Statistica Sinica **28** (2018), 1725-1740
doi:https://doi.org/10.5705/ss.202015.0408

CALIBRATION AND MULTIPLE ROBUSTNESS WHEN DATA ARE MISSING NOT AT RANDOM

Peisong Han

University of Michigan

Abstract: In missing data analysis, multiple robustness is a desirable property resulting from the calibration technique. A multiply robust estimator is consistent if any one of the multiple data distribution models and missingness mechanism models is correctly specified. So far in the literature, multiple robustness has only been established when data are missing at random (MAR). We study how to carry out calibration to construct a multiply robust estimator when data are missing not at random (MNAR). With multiple models available, where each model consists of two components, one for data distribution for complete cases and one for missingness mechanism, our proposed estimator is consistent if any one pair of models are correctly specified.

Key words and phrases: Calibration, empirical likelihood, missing not at random (MNAR), multiple robustness, nonignorable nonresponse.

1. Introduction

Missing data problems are commonly seen in practice. Depending on the nature of missingness, there are three mechanisms that are widely adopted in the literature: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin (1976)). For MCAR, the missingness depends on neither the observed nor the missing values; for MAR, the missingness depends on the observed but not on the missing values; and for MNAR, the missingness depends on both the observed and the missing values. As the missingness mechanism becomes more complex, statistical analysis becomes more difficult. For MCAR, a complete-case analysis ignoring subjects with missing data leads to consistent estimation, and usually gives the best solution in terms of efficiency. Extensive research has been done in the case of MAR, yielding a rich collection of effective methods and interesting results. Much less has been done for MNAR, largely due to the unknown dependence of the missingness on unobserved values, although in many observational studies MNAR is the most likely mechanism.

Calibration is a method originally developed in the sampling survey literature (Deville and Särndal (1992)), where it was used to calibrate the sampling weight so that the weighted average of some auxiliary variables based on the sampled subjects is equal to the known population average. In the survey context calibration has been studied a lot (e.g., Chen and Sitter (1999); Lundström and Särndal (1999); Wu and Sitter (2001); Wu (2003); Chang and Kott (2008); Kim (2009, 2010); Kim and Park (2010); Tan and Wu (2015)). The application of calibration to missing data analysis has attracted considerable research interests recently and has produced many interesting results (e.g., Tan (2006, 2010); Qin and Zhang (2007); Chen, Leung and Qin (2008); Qin, Shao and Zhang (2008); Han and Wang (2013); Chan and Yam (2014); Han (2014, 2016a,b)). In particular, the estimators in Han and Wang (2013), Chan and Yam (2014) and Han (2014, 2016a,b) are multiply robust, in that they are consistent if any one of the multiple missingness mechanism models and/or data distribution models is correctly specified. Such a robustness property is a significant improvement over the well-known double robustness (e.g. Scharfstein, Rotnitzky and Robins (1999); Bang and Robins (2005); Tsiatis (2006)).

So far multiple robustness has been established and studied only when data are MAR. In this paper, for the estimation of the mean of a response variable that is MNAR, we show how to carry out calibration so that multiple robustness can be achieved. Here each model consists of two components, one for the data distribution for complete cases and one for the missingness mechanism. The two components together characterize the whole data distribution. When multiple models are available, our proposed estimator is consistent if any one is correctly specified. Estimating the mean is a common problem in both sampling survey and causal inference, and thus our proposed method is of practical importance.

This paper is organized as follows. Section 2 introduces the notation and gives a review of calibration under MAR. Section 3 covers calibration under MNAR and establishes the multiple robustness property of our proposed estimator. Section 4 contains some simulation results. Some discussion is given in Section 5.

2. Notation and Review of Calibration Under MAR

Let Y denote the response of interest, $\boldsymbol{X}$ a vector of auxiliary variables that are always observed, and R the indicator of observing Y (i.e., $R = 1$ if Y is observed and $R = 0$ if Y is missing). The quantity of interest is $\mu_0 = E(Y)$.

MCAR means that the selection probability $P(R=1|Y,\boldsymbol{X})$ is a constant, MAR means that $P(R=1|Y,\boldsymbol{X})=P(R=1|\boldsymbol{X})$ only depends on the fully observed $\boldsymbol{X}$, and MNAR means that $P(R=1|Y,\boldsymbol{X})$ depends on both Y and $\boldsymbol{X}$. We use $\pi(\boldsymbol{X})$ and $\pi(Y,\boldsymbol{X})$ to denote $P(R=1|Y,\boldsymbol{X})$ under MAR and MNAR, respectively. The observed data are n independent and identically distributed copies of $(R,RY,\boldsymbol{X})$. Let $m=\sum_{i=1}^{n}R_i$ be the number of complete cases. Without loss of generality, assume that these complete cases are $i=1,\ldots,m$.

The original calibration estimator in Deville and Särndal (1992) has a weighting structure $\sum_{i=1}^{m}\hat{w}_iY_i$, with the weight $\hat{w}_i$ derived through

$$\min_{w_1,\ldots,w_m}\sum_{i=1}^{m}\pi(\boldsymbol{X}_i)\{nw_i-\pi(\boldsymbol{X}_i)^{-1}\}^2 \text{ subject to } \sum_{i=1}^{m}w_i\boldsymbol{X}_i=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i.$$

The calibration constraint above makes the weighted average of $\boldsymbol{X}$ based on complete cases equal to the unweighted average of $\boldsymbol{X}$ based on the whole sample, which consistently estimates the population mean of $\boldsymbol{X}$. The weight $\hat{w}_i$ is derived by minimizing the above discrepancy between w_i and the inverse probability weight $1/\{n\pi(\boldsymbol{X}_i)\}$ subject to the calibration constraint. Many variations of calibration have been proposed in the sampling survey literature, some with different optimization criteria (e.g. Chen and Sitter (1999); Kim (2009, 2010); Tan and Wu (2015)) and some with calibration variables being certain functions of $\boldsymbol{X}$ (e.g. Wu and Sitter (2001)). Most of these variations impose two additional constraints: $w_i>0$ and $\sum_{i=1}^{m}w_i=1$.

Calibration in missing data literature has two major variations. The first one derives $\hat{w}_i$ through

$$\max_{w_1,\ldots,w_m}\prod_{i=1}^{m}w_i \quad \text{subject to} \quad w_i>0,\ \sum_{i=1}^{m}w_i=1,\ \sum_{i=1}^{m}w_i\boldsymbol{h}(\boldsymbol{X}_i)=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{h}(\boldsymbol{X}_i), \tag{2.1}$$

where $\boldsymbol{h}(\boldsymbol{X})$ comprises user-specified functions of $\boldsymbol{X}$ (e.g. Qin and Zhang (2007); Qin, Shao and Zhang (2008); Han and Wang (2013); Chan and Yam (2014); Han (2014, 2016a,b)). Due to the positivity and sum-to-one constraints, w_i can be viewed as an empirical likelihood (EL) on complete cases. The formulation in (2.1) is the same as that of an EL problem (Owen (1988, 2001)); Qin and Lawless (1994)). The second major variation considers an additional EL on incomplete cases (Chen, Leung and Qin (2008); Tan (2010)):

$$\max_{w_i's,v_j's}\prod_{i=1}^{m}w_i\prod_{j=m+1}^{n}v_j \qquad \text{subject to}$$

$$w_i > 0,\ \sum_{i=1}^{m} w_i = 1,\ v_j > 0,\ \sum_{j=m+1}^{n} v_j = 1,$$
$$\sum_{i=1}^{m} w_i \boldsymbol{h}(\boldsymbol{X}_i) = \sum_{j=m+1}^{n} v_j \boldsymbol{h}(\boldsymbol{X}_j). \tag{2.2}$$

Han (2014) gave a justification of the compatibility of the constraints in (2.1) and (2.2) in the following way. Let $w(\boldsymbol{X}) = 1/\pi(\boldsymbol{X})$ and $v(\boldsymbol{X}) = 1/\{1-\pi(\boldsymbol{X})\}$. It is easy to verify that

$$E\left(w(\boldsymbol{X})\left[\boldsymbol{h}(\boldsymbol{X}) - E\{\boldsymbol{h}(\boldsymbol{X})\}\right]|R=1\right) = \mathbf{0}, \tag{2.3}$$
$$E\left(v(\boldsymbol{X})\left[\boldsymbol{h}(\boldsymbol{X}) - E\{\boldsymbol{h}(\boldsymbol{X})\}\right]|R=0\right) = \mathbf{0}.$$

Then constraints in (2.1) and (2.2) are simply the empirical version of these equalities with expectations replaced by sample averages.

One interesting result produced by calibration is multiple robustness. Suppose that multiple models $\pi^{(j)}(\boldsymbol{X};\boldsymbol{\alpha}^{(j)})$, $j = 1,\ldots,J$, for $\pi(\boldsymbol{X})$ and multiple models $a^{(k)}(\boldsymbol{X};\boldsymbol{\gamma}^{(k)})$, $k = 1,\ldots,K$, for $E(Y|\boldsymbol{X})$ are postulated. Han and Wang (2013) proposed to derive $\hat{w}_i$ the same way as in (2.1) with $\boldsymbol{h}(\boldsymbol{X}) = \{\pi^{(1)}(\hat{\boldsymbol{\alpha}}^{(1)}),\ldots,\pi^{(j)}(\hat{\boldsymbol{\alpha}}^{(J)}), a^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)}),\ldots,a^{(K)}(\hat{\boldsymbol{\gamma}}^{(K)})\}^{\mathrm{T}}$, where $\hat{\boldsymbol{\alpha}}^{(j)}$ and $\hat{\boldsymbol{\gamma}}^{(k)}$ are estimators of $\boldsymbol{\alpha}^{(j)}$ and $\boldsymbol{\gamma}^{(k)}$, respectively. The resulting estimator of μ_0, denoted by $\hat{\mu}_{\mathrm{mr}}$, is multiply robust, in that it is consistent if any one of the $J+K$ models is correctly specified. Multiple robustness significantly improves over double robustness on protecting estimation consistency against possible model misspecifications, since doubly robust estimators take $J = K = 1$.

In addition to multiple robustness, like other calibration-based estimators with weights derived from (2.1) or (2.2), $\hat{\mu}_{\mathrm{mr}}$ always falls into the parameter space of μ due to it being a convex combination of the observed Y. This is called the sample boundedness property (Robins et al. (2007); Tan (2010)) and is especially desirable when Y is binary. Another merit of $\hat{\mu}_{\mathrm{mr}}$ is its insensitivity to near-zero values of $\pi^{(j)}(\hat{\boldsymbol{\alpha}}^{(j)})$. The maximization in (2.1) makes the occurrence of extreme weights unlikely even if some estimated values of $\pi(\boldsymbol{X})$ are close to zero. Some numerical evidence of the superior performance of $\hat{\mu}_{\mathrm{mr}}$ in this case can be found in Han (2014).

3. Calibration and Multiple Robustness Under MNAR

3.1. Calibration under MNAR

It is seen that, under MAR, the calibration variables used to derive $\hat{\mu}_{\mathrm{mr}}$ are models for $\pi(\boldsymbol{X})$ and models for $E(Y|\boldsymbol{X})$. Under MNAR, models for $\pi(Y, \boldsymbol{X})$ can no longer serve as calibration variables because Y is missing for some subjects, but models for $E(Y|\boldsymbol{X})$ still can. However, the estimator with models for $E(Y|\boldsymbol{X})$ as calibration variables will no longer be consistent even if one model is correctly specified, since the proof of consistency in this case requires MAR assumption (Han and Wang (2013)).

Another look at the calibration in (2.1) when $\boldsymbol{h}(\boldsymbol{X})$ is taken to be models for $E(Y|\boldsymbol{X})$ reveals that the third constraint in (2.1) is essentially

$$\begin{aligned}&\sum_{i=1}^{m} w_i E^{(k)}(Y|\boldsymbol{X}_i, R_i = 1)\\&= \frac{1}{n}\sum_{i=1}^{n}\left\{R_i E^{(k)}(Y|\boldsymbol{X}_i, R_i = 1) + (1 - R_i)E^{(k)}(Y|\boldsymbol{X}_i, R_i = 0)\right\},\end{aligned} \tag{3.1}$$

with $E^{(k)}(\cdot)$ the expectation under the k-th model, because MAR implies that $E(Y|\boldsymbol{X})$, $E(Y|\boldsymbol{X}, R = 1)$ and $E(Y|\boldsymbol{X}, R = 0)$ are all equal. Now under MNAR, we propose to use the calibration constraint as in (3.1).

Modelling $E(Y|\boldsymbol{X}, R = 0)$ as needed by (3.1) is difficult since Y is not observed for subjects with $R = 0$. One possible solution is to use the fact that

$$f(Y|\boldsymbol{X}, R = 0) \propto f(Y|\boldsymbol{X}, R = 1)\frac{1 - \pi(Y, \boldsymbol{X})}{\pi(Y, \boldsymbol{X})}$$

(e.g., Kim and Yu (2011)) to obtain a model for $f(Y|\boldsymbol{X}, R = 0)$ by modelling $f(Y|\boldsymbol{X}, R = 1)$ and $\pi(Y, \boldsymbol{X})$. Suppose there are multiple pairs of models available: $\{f^{(k)}(Y|\boldsymbol{X}, R = 1; \boldsymbol{\gamma}^{(k)}), \pi^{(k)}(Y, \boldsymbol{X}; \boldsymbol{\alpha}^{(k)})\}$, $k = 1, \ldots, K$, each of which determines a model for $f(Y|\boldsymbol{X}, R = 0)$. Here $f^{(k)}(Y|\boldsymbol{X}, R = 1; \boldsymbol{\gamma}^{(k)})$, $k = 1, \ldots, K$, have to be different, because two models with the same $f^{(k)}(Y|\boldsymbol{X}, R = 1; \boldsymbol{\gamma}^{(k)})$ but different $\pi^{(k)}(Y, \boldsymbol{X}; \boldsymbol{\alpha}^{(k)})$ will lead to (3.1) with the same left-hand side but different right-hand sides.

The $\boldsymbol{\gamma}^{(k)}$ can be easily estimated through a complete-case analysis by maximizing $\prod_{i=1}^{m} f^{(k)}(Y_i|\boldsymbol{X}_i, R_i = 1; \boldsymbol{\gamma}^{(k)})$. More attention is needed to the estimation of $\boldsymbol{\alpha}^{(k)}$ due to possible identification problem: the $\boldsymbol{\alpha}^{(k)}$ may not be identifiable from the observed data without further assumptions on the data generating process. One general result ensuring identification is to assume that $\boldsymbol{X}$ contains an "instrumental" or "shadow" variable, that, conditional on the rest of $\boldsymbol{X}$, is associated with Y but independent of R (Wang, Shao and Kim (2014); Miao and Tchetgen Tchetgen (2016); Shao and Wang (2016)). Instead of repeating the detailed result and the exact conditions, which can be found in Wang, Shao and Kim (2014), we make a direct assumption on identification.

Assumption 1. *The $\boldsymbol{\alpha}^{(k)}$ are identifiable.*

Under Assumption 1, the estimation of $\boldsymbol{\alpha}^{(k)}$ can follow some existing methods (Rotnitzky and Robins (1997); Rotnitzky, Robins and Scharfstein (1998); Chang and Kott (2008); Wang, Shao and Kim (2014); Miao and Tchetgen Tchetgen (2016); Shao and Wang (2016)) by solving equations $\sum_{i=1}^{n}\{R_i/\pi_i^{(k)}(\boldsymbol{\alpha}^{(k)})-1\}\boldsymbol{h}(\boldsymbol{X}_i)=\mathbf{0}$ derived based on the fact that $E[\{R/\pi(Y,\boldsymbol{X})-1\}\boldsymbol{h}(\boldsymbol{X})]=\mathbf{0}$. When the user-specified $\boldsymbol{h}(\boldsymbol{X})$ has dimension larger than that of $\boldsymbol{\alpha}^{(k)}$, the equations may be solved by the generalized method of moments (Hansen (1982)) or the EL method.

Let $\hat{\boldsymbol{\gamma}}^{(k)}$ and $\hat{\boldsymbol{\alpha}}^{(k)}$ denote estimators of $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\alpha}^{(k)}$, respectively. Under each $f^{(k)}(Y|\boldsymbol{X},R=1;\hat{\boldsymbol{\gamma}}^{(k)})$, an estimator of $E(Y|\boldsymbol{X},R=1)$, denoted by $a_1^{(k)}(\boldsymbol{X};\hat{\boldsymbol{\gamma}}^{(k)})$, is readily available. On the other hand, although a model $f^{(k)}(Y|\boldsymbol{X},R=0;\hat{\boldsymbol{\gamma}}^{(k)},\hat{\boldsymbol{\alpha}}^{(k)})$ has been completely determined by $\{f^{(k)}(Y|\boldsymbol{X},R=1;\hat{\boldsymbol{\gamma}}^{(k)}),\pi^{(k)}(Y,\boldsymbol{X};\hat{\boldsymbol{\alpha}}^{(k)})\}$, directly deriving a closed-form estimator of $E(Y|\boldsymbol{X},R=0)$ is generally difficult, since $f^{(k)}(Y|\boldsymbol{X},R=0;\hat{\boldsymbol{\gamma}}^{(k)},\hat{\boldsymbol{\alpha}}^{(k)})$ may not be a well-known distribution and calculating its expectation may involve complicated integrals. An easy way to estimate $E(Y|\boldsymbol{X},R=0)$ is to take L random draws $\hat{Y}_0^{(1)},\ldots,\hat{Y}_0^{(L)}$ from $f^{(k)}(Y|\boldsymbol{X},R=0;\hat{\boldsymbol{\gamma}}^{(k)},\hat{\boldsymbol{\alpha}}^{(k)})$ and then use $a_0^{(k)}(\boldsymbol{X};\hat{\boldsymbol{\gamma}}^{(k)},\hat{\boldsymbol{\alpha}}^{(k)},L)=L^{-1}\sum_{l=1}^{L}\hat{Y}_0^{(l)}$ as the k-th estimator. Our proposed estimator of μ_0 is $\hat{\mu}=\sum_{i=1}^{m}\hat{w}_iY_i$, where $\hat{w}_i$ are derived through

$$\max_{w_1,\ldots,w_m}\prod_{i=1}^{m}w_i \qquad \text{subject to} \quad w_i>0,\ \sum_{i=1}^{m}w_i=1,$$
$$\sum_{i=1}^{m}w_ia_1^{(k)}(\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}}^{(k)})=\frac{1}{n}\sum_{i=1}^{n}\Big\{R_ia_1^{(k)}(\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}}^{(k)})+(1-R_i)a_0^{(k)}(\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}}^{(k)},\hat{\boldsymbol{\alpha}}^{(k)},L)\Big\},$$
$$k=1,\ldots,K. \tag{3.2}$$

In general, unlike the MAR case, there is no easy way to justify the compatibility of constraints in (3.2) for a particular set of K models. To avoid this challenging problem, we make the following assumption on compatibility.

Assumption 2. *There exists $w(\boldsymbol{X})>0$, possibly depending on the K pairs of models, such that*

$$E\Big[w(\boldsymbol{X})\{E^{(k)}(Y|\boldsymbol{X},R=1)-E^{(k)}(Y)\}|R=1\Big]=0, \quad k=1,\ldots,K. \tag{3.3}$$

Here (3.3) is an analog of (2.3), now under MNAR. While (2.3) is ensured by simply taking $w(\boldsymbol{X})=1/\pi(\boldsymbol{X})$ (Han (2014)), there does not seem to be an obvious $w(\boldsymbol{X})$ to ensure (3.3) under MNAR. For our proposed method, however, the exact expression for such a $w(\boldsymbol{X})$ is not needed due to the EL formulation, and thus we simply assume its existence as in Assumption 2. Mathematically, (3.3) assumes the existence of a positive function $w(\boldsymbol{X})$ that is orthogonal to K given functions of $\boldsymbol{X}$, $E^{(k)}(Y|\boldsymbol{X}, R=1) - E^{(k)}(Y)$, $k = 1, \ldots, K$, and such an existence is not uncommon in the Hilbert space of all square-integrable functions of $\boldsymbol{X}$. Under Assumption 2, the constraints in (3.2) are empirical versions of (3.3); see (3.6).

Based on the EL theory, the solution to (3.2) is given by

$$\hat{w}_i = \frac{1}{m}\frac{1}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, L)}$$

with

$$1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, L) > 0, \quad i = 1, \ldots, m, \tag{3.4}$$

where $\hat{\boldsymbol{\rho}}$ is the Lagrange multiplier solving

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, L)}{1+\boldsymbol{\rho}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, L)} = \boldsymbol{0}, \tag{3.5}$$

$$\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, L) = \begin{pmatrix} a_1^{(1)}(\boldsymbol{X}; \hat{\boldsymbol{\gamma}}^{(1)}) - \dfrac{1}{n}\sum_{j=1}^{n}\Big\{R_j a_1^{(1)}(\boldsymbol{X}_j; \hat{\boldsymbol{\gamma}}^{(1)}) \\ +(1-R_j)a_0^{(1)}(\boldsymbol{X}_j; \hat{\boldsymbol{\gamma}}^{(1)}, \hat{\boldsymbol{\alpha}}^{(1)}, L)\Big\} \\ \vdots \\ a_1^{(K)}(\boldsymbol{X}; \hat{\boldsymbol{\gamma}}^{(K)}) - \dfrac{1}{n}\sum_{j=1}^{n}\Big\{R_j a_1^{(K)}(\boldsymbol{X}_j; \hat{\boldsymbol{\gamma}}^{(K)}) \\ +(1-R_j)a_0^{(K)}(\boldsymbol{X}_j; \hat{\boldsymbol{\gamma}}^{(K)}, \hat{\boldsymbol{\alpha}}^{(K)}, L)\Big\} \end{pmatrix}.$$

Directly solving (3.5) for $\hat{\boldsymbol{\rho}}$ is not a good option as (3.5) typically has multiple roots, yet it is the $\hat{\boldsymbol{\rho}}$ satisfying (3.4) that is needed. A better way for numerical implementation is as follows. Let $F_n(\boldsymbol{\rho}) = -m^{-1}\sum_{i=1}^{m}\log\{1+\boldsymbol{\rho}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, L)\}$, a strictly convex function of $\boldsymbol{\rho}$. Under (3.3) it can be shown that $F_n(\boldsymbol{\rho})$ has a unique minimizer, at least when n is large (Han (2014)). The stationary equation $\partial F_n(\boldsymbol{\rho})/\partial\boldsymbol{\rho} = \boldsymbol{0}$ for this minimizer turns out to be (3.5), and this minimizer must satisfy (3.4) because of the log function in $F_n(\boldsymbol{\rho})$. Therefore, the $\hat{\boldsymbol{\rho}}$ needed can be derived by minimizing $F_n(\boldsymbol{\rho})$. This is a convex minimization problem and is easily implemented using the Newton-Raphson algorithm. Refer to Chen, Sitter and Wu (2002) and Han (2014) for a detailed description of the algorithm. Chen, Sitter and Wu (2002) also showed that this algorithm always converges.

3.2. Multiple robustness

The proposed estimator $\hat{\mu}$ is multiply robust, in the sense that it is consistent, as $n \to \infty$ and $L \to \infty$, if any one of the K pairs of models $\{f^{(k)}(Y|\boldsymbol{X}, R = 1; \boldsymbol{\gamma}^{(k)}), \pi^{(k)}(Y, \boldsymbol{X}; \boldsymbol{\alpha}^{(k)})\}$ is correctly specified. To see this, let the correct pair be $\{f^{(1)}(Y|\boldsymbol{X}, R = 1; \boldsymbol{\gamma}^{(1)}), \pi^{(1)}(Y, \boldsymbol{X}; \boldsymbol{\alpha}^{(1)})\}$ without loss of generality. As $n \to \infty$ and $L \to \infty$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big\{R_i a_1^{(k)}(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}^{(k)}) + (1-R_i) a_0^{(k)}(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}^{(k)}, \hat{\boldsymbol{\alpha}}^{(k)}, L)\Big\} \\
&\xrightarrow{p} E\Big\{RE^{(k)}(Y|\boldsymbol{X}, R=1) + (1-R)E^{(k)}(Y|\boldsymbol{X}, R=0)\Big\} \\
&= P(R=1)E\Big\{E^{(k)}(Y|\boldsymbol{X}, R=1)|R=1\Big\} \\
&\quad + P(R=0)E\Big\{E^{(k)}(Y|\boldsymbol{X}, R=0)|R=0\Big\} \\
&= P(R=1)E^{(k)}(Y|R=1) + P(R=0)E^{(k)}(Y|R=0) \\
&= E^{(k)}(Y). \qquad (3.6)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\hat{\mu} &= \sum_{i=1}^{m} \hat{w}_i Y_i \\
&= \sum_{i=1}^{m} \hat{w}_i \left\{Y_i - a_1^{(1)}\left(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}^{(1)}\right)\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left\{R_i a_1^{(1)}\left(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}^{(1)}\right) + (1-R_i) a_0^{(1)}\left(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}^{(1)}, \hat{\boldsymbol{\alpha}}^{(1)}, L\right)\right\} \\
&= \frac{1}{m}\sum_{i=1}^{m} \frac{Y_i - a_1^{(1)}\left(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}^{(1)}\right)}{1 + \hat{\boldsymbol{\rho}}^{\mathrm{T}} \hat{\boldsymbol{g}}_i\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, L\right)} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left\{R_i a_1^{(1)}\left(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}^{(1)}\right) + (1-R_i) a_0^{(1)}\left(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}^{(1)}, \hat{\boldsymbol{\alpha}}^{(1)}, L\right)\right\} \\
&\xrightarrow{p} E\left\{\left.\frac{Y - E^{(1)}\left(Y|\boldsymbol{X}, R=1\right)}{1 + \boldsymbol{\rho}_*^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{X})}\right| R=1\right\} + E^{(1)}(Y) = \mu_0,
\end{aligned}$$

where $\boldsymbol{\rho}_*$ is the probability limit of $\hat{\boldsymbol{\rho}}$ and

$$\boldsymbol{g}(\boldsymbol{X}) = \begin{pmatrix} E^{(1)}(Y|\boldsymbol{X}, R=1) - E^{(1)}(Y) \\ \vdots \\ E^{(K)}(Y|\boldsymbol{X}, R=1) - E^{(K)}(Y) \end{pmatrix}.$$

Thus $\hat{\mu}$ is a consistent estimator of μ_0.

Theorem 1. *Under Assumptions* 1 *and* 2 *and the regularity conditions in the Appendix, as $n \to \infty$ and $L \to \infty$, if any one of the K pairs of models $\{f^{(k)}(Y|\boldsymbol{X}, R = 1; \boldsymbol{\gamma}^{(k)}), \pi^{(k)}(Y, \boldsymbol{X}; \boldsymbol{\alpha}^{(k)})\}$ is correctly specified, then $\hat{\mu} \xrightarrow{p} \mu_0$.*

The above multiple robustness result under MNAR is slightly different from that under MAR. Under MAR, a model for $f(Y|\boldsymbol{X})$ (or equivalently for $f(Y|\boldsymbol{X}, R = 1)$) and a model for $\pi(\boldsymbol{X})$ contribute independently to consistency. Under MNAR, a model for $f(Y|\boldsymbol{X}, R = 1)$ and a model for $\pi(Y, \boldsymbol{X})$ work as a pair to make joint contribution to consistency, and only when the whole pair is correctly specified is the resulting estimator consistent. By construction, $\hat{\mu}$ is a convex combination of the observed Y, and thus is always in the parameter space.

Deriving the asymptotic distribution of $\hat{\mu}$ is challenging, since in general we do not know which model is correctly specified. In addition, advanced empirical processes theory is needed to deal with the implicit dependence of the random draws on the nuisance parameters $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\alpha}^{(k)}$ (e.g., Wang and Robins (1998); Robins and Wang (2000)). Dealing with these challenges is beyond the scope of this paper. In practice, the standard error of $\hat{\mu}$ can be calculated by bootstrapping, the effectiveness of which is demonstrated in the next section.

4. Simulation Study

In this section we report on the finite sample performance of the proposed estimator. The construction of a simulation model is a delicate issue in our case. To demonstrate multiple robustness, closed form expressions of $f(Y|\boldsymbol{X}, R = 1)$ and $\pi(Y, \boldsymbol{X})$ need to be available so that we can postulate correct models, but generating data from a distribution corresponding to pre-fixed $f(Y|\boldsymbol{X}, R = 1)$ and $\pi(Y, \boldsymbol{X})$ seems challenging. Therefore, we choose to fix the data generating model first, and then mathematically derive $f(Y|\boldsymbol{X}, R = 1)$ and/or $\pi(Y, \boldsymbol{X})$. The derivation involves calculating integrals and leads to closed form expressions only under carefully chosen data generating models.

We considered an auxiliary variable $X \sim N(0, 1)$. Given X, Y followed a generalized logistic distribution with density

$$f(Y|X) = \frac{2\exp(-(Y - 1 - X - X^2))}{\{1 + \exp(-(Y - 1 - X - X^2))\}^3}, \quad -\infty < Y < \infty,$$

and $E(Y) = 3$. The missingness was generated by $R|Y, X \sim \text{Bernoulli}(\pi(Y, X))$ with $\pi(Y, X) = \{1 + \exp(1 + X + X^2 - Y)\}^{-1}$, for which 33% of the subjects were with Y missing. For this particular data generating process, it is easy to verify that

$$f(Y|X,R=1)=\frac{3\exp(-(Y-1-X-X^2))}{\{1+\exp(-(Y-1-X-X^2))\}^4},\quad -\infty<Y<\infty.$$

We considered two models for $f(Y|X,R=1)$: a correct one

$$f^{(1)}(Y|X,R=1;\boldsymbol{\gamma}^{(1)})=\frac{3\exp\left\{-\left(Y-\gamma_1^{(1)}-\gamma_2^{(1)}X-\gamma_3^{(1)}X^2\right)\right\}}{\left[1+\exp\left\{-\left(Y-\gamma_1^{(1)}-\gamma_2^{(1)}X-\gamma_3^{(1)}X^2\right)\right\}\right]^4}$$

and an incorrect one $f^{(2)}(Y|X,R=1;\boldsymbol{\gamma}^{(2)})$ being the density of a Normal distribution with mean $\gamma_1^{(2)}+\gamma_2^{(2)}X^2$ and standard deviation $\gamma_3^{(2)}$. We also considered two models for $\pi(Y,X)$: a correct one $\pi^{(1)}(Y,X;\boldsymbol{\alpha}^{(1)})=\{1+\exp(\alpha_1^{(1)}+\alpha_2^{(1)}X+\alpha_3^{(1)}X^2+\alpha_4^{(1)}Y)\}^{-1}$ and an incorrect one $\pi^{(2)}(Y,X;\boldsymbol{\alpha}^{(2)})=\{1+\exp(\alpha_1^{(2)}+\alpha_2^{(2)}Y)\}^{-1}$. In our simulation, existing estimation methods (e.g. Rotnitzky, Robins and Scharfstein (1998); Wang, Shao and Kim (2014)) frequently produced $\hat{\boldsymbol{\alpha}}^{(1)}$ and $\hat{\boldsymbol{\alpha}}^{(2)}$ with erroneously large norm, making most of the estimated values of $\pi(Y,X)$ equal to 1. This poor performance might be that, due to cautious selection of the data generating model in order that $f(Y|X,R=1)$ and $\pi(Y,X)$ have closed form expressions, the parameters $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ may be non-identifiable. Thus we fixed $\boldsymbol{\alpha}^{(1)}$ at its true value $(1,1,1,-1)^{\mathrm{T}}$ and $\boldsymbol{\alpha}^{(2)}$ at an arbitrary value $(-2,0.2)^{\mathrm{T}}$. All methods under comparison were derived based on these fixed values. Fixing some parameter values in a model for $\pi(Y,X)$ is a common practice in sensitivity analysis (e.g. Rotnitzky, Robins and Scharfstein (1998)).

In total we had $K=4$ models,

$$\begin{aligned}
&\text{model 1}:\left\{f^{(1)}\left(Y|X,R=1;\boldsymbol{\gamma}^{(1)}\right),\quad \pi^{(1)}\left(Y,X;\boldsymbol{\alpha}^{(1)}\right)\right\},\\
&\text{model 2}:\left\{f^{(2)}\left(Y|X,R=1;\boldsymbol{\gamma}^{(2)}\right),\quad \pi^{(1)}\left(Y,X;\boldsymbol{\alpha}^{(1)}\right)\right\},\\
&\text{model 3}:\left\{f^{(1)}\left(Y|X,R=1;\boldsymbol{\gamma}^{(1)}\right),\quad \pi^{(2)}\left(Y,X;\boldsymbol{\alpha}^{(2)}\right)\right\},\\
&\text{model 4}:\left\{f^{2}\left(Y|X,R=1;\boldsymbol{\gamma}^{2}\right),\quad \pi^{(2)}\left(Y,X;\boldsymbol{\alpha}^{(2)}\right)\right\},
\end{aligned}$$

and model 1 is correctly specified. Since the models used to calculate $\hat{\mu}$ need to have different $f^{(k)}(Y|X,R=1;\boldsymbol{\gamma}^{(k)})$, we had four possible combinations when choosing more than one model from the above four to calculate $\hat{\mu}$: {model 1, model 2}, {model 1, model 4}, {model 2, model 3}, and {model 3, model 4}. To make comparison, we also computed the inverse probability weighted estimator $\hat{\mu}_{\mathrm{ipw}}^{(k)}=n^{-1}\sum_{i=1}^n R_iY_i/\pi_i^{(k)}(\hat{\boldsymbol{\alpha}}^{(k)})$ and the imputation estimator $\hat{\mu}_{\mathrm{im}}^{(k)}=n^{-1}\sum_{i=1}^n\{R_iY_i+(1-R_i)a_0^{(k)}(X_i;\hat{\boldsymbol{\gamma}}^{(k)},\hat{\boldsymbol{\alpha}}^{(k)},L)\}$. We took $n=200$ and $n=500$ and summarized the results based on 1,000 replications. Given X, Y was gen-

Table 1. Simulation results based on $n = 200$ and 1,000 replications. The number in the name of each estimator indicates which one(s) among the 4 models is (are) used. The results have been multiplied by 100.

	$L = 5$			$L = 10$			$L = 20$		
Estimator	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
IPW-1	0	23	15						
IPW-2	41	68	35						
IM-1	−8	18	13	−8	18	12	−8	18	12
IM-2	−50	54	49	−50	54	49	−50	54	50
IM-3	41	44	41	41	44	41	41	44	41
IM-4	65	67	65	65	67	64	65	67	65
MR-1	−7	19	13	−8	19	13	−8	19	13
MR-2	−28	40	26	−27	42	26	−28	40	26
MR-3	41	44	41	41	44	41	41	44	41
MR-4	64	67	65	64	67	64	64	67	65
MR-12	12	31	20	12	31	20	12	32	21
MR-14	−8	23	14	−8	23	14	−8	23	14
MR-23	50	54	50	50	54	50	50	54	50
MR-34	40	43	40	40	43	41	40	43	41

RMSE: root mean square error. MAE: median absolute error. IPW: the inverse probability weighted estimator. IM: the imputation estimator. MR: the proposed estimator.

Table 2. Simulation results based on $n = 500$ and 1,000 replications. The number in the name of each estimator indicates which one(s) among the 4 models is (are) used. The results have been multiplied by 100.

	$L = 5$			$L = 10$			$L = 20$		
Estimator	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
IPW-1	0	15	11						
IPW-2	39	49	36						
IM-1	−8	13	9	−8	13	9	−8	13	9
IM-2	−51	52	50	−51	53	51	−51	52	51
IM-3	41	42	41	41	42	41	41	42	41
IM-4	65	66	65	65	66	65	65	66	65
MR-1	−8	14	9	−8	14	9	−8	13	10
MR-2	−35	44	33	−35	42	33	−35	43	33
MR-3	41	42	41	41	42	41	41	42	41
MR-4	65	66	65	65	66	65	65	66	65
MR-12	12	24	17	13	23	17	13	23	17
MR-14	−8	15	10	−8	15	10	−8	15	10
MR-23	51	53	51	51	53	51	51	53	51
MR-34	41	42	41	42	47	41	41	42	41

RMSE: root mean square error. MAE: median absolute error. IPW: the inverse probability weighted estimator. IM: the imputation estimator. MR: the proposed estimator.

Table 3. Performance of the bootstrapping method based on $n = 200$, $L = 5$ and 500 replications. The bootstrapping resampling size is 100. The number in the name of each estimator indicates which one(s) among the 4 models is (are) used.

Estimator	Bias	SE-EMP	SE-B	CP-B (%)
MR-1	−0.07	0.19	0.18	91.8
MR-2	−0.28	0.28	0.28	84.4
MR-3	0.41	0.17	0.16	24.2
MR-4	0.64	0.20	0.19	6.2
MR-12	0.12	0.27	0.27	92.2
MR-14	−0.07	0.22	0.24	95.6
MR-23	0.51	0.21	0.22	28.6
MR-34	0.41	0.17	0.16	25.8

SE-EMP: empirical standard error; SE-B: averaged bootstrapping standard error; CP-B: coverage probability of the 95% confidence interval based bootstrapping standard errors.

erated using the R package "glogis". Random draws from the distribution $f^{(k)}(Y|X, R = 0; \hat{\boldsymbol{\gamma}}^{(k)}, \hat{\boldsymbol{\alpha}}^{(k)})$ were generated using the R package "distr". Tables 1 and 2 contain the simulation results based on $n = 200$ and $n = 500$, respectively.

Since no parameters were estimated for the inverse probability weighted estimator, IPW-1 and IPW-2 can serve as the benchmark for comparison. Estimators IM-1, MR-1, MR-12 and MR-14 used the correct model 1, and thus were all consistent. This is confirmed by their small bias. The small bias of MR-12 and MR-14 demonstrates the multiple robustness of our proposed estimator. Estimators that do not use the correct model 1 are inconsistent, and this is confirmed by the corresponding large bias. The simulation results also show that the number of random draws L does not have a dramatic impact on the numerical performance.

Table 3 contains results on the performance of the bootstrapping method in calculating the standard error. We took $n = 200$ and the bootstrapping resampling size to be 100. To save computational time, we took $L = 5$ and carried out 500 replications. It is seen that, for each estimator, the empirical standard error and the averaged bootstrapping standard error are very close. In addition, for the consistent estimators MR-1, MR-12 and MR-14 the coverage probabilities of the 95% bootstrapping-based confidence intervals are very close to the nominal level. These observations suggest the effectiveness of the bootstrapping method in calculating the standard error of the proposed estimator.

5. Discussion

Multiple robustness is a desirable property, as it significantly improves the robustness of estimation consistency against possible model misspecifications. In this paper, when data are missing not at random, we have proposed a multiply robust estimator constructed based on calibration.

Theoretically, the number of models postulated has no effect on the multiple robustness, as long as this number stays fixed as the sample size varies. On the other hand, its effect on efficiency is very difficult to study, if not totally impossible, for two reasons. In the current case of MNAR, a closed form expression for the asymptotic variance is difficult to derive; even in the case of MAR where the asymptotic variance is available (e.g., Han and Wang (2013); Han (2014)), the efficiency is affected not only by the number of models but also by the particular functional form of those models. In general, it may be impossible to compare efficiency based on different models because of its complex dependence on them. The numerical performance deteriorates as a large number of models are simultaneously accounted for, since the dimension of the Lagrange multiplier $\boldsymbol{\rho}$ is large in this case. Therefore, we recommend that each model be carefully constructed and used to derive the proposed estimator.

This paper considered estimating the mean of a response. It is of interest to extend the current method to regression analysis with missing response and/or covariates.

Acknowledgment

We wish to thank the Editor, an associate editor and two reviewers for their valuable comments that have helped to greatly improve the quality of this work. Support for this project was partially provided by the Natural Sciences and Engineering Research Council of Canada.

Appendix

Assume that, for $k = 1, \ldots, K$, (1) the parameter spaces $\mathcal{A}^k$ and $\mathcal{G}^k$ for $\boldsymbol{\alpha}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$, respectively, are compact; (2) $\pi^{(k)}(Y, \boldsymbol{X}; \boldsymbol{\alpha}^{(k)})$ and $f^{(k)}(Y|\boldsymbol{X}, R = 1; \boldsymbol{\gamma}^{(k)})$ are continuous in $\boldsymbol{\alpha}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$, respectively; (3) $E\{\log f(Y|\boldsymbol{X}, R = 1)\} < \infty$ and $E\{\sup_{\boldsymbol{\gamma}^{(k)} \in \mathcal{G}^k} |\log f^{(k)}(Y|\boldsymbol{X}, R = 1; \boldsymbol{\gamma}^{(k)})|\} < \infty$; (4) the Kullback–Leibler distance between $f(Y|\boldsymbol{X}, R = 1)$ and $f^{(k)}(Y|\boldsymbol{X}, R = 1; \boldsymbol{\gamma}^{(k)})$, viewed as a function of $\boldsymbol{\gamma}^{(k)}$, has a unique minimum in $\mathcal{G}^k$; (5) $\pi(Y, \boldsymbol{X})$ is bounded away from 0; (6) $E[\sup_{\boldsymbol{\alpha}^{(k)} \in \mathcal{A}^k} ||\{R/\pi^{(k)}(Y, \boldsymbol{X}; \boldsymbol{\alpha}^{(k)}) - 1\}\boldsymbol{h}(\boldsymbol{X})||] < \infty$, where $\boldsymbol{h}(\boldsymbol{X})$ is user-specified for estimating $\boldsymbol{\alpha}^{(k)}$; (7) the quadratic form of $E[\{R/\pi^{(k)}(Y,\boldsymbol{X};\boldsymbol{\alpha}^{(k)})-1\}\boldsymbol{h}(\boldsymbol{X})]$ with the weighting matrix being the inverse of the covariance matrix of $\{R/\pi^{(k)}(Y,\boldsymbol{X};\boldsymbol{\alpha}^{(k)})-1\}\boldsymbol{h}(\boldsymbol{X})$, viewed as a function of $\boldsymbol{\alpha}^{(k)}$, has a unique minimum in $\mathcal{A}^k$; (8) $a_1^{(k)}(\boldsymbol{X};\boldsymbol{\gamma}^{(k)})$ and $a_0^{(k)}(\boldsymbol{X};\boldsymbol{\gamma}^{(k)},\boldsymbol{\alpha}^{(k)},L)$ are continuous in $\boldsymbol{\gamma}^{(k)}$ and $(\boldsymbol{\gamma}^{(k)},\boldsymbol{\alpha}^{(k)})$, respectively; (9) $E\{\sup_{\boldsymbol{\gamma}^{(k)}\in\mathcal{G}^k}|a_1^{(k)}(\boldsymbol{X};\boldsymbol{\gamma}^{(k)})|\}<\infty$ and $E\{\sup_{\boldsymbol{\gamma}^{(k)}\in\mathcal{G}^k,\boldsymbol{\alpha}^{(k)}\in\mathcal{A}^k}|a_0^{(k)}(\boldsymbol{X};\boldsymbol{\gamma}^{(k)},\boldsymbol{\alpha}^{(k)},L)|\}<\infty$; (10) $E[\sup_{\boldsymbol{\rho}\in\mathcal{P}}\log\{1+\boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{\alpha},\boldsymbol{\gamma},L)\}|R=1]<\infty$ where $\mathcal{P}$ is the parameter space for $\boldsymbol{\rho}$ and is compact; (11) $E[\sup_{\boldsymbol{\gamma}^{(k)}\in\mathcal{G}^k,\boldsymbol{\alpha}^{(k)}\in\mathcal{A}^k,\boldsymbol{\rho}\in\mathcal{P}}\{Y-a_1^{(k)}(\boldsymbol{X};\boldsymbol{\gamma}^{(k)})\}/\{1+\boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{\alpha},\boldsymbol{\gamma},L)\}|R=1]<\infty$.

Remark: (1)-(4) ensure the convergence in probability of $\hat{\boldsymbol{\gamma}}^{(k)}$ (White (1982)); (1), (2) and (5)-(7) ensure the convergence in probability of $\hat{\boldsymbol{\alpha}}^{(k)}$ (Hall (2005)); (1), (2) and (8)-(11) ensure the weak law of large numbers needed in the proof of Theorem 1 (Newey and McFadden (1994); Schennach (2007)).

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–972.

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

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95**, 555–571.

Chen, J. and Sitter, R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica* **9**, 385–406.

Chen, J., Sitter, R. R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* **89**, 230–237.

Chen, S. X., Leung, D. H. Y. and Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society, Series B Statistical Methodology* **70**, 803–823.

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

Hall, A. R. (2005). *Generalized Method of Moments.* Oxford University Press.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* **109**, 1159–1173.

Han, P. (2016a). Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scandinavian Journal of Statistics* **43**, 246–260.

Han, P. (2016b). Intrinsic efficiency and multiple robustness in longitudinal studies with dropout. *Biometrika* **103**, 683–700.

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

Hansen, L. P. (1982). Large sample properties of generalized methods of moments estimators. *Econometrica* **50**, 1029–1054.

Kim, J. K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica* **19**, 145–158.

Kim, J. K. (2010). Calibration estimation using exponential tilting in sample surveys. *Survey Methodology* **36**, 145–155.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review* **78**, 21–39.

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

Lundström, S. and Särndal, C. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics* **15**, 305–327.

Miao, W. and Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* **103**, 475–482.

Newey, W. K. and McFadden, D. L. (1994). *Large Sample Estimation and Hypothesis Testing.* Handbook of Econometrics, Vol 4. Amsterdam, The Netherlands: Elsevier Science.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A. (2001). *Empirical Likelihood.* Chapman & Hall/CRC Press, New York.

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

Qin, J., Shao, J. and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association* **103**, 797–810.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society, Series B Statistical Methodology* **69**, 101–122.

Robins, J. M., Sued, M., Gomez-Lei, Q. and Rotnitzky, A. (2007). Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science* **22**, 544–559.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.

Rotnitzky, A. and Robins, J. M. (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine* **16**, 81–102.

Rotnitzky, A., Robins, J. M. and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

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

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

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *the Annals of Statistics* **35**, 634–672.

Shao, J. and Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* **103**, 175–187.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal* *of the American Statistical Association* **101**, 1619–1637.

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

Tan, Z. and Wu, C. (2015). Generalized pseudo empirical likelihood inferences for complex surveys. *Canadian Journal of Statistics* **43**, 1–17.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data.* Springer, New York.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–948.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika* **90**, 937–951.

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

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA.

E-mail: peisong@umich.edu

(Received November 2015; accepted January 2017)