Statistica Sinica Preprint No: SS-2016-0185.R1	
Title	A Further Study of Propensity Score Calibration in Missing Data Analysis
Manuscript ID	SS-2016-0185.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0185
Complete List of Authors	Peisong Han
Corresponding Author	Peisong Han
E-mail	peisonghan@uwaterloo.ca
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

A Further Study of Propensity Score Calibration in Missing Data Analysis

Peisong Han
Department of Statistics and Actuarial Science, University of Waterloo
Waterloo, ON, Canada N2L 3G1
Email: peisonghan@uwaterloo.ca

Abstract

Methods for propensity score (PS) calibration are commonly used in missing data analysis. Most of them are derived based on constrained optimizations where the form of calibration is dictated by the objective function being optimized and the calibration variables used in the constraints. Considerable efforts on pairing an appropriate objective function with the calibration constraints are usually needed to achieve certain efficiency and robustness properties for the final estimators. We consider an alternative approach where the calibration is carried out by solving the empirical version of certain moment equalities. This approach frees us from constructing a particular objective function. Based on this approach, under the setting of estimating the mean of a response, we establish intrinsic, improved and local efficiency and multiple robustness in the presence of multiple data distribution models. A revisit to the generalized pseudo exponential tilting estimator and generalized pseudo empirical likelihood estimator of Tan and Wu (2015) is also provided.

Key Words: Calibration; Efficiency; Empirical likelihood; Missing at random (MAR); Multiple robustness; Propensity score

1 Introduction

For missing-at-random (MAR) (Rubin 1976) data, semiparametric approach through inverse probability weighting (IPW) (Horvitz and Thompson 1952), which weights the observed values by the inverse of the propensity score (Rosenbaum and Rubin 1983), has been widely used. Taking this approach, Robins, Rotnitzky and Zhao (1994, 1995) proposed a large class of estimators, called the augmented IPW (AIPW) estimators, by introducing augmentation terms that were constructed based on models for the data distribution. A particular estimator in this class is locally efficient, in that it attains the semiparametric efficiency bound if both the propensity score and the data distribution are correctly modeled. Scharfstein, Rotnitzky and Robins (1999) noted that consistency of this estimator only requires correctly modeling either the propensity score or the data distribution, but not both. This property is known as double robustness.

Recently, a variety of new estimators based on the IPW approach have been proposed with other nice properties. Estimators in Tan (2006; 2008; 2010), Chen, Leung and Qin (2008), Chan (2012) and Rotnitzky et al. (2012) have intrinsic efficiency: with a correctly specified propensity score model and a fixed user-specified function of observed data, each of these estimators is asymptotically equivalent to the most efficient AIPW estimator among a class of AIPW estimators whose augmentation terms are generated from this fixed function. Estimators in Rubin and van der Laan (2008), Tan (2008; 2010), Cao, Tstiatis and Davidian (2009) and Rotnitzky et al. (2012) have improved efficiency: with a correctly specified propensity score model, each of these estimators is asymptotically equivalent to the most efficient AIPW estimator among a class of AIPW estimators for which the data distribution parameters in the augmentation terms are fixed but arbitrary. Estimators in Han and Wang (2013), Chan and Yam (2014), Han (2014a, 2014b, 2016) and Chen and Haziza (2017) are multiply robust: with multiple models for the propensity score and/or the data distribution, consistency is guaranteed if any one of these models is correctly specified. Estimators in Tan (2006; 2010), Qin and Zhang (2007), Kim (2009; 2010), Chan (2012), Han and Wang (2013), Chan and Yam (2014) and Tan and Wu (2015) are convex combinations of the observed outcomes, and thus always fall within the range of observed values, known as sample boundedness property (Robins et al. 2007).

While the propensity score is crucial for all of the above methods, it is not always incorporated in the same fashion. The IPW and AIPW estimators, as well as many other recently proposed ones, use the inverse of the raw propensity score as the weight. Nice properties of these estimators mainly rely on delicate construction of augmentation terms and/or careful estimation of data distribution parameters. In recent literature, many researchers proposed to use weight derived by modifying the raw propensity score (Tan 2006, 2010; Qin and Zhang 2007; Chen, Leung and Qin 2008; Qin, Shao and Zhang 2008; Kim 2009, 2010; Chan 2012; Han and Wang 2013; Chan and Yam 2014; Han 2014a, 2014b; Tan and Wu 2015; Han 2016). Conceptually, these modifications in essence agree with the idea of calibration in survey sampling literature (Deville and Särndal 1992). The new weight is usually derived by optimizing an objective function subject to certain calibration constraints. Mathematically, such a constrained optimization amounts to fitting a hybrid propensity score model with a separate component that calibrates the raw propensity score. However, in general, considerable efforts on pairing an appropriate objective function with the calibration constraints are needed to achieve certain efficiency and robustness properties for the final estimators (e.g. Kim 2009, 2010; Tan 2010; Tan and Wu 2015).

In this paper, we consider an alternative approach to calibrate the raw propensity score. The calibration is done by solving the empirical version of certain moment equalities rather than by constrained optimization, although the numerical implementation benefits from treating those empirical equations as the first-order conditions of certain objective functions. The spirit is the same as solving estimating equations instead of maximizing a parametric likelihood function for estimation. It frees us from the non-trivial work of constructing an appropriate objective function for optimization in order that the resulting calibrated propensity score leads to desirable properties.

Using the calibrated propensity score based on this alternative approach, we establish intrinsic, improved and local efficiency and multiple robustness for our proposed estimators when multiple models for the data distribution are available. Intrinsic efficiency guarantees that, with a correctly specified propensity score model, the multiple data distribution models are optimally accommodated to maximize efficiency. The efficiency usually increases as the number of models does, except for the case where one data distribution model is correctly specified as well, in which all of our proposed estimators attain the semiparametric efficiency

bound, and thus are locally efficient. Improved efficiency ensures that the parameters in all data distribution models are simultaneously optimally estimated so that the efficiency of our proposed estimators is maximized compared to the same estimators with the data distribution parameters fixed but arbitrary. In addition to efficiency advantages, our proposed estimators are still consistent if the propensity score model is misspecified but one data distribution model is correct. Furthermore, all of our proposed estimators are convex combinations of the observed outcomes, and thus are sample bounded.

To make the paper focused, the proposed approach is only demonstrated in estimating the mean of an outcome. It can certainly be applied to causal inference problems and other more complex missing data problems such as regression analysis. This paper is organized as follows. Section 2 introduces necessary notation and discusses some existing methods. Section 3 investigates the alternative approach to propensity score calibration. Section 4 gives a revisit to the generalized pseudo exponential tilting estimator and generalized pseudo empirical likelihood estimator of Tan and Wu (2015). A numerical study is provided in Section 5. Some discussion is given in Section 6. The Appendix provides some technical details.

2 Notation and Some Existing Methods

Let Y denote an outcome of interest that is subject to missingness, $\boldsymbol{X}$ a vector of covariates that are always observed, and R the indicator of observing Y (i.e., $R = 1$ if Y is observed and $R = 0$ if Y is missing). The observed data are $(R_i, R_iY_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, which are independent and identically distributed. The MAR mechanism in this setting is $P(R = 1|Y, \boldsymbol{X}) = P(R = 1|\boldsymbol{X})$. We use $\pi(\boldsymbol{X})$ to denote this propensity score. Our interest is to estimate $\mu_0 = E(Y)$, the marginal mean of Y.

The IPW method (Horvitz and Thompson 1952) models $\pi(\boldsymbol{X})$ by $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$, where $\boldsymbol{\alpha}$ is a finite-dimensional unknown parameter and may be estimated by $\hat{\boldsymbol{\alpha}}$ that maximizes the Binomial likelihood

$$\prod_{i=1}^{n} \{\pi(\boldsymbol{\alpha}; \boldsymbol{X}_i)\}^{R_i} \{1 - \pi(\boldsymbol{\alpha}; \boldsymbol{X}_i)\}^{1-R_i}. \tag{1}$$

The IPW estimator of μ_0 is $\hat{\mu}_{\text{ipw}} = n^{-1} \sum_{i=1}^{n} R_iY_i/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)$, where the observed Y_i is weighted by the inverse of the raw propensity score $\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)$. $\hat{\mu}_{\text{ipw}}$ is consistent if $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ is a correctly specified model in the sense that $\pi(\boldsymbol{\alpha}_0; \boldsymbol{X}) = \pi(\boldsymbol{X})$ for some $\boldsymbol{\alpha}_0$.

To facilitate the discussion, for now we assume that $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ is a correctly specified model. Since the IPW method does not extract information implied by the dependence of Y on $\boldsymbol{X}$, $\hat{\mu}_{\text{ipw}}$ does not make efficient use of the observed data. Robins, Rotnitzky and Zhao (1994) proposed a class of AIPW estimators

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)}Y_i - \frac{R_i - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)}h(\boldsymbol{X}_i)\right\},$$

where $h(\boldsymbol{X})$ is an arbitrary user-specified function of $\boldsymbol{X}$ and may be constructed based on a model for the data distribution. Within this class, estimators of the form $\hat{\mu}_{\text{aipw}}(\boldsymbol{\gamma})$ with $h(\boldsymbol{X}) = a(\boldsymbol{\gamma}; \boldsymbol{X})$ are of particular interest. Here $a(\boldsymbol{\gamma}; \boldsymbol{X})$ is a model for $E(Y \mid \boldsymbol{X})$ and $\boldsymbol{\gamma}$ is a finite-dimensional unknown parameter. Because $R \perp Y \mid \boldsymbol{X}$ from the MAR mechanism, $\boldsymbol{\gamma}$ is conventionally estimated by $\hat{\boldsymbol{\gamma}}$ based on complete-case analysis. When $a(\boldsymbol{\gamma}; \boldsymbol{X})$ is a reasonable model for $E(Y \mid \boldsymbol{X})$, $\hat{\mu}_{\text{aipw}}(\hat{\boldsymbol{\gamma}})$ usually has better efficiency than $\hat{\mu}_{\text{ipw}}$. When $a(\boldsymbol{\gamma}; \boldsymbol{X})$ is correctly specified in that $a(\boldsymbol{\gamma}_0; \boldsymbol{X}) = E(Y \mid \boldsymbol{X})$ for some $\boldsymbol{\gamma}_0$, $\hat{\mu}_{\text{aipw}}(\hat{\boldsymbol{\gamma}})$ attains the semiparametric efficiency bound.

When $a(\boldsymbol{\gamma}; \boldsymbol{X})$ is incorrectly specified, $\hat{\mu}_{\text{aipw}}(\hat{\boldsymbol{\gamma}})$ can be quite inefficient (Chen, Leung and Qin 2008; Rubin and van der Laan 2008). There have been many recent developments on gaining efficiency in this case. Two main gains have been achieved: intrinsic efficiency and improved efficiency. Estimators that are intrinsically efficient have influence functions of the form

$$\text{Resid}\left\{\frac{R(Y-\mu_0)}{\pi(\boldsymbol{X})}, \frac{R-\pi(\boldsymbol{X})}{\pi(\boldsymbol{X})}h(\boldsymbol{X})\right\}.$$

Hereafter, for any random variable ξ and finite-dimensional random vector $\boldsymbol{\phi}$ with mean zero and finite second moments, $\text{Resid}(\xi, \boldsymbol{\phi}) = \xi - E(\xi\boldsymbol{\phi}^{\mathrm{T}})\{E(\boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}})\}^{-1}\boldsymbol{\phi}$ denotes the residual of the projection of ξ onto $\text{span}\{\boldsymbol{\phi}\}$, the linear space spanned by components of $\boldsymbol{\phi}$. Apparently the projection residual has the smallest variance among the class of influence functions $R(Y-\mu_0)/\pi(\boldsymbol{X}) - c\{R-\pi(\boldsymbol{X})\}h(\boldsymbol{X})/\pi(\boldsymbol{X})$ with an arbitrary c. Various intrinsically efficient estimators have been proposed and studied by Tan (2006; 2008; 2010), Chen, Leung and Qin (2008), Chan (2012) and Rotnitzky et al. (2012). Improved efficiency, on the other hand, is achieved by using an estimator $\tilde{\boldsymbol{\gamma}}$ instead of $\hat{\boldsymbol{\gamma}}$, where $\tilde{\boldsymbol{\gamma}}$ converges in probability to the minimizer of the asymptotic variance of $\hat{\mu}_{\text{aipw}}(\boldsymbol{\gamma})$. Estimators of μ_0 with improved efficiency have been proposed and studied by Rubin and van der Laan (2008), Tan (2008; 2010), Cao,

Tstiatis and Davidian (2009) and Rotnitzky et al. (2012). Many of the above estimators are doubly robust: they are still consistent if $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ is misspecified but $a(\boldsymbol{\gamma}; \boldsymbol{X})$ is not.

Recently, many estimators of the form $\sum_{i=1}^{n} R_i w_i Y_i$ have been proposed where the w_i are derived by optimizing an objective function, such as the empirical likelihood (Tan 2006, 2010; Qin and Zhang 2007; Chen, Leung and Qin 2008; Kim 2009; Han and Wang 2013; Chan and Yam 2014), the exponential tilting (Kim 2010) or some generalizations of them (Tan and Wu 2015), subject to certain constraints on w_i. Different objective functions and/or sets of constraints lead to different w_i, for which the two most common forms are $[\pi(\boldsymbol{\alpha}; \boldsymbol{X}_i) \exp\{\boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{b}(\boldsymbol{X}_i)\}]^{-1}$ or $\{\pi(\boldsymbol{\alpha}; \boldsymbol{X}_i) + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{b}(\boldsymbol{X}_i)\}^{-1}$, where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers and $\boldsymbol{\lambda}$ and $\boldsymbol{b}(\boldsymbol{X})$ are determined by the specific optimization objective function and the particular constraints. In general, $\boldsymbol{\lambda}$ and $\boldsymbol{b}(\boldsymbol{X})$ do not necessarily endow the final estimators of μ_0 with desirable efficiency and robustness properties unless an appropriate objective function is paired with the right constraints (e.g. Tan 2010; Tan and Wu 2015). Because of this difficulty, we propose to circumvent the optimization and directly derive the calibration on $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ by solving certain empirical equations. In this way, we are free to independently choose the calibration constraints and the functional form of $\boldsymbol{b}(\boldsymbol{X})$, so that in combination they lead to certain desirable properties. In particular, we use this approach to construct estimators that have intrinsic, improved and local efficiency and multiple robustness.

3 The Proposed Approach

We first introduce some extra notation. Let $a^k(\boldsymbol{\gamma}^k; \boldsymbol{X})$, $k = 1, \ldots, K$, denote K models for $E(Y \mid \boldsymbol{X})$, $\hat{\boldsymbol{\gamma}}^k$ the estimator of $\boldsymbol{\gamma}^k$ by fitting the k-th model based on complete-case analysis, and $\boldsymbol{\gamma}^k_*$ the probability limit of $\hat{\boldsymbol{\gamma}}^k$. Write $\boldsymbol{\gamma}^{\mathrm{T}} = \{(\boldsymbol{\gamma}^1)^{\mathrm{T}}, \ldots, (\boldsymbol{\gamma}^K)^{\mathrm{T}}\}$, $\hat{\boldsymbol{\gamma}}^{\mathrm{T}} = \{(\hat{\boldsymbol{\gamma}}^1)^{\mathrm{T}}, \ldots, (\hat{\boldsymbol{\gamma}}^K)^{\mathrm{T}}\}$, and $\boldsymbol{\gamma}_*^{\mathrm{T}} = \{(\boldsymbol{\gamma}_*^1)^{\mathrm{T}}, \ldots, (\boldsymbol{\gamma}_*^K)^{\mathrm{T}}\}$. Let $\boldsymbol{S}(\boldsymbol{\alpha}; \boldsymbol{X}, R)$ denote the score function of (1); that is

$$\boldsymbol{S}(\boldsymbol{\alpha}; \boldsymbol{X}, R) = \frac{R - \pi(\boldsymbol{\alpha}; \boldsymbol{X})}{\pi(\boldsymbol{\alpha}; \boldsymbol{X})\{1 - \pi(\boldsymbol{\alpha}; \boldsymbol{X})\}} \pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}; \boldsymbol{X}),$$

where $\pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}; \boldsymbol{X}) = \partial \pi(\boldsymbol{\alpha}; \boldsymbol{X}) / \partial \boldsymbol{\alpha}$. Let $\boldsymbol{\alpha}_*$ denote the probability limit of $\hat{\boldsymbol{\alpha}}$. When $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ is a correctly specified model for $\pi(\boldsymbol{X})$, $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ and write $\boldsymbol{S}(\boldsymbol{X}, R) = \boldsymbol{S}(\boldsymbol{\alpha}_0; \boldsymbol{X}, R)$ and $\pi_{\boldsymbol{\alpha}}(\boldsymbol{X}) = \pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0; \boldsymbol{X})$. For any matrix $\boldsymbol{A}$, let $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$. Hereafter, all linear spaces under our discussion are subspaces of the Hilbert space $\mathcal{H}$ of all mean-zero and finite-variance

functions of $(R, RY, \boldsymbol{X})$ equipped with the inner product $E(\xi_1\xi_2)$ for any $\xi_1, \xi_2 \in \mathcal{H}$. For any $\boldsymbol{\phi}$ whose components are all in $\mathcal{H}$, let $\text{span}\{\boldsymbol{\phi}\}$ denote the linear space spanned by $\boldsymbol{\phi}$.

3.1 The Proposed Estimators and Their Properties

It is easy to see that the moment equalities

$$E\left(\frac{R[\pi(\boldsymbol{\alpha}_*; \boldsymbol{X}) - E\{\pi(\boldsymbol{\alpha}_*; \boldsymbol{X})\}]}{\pi(\boldsymbol{X})}\right) = 0,$$
$$E\left(\frac{R[a^k(\boldsymbol{\gamma}_*^k; \boldsymbol{X}) - E\{a^k(\boldsymbol{\gamma}_*^k; \boldsymbol{X})\}]}{\pi(\boldsymbol{X})}\right) = 0, \qquad (k = 1, \ldots, K) \tag{2}$$

hold. However, the trivial empirical version of (2), namely

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R_i\{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) - \hat{\theta}\}}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} \frac{R_i\{a^k(\hat{\boldsymbol{\gamma}}^k; \boldsymbol{X}_i) - \hat{\eta}^k\}}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} = 0,$$

where $\hat{\theta} = n^{-1}\sum_{i=1}^{n} \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)$ and $\hat{\eta}^k = n^{-1}\sum_{i=1}^{n} a^k(\hat{\boldsymbol{\gamma}}^k; \boldsymbol{X}_i)$, usually does not hold with the observed data, even if $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ is a correctly specified model. Our proposed approach is based on constructing an empirical version of (2) that does hold. Following the form of exponential tilting weight (e.g. Kim 2010; Tan and Wu 2015), consider a calibration of the raw propensity score in the form of $\pi(\boldsymbol{\alpha}; \boldsymbol{X})\exp\{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b}(\boldsymbol{X})\}$, where $\boldsymbol{b}(\boldsymbol{X})$ is a vector of user-specified functions and $\boldsymbol{\lambda}$ is a calibration parameter depending on $\boldsymbol{b}(\boldsymbol{X})$, so that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R_i\{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) - \hat{\theta}\}}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\exp\{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b}(\boldsymbol{X}_i)\}} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} \frac{R_i\{a^k(\hat{\boldsymbol{\gamma}}^k; \boldsymbol{X}_i) - \hat{\eta}^k\}}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\exp\{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{b}(\boldsymbol{X}_i)\}} = 0.$$

It is clear that the calibration here is completely determined by $\boldsymbol{b}(\boldsymbol{X})$ with no optimizations needed. Although the above empirical equations may be solved for a rather arbitrary $\boldsymbol{b}(\boldsymbol{X})$, we consider a particular selection that leads to desirable properties for estimators of μ_0. Write $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}) - \hat{\theta}, a^1(\hat{\boldsymbol{\gamma}}^1; \boldsymbol{X}) - \hat{\eta}^1, \ldots, a^K(\hat{\boldsymbol{\gamma}}^K; \boldsymbol{X}) - \hat{\eta}^K\}^{\mathrm{T}}$. We will take $\boldsymbol{b}(\boldsymbol{X}) = \hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})$. The reason for taking this particular $\boldsymbol{b}(\boldsymbol{X})$ is that it leads to intrinsic efficiency of our proposed estimators.

Let $\hat{\boldsymbol{\lambda}}$ denote the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R_i\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\exp[\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)]} = \boldsymbol{0}. \tag{3}$$

The existence and uniqueness of $\hat{\boldsymbol{\lambda}}$ will be shown in Section 3.2. Our first proposed estimator of μ_0, denoted by $\hat{\mu}_1$, is the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i(Y_i-\mu)}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)]}=0. \tag{4}$$

For now, we assume that $\pi(\boldsymbol{\alpha};\boldsymbol{X})$ is a correctly specified model. This assumption will be relaxed later. To see the consistency of $\hat{\mu}_1$, let θ_*, η_*^k and $\boldsymbol{\lambda}_*$ denote the probability limits of $\hat{\theta}$, $\hat{\eta}^k$ and $\hat{\boldsymbol{\lambda}}$, respectively. It is clear that $\theta_*=E\{\pi(\boldsymbol{\alpha}_*;\boldsymbol{X})\}$ and $\eta_*^k=E\{a^k(\boldsymbol{\gamma}_*^k;\boldsymbol{X})\}$. Write $\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)=\{\pi(\boldsymbol{\alpha}_*;\boldsymbol{X})-\theta_*,a^1(\boldsymbol{\gamma}_*^1;\boldsymbol{X})-\eta_*^1,\ldots,a^K(\boldsymbol{\gamma}_*^K;\boldsymbol{X})-\eta_*^K\}^{\mathrm{T}}$. Since the left-hand side of (3) converges in probability to

$$E\left(\frac{R\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)}{\pi(\boldsymbol{X})\exp[\boldsymbol{\lambda}_*^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)\{1-\pi(\boldsymbol{X})\}/\pi(\boldsymbol{X})]}\right)$$

and $E\{R\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)/\pi(\boldsymbol{X})\}=\mathbf{0}$, we have $\boldsymbol{\lambda}_*=\mathbf{0}$, and thus $\hat{\boldsymbol{\lambda}}=o_p(1)$. Therefore,

$$\hat{\mu}_1=\frac{\frac{1}{n}\sum_{i=1}^{n}R_iY_i/[\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\{1+o_p(1)\}]}{\frac{1}{n}\sum_{i=1}^{n}R_i/[\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\{1+o_p(1)\}]}\xrightarrow{p}E\left\{\frac{R}{\pi(\boldsymbol{X})}Y\right\}=\mu_0.$$

To assess the efficiency of $\hat{\mu}_1$, we need to find its influence function, which is given by the following theorem. The proof is given in the Appendix.

Theorem 1. *When $\pi(\boldsymbol{X})$ is correctly modeled by $\pi(\boldsymbol{\alpha};\boldsymbol{X})$, we have*

$$\sqrt{n}(\hat{\mu}_1-\mu_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{Resid}\left\{\frac{R_i(Y_i-\mu_0)}{\pi(\boldsymbol{X}_i)},\frac{R_i-\pi(\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)}\boldsymbol{\Xi}_1(\boldsymbol{X}_i)\right\}, \tag{5}$$

where $\boldsymbol{\Xi}_1(\boldsymbol{X})=[\boldsymbol{g}(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)^{\mathrm{T}},\pi_{\boldsymbol{\alpha}}(\boldsymbol{X})^{\mathrm{T}}/\{1-\pi(\boldsymbol{X})\}]^{\mathrm{T}}$.

The projection structure of (5) clearly indicates that $\hat{\mu}_1$ is intrinsically efficient and is at least as efficient as, generally more efficient than, $\hat{\mu}_{\mathrm{ipw}}$ and any AIPW estimator whose augmentation term is a linear combination of components of $\{R-\pi(\boldsymbol{X})\}\boldsymbol{g}(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)/\pi(\boldsymbol{X})$. The projection structure also reveals the role the models $a^k(\boldsymbol{\gamma}^k;\boldsymbol{X})$ play in affecting efficiency. Since the larger K is, the larger $\mathrm{span}\{\{R-\pi(\boldsymbol{X})\}\boldsymbol{\Xi}_1(\boldsymbol{X})/\pi(\boldsymbol{X})\}$ is, and thus the smaller the projection residual is, assuming components of $\boldsymbol{\Xi}_1(\boldsymbol{X})$ are linearly independent, to gain efficiency, it is beneficial to postulate multiple models for $E(Y\mid\boldsymbol{X})$. Each model is guaranteed to improve the efficiency.

It is possible to gain efficiency without postulating more models for $E(Y \mid \boldsymbol{X})$. Consider augmenting $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ by adding the components

$$\frac{\pi_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})}{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})} - \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\pi_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)}{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} \right\},$$

and let $\hat{\mu}_1^{\text{aug}}$ denote the resulting estimator. From Theorem 1, the influence function of $\hat{\mu}_1^{\text{aug}}$ is given by (5) with $\boldsymbol{\Xi}_1(\boldsymbol{X})$ replaced by

$$\boldsymbol{\Xi}_1^{\text{aug}}(\boldsymbol{X}) = \left[\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)^{\mathrm{T}}, \frac{\pi_{\boldsymbol{\alpha}}(\boldsymbol{X})^{\mathrm{T}}}{1 - \pi(\boldsymbol{X})} - E\left\{ \frac{\pi_{\boldsymbol{\alpha}}(\boldsymbol{X})^{\mathrm{T}}}{1 - \pi(\boldsymbol{X})} \right\}, \frac{\pi_{\boldsymbol{\alpha}}(\boldsymbol{X})^{\mathrm{T}}}{1 - \pi(\boldsymbol{X})} \right]^{\mathrm{T}}.$$

It is easy to verify that $\text{span}\{\{R - \pi(\boldsymbol{X})\}\boldsymbol{\Xi}_1^{\text{aug}}(\boldsymbol{X})/\pi(\boldsymbol{X})\}$ is the same as $\text{span}\{\{R - \pi(\boldsymbol{X})\}\{1, \boldsymbol{\Xi}_1(\boldsymbol{X})^{\mathrm{T}}\}^{\mathrm{T}}/\pi(\boldsymbol{X})\}$. Therefore, $\hat{\mu}_1^{\text{aug}}$ is in general more efficient than $\hat{\mu}_1$. One exception occurs when $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ is a logistic regression model with intercept, in which case $\pi(\boldsymbol{X})$ is a component of $\pi_{\boldsymbol{\alpha}}(\boldsymbol{X})/\{1 - \pi(\boldsymbol{X})\}$, and thus $\text{span}\{\{R - \pi(\boldsymbol{X})\}\{1, \boldsymbol{\Xi}_1(\boldsymbol{X})^{\mathrm{T}}\}^{\mathrm{T}}/\pi(\boldsymbol{X})\}$ is equal to $\text{span}\{\{R - \pi(\boldsymbol{X})\}\boldsymbol{\Xi}_1(\boldsymbol{X})/\pi(\boldsymbol{X})\}$, which implies that $\hat{\mu}_1^{\text{aug}}$ and $\hat{\mu}_1$ are equally efficient.

Another way to gain efficiency is to increase the dimension of $\pi_{\boldsymbol{\alpha}}(\boldsymbol{X})$, or equivalently the dimension of $\boldsymbol{S}(\boldsymbol{X}, R)$. This may be achieved by including interactions and higher-order terms of components of $\boldsymbol{X}$ when fitting $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$. When $\pi(\boldsymbol{X})$ is completely known and is used in (3) and (4) instead of $\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})$, the projection in (5) becomes onto $\text{span}\{\{R - \pi(\boldsymbol{X})\}\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)/\pi(\boldsymbol{X})\}$ instead. Apparently the new influence function has variance no smaller than that of the previous one. This leads to the counter-intuitive conclusion that, even if $\pi(\boldsymbol{X})$ is completely known, correctly modeling $\pi(\boldsymbol{X})$ may lead to efficiency gain. Refer to Robins, Rotnitzky and Zhao (1995) for more discussion on this observation.

Efficiency cannot be enhanced without a limit. When one model for $E(Y \mid \boldsymbol{X})$ is correctly specified, $E(Y \mid \boldsymbol{X}) - \mu_0$ is a component of $\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$, and thus $\{R - \pi(\boldsymbol{X})\}\{E(Y \mid \boldsymbol{X}) - \mu_0\}/\pi(\boldsymbol{X})$ is in $\text{span}\{\{R - \pi(\boldsymbol{X})\}\boldsymbol{\Xi}_1(\boldsymbol{X})/\pi(\boldsymbol{X})\}$. It is easy to verify that

$$E\left(\left[\frac{R}{\pi(\boldsymbol{X})}(Y - \mu_0) - \frac{R - \pi(\boldsymbol{X})}{\pi(\boldsymbol{X})}\{E(Y \mid \boldsymbol{X}) - \mu_0\} \right] \left\{ \frac{R - \pi(\boldsymbol{X})}{\pi(\boldsymbol{X})} h(\boldsymbol{X}) \right\} \right) = 0$$

for any function $h(\boldsymbol{X})$. Therefore, the influence function of $\hat{\mu}_1$ simplifies to $R(Y - \mu_0)/\pi(\boldsymbol{X}) - \{R - \pi(\boldsymbol{X})\}\{E(Y \mid \boldsymbol{X}) - \mu_0\}/\pi(\boldsymbol{X})$. This is the efficient influence function for estimating μ_0 (Robins, Rotnitzky and Zhao 1994). In other words, $\hat{\mu}_1$ attains the semiparametric efficiency bound in this case. Postulating more models for $E(Y \mid \boldsymbol{X})$, augmenting $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ and/or using $\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})$ instead of $\pi(\boldsymbol{X})$ do not yield further efficiency gain.

Now we study how to achieve improved efficiency when no model for $E(Y \mid \boldsymbol{X})$ is correctly specified. Let

$$\Psi(\boldsymbol{\gamma}) = \text{Resid}\left[\frac{R(Y-\mu_0)}{\pi(\boldsymbol{X})}, \frac{R-\pi(\boldsymbol{X})}{\pi(\boldsymbol{X})}\left\{\boldsymbol{g}(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})^{\mathrm{T}}, \frac{\pi_{\boldsymbol{\alpha}}(\boldsymbol{X})^{\mathrm{T}}}{1-\pi(\boldsymbol{X})}\right\}^{\mathrm{T}}\right]$$

denote the influence function in (5) viewed as a function of $\boldsymbol{\gamma}$. To achieve improved efficiency, in (3) and (4), we need to replace $\hat{\boldsymbol{\gamma}}$ by a $\tilde{\boldsymbol{\gamma}} = \{(\tilde{\boldsymbol{\gamma}}^1)^{\mathrm{T}}, \ldots, (\tilde{\boldsymbol{\gamma}}^K)^{\mathrm{T}}\}^{\mathrm{T}}$ whose probability limit minimizes $\sigma^2(\boldsymbol{\gamma}) = \text{Var}\{\Psi(\boldsymbol{\gamma})\}$. Such a $\tilde{\boldsymbol{\gamma}}$ may be obtained by minimizing a consistent estimator of $\sigma^2(\boldsymbol{\gamma})$. Let $\boldsymbol{g}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}) = [\boldsymbol{g}(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})^{\mathrm{T}}, \pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0;\boldsymbol{X})^{\mathrm{T}}/\{1-\pi(\boldsymbol{\alpha}_0;\boldsymbol{X})\}]^{\mathrm{T}}$,

$$\boldsymbol{L}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}) = E\left\{\frac{R}{\pi(\boldsymbol{\alpha}_0;\boldsymbol{X})}\frac{1-\pi(\boldsymbol{\alpha}_0;\boldsymbol{X})}{\pi(\boldsymbol{\alpha}_0;\boldsymbol{X})}(Y-\mu_0)\boldsymbol{g}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})\right\},$$

$$\boldsymbol{G}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}) = E\left\{\frac{R}{\pi(\boldsymbol{\alpha}_0;\boldsymbol{X})}\frac{1-\pi(\boldsymbol{\alpha}_0;\boldsymbol{X})}{\pi(\boldsymbol{\alpha}_0;\boldsymbol{X})}\boldsymbol{g}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})^{\otimes 2}\right\},$$

we then have

$$\Psi(\boldsymbol{\gamma}) = \frac{R}{\pi(\boldsymbol{X})}(Y-\mu_0) - \frac{R-\pi(\boldsymbol{X})}{\pi(\boldsymbol{X})}\boldsymbol{L}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})^{\mathrm{T}}\boldsymbol{G}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})^{-1}\boldsymbol{g}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}).$$

Simple algebra shows that

$$\sigma^2(\boldsymbol{\gamma}) = \text{Var}(Y) + E\left[\frac{R}{\pi(\boldsymbol{X})}\frac{1-\pi(\boldsymbol{X})}{\pi(\boldsymbol{X})}\{Y-\mu_0-\boldsymbol{L}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})^{\mathrm{T}}\boldsymbol{G}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})^{-1}\boldsymbol{g}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})\}^2\right].$$

Therefore, $\tilde{\boldsymbol{\gamma}}$ may be taken as the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)}\frac{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)}\{Y_i-\hat{\mu}_1-\hat{\boldsymbol{L}}^\flat(\hat{\boldsymbol{\alpha}},\boldsymbol{\gamma})^{\mathrm{T}}\hat{\boldsymbol{G}}^\flat(\hat{\boldsymbol{\alpha}},\boldsymbol{\gamma})^{-1}\hat{\boldsymbol{g}}_i^\flat(\hat{\boldsymbol{\alpha}},\boldsymbol{\gamma})\}^2\right], \tag{6}$$

where $\hat{\boldsymbol{g}}^\flat(\hat{\boldsymbol{\alpha}},\boldsymbol{\gamma})$, $\hat{\boldsymbol{L}}^\flat(\hat{\boldsymbol{\alpha}},\boldsymbol{\gamma})$ and $\hat{\boldsymbol{G}}^\flat(\hat{\boldsymbol{\alpha}},\boldsymbol{\gamma})$ are $\boldsymbol{g}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})$, $\boldsymbol{L}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})$ and $\boldsymbol{G}^\flat(\boldsymbol{\alpha}_0,\boldsymbol{\gamma})$, respectively, with expectations replaced by sample averages, μ_0 replaced by $\hat{\mu}_1$ and $\boldsymbol{\alpha}_0$ replaced by $\hat{\boldsymbol{\alpha}}$. Let $\hat{\mu}_1'$ denote the estimator with $\hat{\boldsymbol{\gamma}}$ in both (3) and (4) replaced by $\tilde{\boldsymbol{\gamma}}$. When $\pi(\boldsymbol{X})$ is correctly modeled, $\hat{\mu}_1'$ is consistent and has intrinsic efficiency and improved efficiency. When $E(Y \mid \boldsymbol{X})$ is also correctly modeled, say, by $a^{k_0}(\boldsymbol{\gamma}^{k_0};\boldsymbol{X})$ so that $a^{k_0}(\boldsymbol{\gamma}_0^{k_0};\boldsymbol{X}) = E(Y \mid \boldsymbol{X})$ for some $\boldsymbol{\gamma}_0^{k_0}$, Lemma 2 in the Appendix shows that $\tilde{\boldsymbol{\gamma}}^{k_0} \xrightarrow{p} \boldsymbol{\gamma}_0^{k_0}$. This implies that $E(Y \mid \boldsymbol{X}) - \mu_0$ is a component of $\boldsymbol{g}(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_{**})$, where $\boldsymbol{\gamma}_{**}$ is the probability limit of $\tilde{\boldsymbol{\gamma}}$. Following the same arguments as those for the local efficiency of $\hat{\mu}_1$, $\hat{\mu}_1'$ is also locally efficient.

We now relax the assumption that $\pi(\boldsymbol{X})$ is correctly modeled; that is $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}_* \neq \boldsymbol{\alpha}_0$. It turns out that both $\hat{\mu}_1$ and $\hat{\mu}_1'$ are still consistent if $a^{k_0}(\boldsymbol{\gamma}^{k_0}; \boldsymbol{X})$ is a correctly specified model for $E(Y \mid \boldsymbol{X})$. To see this for $\hat{\mu}_1$, notice that $\hat{\boldsymbol{\gamma}}^{k_0} \xrightarrow{p} \boldsymbol{\gamma}_0^{k_0}$ and $\hat{\mu}_1 = \varpi_{1,n}/\varpi_{2,n}$, where

$$\varpi_{1,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) \exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)]},$$

$$\varpi_{2,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) \exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)]}.$$

Since

$$\begin{aligned}
\varpi_{1,n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{R_i\{Y_i - a^{k_0}(\hat{\boldsymbol{\gamma}}^{k_0}; \boldsymbol{X}_i) + a^{k_0}(\hat{\boldsymbol{\gamma}}^{k_0}; \boldsymbol{X}_i) - \hat{\eta}^{k_0} + \hat{\eta}^{k_0}\}}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) \exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)]} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{R_i\{Y_i - a^{k_0}(\hat{\boldsymbol{\gamma}}^{k_0}; \boldsymbol{X}_i)\}}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) \exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)]} + \hat{\eta}^{k_0}\varpi_{2,n} \\
&= \hat{\eta}^{k_0}\varpi_{2,n} + o_p(1),
\end{aligned}$$

where the second equality holds because of (3), we have $\hat{\mu}_1 \xrightarrow{p} \mu_0$. As for $\hat{\mu}_1'$, since $\tilde{\boldsymbol{\gamma}}^{k_0} \xrightarrow{p} \boldsymbol{\gamma}_0^{k_0}$ from Lemma 2 in the Appendix, consistency follows exactly the same arguments as above.

The estimators $\hat{\mu}_1$ and $\hat{\mu}_1'$ are based on a multiplicative calibration of the raw propensity score in the form of the exponential tilting weight, $\pi(\boldsymbol{\alpha}; \boldsymbol{X}) \exp\{\boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{b}(\boldsymbol{X})\}$. We may also consider the additive calibration in the form of the empirical likelihood weight, $\pi(\boldsymbol{\alpha}; \boldsymbol{X}) + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{b}(\boldsymbol{X})$, (e.g. Tan 2006, 2010; Qin and Zhang 2007; Chen, Leung and Qin 2008; Kim 2009; Chan 2012; Han and Wang 2013; Chan and Yam 2014; Han 2014a, 2014b; Tan and Wu 2015; Han 2016).

Specifically, now take $\boldsymbol{b}(\boldsymbol{X}) = \hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})\}$. Let $\hat{\mu}_{\mathrm{add},1}$ denote the solution to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R_i(Y_i - \mu)}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) + \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}} = 0, \tag{7}$$

where $\hat{\boldsymbol{\lambda}}$ solves

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R_i \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) + \boldsymbol{\lambda}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}} = \boldsymbol{0}. \tag{8}$$

It is clear that $\hat{\mu}_{\mathrm{add},1}$ is an analogue of $\hat{\mu}_1$ with additive calibration in replacement of multiplicative calibration.

When $\pi(\boldsymbol{X})$ is correctly modeled, the left-hand side of (8) converges in probability to

$$E\left(\frac{R\boldsymbol{g}(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)}{\pi(\boldsymbol{X})+\boldsymbol{\lambda}_*^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)\{1-\pi(\boldsymbol{X})\}}\right).$$

Since $E\{R\boldsymbol{g}(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)/\pi(\boldsymbol{X})\}=\mathbf{0}$, we must have $\boldsymbol{\lambda}_*=\mathbf{0}$. In this case, the difference between $\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X})+\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X})\}$ and $\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X})\exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X})\}/\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X})]$ is a term with order $O_p(n^{-1})$, which does not play a role in the first-order asymptotic results. Therefore, all previous discussion on the asymptotic behavior of $\hat{\mu}_1$ applies to $\hat{\mu}_{\mathrm{add},1}$. Counterparts of $\hat{\mu}_1^{\mathrm{aug}}$ and $\hat{\mu}_1'$ can be similarly defined with the same properties as before.

3.2 Numerical Implementation

Directly solving (3) or (8) is not the ideal way of deriving the calibration parameter $\hat{\boldsymbol{\lambda}}$. For example, (8) may have multiple roots, as shown by Lemma 3 in the Appendix as an illustration when $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$ is one-dimensional. Therefore, we view (3) and (8) as the first-order conditions of certain objective functions, and derive $\hat{\boldsymbol{\lambda}}$ by optimization rather than by solving equations. Note that the construction of objective functions here is only for implementation purpose, which is different from existing methods where the calibration itself is defined through constrained optimization. To facilitate the discussion, let $m=\sum_{i=1}^n R_i$ denote the number of subjects with Y observed, and index these subjects by $i=1,\ldots,m$ without loss of generality.

For (3), define

$$F_1(\boldsymbol{\lambda})=\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}\exp[\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)]}.$$

From (2), it is easy to verify that $E\{\boldsymbol{g}(\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)/\pi(\boldsymbol{X})\mid R=1\}=\mathbf{0}$, which implies that $\mathbf{0}$ is inside the convex hull of $\{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}):\ i=1,\ldots,m\}$, at least when n is large. Using this fact, Lemma 4 in the Appendix shows that $F_1(\boldsymbol{\lambda})$ has a unique and global minimizer. This minimizer must satisfy the first-order condition $\partial F_1(\boldsymbol{\lambda})/\partial\boldsymbol{\lambda}=\mathbf{0}$, which turns out to be (3). On the other hand, due to strict convexity, $F_1(\boldsymbol{\lambda})$ has no other stationary points different from the minimizer. Therefore, (3) always has a solution and the solution is unique.

For (8), define

$$F_{\mathrm{add},1}(\boldsymbol{\lambda})=-\frac{1}{n}\sum_{i=1}^{n}\frac{R_i\log[\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)+\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}]}{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)},$$

which is a strictly convex function on the domain

$$\mathcal{D}_{\text{add},1} = [\boldsymbol{\lambda} : \ \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) + \boldsymbol{\lambda}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\} > 0, \ i = 1, \ldots, m],$$

a non-empty, open and convex set. When $\mathbf{0}$ is inside the convex hull of $\{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) : \ i = 1, \ldots, m\}$, Lemma 5 in the Appendix shows that $F_{\text{add},1}(\boldsymbol{\lambda})$ has a unique and global minimizer inside $\mathcal{D}_{\text{add},1}$. Since $\mathcal{D}_{\text{add},1}$ is an open set, this minimizer must satisfy the first-order condition, which is actually (8). On the other hand, due to strict convexity, $F_{\text{add},1}(\boldsymbol{\lambda})$ has no other stationary points inside $\mathcal{D}_{\text{add},1}$. Therefore, (8) has a unique solution inside $\mathcal{D}_{\text{add},1}$.

The above discussion reveals that, $\hat{\boldsymbol{\lambda}}$ can be found by minimizing $F_1(\boldsymbol{\lambda})$ or $F_{\text{add},1}(\boldsymbol{\lambda})$ instead of directly solving (3) or (8). Such a convex minimization can be easily implemented using the Newton–Raphson algorithm.

Some caution is needed in implementing $\hat{\mu}_1^{\text{aug}}$ and $\hat{\mu}_{\text{add},1}^{\text{aug}}$ when $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ is a logistic regression model with intercept. In this case $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ is a component of $\pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}; \boldsymbol{X})/\{1 - \pi(\boldsymbol{\alpha}; \boldsymbol{X})\}$, and thus $\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}) - \hat{\theta}$ should be removed from $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ to avoid collinearity.

3.3 Some Remarks

Since $\hat{\boldsymbol{\lambda}}$ minimizes $F_1(\boldsymbol{\lambda})$ or $F_{\text{add},1}(\boldsymbol{\lambda})$, the calibrated propensity scores $\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}) \exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})]$ and $\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}) + \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X})\}$ are always positive for each $i = 1, \ldots, m$. Therefore, from (4) and (7), all the proposed estimators are convex combinations of the observed outcomes and thus are sample bounded.

It is easy to see that (2) holds not only for $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ and $a^k(\boldsymbol{\gamma}^k; \boldsymbol{X})$, $k = 1, \ldots, K$, but also for any functions of $\boldsymbol{X}$. In other words, any functions of $\boldsymbol{X}$ may be used to construct components of $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$. $\hat{\mu}_1^{\text{aug}}$ and $\hat{\mu}_{\text{add},1}^{\text{aug}}$ are derived based on this fact using $\pi_{\boldsymbol{\alpha}}(\boldsymbol{X})/\{1 - \pi(\boldsymbol{X})\}$. When $\pi(\boldsymbol{X})$ is correctly modeled, estimators derived this way are intrinsically efficient, and usually have higher efficiency as more functions are used. If $E(Y \mid \boldsymbol{X})$ is not correctly modeled by any single $a^k(\boldsymbol{\gamma}^k; \boldsymbol{X})$ but by a linear combination of these models, the resulting estimators still achieve the semiparametric efficiency bound. In this case, these estimators are still consistent even if $\pi(\boldsymbol{X})$ is incorrectly modeled. However, it is worth pointing out that, even though theoretically the asymptotic efficiency of our estimators is an increase function of K, the small sample behavior may not necessarily be this case. The numerical performance may deteriorate as the dimension of $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ gets too large. Therefore, reasonably modeling

$E(Y \mid \boldsymbol{X})$ is still necessary to balance efficiency gain and numerical performance.

Directly minimizing (6) in the calculation of $\hat{\mu}_1'$ is very challenging, if not infeasible, due to the complicated dependence on $\boldsymbol{\gamma}$. A technique employed by Tan (2008) and Cao, Tsiatis and Davidian (2009) may help simplify the minimization. Define

$$\hat{v}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma}, \boldsymbol{\tau}) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{R_i}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} \frac{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} \{Y_i - \hat{\mu}_1 - \boldsymbol{\tau}^{\mathrm{T}} \hat{\boldsymbol{g}}_i^{\flat}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma})\}^2 \right]. \tag{9}$$

The minimizer of $\hat{v}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma}, \boldsymbol{\tau})$ must satisfy the first-order condition

$$\mathbf{0}^{\mathrm{T}} = \frac{\partial}{\partial \boldsymbol{\tau}^{\mathrm{T}}} \hat{v}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma}, \boldsymbol{\tau}) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{R_i}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} \frac{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} \{Y_i - \hat{\mu}_1 - \boldsymbol{\tau}^{\mathrm{T}} \hat{\boldsymbol{g}}_i^{\flat}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma})\} \hat{\boldsymbol{g}}_i^{\flat}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma})^{\mathrm{T}} \right].$$

It is easy to see that the solution to the above equation at any $\boldsymbol{\gamma}$ is given by $\{\boldsymbol{\gamma}^{\mathrm{T}}, \hat{\boldsymbol{L}}^{\flat}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma})^{\mathrm{T}} \hat{\boldsymbol{G}}^{\flat}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma})^{-1}\}^{\mathrm{T}}$. Therefore, $\tilde{\boldsymbol{\gamma}}$ minimizing (6) is actually the corresponding subvector of $(\tilde{\boldsymbol{\gamma}}^{\mathrm{T}}, \tilde{\boldsymbol{\tau}}^{\mathrm{T}})^{\mathrm{T}}$ minimizing $\hat{v}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma}, \boldsymbol{\tau})$. The latter minimization is relatively straightforward due to the distinctness of $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$. However, this technique does not always work, since $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$ in (9) are not always identifiable. As an illustration, consider a scalar covariate X and a simple linear model for $E(Y \mid X)$, and let $\hat{\boldsymbol{g}}^{\flat}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \gamma(X - \overline{X})$ by ignoring the other components, where $\overline{X}$ is the sample average of X_i over the whole sample. Now (9) becomes

$$\frac{1}{n} \sum_{i=1}^{n} \left[\frac{R_i}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} \frac{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} \{Y_i - \hat{\mu}_1 - \tau\gamma(X_i - \overline{X})\}^2 \right],$$

where τ and γ are apparently not identifiable. Therefore, minimizing (9) may not be as straightforward as what it seems, and achieving improved efficiency based on an intrinsically efficient estimator is in general much more difficult than based on a conventional AIPW estimator $\hat{\mu}_{\mathrm{aipw}}(\boldsymbol{\gamma})$ as in Cao, Tsiatis and Davidian (2009).

As pointed out by one referee, for IPW-type estimators with a correctly specified propensity score model, linear models for $E(Y \mid \boldsymbol{X})$ with polynomial terms of $\boldsymbol{X}$ as regressors often lead to good enough efficiency in practice if those terms catch the overall shape of dependence of $E(Y \mid \boldsymbol{X})$ on $\boldsymbol{X}$. For our proposed estimators, in this special yet important case where $a^k(\boldsymbol{\gamma}^k; \boldsymbol{X})$, $k = 1, \ldots, K$, are all linear models, improved efficiency can be easily achieved. Without loss of generality, assume the linear models are also linear in $\boldsymbol{X}$. Consider $\hat{\mu}_1$ with $\boldsymbol{g}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ replaced by $\boldsymbol{g}(\boldsymbol{\alpha}) = [\pi(\boldsymbol{\alpha}; \boldsymbol{X}) - E\{\pi(\boldsymbol{\alpha}; \boldsymbol{X})\}, \boldsymbol{X}^{\mathrm{T}} - E(\boldsymbol{X}^{\mathrm{T}})]^{\mathrm{T}}$. Then the projection structure of the influence function of $\hat{\mu}_1$ automatically leads to improved efficiency.

4 A Revisit to the GPEL and GPET Estimators

Our approach to propensity score calibration can be applied to moment equalities different from (2). As an example, consider

$$E\{R/\pi(\boldsymbol{X})-1\}=0, \quad E[\{R/\pi(\boldsymbol{X})-1\}\pi(\boldsymbol{\alpha}_*;\boldsymbol{X})]=0,$$
$$E[\{R/\pi(\boldsymbol{X})-1\}a^k(\boldsymbol{\gamma}_*^k;\boldsymbol{X})]=0, \quad (k=1,\ldots,K). \tag{10}$$

Write $\boldsymbol{g}^\natural(\boldsymbol{\alpha},\boldsymbol{\gamma})=\{1,\pi(\boldsymbol{\alpha};\boldsymbol{X}),a^1(\boldsymbol{\gamma}^1;\boldsymbol{X}),\ldots,a^K(\boldsymbol{\gamma}^K;\boldsymbol{X})\}^{\mathrm{T}}$. One can now calibrate $\pi(\boldsymbol{\alpha};\boldsymbol{X})$ to be $\pi(\boldsymbol{\alpha};\boldsymbol{X})\exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\boldsymbol{g}^\natural(\boldsymbol{\alpha},\boldsymbol{\gamma})\{1-\pi(\boldsymbol{\alpha};\boldsymbol{X})\}/\pi(\boldsymbol{\alpha};\boldsymbol{X})]$ where $\hat{\boldsymbol{\lambda}}$ solves

$$\frac{1}{n}\sum_{i=1}^n\left[\left\{\frac{R_i}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)]}-1\right\}\boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\right]=\boldsymbol{0}, \tag{11}$$

an empirical version of (10), and then consider the estimator $\hat{\mu}_2$ solving

$$\frac{1}{n}\sum_{i=1}^n\frac{R_i(Y_i-\mu)}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\exp[\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)]}=0. \tag{12}$$

One may also calibrate $\pi(\boldsymbol{\alpha};\boldsymbol{X})$ to be $\pi(\boldsymbol{\alpha};\boldsymbol{X})+\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\boldsymbol{g}^\natural(\boldsymbol{\alpha},\boldsymbol{\gamma})\{1-\pi(\boldsymbol{\alpha};\boldsymbol{X})\}$ where $\hat{\boldsymbol{\lambda}}$ solves

$$\frac{1}{n}\sum_{i=1}^n\left[\left\{\frac{R_i}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}}-1\right\}\boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\right]=\boldsymbol{0}, \tag{13}$$

and then consider the estimator $\hat{\mu}_{\mathrm{add},2}$ solving

$$\frac{1}{n}\sum_{i=1}^n\frac{R_i(Y_i-\mu)}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)+\hat{\boldsymbol{\lambda}}^{\mathrm{T}}\boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}}=0.$$

The estimators $\hat{\mu}_2$ and $\hat{\mu}_{\mathrm{add},2}$ are actually the generalized pseudo exponential tilting (GPET) estimator and the generalized pseudo empirical likelihood (GPEL) estimator proposed in Tan and Wu (2015); see also Kim (2010) and Tan (2010). In Tan and Wu (2015), the above propensity score calibration for $\hat{\mu}_2$ and $\hat{\mu}_{\mathrm{add},2}$ is derived by minimizing a particular version of the modified forward and backward Kullback–Leibler distances between the desired weight w_i and the propensity score $\pi(\boldsymbol{\alpha};\boldsymbol{X}_i)$ for the complete cases subject to the constraints $w_i>0$, $\sum_{i=1}^m w_i=1$ and $\sum_{i=1}^m w_i\boldsymbol{g}_i^\natural(\boldsymbol{\alpha},\boldsymbol{\gamma})=n^{-1}\sum_{i=1}^n\boldsymbol{g}_i^\natural(\boldsymbol{\alpha},\boldsymbol{\gamma})$. Our approach, in contrast, derives the calibration by solving equations that are the empirical version of (10). While the constrained optimization approach is more principled and fundamental, our approach may

provide a more flexible solution in cases where it is not straightforward to formulate a constraint optimization, such as the calibration in Section 3 done based on moment equalities (2). This is similar to that quasi-likelihood (Wedderburn 1974) and estimating equations are powerful alternatives to the likelihood approach for certain problems where the specification of a parametric distribution is not straightforward, although the principle of maximum likelihood is more fundamental.

Similar to the results in Section 3, $\hat{\mu}_2$ and $\hat{\mu}_{\mathrm{add},2}$ are consistent if either $\pi(\boldsymbol{\alpha};\boldsymbol{X})$ or one of $a^k(\boldsymbol{\gamma}^k;\boldsymbol{X})$ is correctly specified. In addition, both estimators are intrinsically and locally efficient, and have the following asymptotic expansion when $\pi(\boldsymbol{\alpha};\boldsymbol{X})$ is correctly specified:

$$\sqrt{n}(\hat{\mu}_2-\mu_0)=\sqrt{n}(\hat{\mu}_{\mathrm{add},2}-\mu_0)+o_p(1)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{Resid}\left[\frac{R_i(Y_i-\mu_0)}{\pi(\boldsymbol{X}_i)},\frac{R_i-\pi(\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)}\boldsymbol{\Xi}_2(\boldsymbol{X}_i)\right]+o_p(1)$$

where $\boldsymbol{\Xi}_2(\boldsymbol{X})=[\boldsymbol{g}^\natural(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)^{\mathrm{T}},\pi_{\boldsymbol{\alpha}}(\boldsymbol{X})^{\mathrm{T}}/\{1-\pi(\boldsymbol{X})\}]^{\mathrm{T}}$. Due to the asymptotic equivalence between $\hat{\mu}_2$ and $\hat{\mu}_{\mathrm{add},2}$, the following discussion will focus on $\hat{\mu}_2$ only.

From the above asymptotic expansion, the influence function of $\hat{\mu}_2$ has exactly the same structure as that of $\hat{\mu}_1$, but with $\boldsymbol{\Xi}_2(\boldsymbol{X})$ in replacement of $\boldsymbol{\Xi}_1(\boldsymbol{X})$. Simple algebra shows that $\mathrm{span}\{\{R-\pi(\boldsymbol{X})\}\boldsymbol{\Xi}_2(\boldsymbol{X})/\pi(\boldsymbol{X})\}$ is the same as $\mathrm{span}\{\{R-\pi(\boldsymbol{X})\}\boldsymbol{\Xi}_1^{\mathrm{aug}}(\boldsymbol{X})/\pi(\boldsymbol{X})\}$. Therefore, $\hat{\mu}_2$ and $\hat{\mu}_1^{\mathrm{aug}}$ have the same efficiency, and both are in general more efficient than $\hat{\mu}_1$ under the same multiple models for $E(Y\mid\boldsymbol{X})$. Again, one exception occurs when $\pi(\boldsymbol{\alpha};\boldsymbol{X})$ is a logistic regression model with intercept, in which case $\hat{\mu}_2$, $\hat{\mu}_1^{\mathrm{aug}}$ and $\hat{\mu}_1$ are equally efficient. Similar to $\hat{\mu}_1^{\mathrm{aug}}$, $\hat{\mu}_2^{\mathrm{aug}}$ can be defined by adding the components $\pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha};\boldsymbol{X})/\{1-\pi(\boldsymbol{\alpha};\boldsymbol{X})\}$ to $\boldsymbol{g}^\natural(\boldsymbol{\alpha},\boldsymbol{\gamma})$. But it is easy to see that the influence function of $\hat{\mu}_2^{\mathrm{aug}}$ is the same as that of $\hat{\mu}_2$. Hence, $\hat{\mu}_2^{\mathrm{aug}}$ and $\hat{\mu}_2$ have equal efficiency.

To achieve improved efficiency, $\hat{\boldsymbol{\gamma}}$ in (11) and (12) needs to be replaced by a $\tilde{\boldsymbol{\gamma}}$ whose probability limit minimizes the asymptotic variance of $\hat{\mu}_2$. This $\tilde{\boldsymbol{\gamma}}$ can be derived in exactly the same way as that for $\hat{\mu}_1$ in Section 3.

For numerical implementation, from Tan (2010) and Tan and Wu (2015), $\hat{\boldsymbol{\lambda}}$ solving (11) can be derived by minimizing

$$F_2(\boldsymbol{\lambda})=\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i}{\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}\exp[\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{X}_i)]}+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\right\},$$

and $\hat{\boldsymbol{\lambda}}$ solving (13) can be derived by minimizing

$$F_{\text{add},2}(\boldsymbol{\lambda}) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_i \log[\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}_i^{\natural}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}]}{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)} - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}_i^{\natural}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\right)$$

over the region

$$\mathcal{D}_{\text{add},2} = [\boldsymbol{\lambda} : \ \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}_i^{\natural}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\} > 0, \ i = 1, \ldots, m].$$

Under the condition

$$\Lambda_n = \{\boldsymbol{\lambda} : \ \boldsymbol{\lambda} \neq \mathbf{0}, \ \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}_i^{\natural}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \geq 0 \text{ for } i = 1, \ldots, m, \text{ and } \boldsymbol{\lambda}^{\mathrm{T}} n^{-1} \sum_{i=1}^{n} \boldsymbol{g}_i^{\natural}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \leq 0\} = \emptyset,$$

Tan (2010) showed that the above minimizations have unique solutions. Here we give a justification of this condition: Lemma 6 in the Appendix shows that $P(\Lambda_n = \emptyset) \to 1$ as $n \to \infty$.

5 Numerical Studies

Our numerical studies take the simulation setting of Kang and Schafer (2007). The data are generated as $\boldsymbol{X} = \{X^{(1)}, \ldots, X^{(4)}\}^{\mathrm{T}} \sim N(\mathbf{0}, \boldsymbol{I}_4)$, $Y \mid \boldsymbol{X} \sim N\{a(\boldsymbol{X}), 1\}$, and $R \mid \boldsymbol{X} \sim$ Bernoulli$\{\pi(\boldsymbol{X})\}$, where $\boldsymbol{I}_4$ is the 4×4 identity matrix, $\pi(\boldsymbol{X}) = [1 + \exp\{X^{(1)} - 0.5X^{(2)} + 0.25X^{(3)} + 0.1X^{(4)}\}]^{-1}$ and $a(\boldsymbol{X}) = 210 + 27.4X^{(1)} + 13.7\{X^{(2)} + X^{(3)} + X^{(4)}\}$. The true $\pi(\boldsymbol{X})$ leads to approximately 50% of the subjects with missing Y. As in Kang and Schafer (2007), the following variables are calculated: $Z^{(1)} = \exp\{X^{(1)}/2\}$, $Z^{(2)} = X^{(2)}/[1 + \exp\{X^{(1)}\}] + 10$, $Z^{(3)} = \{X^{(1)}X^{(3)}/25 + 0.6\}^3$ and $Z^{(4)} = \{X^{(2)} + X^{(4)} + 20\}^2$. We consider two models for $\pi(\boldsymbol{X})$:

$$\begin{aligned}
\pi^1(\boldsymbol{\alpha}^1; \boldsymbol{X}) &= [1 + \exp\{\alpha_1^1 + \alpha_2^1 X^{(1)} + \alpha_3^1 X^{(2)} + \alpha_4^1 X^{(3)} + \alpha_5^1 X^{(4)}\}]^{-1}, \\
\pi^2(\boldsymbol{\alpha}^2; \boldsymbol{X}) &= [1 + \exp\{\alpha_1^2 + \alpha_2^2 Z^{(1)} + \alpha_3^2 Z^{(2)} + \alpha_4^2 Z^{(3)} + \alpha_5^2 Z^{(4)}\}]^{-1},
\end{aligned}$$

and four models for $E(Y \mid \boldsymbol{X})$:

$$\begin{aligned}
a^1(\boldsymbol{\gamma}^1; \boldsymbol{X}) &= \gamma_1^1 + \gamma_2^1 Z^{(1)} + \gamma_3^1 Z^{(2)}, \\
a^2(\boldsymbol{\gamma}^2; \boldsymbol{X}) &= \gamma_1^2 + \gamma_2^2 Z^{(3)} + \gamma_3^2 Z^{(4)}, \\
a^3(\boldsymbol{\gamma}^3; \boldsymbol{X}) &= \gamma_1^3 + \gamma_2^3 Z^{(1)} + \gamma_3^3 Z^{(2)} + \gamma_4^3 Z^{(3)} + \gamma_5^3 Z^{(4)}, \\
a^4(\boldsymbol{\gamma}^4; \boldsymbol{X}) &= \gamma_1^4 + \gamma_2^4 X^{(1)} + \gamma_3^4 X^{(2)} + \gamma_4^4 X^{(3)} + \gamma_5^4 X^{(4)}.
\end{aligned}$$

It is clear that $\pi^1(\boldsymbol{\alpha}^1;\boldsymbol{X})$ and $a^4(\boldsymbol{\gamma}^4;\boldsymbol{X})$ are correctly specified and the rest are incorrectly specified. Due to the similarity in efficiency and robustness properties between our proposed estimators and Tan and Wu's (2015) GPET and GPEL estimators ($\hat{\mu}_2$, $\hat{\mu}_2^{\text{aug}}$, $\hat{\mu}_{\text{add},2}$ and $\hat{\mu}_{\text{add},2}^{\text{aug}}$), we include all of them in our simulation studies. The simulation results are summarized based on the same 2000 replications, and thus comparisons can be made across different tables.

Table 1 focuses on efficiency assessments under different combinations of models for $E(Y \mid \boldsymbol{X})$ when $\pi(\boldsymbol{X})$ is correctly modeled by $\pi^1(\boldsymbol{\alpha}^1;\boldsymbol{X})$. Compared to $\mu_0 = E(Y) = 210$, all estimators have ignorable bias. When two models $a^1(\boldsymbol{\gamma}^1;\boldsymbol{X})$ and $a^2(\boldsymbol{\gamma}^2;\boldsymbol{X})$ are used instead of $a^2(\boldsymbol{\gamma}^2;\boldsymbol{X})$ only, each estimator has smaller root mean square error (RMSE), consistent with the intrinsic efficiency property. Since $a^1(\boldsymbol{\gamma}^1;\boldsymbol{X})$ and $a^2(\boldsymbol{\gamma}^2;\boldsymbol{X})$ are linear models, improved efficiency can be easily achieved by replacing them with $\boldsymbol{Z}$ in $\boldsymbol{g}(\boldsymbol{\alpha},\boldsymbol{\gamma})$ and $\boldsymbol{g}^\natural(\boldsymbol{\alpha},\boldsymbol{\gamma})$. Indeed, all estimators have noticeably smaller RMSE when $\boldsymbol{Z}$ is used. When all four models for $E(Y \mid \boldsymbol{X})$ are used, all estimators achieve the semiparametric efficiency bound since $a^4(\boldsymbol{\gamma}^4;\boldsymbol{X})$ is correctly specified. This is confirmed by comparing the RMSE of all estimators to that of $\hat{\mu}_{\text{aipw}}$ using $\pi^1(\boldsymbol{\alpha}^1;\boldsymbol{X})$ and $a^4(\boldsymbol{\gamma}^4;\boldsymbol{X})$ (in Table 3), which is known to be semiparametrically efficient. Since $\pi^1(\boldsymbol{\alpha}^1;\boldsymbol{X})$ is a logistic regression with intercept, under the same combination of models for $E(Y \mid \boldsymbol{X})$, all estimators in Table 1 have equal efficiency. Comparison among them indicates that $\hat{\mu}_1^{\text{aug}}$, $\hat{\mu}_{\text{add},1}^{\text{aug}}$, $\hat{\mu}_2^{\text{aug}}$ and $\hat{\mu}_{\text{add},2}^{\text{aug}}$ have better numerical performance than $\hat{\mu}_1$, $\hat{\mu}_{\text{add},1}$, $\hat{\mu}_2$ and $\hat{\mu}_{\text{add},2}$, respectively. There does not seem to be a noticeable difference in performance either between estimators based on different moment equalities (2) and (10), or between the multiplicative and additive calibrations.

Table 2 demonstrates multiple robustness. Here despite that $\pi(\boldsymbol{X})$ is incorrectly modeled by $\pi^2(\boldsymbol{\alpha}^2;\boldsymbol{X})$, each estimator is consistent due to the correct model $a^4(\boldsymbol{\gamma}^4;\boldsymbol{X})$. Indeed, each estimator in Table 2 has ignorable bias. If all four models for $E(Y \mid \boldsymbol{X})$ are used, the RMSE of $\hat{\mu}_1^{\text{aug}}$ and $\hat{\mu}_{\text{add},1}^{\text{aug}}$ is noticeably larger than that of $\hat{\mu}_1$ and $\hat{\mu}_{\text{add},1}$, respectively, when $n = 200$, and is noticeably smaller when $n = 1000$, indicating the sensitivity of $\hat{\mu}_1^{\text{aug}}$ and $\hat{\mu}_{\text{add},1}^{\text{aug}}$ to the number of models when n is not large. Other than this case, $\hat{\mu}_1^{\text{aug}}$, $\hat{\mu}_{\text{add},1}^{\text{aug}}$, $\hat{\mu}_2^{\text{aug}}$ and $\hat{\mu}_{\text{add},2}^{\text{aug}}$ in general have similar or noticeably smaller RMSE than $\hat{\mu}_1$, $\hat{\mu}_{\text{add},1}$, $\hat{\mu}_2$ and $\hat{\mu}_{\text{add},2}$, respectively, especially when $n = 1000$. Estimators based on moment equalities (2) generally have similar or noticeably larger RMSE compared to those based on (10). The comparison between multiplicative and additive calibrations does not seem to yield a superiority of one over the other.

Table 3 contains the comparison of the calibration-based estimators with some existing ones, including $\hat{\mu}_{\text{ipw}}$, $\hat{\mu}_{\text{aipw}}$, $\hat{\mu}_{\text{CTD}}$ (Cao, Tsiatis and Davidian 2009) and $\hat{\mu}_{\text{RLSR}}$ (Rotnitzky et al. 2012). For the particular data generating process under our consideration, the values of $\pi^2(\hat{\boldsymbol{\alpha}}^2; \boldsymbol{X})$ for a few complete cases are erroneously close to zero, yielding extremely large inverse probability weights (Robins et al. 2007). Therefore, in our comparison we use some variants of the IPW and AIPW estimators, still denoted by $\hat{\mu}_{\text{ipw}}$ and $\hat{\mu}_{\text{aipw}}$, where, for models $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$ and $a(\boldsymbol{\gamma}; \boldsymbol{X})$, $\hat{\mu}_{\text{ipw}} = \{\sum_{i=1}^n R_i Y_i/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}/\{\sum_{i=1}^n R_i/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}$ and $\hat{\mu}_{\text{aipw}} = n^{-1}\sum_{i=1}^n a(\hat{\boldsymbol{\gamma}}; \boldsymbol{X}_i) + [\sum_{i=1}^n R_i\{Y_i - a(\hat{\boldsymbol{\gamma}}; \boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)]/\{\sum_{i=1}^n R_i/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}$ (Kang and Schafer 2007). Again, it is seen that $\hat{\mu}_1^{\text{aug}}$, $\hat{\mu}_{\text{add},1}^{\text{aug}}$, $\hat{\mu}_2^{\text{aug}}$ and $\hat{\mu}_{\text{add},2}^{\text{aug}}$ in general have similar or better performance than $\hat{\mu}_1$, $\hat{\mu}_{\text{add},1}$, $\hat{\mu}_2$ and $\hat{\mu}_{\text{add},2}$, respectively, and estimators based on moment equalities (10) occasionally have better performance than those based on (2). The two types of calibration have similar overall performance with neither one dominating the other. Both $\hat{\mu}_{\text{ipw}}$ and $\hat{\mu}_{\text{aipw}}$ have unsatisfactory performance due to their lack of desirable efficiency properties and sensitivity to extreme inverse probability weights. The propensity score calibration greatly reduces the impact of extreme values of $\pi^2(\hat{\boldsymbol{\alpha}}^2; \boldsymbol{X})$ when both $\pi(\boldsymbol{X})$ and $E(Y \mid \boldsymbol{X})$ are incorrectly modeled.

6 Discussion

We have investigated an alternative approach to propensity score calibration. Unlike existing methods where the calibration is derived by constrained optimizations, our approach carries out the calibration by solving the empirical version of certain moment equalities. This approach saves the non-trivial work of constructing an objective function for optimization in order to achieve some desirable properties for the final estimators. We expect that this approach can be generalized to solve many problems more complex than the ones we have considered, such as causal inference problems or regression analysis with missing data.

Numerical performance of the proposed estimators may be unstable when the number of models for $E(Y \mid \boldsymbol{X})$ gets too large, especially if those models lead to collinearity among components of $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ or $\boldsymbol{g}^\natural(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$. One way to avoid collinearity is to check the correlation coefficient (or other quantities that measure the correlation) among the fitted values $a^k(\hat{\boldsymbol{\gamma}}; \boldsymbol{X})$, $k = 1, \ldots, K$, and remove the model that has very high correlation with others or combine

the highly correlated models into a single one. More generally, it is worthwhile to study how to balance the number of models and the numerical performance. With multiple models allowed, the focus is no longer on how well an individual model is fitted, but rather on how well these models could work together to ensure a better performance of the final estimator. More investigation on this is needed.

Acknowledgement

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

Appendix

Proof of Theorem 1. Taking Taylor expansion of the left-hand side of (3) at $(\boldsymbol{\lambda}_*^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}, \boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\gamma}_*^{\mathrm{T}})^{\mathrm{T}}$ and solving for $\sqrt{n}\hat{\boldsymbol{\lambda}}$ leads to

$$\sqrt{n}\hat{\boldsymbol{\lambda}} = \boldsymbol{G}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i - \pi(\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)}\boldsymbol{g}_i(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) - \boldsymbol{B}\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)\right\} + o_p(1),$$

where

$$\boldsymbol{G} = E\left\{\frac{1-\pi(\boldsymbol{X})}{\pi(\boldsymbol{X})}\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)^{\otimes 2}\right\}, \quad \boldsymbol{B} = E\left\{\frac{\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)}{\pi(\boldsymbol{X})}\frac{\partial\pi(\boldsymbol{\alpha}_0; \boldsymbol{X})}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}.$$

Using this result, taking Taylor expansion of the left-hand side of (4) at $(\boldsymbol{\lambda}_*^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}, \boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\gamma}_*^{\mathrm{T}}, \mu_0)^{\mathrm{T}}$ and solving for $\sqrt{n}(\hat{\mu}_1 - \mu_0)$ leads to

$$\begin{aligned}\sqrt{n}(\hat{\mu}_1 - \mu_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{R_i}{\pi(\boldsymbol{X}_i)}(Y_i - \mu_0) - \boldsymbol{L}^{\mathrm{T}}\boldsymbol{G}^{-1}\frac{R_i - \pi(\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)}\boldsymbol{g}_i(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\right\} \\ &\quad -E\left\{\frac{Y - \mu_0 - \boldsymbol{L}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)}{\pi(\boldsymbol{X})}\frac{\partial\pi(\boldsymbol{\alpha}_0; \boldsymbol{X})}{\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(1),\end{aligned}$$

where $\boldsymbol{L} = E[\{1-\pi(\boldsymbol{X})\}(Y-\mu_0)\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)/\pi(\boldsymbol{X})]$. It is easy to verify that

$$\begin{aligned}&E\left\{\frac{Y - \mu_0 - \boldsymbol{L}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)}{\pi(\boldsymbol{X})}\frac{\partial\pi(\boldsymbol{\alpha}_0; \boldsymbol{X})}{\partial\boldsymbol{\alpha}}\right\} \\ =\ & -E\left[\frac{\partial}{\partial\boldsymbol{\alpha}}\left\{\frac{R(Y-\mu_0)}{\pi(\boldsymbol{\alpha}_0; \boldsymbol{X})} - \boldsymbol{L}^{\mathrm{T}}\boldsymbol{G}^{-1}\frac{R - \pi(\boldsymbol{\alpha}_0; \boldsymbol{X})}{\pi(\boldsymbol{\alpha}_0; \boldsymbol{X})}\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\right\}\right] \\ =\ & E\left[\left\{\frac{R(Y-\mu_0)}{\pi(\boldsymbol{X})} - \boldsymbol{L}^{\mathrm{T}}\boldsymbol{G}^{-1}\frac{R - \pi(\boldsymbol{X})}{\pi(\boldsymbol{X})}\boldsymbol{g}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\right\}\boldsymbol{S}(\boldsymbol{X}, R)\right],\end{aligned}$$

where the last equality follows from the generalized information equality (e.g. Lemma 9.1 in Tsiatis 2006). Therefore, from the asymptotic expansion $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = n^{-1/2}\sum_{i=1}^n [E\{\boldsymbol{S}(\boldsymbol{X}, R)^{\otimes 2}\}]^{-1}\boldsymbol{S}(\boldsymbol{X}_i, R_i) + o_p(1)$, we have

$$\sqrt{n}(\hat{\mu}_1 - \mu_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \text{Resid}\left\{\frac{R_i(Y_i - \mu_0)}{\pi(\boldsymbol{X}_i)} - \boldsymbol{L}^{\mathrm{T}}\boldsymbol{G}^{-1}\frac{R_i - \pi(\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)}\boldsymbol{g}_i(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*), \boldsymbol{S}(\boldsymbol{X}_i, R_i)\right\} + o_p(1).$$

On the other hand, it is easy to verify that

$$\frac{R_i(Y_i - \mu_0)}{\pi(\boldsymbol{X}_i)} - \boldsymbol{L}^{\mathrm{T}}\boldsymbol{G}^{-1}\frac{R_i - \pi(\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)}\boldsymbol{g}_i(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) = \text{Resid}\left\{\frac{R_i(Y_i - \mu_0)}{\pi(\boldsymbol{X}_i)}, \frac{R_i - \pi(\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)}\boldsymbol{g}_i(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\right\}.$$

The above facts, together with Lemma 1 below, imply the desired result. □

Lemma 1. $\text{Resid}\{\text{Resid}(\xi, \boldsymbol{\phi}), \boldsymbol{\varphi}\} = \text{Resid}\{\xi, (\boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{\varphi}^{\mathrm{T}})^{\mathrm{T}}\}$ *for any* $\xi \in \mathcal{H}$ *and* $\boldsymbol{\phi}$ *and* $\boldsymbol{\varphi}$ *two finite-dimensional random vectors with components all in* $\mathcal{H}$.

Proof. Let $\mathcal{H}_{\boldsymbol{\phi}} = \text{span}\{\boldsymbol{\phi}\}$, $\mathcal{H}_{\boldsymbol{\varphi}} = \text{span}\{\boldsymbol{\varphi}\}$ and $\mathcal{H}_{\boldsymbol{\phi},\boldsymbol{\varphi}} = \text{span}\{(\boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{\varphi}^{\mathrm{T}})^{\mathrm{T}}\}$. Through the Gram–Schmidt process, we can find three mutually orthogonal subspaces of $\mathcal{H}$, namely $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$, such that $\mathcal{H}_{\boldsymbol{\phi}} = \mathcal{H}_1 \oplus \mathcal{H}_2$, $\mathcal{H}_{\boldsymbol{\varphi}} = \mathcal{H}_2 \oplus \mathcal{H}_3$ and $\mathcal{H}_{\boldsymbol{\phi},\boldsymbol{\varphi}} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \mathcal{H}_3$, where $\oplus$ denotes direct sum. Let $\text{Proj}_{\mathcal{H}_{\boldsymbol{\phi}}}(\xi) = E(\xi\boldsymbol{\phi}^{\mathrm{T}})E(\boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}})^{-1}\boldsymbol{\phi}$ denote the projection of ξ onto $\mathcal{H}_{\boldsymbol{\phi}}$, then $\text{Resid}(\xi, \boldsymbol{\phi}) = \xi - \text{Proj}_{\mathcal{H}_{\boldsymbol{\phi}}}(\xi) = \xi - \text{Proj}_{\mathcal{H}_1}(\xi) - \text{Proj}_{\mathcal{H}_2}(\xi)$. Therefore, we have

$$\begin{aligned}
&\text{Resid}\{\text{Resid}(\xi, \boldsymbol{\phi}), \boldsymbol{\varphi}\} = \text{Resid}(\xi, \boldsymbol{\phi}) - \text{Proj}_{\mathcal{H}_{\boldsymbol{\varphi}}}\{\text{Resid}(\xi, \boldsymbol{\phi})\} \\
= \;& \xi - \text{Proj}_{\mathcal{H}_1}(\xi) - \text{Proj}_{\mathcal{H}_2}(\xi) - \text{Proj}_{\mathcal{H}_2}\{\text{Resid}(\xi, \boldsymbol{\phi})\} - \text{Proj}_{\mathcal{H}_3}\{\text{Resid}(\xi, \boldsymbol{\phi})\} \\
= \;& \xi - \text{Proj}_{\mathcal{H}_1}(\xi) - \text{Proj}_{\mathcal{H}_2}(\xi) - \text{Proj}_{\mathcal{H}_3}(\xi) = \xi - \text{Proj}_{\mathcal{H}_{\boldsymbol{\phi},\boldsymbol{\varphi}}}(\xi) = \text{Resid}\{\xi, (\boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{\varphi}^{\mathrm{T}})^{\mathrm{T}}\},
\end{aligned}$$

where the third equality follows from the facts that projection is a linear operator and $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ are mutually orthogonal. (This proof was provided by an undergraduate student supervised by the author.) □

Lemma 2. *If* $a^{k_0}(\boldsymbol{\gamma}^{k_0}; \boldsymbol{X})$ *is a correctly specified model for* $E(Y \mid \boldsymbol{X})$ such that $a^{k_0}(\boldsymbol{\gamma}_0^{k_0}; \boldsymbol{X}) = E(Y \mid \boldsymbol{X})$ for some $\boldsymbol{\gamma}_0^{k_0}$, and $v(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{\tau})$ defined below has a unique minimizer, then $\tilde{\boldsymbol{\gamma}}^{k_0} \xrightarrow{p} \boldsymbol{\gamma}_0^{k_0}$ *regardless if* $\pi(\boldsymbol{X})$ *is correctly modeled by* $\pi(\boldsymbol{\alpha}; \boldsymbol{X})$.

Proof. Define

$$v(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{\tau}) = E\left[\frac{R}{\pi(\boldsymbol{\alpha}_*; \boldsymbol{X})}\frac{1 - \pi(\boldsymbol{\alpha}_*; \boldsymbol{X})}{\pi(\boldsymbol{\alpha}_*; \boldsymbol{X})}\{Y - \mu_0 - \boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{g}^{\flat}(\boldsymbol{\alpha}_*, \boldsymbol{\gamma})\}^2\right].$$

The minimizer of $v(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{\tau})$ must satisfy the first-order condition

$$\mathbf{0}^{\mathrm{T}} = \frac{\partial}{\partial \boldsymbol{\tau}^{\mathrm{T}}} v(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{\tau}) = E\left[\frac{R}{\pi(\boldsymbol{\alpha}_*; \boldsymbol{X})} \frac{1-\pi(\boldsymbol{\alpha}_*; \boldsymbol{X})}{\pi(\boldsymbol{\alpha}_*; \boldsymbol{X})} \{Y - \mu_0 - \boldsymbol{\tau}^{\mathrm{T}} \boldsymbol{g}^\flat(\boldsymbol{\alpha}_*, \boldsymbol{\gamma})\} \boldsymbol{g}^\flat(\boldsymbol{\alpha}_*, \boldsymbol{\gamma})^{\mathrm{T}}\right].$$

It is easy to see that the solution to the above equation at any $\boldsymbol{\gamma}$ is given by $\{\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{L}^\flat(\boldsymbol{\alpha}_*, \boldsymbol{\gamma})^{\mathrm{T}} \boldsymbol{G}^\flat(\boldsymbol{\alpha}_*, \boldsymbol{\gamma})^{-1}\}^{\mathrm{T}}$. Therefore, the minimizer of $v\{\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{G}^\flat(\boldsymbol{\alpha}_*, \boldsymbol{\gamma})^{-1} \boldsymbol{L}^\flat(\boldsymbol{\alpha}_*, \boldsymbol{\gamma})\}$ is actually a subvector of the minimizer of $v(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{\tau})$. More specifically, since $v\{\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{G}^\flat(\boldsymbol{\alpha}_*, \boldsymbol{\gamma})^{-1} \boldsymbol{L}^\flat(\boldsymbol{\alpha}_*, \boldsymbol{\gamma})\}$ is the probability limit of (6) due to the multiple robustness of $\hat{\mu}_1$, $\boldsymbol{\gamma}_{**}$, the probability limit of $\tilde{\boldsymbol{\gamma}}$, is the subvector of $(\boldsymbol{\gamma}_{**}^{\mathrm{T}}, \boldsymbol{\tau}_{**}^{\mathrm{T}})^{\mathrm{T}}$ minimizing $v(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{\tau})$. When $E(Y \mid \boldsymbol{X})$ is correctly modeled by $a^{k_0}(\boldsymbol{\gamma}^{k_0}; \boldsymbol{X})$, $v(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{\tau})$ attains its minimum 0 by taking $\boldsymbol{\gamma}^{k_0} = \boldsymbol{\gamma}_0^{k_0}$ and $\boldsymbol{\tau}$ with the (k_0+1)-th component 1 and all other components zeros. Therefore, we have $\boldsymbol{\gamma}_{**}^{k_0} = \boldsymbol{\gamma}_0^{k_0}$ assuming that $v(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}, \boldsymbol{\tau})$ has a unique minimizer. □

Lemma 3. *(8) has multiple roots when $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ is one-dimensional.*

Proof. Let $f(\lambda)$ denote the left-hand side of (8). Let $\tilde{\lambda}_i = -\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)/[\hat{g}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}]$, $i = 1, \ldots, m$. We consider the non-trivial case where there are at least three different values among $\tilde{\lambda}_i$, $i = 1, \ldots, m$. Order the $\tilde{\lambda}_i$'s and take three adjacent values $\tilde{\lambda}_j$, $\tilde{\lambda}_l$ and $\tilde{\lambda}_r$ with $\tilde{\lambda}_j < \tilde{\lambda}_l < \tilde{\lambda}_r$. It is easy to see that $\lim_{\lambda \downarrow \tilde{\lambda}_j} f(\lambda) = \infty$ and $\lim_{\lambda \uparrow \tilde{\lambda}_l} f(\lambda) = -\infty$. Therefore, due to the continuity of $f(\lambda)$ on the interval $(\tilde{\lambda}_j, \tilde{\lambda}_l)$, there must exist a root of (8) between $\tilde{\lambda}_j$ and $\tilde{\lambda}_l$. Similarly, there must exist a root between $\tilde{\lambda}_l$ and $\tilde{\lambda}_r$ as well, proving the existence of multiple roots. □

Lemma 4. *$F_1(\boldsymbol{\lambda})$ has a unique and global minimum if $\mathbf{0}$ is inside the convex hull of $\{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) : i = 1, \ldots, m\}$.*

Proof. We only need to show the existence. The uniqueness and globalness then come from the strict convexity of $F_1(\boldsymbol{\lambda})$. Since $\mathbf{0}$ is inside the convex hull of $\{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) : i = 1, \ldots, m\}$, it is also inside the convex hull of $\{\boldsymbol{t}_i : i = 1, \ldots, m\}$ where $\boldsymbol{t}_i = \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1-\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)$. Therefore, for any $\bar{\boldsymbol{\lambda}}$ with $\| \bar{\boldsymbol{\lambda}} \| = 1$, 0 is inside the convex hull of $\{\bar{\boldsymbol{\lambda}}^{\mathrm{T}} \boldsymbol{t}_i : i = 1, \ldots, m\}$, and thus $\max_{i=1,\ldots,m}(-\bar{\boldsymbol{\lambda}}^{\mathrm{T}} \boldsymbol{t}_i) > 0$. Let $\mathcal{S} = \{\bar{\boldsymbol{\lambda}} : \| \bar{\boldsymbol{\lambda}} \| = 1\}$ denote the unit sphere. Due to the compactness of $\mathcal{S}$, there exists $\bar{\boldsymbol{\lambda}}_\dagger \in \mathcal{S}$ such that $\inf_{\bar{\boldsymbol{\lambda}} \in \mathcal{S}} \max_{i=1,\ldots,m}(-\bar{\boldsymbol{\lambda}}^{\mathrm{T}} \boldsymbol{t}_i) = \max_{i=1,\ldots,m}(-\bar{\boldsymbol{\lambda}}_\dagger^{\mathrm{T}} \boldsymbol{t}_i) > 0$. Now let $c = \inf_{\boldsymbol{\lambda}} F_1(\boldsymbol{\lambda})$. Apparently $-\infty < c < \infty$. Let $\{\boldsymbol{\lambda}_j : j \geq 1\}$ be a sequence such that $\lim_{j \to \infty} F_1(\boldsymbol{\lambda}_j) = c$. Without loss of generality, assume $\boldsymbol{\lambda}_j \neq \mathbf{0}$ for

any $j \geq 1$. Write $\boldsymbol{\lambda}_j = l_j \bar{\boldsymbol{\lambda}}_j$, where $l_j = \| \boldsymbol{\lambda}_j \|$ and $\bar{\boldsymbol{\lambda}}_j = \boldsymbol{\lambda}_j / l_j$. If $\limsup_{j\to\infty} l_j = \infty$, then $\limsup_{j\to\infty} \max_{i=1,\ldots,m}(-\boldsymbol{\lambda}_j^{\mathrm{T}} \boldsymbol{t}_i) \geq \limsup_{j\to\infty} l_j \max_{i=1,\ldots,m}(-\bar{\boldsymbol{\lambda}}_\dagger^{\mathrm{T}} \boldsymbol{t}_i) = \infty$, and thus $\limsup_{j\to\infty} F_1(\boldsymbol{\lambda}_j) = \infty$, which contradicts $\lim_{j\to\infty} F_1(\boldsymbol{\lambda}_j) = c < \infty$. Thus we must have $\limsup_{j\to\infty} l_j < \infty$. In other words, $\{\boldsymbol{\lambda}_j : \ j \geq 1\}$ is inside a compact set $\mathcal{D}_1$. Due to the compactness, we can find $\{\boldsymbol{\lambda}_{j'} : \ j' \geq 1\}$, a subsequence of $\{\boldsymbol{\lambda}_j : \ j \geq 1\}$, that converges to $\boldsymbol{\lambda}_\circledast$ and $\boldsymbol{\lambda}_\circledast \in \mathcal{D}_1$. Since $\{F_1(\boldsymbol{\lambda}_{j'}) : \ j' \geq 1\}$ is a subsequence of $\{F_1(\boldsymbol{\lambda}_j) : \ j \geq 1\}$, we must have $F_1(\boldsymbol{\lambda}_\circledast) = F_1(\lim_{j'\to\infty} \boldsymbol{\lambda}_{j'}) = \lim_{j'\to\infty} F_1(\boldsymbol{\lambda}_{j'}) = c$, where the second equality comes from the continuity of $F_1(\boldsymbol{\lambda})$. That is, a minimum of $F_1(\boldsymbol{\lambda})$ exists. □

Lemma 5. *$F_{add,1}(\boldsymbol{\lambda})$ has a unique and global minimum on $\mathcal{D}_{add,1}$ if $\mathbf{0}$ is inside the convex hull of $\{\hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) : \ i = 1, \ldots, m\}$.*

Proof. We only need to show the existence. The uniqueness and globalness then come from the strict convexity of $F_{\mathrm{add},1}(\boldsymbol{\lambda})$. Note that $\mathcal{D}_{\mathrm{add},1} = \{\boldsymbol{\lambda} : \ 1 + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{t}_i > 0, \ i = 1, \ldots, m\}$ with $\boldsymbol{t}_i = \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}/\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)$. From the same arguments as those in the proof of Lemma 3, there exists $\bar{\boldsymbol{\lambda}}_\dagger \in \mathcal{S} = \{\bar{\boldsymbol{\lambda}} : \| \bar{\boldsymbol{\lambda}} \| = 1\}$ such that $\inf_{\bar{\boldsymbol{\lambda}} \in \mathcal{S}} \max_{i=1,\ldots,m}(-\bar{\boldsymbol{\lambda}}^{\mathrm{T}} \boldsymbol{t}_i) = \max_{i=1,\ldots,m}(-\bar{\boldsymbol{\lambda}}_\dagger^{\mathrm{T}} \boldsymbol{t}_i) > 0$. For any $\boldsymbol{\lambda} \in \mathcal{D}_{\mathrm{add},1}$ and $\boldsymbol{\lambda} \neq \mathbf{0}$, write $\boldsymbol{\lambda} = l\bar{\boldsymbol{\lambda}}$, where $l = \| \boldsymbol{\lambda} \|$ and $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda}/l$. Since $1 + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{t}_i = 1 + l\bar{\boldsymbol{\lambda}}^{\mathrm{T}} \boldsymbol{t}_i > 0$, $i = 1, \ldots, m$, we have $1 > l \max_{i=1,\ldots,m}(-\bar{\boldsymbol{\lambda}}^{\mathrm{T}} \boldsymbol{t}_i) \geq l \max_{i=1,\ldots,m}(-\bar{\boldsymbol{\lambda}}_\dagger^{\mathrm{T}} \boldsymbol{t}_i) > 0$, which yields $l \leq \{\max_{i=1,\ldots,m}(-\bar{\boldsymbol{\lambda}}_\dagger^{\mathrm{T}} \boldsymbol{t}_i)\}^{-1} < \infty$. Therefore, $\mathcal{D}_{\mathrm{add},1}$ is a bounded set. Let $c = \inf_{\boldsymbol{\lambda} \in \mathcal{D}_{\mathrm{add},1}} F_{\mathrm{add},1}(\boldsymbol{\lambda})$. Apparently $c < \infty$. Let $\{\boldsymbol{\lambda}_j : \ j \geq 1\}$ be a sequence in $\mathcal{D}_{\mathrm{add},1}$ such that $\lim_{j\to\infty} F_{\mathrm{add},1}(\boldsymbol{\lambda}_j) = c$. Define

$$\mathcal{W}_1 = \{i : \ 1 \leq i \leq m, \ \liminf_{j\to\infty} [\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) + \boldsymbol{\lambda}_j^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}] = 0\},$$

$\mathcal{W}_2 = \{i : \ 1 \leq i \leq m, \ i \notin \mathcal{W}_1\}$, and

$$f_i(\boldsymbol{\lambda}) = -\frac{\log[\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) + \boldsymbol{\lambda}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\}]}{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)}, \quad i = 1, \ldots, m.$$

For any $i \in \mathcal{W}_1$, we have $\limsup_{j\to\infty} f_i(\boldsymbol{\lambda}_j) = \infty$, and for any $i \in \mathcal{W}_2$, we have $-\infty < \liminf_{j\to\infty} f_i(\boldsymbol{\lambda}_j) \leq \limsup_{j\to\infty} f_i(\boldsymbol{\lambda}_j) < \infty$, where the first inequality comes from the boundedness of $\mathcal{D}_{\mathrm{add},1}$. If $\mathcal{W}_1 \neq \emptyset$, then $\limsup_{j\to\infty} F_{\mathrm{add},1}(\boldsymbol{\lambda}_j) = \infty$, which contradicts $\lim_{j\to\infty} F_{\mathrm{add},1}(\boldsymbol{\lambda}_j) = c < \infty$. Therefore, we must have $\mathcal{W}_1 = \emptyset$. This implies that, there exists a $1 < \delta < \infty$, such that for any $j \geq 1$, $\boldsymbol{\lambda}_j \in \mathcal{D}'_{\mathrm{add},1} = [\boldsymbol{\lambda} : \ \delta^{-1} \leq \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i) + \boldsymbol{\lambda}^{\mathrm{T}} \hat{\boldsymbol{g}}_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\{1 - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{X}_i)\} \leq \delta, \ i = 1, \ldots, m] \cap \{\boldsymbol{\lambda} : \| \boldsymbol{\lambda} \| \leq \delta\}$, where the boundedness on the right comes from

the boundedness of $\mathcal{D}_{\text{add},1}$. Since $\mathcal{D}'_{\text{add},1}$ is compact and $F_{\text{add},1}(\boldsymbol{\lambda})$ is continuous, $F_{\text{add},1}(\mathcal{D}'_{\text{add},1})$ is compact, and thus $c = \lim_{j\to\infty} F_{\text{add},1}(\boldsymbol{\lambda}_j) \in F_{\text{add},1}(\mathcal{D}'_{\text{add},1})$ and $c > -\infty$. Due to the compactness of $\mathcal{D}'_{\text{add},1}$ again, we can find $\{\boldsymbol{\lambda}_{j'} : j' \geq 1\}$, a subsequence of $\{\boldsymbol{\lambda}_j : j \geq 1\}$, that converges to $\boldsymbol{\lambda}_\circledast$, and $\boldsymbol{\lambda}_\circledast \in \mathcal{D}'_{\text{add},1}$. Since $\{F_{\text{add},1}(\boldsymbol{\lambda}_{j'}) : j' \geq 1\}$ is a subsequence of $\{F_{\text{add},1}(\boldsymbol{\lambda}_j) : j \geq 1\}$, we must have $F_{\text{add},1}(\boldsymbol{\lambda}_\circledast) = F_{\text{add},1}(\lim_{j'\to\infty} \boldsymbol{\lambda}_{j'}) = \lim_{j'\to\infty} F_{\text{add},1}(\boldsymbol{\lambda}_{j'}) = c$. That is, a minimum of $F_{\text{add},1}(\boldsymbol{\lambda})$ exists. □

Lemma 6. *$P(\Lambda_n = \emptyset) \to 1$ as $n \to \infty$.*

Proof. Noting that

$$\Lambda_n = \Big\{\boldsymbol{\lambda} : \boldsymbol{\lambda} \neq \mathbf{0},\ \boldsymbol{\lambda}^{\mathrm{T}} R_i \boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \geq 0 \text{ for all } i,\ \boldsymbol{\lambda}^{\mathrm{T}} n^{-1} \sum_{i=1}^n R_i \boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\pi(\boldsymbol{X}_i) \geq 0,$$
$$\text{and } \boldsymbol{\lambda}^{\mathrm{T}} n^{-1} \sum_{i=1}^n \boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \leq 0\Big\},$$

we just need to prove that, with probability approaching one, there does not exist $\boldsymbol{\lambda} \neq \mathbf{0}$ that simultaneously satisfies the three inequality restrictions above. For any $\boldsymbol{\lambda} \neq \mathbf{0}$ satisfying the latter two inequalities, since $n^{-1}\sum_{i=1}^n \boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = n^{-1}\sum_{i=1}^n R_i \boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\pi(\boldsymbol{X}_i) + o_p(1)$, we must have (i) $\boldsymbol{\lambda}^{\mathrm{T}} n^{-1}\sum_{i=1}^n R_i \boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\pi(\boldsymbol{X}_i) = o_p(1)$. On the other hand, since the components of $\boldsymbol{g}^\natural(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)/\pi(\boldsymbol{X})$ are linearly independent because the K models for $E(Y \mid \boldsymbol{X})$ are different, we have $P(\boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{g}^\natural(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)/\pi(\boldsymbol{X}) \neq 0 \mid R = 1) > 0$, which, together with $P(R = 1) > 0$, implies that (ii) $P(\boldsymbol{\lambda}^{\mathrm{T}} R \boldsymbol{g}^\natural(\boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)/\pi(\boldsymbol{X}) \neq 0) > 0$. If $\boldsymbol{\lambda}$ also satisfies the first inequality, or equivalently $\boldsymbol{\lambda}^{\mathrm{T}} R_i \boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\pi(\boldsymbol{X}_i) \geq 0$ for all i, then from (ii) we must have $\boldsymbol{\lambda}^{\mathrm{T}} n^{-1}\sum_{i=1}^n R_i \boldsymbol{g}_i^\natural(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\pi(\boldsymbol{X}_i)$ bounded away from zero with probability approaching one, which contradicts (i). □

References

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.

Chan, K. C. G. (2012). Uniform improvement of empirical likelihood for missing response problem. *Electronic Journal of Statistics*, 6:289–302.

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

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, page DOI: https://doi.org/10.1093/biomet/asx007.

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.

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

Han, P. (2014a). A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference*, 148:101–110.

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

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

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

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.

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.

Kim, J. K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, 19:145–157.

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

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

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

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.

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

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.

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.

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.

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

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.

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.

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

Tan, Z. (2008). Comment: Improved local efficiency and double robustness. *The International Journal of Biostatistics*, 4:Article 10.

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

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

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

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61:439–447.

Table 1: Efficiency assessment under different combinations of models for $E(Y \mid \boldsymbol{X})$. Each combination is indicated by the functions inside $\{\,\}$, where a^k is model $a^k(\boldsymbol{\gamma}^k; \boldsymbol{X})$, $k = 1, 2, 3, 4$, and $\{\boldsymbol{Z}\}$ means replacing all models for $E(Y \mid \boldsymbol{X})$ by $\boldsymbol{Z}$. Here $\pi(\boldsymbol{X})$ is correctly modeled by $\pi^1(\boldsymbol{\alpha}^1; \boldsymbol{X})$. The results are summarized based on 2000 replications and have been multiplied by 100. $\mu_0 = E(Y) = 210$.

	$\{a^2\}$			$\{a^1, a^2\}$			$\{\boldsymbol{Z}\}$			$\{a^1, a^2, a^3, a^4\}$		
Estimator	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
						$n = 200$						
$\hat{\mu}_1$	-50	310	213	-7	304	212	-13	275	184	4	261	179
$\hat{\mu}_1^{\text{aug}}$	-43	302	207	5	299	205	5	266	183	4	261	180
$\hat{\mu}_{\text{add},1}$	-49	312	214	-8	306	213	-6	274	185	4	261	179
$\hat{\mu}_{\text{add},1}^{\text{aug}}$	-42	302	207	5	301	205	6	268	183	3	264	179
$\hat{\mu}_2$	-50	310	214	-8	304	211	-11	275	184	4	261	179
$\hat{\mu}_2^{\text{aug}}$	-42	302	207	5	300	204	6	266	183	4	261	179
$\hat{\mu}_{\text{add},2}$	-49	312	216	-8	307	212	-8	274	185	4	261	179
$\hat{\mu}_{\text{add},2}^{\text{aug}}$	-41	303	206	5	301	206	7	266	183	4	261	179
						$n = 1000$						
$\hat{\mu}_1$	-19	139	93	-3	132	88	-4	115	78	-1	112	75
$\hat{\mu}_1^{\text{aug}}$	-15	136	90	2	129	86	2	113	74	-1	112	75
$\hat{\mu}_{\text{add},1}$	-18	140	93	-3	132	89	-2	115	78	-1	112	75
$\hat{\mu}_{\text{add},1}^{\text{aug}}$	-14	136	89	2	129	87	2	113	74	-1	112	75
$\hat{\mu}_2$	-19	139	92	-3	132	88	-3	115	78	-1	112	75
$\hat{\mu}_2^{\text{aug}}$	-15	136	90	2	129	85	3	113	74	-1	112	75
$\hat{\mu}_{\text{add},2}$	-18	140	93	-3	132	88	-2	115	78	-1	112	75
$\hat{\mu}_{\text{add},2}^{\text{aug}}$	-13	136	89	2	129	87	2	113	74	-1	112	75

RMSE: root mean square error. MAE: median absolute error.

Table 2: Robustness assessment where $\pi(\boldsymbol{X})$ is incorrectly modeled by $\pi^2(\boldsymbol{\alpha}^2; \boldsymbol{X})$. Each combination of models for $E(Y \mid \boldsymbol{X})$ is indicated by the functions inside $\{\}$, where a^k is model $a^k(\boldsymbol{\gamma}^k; \boldsymbol{X})$, $k = 1, 2, 3, 4$. The results are summarized based on 2000 replications and have been multiplied by 100. $\mu_0 = E(Y) = 210$.

	$\{a^1, a^4\}$			$\{a^2, a^4\}$			$\{a^3, a^4\}$			$\{a^1, a^2, a^3, a^4\}$		
Estimator	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
						$n = 200$						
$\hat{\mu}_1$	6	273	179	5	262	179	8	279	180	7	275	182
$\hat{\mu}_1^{\text{aug}}$	4	261	179	4	261	179	4	261	180	18	306	186
$\hat{\mu}_{\text{add},1}$	5	263	179	4	261	178	4	261	180	6	271	182
$\hat{\mu}_{\text{add},1}^{\text{aug}}$	4	261	179	4	261	179	4	261	180	16	398	189
$\hat{\mu}_2$	4	261	179	4	262	179	4	261	180	6	273	181
$\hat{\mu}_2^{\text{aug}}$	4	261	179	4	261	179	4	261	180	9	265	183
$\hat{\mu}_{\text{add},2}$	4	261	179	4	261	178	4	261	180	5	262	181
$\hat{\mu}_{\text{add},2}^{\text{aug}}$	4	261	179	4	261	179	4	261	180	9	265	183
						$n = 1000$						
$\hat{\mu}_1$	9	207	75	5	151	76	6	181	75	2	136	76
$\hat{\mu}_1^{\text{aug}}$	-1	112	75	-1	112	75	-1	112	75	0	118	75
$\hat{\mu}_{\text{add},1}$	-1	112	74	1	120	76	1	148	75	5	187	75
$\hat{\mu}_{\text{add},1}^{\text{aug}}$	-1	112	75	-1	112	75	-1	112	75	-1	112	75
$\hat{\mu}_2$	0	129	75	1	134	76	0	129	75	0	129	76
$\hat{\mu}_2^{\text{aug}}$	-1	112	75	-1	112	75	-1	112	75	0	117	75
$\hat{\mu}_{\text{add},2}$	-1	112	75	-1	112	75	-1	112	75	-1	112	75
$\hat{\mu}_{\text{add},2}^{\text{aug}}$	-1	112	75	-1	112	75	-1	112	75	-1	112	75

RMSE: root mean square error. MAE: median absolute error.

Table 3: Comparison of different estimators. All estimators are based on one model for $\pi(\boldsymbol{X})$ and one model for $E(Y \mid \boldsymbol{X})$. Each combination of models is indicated by the functions inside $\{\,\}$, where π^j is model $\pi^j(\boldsymbol{\alpha}^j; \boldsymbol{X})$ and a^k is model $a^k(\boldsymbol{\gamma}^k; \boldsymbol{X})$, $j = 1, 2$ and $k = 1, 2, 3, 4$. The results are summarized based on 2000 replications and have been multiplied by 100. $\mu_0 = E(Y) = 210$.

	$\{\pi^1, a^4\}$			$\{\pi^1, a^3\}$			$\{\pi^2, a^4\}$			$\{\pi^2, a^3\}$		
Estimator	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
						$n = 200$						
$\hat{\mu}_{\text{ipw}}$	-9	388	241	-9	388	241	154	863	315	154	863	315
$\hat{\mu}_{\text{aipw}}$	4	261	179	37	352	234	3	261	179	-416	813	351
$\hat{\mu}_{\text{CTD}}$	4	261	179	0	277	189	5	261	178	-182	351	243
$\hat{\mu}_{\text{RLSR}}$	4	261	179	31	297	202	3	262	178	-170	356	244
$\hat{\mu}_1$	4	261	180	42	308	215	5	263	179	-197	394	259
$\hat{\mu}_1^{\text{aug}}$	4	261	179	17	298	201	4	261	179	-171	350	244
$\hat{\mu}_{\text{add},1}$	4	261	180	38	307	216	4	261	179	-208	378	265
$\hat{\mu}_{\text{add},1}^{\text{aug}}$	4	261	179	15	298	203	4	261	180	-173	351	247
$\hat{\mu}_2$	4	261	180	41	308	215	4	261	179	-201	372	259
$\hat{\mu}_2^{\text{aug}}$	4	261	179	18	298	201	4	261	179	-171	350	244
$\hat{\mu}_{\text{add},2}$	4	261	179	37	308	216	4	261	179	-209	378	265
$\hat{\mu}_{\text{add},2}^{\text{aug}}$	4	261	179	15	298	204	4	261	180	-173	351	247
						$n = 1000$						
$\hat{\mu}_{\text{ipw}}$	-5	176	111	-5	176	111	465	1154	242	465	1154	242
$\hat{\mu}_{\text{aipw}}$	-1	112	75	9	157	107	-1	112	75	-792	1428	485
$\hat{\mu}_{\text{CTD}}$	-1	112	75	-4	115	75	-1	112	75	-194	234	195
$\hat{\mu}_{\text{RLSR}}$	-1	112	75	6	121	84	-1	113	76	-257	324	269
$\hat{\mu}_1$	-1	112	75	11	130	89	10	185	75	-217	308	221
$\hat{\mu}_1^{\text{aug}}$	-1	112	75	2	129	86	-1	112	75	-175	219	175
$\hat{\mu}_{\text{add},1}$	-1	112	75	9	130	89	5	154	75	-235	288	239
$\hat{\mu}_{\text{add},1}^{\text{aug}}$	-1	112	75	2	129	86	-1	112	75	-177	221	178
$\hat{\mu}_2$	-1	112	75	10	130	89	-1	112	75	-221	278	222
$\hat{\mu}_2^{\text{aug}}$	-1	112	75	3	129	87	-1	112	75	-175	220	175
$\hat{\mu}_{\text{add},2}$	-1	112	75	8	130	89	-1	112	74	-239	275	238
$\hat{\mu}_{\text{add},2}^{\text{aug}}$	-1	112	75	2	129	86	-1	112	75	-177	221	178

RMSE: root mean square error. MAE: median absolute error.