Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Multiply Robust Estimation in Regression Analysis With Missing Data

Peisong Han

To cite this article: Peisong Han (2014) Multiply Robust Estimation in Regression Analysis With Missing Data, Journal of the American Statistical Association, 109:507, 1159-1173, DOI: 10.1080/01621459.2014.880058

To link to this article: https://doi.org/10.1080/01621459.2014.880058

Multiply Robust Estimation in Regression Analysis With Missing Data

Peisong HAN

Doubly robust estimators are widely used in missing-data analysis. They provide double protection on estimation consistency against model misspecifications. However, they allow only a single model for the missingness mechanism and a single model for the data distribution, and the assumption that one of these two models is correctly specified is restrictive in practice. For regression analysis with possibly missing outcome, we propose an estimation method that allows multiple models for both the missingness mechanism and the data distribution. The resulting estimator is consistent if any one of those multiple models is correctly specified, and thus provides multiple protection on consistency. This estimator is also robust against extreme values of the fitted missingness probability, which, for most doubly robust estimators, can lead to erroneously large inverse probability weights that may jeopardize the numerical performance. The numerical implementation of the proposed method through a modified Newton–Raphson algorithm is discussed. The asymptotic distribution of the resulting estimator is derived, based on which we study the estimation efficiency and provide ways to improve the efficiency. As an application, we analyze the data collected from the AIDS Clinical Trials Group Protocol 175.

KEY WORDS: Augmented inverse probability weighting (AIPW); Double robustness; Empirical likelihood; Estimating functions; Extreme weights; Missing at random (MAR).

1. INTRODUCTION

Missing-data problems are commonly seen in practical studies, especially when human subjects are involved. There are many reasons that could lead to data collection with missing values; for example, budget restriction, technique failure, and subjects' noncompliance with the protocol or refusal to respond to certain questions. In sample surveys, missing data are often caused by design where the missingness process is determined by the investigator. Missing data can raise significant challenges to estimation and inference, as the subjects with fully observed data oftentimes represent a biased sample from the population, yet it is usually the characteristics of the population that are of main interest. A direct application of methods developed for data without missing values to solving missing-data problems may lead to biased estimation and misleading conclusions. Therefore, special methods are needed.

In this article we consider regression analysis where the outcome is subject to missingness. To fix notation, let Y denote the outcome, $\boldsymbol{X}$ the vector of covariates, and $\boldsymbol{S}$ the vector of auxiliary variables. The full data we intend to collect are n independent and identically distributed copies of $(Y, \boldsymbol{X}^{\mathrm{T}}, \boldsymbol{S}^{\mathrm{T}})$. Our main interest is to estimate the p-dimensional vector $\boldsymbol{\beta}_0$ defined by $E(Y|\boldsymbol{X}) = \mu(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)$, namely the coefficients of the mean regression of Y on $\boldsymbol{X}$. Here, $\mu(\cdot)$ is some known monotone and continuously differentiable link function. Note that the auxiliary variables $\boldsymbol{S}$ are not of direct statistical interest and do not enter the regression model. However, they may help explain the missingness mechanism and improve the estimation efficiency, and thus reduce the impact of missing data on estimation. The collection of auxiliary variables is common in many practical studies, and some examples can be found in Pepe (1992), Pepe, Reilly, and Fleming (1994), and Wang, Rotnitzky, and Lin (2010).

When there are no missing data, $\boldsymbol{\beta}_0$ can be estimated by solving

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = \boldsymbol{0}. \tag{1}$$

Here, $\boldsymbol{U}(Y, \boldsymbol{X}, \boldsymbol{\beta}) = \boldsymbol{D}(\boldsymbol{X}, \boldsymbol{\beta})\{Y - \mu(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}$ is a set of estimating functions, with $\boldsymbol{D}(\boldsymbol{X}, \boldsymbol{\beta})$ being a user-specified p-dimensional vector that may depend on $\boldsymbol{X}$ and $\boldsymbol{\beta}$.

When Y is subject to missingness, let R denote the indicator of observing Y; that is, $R = 1$ if Y is observed and $R = 0$ if Y is missing. The observed data are n independent and identically distributed copies of $(R, RY, \boldsymbol{X}^{\mathrm{T}}, \boldsymbol{S}^{\mathrm{T}})$. To estimate $\boldsymbol{\beta}_0$ in the presence of missing Y, the simplest way is the complete-case analysis, which solves $n^{-1}\sum_{i=1}^{n} R_i\boldsymbol{U}(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = \boldsymbol{0}$ to derive the estimator. However, unless Y is missing completely at random (Little and Rubin 2002) in the sense that $P(R = 1|Y, \boldsymbol{X}, \boldsymbol{S}) = P(R = 1)$, subjects with fully observed data form a biased sample from the population, and thus the complete-case analysis produces a biased estimator of $\boldsymbol{\beta}_0$. A widely used method for correcting the selection bias introduced by missing data is to weight the complete cases by the inverse of their (estimated) selection probabilities (Horvitz and Thompson 1952; Robins, Rotnitzky, and Zhao 1994), which leads to the inverse probability weighted (IPW) estimator. To model the selection probability, the missing at random (MAR) mechanism (Little and Rubin 2002) is usually assumed; that is,

$$P(R = 1|Y, \boldsymbol{X}, \boldsymbol{S}) = P(R = 1|\boldsymbol{X}, \boldsymbol{S}). \tag{2}$$

Denote the above selection probability by $\pi(\boldsymbol{X}, \boldsymbol{S})$, which is also known as the propensity score (Rosenbaum and Rubin 1983). The IPW estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ solves

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\widehat{\pi}(\boldsymbol{X}_i, \boldsymbol{S}_i)}\boldsymbol{U}(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = \boldsymbol{0}, \tag{3}$$

Peisong Han is Assistant Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada (E-mail: *peisonghan@uwaterloo.ca*). We wish to thank the editor, the associate editor, and two reviewers for their valuable comments, which have helped improve the quality of this article.

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2014.880058

where $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$ is some estimated value of $\pi(\boldsymbol{X}, \boldsymbol{S})$. It is easily seen that $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ is consistent only if $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled. In addition, $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ may not have enough estimation efficiency, as it does not fully extract the information contained in the auxiliary variables. To improve the efficiency over $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$, Robins, Rotnitzky, and Zhao (1994) proposed a class of augmented IPW (AIPW) estimators and found that, for a fixed $\boldsymbol{U}(Y, \boldsymbol{X}, \boldsymbol{\beta})$, the maximum efficiency is achieved by the estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$ solving

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i}{\widehat{\pi}(\boldsymbol{X}_i, \boldsymbol{S}_i)}\boldsymbol{U}(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}) - \frac{R_i - \widehat{\pi}(\boldsymbol{X}_i, \boldsymbol{S}_i)}{\widehat{\pi}(\boldsymbol{X}_i, \boldsymbol{S}_i)}\widehat{\boldsymbol{U}}(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\beta})\right\} = \mathbf{0} \tag{4}$$

when both $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$ are correctly modeled. Here, $\widehat{\boldsymbol{U}}(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\beta}) = \boldsymbol{D}(\boldsymbol{X}, \boldsymbol{\beta})\{\widehat{a}(\boldsymbol{X}, \boldsymbol{S}) - \mu(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}$ and $\widehat{a}(\boldsymbol{X}, \boldsymbol{S})$ is some estimated value of $E(Y|\boldsymbol{X}, \boldsymbol{S})$. The second term in (4) is called the augmentation term and extracts more information from the auxiliary variables by modeling $E(Y|\boldsymbol{X}, \boldsymbol{S})$. See also Robins, Rotnitzky, and Zhao (1995), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), van der Laan and Robins (2003), Tsiatis (2006) and Rotnitzky (2008). Scharfstein, Rotnitzky, and Robins (1999) noted that $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$ is doubly robust, in the sense that $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$ is consistent if either $\pi(\boldsymbol{X}, \boldsymbol{S})$ or $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled, but not necessarily both. Due to this double protection on consistency against model misspecification, $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$ was advocated for routine use (Bang and Robins 2005). Recently, a variety of doubly robust estimators have been proposed and studied by Tan (2006, 2008, 2010), Kang and Schafer (2007) and its discussion, Qin and Zhang (2007), Qin, Shao, and Zhang (2008), Rubin and van der Laan (2008), Cao, Tsiatis, and Davidian (2009), Tsiatis, Davidian, and Cao (2011), Han (2012), and Rotnitzky et al. (2012). Most of these developments are based on efficiency concerns. van der Laan and Gruber (2010) established the "collaborative" double robustness property of the estimators that solve the efficient influence curve estimating equation. Their results indicate that consistency can still be achieved under certain combinations of dual misspecifications of the models for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$.

Double robustness, however, does not provide sufficient protection on estimation consistency in many practical studies. It allows only a single model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and a single model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$. With an unknown data-generating process, it is often risky to assume that one of these two models is correctly specified. To increase the likelihood of correct specification, multiple models may be fitted. The question is how to combine these multiple models into estimation. See Robins et al. (2007) for some relevant discussion. In this article, we propose an estimator by solving a set of estimating equations similar to (3), but with the weight different from $1/\pi(\boldsymbol{X}, \boldsymbol{S})$. We allow multiple models for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and multiple models for $E(Y|\boldsymbol{X}, \boldsymbol{S})$. Under the MAR mechanism, the proposed estimator is consistent if any one of those multiple models is correctly specified. Therefore, our estimator provides more protection on estimation consistency. When both $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$ are correctly modeled, our estimator has the same smallest asymptotic variance as that of $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$, without requiring the knowledge of exactly which two models are correct. The development in this article is an extension of Han and Wang (2013) and Han (2014), who only considered the setting of estimating the population mean of a response variable with incomplete data. Our proof of the consistency relies on the empirical likelihood theory (Owen 1988, 2001; Qin and Lawless 1994), which has become a popular tool in solving missing-data problems (Tan 2006, 2010; Qin and Zhang 2007; Chen, Leung, and Qin 2008; Qin, Shao, and Zhang 2008; Qin, Zhang, and Leung 2009; Wang and Chen 2009).

Another desirable property possessed by our proposed estimator is that, the estimator is robust against near-zero values of $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$. When $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$ is close to zero for some subjects whose outcome is observed, the weight $R/\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$ used by the IPW and the AIPW estimators becomes extremely large, leading to skewed and highly variable sampling distribution of those estimators (Robins, Rotnitzky, and Zhao 1995; Scharfstein, Rotnitzky, and Robins 1999; Robins and Wang 2000). There have been some recent developments trying to overcome this difficulty, including Kang and Schafer (2007), Robins et al. (2007), and Cao, Tsiatis, and Davidian (2009). See also Tan (2010) and Rotnitzky et al. (2012). Especially, Kang and Schafer (2007) and Robins et al. (2007) suggested an estimator, denoted by $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW\text{-}B}}$ in this article, by solving

$$\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\boldsymbol{U}}(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\beta}) + \frac{R_i/\widehat{\pi}(\boldsymbol{X}_i, \boldsymbol{S}_i)}{n^{-1}\sum_{i=1}^{n} R_i/\widehat{\pi}(\boldsymbol{X}_i, \boldsymbol{S}_i)} \times \{\boldsymbol{U}(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}) - \widehat{\boldsymbol{U}}(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\beta})\}\right] = \mathbf{0}. \tag{5}$$

Note that the weights in (5) are regularized to sum-to-one to reduce the impact of large values of $R/\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$. Similar to $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW\text{-}B}}$ is also doubly robust. The simulation studies in Robins et al. (2007) demonstrated that $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW\text{-}B}}$ has better performance than $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$ when some values of $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$ are close to zero. Since our proposed estimator uses weight different from $R/\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$, its performance will not be affected dramatically by small $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$.

This article is organized as follows. The proposed multiply robust estimator is introduced in Section 2. Section 3 presents the numerical implementation. Section 4 focuses on the multiple robustness and other properties. The asymptotic distribution and estimation efficiency are derived and discussed in Section 5. Sections 6 and 7 cover the numerical studies and data application, respectively. Some relevant discussions are given in Section 8. The Appendix contains some technical details.

2. THE PROPOSED ESTIMATOR

Hereafter, in our notations we occasionally suppress the dependence on data, which should not cause any confusion. Let $\mathcal{P} = \{\pi^j(\boldsymbol{\alpha}^j) : j = 1, \ldots, J\}$ and $\mathcal{A} = \{a^k(\boldsymbol{\gamma}^k) : k = 1, \ldots, K\}$ denote the two sets of multiple models for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$, respectively, where $\boldsymbol{\alpha}^j$ and $\boldsymbol{\gamma}^k$ are the corresponding parameters. The fact that $E\{E(Y|\boldsymbol{X}, \boldsymbol{S})|\boldsymbol{X}\} = E(Y|\boldsymbol{X}) = \mu(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)$ can serve as a guideline in postulating models for $E(Y|\boldsymbol{X}, \boldsymbol{S})$. We use $\widehat{\boldsymbol{\alpha}}^j$ and $\widehat{\boldsymbol{\gamma}}^k$ to denote the estimators of $\boldsymbol{\alpha}^j$ and $\boldsymbol{\gamma}^k$, respectively. Usually, $\widehat{\boldsymbol{\alpha}}^j$ is taken to be the maximizer of the binomial likelihood

$$\prod_{i=1}^{n}\{\pi_i^j(\boldsymbol{\alpha}^j)\}^{R_i}\{1 - \pi_i^j(\boldsymbol{\alpha}^j)\}^{1-R_i}. \tag{6}$$

From (2), we have $Y \perp R|(\boldsymbol{X}, \boldsymbol{S})$, and thus $E(Y|\boldsymbol{X}, \boldsymbol{S}) = E(Y|R = 1, \boldsymbol{X}, \boldsymbol{S})$. Therefore, $\widehat{\boldsymbol{\gamma}}^k$ can be derived by fitting the model $a^k(\boldsymbol{\gamma}^k)$ based on complete-case analysis. Let $\widehat{\boldsymbol{\beta}}^k$ denote the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{D}(\boldsymbol{X}_i, \boldsymbol{\beta})\left\{R_i Y_i + (1 - R_i)a_i^k(\widehat{\boldsymbol{\gamma}}^k) - \mu\left(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right\} = \boldsymbol{0}.$$

In other words, $\widehat{\boldsymbol{\beta}}^k$ is an estimator of $\boldsymbol{\beta}_0$, and is calculated in a way similar to what we would do had the full data been available, but with the missing outcomes substituted by the fitted values based on model $a^k(\boldsymbol{\gamma}^k)$.

To proceed, let $m = \sum_{i=1}^{n} R_i$ be the number of subjects who have their outcome observed, and index those subjects by $i = 1, \ldots, m$ without loss of generality. In other words, we have $R_1 = \cdots = R_m = 1$ and $R_{m+1} = \cdots = R_n = 0$. Note that, when the data are fully observed, each subject receives equal weight $1/n$ to contribute to the estimating equations in (1), and when Y is subject to missingness, the IPW method weights each subject with fully observed data by $1/\{n\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})\}$ as in (3). The inverse probability weights in the latter case help to recover the population information based on the biased sample $\{i : i = 1, \ldots, m\}$.

Let $w(\boldsymbol{X}, \boldsymbol{S}) = 1/\pi(\boldsymbol{X}, \boldsymbol{S})$. It is easy to verify that

$$E(w(\boldsymbol{X}, \boldsymbol{S})[\pi^j(\boldsymbol{\alpha}^j) - E\{\pi^j(\boldsymbol{\alpha}^j)\}]|R = 1) = 0 \quad (j = 1, \ldots, J), \tag{7}$$

$$E(w(\boldsymbol{X}, \boldsymbol{S})[\boldsymbol{U}^k(\boldsymbol{\beta}, \boldsymbol{\gamma}^k) - E\{\boldsymbol{U}^k(\boldsymbol{\beta}, \boldsymbol{\gamma}^k)\}]|R = 1) = \boldsymbol{0} \quad (k = 1, \ldots, K), \tag{8}$$

where $\boldsymbol{U}^k(\boldsymbol{\beta}, \boldsymbol{\gamma}^k) = \boldsymbol{D}(\boldsymbol{X}, \boldsymbol{\beta})\{a^k(\boldsymbol{\gamma}^k) - \mu(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}$. Therefore, we define weights $w_i, i = 1, \ldots, m$, by imposing the following constraints:

$$\begin{aligned}
&w_i \geq 0 \quad (i = 1, \ldots, m),\\
&\sum_{i=1}^{m} w_i\left\{\pi_i^j(\widehat{\boldsymbol{\alpha}}^j) - \theta^j(\widehat{\boldsymbol{\alpha}}^j)\right\} = 0 \quad (j = 1, \ldots, J),\\
&\sum_{i=1}^{m} w_i\left\{\boldsymbol{U}_i^k(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{\gamma}}^k) - \boldsymbol{\eta}^k(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{\gamma}}^k)\right\} = \boldsymbol{0} \quad (k = 1, \ldots, K),
\end{aligned}$$

where $\theta^j(\boldsymbol{\alpha}^j) = n^{-1}\sum_{i=1}^{n}\pi_i^j(\boldsymbol{\alpha}^j)$ and $\boldsymbol{\eta}^k(\boldsymbol{\beta}, \boldsymbol{\gamma}^k) = n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i^k(\boldsymbol{\beta}, \boldsymbol{\gamma}^k)$. It is clear that the above constraints are the empirical version of (7) and (8). To avoid the trivial case that any w_i satisfying the above constraints still satisfy them after being multiplied by a positive constant, we divide w_i by $\sum_{i=1}^{m} w_i$ as a regularization. Therefore, the final constraints we put on w_i are

$$\begin{aligned}
&w_i \geq 0 \quad (i = 1, \ldots, m), \quad \sum_{i=1}^{m} w_i = 1,\\
&\sum_{i=1}^{m} w_i\left\{\pi_i^j(\widehat{\boldsymbol{\alpha}}^j) - \theta^j(\widehat{\boldsymbol{\alpha}}^j)\right\} = 0 \quad (j = 1, \ldots, J),\\
&\sum_{i=1}^{m} w_i\left\{\boldsymbol{U}_i^k(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{\gamma}}^k) - \boldsymbol{\eta}^k(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{\gamma}}^k)\right\} = \boldsymbol{0} \quad (k = 1, \ldots, K).
\end{aligned} \tag{9}$$

We propose to construct the estimating equations for $\boldsymbol{\beta}_0$ using the w_i that maximize $\prod_{i=1}^{m} w_i$ subject to the constraints in (9).

The derivation of such w_i pertains to a constrained maximization problem. Write $\widehat{\boldsymbol{\alpha}}^{\mathrm{T}} = \{(\widehat{\boldsymbol{\alpha}}^1)^{\mathrm{T}}, \ldots, (\widehat{\boldsymbol{\alpha}}^J)^{\mathrm{T}}\}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{T}} = \{(\widehat{\boldsymbol{\beta}}^1)^{\mathrm{T}}, \ldots, (\widehat{\boldsymbol{\beta}}^K)^{\mathrm{T}}\}$, $\widehat{\boldsymbol{\gamma}}^{\mathrm{T}} = \{(\widehat{\boldsymbol{\gamma}}^1)^{\mathrm{T}}, \ldots, (\widehat{\boldsymbol{\gamma}}^K)^{\mathrm{T}}\}$, and

$$\begin{aligned}
\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})^{\mathrm{T}} = \big[&\pi_i^1(\widehat{\boldsymbol{\alpha}}^1) - \theta^1(\widehat{\boldsymbol{\alpha}}^1), \ldots, \pi_i^J(\widehat{\boldsymbol{\alpha}}^J) - \theta^J(\widehat{\boldsymbol{\alpha}}^J),\\
&\left\{\boldsymbol{U}_i^1(\widehat{\boldsymbol{\beta}}^1, \widehat{\boldsymbol{\gamma}}^1) - \boldsymbol{\eta}^1(\widehat{\boldsymbol{\beta}}^1, \widehat{\boldsymbol{\gamma}}^1)\right\}^{\mathrm{T}}, \ldots, \left\{\boldsymbol{U}_i^K(\widehat{\boldsymbol{\beta}}^K, \widehat{\boldsymbol{\gamma}}^K)\right.\\
&\left. - \boldsymbol{\eta}^K(\widehat{\boldsymbol{\beta}}^K, \widehat{\boldsymbol{\gamma}}^K)\right\}^{\mathrm{T}}\big].
\end{aligned}$$

Using the Lagrange multipliers method, which is routinely employed in the empirical likelihood literature, it is easy to show that the w_i maximizing $\prod_{i=1}^{m} w_i$ subject to the constraints in (9) are given by

$$\widehat{w}_i = \frac{1}{m}\frac{1}{1 + \widehat{\boldsymbol{\rho}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})} \quad (i = 1, \ldots, m),$$

where $\widehat{\boldsymbol{\rho}}^{\mathrm{T}} = (\widehat{\rho}_1, \ldots, \widehat{\rho}_{J+pK})$ is the $(J + pK)$-dimensional Lagrange multipliers solving

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})}{1 + \boldsymbol{\rho}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})} = \boldsymbol{0}. \tag{10}$$

Because of the nonnegativity of $\widehat{w}_i$, $\widehat{\boldsymbol{\rho}}$ must satisfy

$$1 + \widehat{\boldsymbol{\rho}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) > 0 \quad (i = 1, \ldots, m). \tag{11}$$

Our proposed estimator of $\boldsymbol{\beta}_0$, denoted by $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$, is the solution to

$$\sum_{i=1}^{m}\widehat{w}_i\boldsymbol{U}(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = \boldsymbol{0}. \tag{12}$$

Therefore, we follow the same "weighting" approach as the IPW method, but use $\widehat{w}_i$ as our weights for the complete cases instead of $1/\{n\widehat{\pi}(\boldsymbol{X}_i, \boldsymbol{S}_i)\}$.

3. NUMERICAL IMPLEMENTATION

The Lagrange multipliers $\widehat{\boldsymbol{\rho}}$ is essential in the calculation of $\widehat{w}_i$. Directly searching for the root of (10) that satisfies (11) is not the ideal way to find $\widehat{\boldsymbol{\rho}}$. To proceed, define $\boldsymbol{F}_n(\boldsymbol{\rho}) = -n^{-1}\sum_{i=1}^{n} R_i \log\{1 + \boldsymbol{\rho}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})\}$. It is easy to see that $\boldsymbol{F}_n(\boldsymbol{\rho})$ is a strictly convex function. In the Appendix, we show that $\boldsymbol{F}_n(\boldsymbol{\rho})$ has a unique minimizer. This minimizer naturally satisfies (11), and should solve the equation $\partial\boldsymbol{F}_n(\boldsymbol{\rho})/\partial\boldsymbol{\rho} = \boldsymbol{0}$, which turns out to be (10). Therefore, $\widehat{\boldsymbol{\rho}}$ is actually the minimizer of $\boldsymbol{F}_n(\boldsymbol{\rho})$, and this fact significantly simplifies the searching for $\widehat{\boldsymbol{\rho}}$. To be complete, we provide a modified Newton–Raphson algorithm for the calculation, which is similar to the algorithm discussed by Chen, Sitter, and Wu (2002). Hereafter, let $\| \boldsymbol{B} \| = (\sum_{i,j} B_{ij}^2)^{1/2}$ for any matrix $\boldsymbol{B}$.

Algorithm

Step 0. Let $\widehat{\boldsymbol{\rho}}^{(0)} = \boldsymbol{0}$. Set $l = 0$, $\tau_0 = 1$, and $\epsilon = 10^{-8}$.

Step 1. Calculate

$$\boldsymbol{\Delta}_1\{\widehat{\boldsymbol{\rho}}^{(l)}\} = \sum_{i=1}^{m}\frac{\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})}{1 + \{\widehat{\boldsymbol{\rho}}^{(l)}\}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})},$$

$$\boldsymbol{\Delta}_2\{\widehat{\boldsymbol{\rho}}^{(l)}\} = \left(-\sum_{i=1}^{m}\frac{\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{g}}_i^{\mathrm{T}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})}{[1 + \{\widehat{\boldsymbol{\rho}}^{(l)}\}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})]^2}\right)^{-1} \times \boldsymbol{\Delta}_1\{\widehat{\boldsymbol{\rho}}^{(l)}\}.$$

If $\| \boldsymbol{\Delta}_2\{\widehat{\boldsymbol{\rho}}^{(l)}\} \| < \epsilon$, let $\widehat{\boldsymbol{\rho}} = \widehat{\boldsymbol{\rho}}^{(l)}$ and stop the algorithm; otherwise go to Step 2.

Step 2. Calculate $\boldsymbol{\delta}_l = \tau_l \boldsymbol{\Delta}_2\{\widehat{\boldsymbol{\rho}}^{(l)}\}$. If $\widehat{\boldsymbol{\rho}}^{(l)} - \boldsymbol{\delta}_l$ does not satisfy (11) or makes $\boldsymbol{F}_n\{\widehat{\boldsymbol{\rho}}^{(l)} - \boldsymbol{\delta}_l\} > \boldsymbol{F}_n\{\widehat{\boldsymbol{\rho}}^{(l)}\}$, let $\tau_l = \tau_l/2$ and repeat Step 2.

Step 3. Set $\widehat{\boldsymbol{\rho}}^{(l+1)} = \widehat{\boldsymbol{\rho}}^{(l)} - \boldsymbol{\delta}_l$, $l = l + 1$, and $\tau_l = l^{-1/2}$. Go to Step 1.

The proof of the convergence of the above algorithm can be given by following Chen, Sitter, and Wu (2002). The step size $\tau_l = l^{-1/2}$ in Step 3 is for the purpose of proving the convergence. The numerical behavior of the algorithm is not affected much by using $\tau_l = 1$ instead.

4. MULTIPLE ROBUSTNESS AND OTHER PROPERTIES

4.1 Multiple Robustness

Now, suppose that $\mathcal{P}$ contains a correctly specified model for $\pi(\boldsymbol{X}, \boldsymbol{S})$. Without loss of generality, let $\pi^1(\boldsymbol{\alpha}^1)$ be this model, and let $\boldsymbol{\alpha}_0^1$ denote the true value of $\boldsymbol{\alpha}^1$ so that $\pi^1(\boldsymbol{\alpha}_0^1) = \pi(\boldsymbol{X}, \boldsymbol{S})$. If we can show that $\boldsymbol{\beta}_0$ is the solution to (12) as $n \to \infty$, then $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ is a consistent estimator of $\boldsymbol{\beta}_0$. We employ the empirical likelihood theory to give the proof.

First, we build the connection between $\widehat{w}_i$ and another version of empirical likelihood on the biased sample $\{i : i = 1, \ldots, m\}$ by explicitly using the knowledge that $\pi^1(\boldsymbol{\alpha}^1)$ is a correctly specified model. Let p_i denote the conditional empirical probability mass on $(Y_i, \boldsymbol{X}_i, \boldsymbol{S}_i)$ conditional on $R_i = 1, i = 1, \ldots, m$. Based on (7), (8) and the fact that $w(\boldsymbol{X}, \boldsymbol{S}) = 1/\pi^1(\boldsymbol{\alpha}_0^1)$, the most plausible value for p_i should be given through the following constrained optimization:

$$\max_{p_1,\ldots,p_m} \prod_{i=1}^m p_i \quad \text{subject to} \quad p_i \geq 0 \quad (i = 1, \ldots m),$$

$$\sum_{i=1}^m p_i = 1,$$

$$\sum_{i=1}^m p_i \{\pi_i^j(\widehat{\boldsymbol{\alpha}}^j) - \theta^j(\widehat{\boldsymbol{\alpha}}^j)\}/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1) = 0 \quad (j = 1, \ldots, J),$$

$$\sum_{i=1}^m p_i \{\boldsymbol{U}_i^k(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{\gamma}}^k) - \boldsymbol{\eta}^k(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{\gamma}}^k)\}/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1) = \mathbf{0} \quad (k = 1, \ldots, K).$$

Here, the first two constraints make p_i empirical probabilities, and the last two constraints are the empirical version of (7) and (8). Using the Lagrange multipliers method again, we have

$$\widehat{p}_i = \frac{1}{m} \frac{1}{1 + \widehat{\boldsymbol{\lambda}}^{\mathrm{T}} \widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)} \quad (i = 1, \ldots, m),$$

where $\widehat{\boldsymbol{\lambda}}^{\mathrm{T}} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{J+pK})$ is the $(J + pK)$-dimensional Lagrange multipliers solving

$$\frac{1}{m} \sum_{i=1}^m \frac{\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}{1 + \boldsymbol{\lambda}^{\mathrm{T}} \widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)} = \mathbf{0}. \tag{13}$$

Due to the nonnegativity of $\widehat{p}_i$, $\widehat{\boldsymbol{\lambda}}$ must satisfy $1 + \widehat{\boldsymbol{\lambda}}^{\mathrm{T}} \widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1) > 0$ for $i = 1, \ldots, m$. Since

$$\begin{aligned}
&\frac{1}{m} \sum_{i=1}^m \frac{\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}{1 + \boldsymbol{\lambda}^{\mathrm{T}} \widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)} \\
&= \frac{1}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)} \frac{1}{m} \sum_{i=1}^m \frac{\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})}{1 + \frac{\pi_i^1(\widehat{\boldsymbol{\alpha}}^1) - \theta^1(\widehat{\boldsymbol{\alpha}}^1)}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)} + \left\{\frac{\boldsymbol{\lambda}}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)}\right\}^{\mathrm{T}} \widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})} \\
&= \frac{1}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)} \frac{1}{m} \sum_{i=1}^m \frac{\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})}{1 + \left\{\frac{\lambda_1 + 1}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)}, \frac{\lambda_2}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)}, \ldots, \frac{\lambda_{J+pK}}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)}\right\} \widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})},
\end{aligned}$$

the $\widehat{\boldsymbol{\rho}}$ solving (10) is actually given by $\widehat{\rho}_1 = (\widehat{\lambda}_1 + 1)/\theta^1(\widehat{\boldsymbol{\alpha}}^1)$ and $\widehat{\rho}_l = \widehat{\lambda}_l/\theta^1(\widehat{\boldsymbol{\alpha}}^1)$, $l = 2, \ldots, J + pK$. Therefore, we have

$$\widehat{w}_i = \frac{1}{m} \frac{\theta^1(\widehat{\boldsymbol{\alpha}}^1)/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}{1 + \widehat{\boldsymbol{\lambda}}^{\mathrm{T}} \widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)} = \frac{\widehat{p}_i \theta^1(\widehat{\boldsymbol{\alpha}}^1)}{\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}. \tag{14}$$

Thus, when $\pi^1(\boldsymbol{\alpha}^1)$ is a correct model, $\widehat{w}_i$ incorporates the empirical probability $\widehat{p}_i$ as an extra weight compared to $1/\{n\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)\}$ used by the IPW method.

From White (1982), we know that $\widehat{\boldsymbol{\alpha}}^j \xrightarrow{p} \boldsymbol{\alpha}_*^j$, $\widehat{\boldsymbol{\beta}}^k \xrightarrow{p} \boldsymbol{\beta}_*^k$ and $\widehat{\boldsymbol{\gamma}}^k \xrightarrow{p} \boldsymbol{\gamma}_*^k$, and $n^{1/2}(\widehat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}_*^j)$, $n^{1/2}(\widehat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}_*^j)$ and $n^{1/2}(\widehat{\boldsymbol{\gamma}}^j - \boldsymbol{\gamma}_*^j)$ are bounded in probability, where $\boldsymbol{\alpha}_*^j$, $\boldsymbol{\beta}_*^k$ and $\boldsymbol{\gamma}_*^k$ minimize the corresponding Kullback–Leibler distance between the probability distribution based on the postulated models and the one generating the data. We also have $\theta^j(\widehat{\boldsymbol{\alpha}}^j) \xrightarrow{p} \theta_*^j$ and $\boldsymbol{\eta}^k(\widehat{\boldsymbol{\beta}}^k, \widehat{\boldsymbol{\gamma}}^k) \xrightarrow{p} \boldsymbol{\eta}_*^k$, where $\theta_*^j = E\{\pi^j(\boldsymbol{\alpha}_*^j)\}$ and $\boldsymbol{\eta}_*^k = E\{\boldsymbol{U}^k(\boldsymbol{\beta}_*^k, \boldsymbol{\gamma}_*^k)\}$. In general, only when $\pi^j(\boldsymbol{\alpha}^j)$ is a correctly specified model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ do we have $\pi^j(\boldsymbol{\alpha}_*^j) = \pi(\boldsymbol{X}, \boldsymbol{S})$, and only when $a^k(\boldsymbol{\gamma}^k)$ is a correctly specified model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$ do we have $a^k(\boldsymbol{\gamma}_*^k) = E(Y|\boldsymbol{X}, \boldsymbol{S})$. Let $\boldsymbol{\rho}_*$ denote the probability limit of $\widehat{\boldsymbol{\rho}}$, and write $\boldsymbol{\alpha}_*^{\mathrm{T}} = \{(\boldsymbol{\alpha}_*^1)^{\mathrm{T}}, \ldots, (\boldsymbol{\alpha}_*^J)^{\mathrm{T}}\}$, $\boldsymbol{\beta}_*^{\mathrm{T}} = \{(\boldsymbol{\beta}_*^1)^{\mathrm{T}}, \ldots, (\boldsymbol{\beta}_*^K)^{\mathrm{T}}\}$, and $\boldsymbol{\gamma}_*^{\mathrm{T}} = \{(\boldsymbol{\gamma}_*^1)^{\mathrm{T}}, \ldots, (\boldsymbol{\gamma}_*^K)^{\mathrm{T}}\}$.

Because $\widehat{\boldsymbol{\lambda}} \xrightarrow{p} \mathbf{0}$ based on the empirical likelihood theory, the Lemma given in the Appendix yields that $\widehat{\boldsymbol{\lambda}} = O_p(n^{-1/2})$. Using the fact that $m/n \to \theta_*^1$ since $\pi^1(\boldsymbol{\alpha}^1)$ is correctly specified, we have

$$\begin{aligned}
&\sum_{i=1}^m \widehat{w}_i \boldsymbol{U}_i(\boldsymbol{\beta}_0) \\
&= \frac{\theta^1(\widehat{\boldsymbol{\alpha}}^1)}{m} \sum_{i=1}^n \frac{R_i/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}{1 + \widehat{\boldsymbol{\lambda}}^{\mathrm{T}} \widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)} \boldsymbol{U}_i(\boldsymbol{\beta}_0) \\
&= \frac{1}{n} \sum_{i=1}^n \frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_*^1)} \boldsymbol{U}_i(\boldsymbol{\beta}_0) + o_p(1) \xrightarrow{p} E\left\{\frac{R}{\pi(\boldsymbol{X}, \boldsymbol{S})} \boldsymbol{U}(\boldsymbol{\beta}_0)\right\} = \mathbf{0}.
\end{aligned}$$

Therefore, $\boldsymbol{\beta}_0$ is the solution to (12) as $n \to \infty$, and this implies the consistency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$.

Since both $\widehat{p}_i$ and $\widehat{w}_i, i = 1, \ldots, m$, are nonnegative and sum-to-one, both can be considered as an empirical likelihood with support $\{(Y_i, \boldsymbol{X}_i, \boldsymbol{S}_i) : i = 1, \ldots, m\}$. However, $\widehat{p}_i$ is derived assuming that we know $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled by $\pi^1(\boldsymbol{\alpha}^1)$, whereas $\widehat{w}_i$ is derived without distinguishing the correct and incorrect models for $\pi(\boldsymbol{X}, \boldsymbol{S})$. From the imposed constraints on $\widehat{p}_i$ and $\widehat{w}_i$, it can be seen that $\widehat{p}_i$ is an extra weight on subject i in addition to $1/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)$, $i = 1, \ldots, m$, whereas $\widehat{w}_i$ is the sole weight. Since the correct weight $1/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)$ for recovering the population information has been explicitly accounted for, $\widehat{p}_i$

should not deviate much from the discrete uniform likelihood in order that $\widehat{p}_i/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)$ is still a correct weight to be assigned to subject i. Indeed, $\widehat{p}_i = m^{-1}\{1 + O_p(n^{-1/2})\}$ is essentially equal to $1/m$ with a negligible higher-order perturbation. As for $\widehat{w}_i$, although it is derived without knowing which model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly specified, it is capable of automatically distinguishing the correct model and taking it into account, if a correct model is indeed included in $\mathcal{P}$, as illustrated by the numerical equivalence between $\widehat{w}_i$ and $\widehat{p}_i\theta^1(\widehat{\boldsymbol{\alpha}}^1)/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)$ in (14). It is this implicit incorporation of $1/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)$ that ensures the consistency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ when one model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly specified.

Now suppose that $\mathcal{A}$ contains a correctly specified model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$. Without loss of generality, let $a^1(\boldsymbol{\gamma}^1)$ be this model and let $\boldsymbol{\gamma}_0^1$ denote the true value of $\boldsymbol{\gamma}^1$ so that $a^1(\boldsymbol{\gamma}_0^1) = E(Y|\boldsymbol{X}, \boldsymbol{S})$. We then have $\boldsymbol{\gamma}_*^1 = \boldsymbol{\gamma}_0^1$. Note that one of the constraints in (9) is actually

$$\sum_{i=1}^{m} \widehat{w}_i \boldsymbol{U}_i^1(\widehat{\boldsymbol{\beta}}^1, \widehat{\boldsymbol{\gamma}}^1) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_i^1(\widehat{\boldsymbol{\beta}}^1, \widehat{\boldsymbol{\gamma}}^1).$$

In addition, it is easy to see that now $\widehat{\boldsymbol{\beta}}^1$ has the probability limit $\boldsymbol{\beta}_*^1 = \boldsymbol{\beta}_0$. Therefore, the right-hand side of the above constraint, hence the left-hand side as well, converges to $\mathbf{0}$ in probability. Moreover, the left-hand side of (12) evaluated at $\boldsymbol{\beta}_0$ should be close to the left-hand side of the above constraint as n increases, because $a^1(\boldsymbol{\gamma}^1)$ is correctly specified. This fact ensures that $\boldsymbol{\beta}_0$ is the solution to (12) as $n \to \infty$, and thus the consistency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$.

More formally, write

$$\begin{aligned}&\boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)^{\mathrm{T}}\\ &\quad= \big[\pi^1(\boldsymbol{\alpha}_*^1) - \theta_*^1, \ldots, \pi^J(\boldsymbol{\alpha}_*^J) - \theta_*^J,\\ &\qquad\{\boldsymbol{U}^1(\boldsymbol{\beta}_*^1, \boldsymbol{\gamma}_*^1) - \boldsymbol{\eta}_*^1\}^{\mathrm{T}}, \ldots, \{\boldsymbol{U}^K(\boldsymbol{\beta}_*^K, \boldsymbol{\gamma}_*^K) - \boldsymbol{\eta}_*^K\}^{\mathrm{T}}\big]. \quad (15)\end{aligned}$$

Using the fact that $Y \perp R|(\boldsymbol{X}, \boldsymbol{S})$, we have

$$\begin{aligned}&\sum_{i=1}^{m} \widehat{w}_i \boldsymbol{U}_i(\boldsymbol{\beta}_0)\\ &= \sum_{i=1}^{m} \widehat{w}_i\{\boldsymbol{U}_i(\boldsymbol{\beta}_0) - \boldsymbol{U}_i^1(\widehat{\boldsymbol{\beta}}^1, \widehat{\boldsymbol{\gamma}}^1)\} + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_i^1(\widehat{\boldsymbol{\beta}}^1, \widehat{\boldsymbol{\gamma}}^1)\\ &= \frac{1}{m}\sum_{i=1}^{n} \frac{R_i\{\boldsymbol{U}_i(\boldsymbol{\beta}_0) - \boldsymbol{U}_i^1(\widehat{\boldsymbol{\beta}}^1, \widehat{\boldsymbol{\gamma}}^1)\}}{1 + \widehat{\boldsymbol{\rho}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})} + E\{\boldsymbol{U}^1(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0^1)\} + o_p(1)\\ &\xrightarrow{p} \frac{1}{P(R=1)} E\left[\frac{R\{\boldsymbol{U}(\boldsymbol{\beta}_0) - \boldsymbol{U}^1(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0^1)\}}{1 + \boldsymbol{\rho}_*^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)}\right]\\ &= \frac{1}{P(R=1)} E\left(E\left[\frac{R\{\boldsymbol{U}(\boldsymbol{\beta}_0) - \boldsymbol{U}^1(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0^1)\}}{1 + \boldsymbol{\rho}_*^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)} \,\middle|\, R, \boldsymbol{X}, \boldsymbol{S}\right]\right) = \mathbf{0}.\end{aligned}$$

Summarizing all the above results, we have the following theorem, which implies the multiple robustness of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$.

Theorem 1. When $\mathcal{P}$ contains a correctly specified model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ or $\mathcal{A}$ contains a correctly specified model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$, as $n \to \infty$, we have that $\sum_{i=1}^{m} \widehat{w}_i \boldsymbol{U}_i(\boldsymbol{\beta}_0) \xrightarrow{p} \mathbf{0}$.

4.2 Robustness Against Near-Zero $\widehat{\pi}$ $(\boldsymbol{X}, \boldsymbol{S})$

One critique of the IPW and the AIPW estimators is their sensitivity to near-zero values of $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$, which can yield extremely large weight $R/\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$, and thus jeopardize the numerical performance. In this case, through simulation studies, Kang and Schafer (2007) demonstrated that $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$ can have severe bias even if the models for both $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$ are only mildly misspecified. Our estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ does not have this problem. Since $\widehat{w}_i$ maximize $\prod_{i=1}^{m} w_i$ subject to the constraints in (9), the occurrence of extreme values of $\widehat{w}_i$ is prevented. This is because, when w_i are restricted to be nonnegative and sum-to-one, $\prod_{i=1}^{m} w_i$ increases if the values of w_i become more evenly distributed rather than concentrating on a few subjects. Through maximizing $\prod_{i=1}^{m} w_i$, $\widehat{w}_i$ are the most evenly distributed values for w_i that satisfy (9). Therefore, it is unlikely that some subjects receive extremely large weights that dominate others. When $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled so that $\widehat{w}_i$ can be expressed as in (14), it is actually the extra weights $\widehat{p}_i$ that help to mitigate the impact of large values of $1/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)$ on the numerical performance of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$. This is because, the incorporation of $\widehat{p}_i$ ensures that $\widehat{w}_i$ sum to one, implicitly achieving the same goal of regularizing the weights as done in (5). It is worth to mention that the idea of regularizing the weights to improve the numerical performance was also taken by Cao, Tsiatis, and Davidian (2009), who proposed to explicitly impose a constraint $\sum_{i=1}^{m} 1/\widehat{\pi}(\boldsymbol{X}_i, \boldsymbol{S}_i) = n$ to drive the near-zero values of $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$ away from zero. Our proposed method achieves a similar result through maximizing $\prod_{i=1}^{m} w_i$.

5. ASYMPTOTIC DISTRIBUTION AND EFFICIENCY

We derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ by following the typical approach taken to develop the semiparametric theory for missing-data analysis (Robins, Rotnitzky, and Zhao 1994, 1995; Rotnitzky and Robins 1995; Tsiatis 2006); that is, by assuming $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled. In this case, $\widehat{w}_i$ have the form (14). The fact that $\widehat{\boldsymbol{\lambda}} \xrightarrow{p} \mathbf{0}$ and the asymptotic expansion of $n^{1/2}\widehat{\boldsymbol{\lambda}}$ given by the lemma in the Appendix facilitate our derivation, yielding an explicit formula of the asymptotic variance that helps us to assess the efficiency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ and identify possible ways to improve the efficiency. The case where $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled is also of practical importance. In many studies, including the two-stage sampling design studies (e.g., Pepe 1992; Pepe, Reilly, and Fleming 1994), the missingness is determined by the investigator, and thus $\pi(\boldsymbol{X}, \boldsymbol{S})$ is completely known or can be correctly modeled. The asymptotic variance in this case may be used for inference. When no model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly specified, the derivation of the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ requires the asymptotic expansion of $n^{1/2}(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}_*)$, which should be established by taking the Taylor expansion of (10) at $\boldsymbol{\rho}_*$ for the $\boldsymbol{\rho}$ argument. However, since $\boldsymbol{\rho}_*$ is unknown, this Taylor expansion, hence the asymptotic expansion of $n^{1/2}(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}_*)$, is not very informative. As a result, it is difficult to assess the efficiency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ based on the asymptotic distribution. Besides, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ varies in this case depending on which model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is correctly specified. However, if one had such information available, one would estimate $\boldsymbol{\beta}_0$ by directly solving (1) with Y substituted by the fitted value based

on the correct model, due to the high efficiency of this estimator (e.g., Tan 2007). Therefore, there is of little practical interest in deriving the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ when no model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly specified.

Without loss of generality, let $\pi^1(\boldsymbol{\alpha}^1)$ be a correctly specified model for $\pi(\boldsymbol{X}, \boldsymbol{S})$, and let $\boldsymbol{\Psi}(\boldsymbol{\alpha}^1)$ denote the score function of $\boldsymbol{\alpha}^1$ corresponding to the binomial likelihood (6); that is,

$$\boldsymbol{\Psi}(\boldsymbol{\alpha}^1) = \frac{R - \pi^1(\boldsymbol{\alpha}^1)}{\pi^1(\boldsymbol{\alpha}^1)\{1 - \pi^1(\boldsymbol{\alpha}^1)\}} \frac{\partial \pi^1(\boldsymbol{\alpha}^1)}{\partial \boldsymbol{\alpha}^1}.$$

In addition, write $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\alpha}_0^1)$,

$$\begin{aligned} \boldsymbol{L} &= E\left\{ U(\boldsymbol{\beta}_0) \frac{\boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)^{\mathrm{T}}}{\pi^1(\boldsymbol{\alpha}_0^1)} \right\}, \\ \boldsymbol{G} &= E\left\{ \frac{\boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)^{\otimes 2}}{\pi^1(\boldsymbol{\alpha}_0^1)} \right\}, \\ \boldsymbol{Q}(\boldsymbol{\alpha}^1) &= \frac{R}{\pi^1(\boldsymbol{\alpha}^1)} U(\boldsymbol{\beta}_0) - \frac{R - \pi^1(\boldsymbol{\alpha}^1)}{\pi^1(\boldsymbol{\alpha}^1)} \boldsymbol{L}\boldsymbol{G}^{-1} \boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*), \end{aligned} \tag{16}$$

and $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\alpha}_0^1)$, where for any matrix $\boldsymbol{B}$, $\boldsymbol{B}^{\otimes 2} = \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$. Since $\widehat{\boldsymbol{\alpha}}^1$ is the maximizer of the binomial likelihood (6) with $j = 1$, we have the following asymptotic expansion:

$$n^{1/2}(\widehat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}_0^1) = n^{-1/2} \sum_{i=1}^{n} \{E(\boldsymbol{\Psi}^{\otimes 2})\}^{-1} \boldsymbol{\Psi}_i + o_p(1).$$

The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ is given by the following theorem, the proof of which is given in the Appendix.

Theorem 2. When $\mathcal{P}$ contains a correctly specified model for $\pi(\boldsymbol{X}, \boldsymbol{S})$, $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}} - \boldsymbol{\beta}_0)$ has an asymptotic normal distribution with mean $\mathbf{0}$ and variance $\mathrm{var}(\boldsymbol{Z})$, where

$$\boldsymbol{Z} = \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\} \right]^{-1} [\boldsymbol{Q} - \{E(\boldsymbol{Q}\boldsymbol{\Psi}^{\mathrm{T}})\}\{E(\boldsymbol{\Psi}^{\otimes 2})\}^{-1}\boldsymbol{\Psi}].$$

It is interesting to note that $\boldsymbol{Z}$ involves the residual of the projection of $\boldsymbol{Q}$ on $\boldsymbol{\Psi}$. Therefore, one way to reduce $\mathrm{var}(\boldsymbol{Z})$ is to increase the dimension of $\boldsymbol{\Psi}$. For example, interactions and higher-order terms of the components of $\boldsymbol{S}$ and $\boldsymbol{X}$ could be added when fitting the model $\pi^1(\boldsymbol{\alpha}^1)$. When the missingness is caused by design and the known value of $\pi(\boldsymbol{X}, \boldsymbol{S})$ is used to calculate $\widehat{w}_i$, following arguments similar to those in the proof of Theorem 2, it can be shown that the asymptotic variance of $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}} - \boldsymbol{\beta}_0)$ becomes $\mathrm{var}(\widetilde{\boldsymbol{Z}})$, where

$$\widetilde{\boldsymbol{Z}} = \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\} \right]^{-1} \boldsymbol{Q}.$$

Apparently we have $\mathrm{var}(\widetilde{\boldsymbol{Z}}) \geq \mathrm{var}(\boldsymbol{Z})$ in the nonnegative-definite sense. Therefore, the efficiency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ may be improved by correctly modeling $\pi(\boldsymbol{X}, \boldsymbol{S})$ even if its value is completely known. This counter-intuitive phenomenon has been studied by Robins, Rotnitzky, and Zhao (1995) and Rotnitzky and Robins (1995).

Some algebra shows that $\mathrm{var}(\boldsymbol{Z}) = \boldsymbol{V}_{\mathrm{IPW}} - \boldsymbol{V}_{g,1} + \boldsymbol{V}_{g,2} - \boldsymbol{V}_{\boldsymbol{\alpha}}$, where

$$\begin{aligned} \boldsymbol{V}_{\mathrm{IPW}} &= \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\} \right]^{-1} E\left\{ \frac{U(\boldsymbol{\beta}_0)^{\otimes 2}}{\pi^1(\boldsymbol{\alpha}_0^1)} \right\} \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \right\} \right]^{-1}, \\ \boldsymbol{V}_{g,1} &= \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\} \right]^{-1} \boldsymbol{L}\boldsymbol{G}^{-1}[\boldsymbol{G} + E\{\boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)^{\otimes 2}\}] \\ &\quad \times \boldsymbol{G}^{-1}\boldsymbol{L}^{\mathrm{T}} \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \right\} \right]^{-1}, \\ \boldsymbol{V}_{g,2} &= \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\} \right]^{-1} (\boldsymbol{M} + \boldsymbol{M}^{\mathrm{T}}) \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \right\} \right]^{-1} \end{aligned}$$

with $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{G}^{-1} E\left\{ \boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*) U(\boldsymbol{\beta}_0)^{\mathrm{T}} \right\}$, and

$$\begin{aligned} \boldsymbol{V}_{\boldsymbol{\alpha}} &= \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\} \right]^{-1} E(\boldsymbol{Q}\boldsymbol{\Psi}^{\mathrm{T}})\{E(\boldsymbol{\Psi}^{\otimes 2})\}^{-1} E(\boldsymbol{\Psi}\boldsymbol{Q}^{\mathrm{T}}) \\ &\quad \times \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \right\} \right]^{-1}. \end{aligned}$$

It is easy to see that $\boldsymbol{V}_{\mathrm{IPW}}$ is the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ based on the correct model $\pi^1(\boldsymbol{\alpha}^1)$. Note that $\boldsymbol{V}_{\boldsymbol{\alpha}}$ comes from using $\widehat{\boldsymbol{\alpha}}^1$ in the calculation of $\widehat{w}_i$ instead of using the true value $\boldsymbol{\alpha}_0^1$. Since $\boldsymbol{V}_{\boldsymbol{\alpha}}$ is nonnegative definite, it leads to a variance reduction from $\boldsymbol{V}_{\mathrm{IPW}}$, as discussed above. The other models for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and the multiple models for $E(Y|\boldsymbol{X}, \boldsymbol{S})$ overall play a complex role in affecting the efficiency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$. Although $\boldsymbol{V}_{g,1}$ is nonnegative definite, yielding a variance reduction from $\boldsymbol{V}_{\mathrm{IPW}}$, no clear conclusion can be drawn for the overall effect of $\boldsymbol{V}_{g,2} - \boldsymbol{V}_{g,1}$, and thus the effect of $\boldsymbol{V}_{g,2} - \boldsymbol{V}_{g,1} - \boldsymbol{V}_{\boldsymbol{\alpha}}$. Therefore, in general the comparison of efficiency between $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ is inconclusive, similar to the comparison between $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ (e.g., Chen, Leung, and Qin 2008). However, when $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is also correctly modeled, $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ has higher efficiency than $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$, and this will become clear from the following theorem, the proof of which is given in the Appendix.

Theorem 3. When $\mathcal{P}$ contains a correctly specified model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $\mathcal{A}$ contains a correctly specified model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$, $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}} - \boldsymbol{\beta}_0)$ has an asymptotic normal distribution with mean $\mathbf{0}$ and variance $\mathrm{var}(\boldsymbol{Z}_{\mathrm{opt}})$, where

$$\begin{aligned} \boldsymbol{Z}_{\mathrm{opt}} &= \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\} \right]^{-1} \\ &\quad \times \left[\frac{R}{\pi^1(\boldsymbol{\alpha}_0^1)} U(\boldsymbol{\beta}_0) - \frac{R - \pi^1(\boldsymbol{\alpha}_0^1)}{\pi^1(\boldsymbol{\alpha}_0^1)} E\{U(\boldsymbol{\beta}_0)|\boldsymbol{X}, \boldsymbol{S}\} \right]. \end{aligned}$$

Some algebra shows that $\mathrm{var}(\boldsymbol{Z}_{\mathrm{opt}}) = \boldsymbol{V}_{\mathrm{IPW}} - \boldsymbol{V}_{g,3}$, where

$$\begin{aligned} \boldsymbol{V}_{g,3} &= \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\} \right]^{-1} E\left(\frac{1 - \pi^1(\boldsymbol{\alpha}_0^1)}{\pi^1(\boldsymbol{\alpha}_0^1)} [E\{U(\boldsymbol{\beta}_0)|\boldsymbol{X}, \boldsymbol{S}\}]^{\otimes 2} \right) \\ &\quad \times \left[E\left\{ \frac{\partial U(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \right\} \right]^{-1}. \end{aligned}$$

Apparently, we have $\mathrm{var}(\boldsymbol{Z}_{\mathrm{opt}}) \leq \boldsymbol{V}_{\mathrm{IPW}}$ due to the nonnegative definiteness of $\boldsymbol{V}_{g,3}$, and this implies that $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ has higher efficiency than $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ when $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is also correctly modeled. In fact, for a fixed $\boldsymbol{D}(\boldsymbol{X}, \boldsymbol{\beta})$, $\mathrm{var}(\boldsymbol{Z}_{\mathrm{opt}})$ corresponds to the maximum possible efficiency in estimating $\boldsymbol{\beta}_0$ with the observed data (Robins, Rotnitzky, and Zhao 1994; Chen and Breslow 2004; Yu

and Nan 2006). Robins, Rotnitzky, and Zhao (1994) showed that $\widehat{\boldsymbol{\beta}}_{\text{AIPW}}$ achieves this maximum efficiency when both $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$ are correctly modeled. Theorem 3 demonstrates that in this case $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ also achieves the maximum efficiency. However, unlike $\widehat{\boldsymbol{\beta}}_{\text{AIPW}}$, $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ does not require to know exactly which two among the multiple models are correctly specified.

For practical studies where $\pi(\boldsymbol{X}, \boldsymbol{S})$ is known by design, the asymptotic distributions derived in this section may be used for statistical inference. In this case, from Theorem 2 and the subsequent discussion following it, the estimated value of $\pi(\boldsymbol{X}, \boldsymbol{S})$ by model fitting is still preferred over the true value due to possible efficiency improvement. When $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is also correctly modeled, using the estimated value or the true value of $\pi(\boldsymbol{X}, \boldsymbol{S})$ will make no difference, as both ways lead to the maximum efficiency. To make inference, the expectations involved in the asymptotic variances can be estimated by the corresponding sample averages. In the general case where the correct model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ is unknown, we suggest the bootstrapping method as an alternative to calculate the standard error of $\widehat{\boldsymbol{\beta}}_{\text{MR}}$.

6. SIMULATION STUDIES

In this section, we conduct simulation experiments to evaluate the finite-sample performance of the proposed estimator $\widehat{\boldsymbol{\beta}}_{\text{MR}}$. We use the simulation setting in Tsiatis, Davidian, and Cao (2011) with some modifications. This setting produces near-zero values of $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$ when $\pi(\boldsymbol{X}, \boldsymbol{S})$ is incorrectly modeled, and thus enables us to examine the robustness of $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ against small $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$ in addition to the multiple robustness property.

The simulation model has four mutually independent covariates, $X^{(1)} \sim \text{Normal}(5, 1)$, $X^{(2)} \sim \text{Bernoulli}(0.5)$, $X^{(3)} \sim \text{Normal}(0, 1)$ and $X^{(4)} \sim \text{Normal}(0, 1)$. The outcome is generated by $Y = 3.5 + 0.5X^{(1)} + 2X^{(2)} + X^{(3)} + X^{(4)} + \epsilon_Y$. There are three auxiliary variables, $S^{(1)} = 1 + X^{(1)} - X^{(2)} + \epsilon_1$, $S^{(2)} = \mathcal{I}\{S^{(1)} + 0.3\epsilon_2 > 5.8\}$, and $S^{(3)} = \exp[\{S^{(1)}/9\}^2] + \epsilon_3$. Here, $\mathcal{I}(\cdot)$ is the indicator function, $(\epsilon_Y, \epsilon_1, \epsilon_2, \epsilon_3)^{\text{T}} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma}$ is a 4×4 matrix with diagonal entries 2, 2, 1 and 1, (1, 2)-entry and (2, 1)-entry 0.5, and all the other entries 0. The missingness mechanism is set to be $\text{logit}\{\pi(\boldsymbol{X}, \boldsymbol{S})\} = 3.5 - 5.0S^{(2)}$, under which there are approximately 37% of the subjects with missing Y. The correct models for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$ are given by $\text{logit}\{\pi^1(\boldsymbol{\alpha}^1)\} = \alpha_1^1 + \alpha_2^1 S^{(2)}$ and $a^1(\boldsymbol{\gamma}^1) = \gamma_1^1 + \gamma_2^1 X^{(1)} + \cdots + \gamma_5^1 X^{(4)} + \gamma_6^1 S^{(1)}$, respectively. The following two incorrect models are also used in our simulation experiments: $\text{logit}\{\pi^2(\boldsymbol{\alpha}^2)\} = \alpha_1^2 + \alpha_2^2 X^{(1)} + \cdots + \alpha_5^2 X^{(4)} + \alpha_6^2 S^{(1)}$ and $a^2(\boldsymbol{\gamma}^2) = \gamma_1^2 + \gamma_2^2 S^{(1)} + \gamma_3^2 S^{(2)} + \gamma_4^2 S^{(3)}$. We consider two sample sizes $n = 200$ and $n = 800$, and the results are summarized based on 2000 replications. It is easy to see that $\boldsymbol{\beta}_0^{\text{T}} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (3.5, 0.5, 2.0, 1.0, 1.0)$.

The numerical performance of $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ is summarized in Tables 1 and 2. To make comparison, we also calculate estimators $\widehat{\boldsymbol{\beta}}_{\text{IPW}}$, $\widehat{\boldsymbol{\beta}}_{\text{OR}}$, $\widehat{\boldsymbol{\beta}}_{\text{AIPW}}$, and $\widehat{\boldsymbol{\beta}}_{\text{AIPW-B}}$, where $\widehat{\boldsymbol{\beta}}_{\text{OR}}$ solves $n^{-1}\sum_{i=1}^n \boldsymbol{D}(\boldsymbol{X}_i, \boldsymbol{\beta})\{\widehat{a}(\boldsymbol{X}_i, \boldsymbol{S}_i) - \mu(\boldsymbol{X}_i^{\text{T}}\boldsymbol{\beta})\} = \mathbf{0}$. The estimator $\widehat{\boldsymbol{\beta}}_{\text{OR}}$ is usually called the outcome regression estimator, and is consistent only when $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled. Because Y is generated from a normal distribution, we take $\boldsymbol{U}(Y, \boldsymbol{X}, \boldsymbol{\beta}) = \boldsymbol{X}(Y - \boldsymbol{X}^{\text{T}}\boldsymbol{\beta})$ for all the estimators under our comparison. To distinguish all the estimators constructed based on different methods and models, each estimator is assigned a name with the form "method-0000," where each digit of the four-digit number, from left to right, indicates if $\pi^1(\boldsymbol{\alpha}^1)$, $\pi^2(\boldsymbol{\alpha}^2)$, $a^1(\boldsymbol{\gamma}^1)$, or $a^2(\boldsymbol{\gamma}^2)$ is used in the construction ("1" means yes and "0" means no), respectively.

When only one model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ is used, $\widehat{\boldsymbol{\beta}}_{\text{MR-1000}}$ based on the correct model $\pi^1(\boldsymbol{\alpha}^1)$ has almost identical performance to $\widehat{\boldsymbol{\beta}}_{\text{IPW-1000}}$, but $\widehat{\boldsymbol{\beta}}_{\text{MR-0100}}$ based on the incorrect model $\pi^2(\boldsymbol{\alpha}^2)$ clearly has smaller bias and much stabler sampling distribution compared to $\widehat{\boldsymbol{\beta}}_{\text{IPW-0100}}$. When only one model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is used, while $\widehat{\boldsymbol{\beta}}_{\text{MR-0010}}$ based on the correct model $a^1(\boldsymbol{\gamma}^1)$ has slightly larger root mean square errors (RMSE) and median absolute errors (MAE) than $\widehat{\boldsymbol{\beta}}_{\text{OR-0010}}$ as expected, since in our case $\widehat{\boldsymbol{\beta}}_{\text{OR-0010}}$ is essentially the maximum likelihood estimator, $\widehat{\boldsymbol{\beta}}_{\text{MR-0001}}$ based on the incorrect model $a^2(\boldsymbol{\gamma}^2)$ has surprisingly smaller bias, RMSE and MAE than $\widehat{\boldsymbol{\beta}}_{\text{OR-0001}}$. Actually, despite its very slight bias, the performance of $\widehat{\boldsymbol{\beta}}_{\text{MR-0001}}$ is even comparable to $\widehat{\boldsymbol{\beta}}_{\text{OR-0010}}$. Therefore, when using the same single model, $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ is superior to both $\widehat{\boldsymbol{\beta}}_{\text{IPW}}$ and $\widehat{\boldsymbol{\beta}}_{\text{OR}}$, because even if the model is misspecified, $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ still seems to provide a reasonable (at least not too bad) estimate.

When one model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and one model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$ are used, the AIPW estimators $\widehat{\boldsymbol{\beta}}_{\text{AIPW-0110}}$ and $\widehat{\boldsymbol{\beta}}_{\text{AIPW-0101}}$ based on the incorrect model $\pi^2(\boldsymbol{\alpha}^2)$ have disastrous performance due to some extremely large values of $R/\pi^2(\widehat{\boldsymbol{\alpha}}^2)$, even though $\widehat{\boldsymbol{\beta}}_{\text{AIPW-0110}}$ uses the correct model $a^1(\boldsymbol{\gamma}^1)$. Although estimators $\widehat{\boldsymbol{\beta}}_{\text{AIPW-B-0110}}$ and $\widehat{\boldsymbol{\beta}}_{\text{AIPW-B-0101}}$ significantly improve over $\widehat{\boldsymbol{\beta}}_{\text{AIPW-0110}}$ and $\widehat{\boldsymbol{\beta}}_{\text{AIPW-0101}}$, respectively, their performance is still not satisfactory, judging from the RMSE and MAE of $\widehat{\boldsymbol{\beta}}_{\text{AIPW-B-0110}}$ and the bias, RMSE and MAE of $\widehat{\boldsymbol{\beta}}_{\text{AIPW-B-0101}}$. To make things worse, as the sample size increases from 200 to 800, the performance of $\widehat{\boldsymbol{\beta}}_{\text{AIPW-B-0110}}$ and $\widehat{\boldsymbol{\beta}}_{\text{AIPW-B-0101}}$ does not improve much. Our proposed estimators $\widehat{\boldsymbol{\beta}}_{\text{MR-0110}}$ and $\widehat{\boldsymbol{\beta}}_{\text{MR-0101}}$, on the contrary, have excellent behavior. The near-zero values of $\pi^2(\widehat{\boldsymbol{\alpha}}^2)$ do not have any dramatic impact on these two estimators. More surprisingly, despite that $\widehat{\boldsymbol{\beta}}_{\text{MR-0101}}$ uses two incorrect models, it gives estimate nearly as good as $\widehat{\boldsymbol{\beta}}_{\text{MR-1010}}$ based on two correct models, except for the slight bias of the intercept. The above observations demonstrate the superiority of $\widehat{\boldsymbol{\beta}}_{\text{MR}}$, when using the same two models, over both $\widehat{\boldsymbol{\beta}}_{\text{AIPW}}$ and $\widehat{\boldsymbol{\beta}}_{\text{AIPW-B}}$, as $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ is not only more robust against near-zero values of $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$, but also more robust against model misspecification in that it still gives reasonable estimate when the two models are both incorrect. The comparison between $\widehat{\boldsymbol{\beta}}_{\text{MR-1001}}$, $\widehat{\boldsymbol{\beta}}_{\text{AIPW-1001}}$, and $\widehat{\boldsymbol{\beta}}_{\text{AIPW-B-1001}}$ reveals the high efficiency of $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ when $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled but $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is not, a scenario in which it is well known that $\widehat{\boldsymbol{\beta}}_{\text{AIPW}}$ could have low efficiency (e.g., Tan 2006, 2008, 2010; Cao, Tsiatis, and Davidian 2009; Han 2012; Rotnitzky et al. 2012). As our theory predicts, $\widehat{\boldsymbol{\beta}}_{\text{MR-1010}}$ and $\widehat{\boldsymbol{\beta}}_{\text{AIPW-1010}}$ have almost the same RMSE and MAE. Actually these two estimators achieve the maximum possible efficiency under the chosen $\boldsymbol{U}(Y, \boldsymbol{X}, \boldsymbol{\beta}) = \boldsymbol{X}(Y - \boldsymbol{X}^{\text{T}}\boldsymbol{\beta})$.

The multiple robustness of $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ is well demonstrated by the ignorable bias of $\widehat{\boldsymbol{\beta}}_{\text{MR-1100}}$, $\widehat{\boldsymbol{\beta}}_{\text{MR-0011}}$, $\widehat{\boldsymbol{\beta}}_{\text{MR-1110}}$, $\widehat{\boldsymbol{\beta}}_{\text{MR-1101}}$, $\widehat{\boldsymbol{\beta}}_{\text{MR-1011}}$, $\widehat{\boldsymbol{\beta}}_{\text{MR-0111}}$, and $\widehat{\boldsymbol{\beta}}_{\text{MR-1111}}$, for which two models for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and/or two models for $E(Y|\boldsymbol{X}, \boldsymbol{S})$ are allowed. These estimators all have relatively high efficiency judging from their RMSE and MAE. It is especially remarkable to see that $\widehat{\boldsymbol{\beta}}_{\text{MR-1100}}$, which does not include any model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$, has efficiency even comparable to $\widehat{\boldsymbol{\beta}}_{\text{MR-1010}}$. When $n = 200$, $\widehat{\boldsymbol{\beta}}_{\text{MR-1011}}$

Table 1. Comparison of different estimators with $n = 200$ based on 2000 replications. The names of the estimators have the form "method-0000" (or "method-EN" for enlarged models), with each digit of the four-digit number, from left to right, indicating if $\pi^1(\boldsymbol{\alpha}^1)$, $\pi^2(\boldsymbol{\alpha}^2)$, $a^1(\boldsymbol{\gamma}^1)$, or $a^2(\boldsymbol{\gamma}^2)$ is used, respectively. The results have been multiplied by 100

	β_1			β_2			β_3			β_4			β_5		
Estimator	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
IPW-0100	−22	137	81	7	30	18	−9	57	36	−1	27	15	0	27	15
IPW-1000	−2	86	57	0	18	12	0	35	24	0	17	12	0	17	12
IPW-EN	−4	91	58	1	19	12	0	38	26	0	18	12	0	18	12
OR-0001	278	284	279	−35	38	35	−212	212	212	−99	99	99	−99	99	99
OR-0010	0	66	44	0	14	9	0	26	17	0	13	9	0	12	8
AIPW-0101	−108	9430	218	−18	2698	50	291	6316	90	176	2792	40	131	3527	40
AIPW-1001	7	123	77	−1	26	17	−7	46	31	−4	27	18	−3	27	18
AIPW-0110	−26	5663	135	4	1461	32	15	2081	60	19	733	23	−16	841	23
AIPW-1010	−2	85	56	0	17	12	0	35	23	0	17	12	0	17	12
AIPW-EN	−1	96	60	0	20	12	0	40	26	0	21	12	0	21	12
AIPW-B-0101	22	526	170	−1	119	37	−31	150	68	−6	78	35	−6	78	35
AIPW-B-1001	7	123	77	−1	26	17	−7	46	31	−4	27	18	−3	27	18
AIPW-B-0110	−2	308	104	0	73	25	2	97	45	0	47	18	−1	48	18
AIPW-B-1010	−2	85	56	0	17	12	0	35	23	0	17	12	0	17	12
AIPW-B-EN	−1	95	60	0	19	12	0	39	26	0	20	12	0	20	12
MR-1000	−2	86	57	0	18	12	0	35	24	0	17	12	0	17	12
MR-0100	−12	90	57	2	19	13	−3	36	24	0	17	11	0	17	11
MR-0010	2	74	51	0	16	11	0	31	21	0	15	10	0	14	10
MR-0001	7	67	45	−5	15	10	4	27	18	0	14	9	0	13	9
MR-1100	−4	89	58	1	18	12	0	36	25	0	18	12	0	18	12
MR-1010	0	87	57	0	18	12	0	38	25	0	18	12	0	18	12
MR-1001	−1	89	57	0	18	12	−1	38	25	−1	18	12	−1	18	12
MR-0110	1	87	59	0	19	12	1	36	23	0	17	11	0	17	11
MR-0101	−7	90	57	1	19	13	−3	38	25	0	17	11	−1	18	11
MR-0011	3	76	51	−1	16	11	0	32	21	0	15	10	−1	15	10
MR-1110	−1	89	58	0	19	12	1	38	26	0	18	12	0	18	12
MR-1101	−2	91	59	0	19	12	−1	39	27	−1	18	12	−1	18	12
MR-1011	1	91	57	0	19	12	0	40	27	0	18	12	0	18	12
MR-0111	2	88	58	0	19	12	0	38	25	0	17	11	0	17	11
MR-1111	1	93	60	0	19	12	0	41	28	0	18	12	0	19	12
MR-EN	−2	90	59	0	19	12	0	39	26	0	18	12	0	18	12

RMSE: root mean square error. MAE: median absolute error. IPW: inverse probability weighting. OR: outcome regression. AIPW: augmented inverse probability weighting. AIPW-B: AIPW with regularized weights. MR: multiply robust.

and $\widehat{\boldsymbol{\beta}}_{\text{MR-1111}}$ have slightly larger RMSE and MAE compared to $\widehat{\boldsymbol{\beta}}_{\text{MR-1010}}$, due to the larger dimension of $\widehat{\boldsymbol{\rho}}$ for the former two (11 for $\widehat{\boldsymbol{\beta}}_{\text{MR-1011}}$ and 12 for $\widehat{\boldsymbol{\beta}}_{\text{MR-1111}}$ vs. 6 for $\widehat{\boldsymbol{\beta}}_{\text{MR-1010}}$). As n increases to 800, the performance of the three becomes similar, consistent with Theorem 3 that they have the same asymptotic distribution.

One interesting observation is that $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ seems to still provide a reasonable (at least not too bad) estimate of $\boldsymbol{\beta}_0$ even if there is no model correctly specified. See the performance of, for example, $\widehat{\boldsymbol{\beta}}_{\text{MR-0100}}$, $\widehat{\boldsymbol{\beta}}_{\text{MR-0001}}$, and $\widehat{\boldsymbol{\beta}}_{\text{MR-0101}}$. Although a little bit surprising, this observation is not totally unexpected. It is easy to see that

$$E(w(\boldsymbol{X}, \boldsymbol{S})[f(\boldsymbol{X}, \boldsymbol{S}) - E\{f(\boldsymbol{X}, \boldsymbol{S})\}]|R = 1) = 0 \quad (17)$$

is true for an arbitrary function $f(\boldsymbol{X}, \boldsymbol{S})$, where $w(\boldsymbol{X}, \boldsymbol{S}) = 1/\pi(\boldsymbol{X}, \boldsymbol{S})$, and $w(\boldsymbol{X}, \boldsymbol{S})$ is the only weight that possesses this property. Therefore, if we were able to construct constraints on w_i similar to those in (9) but using all possible functions of $\boldsymbol{X}$ and $\boldsymbol{S}$ rather than just the $J + pK$ functions as in (7) and (8), the resulting $\widehat{w}_i$, pretending that it were available, would be $w(\boldsymbol{X}_i, \boldsymbol{S}_i)$ plus some random perturbation due to the finite sample size (more precisely, the normalized value under sum-to-one normalization). In general, $\widehat{w}_i$ should become closer to the normalized value of $w(\boldsymbol{X}_i, \boldsymbol{S}_i)$, albeit some random perturbation, when more functions of $\boldsymbol{X}$ and $\boldsymbol{S}$ are used to construct the constraints, if we leave aside all other concerns, including the numerical issues. In other words, imposing constraints based on (17) facilitates the achievement of consistency regardless of the specific form of the function. When the number of constraints is not too large so that the numerical behavior is not a problem, using more constraints could boost the performance of $\widehat{w}_i$. This is why in our simulation studies $\widehat{\boldsymbol{\beta}}_{\text{MR}}$ perform reasonably well even if all models are misspecified.

As pointed out by one reviewer, the goal of improving the robustness against model misspecifications may also be achieved by the AIPW method through fitting enlarged models that combine the models in $\mathcal{P}$ and/or $\mathcal{A}$. The enlarged model will be correctly specified if one of the small models is correct, with many parameters equal to zero. In Tables 1 and 2, we also include the estimators $\widehat{\boldsymbol{\beta}}^{\text{EN}}_{\text{IPW}}$,

Table 2. Comparison of different estimators with $n = 800$ based on 2000 replications. The names of the estimators have the form "method-0000" (or "method-EN" for enlarged models), with each digit of the four-digit number, from left to right, indicating if $\pi^1(\boldsymbol{\alpha}^1)$, $\pi^2(\boldsymbol{\alpha}^2)$, $a^1(\boldsymbol{\gamma}^1)$, or $a^2(\boldsymbol{\gamma}^2)$ is used, respectively. The results have been multiplied by 100

	β_1			β_2			β_3			β_4			β_5		
Estimator	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
IPW-0100	−32	115	66	11	25	15	−13	44	27	0	23	13	0	23	13
IPW-1000	−2	43	28	0	9	6	0	17	12	0	9	6	0	9	6
IPW-EN	−3	44	30	1	9	6	0	18	12	0	9	6	0	9	6
OR-0001	280	281	279	−34	35	34	−215	215	215	−100	100	100	−100	100	100
OR-0010	−1	33	22	0	7	4	0	13	9	0	6	4	0	6	4
AIPW-0101	−142	3310	258	17	759	58	115	644	101	104	450	49	92	367	48
AIPW-1001	0	62	42	0	13	9	−2	22	15	−1	14	9	−1	13	9
AIPW-0110	−2	1712	146	1	405	35	−3	365	52	10	219	22	−1	175	23
AIPW-1010	−2	42	28	0	9	6	0	17	12	0	9	6	0	9	6
AIPW-EN	−2	44	29	1	9	6	0	18	12	0	9	6	0	9	6
AIPW-B-0101	26	478	164	−2	109	37	−35	116	55	−1	67	27	−1	61	28
AIPW-B-1001	0	62	42	0	13	9	−2	22	15	−1	14	9	−1	13	9
AIPW-B-0110	2	273	93	−1	64	22	0	68	33	1	37	14	−1	35	15
AIPW-B-1010	−2	42	28	0	9	6	0	17	12	0	9	6	0	9	6
AIPW-B-EN	−2	44	30	1	9	6	0	18	12	0	9	6	0	9	6
MR-1000	−2	43	28	0	9	6	0	17	12	0	9	6	0	9	6
MR-0100	−16	49	31	4	10	7	−4	18	12	0	9	6	0	9	6
MR-0010	−2	39	26	0	8	5	0	16	10	0	7	5	0	7	5
MR-0001	8	34	23	−5	8	6	5	14	10	0	6	4	0	6	4
MR-1100	−2	43	28	1	9	6	−1	17	12	0	9	6	0	9	6
MR-1010	−2	43	28	0	9	6	0	17	12	0	9	6	0	9	6
MR-1001	−2	44	30	0	9	6	0	18	12	0	9	6	0	9	6
MR-0110	−2	46	31	1	10	6	0	17	12	0	9	6	0	9	6
MR-0101	−14	49	31	3	10	7	−3	18	12	0	9	6	0	9	6
MR-0011	−2	39	26	0	8	5	−1	16	11	0	7	5	0	7	5
MR-1110	−2	43	28	0	9	6	0	17	12	0	9	6	0	9	6
MR-1101	−2	44	30	1	9	6	−1	18	12	0	9	6	0	9	6
MR-1011	−2	44	29	0	9	6	0	18	12	0	9	6	0	9	6
MR-0111	−2	46	30	1	10	6	0	17	12	0	9	6	0	9	6
MR-1111	−2	44	29	0	9	6	0	18	12	0	9	6	0	9	6
MR-EN	−2	44	29	1	9	6	0	18	12	0	9	6	0	9	6

RMSE: root mean square error. MAE: median absolute error. IPW: inverse probability weighting. OR: outcome regression. AIPW: augmented inverse probability weighting. AIPW-B: AIPW with regularized weights. MR: multiply robust.

$\widehat{\boldsymbol{\beta}}^{\mathrm{EN}}_{\mathrm{AIPW}}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{EN}}_{\mathrm{AIPW\text{-}B}}$, and $\widehat{\boldsymbol{\beta}}^{\mathrm{EN}}_{\mathrm{MR}}$, which are constructed based on the models $\mathrm{logit}\{\pi^{\mathrm{EN}}(\boldsymbol{\alpha}^{\mathrm{EN}})\} = \alpha_1^{\mathrm{EN}} + \alpha_2^{\mathrm{EN}} X^{(1)} + \cdots + \alpha_5^{\mathrm{EN}} X^{(4)} + \alpha_6^{\mathrm{EN}} S^{(1)} + \cdots + \alpha_8^{\mathrm{EN}} S^{(3)}$ and $a^{\mathrm{EN}}(\boldsymbol{\gamma}^{\mathrm{EN}}) = \gamma_1^{\mathrm{EN}} + \gamma_2^{\mathrm{EN}} X^{(1)} + \cdots + \gamma_5^{\mathrm{EN}} X^{(4)} + \gamma_6^{\mathrm{EN}} S^{(1)} + \cdots + \gamma_8^{\mathrm{EN}} S^{(3)}$. These two models combine $\pi^1(\boldsymbol{\alpha}^1)$ with $\pi^2(\boldsymbol{\alpha}^2)$ and $a^1(\boldsymbol{\gamma}^1)$ with $a^2(\boldsymbol{\gamma}^2)$, respectively, by including the main effects of all available explanatory variables. These four estimators are consistent, as illustrated by our simulation results. However, in practice it is not always feasible to fit a large model. For example, when the dimension of $\boldsymbol{S}$ is high, a model that contains all components of $\boldsymbol{S}$ may have too many unknown parameters. Even if these parameters are all identifiable, we may not be able to estimate them well enough based on the collected data. In addition, the $\sqrt{n}$-consistency of the IPW and AIPW estimators requires the parameters in the model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ to be $n^{1/4+\delta}$-consistent for some $\delta > 0$ (Robins, Rotnitzky, and Zhao 1995). This requirement sets a limit on the number of parameters in the enlarged models. Although techniques developed in the variable selection literature can help in this case, different levels of tuning may set different subsets of the parameters to zero, again raising the robustness concern. A more natural approach in case of high-dimensional $\boldsymbol{S}$, as discussed in Robins et al. (2007), is to fit multiple models, each with different subsets of $\boldsymbol{S}$, different orders of interactions, and possibly different link functions. Our proposed method provides an innovative way of combining these multiple models into estimation. Even in cases where fitting enlarged models is feasible, we still recommend using $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ based on the enlarged models. This is because, in observational studies, no matter how complex a model is, it is an approximation to the unknown true data-generating process; that is, "all models are wrong" (Box and Draper 1987), including the enlarged models. Therefore, what matters the most is how a method performs under misspecified models. Our simulation results and previous discussion demonstrate the capability of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ in producing not-too-bad estimates when all the models are misspecified.

To make inference, we need a good estimate of the standard error of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$. In general, we recommend the bootstrapping method. Tables 3 and 4 summarize the performance of this method based on resampling size 200. We provide the mean of bootstrapping-based standard errors and the percentage that

Table 3. Comparison of different ways of calculating the standard error of the multiply robust estimators with $n = 200$ based on 2000 replications. The names of the estimators have the form "MR-0000" (or "MR-EN" for enlarged models), with each digit of the four-digit number, from left to right, indicating if $\pi^1(\boldsymbol{\alpha}^1)$, $\pi^2(\boldsymbol{\alpha}^2)$, $a^1(\boldsymbol{\gamma}^1)$, or $a^2(\boldsymbol{\gamma}^2)$ is used, respectively. The results have been multiplied by 100

	β_1			β_2			β_3			β_4			β_5		
Estimator	EMP	EST	PER	EMP	EST	PER	EMP	EST	PER	EMP	EST	PER	EMP	EST	PER
MR-1000	86	(81)	(92.6%)	18	(17)	(91.9%)	35	(33)	(92.4%)	17	(16)	(92.3%)	17	(16)	(92.6%)
MR-1000	86	{76}	{89.0%}	18	{15}	{88.5%}	35	{32}	{91.5%}	17	{15}	{90.0%}	17	{15}	{89.9%}
MR-0100	89	(80)	(91.5%)	19	(17)	(91.4%)	36	(32)	(91.7%)	17	(15)	(92.4%)	17	(15)	(92.6%)
MR-0010	74	(74)	(94.4%)	16	(16)	(94.0%)	31	(30)	(94.5%)	15	(14)	(94.7%)	14	(14)	(94.6%)
MR-0001	67	(70)	(95.5%)	14	(15)	(93.9%)	26	(27)	(94.9%)	14	(14)	(94.5%)	13	(14)	(95.6%)
MR-1100	88	(83)	(92.4%)	18	(17)	(92.3%)	36	(33)	(92.1%)	18	(16)	(92.9%)	18	(16)	(92.5%)
MR-1100	88	{74}	{87.6%}	18	{15}	{86.9%}	36	{31}	{89.7%}	18	{15}	{88.3%}	18	{15}	{88.3%}
MR-1010	87	(84)	(93.7%)	18	(17)	(93.6%)	38	(35)	(92.5%)	18	(17)	(93.2%)	18	(17)	(93.0%)
MR-1010	87	[80]	[92.2%]	18	[16]	[90.6%]	38	[33]	[91.6%]	18	[16]	[91.7%]	18	[16]	[92.2%]
MR-1001	89	(86)	(93.4%)	18	(18)	(93.4%)	38	(36)	(93.2%)	18	(17)	(93.1%)	18	(17)	(93.6%)
MR-1001	89	{68}	{84.9%}	18	{14}	{83.9%}	38	{29}	{85.2%}	18	{13}	{84.4%}	18	{13}	{84.1%}
MR-0110	87	(83)	(93.9%)	19	(17)	(93.5%)	36	(34)	(93.3%)	17	(16)	(94.3%)	17	(16)	(93.7%)
MR-0101	89	(85)	(93.4%)	19	(18)	(92.8%)	38	(35)	(93.0%)	17	(17)	(92.9%)	17	(16)	(93.5%)
MR-0011	76	(79)	(95.6%)	16	(16)	(94.9%)	32	(32)	(95.5%)	15	(15)	(95.8%)	15	(15)	(95.2%)
MR-1110	89	(86)	(94.0%)	19	(18)	(93.9%)	38	(36)	(93.1%)	18	(17)	(93.4%)	18	(17)	(93.0%)
MR-1110	89	[80]	[91.4%]	19	[16]	[89.5%]	38	[33]	[90.6%]	18	[16]	[91.0%]	18	[16]	[92.2%]
MR-1101	91	(89)	(94.0%)	19	(18)	(93.5%)	39	(37)	(93.4%)	18	(17)	(93.5%)	18	(17)	(93.7%)
MR-1101	91	{66}	{82.9%}	19	{13}	{82.2%}	39	{28}	{83.2%}	18	{13}	{82.9%}	18	{13}	{82.6%}
MR-1011	91	(92)	(95.3%)	19	(19)	(95.0%)	40	(39)	(94.2%)	18	(18)	(95.0%)	18	(18)	(94.5%)
MR-1011	91	[80]	[90.4%]	19	[16]	[89.4%]	40	[33]	[89.6%]	18	[16]	[91.4%]	18	[16]	[91.0%]
MR-0111	88	(90)	(95.7%)	19	(19)	(94.6%)	38	(38)	(95.0%)	17	(18)	(95.1%)	17	(17)	(95.3%)
MR-1111	93	(95)	(95.7%)	19	(19)	(95.3%)	41	(40)	(94.7%)	18	(19)	(95.4%)	19	(19)	(94.9%)
MR-1111	93	[81]	[89.9%]	19	[16]	[88.6%]	41	[33]	[89.0%]	18	[16]	[90.8%]	19	[16]	[90.3%]
MR-EN	90	(88)	(94.0%)	19	(18)	(94.6%)	39	(37)	(93.2%)	18	(17)	(94.2%)	18	(17)	(93.3%)
MR-EN	90	[83]	[90.9%]	19	[17]	[90.4%]	39	[36]	[91.9%]	18	[17]	[91.2%]	18	[17]	[91.0%]

EMP: empirical standard error. EST: mean of estimated standard error. PER: percentage out of 2000 replications that the 95% confidence interval based on the estimated standard error covers the true parameter value. ( ): results based on bootstrapping with 200 resamples. { }: results based on Theorem 2. [ ]: results based on Theorem 3.

the 95% confidence interval constructed using the bootstrapping method covers the true parameter value, both over 2000 replications. We also include the results based on the asymptotic variance estimators given by Theorems 2 and 3, whenever available. The empirical standard error of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ may serve as one criterion for the comparison of different methods. For the bootstrapping method, when at least one model is correctly specified, the percentage that the true parameter value is contained in the 95% confidence interval is reasonably close to 95%, for both $n = 200$ and $n = 800$. Even when no model is correctly specified (i.e., $\widehat{\boldsymbol{\beta}}_{\mathrm{MR\text{-}0100}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{MR\text{-}0001}}$, and $\widehat{\boldsymbol{\beta}}_{\mathrm{MR\text{-}0101}}$), the coverage percentage is still very close to 95%, consistent with our previous discussion that $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ is robust against the complete misspecification of all models. For the method of using the asymptotic normal distribution to calculate the standard errors, the coverage percentage in many cases is substantially smaller than 95% when $n = 200$. Although this percentage gets much closer to 95% when $n = 800$, the coverage based on the bootstrapping method is still more satisfactory.

To further study the robustness of $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ against near-zero values of $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$, we summarize the sampling distributions of weights assigned to subjects with full data ($R = 1$) by different estimators. We use the ratios of the quantiles of the sampling distributions divided by the corresponding median to facilitate the comparison. The results with $n = 200$ are provided in Table 5. It is seen that some values of $R/\pi^2(\widehat{\boldsymbol{\alpha}}^2)$ are unusually large, resulting in some subjects dominating others by receiving erroneously large weights. This observation explains the problematic performance of the corresponding IPW and AIPW estimators in Table 1. The normalization of $R/\pi^2(\widehat{\boldsymbol{\alpha}}^2)$ by its sample mean reduces the occurrence of extreme weights, but apparently not to a satisfactory level, consistent with the improved yet still problematic performance of $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW\text{-}B\text{-}0110}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW\text{-}B\text{-}0101}}$. On the contrary, the weights based on our proposed method, whether $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled or not, have much more regularized values. The observations with $n = 800$ are similar to the above ones, and thus the results are omitted here. It is this elimination of extreme weights that endows our estimator with the robustness against near-zero values of $\widehat{\pi}(\boldsymbol{X}, \boldsymbol{S})$.

7. DATA APPLICATION

As an application of the proposed method, we consider data collected on 2139 HIV-infected subjects enrolled in AIDS Clinical Trials Group Protocol 175 (ACTG 175) (Hammer et al. 1996). This is a randomized trial with patients from 43 AIDS Clinical Trials Units and 9 National Hemophilia Foundation sites in the United States and Puerto Rico. This study evaluates treatment with either a single nucleoside or two nucleosides in HIV-infected subjects whose CD4 cell counts, a measure of

Table 4. Comparison of different ways of calculating the standard error of the multiply robust estimators with $n = 800$ based on 2000 replications. The names of the estimators have the form "MR-0000" (or "MR-EN" for enlarged models), with each digit of the four-digit number, from left to right, indicating if $\pi^1(\boldsymbol{\alpha}^1)$, $\pi^2(\boldsymbol{\alpha}^2)$, $a^1(\boldsymbol{\gamma}^1)$, or $a^2(\boldsymbol{\gamma}^2)$ is used, respectively. The results have been multiplied by 100

	β_1			β_2			β_3			β_4			β_5		
Estimator	EMP	EST	PER	EMP	EST	PER	EMP	EST	PER	EMP	EST	PER	EMP	EST	PER
MR-1000	43	(42)	(94.2%)	9	(9)	(94.1%)	17	(17)	(94.6%)	9	(8)	(93.9%)	9	(8)	(93.9%)
MR-1000	43	{42}	{93.7%}	9	{9}	{93.4%}	17	{17}	{94.5%}	9	{8}	{93.9%}	9	{8}	{93.4%}
MR-0100	46	(43)	(90.8%)	10	(9)	(90.4%)	18	(17)	(92.7%)	9	(8)	(92.3%)	9	(8)	(93.5%)
MR-0010	39	(37)	(93.9%)	8	(8)	(93.5%)	16	(15)	(93.9%)	7	(7)	(93.2%)	7	(7)	(93.7%)
MR-0001	33	(33)	(94.5%)	7	(7)	(89.7%)	13	(13)	(93.0%)	6	(6)	(94.2%)	6	(6)	(94.8%)
MR-1100	43	(42)	(94.0%)	9	(9)	(94.4%)	17	(17)	(94.5%)	9	(8)	(93.7%)	9	(8)	(93.9%)
MR-1100	43	{42}	{93.7%}	9	{8}	{92.7%}	17	{17}	{94.3%}	9	{8}	{93.5%}	9	{8}	{93.1%}
MR-1010	43	(42)	(94.2%)	9	(9)	(93.6%)	17	(17)	(94.3%)	9	(8)	(93.4%)	9	(8)	(93.6%)
MR-1010	43	[42]	[93.9%]	9	[8]	[93.2%]	17	[17]	[94.4%]	9	[8]	[93.5%]	9	[8]	[93.9%]
MR-1001	44	(43)	(93.6%)	9	(9)	(94.1%)	18	(17)	(94.6%)	9	(9)	(94.0%)	9	(9)	(93.7%)
MR-1001	44	{41}	{92.4%}	9	{8}	{92.6%}	18	{17}	{93.5%}	9	{8}	{91.5%}	9	{8}	{91.4%}
MR-0110	46	(42)	(93.0%)	10	(9)	(92.6%)	17	(16)	(94.1%)	9	(8)	(92.7%)	9	(8)	(93.5%)
MR-0101	47	(43)	(91.5%)	10	(9)	(90.9%)	18	(17)	(92.6%)	9	(8)	(92.8%)	9	(8)	(93.2%)
MR-0011	39	(37)	(94.0%)	8	(8)	(93.5%)	16	(15)	(94.5%)	7	(7)	(93.7%)	7	(7)	(94.3%)
MR-1110	43	(42)	(93.9%)	9	(9)	(93.6%)	17	(17)	(94.4%)	9	(8)	(93.1%)	9	(8)	(93.4%)
MR-1110	43	[42]	[93.9%]	9	[8]	[92.9%]	17	[17]	[94.3%]	9	[8]	[93.1%]	9	[8]	[93.5%]
MR-1101	44	(43)	(93.5%)	9	(9)	(93.7%)	18	(17)	(94.4%)	9	(9)	(93.9%)	9	(9)	(93.4%)
MR-1101	44	{40}	{91.8%}	9	{8}	{91.0%}	18	{16}	{92.7%}	9	{8}	{91.6%}	9	{8}	{91.3%}
MR-1011	44	(42)	(93.8%)	9	(9)	(93.8%)	18	(17)	(94.5%)	9	(8)	(93.3%)	9	(8)	(93.0%)
MR-1011	44	[42]	[93.3%]	9	[8]	[93.0%]	18	[17]	[93.6%]	9	[8]	[92.8%]	9	[8]	[92.7%]
MR-0111	46	(42)	(93.2%)	9	(9)	(92.8%)	17	(17)	(94.2%)	9	(8)	(92.9%)	9	(8)	(93.4%)
MR-1111	44	(43)	(93.6%)	9	(9)	(93.9%)	18	(18)	(94.4%)	9	(8)	(93.1%)	9	(8)	(93.0%)
MR-1111	44	[42]	[93.1%]	9	[8]	[92.9%]	18	[17]	[93.3%]	9	[8]	[92.5%]	9	[8]	[92.8%]
MR-EN	44	(42)	(93.5%)	9	(9)	(93.9%)	18	(17)	(94.4%)	9	(9)	(93.9%)	9	(8)	(93.8%)
MR-EN	44	[42]	[93.9%]	9	[9]	[93.7%]	18	[17]	[94.2%]	9	[9]	[93.6%]	9	[8]	[93.8%]

EMP: empirical standard error. EST: mean of estimated standard error. PER: percentage out of 2000 replications that the 95% confidence interval based on the estimated standard error covers the true parameter value. ( ): results based on bootstrapping with 200 resamples. { }: results based on Theorem 2. [ ]: results based on Theorem 3.

immunologic status, are from 200 to 500 per cubic millimeter. The subjects are randomized to four different antiretroviral regimens in equal probability: zidovudine (ZDV) only, ZDV+didanosine (ddI), ZDV+zalcitabine (ddC), and ddI only. Following the analysis in Davidian, Tsiatis, and Leon (2005) and Zhang, Tsiatis, and Davidian (2008), we consider two arms for the treatment: the arm with standard ZDV monotherapy alone and the arm with the other three newer treatments. The two arms have 532 and 1607 subjects, respectively. Our main interest is the treatment arm effect on the CD4 counts measured at 96 ± 5 weeks post baseline ($CD4_{96}$), adjusting for the baseline CD4 counts ($CD4_0$) and certain baseline characteristics, including the continuous covariates age (age, years) and weight (weight, kg) and the binary covariates treatment (trt, 0 = ZDV), race (race, 0 = white), gender (gender, 0 = female), antiretroviral history (history, 0 = naive, 1 = experienced) and whether the subject is off-treatment prior to 96 weeks (offtrt, 0 = no). That is, we want to fit the linear regression model

$$\begin{aligned} \mathrm{CD4}_{96} = \beta_1 + \beta_2 \mathrm{trt} + \beta_3 \mathrm{CD4}_0 + \beta_4 \mathrm{age} + \beta_5 \mathrm{weight} + \beta_6 \mathrm{race} \\ + \beta_7 \mathrm{gender} + \beta_8 \mathrm{history} + \beta_9 \mathrm{offtrt} + \epsilon, \end{aligned}$$

where ϵ has mean zero conditional on all covariates. The data can be found in the R package "speff2trial" (*http://cran.r-project.org/web/packages/speff2trial/speff2trial.pdf*). The average age of the subjects is 35 years old with a standard deviation 8.7 years old. There are 1522 white subjects and 617 nonwhite subjects, and 1171 males and 368 females. Among the patients, 1253 of them have antiretroviral history, and 776 of them are off-treatment before 96 weeks.

Because some subjects drop out of the study, the CD4 counts at 96 ± 5 weeks are missing for 797 subjects (missingness rate 37%). However, at the baseline and during the follow-up, full measurements on additional variables correlated with $CD4_{96}$ are obtained. These include the CD4 counts at 20 ± 5 weeks ($CD4_{20}$) and the CD8 counts (another measure of immunologic status) at both the baseline ($CD8_0$) and 20 ± 5 weeks ($CD8_{20}$). In our analysis, we use them as the auxiliary variables. Despite that it is reasonable to assume $CD4_{96}$ is MAR (Davidian, Tsiatis, and Leon 2005), correctly specifying a model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ is challenging in the presence of an eight-dimensional $\boldsymbol{X}$ and a three-dimensional $\boldsymbol{S}$, even with the help of model selection and diagnostic techniques. The same is true for modeling $E(Y|\boldsymbol{X}, \boldsymbol{S})$. Due to possible model misspecifications, estimation and inference based on doubly robust methods are easily questionable. Therefore, we apply the proposed method, the conclusions from which, although not definitive, may be more trustworthy in the case of model misspecifications, as demonstrated by the simulation studies in Section 6.

We use a logistic regression model for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and a linear regression model for $E(Y|\boldsymbol{X}, \boldsymbol{S})$. To be prudent, both models contain all the main effects of $\boldsymbol{X}$ and $\boldsymbol{S}$. We use the estimating function $\boldsymbol{U}(Y, \boldsymbol{X}, \boldsymbol{\beta}) = \boldsymbol{X}(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$. The results of our

Table 5. Summary of the sampling distributions of different weights assigned to subjects with full data ($R = 1$) based on $n = 200$ and 2000 replications. The names of the weights based on our proposed method have the form "$\widehat{w}_i$-0000," with each digit of the four-digit number, from left to right, indicating if $\pi^1(\boldsymbol{\alpha}^1)$, $\pi^2(\boldsymbol{\alpha}^2)$, $a^1(\boldsymbol{\gamma}^1)$, or $a^2(\boldsymbol{\gamma}^2)$ is used, respectively. The results are the ratios of the corresponding quantiles divided by the median

	Quantiles									
Weight	Min	1%	5%	10%	25%	75%	90%	95%	99%	Max
$R_i/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)$	0.97	0.97	0.98	0.98	0.99	1.02	4.37	5.40	7.14	14.94
$R_i/\pi_i^2(\widehat{\boldsymbol{\alpha}}^2)$	0.90	0.90	0.90	0.91	0.92	1.20	1.84	3.85	24.83	54499.42
$\frac{R_i/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}{n^{-1}\sum_{i=1}^n R_i/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}$	0.82	0.87	0.91	0.93	0.96	1.05	4.40	5.29	6.84	14.34
$\frac{R_i/\pi_i^2(\widehat{\boldsymbol{\alpha}}^2)}{n^{-1}\sum_{i=1}^n R_i/\pi_i^2(\widehat{\boldsymbol{\alpha}}^2)}$	0.00	0.09	0.33	0.48	0.74	1.25	1.80	3.56	22.08	245.03
$\widehat{w}_i$-1000	0.97	0.97	0.98	0.98	0.99	1.02	4.37	5.40	7.14	14.94
$\widehat{w}_i$-0100	0.77	0.88	0.90	0.91	0.93	1.19	1.74	3.21	8.90	48.76
$\widehat{w}_i$-0010	0.14	0.39	0.54	0.62	0.79	1.21	1.58	2.10	4.71	24.58
$\widehat{w}_i$-0001	0.14	0.55	0.71	0.79	0.90	1.10	1.24	1.39	1.99	32.15
$\widehat{w}_i$-1100	0.68	0.94	0.96	0.97	0.98	1.03	3.97	5.35	7.85	45.84
$\widehat{w}_i$-1010	0.23	0.62	0.78	0.86	0.94	1.09	3.36	5.12	9.26	52.62
$\widehat{w}_i$-1001	0.22	0.70	0.81	0.86	0.93	1.10	3.55	5.15	9.18	53.83
$\widehat{w}_i$-0110	0.23	0.57	0.71	0.78	0.88	1.24	1.99	3.50	8.91	45.68
$\widehat{w}_i$-0101	0.26	0.65	0.76	0.81	0.89	1.23	1.92	3.38	9.25	52.22
$\widehat{w}_i$-0011	0.13	0.37	0.51	0.61	0.78	1.28	1.76	2.40	5.25	37.13
$\widehat{w}_i$-1110	0.23	0.60	0.77	0.84	0.93	1.11	3.15	5.04	9.88	49.08
$\widehat{w}_i$-1101	0.18	0.68	0.80	0.85	0.92	1.12	3.28	5.12	9.85	48.82
$\widehat{w}_i$-1011	0.11	0.52	0.68	0.76	0.88	1.20	3.07	5.15	10.83	62.66
$\widehat{w}_i$-0111	0.12	0.49	0.64	0.71	0.84	1.30	2.20	3.78	9.68	51.56
$\widehat{w}_i$-1111	0.09	0.49	0.66	0.75	0.87	1.22	2.97	5.16	11.35	59.74

analysis are summarized in Table 6. Results based on complete-case analysis are also included for comparison. The standard errors are calculated using the bootstrapping method with 1000 replications. From Table 6, it is seen that patients receiving the three newer treatments (trt=1) have significantly higher CD4 counts at 96 ± 5 weeks, adjusting for the baseline CD4 counts and other covariates. In other words, the three newer treatments significantly slow the progression of HIV disease compared to the treatment of ZDV alone. The complete-case analysis draws a similar conclusion, but may overestimate the treatment effect due to ignoring the missingness mechanism. Based on our method, age, weight, and gender do not have substantial impacts on the CD4 counts at 96 ± 5 weeks, whereas race has a marginally significant impact, in that the nonwhite patients have lower CD4 counts, and thus more severe HIV infection, compared to the white patients. Patients with an antiretroviral history have clearly lower CD4 counts at 96 ± 5 weeks compared to those who do have previous antiretroviral therapy. This is because patients with more severe HIV infection are more likely to seek antiretroviral treatment. Patients who stop the treatment during the follow-up have significantly lower CD4 counts at 96 ± 5 weeks.

Table 6. Results of the analysis on the ACTG 175 data ($n = 2139$)

	Our method			Complete-case analysis		
	Estimate	s.e.	p-value	Estimate	s.e.	p-value
Intercept	65.53	34.06	0.054	21.50	30.65	0.483
Trt	52.72	10.34	< 0.001	63.68	8.74	< 0.001
$CD4_0$	0.73	0.05	< 0.001	0.76	0.04	< 0.001
Age	0.14	0.55	0.796	0.10	0.48	0.829
Weight	0.27	0.33	0.417	0.54	0.31	0.084
Race	−18.30	9.66	0.058	−20.60	9.07	0.023
Gender	−16.54	11.34	0.145	−10.73	10.92	0.326
History	−41.45	8.65	< 0.001	−42.02	8.34	< 0.001
Offtrt	−86.87	10.31	< 0.001	−80.72	10.91	< 0.001

8. DISCUSSION

It is seen from Section 5 that, when both $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$ are correctly modeled, $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ achieves the maximum possible efficiency in estimating $\boldsymbol{\beta}_0$ with the observed data under a fixed $\boldsymbol{D}(\boldsymbol{X}, \boldsymbol{\beta})$. When either the models for $\pi(\boldsymbol{X}, \boldsymbol{S})$ or the models for $E(Y|\boldsymbol{X}, \boldsymbol{S})$ are all misspecified, it is difficult in general to quantify the efficiency loss. This is also true for $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$. There have been many recent developments on improving $\widehat{\boldsymbol{\beta}}_{\mathrm{AIPW}}$, and some new estimators always have higher efficiency than $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ even if $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is incorrectly modeled. Refer to, for example, Tan (2006, 2008, 2010), Cao, Tsiatis, and Davidian (2009), Tsiatis, Davidian, and Cao (2011), Han (2012), and Rotnitzky et al. (2012). Inspired by these works, we are trying to improve $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ by incorporating the models for $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$ in a particular way, so that the resulting estimator is multiply robust, is always more efficient than $\widehat{\boldsymbol{\beta}}_{\mathrm{IPW}}$ when $\pi(\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled, and achieves the maximum efficiency when $E(Y|\boldsymbol{X}, \boldsymbol{S})$ is correctly modeled in addition.

The proposed method may have numerical issues when the sample size is small and/or the number of constraints in (9) is large, in which case $\mathbf{0}$ may not be in the convex hull of $\{\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) : i = 1, \ldots, m\}$. The adjusted empirical likelihood

method (Chen, Variyath, and Abraham 2008) seems promising to deal with this problem. Formal theoretical investigation is necessary to study whether this method can yield a multiply robust estimator after it adds an "artificial observation" and assigns a nonzero weight to this observation. Another possible solution is to reduce the number of constraints in (9) by postulating less models for both $\pi(\boldsymbol{X}, \boldsymbol{S})$ and $E(Y|\boldsymbol{X}, \boldsymbol{S})$. We will make future investigation on the numerical issues.

Depending on the specific form of the estimating functions $\boldsymbol{U}(\boldsymbol{\beta})$, there are alternative ways of imposing constraints on w_i other than those in (9). An interesting example is that, when $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{X}(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$, as in our simulation studies and data application, we can impose constraints on w_i without calculating $\widehat{\boldsymbol{\beta}}^k$. Specifically, consider the following constraints:

$$
\begin{aligned}
&w_i \geq 0 \quad (i = 1, \ldots, m), \qquad \sum_{i=1}^{m} w_i = 1,\\
&\sum_{i=1}^{m} w_i \{\pi_i^j(\widehat{\boldsymbol{\alpha}}^j) - \theta^j(\widehat{\boldsymbol{\alpha}}^j)\} = 0 \quad (j = 1, \ldots, J),\\
&\sum_{i=1}^{m} w_i \boldsymbol{X}_i a_i^k(\widehat{\boldsymbol{\gamma}}^k) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i a_i^k(\widehat{\boldsymbol{\gamma}}^k) \quad (k = 1, \ldots, K),\\
&\sum_{i=1}^{m} w_i \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}.
\end{aligned}
$$

We may use the w_i that maximize $\prod_{i=1}^{m} w_i$ subject to the above constraints to derive one estimator of $\boldsymbol{\beta}_0$ as in (12). Following the same arguments presented in this article, the resulting estimator is multiply robust and asymptotically normally distributed. However, now the dimension of the Lagrange multipliers becomes $J + pK + p(p+1)/2$, which is likely to bring numerical challenges in practice.

In this article, we considered the mean regression problem. Another interesting setting is median regression, which is more robust against outliers. More generally, we may consider quantile regression (Koenker and Bassett 1978; He and Shao 1996; Koenker 2005), which provides a systematic way of studying the influence of covariates on the entire distribution of the response variable. There has been only limited research on quantile regression with missing data. See, for example, Lipsitz et al. (1997), Yi and He (2009), and Wei, Ma, and Carroll (2012). We anticipate that our proposed methodology can be generalized to this setting by imposing proper constraints over the weights on the complete cases. This topic is worth future exploration.

APPENDIX

Proof of $\boldsymbol{F}_n(\boldsymbol{\rho})$ having a unique minimizer. Let $\zeta = \| \boldsymbol{\rho} \|$, $\widetilde{\boldsymbol{\rho}} = \boldsymbol{\rho}/\zeta$ and $\mathcal{S} = \{\widetilde{\boldsymbol{\rho}} : \| \widetilde{\boldsymbol{\rho}} \| = 1\}$. Because of (7) and (8), $\mathbf{0}$ is inside the convex hull of $\{\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) : i = 1, \ldots, m\}$, at least when n is sufficiently large. Therefore, we have $\max_{1\leq i\leq m}\{-\widetilde{\boldsymbol{\rho}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})\} > 0$ for any $\widetilde{\boldsymbol{\rho}} \in \mathcal{S}$. Since $\mathcal{S}$ is compact, there exists $\widetilde{\boldsymbol{\rho}}_\dagger \in \mathcal{S}$ such that $\inf_{\widetilde{\boldsymbol{\rho}}\in\mathcal{S}} \max_{1\leq i\leq m}\{-\widetilde{\boldsymbol{\rho}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})\} = \max_{1\leq i\leq m}\{-\widetilde{\boldsymbol{\rho}}_\dagger^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})\} > 0$. Let $\mathcal{D}_n = \{\boldsymbol{\rho} : 1 + \boldsymbol{\rho}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) > 0,\ i = 1, \ldots, m\}$ denote the domain of $\boldsymbol{F}_n(\boldsymbol{\rho})$. For any $\boldsymbol{\rho} \in \mathcal{D}_n$, since $1 + \boldsymbol{\rho}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) = 1 + \zeta\widetilde{\boldsymbol{\rho}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) > 0$, $i = 1, \ldots, m$, it follows that $1 > \zeta \max_{1\leq i\leq m}\{-\widetilde{\boldsymbol{\rho}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})\} \geq \zeta \inf_{\widetilde{\boldsymbol{\rho}}\in\mathcal{S}} \max_{1\leq i\leq m}\{-\widetilde{\boldsymbol{\rho}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})\}$, which yields that $\zeta \leq [\inf_{\widetilde{\boldsymbol{\rho}}\in\mathcal{S}} \max_{1\leq i\leq m}\{-\widetilde{\boldsymbol{\rho}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})\}]^{-1} < \infty$. In other words, $\mathcal{D}_n$ is a bounded set. Therefore, $\boldsymbol{F}_n(\boldsymbol{\rho})$ is bounded below. On the other hand, $\boldsymbol{F}_n(\boldsymbol{\rho})$ increases with no upper bound as $\boldsymbol{\rho}$ approaches the boundary of $\mathcal{D}_n$. Hence, the continuity of $\boldsymbol{F}_n(\boldsymbol{\rho})$ in $\boldsymbol{\rho}$ ensures that $\boldsymbol{F}_n(\boldsymbol{\rho})$ has a minimizer inside $\mathcal{D}_n$. In addition, it is easy to see that $\mathcal{D}_n$ is an open and convex set. This fact, together with the strict convexity of $\boldsymbol{F}_n(\boldsymbol{\rho})$, ensures that the minimizer is unique. □

Lemma. When $\pi^1(\boldsymbol{\alpha}^1)$ is the correctly specified model for $\pi(\boldsymbol{X}, \boldsymbol{S})$, we have

$$
\begin{aligned}
n^{1/2}\widehat{\boldsymbol{\lambda}} = \boldsymbol{G}^{-1}\Bigg(&n^{-1/2}\sum_{i=1}^{n}\left\{\frac{R_i - \pi_i^1(\boldsymbol{\alpha}_0^1)}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\boldsymbol{g}_i(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)\right\}\\
&- E\left[\frac{\boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)}{\pi^1(\boldsymbol{\alpha}_0^1)}\left\{\frac{\partial\pi^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}}\right] n^{1/2}(\widehat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}_0^1)\Bigg) + o_p(1).
\end{aligned}
$$

where $\boldsymbol{G}$ is given by (16) and $\boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)$ is given by (15).

Proof. Let d_j and u_k denote the dimension of $\boldsymbol{\alpha}^j$ and $\boldsymbol{\gamma}^k$, respectively. Taking Taylor expansion of the left-hand side of (13) around $(\mathbf{0}^{\mathrm{T}}, \boldsymbol{\alpha}_*^{\mathrm{T}}, \boldsymbol{\beta}_*^{\mathrm{T}}, \boldsymbol{\gamma}_*^{\mathrm{T}})$ leads to

$$
\begin{aligned}
\mathbf{0} = {}& n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\widehat{\boldsymbol{g}}_i(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)\\
&- \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\frac{\widehat{\boldsymbol{g}}_i(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)^{\otimes 2}}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\right\} n^{1/2}\widehat{\boldsymbol{\lambda}} + \left\{\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\{\pi_i^1(\boldsymbol{\alpha}_0^1)\}^2}\right.\\
&\times\left(\begin{bmatrix}\left\{\dfrac{\partial\pi_i^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}} - \left\{\dfrac{1}{n}\displaystyle\sum_{h=1}^{n}\dfrac{\partial\pi_h^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}}\\ \mathbf{0}_{(J+pK-1)\times d_1}\end{bmatrix}\pi_i^1(\boldsymbol{\alpha}_0^1)\right.\\
&\left.\left. - \widehat{\boldsymbol{g}}_i(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)\left\{\frac{\partial\pi_i^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}}\right)\right\} n^{1/2}(\widehat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}_0^1)\\
&+ \sum_{j=2}^{J}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\begin{bmatrix}\mathbf{0}_{(j-1)\times d_j}\\ \left\{\dfrac{\partial\pi_i^j(\boldsymbol{\alpha}_*^j)}{\partial\boldsymbol{\alpha}^j}\right\}^{\mathrm{T}} - \left\{\dfrac{1}{n}\displaystyle\sum_{h=1}^{n}\dfrac{\partial\pi_h^j(\boldsymbol{\alpha}_*^j)}{\partial\boldsymbol{\alpha}^j}\right\}^{\mathrm{T}}\\ \mathbf{0}_{(J+pK-j)\times d_j}\end{bmatrix}\right)\\
&\times n^{1/2}(\widehat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}_*^j)\\
&+ \sum_{k=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\begin{bmatrix}\mathbf{0}_{\{J+p(k-1)\}\times p}\\ \dfrac{\partial\boldsymbol{U}_i^k(\boldsymbol{\beta}_*^k, \boldsymbol{\gamma}_*^k)}{\partial\boldsymbol{\beta}^k} - \dfrac{1}{n}\displaystyle\sum_{h=1}^{n}\dfrac{\partial\boldsymbol{U}_h^k(\boldsymbol{\beta}_*^k, \boldsymbol{\gamma}_*^k)}{\partial\boldsymbol{\beta}^k}\\ \mathbf{0}_{\{p(K-k)\}\times p}\end{bmatrix}\right)\\
&\times n^{1/2}(\widehat{\boldsymbol{\beta}}^k - \boldsymbol{\beta}_*^k)\\
&+ \sum_{k=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\begin{bmatrix}\mathbf{0}_{\{J+p(k-1)\}\times u_k}\\ \dfrac{\partial\boldsymbol{U}_i^k(\boldsymbol{\beta}_*^k, \boldsymbol{\gamma}_*^k)}{\partial\boldsymbol{\gamma}^k} - \dfrac{1}{n}\displaystyle\sum_{h=1}^{n}\dfrac{\partial\boldsymbol{U}_h^k(\boldsymbol{\beta}_*^k, \boldsymbol{\gamma}_*^k)}{\partial\boldsymbol{\gamma}^k}\\ \mathbf{0}_{\{p(K-k)\}\times u_k}\end{bmatrix}\right)\\
&\times n^{1/2}(\widehat{\boldsymbol{\gamma}}^k - \boldsymbol{\gamma}_*^k) + o_p(1)\\
= {}& n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\widehat{\boldsymbol{g}}_i(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*) - \boldsymbol{G}n^{1/2}\widehat{\boldsymbol{\lambda}}\\
&- E\left[\frac{\boldsymbol{g}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \boldsymbol{\gamma}_*)}{\pi^1(\boldsymbol{\alpha}_0^1)}\left\{\frac{\partial\pi^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}}\right] n^{1/2}(\widehat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}_0^1) + o_p(1).
\end{aligned}
$$

On the other hand, it is easy to check that

$$n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\widehat{\boldsymbol{g}}_i(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\boldsymbol{\gamma}_*) = n^{-1/2}\sum_{i=1}^{n}\frac{R_i-\pi_i^1(\boldsymbol{\alpha}_0^1)}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\boldsymbol{g}_i(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\boldsymbol{\gamma}_*)+o_p(1).$$

Therefore, solving for $\widehat{\boldsymbol{\lambda}}$ from the above Taylor expansion gives the result. □

Proof of Theorem 2. Since $\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}$ satisfies

$$\mathbf{0}=\sum_{i=1}^{m}\widehat{w}_i\boldsymbol{U}_i(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}})=\frac{1}{m}\sum_{i=1}^{n}\frac{R_i\theta^1(\widehat{\boldsymbol{\alpha}}^1)/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}{1+\widehat{\boldsymbol{\lambda}}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}\boldsymbol{U}_i(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}),$$

using the asymptotic expansion of $\widehat{\boldsymbol{\lambda}}$ given by the lemma, the Taylor expansion of the above equation around $(\mathbf{0}^{\mathrm{T}},\boldsymbol{\alpha}_*^{\mathrm{T}},\boldsymbol{\beta}_*^{\mathrm{T}},\boldsymbol{\gamma}_*^{\mathrm{T}},\boldsymbol{\beta}_0^{\mathrm{T}})$ leads to

$$\begin{aligned}
\mathbf{0} &= \frac{1}{m}\left\{\frac{1}{n}\sum_{h=1}^{n}\pi_h^1(\boldsymbol{\alpha}_0^1)\right\}n^{1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\boldsymbol{U}_i(\boldsymbol{\beta}_0)\\
&\quad-\frac{1}{m}\left\{\frac{1}{n}\sum_{h=1}^{n}\pi_h^1(\boldsymbol{\alpha}_0^1)\right\}\left\{\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\boldsymbol{U}_i(\boldsymbol{\beta}_0)\frac{\widehat{\boldsymbol{g}}_i(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\boldsymbol{\gamma}_*)^{\mathrm{T}}}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\right\}n^{1/2}\widehat{\boldsymbol{\lambda}}\\
&\quad+\frac{1}{m}\left(\sum_{i=1}^{n}\frac{R_i}{\{\pi_i^1(\boldsymbol{\alpha}_0^1)\}^2}\boldsymbol{U}_i(\boldsymbol{\beta}_0)\left[\left\{\frac{1}{n}\sum_{h=1}^{n}\frac{\partial\pi_h^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}}\pi_i^1(\boldsymbol{\alpha}_0^1)\right.\right.\\
&\quad\left.\left.-\left\{\frac{1}{n}\sum_{h=1}^{n}\pi_h^1(\boldsymbol{\alpha}_0^1)\right\}\left\{\frac{\partial\pi_i^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}}\right]\right)\\
&\quad\times n^{1/2}(\widehat{\boldsymbol{\alpha}}^1-\boldsymbol{\alpha}_0^1)\\
&\quad+\frac{1}{m}\left\{\frac{1}{n}\sum_{h=1}^{n}\pi_h^1(\boldsymbol{\alpha}_0^1)\right\}\left\{\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\frac{\partial\boldsymbol{U}_i(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\}n^{1/2}\\
&\quad\times(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}-\boldsymbol{\beta}_0)+o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}\boldsymbol{U}_i(\boldsymbol{\beta}_0)-\boldsymbol{L}n^{1/2}\widehat{\boldsymbol{\lambda}}\\
&\quad-E\left[\frac{\boldsymbol{U}(\boldsymbol{\beta}_0)}{\pi^1(\boldsymbol{\alpha}_0^1)}\left\{\frac{\partial\pi^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}}\right]n^{1/2}(\widehat{\boldsymbol{\alpha}}^1-\boldsymbol{\alpha}_0^1)\\
&\quad+E\left\{\frac{\partial\boldsymbol{U}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\}n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}-\boldsymbol{\beta}_0)+o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}\boldsymbol{Q}_i(\boldsymbol{\alpha}_0^1)-E\left[\frac{\boldsymbol{U}(\boldsymbol{\beta}_0)-\boldsymbol{L}\boldsymbol{G}^{-1}\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\boldsymbol{\gamma}_*)}{\pi^1(\boldsymbol{\alpha}_0^1)}\right.\\
&\quad\left.\times\left\{\frac{\partial\pi^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}}\right]n^{-1/2}\sum_{i=1}^{n}\{E(\boldsymbol{\Psi}^{\otimes2})\}^{-1}\boldsymbol{\Psi}_i\\
&\quad+E\left\{\frac{\partial\boldsymbol{U}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\}n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}-\boldsymbol{\beta}_0)+o_p(1).
\end{aligned}$$

Simple algebra and the generalized information equality (e.g., Lemma 9.1 in Tsiatis 2006) give that

$$E\left[\frac{\boldsymbol{U}(\boldsymbol{\beta}_0)-\boldsymbol{L}\boldsymbol{G}^{-1}\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\boldsymbol{\gamma}_*)}{\pi^1(\boldsymbol{\alpha}_0^1)}\left\{\frac{\partial\pi^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}^{\mathrm{T}}\right]=-E\left\{\frac{\partial\boldsymbol{Q}(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^1}\right\}=E(\boldsymbol{Q}\boldsymbol{\Psi}^{\mathrm{T}}).$$

Therefore, we have

$$\mathbf{0}=n^{-1/2}\sum_{i=1}^{n}\left[\boldsymbol{Q}_i-\{E(\boldsymbol{Q}\boldsymbol{\Psi}^{\mathrm{T}})\}\{E(\boldsymbol{\Psi}^{\otimes2})\}^{-1}\boldsymbol{\Psi}_i\right]+E\left\{\frac{\partial\boldsymbol{U}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\}n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}-\boldsymbol{\beta}_0)+o_p(1).$$

Solving for $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{MR}}-\boldsymbol{\beta}_0)$ gives the desired result. □

Proof of Theorem 3. Write

$$\boldsymbol{H}=\frac{R}{\pi^1(\boldsymbol{\alpha}_0^1)}\boldsymbol{U}(\boldsymbol{\beta}_0),\qquad \boldsymbol{A}=\frac{R}{\pi^1(\boldsymbol{\alpha}_0^1)}\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\boldsymbol{\gamma}_*).$$

It is easy to check that $\boldsymbol{L}=E(\boldsymbol{H}\boldsymbol{A}^{\mathrm{T}})$ and $\boldsymbol{G}=E(\boldsymbol{A}^{\otimes2})$. When $\mathcal{A}$ contains a correctly specified model for $E(Y|\boldsymbol{X},\boldsymbol{S})$, $E\{\boldsymbol{U}(\boldsymbol{\beta}_0)|\boldsymbol{X},\boldsymbol{S}\}$ is a component of $\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\boldsymbol{\gamma}_*)$, and thus $E\{\boldsymbol{U}(\boldsymbol{\beta}_0)|\boldsymbol{X},\boldsymbol{S}\}R/\pi^1(\boldsymbol{\alpha}_0^1)$ is in the linear space spanned by $\boldsymbol{A}$. Since

$$E\left(\left[\boldsymbol{H}-\frac{R}{\pi^1(\boldsymbol{\alpha}_0^1)}E\{\boldsymbol{U}(\boldsymbol{\beta}_0)|\boldsymbol{X},\boldsymbol{S}\}\right]\left\{\frac{R}{\pi^1(\boldsymbol{\alpha}_0^1)}f(\boldsymbol{X},\boldsymbol{S})\right\}\right)=\mathbf{0}$$

for any function $f(\boldsymbol{X},\boldsymbol{S})$ and all components of $\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\boldsymbol{\gamma}_*)$ are functions of $\boldsymbol{X}$ and $\boldsymbol{S}$ only, we have that

$$\boldsymbol{L}\boldsymbol{G}^{-1}\boldsymbol{A}=E(\boldsymbol{H}\boldsymbol{A}^{\mathrm{T}})\{E(\boldsymbol{A}^{\otimes2})\}^{-1}\boldsymbol{A}=\frac{R}{\pi^1(\boldsymbol{\alpha}_0^1)}E\{\boldsymbol{U}(\boldsymbol{\beta}_0)|\boldsymbol{X},\boldsymbol{S}\}.$$

This fact yields that $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\boldsymbol{\gamma}_*)=E\{\boldsymbol{U}(\boldsymbol{\beta}_0)|\boldsymbol{X},\boldsymbol{S}\}$, and thus

$$\boldsymbol{Q}=\frac{R}{\pi^1(\boldsymbol{\alpha}_0^1)}\boldsymbol{U}(\boldsymbol{\beta}_0)-\frac{R-\pi^1(\boldsymbol{\alpha}_0^1)}{\pi^1(\boldsymbol{\alpha}_0^1)}E\{\boldsymbol{U}(\boldsymbol{\beta}_0)|\boldsymbol{X},\boldsymbol{S}\}.$$

On the other hand, simple calculation shows that now $E(\boldsymbol{Q}\boldsymbol{\Psi}^{\mathrm{T}})=\mathbf{0}$. The desired result then follows from Theorem 2. □

[Received July 2013. Revised December 2013.]

REFERENCES

Bang, H., and Robins, J. M. (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972. [1160]

Box, G. E. P., and Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*, New York: Wiley. [1167]

Cao, W., Tsiatis, A. A., and Davidian, M. (2009), "Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean With Incomplete Data," *Biometrika*, 96, 723–734. [1160,1163,1165,1170]

Chen, J., and Breslow, N. E. (2004), "Semiparametric Efficient Estimation for the Auxiliary Outcome Problem With Conditional Mean Model," *Canadian Journal of Statistics*, 32, 359–372. [1164]

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. [1161,1162]

Chen, J., Variyath, A. M., and Abraham, B. (2008), "Adjusted Empirical Likelihood and Its Properties," *Journal of Computational and Graphical Statistics*, 17, 426–443. [1160]

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, 70, 803–823. [1164,1171]

Davidian, M., Tsiatis, A. A., and Leon, S. (2005), "Semiparametric Estimation of Treatment Effect in a Pretest–Posttest Study With Missing Data," *Statistical Science*, 20, 261–301. [1169]

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundaker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C. (1996), "A Trial Comparing Nucleoside Monotherapy With Combination Therapy in HIV-Infected Adults With CD 4 Cell Counts From 200 to 500 Per Cubic Millimeter," *The New England Journal of Medicine*, 335, 1081–1089. [1168]

Han, P. (2012), "A Note on Improving the Efficiency of Inverse Probability Weighted Estimator Using the Augmentation Term," *Statistics and Probability Letters*, 82, 2221–2228. [1160,1165,1170]

——— (2014), "A Further Study of the Multiply Robust Estimator in Missing Data Analysis," *Journal of Statistical Planning and Inference*, 148, 101–110. [1160]

Han, P., and Wang, L. (2013), "Estimation With Missing Data: Beyond Double Robustness," *Biometrika*, 100, 417–430. [1160]

He, X., and Shao, Q. M. (1996), "A General Bahadur Representation of M-Estimators and its Application to Linear Regression With Nonstochastic Designs," *The Annals of Statistics*, 24, 2608–2630. [1171]

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685. [1159]

Kang, J. D. Y., and Schafer, J. L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean From Incomplete Data" (with discussion), *Statistical Science*, 22, 523–539. [1160,1163]

Koenker, R. (2005), *Quantile Regression*, Cambridge: Cambridge University Press. [1171]

Koenker, R., and Bassett, G. J. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [1159,1171]

Lipsitz, S., Fitzmaurice, G., Molenberghs, G., and Zhao, L. P. (1997), "Quantile Regression Methods for Longitudinal Data With Drop-Outs: Application to CD4 Cell Counts of Patients Infected With the Human Immunodeficiency Virus," *Applied Statistics*, 46, 463–476. [1171]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2 ed.), New York: Wiley. [1159]

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for A Single Functional," *Biometrika*, 75, 237–249. [1160]

——— (2001), *Empirical Likelihood*, New York: Chapman & Hall/CRC Press. [1160]

Pepe, M. S. (1992), "Inference Using Surrogate Outcome Data and a Validation Sample," *Biometrika*, 79, 355–365. [1163]

Pepe, M. S., Reilly, M., and Fleming, T. R. (1994), "Auxiliary Outcome Data and the Mean Score Method," *Journal of Statistical Planning and Inference*, 42, 137–160. [1159,1163]

Qin, J., and Lawless, J. (1994), "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22, 300–325. [1160,1164]

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. [1160]

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, 69, 101–122. [1160,1165]

Qin, J., Zhang, B., and Leung, D. H. Y. (2009), "Empirical Likelihood in Missing Data Problems," *Journal of the American Statistical Association*, 104, 1492–1503. [1160]

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regresion Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129. [1160]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. [1159,1163,1165]

——— (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121. [1160,1163,1164,1167]

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. [1160,1167]

Robins, J. M., and Wang, N. (2000), "Inference for Imputation Estimators," *Biometrika*, 87, 113–124. [1160]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [1159]

Rotnitzky, A. (2008), "Inverse Probability Weighted Methods," in *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, eds. G. Fitzmaurice M. Davidian G. Verbeke and G. Molenberghs pp. 453–476, New York: Chapman & Hall/CRC press. [1160]

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012), "Improved Double-Robust Estimation in Missing Data and Causal Inference Models," *Biometrika*, 99, 439–456. [1160,1165,1170]

Rotnitzky, A., and Robins, J. M. (1995), "Semiparametric Regression Estimation in the Presence of Dependent Censoring," *Biometrika*, 82, 805–820. [1160,1163,1164]

Rubin, D. B., and van der Laan, M. J. (2008), "Empirical Efficiency Maximization: Improved Locally Efficient Covariate Adjustment in Randomized Experiments and Survival Analysis," *International Journal of Biostatistics*, 4, article 5. [1160]

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. [1160]

Tan, Z. (2006), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of the American Statistical Association*, 101, 1619–1637. [1160,1170]

——— (2007), "Comment: Understanding OR, PS and DR," *Statistical Science*, 22, 560–568. [1164]

——— (2008), "Comment: Improved Local Efficiency and Double Robustness," *The International Journal of Biostatistics*, 4, Article 10. [1160,1165,1170]

——— (2010), "Bounded, Efficient and Doubly Robust Estimation With Inverse Weighting," *Biometrika*, 97, 661–682. [1160,1165,1170]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [1160,1163,1172]

Tsiatis, A. A., Davidian, M., and Cao, W. (2011), "Improved Doubly Robust Estimation When Data are Monotonely Coarsened, With Application to Longitudinal Studies With Dropout," *Biometrics*, 67, 536–545. [1160,1170]

van der Laan, M. J., and Gruber, S. (2010), "Collaborative Double Robust Targeted Maximum Likelihood Estimation," *The International Journal of Biostatistics*, 6, Article 17. [1160]

van der Laan, M. J., and Robins, J. M. (2003), *Unified Methods for Censored and Longitudinal Data and Causality*, New York: Springer. [1160]

Wang, D., and Chen, S. X. (2009), "Empirical Likelihood for Estimating Equations With Missing Values," *The Annals of Statistics*, 37, 490–517. [1160]

Wang, L., Rotnitzky, A., and Lin, X. (2010), "Nonparametric Regression With Missing Outcomes Using Weighted Kernel Estimating Equations," *Journal of the American Statistical Association*, 105, 1135–1146. [1159]

Wei, Y., Ma, Y., and Carroll, R. J. (2012), "Multiple Imputation in Quantile Regression," *Biometrika*, 99, 423–438. [1171]

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [1162]

Yi, G., and He, W. (2009), "Median Regression Models for Longitudinal Data With Dropouts," *Biometrics*, 65, 618–625. [1171]

Yu, M., and Nan, B. (2006), "A Revisit of Semiparametric Regression Models With Missing Data," *Statistca Sinica*, 16, 1193–1212. [1165]

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008), "Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates," *Biometrics*, 64, 707–715. [1169]