

Doubly-Robust Parametric Estimation in Moment Conditions Models with Missing-At-Random Data*

Saraswata Chaudhuri[†] and Hye-Young Min[‡]

April 30, 2012

Abstract

We consider estimation of finite dimensional parameters defined by a set of moment restrictions when observations on key variables are missing-at-random. We consider the verify-in-sample and verify-out-of-sample cases of Chen, Hong and Tarozzi (2008) and present parametric alternatives to their semiparametric estimators. These alternatives are extension of the Augmented Inverse Probability Weighted (AIPW) estimator of Robins, Rotnitzky and Zhao (1994). We also consider a modification proposed by Cao, Tsiatis and Davidian (2009) and present it as a one-step update of the AIPW estimators. Compared to other parametric estimators, the AIPW estimator and its modification provide additional protection against inconsistency due to parametric misspecification. They are also locally efficient in the absence of misspecification. Simulation results in the context of missing instrumental variables suggest that both estimators are likely to be useful when the researcher is reasonably certain about key components of the parametric specifications.

JEL Classification: C12; C13; C30

Keywords: Missing data; Inverse probability weighting; Doubly-robust estimator

*We thank M. Caner, D. Frazier, M. Kejriwal, S. Park and E. Renault for carefully reading the manuscript and giving us helpful comments. We also benefited from discussions with R. Ashley, D. Guilkey, J. Hill, K. Peter, B. McManus, K. Mumford, E. Rose, A. Spanos, B. Tsang, M. Wiswall and the seminar participants at Purdue and Virginia Tech on various parts of our paper. All errors are solely our responsibility.

[†]Department of Economics, CB 3305, University of North Carolina, Chapel Hill, NC 27519. Telephone: 919-966-3962. Fax: 919-966-4986. Email: saraswata_chaudhuri@unc.edu.

[‡]Korea Fair Trade Commission, 217 BanPo DaeRo, SeoCho-Gu, Seoul, 137-966. Email: hymin714@gmail.com.

1 Introduction

We present parametric alternatives to Chen et al. (2008)’s semiparametric estimators for finite-dimensional parameters defined by a set of moment restrictions when observations on key variables are missing from the sample. The parametric estimators are based on two approaches originally proposed by Robins et al. (1994) and Cao et al. (2009). Such estimators, while common and are found useful in the biostatistics and statistics literature, have been largely overlooked in economics.¹ Our paper seeks to fill this gap.

Consider a sample $\{R_i, Z_i := (Y_i, X_i)\}_{i=1}^n$ from $\{R, Z := (Y, X)\}$. Y is a key variable and Y_i can be missing. The binary variable R_i is 1 if Y_i is missing, and is 0 otherwise. X denotes the other variables and X_i is always observed. We refer to the part of the sample with missing Y_i ’s (i.e., $R_i = 1$) as the *primary* sample, and the other part (i.e., $R_i = 0$) as the *auxiliary* sample.²

The parameter value of interest, $\theta^0 \in \Theta \subset \mathbb{R}^d$, is uniquely defined by a set of moment restrictions. We consider the following two distinct cases for defining θ^0 :

$$\text{“verify-in-sample”}: \quad E[g(Z; \theta)] = 0 \text{ if and only if } \theta = \theta^0, \quad (1)$$

$$\text{“verify-out-of-sample”}: \quad E[g(Z; \theta)|R = 1] = 0 \text{ if and only if } \theta = \theta^0. \quad (2)$$

Chen et al. (2004) present five important examples where (1) or (2) occur. The latter has received less attention, at least in the generality of (2). Following the general framework of Chen et al. (2008) we consider both set of restrictions in a unified way while describing the parametric estimators.

With missing observations, identification of θ^0 has traditionally been achieved by the “missing-at-random”(MAR) assumption.³ MAR ensures that the distribution of Y , conditional on X , is same in the primary and the auxiliary samples. This is the maintained assumption in our paper.

Our goal is consistent and, when possible, efficient parametric estimation of θ^0 under (1) or (2). The parametric alternatives presented here belong to the class of the doubly-robust locally efficient Augmented Inverse Probability Weighted (AIPW) estimators and are extensions of the original work by Robins et al. (1994) and Cao et al. (2009). We also draw heavily from Chen et al. (2008).

¹Exceptions include Wooldridge (2007), Busso et al. (2009, 2011), Graham (2011) and Graham et al. (2011a).

²This terminology is slightly different from Chen et al. (2008), but the rest of the paper is faithful to their exposition.

³See, among (many) others, Rubin (1976), Rosenbaum and Rubin (1983), Robins et al. (1994), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995), Hahn (1998), Hirano et al. (2003), Chen et al. (2008), and Graham et al. (2011a).

Related literature

Two key quantities involved in the estimation of θ^0 are $q_0(X; \theta) := E[g(Z; \theta)|X]$ and $p_0(X) := P(R = 1|X) \equiv E[R|X]$. Estimators are characterized by their treatment of these two quantities.

Chen et al. (2008) propose two semiparametric estimators, conditional expectation projection generalized method of moments (CEP-GMM) and IPW-GMM, of estimation of θ^0 .⁴ The estimation consists of two steps. The first step is a nonparametric sieve estimation of either $q_0(X; \theta)$ (for CEP-GMM) or $p_0(X)$ (for IPW-GMM). The second step is a GMM estimation of θ^0 based on a properly weighted moment vector involving the first-step nonparametric estimates. These estimators should be preferred in relatively large samples because under minimal assumptions both are consistent for θ^0 and their asymptotic variances attain suitable semiparametric efficiency bounds (SEB) established in Chen et al. (2008). However, depending on the smoothness of the functions $q_0(X; \theta)$ and $p_0(X)$, and the dimension of the proxy variables X , relatively small samples may not allow for the presence of enough terms in the basis functions that would adequately approximate $q_0(X; \theta)$ and/or $p_0(X)$.⁵

For possible better finite-sample behavior, researchers often postulate parametric models for $q_0(X; \theta)$ or $p_0(X)$. The former gives the parametric imputations (PI) estimators and the latter gives the IPW estimators [see, for e.g., Qin et al. (2008), Wooldridge (2002, 2007, 2010)]. These estimators are inconsistent if the postulated parametric models (PPMs) are incorrect. Even with correct PPMs, their asymptotic variance may not attain a suitable SEB under standard data generating processes.

Robins et al. (1994) consider estimation of parameters (θ^0) in a regression function with missing regressors and propose a method that uses PPMs for both $q_0(X; \theta)$ and $p_0(X)$. This gives the AIPW estimator that is consistent if at least one of the PPMs is correct, and hence is doubly-robust (DR) to parametric misspecifications [see Scharfstein et al. (1999)]. The asymptotic variance of the estimator also attains a suitable SEB when both PPMs are correct. This is the local efficiency (LE) property. Numerous papers have since considered the original AIPW estimator, its extensions, and various other DR estimators for the estimation of population or sub-population averages [see, for e.g., Hirano and Imbens (2001), *Statistical Science* (2007:22), Cao et al. (2009), Tan (2010, 2011)].

⁴The term “semiparametric estimator” for θ^0 is used to refer to two-step estimators when the first step involves fully nonparametric estimation of unknown (infinite dimensional) nuisance parameters. Other estimators are referred to as “parametric” although they do not require a fully specified (conditional) likelihood function.

⁵See Assumption 5(5) in Chen et al. (2004), Assumption 4(5) in Chen et al. (2008) and the simulation results for Designs 2, 3 and 4 (last 3 rows) in Graham et al. (2011b) [also see, condition (v) in Theorem 6 of Hahn (1998) and Assumption 5 of Hirano et al. (2003)].

Our paper

We follow Robins et al. (1994) and present AIPW estimators for θ^0 defined by moment restrictions (1) or (2). Our goal is more general than estimation of parameters that can be explicitly expressed as population or sub-population averages, as has been the focus of the limited use of AIPW estimators in economics. We are interested in parameter values implicitly defined by a set of moment restrictions (not necessarily for regression, as has been the focus in biostatistics). This extension seems natural but, to our knowledge, the parametric methods mentioned hitherto have not been applied to our general setup in (1) and (2) with the exception of Wooldridge (2007) and Graham et al. (2011a) (both consider (1)). Given the importance of the setup, as highlighted in Chen et al. (2004), this is an important and practical gap in the literature. Our paper seeks to fill this gap.

However, as noted by Kang and Schafer (2007), the AIPW estimators may not be better than the other parametric estimators when either PPM is incorrect. Part of the drawback is addressed by Cao et al. (2009) who propose a modification of the AIPW estimator in the context of moment restrictions (1) with scalar valued $g(Z, \theta) = Y - \theta$.⁶ We extend this modification to our setup in (1) and (2). The extension is nontrivial given the generality of our setup that includes as a special case the setup of Cao et al. (2009) and others.⁷ These modified AIPW estimators are also DR to parametric misspecifications. Their asymptotic variances attain suitable SEBs when PPMs for both $p_0(X)$ and $q_0(X; \theta)$ are correct. So they are also LE. Moreover, we design the estimators to be *A-optimal* (made precise in Section 4.1) in the class of DR-LE-AIPW estimators when the PPM for $p_0(X)$ is correct but that for $q_0(X; \theta)$ is not. To reduce computation, we present them as two-step estimators. The first step is the AIPW estimation whereas the second step is an updating to achieve A-optimality.

We anticipate that the AIPW and the modified AIPW estimators will be useful in practice when the researcher is reasonably confident that at least one of the PPMs is correct, and when semiparametric approaches are deemed less attractive due to relatively small sample size or other reasons. We take the PPMs as given (by experts) and discuss the consequences of them being correct or wrong on the asymptotic behavior of the AIPW and the modified AIPW estimators.

⁶Tan (2011) provides another interesting solution in the context of (1). We do not pursue it in the current paper.

⁷It also imposes a cost in terms of the generality of our exposition. Unlike Chen et al. (2008) (but like Wooldridge (2007) and Graham et al. (2011a)), we are restricted to just-identified models. Extension to over-identified models is straightforward for the AIPW estimators, but not for the modified AIPW estimators.

The paper is organized as follows. Section 2 describes the theoretical framework. Sections 3 and 4 present the AIPW and the modified AIPW estimators respectively. Section 5 discusses possible extensions. Section 6 presents a small Monte-Carlo experiment studying the finite-sample behavior of various parametric estimators in the context of a missing instrumental variables regression. We conclude in Section 7. The paper is, unfortunately, heavy in notations. We collect all the notations in Section A of the Appendix for the convenience of readers. Proofs of the stated results are standard and mechanical, and hence relegated to Section B of the Appendix.

2 Framework

First we state our maintained assumption M that describes the given framework.

Assumption M:

- (1) (R, Y, X) are random variables such that $\theta^0 \in \text{int}(\Theta)$ satisfies the moment restrictions described in (1) or (2). $\Theta \subset \mathbb{R}^d$ is compact.
- (2) (i) $g : \mathcal{Z} \times \Theta \mapsto \mathbb{R}^d$. \mathcal{Z} is the support of $Z := (Y, X)$. (ii) $\|g(z; \theta)\|^2 \leq b(z)$ for all $z \in \mathcal{Z}$ and $\theta \in \Theta$ where $b(z) \geq 0$ and $E[b(Z)] < \infty$. (iii) $g(z; \theta)$ is continuous in $\theta \in \Theta$ for each $z \in \mathcal{Z}$. (iv) $g(z; \theta)$ is continuously differentiable in $\theta \in \text{int}(\Theta)$ for each $z \in \mathcal{Z}$. There exists an open neighborhood $\mathcal{N}(\theta^0) \subset \Theta$ containing θ^0 such that for all $\theta \in \mathcal{N}(\theta^0)$, $E[\|G(Z; \theta)\|] < \infty$ and $E[\|G(Z; \theta)\| | R = 1] < \infty$ where $G(z; \theta) := \frac{\partial}{\partial \theta'} g(z; \theta)$. $G_{[1]} := E[G(Z; \theta^0)]$ and $G_{[2]} := E[G(Z; \theta^0) | R = 1]$ have full rank d .
- (3) $(R_i, Y_i, X_i)_{i=1}^n$ is an i.i.d. sample from (R, Y, X) , but Y_i is missing if and only if $R_i = 1$. There exist constants κ_*, κ^* such that $0 < \kappa_* \leq p_0(X) := P(R = 1 | X) \leq \kappa^* < 1$ a.s. X .
- (4) $Y \perp R | X$.

Remarks: M(1) defines the parameter value of interest θ^0 and the parameter space. M(2) describes the moment vector. M(3) defines the propensity score of missingness, $p_0(X)$, and imposes bounds on it. This is a technical requirement.⁸ Naturally this implies that $p_0 := P(R = 1) \in [\kappa_*, \kappa^*]$. M(4) defines the missing-at-random (MAR) setup and is crucial for identification of θ^0 . ■

⁸This is the strong-overlap assumption. Frolich (2004) and Busso et al. (2009, 2011) present systematic simulation studies of the consequences of its violation in the context of estimation of average treatment effect (and on the treated).

Now we consider three toy examples to illustrate the consequences of MAR on using only the non-missing observations (the so-called “complete-case estimator”), a practice that is common in applications. The first two examples are well known. But the third one has received little attention and hence will be the subject of the simulation study in Section 6. We note the importance of using the inverse probability weighted moment vector, and subsequently motivate the AIPW estimator. For brevity, we only consider the moment restrictions in (1) for this illustration. See Chapter 19.8 of Wooldridge (2010) for a textbook treatment of inverse probability weighting.

Toy example - 1: Missing outcome variable in ordinary least squares regression

Consider a model $Y = X\theta^0 + \epsilon$ (all scalar). Let $E[X\epsilon] = E[\epsilon] = 0$. $\theta^0 := E[XY]/E[X^2]$ is defined by (1) with $g(Z; \theta) = X(Y - X\theta)$. Using only the non-missing observations gives the estimating equations $\sum_{i=1}^n (1 - R_i)X_i(Y_i - X_i\hat{\theta}) = 0$. The estimator $\hat{\theta} \xrightarrow{P} E[(1 - R)XY]/E[(1 - R)X^2] = E[(1 - p_0(X))XY]/E[(1 - p_0(X))X^2] \neq \theta^0$ (generally) unless a stronger assumption $E[\epsilon|X] = E[\epsilon] = 0$ (i.e. $E[Y|X] = X\theta^0$) holds. On the other hand, using an inverse probability weighted moment vector $g(Z; \theta)/(1 - p_0(X))$ restores balance between the primary and the auxiliary samples. This gives the estimating equations $\sum_{i=1}^n \{(1 - R_i)/(1 - p_0(X_i))\}X_i(Y_i - X_i\hat{\theta}) = 0$ and the estimator $\hat{\theta} \xrightarrow{P} E[\frac{1-R}{1-p_0(X)}XY]/E[\frac{1-R}{1-p_0(X)}X^2] = \theta^0$ by using M(4). ■

Toy example - 2: Missing explanatory variable in ordinary least squares regression

This is similar to toy example 1 with the roles of Y and X interchanged. ■

Toy example - 3: Missing instrumental variable in instrumental variables regression

Consider a model $X_1 = X_2\theta^0 + \epsilon$. Let $E[X_2\epsilon] \neq 0$, $E[\epsilon] = 0$, $E[Y\epsilon] = 0$ and $E[YX_2] \neq 0$. $\theta^0 := E[YX_1]/E[YX_2]$ is defined by (1) with $g(Z; \theta) = Y(X_1 - X_2\theta)$. Using only the non-missing observations gives the estimating equations $\sum_{i=1}^n (1 - R_i)Y_i(X_{1i} - X_{2i}\hat{\theta}) = 0$. The estimator $\hat{\theta} \xrightarrow{P} E[(1 - R)YX_1]/E[(1 - R)YX_2] = E[(1 - p_0(X))YX_1]/E[(1 - p_0(X))YX_2] \neq \theta^0$ generally (and not even under the stronger assumption $E[\epsilon|Y] = E[\epsilon] = 0$). But, the inverse probability weighted moment vector $g(Z; \theta)/(1 - p_0(X))$ gives the estimating equations $\sum_{i=1}^n \{(1 - R_i)/(1 - p_0(X_i))\}Y_i(X_{1i} - X_{2i}\hat{\theta}) = 0$ and the estimator $\hat{\theta} \xrightarrow{P} E[\frac{1-R}{1-p_0(X)}YX_1]/E[\frac{1-R}{1-p_0(X)}YX_2] = \theta^0$ by using M(4).⁹ Since this example is apparently at odds with the literature, we present a simulation study of it in Section 6. ■

So, under MAR, inverse probability weighting of the moment vector by the unknown $p_0(X) :=$

⁹The definition of MAR in the context of toy example 3 is at odds with the more conventional use of it by Mogstad and Wiswall (2011) or Abrevaya and Donald (2011) who found consistency using only the non-missing observations.

$P(R = 1|X)$ leads to consistency. The parametric IPW estimator uses an estimator of $p_0(X)$ based on a PPM. However, it is also clear that the parametric IPW estimator is inconsistent if the PPM is incorrect. Even otherwise, its asymptotic variance does not generally attain a suitable SEB [see Hahn (1998), Hirano et al. (2003), Wooldridge (2007) and footnote 11 of Graham et al. (2011a)].

The AIPW and modified AIPW estimators are precisely meant to address these two problems of the parametric IPW estimators. These estimators work directly with a feasible version of the efficient influence function by replacing the unknown infinite dimensional nuisance parameters, such as $q_0(X; \theta)$ and $p_0(X)$, by PPMs involving unknown finite dimensional nuisance parameters. See Chen et al. (2008) for a unified treatment of the efficient influence functions and their variances, i.e., the SEBs [also see Robins et al. (1994) and Hahn (1998).] We list them below.

Under moment restrictions (1), the efficient influence function is $G_{[1]}^{-1}\psi_{[1]}^{\text{inf}}(\theta^0)$, where

$$\psi_{[1]}^{\text{inf}}(\theta) := \frac{1 - R}{1 - p_0(X)} [g(Z; \theta) - q_0(X; \theta)] + q_0(X; \theta). \quad (3)$$

The superscript “inf” stands for infeasible because (3) depends on possibly infinite dimensional unknown nuisance parameters $p_0(X)$ and $q_0(X; \theta^0)$. On the other hand, the efficient influence functions under moment restrictions (2) are different based on the propensity score $p_0(X)$ being completely unknown, partially unknown (up to finite dimensional unknown nuisance parameters), or completely known. [We do not consider the case where $p_0(X)$ is completely known because our focus is DR estimators.] If $p_0(X)$ is completely unknown, the efficient influence function is $G_{[2]}^{-1}\psi_{[2]-\text{cu}}^{\text{inf}}(\theta^0)$ where

$$\psi_{[2]-\text{cu}}^{\text{inf}}(\theta) := \frac{p_0(X)}{p_0} \psi_{[1]}^{\text{inf}}(\theta) + \frac{R - p_0(X)}{p_0} q_0(X; \theta) \quad (4)$$

and the subscript “cu” stands for completely unknown. If $p_0(X)$ is partially unknown (denoted by subscript “pu”) up to finite dimensional unknown nuisance parameters, say γ , the efficient influence function is $G_{[2]}^{-1}\psi_{[2]-\text{pu}}^{\text{inf}}(\theta^0)$, where

$$\psi_{[2]-\text{pu}}^{\text{inf}}(\theta) := \frac{p_0(X)}{p_0} \psi_{[1]}^{\text{inf}}(\theta) + \Pi \left(\frac{R - p_0(X)}{p_0} q_0(X; \theta) | S_\gamma(\gamma^0) \right). \quad (5)$$

$S_\gamma(\gamma^0)$ is the score vector (defined below in PS(2)) for γ evaluated at γ^0 , which is the unknown true value of the finite-dimensional nuisance parameters γ such that $p(X, \gamma^0) = p_0(X)$ a.s. X . $\Pi(\cdot|.)$ is

used to denote the population least squares projection.¹⁰

The corresponding SEBs, i.e., the asymptotic variances of the efficient influence functions, are

$$\text{SEB}_{[1]} := G_{[1]}^{-1} E \left[\psi_{[1]}^{\text{inf}}(\theta^0) \psi_{[1]}^{\text{inf}'}(\theta^0) \right] G_{[1]}^{-1'}, \quad (6)$$

$$\text{SEB}_{[2]-\text{cu}} := G_{[2]}^{-1} E \left[\psi_{[2]-\text{cu}}^{\text{inf}}(\theta^0) \psi_{[2]-\text{cu}}^{\text{inf}'}(\theta^0) \right] G_{[2]}^{-1'}, \quad (7)$$

$$\text{SEB}_{[2]-\text{pu}} := G_{[2]}^{-1} E \left[\psi_{[2]-\text{pu}}^{\text{inf}}(\theta^0) \psi_{[2]-\text{pu}}^{\text{inf}'}(\theta^0) \right] G_{[2]}^{-1'}. \quad (8)$$

These were established by Chen et al. (2008), but their expressions are not central to our discussion.

In the sequel we use the following abbreviations: “[1]” for “verify-in-sample”; “[2]-cu” for “verify-out-of-sample” with propensity score completely unknown; and “[2]-pu” for “verify-out-of-sample” with propensity score partially unknown (i.e., known up to finite dimensional nuisance parameters γ).

The key components for the AIPW and modified AIPW estimators are the PPMs for $q_0(X; \theta)$ and $p_0(X)$. The generic PPMs are defined in Model-CE and Model-PS.

Model-CE: For each $\theta \in \Theta$, the PPM for $q_0(X; \theta)$ is $q(X; \theta, \beta)$ where β is a $d_\beta \times 1$ vector of unknown nuisance parameters belonging to a parameter space $\mathcal{B}(\theta)$ that is a compact subspace of \mathbb{R}^{d_β} and is possibly dependent on θ . For each $\theta \in \Theta$, there exists a unique $\beta^0(\theta) \in \text{int}(\mathcal{B}(\theta))$ such that

$$\beta^0(\theta) := \arg \min_{\beta \in \mathcal{B}(\theta)} Q(\theta, \beta), \text{ where} \quad (9)$$

$$Q(\theta, \beta) := \frac{1}{2} E \left[(1 - p_0(X))(g(Z; \theta) - q(X; \theta, \beta))' (g(Z; \theta) - q(X; \theta, \beta)) \right]. \quad (10)$$

Model-PS: The PPM for $p_0(X)$ is $p(X; \gamma)$ where γ is a $d_\gamma \times 1$ vector of unknown nuisance parameters belonging to a parameter space Γ that is a compact subspace of \mathbb{R}^{d_γ} . There exists a unique $\gamma^0 \in \text{int}(\Gamma)$ such that

$$\gamma^0 := \arg \max_{\gamma \in \Gamma} L(\gamma), \text{ where} \quad (11)$$

$$L(\gamma) := E [R \log(p(X, \gamma)) + (1 - R) \log(1 - p(X; \gamma))]. \quad (12)$$

The properties of the AIPW and the modified AIPW estimators depend on the assumption PPM below defining the correctness of the PPMs. These assumptions may or may not hold.

¹⁰For any two random vectors Z_A and Z_B with mean zero and bounded second moment such that $E[Z_B Z_B']$ is nonsingular, the population least squares projection of Z_A on Z_B is $\Pi(Z_A|Z_B) := E[Z_A Z_B'] \{E[Z_B Z_B']\}^{-1} Z_B$.

Assumption PPM:

(CE) $\beta^0(\theta) \in \mathcal{B}(\theta)$ defined in (9) uniquely satisfies $\Delta_q(X; \theta, \beta) = 0$ a.s. X for each $\theta \in \Theta$ where

$$\Delta_q(X; \theta, \beta) = q_0(X; \theta) - q(X; \theta, \beta).$$

(PS) $\gamma^0 \in \Gamma$ defined in (11) uniquely satisfies $\Delta_p(X; \gamma) := p_0(X) - p(X; \gamma) = 0$ a.s. X .

For notational convenience define $\Delta_q(X; \theta) := \Delta_q(X; \theta, \beta^0(\theta))$ and $\Delta_p(X) := \Delta_p(X; \gamma^0)$.

Remarks: (10) is unconventional because the RHS contains the scalar multiple $1 - p_0(X)$. This does not affect the main results related to the AIPW estimators. A more conventional representation would require the objective function to be inverse weighted by the estimated $1 - p_0(X)$. When PPM-(CE) is not true, $\beta^0(\theta^0)$ is usually different depending on whether it is inverse weighted.¹¹ The modified AIPW estimators exploit this dependence to improve upon the AIPW estimators.

Typically $\beta^0(\theta)$ and γ^0 are estimated by nonlinear least squares (NLS) and (conditional) quasi maximum likelihood (QML) respectively. Following White (1981) and White (1982) we list below the standard assumptions required for consistent and asymptotically normally distributed NLS and QML estimators. For brevity we write the assumptions in a compact form that may not be conducive to appreciate their significance (but these assumptions are already well known).

Assumption CE:

- (1) $q(x; \theta, \beta)$ is continuous in $\theta \in \Theta$ and $\beta \in \mathcal{B}(\theta)$ for all $x \in \mathcal{X}$ (support of X). $q(x; \theta, \beta)$ is twice continuously differentiable in $\beta \in \text{int}(\mathcal{B}(\theta))$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$. The derivatives are denoted by $q_\beta(X; \theta, \beta) := \frac{\partial}{\partial \beta} q(X; \theta, \beta)$, and for each $j = 1, \dots, d$, $q_{\beta\beta, j}(X; \theta, \beta) := \frac{\partial^2}{\partial \beta \partial \beta'} q_j(X; \theta, \beta)$. $q(x; \theta, \beta)$ and $q_\beta(X; \theta, \beta)$ are continuously differentiable in $\theta \in \text{int}(\Theta)$ for all $x \in \mathcal{X}$ and for all $\beta \in \mathcal{B}(\theta)$ and $\text{int}(\mathcal{B}(\theta))$ respectively. The derivatives are denoted by $q_\theta(X; \theta, \beta) := \frac{\partial}{\partial \theta} q(X; \theta, \beta)$, and for each $j = 1, \dots, d$, $q_{\theta\beta, j}(X; \theta, \beta) := \frac{\partial^2}{\partial \theta \partial \beta'} q_j(X; \theta, \beta)$. The following dominance conditions hold — (a) $\sup_{\theta \in \Theta} \sup_{\beta \in \mathcal{B}(\theta)} \|q(x; \theta, \beta)\|^2 < b(x)$, (b) $\sup_{\theta \in \Theta} \sup_{\beta \in \text{int}(\mathcal{B}(\theta))} \|q_\beta(x; \theta, \beta)\|^2 < b(x)$, (c) $\sup_{\theta \in \Theta} \sup_{\beta \in \mathcal{N}(\beta^0(\theta))} \|q_{\beta\beta, j}(x; \theta, \beta)\| < b(x)$, (d) $\sup_{\theta \in \mathcal{N}(\theta^0)} \sup_{\beta \in \mathcal{N}(\beta^0(\theta))} \{\|q_\theta(x; \theta, \beta)\| + \|q_{\theta\beta, j}(x; \theta, \beta)\|\} < b(x)$ — for all $x \in \mathcal{X}$ where $b(X) \geq 0$ and $E[b(X)] < \infty$, and $\mathcal{N}(\beta^0(\theta)) \subset \mathcal{B}(\theta)$ is some open neighborhood containing $\beta^0(\theta)$ for $\theta \in \Theta$.

¹¹This should not be surprising because Toy example 1 showed that inverse probability weighting does not matter (for consistency) when the regression function is a conditional expectation, but does matter otherwise. However, in informal simulations (not reported here) based on the design of Kang and Schafer (2007), we found that such inverse probability weighting, as advocated by Hirano and Imbens (2001), provides remarkable improvement in the finite-sample behavior of the AIPW estimator irrespective of PPM-(PS) or PPM-(CE) being correct.

(2) $\beta^0(\theta)$ defined in (9) is such that $\sup_{\theta \in \mathcal{N}(\theta^0)} \left\| \frac{\partial}{\partial \theta'} \beta^0(\theta) \right\| < b$ for a fixed $b \geq 0$.

(3) $A^{(\beta)}(\theta, \beta) := -E \left[\frac{\partial^2}{\partial \beta \partial \beta'} Q_i(\theta, \beta) \right]$ and $B^{(\beta)}(\theta, \beta) := E \left[\frac{\partial}{\partial \beta} Q_i(\theta, \beta) \frac{\partial}{\partial \beta'} Q_i(\theta, \beta) \right]$ are continuous in $\beta \in \mathcal{N}(\beta^0(\theta))$ and in $\theta \in \Theta$, and are nonsingular at $\beta = \beta^0(\theta)$ for $\theta \in \Theta$.

Assumption PS:

(1) $p(x; \gamma)$ is continuous in $\gamma \in \Gamma$ for all $x \in \mathcal{X}$. $p(x; \gamma)$ is twice continuously differentiable in $\gamma \in \text{int}(\Gamma)$ for all $x \in \mathcal{X}$. The derivatives are denoted by $p_\gamma(X; \gamma) := \frac{\partial}{\partial \gamma} p(X; \gamma)$ and $p_{\gamma\gamma}(X; \gamma) := \frac{\partial}{\partial \gamma'} p'_\gamma(X; \gamma)$. The following dominance conditions hold: $\sup_{\gamma \in \text{int}(\Gamma)} \|p_\gamma(x; \gamma)\|^2 + \sup_{\gamma \in \mathcal{N}(\gamma^0)} \|p_{\gamma\gamma}(x; \gamma)\| < b(x)$ for all $x \in \mathcal{X}$ where $b(X) \geq 0$ and $E[b(X)] < \infty$, and $\mathcal{N}(\gamma^0) \subset \Gamma$ is an open neighborhood containing γ^0

(2) $A^{(\gamma)}(\gamma) := -E \left[\frac{\partial}{\partial \gamma'} S_{\gamma,i}(\gamma) \right]$ and $B^{(\gamma)}(\gamma) := E \left[S_{\gamma,i}(\gamma) S'_{\gamma,i}(\gamma) \right]$ are continuous in $\gamma \in \mathcal{N}(\gamma^0)$, and are nonsingular at $\gamma = \gamma^0$. [$S_{\gamma,i}(\gamma) := \frac{\partial}{\partial \gamma} L_i(\gamma) = \frac{R_i - p(X_i; \gamma)}{p(X_i; \gamma)(1 - p(X_i; \gamma))} p'_\gamma(X; \gamma)$.]

Remarks: CE(2) imposes smoothness that will be sufficient for our results.¹² We note here that this is a critical assumption that allows us to consider the various estimating equations in the subsequent sections separately. Rest of the assumptions are standard and will also be maintained for the discussion of the AIPW and the modified AIPW estimators. In terms of notation, $Q_i(\theta, \beta)$ used in CE(3) and $L_i(\gamma)$ used in PS(2) are defined in (20) and (22) respectively (and are related to the sample counterparts of (10) and (12)). ■

3 AIPW estimators under moment restrictions (1) and (2)

In this section we define the AIPW estimators for all three cases: [1], [2]-cu and [2]-pu. We provide intuitions behind them and then state their asymptotic properties in Theorem-3.2.

3.1 Definition

To define the three AIPW estimators, first we list the notations to be used. For any θ , γ and β , define the postulated parametric versions of the efficient influence functions in (3), (4) and (5) (without the

¹²Weaker conditions, such as $\|\beta^0(\theta_1) - \beta^0(\theta_2)\| \leq b\|\theta_1 - \theta_2\|$ for any $\theta_1, \theta_2 \in \mathcal{N}(\theta^0)$, should serve the same purpose with minor modifications in the proofs of our results.

pre-multiples $G_{[1]}^{-1}$ or $G_{[2]-\text{cu}}^{-1}$ or $G_{[2]-\text{pu}}^{-1}$) as

$$\psi_{[1],i}(\theta, \gamma, \beta) := \frac{1 - R_i}{1 - p(X_i; \gamma)} [g(Z_i; \theta) - q(X_i; \theta, \beta)] + q(X_i; \theta, \beta), \quad (13)$$

$$\psi_{[2]-\text{cu},i}(\theta, \gamma, \beta) := \frac{p(X_i; \gamma)}{p_0} \psi_{[1],i}(\theta, \gamma, \beta) + \frac{R_i - p(X_i; \gamma)}{p_0} q(X_i; \theta, \beta), \quad (14)$$

$$\psi_{[2]-\text{pu},i}(\theta, \gamma, \beta) := \frac{p(X_i; \gamma)}{p_0} \psi_{[1],i}(\theta, \gamma, \beta) + \Pi \left(\frac{R_i - p(X_i; \gamma)}{p_0} q(X_i; \theta, \beta) | S_{\gamma,i}(\gamma) \right). \quad (15)$$

(14) and (15) are still infeasible given θ, γ and β because of the presence of the unknown p_0 and the population projection $\Pi(\cdot)$. So define the corresponding estimated versions $\widehat{\psi}_{[1],i}(\theta, \gamma, \beta) := \psi_{[1],i}(\theta, \gamma, \beta)$, $\widehat{\psi}_{[2]-\text{cu},i}(\theta, \gamma, \beta)$ and $\widehat{\psi}_{[2]-\text{pu},i}(\theta, \gamma, \beta)$ by replacing p_0 with $\widehat{p}_0 := n_p/n$ (where $n_p = \sum_{i=1}^n R_i$ is the size of the primary sample), and $\Pi(\cdot)$ with its sample analogue $\widehat{\Pi}(\cdot)$ which, similar to $\Pi(\cdot)$, is defined as $\widehat{\Pi}(Z_{A,i} | Z_{B,i}) := \left(\sum_{j=1}^n Z_{A,j} Z'_{B,j} \right) \left(\sum_{j=1}^n Z_{B,j} Z'_{B,j} \right)^{-1} Z_{B,i}$. In principle, there should be a subscript n in the $\widehat{\psi}$'s because these are triangular arrays. We omit it to avoid further notational clutter.

The AIPW estimators for θ^0 in the three cases – [1], [2]-cu and [2]-pu – are defined as follows:

$$\text{Case - [1]} : \widehat{\theta}_{[1]} \text{ solves } \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{[1],i} \left(\widehat{\theta}_{[1]}, \widehat{\gamma}, \widehat{\beta}(\widehat{\theta}_{[1]}) \right) = 0, \quad (16)$$

$$\text{Case - [2]-cu} : \widehat{\theta}_{[2]-\text{cu}} \text{ solves } \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{[2]-\text{cu},i} \left(\widehat{\theta}_{[2]-\text{cu}}, \widehat{\gamma}, \widehat{\beta}(\widehat{\theta}_{[2]-\text{cu}}) \right) = 0, \quad (17)$$

$$\text{Case - [2]-pu} : \widehat{\theta}_{[2]-\text{pu}} \text{ solves } \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{[2]-\text{pu},i} \left(\widehat{\theta}_{[2]-\text{pu}}, \widehat{\gamma}, \widehat{\beta}(\widehat{\theta}_{[2]-\text{pu}}) \right) = 0. \quad (18)$$

For each $\theta \in \Theta$, $\widehat{\beta}(\theta)$ is the NLS estimator defined as

$$\widehat{\beta}(\theta) := \arg \min_{\beta \in \mathcal{B}(\theta)} \frac{1}{n} \sum_{i=1}^n Q_i(\theta, \beta), \text{ where} \quad (19)$$

$$Q_i(\theta, \beta) := \frac{1}{2} (1 - R_i) (g(Z_i; \theta) - q(X_i; \theta, \beta))' (g(Z_i; \theta) - q(X_i; \theta, \beta)). \quad (20)$$

$\widehat{\gamma}$ is the QML estimator defined as

$$\widehat{\gamma} := \arg \max_{\gamma \in \Gamma} \frac{1}{n} \sum_{i=1}^n L_i(\gamma), \text{ where} \quad (21)$$

$$L_i(\gamma) := R_i \log(p(X_i, \gamma)) + (1 - R_i) \log(1 - p(X_i, \gamma)). \quad (22)$$

3.2 Double-robustness (DR) and Local efficiency (LE)

AIPW estimators work directly with the influence functions in a parametric sub-model which, when correct, should have no effect on the asymptotic variances. Since they actually use the efficient influence function, the asymptotic variances attain the suitable SEBs when the PPMs are correct, meaning the estimators are LE.¹³ Less apparent is the DR property, and hence we consider this first.¹⁴ The idea is the same for all three cases (and also for the modified AIPW estimators). So, for the purpose of brevity, focus on case [1], i.e., on the estimator $\hat{\theta}_{[1]}$.

The expectation of (13), i.e., the population estimating equation for a given θ, γ^0 and $\beta^0(\theta)$ can be expressed as

$$\begin{aligned} E[\psi_{[1]}(\theta, \gamma^0, \beta^0(\theta))] &= E[g(Z; \theta)] - E\left[\left(1 - \frac{1-R}{1-p(X; \gamma^0)}\right) [g(Z; \theta) - q(X; \theta, \beta^0(\theta))]\right] \\ &= E[g(Z; \theta)] - E\left[\frac{\Delta_p(X)\Delta_q(X; \theta)}{1-p(X; \gamma^0)}\right] \end{aligned}$$

by using M(4). If PPM-(CE) or PPM-(PS) holds, i.e., if $\Delta_q(X; \theta) = 0$ for each θ or $\Delta_p(X) = 0$ a.s. X , then, by M(1), the above expectation is zero if and only if $\theta = \theta^0$.¹⁵ So consistency depends only on the validity of moment restrictions (1) (i.e., on M(1)) as long as one of the PPMs is correct. This is what leads to the so-called DR property [see Scharfstein et al. (1999)].¹⁶

For the intuition behind LE, consider any $\theta \in \text{int}(\Theta)$. A mean-value expansion gives the following useful decomposition

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\psi}_{[1],i}(\theta, \hat{\gamma}, \hat{\beta}(\theta)) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{[1],i}^{\text{inf}}(\theta) + \frac{1}{\sqrt{n}} \sum_{i=1}^n T_{[1],i}^{(1)}(\theta) + \frac{1}{\sqrt{n}} \sum_{i=1}^n T_{[1],i}^{(2)}(\theta) + \frac{1}{\sqrt{n}} \sum_{i=1}^n T_{[1],i}^{(3)}(\theta) \\ &\quad + \Psi_{\gamma}(\theta)\sqrt{n}(\hat{\gamma} - \gamma^0) + \Psi_{\beta}(\theta)\sqrt{n}(\hat{\beta}(\theta) - \beta^0(\theta)) + o_P(1), \end{aligned} \tag{23}$$

¹³Theorem 2.1 and its neighborhood in Graham (2011) give an insightful discussion of how the apparently much more informative MAR (i.e. M(4)) assumption under case [1] is captured by the form of the efficient influence function.

¹⁴Busso et al. (2009) (Section II.C) also stress on the relative importance of the DR property (to economists) in their discussion of estimators for average treatment effect (and on the treated).

¹⁵This also explains why, unlike Assumption 2.1 of Graham et al. (2011a), we do not restrict our assumption PPM-(CE) to only $\theta = \theta^0$. Doing so in our context would require imposing (not intuitive) additional assumptions for the above expectation to be zero uniquely at $\theta = \theta^0$.

¹⁶While DR could directly be shown for all three cases using the decomposition in (34) (that we actually use in the proof), we do not do so here to avoid referring the readers to the notational appendix for this crucial DR argument.

where the (new) quantities in the second and the third line in the RHS of (23) are

$$\begin{aligned}
T_{[1],i}^{(1)}(\theta) &:= -\frac{R_i - p_0(X_i)}{1 - p(X_i; \gamma^0)} \Delta_q(X_i; \theta), \\
T_{[1],i}^{(2)}(\theta) &:= -\frac{1 - R_i}{1 - p_0(X_i)} \frac{\Delta_p(X_i)}{1 - p(X_i; \gamma^0)} [g(Z_i; \theta) - q_0(X_i; \theta)], \\
T_{[1],i}^{(3)}(\theta) &:= -\frac{\Delta_p(X_i)}{1 - p(X_i; \gamma^0)} \Delta_q(X_i; \theta), \\
\Psi_\gamma(\theta) &:= E \left[\frac{\partial}{\partial \gamma'} \psi_{[1],i}(\theta, \gamma^0, \beta^0(\theta)) \right] = E \left[\frac{1 - p_0(X_i)}{1 - p(X_i; \gamma^0)} \Delta_q(X_i; \theta) \frac{p_\gamma(X_i; \gamma^0)}{1 - p(X_i; \gamma^0)} \right], \\
\Psi_\beta(\theta) &:= E \left[\frac{\partial}{\partial \beta'} \psi_{[1],i}(\theta, \gamma^0, \beta^0(\theta)) \right] = E \left[\frac{\Delta_p(X_i)}{1 - p(X_i; \gamma^0)} q_\beta(X_i; \theta, \beta^0(\theta)) \right].
\end{aligned}$$

$\sqrt{n}(\widehat{\beta}(\theta) - \beta^0(\theta)) = O_P(1)$ and $\sqrt{n}(\widehat{\gamma} - \gamma^0) = O_P(1)$ under assumptions CE and PS [see White (1981) and White (1982) or Lemma 3.1 below]. Now consider the scenario where both PPM-(CE) and PPM-(PS) hold, i.e., $\Delta_q(X; \theta) = 0$ for each θ and $\Delta_p(X) = 0$ a.s. X . Therefore $T_{[1],i}^{(1)}(\theta) = T_{[1],i}^{(2)}(\theta) = T_{[1],i}^{(3)}(\theta) = 0$ a.s. X , $\Psi_\gamma(\theta) = 0$ and $\Psi_\beta(\theta) = 0$. This implies

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{\psi}_{[1],i}(\theta^0, \widehat{\gamma}, \widehat{\beta}(\theta^0)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{[1],i}^{\text{inf}}(\theta^0) + o_P(1), \text{ and} \\
\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta'} \widehat{\psi}_{[1],i}(\theta^0, \widehat{\gamma}, \widehat{\beta}(\theta^0)) &= G_{[1]} + o_P(1),
\end{aligned}$$

and hence the asymptotic variance of the AIPW estimator under case [1] attains the SEB in (6). Additionally, the decomposition in (23) also shows that when PPM-(CE) is correct, estimation of γ does not affect the asymptotic variance of the AIPW estimators. Similarly, when PPM-(PS) is correct, estimation of β does not affect the asymptotic variance of the AIPW estimators.

The same intuition behind DR, LE and asymptotic variance (in general) holds for the AIPW estimators defined in (17), (18) and the modified AIPW estimators to be defined in the next section.

3.3 Asymptotic properties

An useful intermediate result for the asymptotic properties of the AIPW and the modified AIPW estimators is stated in Lemma-3.1. Then the main result of the paper is presented in Theorem 3.2.

Lemma 3.1 *Let assumptions M, CE, PS hold. Then the following results hold for $\widehat{\gamma}$ and $\widehat{\beta}(\theta)$ defined in (21) and (19) respectively as $n \rightarrow \infty$:*

$$(a1) \hat{\gamma} \xrightarrow{P} \gamma^0.$$

$$(a2) \sqrt{n} (\hat{\gamma} - \gamma^0) \xrightarrow{d} N \left(0, [A^{(\gamma)}]^{-1} B^{(\gamma)} [A^{(\gamma)}]^{-1'} \right).$$

$$(b1) \hat{\beta}(\theta) \xrightarrow{P} \beta^0(\theta) \text{ uniformly in } \theta \in \Theta.$$

$$(b2) \text{ For any } \theta \in \Theta, \sqrt{n} \left(\hat{\beta}(\theta) - \beta^0(\theta) \right) \xrightarrow{d} N \left(0, [A^{(\beta)}(\theta, \beta^0(\theta))]^{-1} B^{(\beta)}(\theta, \beta^0(\theta)) [A^{(\beta)}(\theta, \beta^0(\theta))]^{-1'} \right).$$

Expressions for $A^{(\gamma)}$, $B^{(\gamma)}$, $A^{(\beta)}(\theta, \beta^0(\theta))$ and $B^{(\beta)}(\theta, \beta^0(\theta))$ are given in Section A of the Appendix.

These results are well known since White (1981) and White (1982); the only apparent difference being the underlying notion of uniformity in θ . The following theorem describes the properties of the AIPW estimators for the three cases: [1], [2]-cu and [2]-pu. This is the main result of the paper.

Theorem 3.2 *Let assumptions M, CE, PS hold. The postulated models Model-CE and Model-PS are given. Then the following results hold for the AIPW estimators $\hat{\theta}_{[1]}$, $\hat{\theta}_{[2]-cu}$ and $\hat{\theta}_{[2]-pu}$ defined in (16), (17) and (18) respectively, as $n \rightarrow \infty$:*

(i) *If PPM-(CE) or (and) PPM-(PS) holds, then $\hat{\theta}_{[1]}$, $\hat{\theta}_{[2]-cu}$ and $\hat{\theta}_{[2]-pu}$ are consistent for θ^0 .*

(ii) (a) *If both PPM-(CE) and PPM-(PS) hold, then $\hat{\theta}_{[1]}$, $\hat{\theta}_{[2]-cu}$ and $\hat{\theta}_{[2]-pu}$ converge in distribution to normal variables with mean θ^0 and variance equal to the SEBs defined in (6), (7) and (8).*

(ii) (b) *If PPM-(PS) holds but PPM-(CE) does not, then $\hat{\theta}_{[1]}$, $\hat{\theta}_{[2]-cu}$ and $\hat{\theta}_{[2]-pu}$ converge in distribution to normal variables with mean θ^0 and variance given by a generic form:*

$$SEB + G^{-1} \text{Var} \left[\hat{\xi}(\theta^0, \beta^0(\theta^0)) \right] G^{-1'}$$

where $\hat{\xi}(\theta, \beta) := \hat{T}^{(1)}(\theta, \beta) - \Pi \left(\hat{T}^{(1)}(\theta, \beta) | S_{\gamma}(\gamma^0) \right)$. The expressions of $\hat{T}^{(1)}(\theta, \beta)$ for all three cases are given in Section A of the Appendix. $G = G_{[1]}$ for case [1] and $G_{[2]}$ for cases [2]-cu and [2]-pu. SEBs for the three cases are defined in (6), (7) and (8) respectively.

(ii) (c) *If PPM-(CE) holds but PPM-(PS) does not, then $\hat{\theta}_{[1]}$ and $\hat{\theta}_{[2]-cu}$ converge in distribution to normal variables with mean θ^0 and finite variance.¹⁷*

¹⁷Unlike in (ii.a) and (ii.b), the expression for the asymptotic variance in (ii.c) does not lead to any insight about the behavior of the AIPW estimator and hence is not reported. It should be noted that case [2]-pu is not relevant for (ii.c) because it imposes PPM-(PS).

Remarks: (i) is the DR property of the AIPW estimators. Unlike other parametric estimators such as PI or IPW that are inconsistent when, respectively, PPM-(CE) or PPM-(PS) is incorrect, the AIPW estimators are consistent as long as (at least) one is correct. (ii) shows that if PPM-(CE) or PPM-(PS) is correct, then AIPW estimators are asymptotically normal with mean θ^0 and finite variance. Part (a) shows that the asymptotic variance attains the SEB. This is not necessarily true for the PI estimator [see Qin et al. (2008)] or the IPW estimator [see Wooldridge (2007) and footnote 11 of Graham et al. (2011a)]. Part(b) quantifies the inefficiency of the AIPW estimators. It is this inefficiency that the modified AIPW estimators in the next section tries to minimize. ■

4 Modified AIPW estimators

We present the modified AIPW estimator as a two-step estimator where the first step is the AIPW estimation and the second step involves updating to achieve a certain optimality. We focus on the updating step to avoid repetition. Generic quantities and statements that are valid for all three cases are reported without the subscripts “[1]”, “[2]-cu” and “[2]-pu” to avoid notational clutter.

4.1 Modification for A-optimality

The form of the asymptotic variance of the AIPW estimators given in Theorem-3.2(ii.b) motivates the modified AIPW estimators in the spirit of Cao et al. (2009). Under this scenario one may wish to estimate the parameters β in $q(X; \theta^0, \beta)$ such that it converges in probability to some $\beta^*(\theta^0) \in \text{int}(\mathcal{B}(\theta^0))$ satisfying $G^{-1} \text{Var} [\hat{\xi}(\theta^0, \beta)] G^{-1'} - G^{-1} \text{Var} [\hat{\xi}(\theta^0, \beta^*(\theta^0))] G^{-1'}$ is p.s.d. for all $\beta \in \mathcal{B}(\theta^0)$. This would minimize the variance (extra over SEB) of the AIPW estimator under the conditions of Theorem-3.2(ii.b). However, we are unable to provide a simple method that would minimize the variance and still retain the other properties of the AIPW estimator unless θ is a scalar parameter.¹⁸

A compromise is to modify the goal and induce a probability limit $\beta^*(\theta^0) \in \text{int}(\mathcal{B}(\theta^0))$ for the estimator of β such that Theorem-3.2 (i), (ii.a) and (ii.c) remain correct and, at the same time, under the conditions of Theorem-3.2(ii.b),

$$\text{Trace} \left(G^{-1} \text{Var} [\hat{\xi}(\theta^0, \beta)] G^{-1'} - G^{-1} \text{Var} [\hat{\xi}(\theta^0, \beta^*(\theta^0))] G^{-1'} \right) \geq 0 \quad (24)$$

¹⁸Over-identified models can be easily accommodated with modified AIPW estimators if θ is a scalar.

for all $\beta \in \mathcal{B}(\theta^0)$. This is in the spirit of *A-optimality* in the design of experiments literature.¹⁹ Under suitable conditions (Lemma 3.6 of Newey and McFadden (1994)), and for a given θ , $\beta^*(\theta)$ solves the first-order condition of the optimization problem in (24), i.e.,

$$\begin{aligned}
0 &= \frac{\partial}{\partial \beta} E \left[\widehat{\xi}'(\theta, \beta) G^{-1'} G^{-1} \widehat{\xi}(\theta, \beta) \right], \\
\Rightarrow 0 &= E \left[\left(\frac{\partial}{\partial \beta} \widehat{T}^{(1)'}(\theta, \beta) \right) G^{-1'} G^{-1} \left\{ \widehat{T}^{(1)}(\theta, \beta) - \Pi \left(\widehat{T}^{(1)}(\theta, \beta) | S_\gamma(\gamma^0) \right) \right\} \right] \\
&\quad - E \left[\left(\frac{\partial}{\partial \beta} \Pi \left(\widehat{T}^{(1)}(\theta, \beta) | S_\gamma(\gamma^0) \right) \right)' G^{-1'} G^{-1} \left\{ \widehat{T}^{(1)}(\theta, \beta) - \Pi \left(\widehat{T}^{(1)}(\theta, \beta) | S_\gamma(\gamma^0) \right) \right\} \right] \\
\Rightarrow 0 &= E \left[\left(\frac{\partial}{\partial \beta} \widehat{T}^{(1)'}(\theta, \beta) \right) G^{-1'} G^{-1} \left\{ \widehat{T}^{(1)}(\theta, \beta) - \Pi \left(\widehat{T}^{(1)}(\theta, \beta) | S_\gamma(\gamma^0) \right) \right\} \right]. \tag{25}
\end{aligned}$$

The term on the second last line above is zero and gives the simplification in the last line of (25).²⁰ Since γ^0 is consistently estimable and so is G (given θ^0 and γ^0), treat them as known for now. $\Pi \left(\widehat{T}^{(1)}(\theta, \beta) | S_\gamma(\gamma^0) \right)$ in (25) is just a least squares projection of $\widehat{T}^{(1)}(\theta, \beta)$ on $S_\gamma(\gamma^0)$.

Now recall that when PPM-(CE) holds, we would still require $\beta^*(\theta^0) = \beta^0(\theta^0)$ so that the results in Theorem-3.2 (i), (ii.a) and (ii.c) remain correct. To incorporate this, we follow Cao et al. (2009) and define $\beta^*(\theta)$, for a given θ , as the solution of the following augmented system. For a given $\theta \in \Theta$, let $\beta^*(\theta)$ and $\Lambda^0 := [\lambda_1^0, \dots, \lambda_d^0]'$ solve the following $d_\beta + d \times d_\gamma$ equations

$$0 = E \left[\left(\frac{\partial}{\partial \beta} \widehat{T}^{(1)'}(\theta, \beta) \right) G^{-1'} G^{-1} \left\{ \widehat{T}^{(1)}(\theta, \beta) - \Lambda^0 S_\gamma(\gamma^0) \right\} \right] \tag{26}$$

$$0 = E \left[S_\gamma(\gamma^0) \left\{ \widehat{T}_j^{(1)}(\theta, \beta) - S_\gamma'(\gamma^0) \lambda_j \right\} \right] \text{ for } j = 1, \dots, d. \tag{27}$$

Remarks:

- (i) For a given β (and θ), (27) gives $\lambda_j^0(\beta) = \{E [S_\gamma(\gamma^0) S_\gamma'(\gamma^0)]\}^{-1} E [S_\gamma(\gamma^0) \widehat{T}_j^{(1)}(\theta, \beta)]$ for

¹⁹In our context, this seems to be a more relevant goal than the related concepts of *D-optimality* (minimizing the log of the determinant) or *E-optimality* (minimizing the maximum eigen value).

²⁰To see the zero consider any $j = 1 \dots, d_\beta$ and note that

$$\begin{aligned}
&\frac{\partial}{\partial \beta_j} E \left[S_\gamma'(\gamma^0) \{B^\gamma(\gamma^0)\}^{-1} E \left[S_\gamma(\gamma^0) \widehat{T}^{(1)'}(\theta, \beta) \right] G^{-1'} G^{-1} \left\{ \widehat{T}^{(1)}(\theta, \beta) - \Pi \left(\widehat{T}^{(1)}(\theta, \beta) | S_\gamma(\gamma^0) \right) \right\} \right] \\
&= \frac{\partial}{\partial \beta_j} E \left[\text{Trace} \left(S_\gamma'(\gamma^0) \{B^\gamma(\gamma^0)\}^{-1} E \left[S_\gamma(\gamma^0) \widehat{T}^{(1)'}(\theta, \beta) \right] G^{-1'} G^{-1} \left\{ \widehat{T}^{(1)}(\theta, \beta) - \Pi \left(\widehat{T}^{(1)}(\theta, \beta) | S_\gamma(\gamma^0) \right) \right\} \right) \right] \\
&= \text{Trace} \left(\{B^\gamma(\gamma^0)\}^{-1} \frac{\partial}{\partial \beta_j} E \left[S_\gamma(\gamma^0) \widehat{T}^{(1)'}(\theta, \beta) \right] G^{-1'} G^{-1} E \left[\left\{ \widehat{T}^{(1)}(\theta, \beta) - \Pi \left(\widehat{T}^{(1)}(\theta, \beta) | S_\gamma(\gamma^0) \right) \right\} S_\gamma'(\gamma^0) \right] \right) \\
&= 0
\end{aligned}$$

because $E \left[\left\{ \widehat{T}^{(1)} - \Pi \left(\widehat{T}^{(1)} | S_\gamma \right) \right\} S_\gamma' \right] = 0$ by the definition of least squares projection.

$j = 1, \dots, d$. PS(2) implies that the solution is unique. Plugging $\Lambda^0(\beta) := [\lambda_1^0(\beta), \dots, \lambda_d^0(\beta)]'$ in (26) gives (25). As explained above, this leads to the optimality defined in (24).

(ii) Now suppose PPM-(CE) holds. M(4) implies that (27) can be equivalently written as $0 = E \left[S_\gamma(\gamma^0) \left\{ T_j^{(1)}(\theta, \beta) - S'_\gamma(\gamma^0) \lambda_j \right\} \right]$ for $j = 1, \dots, d$, by using the law of iterated expectation. But $T^{(1)}(\theta, \beta^0(\theta)) = 0$ when PPM-(CE) holds. Therefore, under PPM-(CE) and PS(2), the unique solution for (27) is $\Lambda^0 = 0_{d \times d_\gamma}$. Plugging this in (26) and using M(3) along with the same arguments as in Lemma-3.1, it follows that the unique solution to the system (26) and (27) is $\beta^*(\theta) = \beta^0(\theta)$ and $\Lambda^0 = 0_{d \times d_\gamma}$. Therefore, when PPM-(CE) holds the system gives $\beta^*(\theta) = \beta^0(\theta)$ defined in (9). ■

4.2 Definition and discussion of the modified AIPW estimator

For the sake of reference, first we list the functional forms of $\widehat{T}^{(1)}(\theta, \beta)$. For case [2]-pu we list instead a useful simplification of $\widehat{T}^{(1)}(\theta, \beta)$ from (46) in Section B of the Appendix. For all, we ignore the negative sign (in front) and the constant multiple $1/p_0$ because they do not affect our purpose here.

$$\begin{aligned} \text{Case - [1] and Case - [2]-cu:} & \quad \frac{R - p_0(X)}{1 - p(X; \gamma^0)} [g(Z; \theta) - q(X; \theta, \beta)], \\ \text{Case - [3]-pu:} & \quad p_0(X) \frac{R - p_0(X)}{1 - p_0(X)} [g(Z; \theta) - q(X; \theta, \beta)]. \end{aligned}$$

This gives, for a given θ and γ , the following unified representation for the sample analogs of (26) and (27), i.e., the sample estimating equations for β and $\Lambda = (\lambda_1, \dots, \lambda_d)'$

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{i=1}^n \frac{1 - R_i}{1 - p(X_i; \gamma)} \frac{p(X_i; \gamma)}{1 - p(X_i; \gamma)} \times \\ & \times \left\{ \begin{array}{l} c(X_i; \gamma) q'_\beta(X_i; \theta, \beta) \widehat{G}^{-1}(\theta, \gamma) \widehat{G}^{-1}(\theta, \gamma) \left[c(X_i; \gamma) (g(Z_i; \theta) - q(X_i; \theta, \beta)) - \Lambda \frac{p'_\gamma(X_i; \gamma)}{p(X_i; \gamma)} \right] \\ \frac{p'_\gamma(X_i; \gamma)}{p(X_i; \gamma)} \left[c(X_i; \gamma) (g_1(Z_i; \theta) - q_1(X_i; \theta, \beta)) - \frac{p_\gamma(X_i; \gamma)}{p(X_i; \gamma)} \lambda_1 \right] \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \frac{p'_\gamma(X_i; \gamma)}{p(X_i; \gamma)} \left[c(X_i; \gamma) (g_d(Z_i; \theta) - q_d(X_i; \theta, \beta)) - \frac{p_\gamma(X_i; \gamma)}{p(X_i; \gamma)} \lambda_d \right] \end{array} \right\}. \end{aligned} \quad (28)$$

$c(X_i; \gamma) := 1$ in cases [1] and [2]-cu and $c(X_i; \gamma) := p(X_i; \gamma)$ in case [2]-pu. On the other hand, $\widehat{G}(\theta, \gamma) := \frac{1}{n} \sum_{j=1}^n \frac{1 - R_j}{1 - p(X_j; \gamma)} G(Z_j; \theta)$ in case [1] and $\widehat{G}(\theta, \gamma) := \frac{1}{n} \sum_{j=1}^n \frac{1 - R_j}{1 - p(X_j; \gamma)} \frac{p(X_j; \gamma)}{\widehat{p}_0} G(Z_j; \theta)$ in cases [2]-cu and [2]-pu where $\widehat{p}_0 := n_p/n$.

Remarks:

(i) The last $d \times d_\gamma$ estimating equations in (28) for $\lambda_1, \dots, \lambda_d$ respectively are easy to follow at γ^0 and at given θ and β . Under standard regularity conditions, the estimated $\lambda_j(\theta, \beta)$ converges in probability to $\left\{ E \left[\frac{1-p_0(X)}{1-p(X;\gamma^0)} \frac{p'_\gamma(X;\gamma^0)p_\gamma(X;\gamma^0)}{p(X;\gamma^0)(1-p(X;\gamma^0))} \right] \right\}^{-1} E \left[\frac{1-p_0(X)}{1-p(X;\gamma^0)} \frac{p'_\gamma(X;\gamma^0)}{1-p(X;\gamma^0)} c(X; \gamma^0) \Delta_{q,d}(\theta, \beta) \right]$ which, when considered under PPM-(PS) and PPM-(CE), conforms with the intuition in the last subsection.

(ii) Under PPM-(PS) and suitable dominance conditions, RHS of the first d_β estimating equations in (28) at $\gamma = \gamma^0$ converge uniformly to the population estimating functions in (26) (once the omitted p_0 and the signs are reconsidered). To see this, consider, for e.g., case [1] and assume PPM-(PS) holds so that we can safely focus at $\gamma = \gamma^0$. The estimating functions in (26), using M(4) are

$$\begin{aligned} & E \left[\frac{R - p_0(X)}{1 - p_0(X)} q'_\beta(X; \theta, \beta) G^{-1'} G^{-1} \left\{ \frac{R - p_0(X)}{1 - p_0(X)} [g(Z; \theta) - q(X; \theta, \beta)] - \Lambda \frac{R - p_0(X)}{1 - p_0(X)} \frac{p'_\gamma(X; \gamma^0)}{p_0(X)} \right\} \right] \\ = & E \left[\frac{p_0(X)}{1 - p_0(X)} q'_\beta(X; \theta, \beta) G^{-1'} G^{-1} \left\{ [g(Z; \theta) - q(X; \theta, \beta)] - \Lambda \frac{p'_\gamma(X; \gamma^0)}{p_0(X)} \right\} \right], \end{aligned}$$

which, under standard regularity conditions and because $c(X; \gamma) = 1$, is the probability limit of the first d_β estimating equations in (28). This conforms with the intuition in the last subsection. ■

Equipped with the intuition behind the updating step, now we present the modified AIPW estimators extending the original idea of Cao et al. (2009), and remark on its asymptotic properties.

The modified AIPW estimators for the three cases [1], [2]-cu and [2]-pu are defined as follows:

$$\text{Case - [1] : } \tilde{\theta}_{[1]} \text{ solves } \frac{1}{n} \sum_{i=1}^n \hat{\psi}_{[1],i}(\tilde{\theta}_{[1]}, \hat{\gamma}, \tilde{\beta}(\tilde{\theta}_{[1]}, \hat{\gamma})) = 0, \quad (29)$$

$$\text{Case - [2]-cu : } \tilde{\theta}_{[2]-\text{cu}} \text{ solves } \frac{1}{n} \sum_{i=1}^n \hat{\psi}_{[2]-\text{cu},i}(\tilde{\theta}_{[2]-\text{cu}}, \hat{\gamma}, \tilde{\beta}(\tilde{\theta}_{[2]-\text{cu}}, \hat{\gamma})) = 0, \quad (30)$$

$$\text{Case - [2]-pu : } \tilde{\theta}_{[2]-\text{pu}} \text{ solves } \frac{1}{n} \sum_{i=1}^n \hat{\psi}_{[2]-\text{pu},i}(\tilde{\theta}_{[2]-\text{pu}}, \hat{\gamma}, \tilde{\beta}(\tilde{\theta}_{[2]-\text{pu}}, \hat{\gamma})) = 0, \quad (31)$$

where $\hat{\gamma}$ is given by (21). $\hat{\theta}_{[1]}, \hat{\theta}_{[2]-\text{cu}}$ and $\hat{\theta}_{[2]-\text{pu}}$ are the AIPW estimators defined in (16), (17) and (18) respectively. $\tilde{\beta}(\hat{\theta}_{[1]}, \hat{\gamma}), \tilde{\beta}(\hat{\theta}_{[2]-\text{cu}}, \hat{\gamma}),$ and $\tilde{\beta}(\hat{\theta}_{[2]-\text{pu}}, \hat{\gamma})$ are solutions for β from (28) with $\gamma = \hat{\gamma}$ and, respectively, $\theta = \hat{\theta}_{[1]}, \theta = \hat{\theta}_{[2]-\text{cu}}$ and $\theta = \hat{\theta}_{[2]-\text{pu}}$ for cases [1], [2]-cu and [2]-pu.

Remarks:

(i) The modified AIPW estimator is in essence obtained in two steps. The first step obtains the

AIPW estimator of θ^0 (along with $\hat{\gamma}$). The second step revises the estimator of β based on (28) by using $\hat{\gamma}$ and the first step AIPW estimator; and finally obtains the modified AIPW estimators of θ^0 by solving for θ from (29) (or (30) or (31), depending on the case) after plugging in $\hat{\gamma}$ and the revised estimator of β . Given the computational burden, such a two step method is likely to be more useful rather than simultaneously solving for θ and β from (28) and (29) (or (30) or (31)). It also seems natural since one may be inclined to report both the AIPW and the modified AIPW estimators.

(ii) In the spirit of a one-step Newton update of the AIPW estimators $(\hat{\theta})$ in (16), (17) and (18), we can alternatively express the modified AIPW estimators $(\tilde{\theta})$ in (29), (30) and (31) respectively in asymptotically equivalent forms. A generic form of the update for all three cases is

$$\tilde{\theta} = \hat{\theta} - \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta'} \hat{\psi}_i \left(\hat{\theta}, \hat{\gamma}, \tilde{\beta}(\hat{\theta}, \hat{\gamma}) \right) \right]^{-1} \frac{1}{n} \sum_{i=1}^n \hat{\psi}_i \left(\hat{\theta}, \hat{\gamma}, \tilde{\beta}(\hat{\theta}, \hat{\gamma}) \right). \quad (32)$$

To get a heuristic idea of the essential feature of this one-step Newton update note that (29) (or (30) or (31)) implies that for some mean-value $\bar{\beta}$ (element-by-element) the following relationship holds:

$$\begin{aligned} \sqrt{n} (\tilde{\theta} - \hat{\theta}) &= - \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta'} \hat{\psi}_i \left(\hat{\theta}, \hat{\gamma}, \bar{\beta}(\hat{\theta}, \hat{\gamma}) \right) \right]^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\psi}_i \left(\hat{\theta}, \hat{\gamma}, \bar{\beta}(\hat{\theta}) \right) \right. \\ &\quad \left. + \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta'} \hat{\psi}_i \left(\hat{\theta}, \hat{\gamma}, \bar{\beta} \right) \right\} \sqrt{n} \left(\bar{\beta}(\hat{\theta}, \hat{\gamma}) - \bar{\beta}(\hat{\theta}) \right) \right]. \end{aligned}$$

$\sum_{i=1}^n \hat{\psi}_i \left(\hat{\theta}, \hat{\gamma}, \bar{\beta}(\hat{\theta}) \right) = 0$ by definition of the AIPW estimators. Under the generality of our setup, $\sqrt{n} \left(\bar{\beta}(\hat{\theta}, \hat{\gamma}) - \bar{\beta}(\hat{\theta}) \right) = O_P(1)$ (and, crucially, not $o_P(1)$) when PPM-(CE) holds. $\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta'} \hat{\psi}_i \left(\hat{\theta}, \hat{\gamma}, \bar{\beta} \right) \xrightarrow{P} \Psi_{\beta}(\theta^0, \beta^0(\theta^0))$ (defined in Section A of the Appendix) when PPM-(CE) holds because $\bar{\beta}(\hat{\theta}, \hat{\gamma}) \xrightarrow{P} \beta^0(\theta)$ under PPM-(CE) (see Remark(ii) following (27)) and because $\bar{\beta}(\hat{\theta}) \xrightarrow{P} \beta^0(\theta)$ always (see Lemma-3.1(b1)). But $\Psi_{\beta}(\theta^0, \beta^0(\theta^0)) = 0$ when PPM-(PS) holds. So there is asymptotically no update due to the revised estimator of β only when both PPM-(PS) and PPM-(CE) hold. Updating in (32) happens under all other scenarios.²¹

(iii) The revised estimator of β is constructed with the following goals: (a) the modified AIPW

²¹In this sense the modified AIPW estimator is not quite similar in spirit to the estimator of Tan (2011) that is asymptotically equivalent to the IPW estimator when PPM-(PS) holds, and to the AIPW estimator when PPM-(PS) does not hold. For the modified AIPW estimators there can be (a non-standard) updating even when PPM-(CE) holds (if PPM-(PS) does not hold), because $\sqrt{n} \left(\bar{\beta}(\hat{\theta}, \hat{\gamma}) - \bar{\beta}(\hat{\theta}) \right) = O_P(1)$ and NOT $o_P(1)$.

estimators should be DR, (b) the modified AIPW estimator should have an asymptotic variance that equals the SEB whenever PPM-(PS) and PPM-(CE) hold, and (c) the modified AIPW estimator should be optimal in the sense of (24) among the DR-LE-(i.e. satisfying (a) and (b))-AIPW estimators whenever PPM-(PS) holds but PPM-(CE) does not. Therefore, the modified AIPW estimator has an advantage over the AIPW estimator in case [2]-pu, which imposes PPM-(PS). In general, since both estimators are consistent and asymptotically normal, optimality in the sense of (24) is a major advantage, especially if θ is scalar valued.

(iv) The only additional theoretical considerations are related to the asymptotic properties of the revised estimator of β . Consistency would simply require some additional dominance conditions beyond those already stated.²² For this, unlike in CE(1), we can focus at θ^0 . Further uniformity with respect to θ is not required because under the relevant scenarios, the AIPW estimator used for the revision of β is already consistent. The general form of the asymptotic variance follows from Section 6 of Newey and McFadden (1994). However, given the three stated goals in Remark (iii), none of these should be a major concern once we note from (23) that the asymptotic distribution of AIPW estimators is not affected by that of the estimators of β (when PPM-(PS) holds) and γ (when PPM-(CE) holds). ■

5 Related method and possible extensions

Graham et al. (2011a) consider estimation of θ^0 in just-identified models under case [1] and propose an estimator, called the IPT estimator, based on augmenting the original moment vector. Their work is closely related to our paper and its possible extensions. So we briefly discuss their work here.

The IPT estimator is computed from an augmented system of moment restrictions obtained by using additional restrictions based on PPMs $p_0(X) = p(r(X)\gamma^0)$ and $q_0(X; \theta^0) = t(X)\beta^0$ a.s. X . where $r(X)$ and $t(X)$ are known random vectors, and $p(v)$ is an increasing function with $\lim_{v \rightarrow -\infty} p(v) = 0$ and $\lim_{v \rightarrow \infty} p(v) = 1$. The IPT estimator is consistent when either PPM is correct, and attains the SEB in (6) when both are correct. The functions $r(X)$ and $t(X)$ have their first element as 1, and the remaining elements (rows) are possibly some finite truncation of a basis function for the function space of X . Depending on the smoothness of $p_0(X)$ and $q_0(X, \theta^0)$, and the size of the sample, these

²²Throughout we have used Theorem 2 of Jenrich (1969) to establish uniform convergence in probability of the sample estimating functions to the population estimating functions, that is required for the consistency of the estimators.

may or may not be the preferred or flexible enough or parsimonious functional forms. Nevertheless, IPT is a novel and useful method for estimation with MAR data. The IPT estimator always lies in the convex-hull of the data. The AIPW and the modified AIPW estimators do not possess this nice property. The IPT estimator also exactly balances the primary and the auxiliary samples.²³ The IPW, AIPW and modified AIPW estimators can, however, be modified to do the same, for e.g., (i) by following Cao et al. (2009)'s proposal of enhanced projection or (ii) by introducing an extra parameter, say, $\bar{\gamma}$ in the PPM for $p_0(X)$ as $\bar{\gamma} + p(X; \gamma)$ and estimating $\bar{\gamma}$ and γ by QML under the restriction that the estimated propensity score lies between 0 and 1. [Model-PS allows for this.]

Returning to the general idea of augmentation for possible improvements under parametric misspecifications, we note that one can also augment the sample moment vector with additional DR moment restrictions. For e.g., in the context of case [1] (i.e., moment restrictions (1)) such augmenting moment restrictions can be based on moment vectors of the form:

$$\begin{aligned} h_i(\theta, \beta, \gamma) &= (R_i - p(X_i; \gamma))(1 - R_{i-1})[g(Z_{i-1}, \theta) - q(X_{i-1}; \beta)], \\ h_i(\theta, \beta, \gamma) &= (R_{i-1} - p(X_{i-1}; \gamma))(1 - R_i)[g(Z_i, \theta) - q(X_i; \beta)] \end{aligned}$$

for $i = 2, \dots, n$. One can also use their inverse probability weighted versions like the original moment vector. Such augmentation will never lead to inconsistency (when the original estimators were consistent) and can only decrease the asymptotic variance when any PPM is incorrect. Exploring the benefits of such moment augmentation under parametric misspecifications, relaxation of the strict-overlap in M(3), and extension to over-identified models are the topic of our future research.

6 Finite-sample behavior: Monte-Carlo experiment with MAR IV

In this section we conduct a simulation experiment based on Toy example - 3 and demonstrate the inconsistency due to ignoring missing observations when the instrumental variable is MAR, the correction for it due to the IPW estimator, and finally the refinement due to the AIPW and the modified AIPW estimators. This example of MAR instrumental variables has not received much

²³As noted in Section III of Busso et al. (2011), it also means that the IPT estimator is not computable when such balancing is not feasible. This happened in roughly 40 percent of their replication data set associated with their designs 1 and 5 (and roughly 5 percent for designs 2, 3 and 4).

attention in the literature. To our knowledge the only papers considering a similar setup but with a fundamental difference in the MAR assumption and, hence, its consequences are Mogstad and Wiswall (2011) and Abrevaya and Donald (2011).

6.1 Simulation design and the estimators

Consider a random sample $\{Y_i, X_{1i}, X_{2i}\}_{i=1}^n$ drawn from the following model:

$$X_1 = X_2\theta^0 + (u + v), \quad X_2 = Y + v,$$

where $(u, v) \sim N(0, I_2)$ are the model errors. We consider two designs that differ in the specification for the unconditional distribution of the instrument Y . In Design-I, $Y \sim N(0, 1)$ whereas in Design-II, $Y \sim \text{Bin}(1, .5)$. In both cases Y is independent of u, v . The parameter of interest is θ . Its value $\theta^0 (= -1)$ is defined by the moment restriction (1) with $g(Z; \theta) = Y(X_1 - X_2\theta)$. The full-observation IV estimator $\hat{\theta}_{\text{Full-Obsn.}} = \sum_{i=1}^n Y_i X_{1i} / \sum_{i=1}^n Y_i X_{2i}$ is consistent for θ^0 . This is the benchmark.

Now we construct the MAR-sample by drawing a random sample $\{R_i\}_{i=1}^n$ from R where

$$R|X_1, X_2 \sim \text{Bin}\left(1, p_0(X) = \frac{1}{4} + \frac{1}{\pi} \arctan(X_1^2)\right), \quad (33)$$

and then by deleting the Y_i 's corresponding to $R_i = 1$. The choice of $p_0(X)$ is influenced by the strict-overlap assumption in M(3). Being overly cautious, we use the additive factor $1/4$ instead of $1/2$ unlike the standard Cauchy distribution function.

With missing Y_i 's, the full-observation estimator, our benchmark, is no longer feasible. We consider the following parametric estimators based on the available observations as alternatives to the benchmark: the complete-case estimator $\hat{\theta}_{\text{Comp-Case}} = \sum_{i=1}^n (1 - R_i) Y_i X_{1i} / \sum_{i=1}^n (1 - R_i) Y_i X_{2i}$; the IPW, the AIPW, the modified AIPW (MAIPW) and the IPT estimators.

For the conditional expectation $q_0(X; \theta) = E[Y|X_1, X_2](X_1 - X_2\theta)$, the postulated model (PPM-CE) used is $q(X; \theta, \beta) = t(X)\beta(X_1 - X_2\theta)$ where $t(X) = [1, X_1, X_2]$. The PPM-CE is correct for Design-I.

For the propensity score $p_0(X)$, we consider three options for the postulated model: (i) the infeasible true $p_0(X)$, (ii) a correct PPM-PS $p(X; \gamma) = \frac{1}{4} + \frac{1}{\pi} \arctan(\gamma \times X_1^2)$, and (iii) two commonly

used but wrong PPM-PS $p(X; \gamma) = \exp(r(X)\gamma)/[1 + \exp(r(X)\gamma)]$ where $r(X) = [1, X_1, X_2]$ (PPM-PS-1) and $r(X) = [1, X_1, X_2, X_1^2, X_2^2, X_1X_2]$ (PPM-PS-2).²⁴ [We do not estimate the propensity score in option (i).]

A word on the computation of the IPT estimator. We greatly benefit from using the codes generously made available by Graham et al. (2011a) on the first author's website. Their code carefully incorporates the computational details described in Appendix A of Graham et al. (2011a). In particular, we use the function files IPT.LOGIT.m, IPT.LOGIT.CRIT.m and IPT.LOGIT.PHI.m. Only PPM-PS-1 and PPM-PS-2 are considered for computation of the IPT estimator.

6.2 Summary of results

Based on 10000 simulations we estimate the mean bias, absolute bias, standard deviation and interquartile range of all the estimators for sample sizes $n = 500$ and 1000 , and report them in Tables 1 (Design-1, $n = 500$), 2 (Design-1, $n = 1000$), 3 (Design-2, $n = 500$) and 4 (Design-2, $n = 1000$). More results and the Matlab code for the reported results are available from the authors.²⁵

The simulation results conform with the theoretical discussion. Ignoring the MAR instruments leads to bias in the complete-case estimator and the bias does not vanish as sample size increases. This is different from Mogstad and Wiswall (2011) and Abrevaya and Donald (2011). The bias is removed by the IPW, the AIPW and the modified AIPW estimators using the infeasible true $p_0(X)$ or the estimated correct PPM-PS.

The IPW estimator with the wrong PPM-PS-1 or the mildly wrong PPM-PS-2 is badly biased. On the other hand, the AIPW and the modified AIPW estimators, by virtue of their double-robustness property, avoid this problem whenever PPM-CE is correct. In fact, even with the mildly wrong PPM-CE (i.e., in Design-II), the AIPW and the modified AIPW estimators enjoy almost no bias.

As expected, the AIPW and the modified AIPW estimators are more precise than the IPW estimator in all cases considered here.²⁶ The only theoretical result that was not borne out by our

²⁴Parametrically estimating binary choice models based on logit or probit is unlikely to approximate the true $p_0(X)$ in (33) perfectly. Although, thicker tail makes logit preferable to probit. Hence our choice (iii). Still the tail is not thick enough and we trim the estimated $p(X; \hat{\gamma})$ at .05 (not important) and .95 to avoid instability. Column 1 of Tables 1 and 2 reports that the average percentage (of sample size) trimmed is negligible and decreases with sample size.

²⁵The IPT estimator appears in only two parametric specifications. So we discuss it separately after describing the simulation results for the other estimators that are directly related to the theoretical discussion in our paper.

²⁶The smaller standard deviation of the IPW estimator due to the use of the estimated correct PPM-PS rather than the true $p_0(X)$ is already a well known fact [see, for e.g., Wooldridge (2007) and Graham (2011)].

simulations is that the modified AIPW estimator using the infeasible true $p_0(X)$ or the estimated correct PPM-PS is actually not more precise than the AIPW estimator when PPM-CE is wrong (i.e., in Design-II). As documented in Table-5, such improvement in precision is, however, obtained when $q(X; \theta, \beta)$ directly models $q_0(X; \theta)$ following Wang and Chen (2009) (paragraph after equation 3 on page 493).²⁷ On the other hand, with the mildly wrong PPM-PS-2, the standard deviation of the modified AIPW estimator is smaller than that of the AIPW estimator. This is a useful observation because the theory in our paper is silent about the gains in precision when PPM-PS is wrong.

In all aspects – mean bias, absolute bias, standard deviation and interquartile range – the AIPW and the modified AIPW estimators are closest to the infeasible full-observation IV estimator and hence seem desirable in terms of their behavior in finite samples. The same conclusion holds in the unreported simulation results under similar data generating processes.

Now we discuss the finite-sample behavior of the IPT estimator based on the same simulation experiment. In Design-I and with PPM-PS-1, the IPT estimator is poor and is similar to the corresponding IPW estimator. With PPM-PS-2 it improves but still is slightly worse than the modified AIPW estimator. In Design-II and with PPM-PS-1, the IPT estimator is good in terms of precision but is badly biased compared to the AIPW or the modified AIPW estimators. With PPM-PS-2 the IPT estimator is better than the rest. The relatively good performance of the IPT estimator with PPM-PS-2 is due to the richness of this specification for the propensity score.

Comparative studies of the finite-sample behavior of many of these estimators under specific settings can also be found in Frolich (2004), Kang and Schafer (2007), Cao et al. (2009), Busso et al. (2009, 2011), Qin et al. (2009) and Graham et al. (2011b) (and many others). Finite-sample behavior depends strongly on $p_0(X)$, $q_0(X; \theta)$ and the PPMs used. For e.g., in each of the cited papers above, a different estimator appears to be the preferred one based on the behavior in finite samples (although either the AIPW or the modified AIPW or both were actually competitive when considered). Needless to say that it is not difficult to find a setting where the AIPW and the modified AIPW estimators perform poorly. Generally this is related to the characteristics of the propensity score function (for e.g., violation of M(3)). However, in such cases the performance of the IPW estimator is also typically not much better. In summary, based on our experience with simulations,

²⁷It was surprising to notice this difference in results simply due to the intermediate step of modeling $E[Y|X]$ that was done for the convenience of freeing the computation of the estimator of θ from the nuisance parameters β .

we agree with Busso et al. (2011)’s conclusion “At a minimum, we recommend that researchers ... present results from a variety of approaches ...”.

7 Conclusion

We consider the missing data problem and present parametric alternatives to Chen et al. (2008)’s semiparametric estimation of θ^0 defined by the moment restrictions in (1) or (2). The two parametric alternatives – the AIPW and the modified AIPW estimators – are extensions of the original estimators proposed by Robins et al. (1994) and Cao et al. (2009) (under special cases of our setup). We consider the verify-in-sample and verify-out-of-sample setup of Chen et al. (2008) and discuss the asymptotic behavior of these estimators under the MAR assumption and demonstrate their double-robustness and local efficiency properties. A small Monte-Carlo experiment illustrates the nice finite-sample behavior of these estimators in the context of a linear IV regression with MAR instrumental variable.

In theory, the AIPW estimators are not difficult to compute. They are presented in a way so that one can recursively solve a set of estimating equations for the parameters of interest θ^0 and another set of estimating equations for the nuisance parameters β . The modified AIPW estimators do not add much to the computational burden because they only involve one additional step of Newton-updating after the computation of the AIPW estimators.

We anticipate that the AIPW and the modified AIPW estimators will be useful in practice when the researcher is reasonably certain about key components of the parametric specifications, and when semiparametric approaches are deemed less attractive due to relatively small sample size or other reasons. The semiparametric estimators of Chen et al. (2008) are consistent and semiparametrically efficient under minimal assumptions on $p_0(X)$ and $q_0(X; \theta)$. Needless to say that these estimators are preferable when sample size is “relatively” large. On the other hand, the computationally (relatively) convenient AIPW and the modified AIPW estimators are based on parametric specifications for $p_0(X)$ and $q_0(X; \theta)$, which, when correct, can lead to better behavior in “relatively” small samples. Importantly, unlike many other parametric estimators, the AIPW and the modified AIPW estimators, by virtue of their double-robustness property, provide an additional guard against inconsistency due to parametric misspecification, and are also locally efficient in the absence of misspecification.

A Appendix: Notations

The notations are listed in the following order. First we collect the notations related to the PPMs Model-PS and Model-CE. Then, we list the notations related to our main result. Also, throughout, all the population-related notations involving γ are expressed at $\gamma = \gamma^0$.

$$\begin{aligned}
\Delta_p(X_i, \gamma) &:= p_0(X_i) - p(X_i; \gamma) \\
\Delta_p(X_i) &:= \Delta_p(X_i; \gamma^0) \\
p_\gamma(X_i; \gamma) &:= \frac{\partial}{\partial \gamma'} p(X_i; \gamma), \\
p_{\gamma\gamma}(X_i; \gamma) &:= \frac{\partial^2}{\partial \gamma \partial \gamma'} p(X_i; \gamma) \\
S_{\gamma,i}(\gamma) &:= \frac{\partial}{\partial \gamma} L_i(\gamma) = \frac{R_i - p(X_i; \gamma)}{p(X_i; \gamma)(1 - p(X_i; \gamma))} p'_\gamma(X_i; \gamma) \\
A^{(\gamma)} &:= -E \left[\frac{\partial}{\partial \gamma'} S_{\gamma,i}(\gamma^0) \right] \\
&= E \left[\frac{p'_\gamma(X; \gamma^0) p_\gamma(X; \gamma^0) \left(1 + \Delta_p(X) \frac{1 - 2p(X; \gamma^0)}{p(X; \gamma^0)(1 - p(X; \gamma^0))} \right) - \Delta_p(X) p_{\gamma\gamma}(X; \gamma^0)}{p(X; \gamma^0)(1 - p(X; \gamma^0))} \right] \\
B^{(\gamma)} &:= E [S_{\gamma,i}(\gamma^0) S'_{\gamma,i}(\gamma^0)] \\
&= E \left[\frac{p'_\gamma(X; \gamma^0) p_\gamma(X; \gamma^0) (p_0(X)(1 - p(X; \gamma^0)) - p(X; \gamma^0) \Delta_p(X))}{p^2(X; \gamma^0)(1 - p(X; \gamma^0))^2} \right] \\
\Delta_q(X_i, \theta, \beta) &:= q_0(X_i; \theta) - q(X_i; \theta, \beta) \\
\widehat{\Delta}_q(X_i, \theta, \beta) &:= g(Z_i; \theta) - q(X_i; \theta, \beta) \\
\Delta_q(X_i; \theta) &:= \Delta_q(X_i; \theta, \beta^0(\theta)) \\
q_\beta(X_i; \theta, \beta) &:= \frac{\partial}{\partial \beta'} q(X_i; \theta, \beta) \\
q_\theta(X_i; \theta, \beta) &:= \frac{\partial}{\partial \theta'} q(X_i; \theta, \beta) \\
q_{\beta\beta,j}(X_i; \theta, \beta) &:= \frac{\partial^2}{\partial \beta \partial \beta'} q_j(X_i; \theta, \beta), \text{ for } j = 1, \dots, d, \text{ where } q(X_i; \theta, \beta) = [q_1(\cdot), q_2(\cdot), \dots, q_d(\cdot)]', \\
q_{\theta\beta,j}(X_i; \theta, \beta) &:= \frac{\partial^2}{\partial \theta \partial \beta'} q_j(X_i; \theta, \beta) \text{ for } j = 1, \dots, d \\
A^{(\beta)}(\theta, \beta) &:= -E \left[\frac{\partial^2}{\partial \beta \partial \beta'} Q_i(\beta, \theta) \right] \\
&= E \left[(1 - p_0(X)) \left\{ q'_\beta(X; \theta, \beta) q_\beta(X; \theta, \beta) - \sum_{j=1}^d q_{\beta\beta,j}(X; \theta, \beta) \Delta_{q,j}(X; \theta, \beta) \right\} \right]
\end{aligned}$$

$$\begin{aligned}
B^{(\beta)}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \beta} Q_i(\beta, \theta) \frac{\partial}{\partial \beta'} Q_i(\beta, \theta) \right] \\
&= E \left[(1 - p_0(X)) q'_\beta(X; \theta, \beta) \{ V(g(Z; \theta) | X) + \Delta_q(X; \theta, \beta) \Delta'_q(X; \theta, \beta) \} q_\beta(X; \theta, \beta) \right] \\
C^{(\beta)}(\theta, \beta) &:= E \left[\frac{\partial^2}{\partial \beta \partial \theta'} Q_i(\beta, \theta) \right] \\
&= \sum_{j=1}^d E \left[(1 - p_0(X)) q_{\beta\theta, j}(X; \theta, \beta) \Delta_{q, j}(X; \theta, \beta) \right] \\
&\quad + E \left[(1 - p_0(X)) q'_\beta(X; \theta, \beta) (E[G(Z; \theta) | X] - q_\theta(X; \theta, \beta)) \right] \\
J^{(\beta)}(\theta, \beta) &:= \Psi_\theta(\theta, \beta) + \Psi_\beta(\theta, \beta) \left[A^{(\beta)}(\theta, \beta) \right]^{-1} C^{(\beta)}(\theta, \beta),
\end{aligned}$$

where the expressions for $\Psi_\theta(\theta, \beta)$, $\Psi_\beta(\theta, \beta)$, and related quantities for all 3 cases are given below.

Expressions for case [2]-pu impose $p_0(X) = p(X; \gamma^0)$, i.e., $\Delta_p(X) = 0$ a.s. X .

$$\begin{aligned}
\Psi_{\theta, [1]}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \theta'} \psi_{[1], i}(\theta, \gamma^0, \beta) \right] = E \left[\frac{1 - p_0(X)}{1 - p(X; \gamma^0)} G(Z; \theta) + \frac{\Delta_p(X)}{1 - p(X; \gamma^0)} q_\theta(X; \theta, \beta) \right] \\
\Psi_{\gamma, [1]}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \gamma'} \psi_{[1], i}(\theta, \gamma^0, \beta) \right] = E \left[\frac{1 - p_0(X)}{(1 - p(X; \gamma^0))^2} \Delta_q(X; \theta, \beta) p_\gamma(X; \gamma^0) \right] \\
\Psi_{\beta, [1]}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \beta'} \psi_{[1], i}(\theta, \gamma^0, \beta) \right] = E \left[\frac{1}{1 - p(X; \gamma^0)} \Delta_p(X) q_\beta(X; \theta, \beta) \right] \\
\Psi_{\theta, [2]-cu}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \theta'} \psi_{[2]-cu, i}(\theta, \gamma^0, \beta) \right] = E \left[\frac{p(X; \gamma^0)}{p_0} \left\{ \frac{1 - p_0(X)}{1 - p(X; \gamma^0)} G(Z; \theta) + \right. \right. \\
&\quad \left. \left. + \frac{\Delta_p(X)}{1 - p(X; \gamma^0)} q_\theta(X; \theta, \beta) \right\} \right] \\
\Psi_{\gamma, [2]-cu}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \gamma'} \psi_{[2]-cu, i}(\theta, \gamma^0, \beta) \right] = \frac{1}{p_0} \Psi_{\gamma, [1]}(\theta, \beta) \\
\Psi_{\beta, [2]-cu}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \beta'} \psi_{[2]-cu, i}(\theta, \gamma^0, \beta) \right] = \frac{1}{p_0} \Psi_{\beta, [1]}(\theta, \beta) \\
\Psi_{\theta, [2]-pu}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \theta'} \psi_{[2]-pu, i}(\theta, \gamma^0, \beta) \right] = E \left[\frac{p_0(X)}{p_0} G(Z; \theta) \right] = E[G(Z; \theta) | R = 1] \\
\Psi_{\gamma, [2]-pu}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \gamma'} \psi_{[2]-pu, i}(\theta, \gamma^0, \beta) \right] = \Psi_{\gamma, [2]-cu}(\theta, \beta) + \frac{1}{p_0} E[\Delta_q(X; \theta, \beta) p_\gamma(X; \gamma^0)] \times \\
&\quad \times \left(I_{d_\gamma} + \{E[S_\gamma(\gamma^0) S'_\gamma(\gamma^0)]\}^{-1} E \left[\frac{\partial}{\partial \gamma'} S_\gamma(\gamma^0) \right] \right) \\
&= \frac{1}{p_0} E \left[\frac{1}{1 - p_0(X)} \Delta_q(X; \theta, \beta) p_\gamma(X; \gamma^0) \right] \\
\Psi_{\beta, [2]-pu}(\theta, \beta) &:= E \left[\frac{\partial}{\partial \beta'} \psi_{[2]-pu, i}(\theta, \gamma^0, \beta) \right] = \frac{1}{p_0} E \left[\frac{p(X; \gamma^0)}{1 - p(X; \gamma^0)} \Delta_p(X) q_\beta(X; \theta, \beta) \right] = 0.
\end{aligned}$$

The following is an useful decomposition of the infeasible postulated parametric versions of the efficient influence functions, defined in (13), (14) and (15), evaluated at $\gamma^0, \beta^0(\theta)$:

$$\psi_i(\theta, \gamma^0, \beta^0(\theta)) = \psi_i^{\text{inf}}(\theta) + T_i^{(1)}(\theta, \beta^0(\theta)) + T_i^{(2)}(\theta, \beta^0(\theta)) + T_i^{(3)}(\theta, \beta^0(\theta)). \quad (34)$$

For each of the three cases, the form of the function $\psi_i^{\text{inf}}(\theta)$ is defined in (3), (4) and (5) respectively. The other quantities on the RHS of (34) are listed below for each of the three cases, [1], [2]-cu and [2]-pu. The expression for case [2]-pu imposes $p_0(X) = p(X; \gamma^0)$, i.e., $\Delta_p(X) = 0$ a.s. X .

$$\begin{aligned} T_{[1],i}^{(1)}(\theta, \beta) &:= -\frac{R_i - p_0(X_i)}{1 - p(X_i; \gamma^0)} \Delta_q(X_i; \theta, \beta) \\ T_{[1],i}^{(2)}(\theta, \beta) &:= -\frac{1 - R_i}{1 - p_0(X_i)} \frac{\Delta_p(X_i)}{1 - p(X_i; \gamma^0)} \{g(Z_i; \theta) - q_0(X_i; \theta)\} \\ T_{[1],i}^{(3)}(\theta, \beta) &:= -\frac{\Delta_p(X_i)}{1 - p(X_i; \gamma^0)} \Delta_q(X_i; \theta, \beta) \\ T_{[2]-\text{cu},i}^{(1)}(\theta, \beta) &:= \frac{1}{p_0} T_{[1],i}^{(1)}(\theta) \\ T_{[2]-\text{cu},i}^{(2)}(\theta, \beta) &:= \frac{1}{p_0} T_{[1],i}^{(2)}(\theta) \\ T_{[2]-\text{cu},i}^{(3)}(\theta, \beta) &:= \frac{1}{p_0} T_{[1],i}^{(3)}(\theta) \\ T_{[2]-\text{pu},i}^{(1)}(\theta, \beta) &:= -\frac{R_i - p_0(X_i)}{1 - p_0(X_i)} \frac{p_0(X_i)}{p_0} \Delta_q(X_i; \theta, \beta) - \Pi \left(\frac{R_i - p_0(X_i)}{p_0} \Delta_q(X_i; \theta, \beta) | S_{\gamma,i}(\gamma^0) \right) \\ T_{[2]-\text{pu},i}^{(2)}(\theta, \beta) &:= 0 \\ T_{[2]-\text{pu},i}^{(3)}(\theta, \beta) &:= 0. \end{aligned}$$

The quantities $\widehat{T}^{(1)}(\theta, \beta)$ (for all the three cases [1], [2]-cu and [2]-pu) and $\widehat{\xi}(\theta, \beta)$ from Theorem-3.2(ii.b) and used throughout for the modified AIPW estimators are

$$\begin{aligned} \xi_i(\theta, \beta) &:= T_i^{(1)}(\theta, \beta) - \Pi \left(T_i^{(1)}(\theta, \beta) | S_{\gamma,i}(\gamma^0) \right) \\ \widehat{\xi}_i(\theta, \beta) &:= \widehat{T}_i^{(1)}(\theta, \beta) - \Pi \left(\widehat{T}_i^{(1)}(\theta, \beta) | S_{\gamma,i}(\gamma^0) \right) \text{ where} \\ \widehat{T}_{[1],i}^{(1)}(\theta, \beta) &:= -\frac{R_i - p_0(X_i)}{1 - p(X_i; \gamma^0)} \widehat{\Delta}_q(X_i; \theta, \beta) \\ \widehat{T}_{[2]-\text{cu},i}^{(1)}(\theta, \beta) &:= \frac{1}{p_0} \widehat{T}_{[1],i}^{(1)}(\theta) \\ \widehat{T}_{[2]-\text{pu},i}^{(1)}(\theta, \beta) &:= -\frac{R_i - p_0(X_i)}{1 - p_0(X_i)} \frac{p_0(X_i)}{p_0} \widehat{\Delta}_q(X_i; \theta, \beta) - \Pi \left(\frac{R_i - p_0(X_i)}{p_0} \widehat{\Delta}_q(X_i; \theta, \beta) | S_{\gamma,i}(\gamma^0) \right). \end{aligned}$$

B Appendix: Proofs

The proof for Lemma 3.1 is standard at least since White (1981, 1982). Hence we only give a sketch of it. However, since we need – (i) $\widehat{\beta}(\theta) \xrightarrow{P} \beta^0(\theta)$ uniformly in $\theta \in \Theta$, and (ii) the influence function representation of the estimators – for the proof of Theorem 3.2, we highlight them below.

Proof of Lemma 3.1:

(a) Reference: Theorem 2.2 (consistency) and Theorem 3.2 (asymptotic normality) of White (1982).

(a1) For each $i = 1, \dots, n$, $L_i(\gamma)$ is continuous in $\gamma \in \Gamma$ by PS(1) and M(3). Also, $|L_i(\gamma)| < |\log(1 - \kappa_*)| + |\log(\kappa^*)| < \infty$ by

(a2) $\gamma^0 \in \text{int}(\Gamma)$. So (a1) and PS(1) imply that with probability approaching 1, $\widehat{\gamma}$ is such that

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \gamma} L_i(\widehat{\gamma}) = \frac{1}{n} \sum_{i=1}^n S_{\gamma,i}(\widehat{\gamma}). \quad (35)$$

Therefore, following the same steps as in Theorem 3.2 of White (1982), and noting that – (i) the probability limit γ^0 solves the population first order conditions, $E[S_{\gamma}(\gamma)] = 0$,²⁸ (ii) $\widehat{\gamma} \in \mathcal{N}(\gamma^0)$ with probability approaching 1 – assumptions in PS directly give

$$\begin{aligned} \sqrt{n}(\widehat{\gamma} - \gamma^0) &= \left[A(\gamma) \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n S_{\gamma,i}(\gamma^0) + o_P(1), \\ &\xrightarrow{d} N\left(0, \left[A(\gamma) \right]^{-1} B(\gamma) \left[A(\gamma) \right]^{-1'}\right). \end{aligned} \quad (36)$$

(b) Reference: Theorem 2.1 (consistency) and Theorem 3.3 (asymptotic normality) of White (1981).

(b1) First note that for each $i = 1, \dots, n$, $Q_i(\theta, \beta)$ is continuous in $\beta \in \mathcal{B}(\theta)$ and $\theta \in \Theta$ by CE(1). Also $|Q_i(\theta, \beta)| = (1 - R_i) \|g(Z_i; \theta) - q(X_i; \theta, \beta)\|^2 \leq 2[b(Z_i) + b(X_i)]$ by M(2)(ii) and CE(1a). Therefore, under our assumptions, $\frac{1}{n} \sum_{i=1}^n Q_i(\theta, \beta) \xrightarrow{P} Q(\theta, \beta)$ uniformly in $\beta \in \mathcal{B}(\theta)$ and $\theta \in \Theta$ by Theorem 2 of Jenrich (1969). Second, note that Model-CE states that $\beta^0(\theta)$ is the unique minimizer of $Q(\theta, \beta)$ for any $\theta \in \Theta$. Also $Q(\theta, \beta)$ is continuous in $\beta \in \mathcal{B}(\theta)$ and $\theta \in \Theta$ by CE(1). Therefore, $\widehat{\beta}(\theta) \xrightarrow{P} \beta^0(\theta)$ uniformly in $\theta \in \Theta$.

(b2) $\beta^0(\theta) \in \text{int}(\mathcal{B}(\theta))$. So it follows under our assumptions that with probability approaching 1,

²⁸M(3) and PS(1) ensure that the conditions in Lemma 3.6 of Newey and McFadden (1994) hold.

$\widehat{\beta}(\theta)$ is such that

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta} Q_i(\theta, \widehat{\beta}(\theta)) = \frac{1}{n} \sum_{i=1}^n (1 - R_i) q'_\beta(X_i; \theta, \widehat{\beta}(\theta)) \left[g(Z_i; \theta) - q(X_i; \theta, \widehat{\beta}(\theta)) \right]. \quad (37)$$

Therefore, following the same steps as in Theorem 3.2 of White (1981), and noting that – (i) the probability limit $\beta^0(\theta)$ solves the population first order conditions, $E \left[\frac{\partial}{\partial \beta} Q_i(\theta, \beta) \right] = 0$,²⁹ (ii) $\widehat{\beta}(\theta) \in \mathcal{N}(\beta^0(\theta))$ – assumptions in CE(1) and CE(3) directly give

$$\begin{aligned} \sqrt{n} \left(\widehat{\beta}(\theta) - \beta^0(\theta) \right) &= \left[A^{(\beta)}(\theta, \beta^0(\theta)) \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \beta} Q_i(\theta, \beta^0(\theta)) + o_P(1), \\ &\xrightarrow{d} N \left(0, \left[A^{(\beta)}(\theta, \beta^0(\theta)) \right]^{-1} B^{(\beta)}(\theta, \beta^0(\theta)) \left[A^{(\beta)}(\theta, \beta^0(\theta)) \right]^{-1'} \right). \blacksquare \end{aligned} \quad (38)$$

Proof of Theorem 3.2: We write the proof omitting the subscripts “[1]”, “[2]-cu” and “[2]-pu” because the arguments are the same for all three cases. At each step of the proof we specify the forms of the generic expressions specializing for each of the three cases.

Part (i) CONSISTENCY

The arguments for the consistency proof of Z-estimators in Theorem 5.9 of van der Vaart (1998) are used. The only difference is that the sample estimating function $\frac{1}{n} \sum_{i=1}^n \widehat{\psi}_i(\theta, \widehat{\gamma}, \widehat{\beta}(\theta))$ involves estimated finite dimensional (\widehat{p}_0 , $\widehat{\gamma}$ and $\widehat{\beta}$) and infinite dimensional ($\widehat{\Pi}(\cdot|\cdot)$) nuisance parameters.

Step-1: We establish that the sample estimating function $\frac{1}{n} \sum_{i=1}^n \widehat{\psi}_i(\theta, \widehat{\gamma}, \widehat{\beta}(\theta))$ converges in probability to a population estimating function $E [\psi_i(\theta, \gamma^0, \beta^0(\theta))]$ uniformly in $\theta \in \Theta$. For this, we take the following intermediate steps.

Step-1A: A mean-value expansion of the sample estimating function gives

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \widehat{\psi}_i(\theta, \widehat{\gamma}, \widehat{\beta}(\theta)) \\ &= \frac{1}{n} \sum_{i=1}^n \psi_i(\theta, \gamma^0, \beta^0(\theta)) + \left\{ \frac{1}{n} \sum_{i=1}^n \psi_{\gamma,i}(\theta, \bar{\gamma}, \bar{\beta}(\theta)) \right\} (\widehat{\gamma} - \gamma^0) \\ &\quad + \left\{ \frac{1}{n} \sum_{i=1}^n \psi_{\beta,i}(\theta, \bar{\gamma}, \bar{\beta}(\theta)) \right\} (\widehat{\beta}(\theta) - \beta^0(\theta)) + \frac{1}{n} \sum_{i=1}^n \left[\widehat{\psi}_i(\theta, \widehat{\gamma}, \widehat{\beta}(\theta)) - \psi_i(\theta, \bar{\gamma}, \bar{\beta}(\theta)) \right], \end{aligned} \quad (39)$$

where $\bar{\gamma}$ and $\bar{\beta}(\theta)$ are mean-values, varying row by row, such that $\|\bar{\gamma} - \gamma^0\| \leq \|\widehat{\gamma} - \gamma^0\| = O_P(1/\sqrt{n})$

²⁹M(2)(ii), CE(1)(a) and CE(1)(b) ensure that the conditions in Lemma 3.6 of Newey and McFadden (1994) hold.

and $\|\bar{\beta}(\theta) - \beta^0(\theta)\| \leq \|\widehat{\beta}(\theta) - \beta^0(\theta)\| = O_P(1/\sqrt{n})$ by using Lemmas 3.1(b) and 3.1(a). In addition, by Lemma-3.1, $\bar{\gamma} \in \mathcal{N}(\gamma^0)$ and $\bar{\beta}(\theta) \in \mathcal{N}(\beta^0(\theta))$ with probability approaching 1.

Step-1B: The first term on the RHS of (39) converges in probability to $E[\psi_i(\theta, \gamma^0, \beta^0(\theta))]$ uniformly in $\theta \in \Theta$. This can be proved by verifying the conditions in Theorem 2 of Jenrich (1969) as follows. (1) Θ is compact. (2) Under our assumptions, $\psi_i(\theta, \gamma^0, \beta^0(\theta))$ is a continuous function of θ at each $z \in \mathcal{Z}$ and $R_i = \{0, 1\}$. This follows from M(2)(ii), CE(1), PS(1), and by noting M(4) and PS(2).³⁰ (3) $\|\psi_i(\theta, \gamma^0, \beta^0(\theta))\| < c(Z_i)$ where $E[c(Z)] < \infty$. Non-sharp (and needlessly complicated) dominating functions $c(Z)$ for the three cases are as follows. For case [1], $c(Z) = \sqrt{b(Z)}/(1 - \kappa^*) + \sqrt{b(X)}(2 - \kappa^*)/(1 - \kappa^*)$. For case [2]-cu, $c(Z) = 1/[\kappa_*(1 - \kappa^*)] [\kappa^* \sqrt{b(Z)} + \sqrt{b(X)}]$. For case [2]-pu, $c(Z) = (\kappa^*/\kappa_*) [\sqrt{b(Z)}/(1 - \kappa^*) + \sqrt{b(X)}(2 - \kappa^*)/(1 - \kappa^*)] + E\left(\frac{R-p(X;\gamma)}{p_0} q(X; \theta, \beta) S'_\gamma(\gamma)\right) (E[S_\gamma(\gamma) S'_\gamma(\gamma)])^{-1} \sqrt{b(X)}/2$.³¹ Hence the first term on the RHS of (39) $\xrightarrow{P} E[\psi_i(\theta, \gamma^0, \beta^0(\theta))]$ uniformly in $\theta \in \Theta$.

Step-1C: The second and third terms on the RHS of (39) are $o_P(1)$ uniformly in $\theta \in \Theta$ because of the following reasons. (1) Under our assumptions, the terms inside the curly brackets are $O_P(1)$ uniformly in $\theta \in \Theta$ because we have already noted before that the mean values are such that $\bar{\gamma} \in \mathcal{N}(\gamma^0)$ and $\bar{\beta}(\theta) \in \mathcal{N}(\beta^0(\theta))$ with probability approaching 1. (2) Lemma 3.1(a1) implies $\widehat{\gamma} - \gamma^0 = o_P(1)$ (not depending on θ). (3) Lemma 3.1(b1) implies $\widehat{\beta}(\theta) - \beta^0(\theta) = o_P(1)$ uniformly in $\theta \in \Theta$.

Step-1D: Now consider the last term on the RHS of (39). For case [1], this is identically 0. For case [2]-cu,

$$\frac{1}{n} \sum_{i=1}^n \left[\widehat{\psi}_i(\theta, \widehat{\gamma}, \widehat{\beta}(\theta)) - \psi_i(\theta, \widehat{\gamma}, \widehat{\beta}(\theta)) \right] = \left\{ \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{[2]-\text{cu}, i}(\theta, \widehat{\gamma}, \widehat{\beta}(\theta)) \right\} \left[\frac{p_0 - \widehat{p}_0}{p_0} \right], \quad (40)$$

where $\widehat{p}_0 = n_p/n$ (bounded away from 0 a.s. (R, Z)). On the RHS, the term inside the curly bracket is $O_P(1)$ uniformly in $\theta \in \Theta$ by the same logic as before after noting that $\widehat{\gamma} \in \mathcal{N}(\gamma^0)$ and $\widehat{\beta}(\theta) \in \mathcal{N}(\beta^0(\theta))$ with probability approaching 1. The term inside the square brackets on the RHS is $o_P(1)$ (not depending on θ). Therefore, the last term on the RHS of (39) is $o_P(1)$ uniformly in $\theta \in \Theta$ for case [2]-cu. For case [2]-pu, first note that for a generic $Z_{A,i}$, $\sum_{i=1}^n \Pi(Z_{A,i} | S_{\gamma,i}(\widehat{\gamma})) =$

³⁰ $E[S_\gamma(\gamma^0) S'_\gamma(\gamma^0)]$ is non-singular and continuity is preserved under inverse.

³¹ $E\left(\frac{R-p(X;\gamma)}{p_0} q(X; \theta, \beta) S'_\gamma(\gamma)\right) < \infty$ by CE (1a) and PS(1) by using Cauchy-Schwartz inequality.

$\sum_{i=1}^n \widehat{\Pi}(Z_{A,i}|S_{\gamma,i}(\widehat{\gamma})) = 0$ because $\sum_{i=1}^n S_{\gamma,i}(\widehat{\gamma}) = 0$ by definition of $\widehat{\gamma}$ in (35). Therefore, we obtain,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \left[\widehat{\psi}_i(\theta, \widehat{\gamma}, \widehat{\beta}(\theta)) - \psi_i(\theta, \widehat{\gamma}, \widehat{\beta}(\theta)) \right] &= \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1 - R_i}{1 - p(X_i; \gamma)} \frac{p(X_i; \gamma)}{\widehat{p}_0} [g(Z_i; \theta) - q(X_i; \theta, \beta)] \right. \\ &\quad \left. + \frac{p(X_i; \gamma)}{\widehat{p}_0} q(X_i; \theta; \beta) \right\} \left[\frac{p_0 - \widehat{p}_0}{p_0} \right] \\ &= \left\{ \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_{[2]-\text{pu},i}(\theta, \widehat{\gamma}, \widehat{\beta}(\theta)) \right\} \left[\frac{p_0 - \widehat{p}_0}{p_0} \right]. \end{aligned} \quad (41)$$

Similar to case [2]-cu, on the RHS, the term inside the curly bracket is $O_P(1)$ uniformly in $\theta \in \Theta$, whereas that inside the square brackets is $o_P(1)$ (not depending on θ). Therefore, the last term on the RHS of (39) is $o_P(1)$ uniformly in $\theta \in \Theta$ for case [2]-pu.

Combining Step-1A, Step-1B, Step-1C and Step-1D, we get from (39) that the sample estimating function $\frac{1}{n} \sum_{i=1}^n \widehat{\psi}_i(\theta, \widehat{\gamma}, \widehat{\beta}(\theta))$ converges in probability to the population estimating function $E[\psi_i(\theta, \gamma^0, \beta^0(\theta))]$ uniformly in $\theta \in \Theta$.

Step-2: Now we show that if either PPM-(CE) or PPM-(PS) is correct, then $E[\psi_i(\theta, \gamma^0, \beta^0(\theta))] = 0$ iff $\theta = \theta^0$. First note that $E[\psi_i(\theta, \gamma^0, \beta^0(\theta))] = E[\psi_i^{\text{inf}}(\theta)] + E[T_i^{(1)}(\theta)] + E[T_i^{(2)}(\theta)] + E[T_i^{(3)}(\theta)]$ (identity). For the three cases, $\psi^{\text{inf}}(\theta)$ is defined in (3), (4) and (5). $T_i^{(1)}(\theta)$, $T_i^{(2)}(\theta)$ and $T_i^{(3)}(\theta)$ are long expressions and are defined in Section A of the Appendix. Under assumption M, $E[T_i^{(1)}(\theta)] = E[T_i^{(2)}(\theta)] = 0$. $E[T_i^{(3)}(\theta)] = 0$ if either PPM-(CE) or PPM-(PS) is correct.³² On the other hand, $E[\psi^{\text{inf}}(\theta)] = 0$ iff $\theta = \theta^0$, and therefore, θ^0 is the unique solution of the population estimating equations. Finally note that $E[\psi^{\text{inf}}(\theta)]$ is continuous in θ (follows from M(2)(iii), M(2)(iv) and M(3)). Therefore, if either PPM-(CE) or PPM-(PS) is correct, then compactness of Θ implies that for any arbitrary $\epsilon > 0$, $\inf_{\theta: |\theta - \theta^0| \geq \epsilon} \|E[\psi_i(\theta, \gamma^0, \beta^0(\theta))]\| = \inf_{\theta: |\theta - \theta^0| \geq \epsilon} \|E[\psi^{\text{inf}}(\theta)]\| > 0$. So the second condition in Theorem 5.9 of van der Vaart (1998) is also satisfied.

Hence Step-1 and Step-2 give that $\widehat{\theta} \xrightarrow{P} \theta^0$ if either PPM-(CE) or PPM-(PS) is correct.

Part (ii) ASYMPTOTIC NORMALITY

Recall from Lemma-3.1 that $\widehat{\gamma} \in \mathcal{N}(\gamma^0)$, and for any $\theta \in \Theta$, $\widehat{\beta}(\theta) \in \mathcal{N}(\beta^0(\theta))$ with probability approaching one. From Part(i) we know that $\widehat{\theta} \in \mathcal{N}(\theta^0)$ with probability approaching one. Therefore,

³²For case [2]-pu PPM-(PS) has been imposed.

from (39), the definition of $\widehat{\theta}$ and under the maintained assumptions of the theorem, it follows that

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{\psi}_i(\widehat{\theta}, \widehat{\gamma}, \widehat{\beta}(\widehat{\theta})) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i(\theta^0, \gamma^0, \beta^0(\theta^0)) + \left[\Psi_\theta(\theta^0, \beta^0(\theta^0)) + \Psi_\beta(\theta^0, \beta^0(\theta^0)) \frac{\partial}{\partial \theta'} \beta^0(\theta^0) \right] \sqrt{n}(\widehat{\theta} - \theta^0) \\
&\quad + \Psi_\gamma(\theta^0, \beta^0(\theta^0)) \sqrt{n}(\widehat{\gamma} - \gamma^0) + \Psi_\beta(\theta^0, \beta^0(\theta^0)) \sqrt{n}(\widehat{\beta}(\widehat{\theta}) - \beta^0(\widehat{\theta})) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\widehat{\psi}_i(\widehat{\theta}, \widehat{\gamma}, \widehat{\beta}(\widehat{\theta})) - \psi_i(\widehat{\theta}, \widehat{\gamma}, \widehat{\beta}(\widehat{\theta})) \right] + o_P(1). \tag{42}
\end{aligned}$$

This is where we use CE(2). $\Psi_\theta(\theta, \beta)$, $\Psi_\beta(\theta, \beta)$ and $\Psi_\gamma(\theta, \beta)$ are long expressions and hence are defined in Section A of the Appendix for each of the three cases [1], [2]-cu and [2]-pu.

To proceed, first note that the first term on the RHS of the last line of (42) is identically zero for cases [1], [2]-cu and [2]-pu. For case [1], this is obvious. For case [2]-cu this follows from the definition of $\widehat{\theta}$ and (40). For case [2]-pu this follows from the definition of $\widehat{\theta}$, (41) and case [2]-cu.

Second, consider the last term on the second to last line of the RHS of (42). From (38), using the continuity of $A^{(\beta)}(\theta, \beta^0(\theta))$ and noting that continuity is preserved under inverse, we obtain via a mean-value expansion that

$$\begin{aligned}
\sqrt{n}(\widehat{\beta}(\widehat{\theta}) - \beta^0(\widehat{\theta})) &= \left[A^{(\beta)}(\theta^0, \beta^0(\theta^0)) \right]^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \beta} Q_i(\theta^0, \beta^0(\theta^0)) + C^{(\beta)}(\theta^0, \beta^0(\theta^0)) \sqrt{n}(\widehat{\theta} - \theta^0) \right. \\
&\quad \left. - A^{(\beta)}(\theta^0, \beta^0(\theta^0)) \frac{\partial}{\partial \theta'} \beta^0(\theta^0) \sqrt{n}(\widehat{\theta} - \theta^0) \right] + o_P(1), \\
&= \left[A^{(\beta)}(\theta^0, \beta^0(\theta^0)) \right]^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \beta} Q_i(\theta^0, \beta^0(\theta^0)) + C^{(\beta)}(\theta^0, \beta^0(\theta^0)) \sqrt{n}(\widehat{\theta} - \theta^0) \right] \\
&\quad - \frac{\partial}{\partial \theta'} \beta^0(\theta^0) \sqrt{n}(\widehat{\theta} - \theta^0) + o_P(1).
\end{aligned}$$

($A^{(\beta)}(\cdot)$ and $C^{(\beta)}(\cdot)$ are defined in Section A of the Appendix. In addition, note that here we use CE(2) again.) Therefore, from (42), and using (38) and (36), we obtain

$$\begin{aligned}
\sqrt{n}(\widehat{\theta} - \theta) &= - \left[J^{(\beta)}(\theta^0, \beta^0(\theta^0)) \right]^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i(\theta^0, \gamma^0, \beta^0(\theta^0)) + \Psi_\gamma(\theta^0, \beta^0(\theta^0)) \sqrt{n}(\widehat{\gamma} - \gamma^0) \right. \\
&\quad \left. + \Psi_\beta(\theta^0, \beta^0(\theta^0)) \sqrt{n}(\widehat{\beta}(\theta^0) - \beta^0(\theta^0)) \right] + o_P(1)
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \sqrt{n}(\hat{\theta} - \theta) &= - \left[J^{(\beta)}(\theta^0, \beta^0(\theta^0)) \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i(\theta^0, \gamma^0, \beta^0(\theta^0)) \\
&\quad - \left[J^{(\beta)}(\theta^0, \beta^0(\theta^0)) \right]^{-1} \Psi_{\gamma}(\theta^0, \beta^0(\theta^0)) \left[A^{(\gamma)} \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n S_{\gamma,i}(\gamma^0) \\
&\quad - \left[J^{(\beta)}(\theta^0, \beta^0(\theta^0)) \right]^{-1} \Psi_{\beta}(\theta^0, \beta^0(\theta^0)) \left[A^{(\beta)}(\theta^0, \beta^0(\theta^0)) \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \beta} Q_i(\theta^0, \beta^0(\theta^0)) \\
&\quad + o_P(1)
\end{aligned} \tag{43}$$

where $J^{(\beta)}(\theta, \beta) := \Psi_{\theta}(\theta, \beta) + \Psi_{\beta}(\theta, \beta) \left[A^{(\beta)}(\theta, \beta) \right]^{-1} C^{(\beta)}(\theta, \beta)$.

The influence function representation in (43) is crucial. Let us now consider the 3 scenarios – (a) both PPM-PS and PPM-CE hold, (b) PPM-PS holds but not PPM-CE, and (c) PPM-CE holds but not PPM-PS – for all the 3 cases: [1], [2]-cu and [2]-pu.

(a) When both PPM-PS and PPM-CE hold, using the expressions (please see the Notations section, i.e, Section A of the Appendix) of the involved quantities in (43) we get, for all 3 cases, that: $\psi_i(\theta^0, \gamma^0, \beta^0(\theta^0)) = \psi_i^{\text{inf}}(\theta^0)$, $\Psi_{\gamma}(\theta^0, \beta^0(\theta^0)) = 0$, $\Psi_{\beta}(\theta^0, \beta^0(\theta^0)) = 0$. In addition, $J^{(\beta)}(\theta^0, \beta^0(\theta^0)) = \Psi_{\theta}(\theta^0, \beta^0(\theta^0))$. Under scenario (a) and in case [1], $\Psi_{\theta}(\theta^0, \beta^0(\theta^0)) = E[G(Z; \theta^0)] = G_{[1]}$. Under scenario (a) and in cases [2]-cu and [2]-pu, $\Psi_{\theta}(\theta^0, \beta^0(\theta^0)) = E \left[\frac{p_0(X)}{p_0} G(Z; \theta^0) \right] = E[G(Z; \theta^0) | R = 1] = G_{[2]}$. Therefore, $\sqrt{n}(\hat{\theta} - \theta) = G^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i^{\text{inf}}(\theta^0) + o_P(1)$. Hence the mean-zero normal variable, to which $\sqrt{n}(\hat{\theta} - \theta^0)$ converges in distribution, has a variance matrix equal to the SEBs: $\text{SEB}_{[1]}$, $\text{SEB}_{[2]-\text{cu}}$ and $\text{SEB}_{[2]-\text{pu}}$ defined in (6), (7) and (8) respectively.

(b) Now consider the scenario when PPM-PS holds but PPM-CE does not. The involved quantities, for all three cases – [1], [2]-cu and [2]-pu – are: $\psi_i(\theta^0, \gamma^0, \beta^0(\theta^0)) = \psi_i^{\text{inf}}(\theta^0) + T_i^{(1)}(\theta^0, \beta^0(\theta^0))$, $\Psi_{\beta}(\theta^0, \beta^0(\theta^0)) = 0$, $\Psi_{\theta, [1]}(\theta^0, \beta^0(\theta^0)) = E[G(Z; \theta^0)] = G_{[1]}$, $\Psi_{\theta, [2]-\text{cu}}(\theta^0, \beta^0(\theta^0)) = \Psi_{\theta, [2]-\text{pu}}(\theta^0, \beta^0(\theta^0)) = E \left[\frac{p_0(X)}{p_0} G(Z; \theta^0) \right] = E[G(Z; \theta^0) | R = 1] = G_{[2]}$, $\Psi_{\gamma, [1]}(\theta^0, \beta^0(\theta^0)) = E \left[\frac{1}{1-p_0(X)} \Delta_q(X; \theta^0) p_{\gamma}(X; \gamma^0) \right]$ and $\Psi_{\gamma, [2]-\text{cu}}(\theta^0, \beta^0(\theta^0)) = \Psi_{\gamma, [2]-\text{pu}}(\theta^0, \beta^0(\theta^0)) = \frac{1}{p_0} \Psi_{\gamma, [1]}(\theta^0, \beta^0)$. Furthermore, in scenario (b) $A^{(\gamma)} = B^{(\gamma)}$. Therefore, with these specialized expressions for the three cases, (43) gives

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta^0) &= -\Psi_{\theta}^{-1}(\theta^0, \beta^0(\theta^0)) \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i^{\text{inf}}(\theta^0) + \frac{1}{\sqrt{n}} \sum_{i=1}^n T_i^{(1)}(\theta^0, \beta^0(\theta^0)) \right. \\
&\quad \left. + \Psi_{\gamma}(\theta^0, \beta^0(\theta^0)) \left[B^{(\gamma)} \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n S_{\gamma,i}(\gamma^0) \right] + o_P(1).
\end{aligned}$$

By noting that $E \left[T_i^{(1)}(\theta^0, \beta^0(\theta^0)) S'_{\gamma,i}(\gamma^0) \right] = -\Psi_\gamma(\theta^0, \beta^0(\theta^0))$ (please see the notations in Section A of the Appendix) under scenario(b), this gives

$$\sqrt{n}(\widehat{\theta} - \theta^0) = -\Psi_\theta^{-1}(\theta^0, \beta^0(\theta^0)) \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i^{\text{inf}}(\theta^0) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i(\theta^0, \beta^0(\theta^0)) \right] + o_P(1) \quad (44)$$

where, for all the three cases, $\xi_i(\theta, \beta) := T_i^{(1)}(\theta, \beta) - \Pi \left(T_i^{(1)}(\theta, \beta) | S_{\gamma,i}(\gamma^0) \right)$. For cases [1] and [2]-cu this representation is obvious. For case [2]-pu, it follows by noting that

$$\begin{aligned} T_{[2]-\text{pu},i}^{(1)}(\theta, \beta) &:= -\frac{R_i - p_0(X_i)}{1 - p_0(X_i)} \frac{p_0(X_i)}{p_0} \Delta_q(X_i; \theta, \beta) - \Pi \left(\frac{R_i - p_0(X_i)}{p_0} \Delta_q(X_i; \theta, \beta) | S_{\gamma,i}(\gamma^0) \right) \\ &= T_{[2]-\text{cu},i}^{(1)}(\theta, \beta) + \Pi^\perp \left(\frac{R_i - p_0(X_i)}{p_0} \Delta_q(X_i; \theta, \beta) | S_{\gamma,i}(\gamma^0) \right) \end{aligned} \quad (45)$$

where for any two generic Z_A and Z_B , $\Pi^\perp(Z_A | Z_B) := Z_A - \Pi(Z_A | Z_B)$, and hence the second term on the RHS of (45) is uncorrelated to $S_{\gamma,i}(\gamma^0)$.

Note that both terms inside the square bracket on the RHS of (44) converge in distribution to mean zero normal variables.³³ Furthermore, in all three cases, $\text{Cov}(\psi_i^{\text{inf}}(\theta^0), T_i^{(1)}(\theta^0, \beta^0(\theta^0))) = \text{Cov}(\psi_i^{\text{inf}}(\theta^0), S_{\gamma,i}(\gamma^0)) = 0$. Therefore, it follows directly from (44) that under scenario (b) the mean-zero normal variable, to which $\sqrt{n}(\widehat{\theta} - \theta^0)$ converges in distribution, has a variance matrix

$$\text{SEB} + \Psi_\theta^{-1}(\theta^0, \beta^0(\theta^0)) \text{Var} [\xi_i(\theta^0, \beta^0(\theta^0))] \Psi_\theta^{-1'}(\theta^0, \beta^0(\theta^0))$$

that is larger than the SEB (in a positive semi-definite sense). Finally, since $\text{Var} [\xi_i(\theta^0, \beta^0(\theta^0))] = \text{Var} [\widehat{\xi}_i(\theta^0, \beta^0(\theta^0))]$ (by the law of iterated variance), $\Psi_{\theta,[1]}(\theta^0, \beta^0(\theta^0)) = G_{[1]}$ and $\Psi_{\theta,[2]-\text{cu}}(\theta^0, \beta^0(\theta^0)) = \Psi_{\theta,[2]-\text{pu}}(\theta^0, \beta^0(\theta^0)) = G_{[2]}$ in this scenario, the asymptotic variance can be written equivalently as $\text{SEB} + G^{-1} \text{Var} [\widehat{\xi}_i(\theta^0, \beta^0(\theta^0))] G^{-1'}$.

(c) Now consider the scenario when PPM-CE holds but PPM-PS does not. Note that this scenario does not apply to case [2]-pu. So we focus on cases [1] and [2]-cu. The only simplifications of the involved quantities, for all three cases – [1], [2]-cu and [2]-pu – are: $\psi_i(\theta^0, \gamma^0, \beta^0(\theta^0)) =$

³³The mean zero part is true for $\xi_i(\theta^0, \beta^0(\theta^0))$ because both $T_i^{(1)}(\theta^0, \beta^0(\theta^0))$ and $S_{\gamma,i}(\gamma^0)$ have mean zero.

$\psi_i^{\text{inf}}(\theta^0) + T_i^{(2)}(\theta^0, \beta^0(\theta^0)), \Psi_\gamma(\theta^0, \beta^0(\theta^0)) = 0$. Therefore, (43) gives

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta^0) &= - \left[J^{(\beta)}(\theta^0, \beta^0(\theta^0)) \right]^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i^{\text{inf}}(\theta^0) + \frac{1}{\sqrt{n}} \sum_{i=1}^n T_i^{(2)}(\theta^0, \beta^0(\theta^0)) \right. \\ &\quad \left. + \Psi_\beta(\theta^0, \beta^0(\theta^0)) \left[A^{(\beta)}(\theta^0, \beta^0(\theta^0)) \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \beta} Q_i(\theta^0, \beta^0(\theta^0)) \right] + o_P(1). \end{aligned}$$

Under our assumptions, $\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i^{\text{inf}}(\theta^0)$, $\frac{1}{\sqrt{n}} \sum_{i=1}^n T_i^{(2)}(\theta^0, \beta^0(\theta^0))$ and $\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \beta} Q_i(\theta^0, \beta^0(\theta^0))$ jointly converge in distribution to mean-zero normal variables. Therefore, $\sqrt{n}(\widehat{\theta} - \theta^0)$ also converges in distribution to a mean-zero normal variable. However, the asymptotic variance is not equal to the SEB unlike in scenario (a), nor does it simplify to a meaningful form unlike in scenario (b). ■

Remark: [A simplification used for the modified AIPW estimator under case [2]-pu (45) gives an useful simplification under scenario (b) for case [2]-pu that results in

$$\begin{aligned} \xi_{[2]\text{-pu},i}(\theta, \beta) &:= T_{[2]\text{-pu},i}^{(1)}(\theta, \beta) - \Pi \left(T_{[2]\text{-pu},i}^{(1)}(\theta, \beta) | S_{\gamma,i}(\gamma^0) \right) \\ &= \Pi^\perp \left(T_{[2]\text{-pu},i}^{(1)}(\theta, \beta) | S_{\gamma,i}(\gamma^0) \right) \\ &= \Pi^\perp \left(T_{[2]\text{-cu},i}^{(1)}(\theta, \beta) + \frac{R_i - p_0(X_i)}{p_0} \Delta_q(X_i; \theta, \beta) | S_{\gamma,i}(\gamma^0) \right) \\ &= \Pi^\perp \left(\frac{R_i - p_0(X_i)}{1 - p_0(X_i)} \frac{p_0(X_i)}{p_0} \Delta_q(X_i; \theta, \beta) | S_{\gamma,i}(\gamma^0) \right) \end{aligned} \quad (46)$$

and similarly for $\widehat{\xi}_{[2]\text{-pu},i}(\theta, \beta)$.

References

- Abrevaya, J. and Donald, S. G. (2011). A GMM approach for dealing with missing data on regressors and instruments. Mimeo.
- Busso, M., DiNardo, J., and McCrary, J. (2009). Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects. Mimeo.
- Busso, M., DiNardo, J., and McCrary, J. (2011). New Evidence on the finite Sample Properties of Propensity Score Reweighting and Matching Estimators. Mimeo.
- Cao, W., Tsiatis, 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.
- Chen, X., Hong, H., and Tarozzi, A. (2004). Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects. Mimeo.
- Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric Efficiency in GMM Models with Auxiliary Data. *Annals of Statistics*, 36: 808–843.
- Frolich, M. (2004). Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators. *Review of Economics and Statistics*, 86: 77–90.
- Graham, B. S. (2011). Efficiency Bounds for Missing Data Models with Semiparametric Restrictions. *Econometrica*, 79: 437 – 452.
- Graham, B. S., Pinto, C., and Egel, D. (2011a). Inverse Probability Tilting for Moment Condition Models with Missing Data. Mimeo.
- Graham, B. S., Pinto, C., and Egel, D. (2011b). Inverse Probability Tilting for Moment Condition Models with Missing Data. Supplemental Material.
- Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66: 315–331.
- Hirano, K. and Imbens, G. (2001). Estimation of Causal Effects using Propensity Score Weighting : An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology*, 2: 259–278.
- Hirano, K., Imbens, G., and Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores. *Econometrica*, 71: 1161–1189.
- Jenrich, R. L. (1969). Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics*, 40: 633–643.
- Kang, J. and Schafer, J. (2007). Demystifying Double Robustness :A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22: 523–539.
- Mogstad, M. and Wiswall, M. (2011). Instrumental variables estimation with partially missing instruments. *Economics Letters*, <http://dx.doi.org/10.1016/j.econlet.2011.10.013>.

- Newey, W. K. and McFadden, D. L. (1994). Large Sample Estimation and Hypothesis Testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume IV, chapter 36, pages 2212–2245. Elsevier Science Publisher.
- Qin, J., Shao, J., and Zhang, B. (2008). Efficient and Doubly Robust Imputation for Covariate-Dependent Missing Responses. *Journal of the American Statistical Association*, 108: 797–810.
- Qin, J., Zhang, B., and Leung, D. (2009). Empirical-likelihood in Missing Data Problems. *Journal of the American Statistical Association*, pages 1492 – 1503.
- Robins, J. and Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of American Statistical Association*, 90: 122–129.
- Robins, M., Rotnitzky, A., and Zhao, L. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of American Statistical Association*, 427: 846–866.
- Rosenbaum, P. and Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70: 41–55.
- Rotnitzky, A. and Robins, J. (1995). Semiparametric Regression Estimation in the Presence of Dependent Censoring. *Biometrika*, 82: 805–820.
- Rubin, D. (1976). Inference and Missing Data. *Biometrika*, 63: 581–592.
- Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94: 1096–1146.
- Tan, Z. (2010). Bounded, Efficient and Doubly Robust Estimation with Inverse Weighting. *Biometrika*, 97: 661–682.
- Tan, Z. (2011). Efficient Restricted Estimators for Conditional Mean Models with Missing Data. *Biometrika*, Forthcoming.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Wang, D. and Chen, S. X. (2009). Empirical Likelihood for Estimating Equation with Missing Values. *Annals of Statistics*, 37: 490–517.
- White, H. (1981). Consequence and Detection of Misspecified Nonlinear Regression Models. *Journal of the American Statistical Association*, 76: 419–433.
- White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50: 1–25.
- Wooldridge, J. (2002). Inverse Probability Weighted M-Estimation for Sample Selection, Attrition, and Stratification. *Portuguese Economic Journal*, 1: 117–139.
- Wooldridge, J. (2007). Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics*, 141(2): 1281–1301.
- Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Tables: Results for Section 6

% trimmed	Parametric IV-Estimators	Mean Bias	Abs. Bias	Stdev	IQR
	Full-Obsn	-0.0016	0.0504	0.0632	0.0847
	Comp-Case	0.1953	0.1968	0.0870	0.1140
	PS-model: infeasible true $p_0(X)$				
	IPW	-0.0022	0.0817	0.1035	0.1357
	AIPW	-0.0017	0.0607	0.0766	0.1021
	MAIPW	-0.0018	0.0608	0.0768	0.1024
	PS-model: $p(X; \hat{\gamma})$ from true $p_0(X)$				
	IPW	-0.0012	0.0817	0.1034	0.1353
	AIPW	-0.0017	0.0607	0.0766	0.1022
	MAIPW	-0.0017	0.0613	0.0775	0.1027
	PS-model: $p(X; \hat{\gamma})$ from PPM-PS-1				
	IPW	0.1951	0.1967	0.0875	0.1145
lower tail: 0	AIPW	-0.0019	0.0596	0.0753	0.1006
upper tail: 0	MAIPW	-0.0016	0.0607	0.0768	0.1019
	IPT	0.1938	0.1955	0.0876	0.1147
	PS-model: $p(X; \hat{\gamma})$ from PPM-PS-2				
	IPW	-0.1288	0.1443	0.1291	0.1663
lower tail: 0	AIPW	-0.0014	0.0718	0.0914	0.1180
upper tail: 0.7	MAIPW	-0.0020	0.0625	0.0791	0.1039
	IPT	-0.0018	0.0628	0.0793	0.1048

Table 1: Design-I (MAR instrument $Y \sim N(0, 1)$): Sample size $n = 500$. Results are based on averages over 10000 replications in Matlab with seed 0 for random number generation.

% trimmed	Parametric IV-Estimators	Mean Bias	Abs. Bias	Stdev	IQR
	Full-Obsn	-0.0011	0.0357	0.0449	0.0601
	Comp-Case	0.1965	0.1966	0.0612	0.0816
	PS-model: infeasible true $p_0(X)$				
	IPW	-0.0016	0.0581	0.0731	0.0982
	AIPW	-0.0015	0.0431	0.0544	0.0727
	MAIPW	-0.0015	0.0432	0.0545	0.0729
	PS-model: $p(X; \hat{\gamma})$ from true $p_0(X)$				
	IPW	-0.0011	0.0581	0.0731	0.0980
	AIPW	-0.0015	0.0431	0.0544	0.0729
	MAIPW	-0.0015	0.0433	0.0547	0.0730
	PS-model: $p(X; \hat{\gamma})$ from PPM-PS-1				
	IPW	0.1963	0.1964	0.0614	0.0820
lower tail: 0	AIPW	-0.0013	0.0422	0.0533	0.0715
upper tail: 0	MAIPW	-0.0011	0.0431	0.0543	0.0734
	IPT	0.1956	0.1957	0.0615	0.0821
	PS-model: $p(X; \hat{\gamma})$ from PPM-PS-2				
	IPW	-0.1384	0.1416	0.0930	0.1227
lower tail: 0	AIPW	-0.0011	0.0515	0.0652	0.0857
upper tail: 0.6	MAIPW	-0.0014	0.0443	0.0558	0.0746
	IPT	-0.0012	0.0446	0.0562	0.0748

Table 2: Design-I (MAR instrument $Y \sim N(0,1)$): Sample size $n = 1000$. Results are based on averages over 10000 replications in Matlab with seed 0 for random number generation.

% trimmed	Parametric IV-Estimators	Mean Bias	Abs. Bias	Stdev	IQR
	Full-Obsn	-0.0048	0.0728	0.0915	0.1219
	Comp-Case	0.2438	0.2462	0.1090	0.1420
	PS-model: infeasible true $p_0(X)$				
	IPW	-0.0067	0.1123	0.1428	0.1853
	AIPW	-0.0049	0.0861	0.1090	0.1456
	MAIPW	-0.0052	0.0863	0.1091	0.1456
	PS-model: $p(X; \hat{\gamma})$ from true $p_0(X)$				
	IPW	-0.0052	0.1119	0.1423	0.1856
	AIPW	-0.0048	0.0861	0.1090	0.1455
	MAIPW	-0.0054	0.0872	0.1101	0.1466
	PS-model: $p(X; \hat{\gamma})$ from PPM-PS-1				
	IPW	0.1361	0.1487	0.1062	0.1408
lower tail: 0	AIPW	0.0026	0.0844	0.1065	0.1412
upper tail: 0	MAIPW	0.0043	0.0873	0.1110	0.1460
	IPT	0.0982	0.1207	0.1052	0.1400
	PS-model: $p(X; \hat{\gamma})$ from PPM-PS-2				
	IPW	-0.1578	0.1824	0.1715	0.2216
lower tail: 0	AIPW	0.0139	0.0939	0.1186	0.1523
upper tail: 0.5	MAIPW	0.0013	0.0866	0.1094	0.1462
	IPT	0.0047	0.0862	0.1088	0.1441

Table 3: Design-II (MAR instrument $Y \sim Bin(1, .5)$): Sample size $n = 500$. Results are based on averages over 10000 replications in Matlab with seed 0 for random number generation.

% trimmed	Parametric IV-Estimators	Mean Bias	Abs. Bias	Stdev	IQR
	Full-Obsn	-0.0026	0.0504	0.0632	0.0849
	Comp-Case	0.2465	0.2466	0.0750	0.1007
	PS-model: infeasible true $p_0(X)$				
	IPW	-0.0041	0.0776	0.0982	0.1289
	AIPW	-0.0033	0.0600	0.0754	0.1011
	MAIPW	-0.0034	0.0600	0.0755	0.1012
	PS-model: $p(X; \hat{\gamma})$ from true $p_0(X)$				
	IPW	-0.0033	0.0774	0.0980	0.1292
	AIPW	-0.0033	0.0600	0.0754	0.1011
	MAIPW	-0.0035	0.0604	0.0759	0.1013
	PS-model: $p(X; \hat{\gamma})$ from PPM-PS-1				
	IPW	0.1381	0.1405	0.0732	0.0983
lower tail: 0	AIPW	0.0041	0.0587	0.0735	0.0985
upper tail: 0	MAIPW	0.0065	0.0611	0.0759	0.1020
	IPT	0.0999	0.1066	0.0726	0.0967
	PS-model: $p(X; \hat{\gamma})$ from PPM-PS-2				
	IPW	-0.1671	0.1728	0.1199	0.1587
lower tail: 0	AIPW	0.0180	0.0671	0.0825	0.1092
upper tail: 0.4	MAIPW	0.0027	0.0604	0.0758	0.1015
	IPT	0.0086	0.0602	0.0749	0.1005

Table 4: Design-II (MAR instrument $Y \sim Bin(1, .5)$): Sample size $n = 1000$. Results are based on averages over 10000 replications in Matlab with seed 0 for random number generation.

Sample size	Parametric IV-Estimators	Mean Bias	Abs. Bias	Stdev	IQR
500	PS-model: infeasible true $p_0(X)$				
	AIPW	-0.0030	0.0943	0.1189	0.1576
	MAIPW	-0.0028	0.0932	0.1176	0.1549
	PS-model: $p(X; \hat{\gamma})$ from true $p_0(X)$				
	AIPW	-0.0022	0.0939	0.1184	0.1563
	MAIPW	-0.0034	0.0921	0.1164	0.1533
1000	PS-model: infeasible true $p_0(X)$				
	AIPW	-0.0022	0.0651	0.0820	0.1085
	MAIPW	-0.0019	0.0641	0.0809	0.1079
	PS-model: $p(X; \hat{\gamma})$ from true $p_0(X)$				
	AIPW	-0.0017	0.0648	0.0817	0.1096
	MAIPW	-0.0025	0.0634	0.0800	0.1064

Table 5: Design-II (MAR instrument $Y \sim Bin(1, .5)$). Results are based on averages over 10000 replications in Matlab with seed 0 for random number generation. Improvement in precision due to modified AIPW over AIPW when PPM-PS is correct but PPM-CE is wrong is evident here. The only thing done differently is that we use a PPM-CE that directly models $E[Y(X_1 - X_2\theta)|X]$ without the intermediate step of modeling $E[Y|X]$ first. [See the paragraph after equation 3 on page 493 in Wang and Chen (2009) for a related discussion.] The resulting estimators are worse than the corresponding AIPW and modified AIPW estimators reported in Tables 3 and 4. However, they are still better than the corresponding IPW estimators reported in Tables 3 and 4.