Journal of Statistical Planning and Inference 207 (2020) 246–254

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Robust estimation for moment condition models with data missing not at random

Wei Li [a], Shu Yang [b,*], Peisong Han [c]

[a] Department of Mathematics, Syracuse University, Syracuse, NY 13244, USA
[b] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
[c] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 4 July 2018
Received in revised form 10 July 2019
Accepted 2 January 2020
Available online 13 January 2020

Keywords:
Identification
Empirical likelihood
Missing not at random
Multiple robustness
Semiparametric maximum likelihood estimator

ABSTRACT

We consider estimation for parameters defined through moment conditions when data are missing not at random. The missingness mechanism cannot be determined from the data alone, and inference under missingness not at random may be sensitive to unverifiable assumptions about the missingness mechanism. To add protection against model misspecification, we posit multiple models for the response probability and propose a weighting estimator with calibrated weights. Assuming the conditional distribution of the outcome given covariates is correctly modeled, we show that if any one of the multiple models for the response probability is correctly specified, the proposed estimator is consistent for the true value. A simulation study confirms that our estimator has multiple robustness when the outcome data is missing not at random. The method is also applied to an application.

1. Introduction

Missing data analyses have received much attention in statistics. Data are missing at random (Rubin, 1976) if the missingness depends on the observed but not on the missing values; whereas data are missing not at random if the missingness depends on both the observed and missing values. Under missingness not at random, the full data distribution is not identifiable without further assumptions. Researchers have considered parametric assumptions, such as pattern-mixture models (Little, 1993) and sample selection models (Heckman, 1979) and assumptions based on instrumental variables (Tang et al., 2003; D'Haultfoeuille, 2010; Wang et al., 2014) or shadow variables (Miao and Tchetgen Tchetgen, 2016; Kott and Liao, 2017).

Under a fully parametric model, identification of parameters can be achieved by sufficiently stringent modeling restrictions, which then invokes either the likelihood or Bayesian method; however, fully parametric approaches are sensitive to model misspecification. Researchers have also developed semiparametric methods for which one of the outcome model and the missing data mechanism model is parametric and the other is nonparametric. Among these, Tang et al. (2003) and Zhao and Shao (2015) proposed maximum pseudo likelihood estimators without modeling the non-response mechanism, and D'Haultfoeuille (2010) considered a regression analysis using a nonparametric nonresponse model. Scharfstein et al. (1999) and Shao and Wang (2016) proposed semiparametric nonresponse models and inverse probability weighted estimation. Qin et al. (2002), Kim and Yu (2011), Tang et al. (2014), and Zhao et al. (2017) used semiparametric or empirical likelihood approaches with parametric assumptions on the missing data mechanism.

* Corresponding author.
E-mail addresses: wli169@syr.edu (W. Li), syang24@ncsu.edu (S. Yang), peisong@umich.edu (P. Han).

https://doi.org/10.1016/j.jspi.2020.01.001
0378-3758/

In general, the missingness mechanism cannot be determined from the data alone, and inference under missingness not at random may be sensitive to unverifiable assumptions about the missingness mechanism. For identification, one common approach is to choose among the covariates a valid nonresponse instrument, a variable that is related to the outcome and can be excluded from the nonresponse model when the outcome and other covariates are included. The selection of a valid nonresponse instrument can be challenging. To address this issue, researchers have developed sensitivity analysis methods (e.g. Robins et al., 2000). Although widely used in practice, sensitivity analysis cannot provide point identification of parameter of interest.

In this article, we focus on estimation for general parameters defined through moment conditions and develop a robust estimation method with multiple working models for the response mechanisms. These response models can be based on different assumptions of the mechanism. For example, models for both missingness at random and missingness not at random with different identification conditions can be simultaneously considered. With multiple nonresponse models, we consider weights on the complete cases derived based on a set of calibration constraints. Construction of these calibration constraints is essential. Under missingness at random, calibration constraints imposed only on covariates are sufficient to eliminate selection bias after reweighting the complete cases (Han, 2014). However, this is not clear in that paper whether a similar approach can be applied under missingness not at random. This paper fills in this gap by proposing to construct calibration constraints directly on the score equations for the parameter of interest under multiple working models. Using such calibration weights, the proposed estimator is multiply robust, in the sense that, under a correct specification of the conditional distribution of the response given covariates, it is consistent if any one of the multiple response models is correctly specified. Such a robustness property renders the estimator more protection against misspecification of the response model. The multiple robustness has been studied in Han and Wang (2013), Chan and Yam (2014) and Han (2014), under missingness at random and in Han (2017), for the mean of outcome under missingness not at random. Our contribution is to develop multiple robust estimators for general parameters defined through moment conditions with multiple response models under missingness not at random.

The multiple-robustness property discussed in this article is different from that in some other articles. Consider a classical setting where likelihood function can be factorized into non-response and outcome mechanisms, each specified with a working model. The well-known "Double robustness" refers to the consistency property that allows misspecification of either one of the two models. Some articles generalize this classical concept in a likelihood model with multiple components. For instance, Molina et al. (2017) studied factorized likelihood models, where both nonresponse and outcome mechanisms can be further factorized into several components. A working model is specified for each of these components. Multiply robust estimators, according to their definition, are those that are consistent when some (not necessarily all) of these models are correct. In other words, their multiple robustness offers protection against misspecification of two or more than two components of the likelihood function. Wang and Tchetgen Tchetgen (2018) interpreted multiple robustness in a similar way. In contrast, we specify multiple models for the nonresponse mechanism. The multiple robustness discussed here refers to the consistency property that requires correct specification of only one of multiple models for the nonresponse mechanism.

The rest of this article is organized as follows. In Section 2, we introduce the setup, discuss assumptions on the response mechanism, and provide estimation methods. In Sections 3 and 4, we derive the proposed multiply robust estimator for general parameters under multiple response models and its consistency. In Section 5, we evaluate the finite sample performance of the proposed estimators via simulations. In Section 6, we apply the method to an application. We then end with a brief discussion in Section 7.

2. Setup, models and estimation

Let $\{(x_i, y_i, \delta_i) : i = 1, \ldots, n\}$ denote n realizations of (X, Y, δ), where X is a d-dimensional vector of covariates that is always observed, Y is an outcome variable that has missing values, and δ is the response indicator of Y, i.e., $\delta = 1$ if Y is observed and 0 if it is missing. Define the parameter of interest θ_0 through some moment condition $E\{U(\theta_0; X, Y)\} = 0$. For example, if $U(\theta; X, Y) = Y - \theta$, θ is the population mean of Y. If $U(\theta; X, Y) = \{Y - \mu(X; \theta)\}\partial\mu(X; \theta)/\partial\theta$, θ is the parameter that governs a model for the mean of Y given X.

We assume that the missingness of Y may depend on Y itself and make the following assumption for the response probability.

Assumption 1 (*Positivity*)**.** There exists a positive number C such that $P(\delta = 1 \mid X, Y) = \pi(X, Y) > C > 0$ almost surely.

Under missingness not at random, the response model may not be identifiable without further assumptions. If, for example, one can obtain a valid instrument – a variable that is correlated with the outcome and conditionally independent of the response indicator given other variables and the outcome – then identification of the parameters in the response model is possible (Wang et al., 2014). Suppose for a given model $\pi(X, Y; \alpha)$ for $P(\delta = 1 \mid X, Y)$, under certain identification conditions, let $S(\alpha; X, Y, \delta)$ be the score function of α:

$$S(\alpha; X, Y, \delta) = \frac{\delta - \pi(X, Y; \alpha)}{\pi(X, Y; \alpha)\{1 - \pi(X, Y; \alpha)\}} \frac{\partial \pi(X, Y; \alpha)}{\partial \alpha}. \tag{1}$$

The semiparametric efficient estimator $\hat{\alpha}$ can be obtained by solving

$$\sum_{i=1}^{n}\left\{\frac{\delta_i}{\pi(x_i,y_i;\alpha)}-1\right\}h(x_i;\alpha)=0 \tag{2}$$

where

$$h(X;\alpha)=\frac{E\{S(\alpha;X,Y,\delta=0)O(X,Y;\alpha)\mid X\}}{E\{O(X,Y;\alpha)\mid X\}} \tag{3}$$

with $O(X,Y;\alpha)=P(\delta=0\mid X,Y;\alpha)/P(\delta=1\mid X,Y;\alpha)=\pi(X,Y;\alpha)^{-1}-1$. See, for example, Robins and Rotnitzky (1997).

In what follows, we use $f_d(Y\mid X)$ to denote $f(Y\mid X,\delta=d)$, and $E_d(\cdot\mid X)=E(\cdot\mid X,\delta=d)$ for $d=0,1$, and $E_1(\cdot)=E(\cdot\mid\delta=1)$. For computation of (3), we note that

$$\begin{aligned}h(X;\alpha)&=\frac{E_1\left\{S(\alpha;X,Y,\delta=0)\pi(X,Y;\alpha)^{-1}O(X,Y;\alpha)\mid X\right\}}{E_1\left\{\pi(X,Y;\alpha)^{-1}O(X,Y;\alpha)\mid X\right\}}\\&=\frac{\int S(\alpha;X,y,\delta=0)\pi(X,y;\alpha)^{-1}O(X,y;\alpha)f_1(y\mid X)\mathrm{d}y}{\int\pi(X,y;\alpha)^{-1}O(X,y;\alpha)f_1(y\mid X)\mathrm{d}y}.\end{aligned} \tag{4}$$

To approximate the above integrals, we can substitute a nonparametric estimator $\hat{f}_1(Y\mid X)$ for $f_1(Y\mid X)$. The drawback of the nonparametric approach is that it is subject to the curse of dimensionality and often has a poor performance when the dimension of X is large. As an alternative, we consider a parametric working model $f_1(Y\mid X;\gamma)$ for $f_1(Y\mid X)$ and obtain an estimator $\hat{\gamma}$ from the respondents. Then, we can apply a numerical approximation technique to obtain the integrals in (4). If $f_1(y\mid x)$ is correctly specified, the solution to (2) is optimal for α when the true response mechanism follows $\pi(X,Y;\alpha)$. If the model $f_1(Y\mid X;\gamma)$ is incorrectly specified but $\pi(X,Y;\alpha)$ is correctly specified, the solution to (2) is still consistent, based on the fact that $E\left[\left\{\pi(X,Y;\alpha)^{-1}\delta-1\right\}h(X)\right]=0$ for any squared integrable function $h(X)$.

Once we obtain the parameter estimate $\hat{\alpha}$, the standard inverse probability weighting estimator $\hat{\theta}_{\mathrm{ipw}}$ of θ can be obtained by solving the weighted estimating equation

$$\sum_{i=1}^{n}\delta_i\hat{\omega}_iU(\theta;x_i,y_i)=0, \tag{5}$$

where $\hat{\omega}_i=\{\pi(x_i,y_i;\hat{\alpha})\}^{-1}$. It is well known that $\hat{\theta}_{\mathrm{ipw}}$ has a large variance if the estimated probability is close to zero and its consistency relies on the correct specification of the response probability.

3. Multiple robust estimation

The method to be discussed allows for multiple specifications of the response probability. As mentioned, for identification under missing not at random, one common approach would be to choose among the set of covariates a valid nonresponse instrument X_2, which satisfies (i) $f(Y\mid X_1,X_2=a)\neq f(Y\mid X_1,X_2=b)$ for any $a\neq b$; and (ii) $P(\delta=1\mid X,Y)=P(\delta=1\mid X_1,Y)$. The selection of a valid nonresponse instrument can be challenging. In this section, we develop a multiply robust estimation method that can accommodate multiple models for the response mechanism. The advantage is that in each model, we can make different assumptions, such as missingness at random and missingness not at random with different identification conditions, such as with different nonresponse instruments

Let the set of K specifications of the response probability be $\{\pi^k(X,Y;\alpha^k):k=1,\ldots,K\}$, where $\pi^k(X,Y;\alpha^k)$ is the kth model for $\pi(X,Y)$, known up to the parameter α^k, $k=1,\ldots,K$. For each model, α^k can be estimated as the solution to (2) with the corresponding score function under model $\pi^k(X,Y;\alpha^k)$, denoted as $\hat{\alpha}^k$. From the standard Z-estimation theory (e.g., van der Vaart, 2000), these estimators are well-defined and converge to some values in probability under regularity conditions.

For convenience, we assume $\delta_1=\cdots=\delta_m=1$ and $\delta_{m+1}=\cdots=\delta_n=0$. Let $\omega(X,Y)=\pi(X,Y)^{-1}$ and $\omega_i=\omega(x_i,y_i)$. We then have the following lemma.

Lemma 1. *For any $g(X,Y)$, $E_1\left(\omega(X,Y)\delta[g(X,Y)-E\{g(X,Y)\}]\right)=0$.*

Under missingness at random, Han (2014) defined positive weights ω_i, $i=1,\ldots,m$, by maximizing $L(\omega_1,\ldots,\omega_m)=\prod_{i=1}^{m}\omega_i$ subject to

$$\sum_{i=1}^{m}\omega_i=1,\ \sum_{i=1}^{m}\omega_i\{g^k(x_i)-n^{-1}\sum_{i=1}^{n}g^k(x_i)\}=0\ (k=1,\ldots,K), \tag{6}$$

where (6) is the sample version of the moment equality in Lemma 1 with $g(X,Y)$ being $g^k(X)$ and $n^{-1}\sum_{i=1}^{n}g^k(x_i)$ estimating $E\{g^k(X)\}$. Let $g^k(X)$ be the kth estimated response probability $\pi^k(X;\hat{\alpha}^k)$, Han (2014) showed that with (6),

the weighting estimator of θ, solving (5) with $\hat{\omega}_i$ being the resulting weight from above procedure is multiply robust in the sense that if any one of models $\{\pi^k(X;\alpha^k): k=1,\ldots,K\}$ is correctly specified, the weighting estimator is consistent for the true value θ_0. Chen and Haziza (2017) considered a similar problem under missingness at random.

Under missingness not at random, imposing constraints implied by Lemma 1 is not feasible, because $g(x_i,y_i)$ is not available for individuals with $\delta_i=0$, and therefore $E\{g(X,Y)\}$ cannot be simply estimated by the sample average of $g(x_i,y_i)$. To overcome this difficulty, note that

$$E\{g(X,Y)\}=E\left[\delta g(X,Y)+(1-\delta)E_0\{g(X,Y)\mid X\}\right]. \tag{7}$$

To compute $E_0\{g(X,Y)\mid X\}$, by the Bayes rule, we can derive

$$f_0(Y\mid X)=f_1(Y\mid X)\frac{O(X,Y)}{E_1\{O(X,Y)\mid X\}}. \tag{8}$$

and therefore,

$$E_0\{g(X,Y)\mid X\}=\frac{\int g(X,y)O(X,y)f_1(y\mid X)\mathrm{d}y}{\int O(X,y)f_1(y\mid X)\mathrm{d}y}=\frac{E_1\{g(X,Y)O(X,Y)\mid X\}}{E_1\{O(X,Y)\mid X\}}.$$

Because $f(Y\mid X)$ is often of primary scientific interest and is usually assumed to be correctly modeled based on subject matter knowledge or existing literature, we denote $f(Y\mid X,\beta)$ as this correct model. Then, based on the kth model of $\pi(X,Y)$, we can specify $f_1^k(Y\mid X;\gamma^k)$ through

$$f_1^k(Y\mid X;\gamma^k)=\frac{f(Y\mid X;\beta)\pi^k(X,Y;\alpha^k)}{\int f(y\mid X;\beta)\pi^k(X,y;\alpha^k)\mathrm{d}y},$$

where $\gamma^k=(\alpha^k,\beta)$. Let $\hat{\gamma}^k=(\hat{\alpha}^k,\hat{\beta}^k)$, where $\hat{\beta}^k$ can be obtained using a weighted analysis among the respondents with weights $\pi^k(x_i,y_i;\hat{\alpha}^k)^{-1}$. Then, $E_0\{g(X,Y)\mid X\}$ can be estimated by

$$\hat{E}_0^k\{g(X,Y)\mid X;\hat{\gamma}^k\}=\frac{\int g(X,y)O^k(X,y;\hat{\alpha}^k)\hat{f}_1^k(y\mid X;\hat{\gamma}^k)\mathrm{d}y}{\int O^k(X,y;\hat{\alpha}^k)\hat{f}_1^k(y\mid X;\hat{\gamma}^k)\mathrm{d}y}, \tag{9}$$

where $O^k(X,Y;\alpha^k)=\pi^k(X,Y;\alpha^k)^{-1}-1$. Therefore, (7) can be estimated by

$$\hat{E}^k\{g(X,Y);\hat{\gamma}^k\}=\frac{1}{n}\sum_{i=1}^{n}\left[\delta_i g(x_i,y_i)+(1-\delta_i)\hat{E}_0^k\{g(X,Y)\mid X=x_i;\hat{\gamma}^k\}\right]. \tag{10}$$

Let $g(X,Y)=U(\theta;X,Y)$. We then propose the set of positive weights $\hat{\omega}_i$, $i=1,\ldots,m$, which maximizes $L(\omega_1,\ldots,\omega_m)$ subject to the constraints

$$\sum_{i=1}^{m}\omega_i=1,\ \sum_{i=1}^{m}\omega_i U(\hat{\theta}^k;X_i,Y_i)=0\ (k=1,\ldots,K), \tag{11}$$

where $\hat{\theta}^k$ solves $\bar{U}^k(\theta)=n^{-1}\sum_{i=1}^{n}[\delta_i U(\theta;x_i,y_i)+(1-\delta_i)\hat{E}_0^k\{U(\theta;X,Y)\mid X=x_i;\hat{\gamma}^k\}]=0$, i.e., the estimating equation approach applied to the imputed dataset with the missing outcomes filled in by $\{\hat{E}_0^k(Y\mid X=x_i;\hat{\gamma}^k):\delta_i=0\}$.

Once we obtain the set of weights, the proposed estimator $\hat{\theta}_{\mathrm{MR}}$ of θ is obtained by solving

$$\sum_{i=1}^{m}\hat{\omega}_i U(\theta;x_i,y_i)=0. \tag{12}$$

In summary, the proposed estimation proceeds as follows:

Step 1. Posit a set of multiple parametric working models for $\pi(X,Y)$ and $f(Y\mid X)$, $\mathcal{P}=\{\pi^k(X,Y;\alpha^k),f(Y\mid X;\beta):k=1,\ldots,K\}$. For the kth model, estimate α^k by $\hat{\alpha}^k$ which solves $\sum_{i=1}^{n}H^k(\alpha^k;x_i,y_i,\delta_i)=0$ for α^k, where

$$H^k(\alpha^k;X,Y,\delta)=\left\{\frac{\delta}{\pi^k(X,Y;\alpha^k)}-1\right\}\times\frac{\hat{E}_1^k\left\{S(\alpha^k;X,Y,\delta=0)\pi^k(X,Y;\alpha^k)^{-1}O^k(X,Y;\alpha^k)\mid X\right\}}{\hat{E}_1^k\left\{\pi^k(X,Y;\alpha^k)^{-1}O^k(X,Y;\alpha^k)\mid X\right\}},$$

and $O^k(X,Y;\alpha^k)=\pi^k(X,Y;\alpha^k)^{-1}-1$ and $\hat{E}_1^k(\cdot\mid X,\delta=1)$ is taken with respect to the working model $\hat{f}_1^k(Y\mid X)$. Also, the coefficient β is estimated by $\hat{\beta}^k$ which is obtained by applying a weighted analysis among the respondents with weight $\pi^k(x_i,y_i;\hat{\alpha}^k)^{-1}$.

Step 2. Let $\bar{U}^k(\theta)=\hat{E}^k\{U(\theta;X,Y);\hat{\gamma}^k\}$ with $\hat{E}^k(\cdot)$ defined in (10). Let $\hat{\theta}^k$ solve $\bar{U}^k(\theta)=0$. Then, impose the constraints (11).

Step 3. Obtain the set of weights $\{\hat{\omega}_i : i = 1, \ldots, m\}$ by maximizing $L(\omega_1, \ldots, \omega_m)$ subject to the constraints in (11). Denote $\hat{\alpha} = (\hat{\alpha}^1, \ldots, \hat{\alpha}^K)^{\mathrm{T}}$, and $\hat{c}(X, Y) = [U(\hat{\theta}^1; X, Y) - \bar{U}^1(\hat{\theta}^1), \ldots, U(\hat{\theta}^K; X, Y) - \bar{U}^K(\hat{\theta}^K)]^{\mathrm{T}}$, where $\bar{U}^k(\hat{\theta}^k) = 0$ for all k by the way how $\hat{\theta}^k$ is obtained. By the Lagrange multipliers technique, we have $\hat{\omega}_i = m^{-1}\{1 + \hat{\rho}^{\mathrm{T}}\hat{c}(x_i, y_i)\}^{-1}$ $(i = 1, \ldots, m)$, where $\hat{\rho} = (\hat{\rho}_1, \ldots, \hat{\rho}_K)^{\mathrm{T}}$ satisfies

$$\sum_{i=1}^{m}\{1 + \hat{\rho}^{\mathrm{T}}\hat{c}(x_i, y_i)\}^{-1}\hat{c}(x_i, y_i) = 0.$$

Step 4. The proposed estimator $\hat{\theta}_{\mathrm{MR}}$ of θ is obtained by solving (12).

Remark 3.1. Because $\hat{\omega}_i$ should be non-negative, $\hat{\rho}$ must satisfy $1 + \hat{\rho}^{\mathrm{T}}\hat{c}(x_i, y_i) > 0$ for $i = 1, \ldots, m$. To ensure the non-negativity of $\hat{\omega}_i$, the modified Newton–Raphson algorithm proposed by Chen et al. (2002) to obtain range restricted weights can be applied.

Remark 3.2. There are some other proposals for $\hat{\theta}^k$. For example, one can obtain $\hat{\theta}^k$ using a weighted regression analysis among the respondents with weights $\pi^k(x_i, y_i; \hat{\alpha}^k)^{-1}$. In these cases, $\bar{U}^k(\hat{\theta}^k)$ in Step 3 may not necessarily be zero. These variations would not affect the multiple robustness property of the proposed estimator.

4. Main result

In this section, we show the consistency of the proposed estimator $\hat{\theta}_{\mathrm{MR}}$ and its multiple robustness. We first state the following regularity conditions. For each $k = 1, \ldots, K$,

(C1) the parameter space $\mathcal{A}^k$, $\mathcal{B}^k$ and Θ^k for α^k, β^k and θ^k are compact;
(C2) $\pi^k(X, Y; \alpha^k)$ and $h(X; \alpha^k)$ are continuous in α^k; $U(\theta^k; X, Y)$ is continuous in θ^k; $E_0^k\{U(\theta; X, Y) \mid X; \alpha^k, \beta^k\}$ is continuous in α^k and β^k;
(C3) the function

$$q(X, Y; \alpha^k, \beta^k, \theta^k) = \begin{bmatrix} \{\delta/\pi^k(X, Y; \alpha^k) - 1\}h(X; \alpha^k) \\ \delta/\pi^k(X, Y; \alpha^k)U(\theta^k; X, Y) \\ \delta U(\theta; X, Y) + (1 - \delta)E_0^k\{U(\theta^k; X, Y) \mid X; \alpha^k, \beta^k\} \end{bmatrix}$$

satisfies that $E\left\{\sup_{(\alpha^k, \beta^k, \theta^k) \in \mathcal{A}^k \times \mathcal{B}^k \times \Theta^k} \|q(X, Y; \alpha^k, \beta^k, \theta^k)\|\right\} < \infty$;
(C4) $E\{q(X, Y; \alpha^k, \beta^k, \theta^k)\} = 0$ has a unique solution in $\mathcal{A}^k \times \mathcal{B}^k \times \Theta^k$;
(C5) $E[\sup_{\rho \in \wp} \log\{1 + \rho^T c(X, Y)\} \mid \delta = 1] < \infty$ where $\wp$ is the parameter space for ρ and is compact;
(C6) $E[\sup_{\theta^k \in \Theta^k, \rho \in \wp}\{U(\theta_0; X, Y) - U(\theta^k; X, Y)\}/\{1 + \rho^T c(X, Y)\} \mid \delta = 1] < \infty$.

Conditions (C1)–(C5) are standard assumptions for the consistency of the estimators $\hat{\alpha}^k$ and $\hat{\beta}^k$ and $\hat{\theta}^k$; see for instance Newey and McFadden (1994). Conditions (C5)–(C6) are needed for the law of large number in the proof; see for instance Han (2017). We state the main result in the following theorem.

Theorem 4.1 (*Multiple Robustness*). *Under Assumption 1 and regularity conditions (C1)–(C6), if any one working model is correctly specified for the response probability* $\pi(X, Y)$, $\hat{\theta}_{\mathrm{MR}} \to \theta_0$ *in probability as* $n \to \infty$.

Proof of Theorem 4.1. Assume that $\pi^1(X, Y; \alpha^1)$ and $f(Y \mid X; \beta)$ are correctly specified models for $\pi(X, Y)$ and $f(Y \mid X)$ with true parameter values α_0^1 and β_0; i.e., $\pi^1(X, Y; \alpha_0^1) = \pi(X, Y)$ and $f(Y \mid X; \beta_0) = f(Y \mid X)$. Let $\hat{\alpha}^1$ be the solution to (2) under $\pi^1(X, Y; \alpha^1)$ and $\hat{\beta}^1$ be a weighted estimator of β^1 using the respondents with weight $\pi^1(x_i, y_i; \hat{\alpha}^1)^{-1}$. Then, under above regularity conditions, $\hat{\alpha}^1$ and $\hat{\beta}^1$ are root-n consistent for α_0^1 and β_0.

To show $\hat{\theta}_{\mathrm{MR}}$ is consistent for θ_0, it suffices to show that $\sum_{i=1}^m \hat{\omega}_i U(\theta_0; x_i, y_i) \to 0$ in probability. Firstly, because $\pi^1(X, Y; \alpha^1)$ is correctly specified, we have $\hat{\theta}^1 \to \theta_0$ in probability. By constraint (11) with $k = 1$, we have

$$\begin{aligned}\sum_{i=1}^{m}\hat{\omega}_i U(\theta_0; x_i, y_i) &= \sum_{i=1}^{m}\hat{\omega}_i\left\{U(\theta_0; x_i, y_i) - U(\hat{\theta}^1; x_i, y_i)\right\} + \bar{U}^1(\hat{\theta}^1) \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{U(\theta_0; x_i, y_i) - U(\hat{\theta}^1; x_i, y_i)}{1 + \hat{\rho}^{\mathrm{T}}\hat{c}(x_i, y_i)} \\ &\to E\left\{\frac{U(\theta_0; X, Y) - U(\theta_0; X, Y)}{1 + \bar{\rho}^{\mathrm{T}}\bar{c}(X, Y)}\Big|\delta = 1\right\} = 0,\end{aligned}$$

Table 1
Bias, standard deviation, estimated standard error based on 100 bootstrap replicates, and coverage of 95% confidence intervals over 2000 simulated datasets. All numbers are on the scale 10^{-2}.

	Bias	s.d.	s.e.	Coverage	Bias	s.d.	s.e.	Coverage
	$\beta_1 = .5$				$\beta_2 = 1$			
IPW10	0.56	9.19	11.3	96.8	−0.36	13.2	13.6	94.0
IPW01	13.2	7.19	7.03	53.5	−10.3	11.9	11.0	81.4
MR10	0.29	9.16	10.8	97.5	0.23	11.8	12.3	94.3
MR01	13.0	7.04	6.9	53.3	−9.94	10.9	10.4	82.1
MR11	0.29	9.16	10.8	97.5	0.23	11.8	12.3	94.3
	$\beta_3 = 1$				$\beta_4 = 1$			
IPW10	0.33	11.3	11.5	94.2	0.11	11.6	11.5	94.6
IPW01	−3.53	10.4	10.0	92.0	−3.57	10.7	10.0	91.2
MR10	0.20	10.0	10.0	94.2	0.24	10.2	10.0	94.6
MR01	−3.38	10.1	9.9	92.0	−3.58	10.4	9.85	91.4
MR11	0.20	10.0	10.0	94.2	.024	10.2	10.0	94.6

where $\bar{\rho}$ is the probability limit of $\hat{\rho}$, and

$$\bar{c}(X, Y) = \begin{pmatrix} U(\bar{\theta}^1; X, Y) \\ \vdots \\ U(\bar{\theta}^K; X, Y) \end{pmatrix},$$

and $\bar{\theta}^k$ is the probability limit of $\hat{\theta}^k$, for $k = 1, \ldots, K$. This completes the proof.

Remark 4.1. From Theorem 4.1, consistency of the proposed estimator is guaranteed if one response probability model is correctly specified. Both the number of the posited models and their functional forms can affect the efficiency of the proposed estimator in a very complex way. In addition, with a finite sample size, the numerical performance can be unstable if there are a large number of working models. In particular, Step 3 may not have a solution when some of the models are poorly constructed. In this case, although some adjustments can be made to ensure existence of a solution (e.g., Chen et al., 2008; Emerson and Owen, 2009; Tsao and Wu, 2013), the implementation of these adjustments is complicated. To reduce the chance of running into numerical issues, we suggest positing a few well-constructed working models instead of a large number of poorly built ones.

The proposed estimator can be viewed as a Z-estimator (van der Vaart, 2000), solving a set of estimating equations, and its confidence interval can be constructed via the bootstrap.

5. A simulation study

In this section, we evaluate the finite-sample performance of the proposed estimator for its robustness compared to existing methods.

We generate samples of size $n = 300$. The covariates X_1, X_2 and X_3 are identically and independently generated from Normal(0, 0.5). The outcome variable is $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \epsilon$ with $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 1, 1, 1)$ and $\epsilon \sim$ Normal(0, 1). The missing indicator δ, is generated from Bernoulli(π), where $\text{logit}(\pi) = 1 + 0.5y + x_1$, and therefore the missingness is missing not at random. The average response rate is about 72%. The parameters of interest are the regression coefficients β_1,..., β_4.

We consider two response models. The correct model for $\pi(x, y)$ is specified as $\text{logit}\{\pi^1(x, y; \alpha^1)\} = \alpha_1^1 + \alpha_2^1 y + \alpha_3^1 x_1$ and the incorrect model is specified as $\text{logit}\{\pi^2(x, y; \alpha^2)\} = \alpha_1^2 + \alpha_2^2 x_1 + \alpha_3^2 x_2 + \alpha_4^2 x_3$. Each estimator is assigned a name with the form "method-00," where each digit of the two-digit number, from left to right, indicates if $\pi^1(x, y; \alpha^1)$ or $\pi^2(x, y; \alpha^2)$ is used in the construction with "1" meaning yes and "0" no, respectively. For example, "IPW10" is the inverse probability weighting estimator with the response model $\pi^1(x, y; \alpha^1)$ and "MR11" is the proposed estimator based on the response models $\pi^1(x, y; \alpha^1)$ and $\pi^2(x, y; \alpha^2)$.

Table 1 and Fig. 1 contain the comparison of different estimators. From these results, one can notice that when $\pi(x, y)$ is correctly modeled, IPW10 and MR10 have small biases, and MR10 improves the efficiency over IPW10 for all parameters. When $\pi(x, y)$ is incorrectly modeled, IPW01 and MR01 have large biases. The proposed estimator has small biases for all parameters when both response models are used in calibration. These observations confirm our theoretical results and shows that the proposed estimator has improved robustness over the existing weighting estimator.

6. Application

In this section, we apply our method to an application. In the Stroke Recovery in Underserved Populations 2005–2006 (U.S.) study, several metrics for stroke patients' emotional and physiological wellness were measured at four

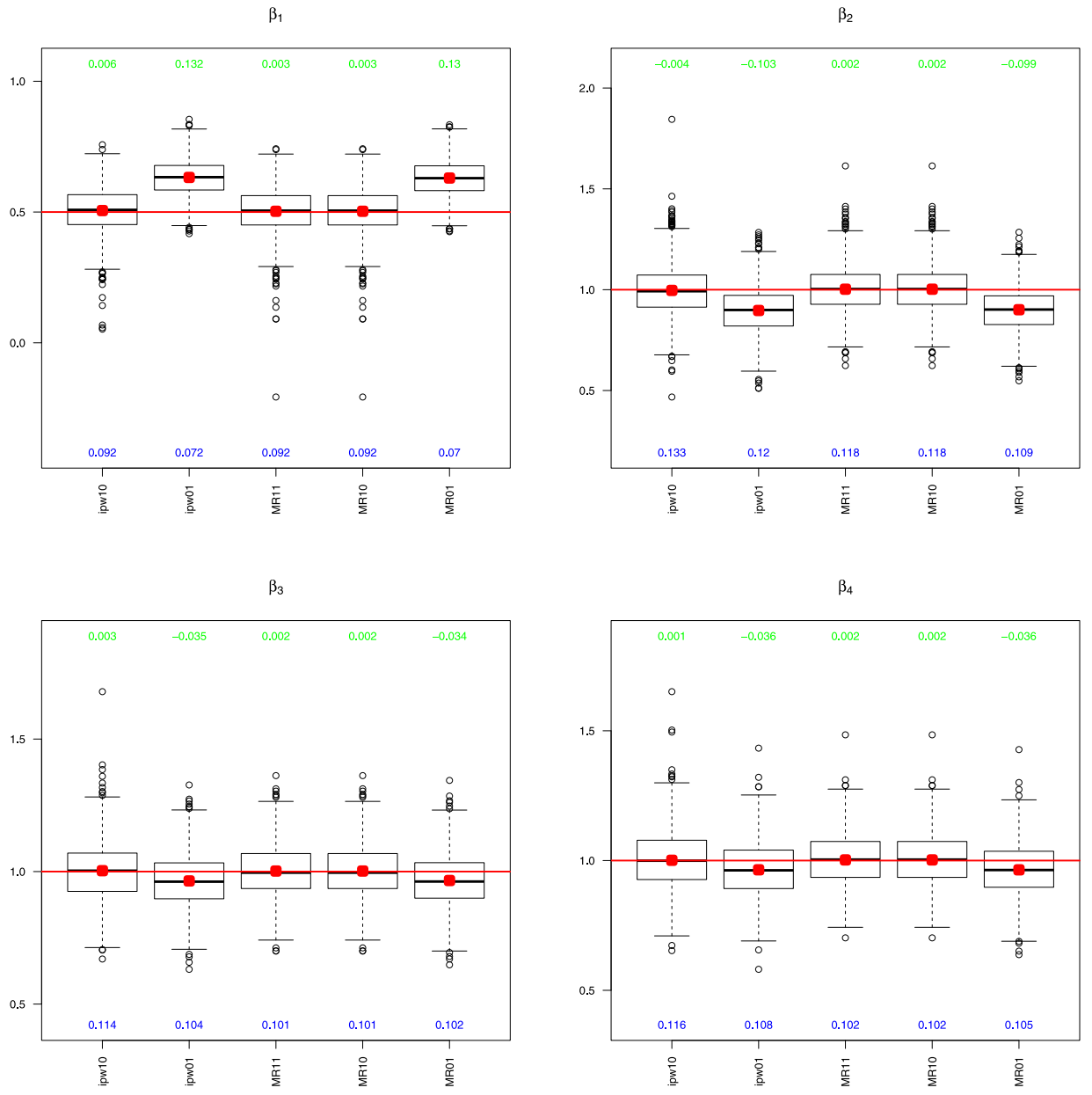

Fig. 1. Simulation results for $\beta_1, \beta_2, \beta_3, \beta_4$: green numbers are biases, blue numbers are standard deviations, and red dots are the Monte Carlo averages of the estimates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

time points at admission and discharge from rehabilitation facility, and 3 months and 12 months after discharge. Patients were also followed up with questions regarding their health maintenance type — own care, unpaid person (or family), or paid professional. Simple summary statistics show that patients with own care tend to have better emotional health condition in the 12 month follow-up than those with health aids (paid or unpaid). It would be interesting to see whether such effects remain significant if other variables such as emotional wellness and functional recovery status at discharge are controlled for.

In our analysis, the variable Y is the depression scale (CESD) at 12 month follow-up. A high score of Y indicates high symptoms of depression. Covariates of interest are base-line depression scale at discharge, and physiological wellness — Functional Independence Measure (FIM) at discharge, and whether a patient self cares. Many demographic variables do not indicate statistical significance. We include only age in our study. Specifically, we let X_1 be the age of patients, X_2 is the depression scale at discharge, X_3 is the FIM score at discharge and X_4 the binary variable indicating whether a patient self-cares or not. The dataset we study have Y missing for 172 cases out of 1005 total cases. Two response

Table 2
Dependent variable: CESD at 12 month follow up. Standard errors are computed based on 100 bootstrap samples.

Variable	IPW10	IPW01	MR11	MR10	MR01
Intercept	8.905	8.368	8.680	8.704	8.130
	(2.49)	(2.21)	(1.89)	(1.88)	(2.22)
Age	−0.035	−0.032	−0.033	−0.033	−0.031
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
CESD-base	0.361	0.332	0.318	0.319	0.333
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
FIM-base	0.010	0.011	0.013	0.012	0.013
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
Self care	−2.949	−2.647	−2.786	−2.679	−2.681
	(0.72)	(0.64)	(0.73)	(0.60)	(0.65)

models are considered: the first one being $\text{logit}\{\pi^1(x, y; \alpha^1)\} = \alpha_1^1 + \alpha_2^1 y + \alpha_3^1 x_3$ and second one $\text{logit}\{\pi^1(x, y; \alpha^1)\} = \alpha_1^1 + \alpha_2^1 x_1 + \alpha_3^1 x_2 + \alpha_3^1 x_3$.

Results are given in Table 2. One can see that all estimators for coefficients for X_2 and X_4 are statistically significant. For comparison, a complete case analysis yields estimates 0.332 and −2.657 for coefficients for X_2 and X_4 respectively. They are close to estimates that are based on MAR assumption. On the other hand, our MR11 estimate is closer to MR10 than to MR01. In particular, for MR11 estimation, the effect of self-care on the depression level y is larger and the effect of base line depression level on y is smaller. In all cases, after base-line emotional and functional health conditions are controlled for, opting for self care tends to improve a patient's emotional well-being.

7. Discussion

We have developed multiple robust estimators for parameters defined by moment conditions that allow multiple response models in the presence of missing not at random data. The improved robustness comes from multiple model specifications for the nonresponse mechanism. Our proposed method provides multiple protections to model misspecification and therefore is an attractive alternative to existing inverse probability weighting estimators.

Missing data often arise in survey sampling. In complex surveys, the challenge is to take design information or design weights into account when developing propensity score methods for handling missing data. The proposed weighting method can be combined with sampling weights for an integrated solution to handle missing not at random data and sampling designs. This extension will be pursued in a separate paper.

Acknowledgment

Dr. Yang is partially supported by NSF DMS 1811245 and NCI P01 CA142538.

References

Chan, K.C.G., Yam, S.C.P., 2014. Oracle, multiple robust and multipurpose calibration in a missing response problem. Statist. Sci. 29, 380–396.
Chen, S., Haziza, D., 2017. Multiply robust imputation procedures for the treatment of item nonresponse in surveys. Biometrika 104, 439–453.
Chen, J., Sitter, R., Wu, C., 2002. Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. Biometrika 89, 230–237.
Chen, J., Variyath, A.M., Abraham, B., 2008. Adjusted empirical likelihood and its properties. J. Comput. Graph. Statist. 17, 426–443.
D'Haultfoeuille, X., 2010. A new instrumental method for dealing with endogenous selection. J. Econometrics 154, 1–15.
Emerson, S.C., Owen, A.B., 2009. Calibration of the empirical likelihood method for a vector mean. Electron. J. Stat. 3, 1161–1192.
Han, P., 2014. Multiply robust estimation in regression analysis with missing data. J. Amer. Statist. Assoc. 109 (507), 1159–1173.
Han, P., 2017. Calibration and multiple robustness when data are missing not at random. Statist. Sinica accepted.
Han, P., Wang, L., 2013. Estimation with missing data: beyond double robustness. Biometrika 100, 417–430.
Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.
Kim, J.K., Yu, C.L., 2011. A semiparametric estimation of mean functionals with nonignorable missing data. J. Amer. Statist. Assoc. 106, 157–165.
Kott, P.S., Liao, D., 2017. Calibration weighting for nonresponse that is not missing at random: allowing more calibration than response-model variables. J. Surv. Stat. Methodol. 5, 159–174.
Little, R.J., 1993. Pattern-mixture models for multivariate incomplete data. J. Amer. Statist. Assoc. 88, 125–134.
Miao, W., Tchetgen Tchetgen, E.J., 2016. On varieties of doubly robust estimators under missingness not at random with a shadow variable. Biometrika 2, 475–482.
Molina, J., Rotnitzky, A., Sued, M., Robins, J., 2017. Multiple robustness in factorized likelihood models. Biometrika 104, 561–581.
Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. Handb. Econom. 4, 2111–2245.
Qin, J., Leung, D., Shao, J., 2002. Estimation with survey data under nonignorable nonresponse or informative sampling. J. Amer. Statist. Assoc. 97, 193–200.
Robins, J., Rotnitzky, A., 1997. Analysis of semi-parametric regression models with non-ignorable non-response. Stat. Med. 16, 81–102.
Robins, J.M., Rotnitzky, A., Scharfstein, D.O., 2000. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Statistical Models in Epidemiology, the Environment, and Clinical Trials. Springer, New York, pp. 1–94.
Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

Scharfstein, D.O., Rotnitzky, A., Robins, J.M., 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J. Amer. Statist. Assoc. 94, 1096–1120.
Shao, J., Wang, L., 2016. Semiparametric inverse propensity weighting for nonignorable missing data. Biometrika 103, 175–187.
Tang, G., Little, R.J., Raghunathan, T.E., 2003. Analysis of multivariate missing data with nonignorable nonresponse. Biometrika 90, 747–764.
Tang, N., Zhao, P., Zhu, H., 2014. Empirical likelihood for estimating equations with nonignorably missing data. Statist. Sinica 24, 723–747.
Tsao, M., Wu, F., 2013. Empirical likelihood on the full parameter space. Ann. Statist. 41 (4), 2176–2196.
van der Vaart, 2000. Asymptotic Statistics, Vol. 3. Cambridge university press, Cambridge: Cambridge University Press.
Wang, S., Shao, J., Kim, J.K., 2014. An instrumental variable approach for identification and estimation with nonignorable nonresponse. Statist. Sinica 24, 1097–1116.
Wang, L., Tchetgen Tchetgen, E., 2018. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. J. R. Stat. Soc. Ser. B Stat. Methodol..
Zhao, J., Shao, J., 2015. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. J. Amer. Statist. Assoc. 110, 1577–1590.
Zhao, P., Tang, N., Qu, A., Jiang, D., 2017. Semiparametric estimating equations inference with nonignorable missing data. Statist. Sinica 27 (1), 89–113.