A^{-1}(\boldsymbol{\zeta}_0)\Sigma(\boldsymbol{\zeta}_0)A'^{-1}(\boldsymbol{\zeta}_0)),$$

where $A(\boldsymbol{\zeta}) = \partial\boldsymbol{\eta}(\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}$. See Theorem 5.21 in Van der Vaart (2000).

Since $T_n \to T$ uniformly by [C1] and both mappings are one-to-one functions, we can state that $T_n^{-1} \to T^{-1}$ uniformly. Thus,

$$\left.\frac{\partial T_n^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0} \xrightarrow{P} \left.\frac{\partial T^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0} = A^{-1}(\boldsymbol{\zeta}_0).$$

Also, by [C7], we have $\hat{\Sigma} \xrightarrow{P} \Sigma(\hat{\boldsymbol{\zeta}})$ and $\hat{\boldsymbol{\zeta}} \xrightarrow{P} \boldsymbol{\zeta}_0$. By the Lipschitz continuity of $\Sigma(\zeta)$, we can conclude that $\Sigma(\hat{\boldsymbol{\zeta}}) \xrightarrow{P} \Sigma(\boldsymbol{\zeta}_0)$. Thus, $\hat{\Sigma} \xrightarrow{P} \Sigma(\boldsymbol{\zeta}_0)$

and

$$\begin{aligned} nVar(\boldsymbol{\zeta}^*|\hat{\boldsymbol{\zeta}}) - nVar(\hat{\boldsymbol{\zeta}}) \quad &= \left.\frac{\partial T_n^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0} \hat{\Sigma} \left\{ \left.\frac{\partial T_n^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}=0} \right\}' \\ &\quad -A^{-1}(\boldsymbol{\zeta}_0)\Sigma(\boldsymbol{\zeta}_0)A'^{-1}(\boldsymbol{\zeta}_0) \xrightarrow{P} 0, \end{aligned} \tag{B.8}$$

by the continuous mapping theorem. Combining the previous conclusions (B.4), (B.7) and (B.8), we can use Slutsky's theorem to get

$$\left\{Var(\hat{\boldsymbol{\zeta}})\right\}^{-1/2}(\boldsymbol{\zeta}^* - \hat{\boldsymbol{\zeta}})|\hat{\boldsymbol{\zeta}} \xrightarrow{d} N(0, \boldsymbol{I}_p),$$

which proves (14).

Step VI

Let $\alpha \in (0, 1)$, and define

$$C_{n,\alpha} = \left\{\boldsymbol{\zeta}^* : (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*)' \left\{Var(\hat{\boldsymbol{\zeta}})\right\}^{-1} (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*) \leq \chi_p^2(\alpha)\right\},$$

where the $\chi_p^2(\alpha)$ is the α quantile of Chi-square distribution with p degrees of freedom.

Furthermore, from a property of the Raylei quotient (Horn and Johnson, 1985), there exists a matrix O such that

$$O\left\{Var^{-1}(\hat{\boldsymbol{\zeta}})/n\right\}O^T = \text{diag}\left\{\lambda_1, \cdots, \lambda_p\right\},$$

where $OO^T = \boldsymbol{I}_p$ and $0 < \lambda_1 \leq \lambda_2, \cdots, \leq \lambda_p$. Thus we obtain

$$\boldsymbol{x}^T \left\{ Var^{-1}(\hat{\boldsymbol{\zeta}}) \right\} \boldsymbol{x} \geq n\lambda_1 \boldsymbol{x}^T \boldsymbol{x}. \tag{B.9}$$

Also, we can apply the conclusion (B.9) to get

$$\|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*\| \leq \lambda_1^{-1/2} \sqrt{(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*)' \left\{ Var(\hat{\boldsymbol{\zeta}}) \right\}^{-1} (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*)/n} \leq \lambda_1^{-1/2} \sqrt{\chi_p^2(\alpha)/n}$$

for all $\boldsymbol{\zeta}^* \in C_{n,\alpha}$.

Similarly, by the asymptotic normality of the estimator $\hat{\boldsymbol{\zeta}}$ and applying the conclusion (B.9),

$$\|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\| \leq \lambda_1^{-1/2} \sqrt{(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)^T \left\{ Var(\hat{\boldsymbol{\zeta}}) \right\}^{-1} (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)} \leq \lambda_1^{-1/2} \sqrt{\chi_p^2(\alpha)/n}. \tag{B.10}$$

Next, from (B.10), we can conclude that

$$\lim_{n \longrightarrow \infty} Pr\left(\|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\| \leq \lambda_1^{-1/2} \sqrt{\chi_p^2(\alpha)/n} \right) \geq \alpha.$$

By the inequality $\|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}_0\| \leq \|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*\| + \|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\|$, we obtain

$$\lim_{n \longrightarrow \infty} Pr\left(\forall \boldsymbol{\zeta}^* \in C_{n,\alpha}, \quad \|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}_0\| \leq 2\lambda_1^{-1/2} \sqrt{\chi_p^2(\alpha)/n} \right) \geq \alpha. \tag{B.11}$$

Since we have defined $N_n(\boldsymbol{\zeta}_0)$ in a neighborhood with center $\boldsymbol{\zeta}_0$ and radius r_n, where r_n satisfies $r_n \to 0$ and $\sqrt{n} r_n \to \infty$. From (B.11),

$$\lim_{n \longrightarrow \infty} Pr(\forall \boldsymbol{\zeta}^* \in C_{n,\alpha}, \quad \|\boldsymbol{\zeta}^* - \boldsymbol{\zeta}_0\| \leq r_n) \geq \alpha,$$

$$\lim_{n \longrightarrow \infty} Pr(C_{n,\alpha} \subset N_n(\boldsymbol{\zeta}_0)) \geq \alpha.$$

Therefore,

$$\lim_{n \longrightarrow \infty} Pr\left(\int_{N_n(\boldsymbol{\zeta}_0)} \phi_{\hat{\boldsymbol{\zeta}}, Var(\hat{\boldsymbol{\zeta}})}(\boldsymbol{\zeta}^*) d\boldsymbol{\zeta}^* \geq \int_{C_{n,\alpha}} \phi_{\hat{\boldsymbol{\zeta}}, Var(\hat{\boldsymbol{\zeta}})}(\boldsymbol{\zeta}^*) d\boldsymbol{\zeta}^* \right) \geq \alpha.$$

This is equivalent to

$$\lim_{n \longrightarrow \infty} Pr\left(\int_{N_n(\boldsymbol{\zeta}_0)} \phi_{\hat{\boldsymbol{\zeta}}, Var(\hat{\boldsymbol{\zeta}})}(\boldsymbol{\zeta}^*) d\boldsymbol{\zeta}^* \geq \alpha \right) \geq \alpha.$$

The above conclusion holds for any $\alpha \in (0, 1)$. Thus

$$\lim_{n \longrightarrow \infty} \int_{N_n(\boldsymbol{\zeta}_0)} \phi_{\hat{\boldsymbol{\zeta}}, Var(\hat{\boldsymbol{\zeta}})}(\boldsymbol{\zeta}^*) d\boldsymbol{\zeta}^* = 1 \quad \text{in probability.}$$

C. Computational Details for the Metropolis-Hastings Algorithm

Implementing the optimal Bayesian propensity score (OBPS) method is done through the following algorithm.

1. Choose the initial value for $\boldsymbol{\psi}$ and denote it as $\boldsymbol{\psi}_0$.

2. For iteration t, given the current parameter value $\boldsymbol{\psi}_t$, generate $\Delta\boldsymbol{\psi}$ from $N(\mathbf{0}, V)$, where V is a tunning parameter obtainable by the data-driven method discussed below. Let the candidate value be $\boldsymbol{\psi}^* = \boldsymbol{\psi}_t + \Delta\boldsymbol{\psi}$.

3. Compute the acceptance probability
$$\alpha = \alpha(\boldsymbol{\psi}^*|\boldsymbol{\psi}_t) = \min\left\{1, \frac{g\left(U_n|\boldsymbol{\psi}^*\right)\pi(\boldsymbol{\psi}^*)}{g\left(U_n|\boldsymbol{\psi}_t\right)\pi(\boldsymbol{\psi}_t)}\right\}.$$

4. Generate u from Uniform $(0,1)$ distribution. If $u < \alpha$, accept the candidate $\boldsymbol{\psi}_{t+1} = \boldsymbol{\psi}^*$. Otherwise let $\boldsymbol{\psi}_{t+1} = \boldsymbol{\psi}_t$.

5. For burning in period, discard the values from the first B iterations. Then collect M values. These M values can be treated as values generated from the target posterior distribution.

For the choice of the initial value for $\boldsymbol{\psi}$, we can use the solution to
$$(U_1'(\phi), U_2(\phi, \theta), U_4'(\boldsymbol{\mu}_x))' = \mathbf{0}.$$

In Metropolis-Hastings algorithm, the value of V for the random walk will directly affect the speed of convergence of the Markov chain and the acceptance rate. We recommend a data-driven method to set V. A data-driven choice of V can be obtained from the posterior variance of the Monte Carlo samples from $p\{\boldsymbol{\mu}_x, \phi, \theta|(U_1, U_2, U_4) = \mathbf{0}\}$.

To compute the acceptance probability, we need to compute the ratio
$$\frac{g\left(U_n|\boldsymbol{\psi}^*\right)}{g\left(U_n|\boldsymbol{\psi}_t\right)} = \frac{|\Sigma(\boldsymbol{\psi}^*)|^{-1/2}\exp\left\{-\frac{n}{2}U_n'(\boldsymbol{\psi}^*)\Sigma^{-1}(\boldsymbol{\psi}^*)U_n(\boldsymbol{\psi}^*)\right\}}{|\Sigma(\boldsymbol{\psi}_t)|^{-1/2}\exp\left\{-\frac{n}{2}U_n'(\boldsymbol{\psi}_t)\Sigma^{-1}(\boldsymbol{\psi}_t)U_n(\boldsymbol{\psi}_t)\right\}},$$

which can be approximated by

$$\exp\left\{-\frac{n}{2}U_n'(\boldsymbol{\psi}*)\hat{\Sigma}^{-1}U_n(\boldsymbol{\psi}*)+\frac{n}{2}U_n'(\boldsymbol{\psi}_t)\hat{\Sigma}^{-1}U_n(\boldsymbol{\psi}_t)\right\},$$

where

$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} s(\hat{\phi};\mathbf{x}_i) \\ \delta_i\hat{\pi}_i^{-1}U(\hat{\theta};\mathbf{x}_i,y_i) \\ \delta_i\hat{\pi}_i^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_x) \\ (\mathbf{x}_i-\hat{\boldsymbol{\mu}}_x)\end{pmatrix}^{\otimes 2},$$

and $(\hat{\boldsymbol{\mu}}_x,\hat{\phi},\hat{\theta})$ are the consistent estimators.

D. Fractional imputation algorithm in simulation study two

Let $s(\phi;\delta_i,\mathbf{x}_1,y)$ be the score function for the response model. In additional to the response model, we also assume $f(y\mid x,\delta=1;\gamma)$. To solve the observed score function of ϕ, that is

$$\bar{S}(\phi)=\sum_{i=1}^{n}\left[\delta_i s(\phi;\delta_i,x_i,y_i)+(1-\delta_i)E\left\{s(\phi;\delta_i,x_i,y)|x_i,\delta_i=0\right\}\right]=0, \quad \text{(D.1)}$$

where the conditional expectation is with respect to the prediction model in (20). The estimate $\hat{\gamma}$ of γ can be obtained by using the observed data to fit the model $f(y\mid x,\delta=1;\gamma)$. To compute the solution $\hat{\phi}$ to (D.1), EM algorithm using fractional imputation (Kim, 2011) can be used. The algorithm is described as followings:

E-step: For each unit i with $\delta_i=0$, generate y_{ij}^* from $f_1(y\mid x_i,\delta_i=1;\hat{\gamma})$ for $j=1,2,\cdots,b$. Given the current value of ϕ_1, compute the fractional weights of y_{ij} as

$$w_{ij}^*\propto O(\phi;y_{ij}^*)=\frac{1-Pr(\delta_i=1\mid x_i,y_{ij}^*)}{Pr(\delta_i=1\mid x_i,y_{ij}^*)}\propto\exp(-\phi_1 y_{ij}^*),$$

subject to $\sum_{j=1}^{b}w_{ij}^*=1$.

M-step: Update ϕ by solving

$$\bar{S}(\phi)=\sum_{i=1}^{n}\left[\delta_i s(\phi;\delta_i,x_i,y_i)+(1-\delta_i)\sum_{j=1}^{b}w_{ij}^* s(\phi;\delta_i,x_i,y_{ij}^*)\right]=0.$$

Repeat **E-step** and **M-step** iteratively until convergence. After convergence, the final estimator of $\theta = E(Y)$ is constructed by

$$\hat{\theta}_{FI} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i y_i + (1 - \delta_i) \sum_{j=1}^{b} w_{ij}^{*} y_{ij}^{*} \right\}.$$

References

An, W. (2010). Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology 40*(1), 151–189.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate Bayesian Computation in population genetics. *Genetics 162*(4), 2025–2035.

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*(3), 723–734.

Chen, M.-H. and Q.-M. Shao (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics 8*(1), 69–92.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician 49*(4), 327–335.

Flanders, W. D. and S. Greenland (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine 10*(5), 739–747.

Horn, R. A. and C. R. Johnson (1985). Matrix Analysis Cambridge University Press. *New York*.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician 50*(2), 120–126.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 243–263.

Ishwaran, H. and J. S. Rao (2005). Spike and Slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics 33*(2), 730–773.

Kaplan, D. and J. Chen (2012). A two-step Bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika 77*(3), 581–609.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika 98*(1), 119–132.

Kim, J. K. and J. J. Kim (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics 35*(4), 501–514.

Kim, J. K. and J. Shao (2013). *Statistical Methods for Handling Incomplete Data*. CRC Press.

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*(493), 157–165.

Kott, P. S. and T. Chang (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association 105*(491), 1265–1275.

Little, R. J. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics. *Journal of Official Statistics 28*(3), 309.

McCandless, L. C., P. Gustafson, and P. C. Austin (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine 28*(1), 94–112.

Morgan, S. L. and C. Winship (2014). *Counterfactuals and Causal inference*. Cambridge University Press.

Riddles, M. K., J. K. Kim, and J. Im (2016). A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology 4*(2), 215.

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.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association 82*(398), 387–394.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Soubeyrand, S. and E. Haon-Lasportes (2015). Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in abc. *Statistics & Probability Letters 107*, 84–92.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association 82*(398), 528–540.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Volume 3. Cambridge university press.

Wang, S., J. Shao, and J. K. Kim (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica 24*, 1097–1116.

Zhou, M. and J. K. Kim (2012). An efficient method of estimation for longitudinal surveys with monotone missing data. *Biometrika 99*(3), 631–648.