Biometrika (2013), *pp.* 1–14
doi: 10.1093/biomet/ass087

Printed in Great Britain

Estimation with missing data: beyond double robustness

BY PEISONG HAN AND LU WANG

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109, U.S.A.

peisong@umich.edu luwang@umich.edu

SUMMARY

We propose an estimator that is more robust than doubly robust estimators, based on weighting complete cases using weights other than inverse probability when estimating the population mean of a response variable subject to ignorable missingness. We allow multiple models for both the propensity score and the outcome regression. Our estimator is consistent if any of the multiple models is correctly specified. Such multiple robustness against model misspecification is a significant improvement over double robustness, which allows only one propensity score model and one outcome regression model. Our estimator attains the semiparametric efficiency bound when one propensity score model and one outcome regression model are correctly specified, without requiring knowledge of which models are correct.

Some key words: Augmented inverse probability weighting; Causal inference; Empirical likelihood; Missing at random; Semiparametric efficiency.

1. INTRODUCTION

We consider semiparametric estimation of the population mean of a response variable that is subject to ignorable missingness (Little & Rubin, 2002). This problem arises frequently in biomedical and social science studies, where the missingness could be due to noncompliance or nonresponse. Also, in causal inference, the counterfactual outcomes can be treated as missing when estimating the average treatment effects (Rubin, 1974). Apart from its practical importance, estimation of a population mean when there are missing responses provides a simple setting to explore the fundamental issues associated with missing data (Tsiatis & Davidian, 2007).

We assume that the missingness depends on certain fully observed covariates that can be correlated with the response, the so-called missing at random mechanism (Little & Rubin, 2002). A natural approach is to fit a regression model to the response given the covariates and then to take the mean of the fitted values from all subjects as an estimator. Another popular approach is to weight the complete cases using the inverse of their selection probabilities or estimated values, called the propensity score (Rosenbaum & Rubin, 1983), to correct for the selection bias due to missingness (Horvitz & Thompson, 1952). Although easy to implement, both approaches fail to produce a consistent estimator if the outcome regression or the propensity score model is misspecified. Robins et al. (1994) combined the two approaches and proposed the augmented inverse probability weighting method, which uses the outcome regression model as an augmentation term to the inverse probability weighting method. The augmented inverse probability weighted estimator attains the semiparametric efficiency bound when both the propensity score model and the outcome regression model are correctly specified (Robins & Rotnitzky, 1995). Scharfstein et al. (1999) noted that this locally efficient estimator has the double robustness

property, which guarantees estimation consistency if either of the two models is correctly specified; see also Bang & Robins (2005). Recently, various doubly robust estimators have been proposed and studied; see, for example, Tan (2006, 2008, 2010), Kang & Schafer (2007) and its discussion, Qin & Zhang (2007), Qin et al. (2008), Rubin & van der Laan (2008), Cao et al. (2009), Han (2012), and Rotnitzky et al. (2012). Most of these developments are based on efficiency concerns.

Double robustness does not provide sufficient protection for estimation consistency, as it allows only one model for the propensity score and one for the outcome regression. With an unknown data-generating process, it is often risky to assume that one of these two models is correctly specified. Therefore, multiple models may be fitted in practice, each involving different subsets of covariates and possibly different link functions, with none of them ruling out the possibility of others. Such multiple models increase the likelihood of correct specification. The question then is how to combine them. See Robins et al. (2007) for some related discussion.

In this paper, we construct a novel estimator by making the weighted complete-case mean of each postulated model and the corresponding unweighted sample mean equal. As a result, our estimator takes multiple models into account. When appropriately calculated, the proposed estimator is consistent if any one of those multiple models, for either the propensity score or the outcome regression, is correctly specified. This multiple robustness property appears superior to the double robustness property. Our proof of consistency of the estimator makes intensive use of empirical likelihood theory (Owen, 1988, 2001; Qin & Lawless, 1994), which has become a popular tool in the missing data literature; see, for example, Tan (2006, 2010), Qin & Zhang (2007), Qin et al. (2008) and Chen et al. (2008). We derive the asymptotic distribution of the proposed estimator and suggest ways to improve its efficiency. Our estimator attains the semiparametric efficiency bound if one propensity score model and one outcome regression model are correctly specified, without requiring knowledge of which two models are correct. Some caution is needed in the numerical implementation, as the calculation of our estimator requires choosing one root from possibly multiple roots of an equation, and only when the appropriate root is chosen does our estimator have the multiple robustness property. In § 5 we discuss this in detail and provide suggestions for numerical implementation.

2. THE PROPOSED ESTIMATOR

Let Y denote the response variable, X the vector of covariates, and n the sample size. Let $R = 1$ if Y is observed and $R = 0$ if Y is missing. We assume that $\mathrm{pr}(R = 1 \mid Y = y, X = x) = \mathrm{pr}(R = 1 \mid X = x)$ and denote this probability by $\pi(x)$, which is called the propensity score function. The observed data are independently and identically distributed triples $(R_i Y_i, X_i, R_i)$ $(i = 1, \ldots, n)$. Let $m = \sum_{i=1}^{n} R_i$ be the number of subjects who have their response observed, and index those subjects by $i = 1, \ldots, m$ without loss of generality. We wish to estimate the quantity of interest, $\mu_0 = E(Y)$, using a weighted average of the observed responses, $\sum_{i=1}^{m} w_i Y_i$, where w_i is the weight assigned to subject i. Because the subjects $i = 1, \ldots, m$ form a biased sample from the underlying population, the weights should be chosen so that population moments can be recovered based on this sample. When $\pi(x)$ is known or can be correctly modelled, the inverse probability weighted estimator $\hat{\mu}_{\mathrm{ipw}} = n^{-1} \sum_{i=1}^{n} R_i Y_i / \hat{\pi}(X_i)$ takes w_i to be $1/\{n\hat{\pi}(X_i)\}$, where $\hat{\pi}(x)$ is the true or estimated value of $\pi(x)$. The augmented inverse probability weighted estimator $\hat{\mu}_{\mathrm{aipw}} = n^{-1} \sum_{i=1}^{n} [R_i Y_i / \hat{\pi}(X_i) - \{R_i / \hat{\pi}(X_i) - 1\}\hat{a}(X_i)]$ introduces an extra term that involves an outcome regression model $\hat{a}(x)$ for $a(x) = E(Y \mid X = x)$. The estimator $\hat{\mu}_{\mathrm{aipw}}$ is consistent when either $\pi(x)$ or $a(x)$ is correctly modelled; this is known as the double robustness

property. When neither $\pi(x)$ nor $a(x)$ is correctly modelled, $\hat{\mu}_{\text{aipw}}$ loses consistency. The same is true of all existing doubly robust estimators.

To improve on double robustness, we postulate multiple models $\mathcal{P} = \{\pi^j(\alpha^j; x) : j = 1, \ldots, J\}$ for $\pi(x)$ and multiple models $\mathcal{A} = \{a^k(\gamma^k; x) : k = 1, \ldots, K\}$ for $a(x)$. Here the α^j and γ^k are the corresponding parameters, and we use $\hat{\alpha}^j$ and $\hat{\gamma}^k$ to denote their estimators. Usually, each $\hat{\alpha}^j$ is taken to be the maximizer of the binomial likelihood

$$\prod_{i=1}^{n} \{\pi^j(\alpha^j; X_i)\}^{R_i} \{1 - \pi^j(\alpha^j; X_i)\}^{1-R_i}, \tag{1}$$

and each $\hat{\gamma}^k$ is taken to be the regression coefficient of a generalized linear model for $a(x)$ based on complete-case analysis. Define $\hat{\theta}^j = n^{-1} \sum_{i=1}^{n} \pi^j(\hat{\alpha}^j; X_i)$ and $\hat{\eta}^k = n^{-1} \sum_{i=1}^{n} a^k(\hat{\gamma}^k; X_i)$. In order to recover population moments from the biased sample $\{i : i = 1, \ldots, m\}$, we impose the following constraints on the weights w_i:

$$\begin{aligned} &\sum_{i=1}^{m} w_i = 1, \quad \sum_{i=1}^{m} w_i \pi^j(\hat{\alpha}^j; X_i) = \hat{\theta}^j \quad (j = 1, \ldots, J), \\ &\sum_{i=1}^{m} w_i a^k(\hat{\gamma}^k; X_i) = \hat{\eta}^k \quad (k = 1, \ldots, K). \end{aligned} \tag{2}$$

Here, the first constraint is imposed for regularization, and the second and third constraints equate the weighted average of each postulated parametric function evaluated at the biased sample to the corresponding unweighted sample mean, which consistently estimates the population mean. The constraints in (2) are linear in w_i and are similar to those used when calibrating sample surveys (Deville & Särndal, 1992; Wu & Sitter, 2001). A sufficient condition for the existence of a solution to (2) is that $J + K + 1 \leqslant m$.

To introduce weights that satisfy (2), write

$$\hat{\alpha}^{\mathrm{T}} = \{(\hat{\alpha}^1)^{\mathrm{T}}, \ldots, (\hat{\alpha}^J)^{\mathrm{T}}\}, \quad \hat{\gamma}^{\mathrm{T}} = \{(\hat{\gamma}^1)^{\mathrm{T}}, \ldots, (\hat{\gamma}^K)^{\mathrm{T}}\},$$

$$\hat{g}_i(\hat{\alpha}, \hat{\gamma}) = \{\pi^1(\hat{\alpha}^1; X_i) - \hat{\theta}^1, \ldots, \pi^J(\hat{\alpha}^J; X_i) - \hat{\theta}^J, a^1(\hat{\gamma}^1; X_i) - \hat{\eta}^1, \ldots, a^K(\hat{\gamma}^K; X_i) - \hat{\eta}^K\}^{\mathrm{T}}.$$

If $\hat{\rho}^{\mathrm{T}} = (\hat{\rho}_1, \ldots, \hat{\rho}_{J+K})$ is a $(J + K)$-dimensional vector satisfying the equation

$$\sum_{i=1}^{m} \frac{\hat{g}_i(\hat{\alpha}, \hat{\gamma})}{1 + \rho^{\mathrm{T}} \hat{g}_i(\hat{\alpha}, \hat{\gamma})} = 0, \tag{3}$$

then it is easy to check that

$$\hat{w}_i = \frac{1}{m} \frac{1}{1 + \hat{\rho}^{\mathrm{T}} \hat{g}_i(\hat{\alpha}, \hat{\gamma})} \Bigg/ \left\{ \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 + \hat{\rho}^{\mathrm{T}} \hat{g}_i(\hat{\alpha}, \hat{\gamma})} \right\} \quad (i = 1, \ldots, m) \tag{4}$$

satisfy (2). Our proposed estimator is $\hat{\mu} = \sum_{i=1}^{m} \hat{w}_i Y_i$.

In general, multiple roots exist for (3), and not all of them yield $\hat{w}_i$ that endow $\hat{\mu}$ with the desired properties. Nevertheless, by carefully choosing $\hat{\rho}$, $\hat{\mu}$ will be consistent if either $\mathcal{P}$ contains a correctly specified model for $\pi(x)$ or $\mathcal{A}$ contains a correctly specified model for $a(x)$, as will be discussed in the next section.

3. MULTIPLE ROBUSTNESS

3·1. *Consistency when $\pi(x)$ is correctly modelled*

Suppose that $\mathcal{P}$ contains a correctly specified model for $\pi(x)$, say $\pi^1(\alpha^1; x)$ without loss of generality. Let α_0^1 denote the true value of α^1 such that $\pi^1(\alpha_0^1; x) = \pi(x)$, and let $\hat{\alpha}^1$ denote the maximizer of (1) with $j = 1$. We employ empirical likelihood theory to prove the consistency of $\hat{\mu}$.

Define the empirical probability of (Y_i, X_i) conditional on $R_i = 1$ to be p_i $(i = 1, \ldots, m)$. Conditional on $R = 1$, an argument in the Appendix yields

$$E\left[\frac{\pi^j(\alpha^j; X) - E\{\pi^j(\alpha^j; X)\}}{\pi(X)} \,\middle|\, R = 1\right] = 0 \quad (j = 1, \ldots, J),$$
$$E\left[\frac{a^k(\gamma^k; X) - E\{a^k(\gamma^k; X)\}}{\pi(X)} \,\middle|\, R = 1\right] = 0 \quad (k = 1, \ldots, K) \tag{5}$$

so the most plausible value for p_i should be defined through the constrained optimization problem

$$\max_{p_1,\ldots,p_m} \prod_{i=1}^m p_i \quad \text{subject to} \quad p_i \geqslant 0 \ (i = 1, \ldots, m), \quad \sum_{i=1}^m p_i = 1,$$
$$\sum_{i=1}^m p_i \frac{\pi^j(\hat{\alpha}^j; X_i) - \hat{\theta}^j}{\pi^1(\hat{\alpha}^1; X_i)} = 0 \quad (j = 1, \ldots, J),$$
$$\sum_{i=1}^m p_i \frac{a^k(\hat{\gamma}^k; X_i) - \hat{\eta}^k}{\pi^1(\hat{\alpha}^1; X_i)} = 0 \quad (k = 1, \ldots, K). \tag{6}$$

The first two constraints make the p_i empirical probabilities, and the last two are the empirical versions of the statements in (5), using the knowledge that $\pi^1(\alpha^1; x)$ is correctly specified.

Applying Lagrange multipliers, we have

$$\hat{p}_i = \frac{1}{m} \frac{1}{1 + \hat{\lambda}^{\mathrm{T}} \hat{g}_i(\hat{\alpha}, \hat{\gamma})/\pi^1(\hat{\alpha}^1; X_i)},$$

where $\hat{\lambda}^{\mathrm{T}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_{J+K})$ is the $(J + K)$-dimensional vector of Lagrange multipliers that solves the equation

$$\sum_{i=1}^m \frac{\hat{g}_i(\hat{\alpha}, \hat{\gamma})/\pi^1(\hat{\alpha}^1; X_i)}{1 + \lambda^{\mathrm{T}} \hat{g}_i(\hat{\alpha}, \hat{\gamma})/\pi^1(\hat{\alpha}^1; X_i)} = 0. \tag{7}$$

Comparison of equations (3) and (7) reveals that a solution to (3) is given by $\hat{\rho}^\pi$ with $\hat{\rho}_1^\pi = (\hat{\lambda}_1 + 1)/\hat{\theta}^1$ and $\hat{\rho}_l^\pi = \hat{\lambda}_l/\hat{\theta}^1$ for $l = 2, \ldots, J + K$. Inserting $\hat{\rho}^\pi$ into (4) yields

$$\hat{w}_i = \frac{1}{m} \frac{\hat{\theta}^1/\pi^1(\hat{\alpha}^1; X_i)}{1 + \hat{\lambda}^{\mathrm{T}} \hat{g}_i(\hat{\alpha}, \hat{\gamma})/\pi^1(\hat{\alpha}^1; X_i)} \Bigg/ \left\{\frac{1}{m} \sum_{i=1}^m \frac{\hat{\theta}^1/\pi^1(\hat{\alpha}^1; X_i)}{1 + \hat{\lambda}^{\mathrm{T}} \hat{g}_i(\hat{\alpha}, \hat{\gamma})/\pi^1(\hat{\alpha}^1; X_i)}\right\}. \tag{8}$$

On the other hand, from (6) we have

$$1 - \frac{1}{m} \sum_{i=1}^m \frac{\hat{\theta}^1/\pi^1(\hat{\alpha}^1; X_i)}{1 + \hat{\lambda}^{\mathrm{T}} \hat{g}_i(\hat{\alpha}, \hat{\gamma})/\pi^1(\hat{\alpha}^1; X_i)} = 1 - \sum_{i=1}^m \frac{\hat{p}_i \hat{\theta}^1}{\pi^1(\hat{\alpha}^1; X_i)} = \sum_{i=1}^m \hat{p}_i \frac{\pi^1(\hat{\alpha}^1; X_i) - \hat{\theta}^1}{\pi^1(\hat{\alpha}^1; X_i)} = 0.$$

Therefore, when $\pi^1(\alpha^1; x)$ is correctly specified, $\hat{w}_i$ given by (4) with $\hat{\rho} = \hat{\rho}^{\pi}$ is equal to $\hat{p}_i\hat{\theta}^1/\pi^1(\hat{\alpha}^1; X_i)$, and our estimator becomes $\hat{\mu} = \sum_{i=1}^n R_i\hat{p}_i\hat{\theta}^1 Y_i/\pi^1(\hat{\alpha}^1; X_i)$. Compared to $\hat{\mu}_{\mathrm{ipw}}$, $\hat{\mu}$ incorporates the empirical probabilities $\hat{p}_i$ as extra weights.

It is apparent that the weights $\hat{w}_i = \hat{p}_i\hat{\theta}^1/\pi^1(\hat{\alpha}^1; X_i)$ $(i = 1, \ldots, m)$ are all positive. In fact, when $\pi^1(\alpha^1; x)$ is correctly specified, these weights can be derived by directly maximizing $\prod_{i=1}^m w_i$ subject to a nonnegativity constraint on the w_i in addition to the constraints in (2). This derivation includes that of Qin & Zhang (2007) as a special case.

Using the results of White (1982), we know that for $j = 1, \ldots, J$ and $k = 1, \ldots, K$, $\hat{\alpha}^j \to \alpha_*^j$ and $\hat{\gamma}^k \to \gamma_*^k$ in probability, where α_*^j and γ_*^k are the least false values that minimize the corresponding Kullback–Leibler distance between the probability distribution based on the postulated model and the one that generates the data. Write $\alpha_*^{\mathrm{T}} = \{(\alpha_*^1)^{\mathrm{T}}, \ldots, (\alpha_*^J)^{\mathrm{T}}\}$ and $\gamma_*^{\mathrm{T}} = \{(\gamma_*^1)^{\mathrm{T}}, \ldots, (\gamma_*^K)^{\mathrm{T}}\}$. In general, $\pi^j(\alpha_*^j; x) \neq \pi(x)$ unless $\pi^j(\alpha^j; x)$ is a correctly specified model for $\pi(x)$, and $a^k(\gamma_*^k; x) \neq a(x)$ unless $a^k(\gamma^k; x)$ is a correctly specified model for $a(x)$. Since $\pi^1(\alpha^1; x)$ is assumed to be correctly specified, we have $\alpha_*^1 = \alpha_0^1$. In addition, we have $\hat{\theta}^j \to \theta_*^j$ and $\hat{\eta}^k \to \eta_*^k$ in probability under suitable regularity conditions, where $\theta_*^j = E\{\pi^j(\alpha_*^j; X)\}$ and $\eta_*^k = E\{a^k(\gamma_*^k; X)\}$. Because $\hat{\lambda} \to 0$ in probability based on empirical likelihood theory, Lemma A1 of the Appendix yields $\hat{\lambda} = O_p(n^{-1/2})$. Therefore, since $\hat{w}_i$ can be expressed as (8), simple algebra gives $\hat{\mu} = n^{-1}\sum_{i=1}^n\{R_i Y_i/\pi(X_i)\} + o_p(1) \to \mu_0$ in probability. Hence, we have the following result on the consistency of $\hat{\mu}$.

THEOREM 1. *When $\mathcal{P}$ contains a correctly specified model for $\pi(x)$ and (4) is evaluated at $\hat{\rho} = \hat{\rho}^{\pi}$, $\hat{\mu}$ is a consistent estimator of μ_0 as $n \to \infty$.*

3·2. *Consistency when $a(x)$ is correctly modelled*

Now suppose that $\mathcal{A}$ contains a correctly specified model for $a(x)$, say $a^1(\gamma^1; x)$ without loss of generality. Let γ_0^1 denote the true value of γ^1 such that $a^1(\gamma_0^1; x) = a(x)$; then $\gamma_*^1 = \gamma_0^1$. Unless $\mathcal{P}$ also contains a correctly specified model for $\pi(x)$, the derivation of the w_i given in § 3·1 is not applicable here. To prove consistency of $\hat{\mu}$ in this case, suppose that there exists a solution $\hat{\rho}^a$ to (3) which has a finite probability limit ρ_*^a and that (4) is evaluated at $\hat{\rho}^a$. Under suitable regularity conditions, some algebra shows that

$$\begin{aligned}\hat{\mu} &= \sum_{i=1}^m \hat{w}_i\{Y_i - a^1(\hat{\gamma}^1; X_i)\} + \frac{1}{n}\sum_{i=1}^n a^1(\hat{\gamma}^1; X_i)\\ &\to E\left[\frac{R\{Y - a(X)\}}{1 + (\rho_*^a)^{\mathrm{T}} g(\alpha_*, \gamma_*; X)}\right] \Big/ E\left\{\frac{R}{1 + (\rho_*^a)^{\mathrm{T}} g(\alpha_*, \gamma_*; X)}\right\} + \mu_0 = \mu_0\end{aligned}$$

in probability, where

$$\begin{aligned}&g(\alpha_*, \gamma_*; X)\\ &\quad = \{\pi^1(\alpha_*^1; X) - \theta_*^1, \ldots, \pi^J(\alpha_*^J; X) - \theta_*^J, a^1(\gamma_*^1; X) - \eta_*^1, \ldots, a^K(\gamma_*^K; X) - \eta_*^K\}^{\mathrm{T}}. \quad (9)\end{aligned}$$

Hence, we have the following result on the consistency of $\hat{\mu}$.

THEOREM 2. *When $\mathcal{A}$ contains a correctly specified model for $a(x)$ and (4) is evaluated at $\hat{\rho} = \hat{\rho}^a$, where $\hat{\rho}^a$ solves (3) and has a finite probability limit, $\hat{\mu}$ is a consistent estimator of μ_0 as $n \to \infty$.*

In some applications, rather than postulating parametric models, several candidate functions for $a(x)$ may be suggested directly. In other words, we have $\mathcal{A} = \{a^k(x) : k = 1, \ldots, K\}$ where no unknown parameters are to be estimated. If this is the case, the following corollary states that the consistency of $\hat{\mu}$ may be achieved even if $\mathcal{A}$ does not contain $a(x)$. The proof is omitted, as it is similar to that of Theorem 2.

COROLLARY 1. *When $a(x)$ is a linear combination of functions in $\mathcal{A} = \{a^k(x) : k = 1, \ldots, K\}$ and (4) is evaluated at $\hat{\rho} = \hat{\rho}^a$, where $\hat{\rho}^a$ solves (3) and has a finite probability limit, $\hat{\mu}$ is a consistent estimator of μ_0 as $n \to \infty$.*

We can specify the functions $a^1(x), \ldots, a^K(x)$ to be linearly independent without affecting the above result. When $a(x)$ is a linear combination of $a^1(x), \ldots, a^K(x)$, Corollary 1 implies that the consistency of $\hat{\mu}$ can be achieved without estimating the coefficients of the linear combination, as long as all the components of $a(x)$ are used in (2). In practice, however, we do not recommend this approach, as it could significantly increase the number of constraints in (2) and thus the dimension of equation (3), causing numerical problems.

4. ASYMPTOTIC DISTRIBUTION AND EFFICIENCY

The asymptotic distribution of $\hat{\mu}$ depends on which of the $J + K$ models is correctly specified. Although the propensity score models and the outcome regression models are treated equally when constructing $\hat{\mu}$ in § 2, exploration of their asymmetry helps us derive the asymptotic distribution. As pointed out by Tan (2007), it is more constructive to view doubly robust estimation in the inverse probability weighting approach by incorporating an outcome regression model. Indeed, semiparametric theory for estimation with missing data has been developed from such a point of view (Robins et al., 1994; Robins & Rotnitzky, 1995; Tsiatis, 2006). In addition, under the missing at random assumption, propensity score models are fitted based on fully observed data, so their goodness-of-fit can be tested. This is not true of the outcome regression models, which are thus more likely to be misspecified. Therefore, to derive the asymptotic distribution of $\hat{\mu}$, we assume that one model for $\pi(x)$ is correctly specified, and take it to be $\pi^1(\alpha^1; x)$ without loss of generality.

Let $S(\alpha^1; X, R)$ denote the score function corresponding to the binomial likelihood (1); that is,

$$S(\alpha^1; X, R) = \frac{R - \pi^1(\alpha^1; X)}{\pi^1(\alpha^1; X)\{1 - \pi^1(\alpha^1; X)\}} \frac{\partial \pi^1(\alpha^1; X)}{\partial \alpha^1}.$$

In addition, write $S = S(\alpha_0^1; X, R)$, $L = E\{(Y - \mu_0)g(\alpha_*, \gamma_*; X)/\pi(X)\}$, $G = E\{g(\alpha_*, \gamma_*; X)^{\otimes 2}/\pi(X)\}$, and

$$U = \frac{R(Y - \mu_0)}{\pi(X)} - \frac{R - \pi(X)}{\pi(X)} L^{\mathrm{T}} G^{-1} g(\alpha_*, \gamma_*; X),$$

where, for any matrix B, $B^{\otimes 2} = BB^{\mathrm{T}}$. The following theorem gives the asymptotic distribution of $\hat{\mu}$.

THEOREM 3. *When $\pi^1(\alpha^1; x)$ is a correctly specified model for $\pi(x)$ and (4) is evaluated at $\hat{\rho} = \hat{\rho}^\pi$, $n^{1/2}(\hat{\mu} - \mu_0)$ has an asymptotic normal distribution with mean 0 and variance* $\mathrm{var}(Z)$*, where*

$$Z = U - E(US^{\mathrm{T}})\{E(SS^{\mathrm{T}})\}^{-1} S. \tag{10}$$

It is clear from (10) that Z is the residual of the projection of U onto S, so $\mathrm{var}(Z) \leqslant \mathrm{var}(U)$. Therefore, one way to improve the efficiency of $\hat{\mu}$ is to augment the model $\pi^1(\alpha^1; x)$. For example, interactions and higher-order terms of the components of x can be added when fitting $\pi^1(\alpha^1; x)$, in which case the dimension of α^1 increases, and so does the dimension of $S(\alpha^1; X, R)$. Consequently, the residual Z is likely to have smaller variance. When missing responses are caused by study design, as in the case of two-stage design studies (Pepe, 1992; Pepe et al., 1994), $\pi(x)$ is known. If the known $\pi(x)$ is used in calculating the $\hat{w}_i$, following arguments similar to those in the proof of Theorem 3, it can be shown that the asymptotic variance of $n^{1/2}(\hat{\mu} - \mu_0)$ is $\mathrm{var}(U)$. Since $\mathrm{var}(U) \geqslant \mathrm{var}(Z)$, the efficiency of $\hat{\mu}$ can be improved by modelling $\pi(x)$ even when $\pi(x)$ is known. This counter-intuitive fact has been studied by Robins et al. (1995).

Write $H = R(Y - \mu_0)/\pi(X)$ and $Q = Rg(\alpha_*, \gamma_*; X)/\pi(X)$. It is clear that $L^{\mathrm{T}}G^{-1} = E(HQ^{\mathrm{T}})\{E(QQ^{\mathrm{T}})\}^{-1}$ is the coefficient of the projection of H onto Q. When $\mathcal{A}$ contains a correctly specified model for $a(x)$, i.e., when $a(X) - \mu_0$ is a component of $g(\alpha_*, \gamma_*; X)$, it is easy to verify that the projection of H onto Q is $R\{a(X) - \mu_0\}/\pi(X)$, which gives $L^{\mathrm{T}}G^{-1}g(\alpha_*, \gamma_*; X) = a(X) - \mu_0$. In this case, we have

$$U = U_{\mathrm{opt}} = \frac{R(Y - \mu_0)}{\pi(X)} - \frac{R - \pi(X)}{\pi(X)}\{a(X) - \mu_0\}.$$

A simple calculation shows that $E(U_{\mathrm{op}}S^{\mathrm{T}}) = 0$, so the asymptotic variance of $n^{1/2}(\hat{\mu} - \mu_0)$ is $\mathrm{var}(U_{\mathrm{opt}})$, which is equal to the semiparametric efficiency bound (Robins & Rotnitzky, 1995; Tsiatis, 2006). This yields the following theorem on the efficiency of $\hat{\mu}$.

THEOREM 4. *When $\mathcal{P}$ contains a correctly specified model for $\pi(x)$ and $\mathcal{A}$ contains a correctly specified model for $a(x)$, if (4) is evaluated at $\hat{\rho} = \hat{\rho}^{\pi}$, then the asymptotic variance of $\hat{\mu}$ attains the semiparametric efficiency bound.*

The efficiency given in Theorem 4 is different from the local efficiency of the augmented inverse probability weighted estimator, as it is achieved without knowing which two among the $J + K$ models are correctly specified. This is similar to the oracle property (Fan & Li, 2001) in the variable selection literature, so we call it local oracle efficiency. When $\mathcal{A} = \{a^k(x) : k = 1, \ldots, K\}$, oracle efficiency of $\hat{\mu}$ may be achieved even if $\mathcal{A}$ does not contain $a(x)$, as implied by the following corollary, the proof of which is omitted due to its similarity to that of Theorem 4.

COROLLARY 2. *When $\mathcal{P}$ contains a correctly specified model for $\pi(x)$, $\mathcal{A} = \{a^k(x) : k = 1, \ldots, K\}$ and $a(x)$ is a linear combination of functions in $\mathcal{A}$, if (4) is evaluated at $\hat{\rho} = \hat{\rho}^{\pi}$, then the asymptotic variance of $\hat{\mu}$ attains the semiparametric efficiency bound.*

In practice, the asymptotic distributions derived in this section can be utilized for statistical inference when $\pi(x)$ is known by design. In this case, the estimated value of $\pi(x)$ based on model fitting is still preferred to the true $\pi(x)$ because of possible efficiency improvements. The expectations involved in the asymptotic variance can be estimated by corresponding sample averages. In the more general case where the correct models for $\pi(x)$ and $a(x)$ are both unknown, the results in this section cannot be employed for inference. As an alternative, we suggest bootstrapping to calculate the standard error of $\hat{\mu}$.

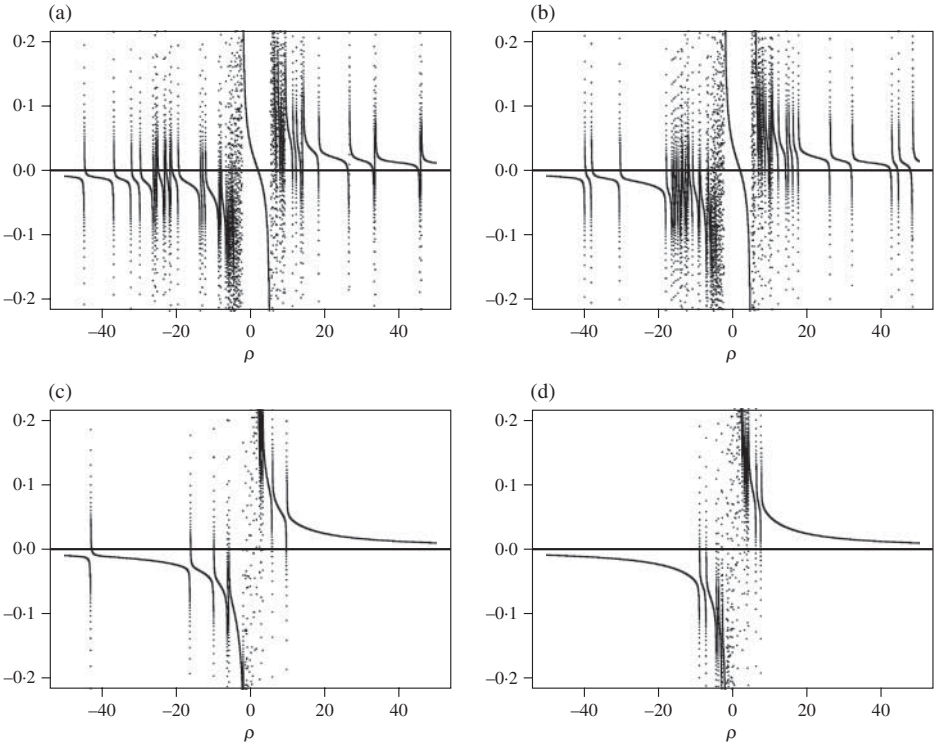

Fig. 1. Plots of $n^{-1} f_n(\rho)$ with $n = 300$ in four scenarios: (a) $\hat{g}(\hat{\alpha}, \hat{\gamma}) = \pi^1(\hat{\alpha}^1; X) - \hat{\theta}^1$, correct propensity score model; (b) $\hat{g}(\hat{\alpha}, \hat{\gamma}) = \pi^2(\hat{\alpha}^2; X) - \hat{\theta}^2$, incorrect propensity score model; (c) $\hat{g}(\hat{\alpha}, \hat{\gamma}) = a^1(\hat{\gamma}^1; X) - \hat{\eta}^1$, correct outcome regression model; (d) $\hat{g}(\hat{\alpha}, \hat{\gamma}) = a^2(\hat{\gamma}^2; X) - \hat{\eta}^2$, incorrect outcome regression model.

5. NUMERICAL IMPLEMENTATION

There is a subtle yet important numerical issue concerning the multiple robustness implied by Theorems 1 and 2: how to choose an appropriate $\hat{\rho}$ among the possibly multiple roots of (3) to evaluate (4). We conduct a simple numerical experiment. Let $X \sim \mathrm{Un}(-2{\cdot}5, 2{\cdot}5)$, $Y \mid X = x \sim N\{a(x), 4x^2 + 2\}$ and $R \mid X = x \sim \mathrm{Ber}\{\pi(x)\}$, where $a(x) = 1 + 2x + 3x^2$ and $\mathrm{logit}\{\pi(x)\} = -0{\cdot}8 - 0{\cdot}5x + 0{\cdot}3x^2$. Let $\pi^1(\alpha^1; x)$ be correctly specified for $\pi(x)$ and let $a^1(\gamma^1; x)$ be correctly specified for $a(x)$. We also postulate two incorrectly specified models, $\pi^2(\alpha^2; x) = 1 - \exp[-\exp\{\alpha_1^2 + \alpha_2^2 x + \alpha_3^2 \exp(x)\}]$ for $\pi(x)$ and $a^2(\gamma^2; x) = \gamma_1^2 + \gamma_2^2 x + \gamma_3^2 \exp(x)$ for $a(x)$. For illustration, we consider only the four scenarios where ρ is a scalar, corresponding to $\hat{g}(\hat{\alpha}, \hat{\gamma})$ being equal to $\pi^1(\hat{\alpha}^1; X) - \hat{\theta}^1$, $\pi^2(\hat{\alpha}^2; X) - \hat{\theta}^2$, $a^1(\hat{\gamma}^1; X) - \hat{\eta}^1$ and $a^2(\hat{\gamma}^2; X) - \hat{\eta}^2$, respectively.

Let $f_n(\rho)$ denote the function on the left-hand side of (3). Figure 1 shows plots of $n^{-1} f_n(\rho)$ in the four scenarios when $n = 300$. In each plot, the branches in the middle resemble the cotangent function, while the branches at the ends resemble the reciprocal function with the x-axis as one asymptote. It is clear from the multiple crossings with the x-axis that, for our simulated data, (3) has multiple roots in all four scenarios. As n increases, under suitable regularity conditions, the probability limit of the root(s) of (3) must be the root of

$$E\left\{\frac{Rg(\alpha_*, \gamma_*; X)}{1 + \rho^{\mathrm{T}} g(\alpha_*, \gamma_*; X)}\right\} = 0. \tag{11}$$

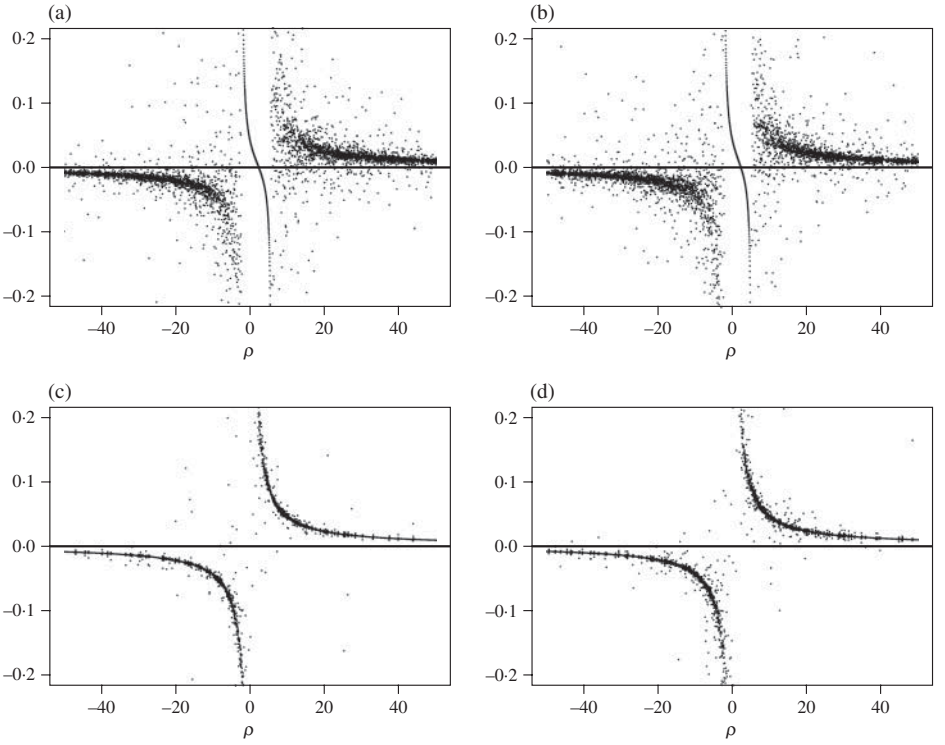

Fig. 2. Plots of $n^{-1} f_n(\rho)$ with $n = 30\,000$ in four scenarios: (a) $\hat{g}(\hat{\alpha}, \hat{\gamma}) = \pi^1(\hat{\alpha}^1; X) - \hat{\theta}^1$, correct propensity score model; (b) $\hat{g}(\hat{\alpha}, \hat{\gamma}) = \pi^2(\hat{\alpha}^2; X) - \hat{\theta}^2$, incorrect propensity score model; (c) $\hat{g}(\hat{\alpha}, \hat{\gamma}) = a^1(\hat{\gamma}^1; X) - \hat{\eta}^1$, correct outcome regression model; (d) $\hat{g}(\hat{\alpha}, \hat{\gamma}) = a^2(\hat{\gamma}^2; X) - \hat{\eta}^2$, incorrect outcome regression model.

Equation (11) is the population version of (3). When $\pi^1(\alpha^1; x)$ is correctly specified, it is easy to check that $\rho_*^{\pi} = (1/\theta_0^1, 0, \ldots, 0)^{\mathrm{T}}$ solves (11), in addition to the trivial root ∞, where $\theta_0^1 = E\{\pi(X)\} = \mathrm{pr}(R = 1)$. This can be seen in Fig. 2(a), which contains the same plots as in Fig. 1 but with $n = 30\,000$; observe that the number of branches seems to have reduced to three, with the middle branch having an obvious crossing with the x-axis at around $\rho = n/m \approx 2{\cdot}22$, which is an estimated value of $1/\theta_0^1$, and the other two branches having the x-axis as the asymptote. When no model in $\mathcal{P}$ is correctly specified, however, (11) may or may not have a nontrivial root. For the scenario where $\hat{g}(\hat{\alpha}, \hat{\gamma})$ is based on $\pi^2(\alpha^2; x)$, Fig. 2(b) seems to indicate that a nontrivial root exists. Although this plot looks similar to Fig. 2(a), the nontrivial root is different from $1/\theta_0^1$. For the two scenarios where $\hat{g}(\hat{\alpha}, \hat{\gamma})$ is based on $a^1(\gamma^1; x)$ or $a^2(\gamma^2; x)$, Figs. 2(c) and (d) seem to indicate that (11) has only the trivial root ∞. These observations demonstrate that although (3) may have multiple roots, their convergence in probability to a finite limit may not be guaranteed when $\pi(x)$ is incorrectly modelled. This explains why, in Theorem 2 and Corollary 1, we explicitly assumed that a finite probability limit exists or, equivalently, that a finite solution to (11) exists.

Since it is hard, in general, to study the behaviour of the left-hand side of (11) as a function of ρ when $\pi(x)$ is not correctly modelled, the existence of a finite solution to (11), and hence the assumption made in Theorem 2 and Corollary 1, is difficult to ascertain. Even if such a finite solution exists, we may still lack an effective way to identify the $\hat{\rho}^a$ that converges to it in probability among multiple roots of (3). However, when $\pi^1(\alpha^1; x)$ is correctly specified, $\hat{\rho}^{\pi}$

is easy to identify, because it must be close to $(n/m, 0, \ldots, 0)^{\mathrm{T}}$, owing to the fact that both $\hat{\rho}^{\pi}$ and $(n/m, 0, \ldots, 0)^{\mathrm{T}}$ have probability limit ρ_*^{π}. Therefore, we recommend implementing our proposed estimator in a similar fashion to implementation of a doubly robust estimator. We start by carefully selecting multiple propensity score models so that the true one is likely to be included in $\mathcal{P}$; then we incorporate multiple outcome regression models in the hope of making efficiency improvements. Since we do not know which propensity score model is correctly specified, it is better to try different initial values ne^j/m $(j = 1, \ldots, J)$ when solving (3), where each e^j is a $(J + K)$-dimensional vector with jth component equal to 1 and all other components equal to 0. These initial values may lead to different points of algorithmic convergence, corresponding to different roots of (3). If one converging point yields $\hat{w}_i$ that are all positive, these $\hat{w}_i$ should be used to calculate $\hat{\mu}$. Otherwise, one may calculate the $\hat{w}_i$ by simply using the converging point that makes $f_n(\rho)$ the closest to 0.

The aim of postulating multiple models is to best approximate the true but unknown model that generates the data. In general, each model should be constructed to be as close to the true one as possible. Although the aggregate effect suffices for our theoretical results under the special scenario considered in the two corollaries, we still recommend achieving good individual-model approximation to control the total number of models $J + K$, as too many models will jeopardize the numerical performance of our estimator.

6. Simulation study

In this section, we study the finite-sample performance of $\hat{\mu}$ and make comparisons with $\hat{\mu}_{\mathrm{aipw}}$. The data are generated with $X \sim \mathrm{Un}(-2{\cdot}5, 2{\cdot}5)$, $Y \mid X = x \sim N\{a(x), 4x^2 + 2\}$ and $R \mid X = x \sim \mathrm{Ber}\{\pi(x)\}$ under four $\{\pi(x), a(x)\}$ combinations, where $\pi(x) = \{1 + \exp(0{\cdot}8 + 0{\cdot}5x - 0{\cdot}3x^2)\}^{-1}$ or $1 - \exp[-\exp\{0{\cdot}5 + 0{\cdot}5x - 0{\cdot}3\exp(x)\}]$ and $a(x) = 1 + 2x + 3x^2$ or $1 + 2x + 3\exp(x)$. For each combination, we postulate two propensity score models, $\pi^1(\alpha^1; x) = \{1 + \exp(\alpha_1^1 + \alpha_2^1 x + \alpha_3^1 x^2)\}^{-1}$ and $\pi^2(\alpha^2; x) = 1 - \exp[-\exp\{\alpha_1^2 + \alpha_2^2 x + \alpha_3^2 \exp(x)\}]$, and two outcome regression models, $a^1(\gamma^1; x) = \gamma_1^1 + \gamma_2^1 x + \gamma_3^1 x^2$ and $a^2(\gamma^2; x) = \gamma_1^2 + \gamma_2^2 x + \gamma_3^2 \exp(x)$. We use a four-digit subscript to distinguish estimators constructed using different postulated models; from left to right, each digit indicates whether or not $\pi^1(\alpha^1; x)$, $\pi^2(\alpha^2; x)$, $a^1(\gamma^1; x)$ or $a^2(\gamma^2; x)$ is used, respectively. For example, $\hat{\mu}_{\mathrm{aipw},1001}$ denotes the augmented inverse probability weighted estimator based on $\pi^1(\alpha^1; x)$ and $a^2(\gamma^2; x)$, and $\hat{\mu}_{1101}$ denotes our proposed estimator based on $\pi^1(\alpha^1; x)$, $\pi^2(\alpha^2; x)$ and $a^2(\gamma^2; x)$, for which the $\hat{w}_i$ $(i = 1, \ldots, m)$ are calculated as

$$\hat{w}_i = \frac{[1 + \hat{\rho}_1\{\pi^1(\hat{\alpha}^1, x_i) - \hat{\theta}^1\} + \hat{\rho}_2\{\pi^2(\hat{\alpha}^2, x_i) - \hat{\theta}^2\} + \hat{\rho}_3\{a^2(\hat{\gamma}^2, x_i) - \hat{\eta}^2\}]^{-1}}{\sum_{i=1}^m [1 + \hat{\rho}_1\{\pi^1(\hat{\alpha}^1, x_i) - \hat{\theta}^1\} + \hat{\rho}_2\{\pi^2(\hat{\alpha}^2, x_i) - \hat{\theta}^2\} + \hat{\rho}_3\{a^2(\hat{\gamma}^2, x_i) - \hat{\eta}^2\}]^{-1}},$$

where $\hat{\rho} = (\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3)$ solves the equation

$$\sum_{i=1}^m \frac{\{\pi^1(\hat{\alpha}^1, x_i) - \hat{\theta}^1, \pi^2(\hat{\alpha}^2, x_i) - \hat{\theta}^2, a^2(\hat{\gamma}^2, x_i) - \hat{\eta}^2\}^{\mathrm{T}}}{1 + \rho_1\{\pi^1(\hat{\alpha}^1, x_i) - \hat{\theta}^1\} + \rho_2\{\pi^2(\hat{\alpha}^2, x_i) - \hat{\theta}^2\} + \rho_3\{a^2(\hat{\gamma}^2, x_i) - \hat{\eta}^2\}} = 0.$$

We conducted 5000 replications for the simulation study; the results are summarized in Table 1. In the following, we refer to the four data-generating models in Table 1, from left to right, as the first, second, third and fourth model.

Table 1. *Simulation results based on* 5000 *Monte Carlo replications; values have been multiplied by* 100. *The two propensity score models lead to approximately* 55% *and* 44% *of subjects with missing responses, respectively*

	$\pi(x) = \{1 + \exp(0{\cdot}8 + 0{\cdot}5x - 0{\cdot}3x^2)\}^{-1}$						$\pi(x) = 1 - \exp[-\exp\{0{\cdot}5 + 0{\cdot}5x - 0{\cdot}3\exp(x)\}]$					
	$a(x) = 1 + 2x + 3x^2$			$a(x) = 1 + 2x + 3\exp(x)$			$a(x) = 1 + 2x + 3x^2$			$a(x) = 1 + 2x + 3\exp(x)$		
	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
						$n = 300$						
$\hat{\mu}_{\text{aipw},1010}$	0	46	31	0	74	50	1	47	31	−22	77	52
$\hat{\mu}_{\text{aipw},1001}$	0	49	32	0	73	49	28	59	40	1	74	50
$\hat{\mu}_{\text{aipw},0110}$	0	46	31	−11	74	51	1	48	32	0	75	51
$\hat{\mu}_{\text{aipw},0101}$	16	52	34	0	73	49	3	51	35	1	74	50
$\hat{\mu}_{1110}$	0	46	30	−5	77	50	1	49	32	−4	76	50
$\hat{\mu}_{1101}$	1	47	31	0	74	49	4	50	33	1	77	51
$\hat{\mu}_{1011}$	0	47	31	0	74	49	0	62	34	−1	105	54
$\hat{\mu}_{0111}$	0	50	31	0	79	49	1	56	33	0	87	52
$\hat{\mu}_{1111}$	0	47	31	0	73	49	1	50	33	−1	81	51
						$n = 1000$						
$\hat{\mu}_{\text{aipw},1010}$	0	25	17	0	40	28	0	25	17	−22	45	31
$\hat{\mu}_{\text{aipw},1001}$	0	26	17	0	40	27	26	38	27	0	40	27
$\hat{\mu}_{\text{aipw},0110}$	0	25	17	−12	42	28	0	25	17	0	41	27
$\hat{\mu}_{\text{aipw},0101}$	15	31	21	0	40	27	0	27	18	0	40	27
$\hat{\mu}_{1110}$	0	25	17	−3	43	27	−1	30	17	−2	42	28
$\hat{\mu}_{1101}$	0	25	17	0	41	27	1	27	18	0	42	27
$\hat{\mu}_{1011}$	0	25	17	0	40	27	0	36	18	−1	68	30
$\hat{\mu}_{0111}$	0	29	17	0	41	27	0	27	17	0	46	28
$\hat{\mu}_{1111}$	0	25	17	0	40	27	0	26	17	0	43	28

RMSE, root mean square error; MAE, median absolute error.

The multiple robustness of $\hat{\mu}$ implies that the estimators $\hat{\mu}_{1110}$, $\hat{\mu}_{1101}$, $\hat{\mu}_{1011}$, $\hat{\mu}_{0111}$ and $\hat{\mu}_{1111}$ are consistent under all four data-generating models. This is well demonstrated by inspecting their ignorable bias under all models. Neither $\hat{\mu}_{\text{aipw}}$ nor any existing doubly robust estimator can achieve such robustness. To assess the efficiency of our proposed estimators, we use the root mean square error of $\hat{\mu}_{\text{aipw,opt}}$ as the benchmark, where $\hat{\mu}_{\text{aipw,opt}}$ denotes the corresponding augmented inverse probability weighted estimator that attains the semiparametric efficiency bound under each specific data-generating model; that is, $\hat{\mu}_{\text{aipw},1010}$ under the first model, $\hat{\mu}_{\text{aipw},1001}$ under the second model, $\hat{\mu}_{\text{aipw},0110}$ under the third model, and $\hat{\mu}_{\text{aipw},0101}$ under the fourth model. The estimators $\hat{\mu}_{1110}$ and $\hat{\mu}_{1101}$ both have excellent efficiency under all four models. Estimator $\hat{\mu}_{1011}$ has root mean square error almost identical to that of $\hat{\mu}_{\text{aipw,opt}}$ under the first two models, which is in full agreement with our asymptotic theory that $\hat{\mu}_{1011}$ attains the semiparametric efficiency bound in these two cases. Under the last two models, $\hat{\mu}_{1011}$ has relatively low efficiency. Estimator $\hat{\mu}_{0111}$ can be seen to have high efficiency by inspecting its root mean square error when $n = 1000$, although its numerical performance under the last model is not satisfactory when $n = 300$. According to our theory, $\hat{\mu}_{1111}$ attains the semiparametric efficiency bound under all four data-generating models; this is confirmed by comparing its root mean square error to that of $\hat{\mu}_{\text{aipw,opt}}$.

Acknowledgement

The authors wish to thank the editor, the associate editor and two reviewers for their helpful comments.

APPENDIX

Proof of (5). The case of $\pi^j(\alpha^j; X)$ with arbitrary j follows from the fact that

$$\begin{aligned}
0 &= E\left(\frac{R}{\pi(X)}[\pi^j(\alpha^j; X) - E\{\pi^j(\alpha^j; X)\}]\right) \\
&= \mathrm{pr}(R=1)\, E\left(\frac{R}{\pi(X)}[\pi^j(\alpha^j; X) - E\{\pi^j(\alpha^j; X)\}] \,\middle|\, R=1\right) \\
&\quad + \mathrm{pr}(R=0)\, E\left(\frac{R}{\pi(X)}[\pi^j(\alpha^j; X) - E\{\pi^j(\alpha^j; X)\}] \,\middle|\, R=0\right) \\
&= \mathrm{pr}(R=1)\, E\left[\frac{\pi^j(\alpha^j; X) - E\{\pi^j(\alpha^j; X)\}}{\pi(X)} \,\middle|\, R=1\right].
\end{aligned}$$

The case of $a^k(\gamma^k; X)$ with arbitrary k follows from a similar argument. □

In the following, to simplify notation, we write $\pi_i^j(\alpha^j) = \pi^j(\alpha^j; X_i)$, $a_i^k(\gamma^k) = a^j(\gamma^k; X_i)$ and $g(X) = g(\alpha_*, \gamma_*; X)$, which is given by (9).

LEMMA A1. *When $\pi^1(\alpha^1; x)$ is a correctly specified model for $\pi(x)$, we have*

$$n^{1/2}\hat{\lambda} = G^{-1}\left\{n^{-1/2}\sum_{i=1}^n \frac{R_i - \pi(X_i)}{\pi(X_i)} g(X_i) - An^{1/2}(\hat{\alpha}^1 - \alpha_0^1)\right\} + o_p(1),$$

where $A = E[\{g(X)/\pi(X)\}\{\partial\pi^1(\alpha_0^1; X)/\partial\alpha^{1}\}^{\mathrm{T}}]$.

Proof of Lemma A1. Let d_j and u_k denote the dimensions of α^j and γ^k, respectively. Taylor expansion of the left-hand side of (7) about $(0^{\mathrm{T}}, \alpha_*^{\mathrm{T}}, \gamma_*^{\mathrm{T}})$ leads to

$$\begin{aligned}
0 &= n^{-1/2}\sum_{i=1}^n \frac{R_i}{\pi_i^1(\alpha_0^1)}\hat{g}_i(\alpha_*, \gamma_*) - \left\{\frac{1}{n}\sum_{i=1}^n \frac{R_i}{\pi_i^1(\alpha_0^1)}\frac{\hat{g}_i(\alpha_*, \gamma_*)^{\otimes 2}}{\pi_i^1(\alpha_0^1)}\right\} n^{1/2}\hat{\lambda} \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^n \frac{R_i}{\{\pi_i^1(\alpha_0^1)\}^2}\left[\left\{\begin{array}{c}\dfrac{\partial\pi_i^1(\alpha_0^1)}{\partial\alpha^{1,\mathrm{T}}} - \dfrac{1}{n}\displaystyle\sum_{h=1}^n \dfrac{\partial\pi_h^1(\alpha_0^1)}{\partial\alpha^{1,\mathrm{T}}} \\ 0_{(J+K-1)\times d_1}\end{array}\right\}\pi_i^1(\alpha_0^1) - \hat{g}_i(\alpha_*, \gamma_*)\frac{\partial\pi_i^1(\alpha_0^1)}{\partial\alpha^{1,\mathrm{T}}}\right]\right) \\
&\quad \times n^{1/2}(\hat{\alpha}^1 - \alpha_0^1) \\
&\quad + \sum_{j=2}^J \left[\frac{1}{n}\sum_{i=1}^n \frac{R_i}{\pi_i^1(\alpha_0^1)}\left\{\begin{array}{c}0_{(j-1)\times d_j} \\ \dfrac{\partial\pi_i^j(\alpha_*^j)}{\partial\alpha^{j,\mathrm{T}}} - \dfrac{1}{n}\displaystyle\sum_{h=1}^n \dfrac{\partial\pi_h^j(\alpha_*^j)}{\partial\alpha^{j,\mathrm{T}}} \\ 0_{(J+K-j)\times d_j}\end{array}\right\}\right] n^{1/2}(\hat{\alpha}^j - \alpha_*^j) \\
&\quad + \sum_{k=1}^K \left[\frac{1}{n}\sum_{i=1}^n \frac{R_i}{\pi_i^1(\alpha_0^1)}\left\{\begin{array}{c}0_{(J+k-1)\times u_k} \\ \dfrac{\partial a_i^k(\gamma_*^k)}{\partial\gamma^{k,\mathrm{T}}} - \dfrac{1}{n}\displaystyle\sum_{h=1}^n \dfrac{\partial a_h^k(\gamma_*^k)}{\partial\gamma^{k,\mathrm{T}}} \\ 0_{(K-k)\times u_k}\end{array}\right\}\right] n^{1/2}(\hat{\gamma}^k - \gamma_*^k) + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^n \frac{R_i - \pi(X_i)}{\pi(X_i)} g(X_i) - Gn^{1/2}\hat{\lambda} - An^{1/2}(\hat{\alpha}^1 - \alpha_0^1) + o_p(1).
\end{aligned}$$

Solving for $\hat{\lambda}$ gives the result. □

Proof of Theorem 3. Taylor expansion of $n^{1/2}(\hat{\mu}-\mu_0)$ about $(0^{\mathrm{T}}, \alpha_*^{\mathrm{T}}, \gamma_*^{\mathrm{T}})$ leads to

$$
\begin{aligned}
n^{1/2}(\hat{\mu}-\mu_0) &= n^{1/2}\sum_{i=1}^{n} R_i\hat{w}_i(Y_i-\mu_0)\\
&= n^{1/2}\frac{1}{m}\sum_{i=1}^{n}\frac{R_i(Y_i-\mu_0)\hat{\theta}^1/\pi_i^1(\hat{\alpha}^1)}{1+\hat{\lambda}^{\mathrm{T}}\hat{g}_i(\hat{\alpha},\hat{\gamma})/\pi_i^1(\hat{\alpha}^1)}\\
&= n^{1/2}\frac{1}{m}\left\{\frac{1}{n}\sum_{h=1}^{n}\pi_h^1(\alpha_0^1)\right\}\sum_{i=1}^{n}\frac{R_i(Y_i-\mu_0)}{\pi_i^1(\alpha_0^1)}\\
&\quad -\frac{1}{m}\left\{\frac{1}{n}\sum_{h=1}^{n}\pi_h^1(\alpha_0^1)\right\}\left\{\sum_{i=1}^{n}\frac{R_i(Y_i-\mu_0)}{\pi_i^1(\alpha_0^1)}\frac{\hat{g}_i^{\mathrm{T}}(\alpha_*,\gamma_*)}{\pi_i^1(\alpha_0^1)}\right\}n^{1/2}\hat{\lambda}\\
&\quad +\frac{1}{m}\sum_{i=1}^{n}\frac{R_i(Y_i-\mu_0)}{\{\pi_i^1(\alpha_0^1)\}^2}\left[\left\{\frac{1}{n}\sum_{h=1}^{n}\frac{\partial\pi_h^1(\alpha_0^1)}{\partial\alpha^{1,\mathrm{T}}}\right\}\pi_i^1(\alpha_0^1)-\left\{\frac{1}{n}\sum_{h=1}^{n}\pi_h^1(\alpha_0^1)\right\}\frac{\partial\pi_i^1(\alpha_0^1)}{\partial\alpha^{1,\mathrm{T}}}\right]\\
&\qquad \times n^{1/2}(\hat{\alpha}^1-\alpha_0^1)+o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}\left\{\frac{R_i(Y_i-\mu_0)}{\pi(X_i)}-\frac{R_i-\pi(X_i)}{\pi(X_i)}L^{\mathrm{T}}G^{-1}g(X_i)\right\}\\
&\quad -E\left\{\frac{Y-\mu_0-L^{\mathrm{T}}G^{-1}g(X)}{\pi(X)}\frac{\partial\pi^1(\alpha_0^1;X)}{\partial\alpha^{1,\mathrm{T}}}\right\}n^{1/2}(\hat{\alpha}^1-\alpha_0^1)+o_p(1).
\end{aligned}
$$

It is easy to verify that

$$
E\left\{\frac{Y-\mu_0-L^{\mathrm{T}}G^{-1}g(X)}{\pi(X)}\frac{\partial\pi^1(\alpha_0^1;X)}{\partial\alpha^1}\right\}=-E\left[\partial\left\{\frac{R(Y-\mu_0)}{\pi^1(\alpha_0^1;X)}-\frac{R-\pi^1(\alpha_0^1;X)}{\pi^1(\alpha_0^1;X)}L^{\mathrm{T}}G^{-1}g(X)\right\}\Big/\partial\alpha^1\right].
$$

Therefore, the desired result follows from the generalized information equality (e.g. Tsiatis, 2006, Lemma 9.1) and the asymptotic expansion $n^{1/2}(\hat{\alpha}^1-\alpha_0^1)=n^{-1/2}\sum_{i=1}^{n}\{E(S^{\otimes 2})\}^{-1}S(\alpha_0^1;X_i,R_i)$. □

References

Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–72.

Cao, W., Tsiatis, A. A. & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–34.

Chen, S. X., Leung, D. H. Y. & Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *J. R. Statist. Soc.* B **70**, 803–23.

Deville, J. & Särndal, C. (1992). Calibration estimators in survey sampling. *J. Am. Statist. Assoc.* **87**, 376–82.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Han, P. (2012). A note on improving the efficiency of inverse probability weighted estimator using the augmentation term. *Statist. Prob. Lett.* **82**, 2221–8.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663–85.

Kang, J. D. Y. & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with Discussion). *Statist. Sci.* **22**, 523–39.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley, 2nd ed.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–49.

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

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355–65.

Pepe, M. S., Reilly, M. & Fleming, T. R. (1994). Auxiliary outcome data and the mean score method. *J. Statist. Plan. Infer.* **42**, 137–60.

Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–25.

QIN, J., SHAO, J. & ZHANG, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *J. Am. Statist. Assoc.* **103**, 797–810.
QIN, J. & ZHANG, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Statist. Soc.* B **69**, 101–22.
ROBINS, J. M. & ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Assoc.* **90**, 122–9.
ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.
ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.* **90**, 106–21.
ROBINS, J. M., SUED, M., GOMEZ-LEI, Q. & ROTNITZKY, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statist. Sci.* **22**, 544–59.
ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
ROTNITZKY, A., LEI, Q., SUED, M. & ROBINS, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**, 439–56.
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.
RUBIN, D. B. & VAN DER LAAN, M. J. (2008). Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Int. J. Biostat.* **4**, Article 5.
SCHARFSTEIN, D. O., ROTNITZKY, A. & ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Statist. Assoc.* **94**, 1096–120.
TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Am. Statist. Assoc.* **101**, 1619–37.
TAN, Z. (2007). Comment: Understanding OR, PS and DR. *Statist. Sci.* **22**, 560–8.
TAN, Z. (2008). Comment: Improved local efficiency and double robustness. *Int. J. Biostat.* **4**, Article 10.
TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97**, 661–82.
TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
TSIATIS, A. A. & DAVIDIAN, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22**, 569–73.
WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
WU, C. & SITTER, R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Statist. Assoc.* **96**, 185–93.

[*Received August* 2012. *Revised December* 2012]