

Inverse probability tilting for moment condition models with missing data¹

Bryan S. Graham[†], Cristine Campos de Xavier Pinto[◇] and Daniel Egel⁺

INITIAL DRAFT: July 2006 THIS DRAFT: June 17, 2011

Abstract

We propose a new inverse probability weighting (IPW) estimator for moment condition models with missing data. Our estimator is easy to implement and compares favorably with existing IPW estimators, including augmented inverse probability weighting (AIPW) estimators, in terms of efficiency, robustness, and higher order bias. We illustrate our method with a study of the relationship between early Black-White differences in cognitive achievement and subsequent differences in adult earnings. In our dataset the early childhood achievement measure, the main regressor of interest, is missing for many units.

JEL CLASSIFICATION: C14, C21, C23, J15, J70

KEY WORDS: MISSING DATA, SEMIPARAMETRIC EFFICIENCY, DOUBLE ROBUSTNESS, (AUGMENTED) INVERSE PROBABILITY WEIGHTING (IPW), HIGHER-ORDER COMPARISONS, BLACK-WHITE GAP, CAUSAL INFERENCE, AVERAGE TREATMENT EFFECT (ATE)

¹We would like to thank David Card, Stephen Cosslett, Jinyong Hahn, Patrick Kline, Justin McCrary, Richard Smith, Tom Rothenberg, members of the Berkeley Econometrics Reading Group and especially Michael Jansson for helpful discussions. We are particularly grateful to Gary Chamberlain, Guido Imbens, Geert Ridder, Enrique Sentana and three anonymous referees for detailed comments on earlier drafts. We also acknowledge feedback and suggestions from participants in seminars at the University of Pittsburgh, Ohio State University, University of Southern California, University of California - Riverside, University of California - Davis, University of Maryland, Georgetown University, Duke University, the University of California - Berkeley, CEMFI (Madrid), Harvard University, Pontificia Universidade Católica do Rio de Janeiro and the 2009 Latin American Meetings of the Econometric Society. This is a heavily revised and extended version of *NBER Working Paper w13981* titled “Inverse probability tilting and missing data problems”. Previous versions of this paper also circulated under the title “A new method of estimating moment condition models with missing data when selection is on observables.” Material in Section 4 of the initial NBER paper is not included in this version of the paper, but may be found in the companion paper “Efficient estimation of data combination problems by the method of auxiliary-to-study tilting” (*NBER Working Paper w16928*). All the usual disclaimers apply.

[†]Department of Economics, New York University, 19 West 4th Street, 6FL, New York, NY 10012 and NBER. E-MAIL: bryan.graham@nyu.edu, WEB: <https://files.nyu.edu/bsg1/public/>.

[◇]Escola de Economia de São Paulo, FGV, Rua Itapeva 474, sala 1010, CEP: 01332-000. E-MAIL: cristinepinto@gmail.com. WEB: <http://sites.google.com/site/cristinepinto/>.

⁺RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202-5050. E-MAIL: Daniel_Egel@rand.org. WEB: <http://www.egels.org/daniel/Welcome.html>.

Missing data are ubiquitous in applied econometric research. When data are missing at random (MAR), or selection is on observables, a simple consistent procedure is to (i) reweight those units without any missing data by the inverse of the probability of selection or the propensity score, and (ii) apply standard estimation methods to this reweighted subsample (e.g., Wooldridge, 2007). Inverse probability weighting (IPW) is widely-used to address attrition in panel data (e.g., Abowd, Crépon and Kramarz, 2001), program evaluation under exogenous treatment assignment (e.g., Hirano, Imbens and Ridder, 2003), and to control biases caused by missing and/or mismeasured regressors (e.g., Robins, Rotnitzky and Zhao, 1994). Chen, Hong and Tarozzi (2004) and Wooldridge (2007) survey additional applications of IPW.

In this paper we propose a modified version of inverse probability weighting, which we call inverse probability tilting (IPT). Our procedure coincides with the IPW estimator of, for example, Wooldridge (2007), except that we replace the conditional maximum likelihood estimate (CMLE) of the propensity score with an alternative method of moments estimate. We show that if the unconditional moments used to estimate the propensity score parameter are appropriately chosen our procedure (i) is locally efficient and (ii) remains consistent even if the propensity score is misspecified. These properties, local efficiency and double robustness, which we carefully define below, are not shared by the standard IPW estimator.²

A key appeal of IPW is its conceptual and operational simplicity. Inverse probability tilting preserves this advantage, while offering improvements in terms of estimator efficiency and robustness. However other modifications of IPW exist. A leading one, which shares IPT's local efficiency and double robustness properties, is the augmented inverse probability weighting (AIPW) estimator introduced by Robins, Rotnitzky and Zhao (1994).³ We characterize the N^{-1} order asymptotic bias of IPT and a class of AIPW estimators under conditions where they are first order equivalent. We find that IPT has smaller bias than AIPW in this setting. To our knowledge these are the first higher-order comparisons in the missing data literature.

In an illustrative empirical application we revisit Johnson and Neal's (1998) analy-

²To be more specific, IPW is locally efficient at a rather peculiar data generating process (DGP). Unfortunately this DGP is difficult to interpret and a priori implausible. We discuss this point below.

³While perhaps less familiar to econometricians, although Hirano and Imbens (2001), Imbens (2004), and Wooldridge (2007) are notable exceptions, AIPW methods are widely-studied (and used) by statisticians. Tsiatis (2006) provides a book length treatment.

sis of the Black-White wage gap for young men in the United States. They find that approximately 60 percent of the Black-White gap can be predicted by group differences in cognitive skills acquired prior to labor market entry at age 18. We study the predictive value of group differences in skills acquired prior to adolescence (i.e., by age 12). We find that pre-adolescent skill differences can account for about 40 percent of the overall wage gap and two thirds of pre-market effect found by Johnson and Neal (1998).

Our analysis is complicated by the fact that a pre-adolescence test score is available for just 11 percent of respondents.⁴ In addition to being few in number, these complete cases are unrepresentative of the sample as a whole. An analysis which ignores these facts may be both inconsistent and imprecise. The IPT estimate of the wage gap conditional on the preadolescence test score corrects for the unrepresentativeness of the complete cases. The IPT point estimate is also precisely determined. Its standard error is, respectively, one third and one half, the length of the corresponding unweighted complete case and IPW ones. Our application provides a concrete example of the type of efficiency gains IPT can provide. These gains arise despite the fact that we implement IPW with a heavily overparameterized propensity score model, which theory suggests should lead to a precisely determined point estimate (Hirano, Imbens and Ridder, 2003; Wooldridge, 2007).

The next section formally defines the class of problems to which our IPT procedure applies. In Section 2 we present our estimator and characterize its large sample properties. Section 3 compares the higher order bias of IPT with that of the class of AIPW estimators introduced by Robins, Rotnitzky and Zhao (1994). Section 4 presents the empirical application. Section 5 ends with some suggestions for further research. Selected proofs are collected in the Appendix, which also includes details on computation. Additional proofs, further details on the empirical application, and a full set of Monte Carlo experiments can be found in the Supplemental Web Appendix. Software implementing our procedure is available online at <https://files.nyu.edu/bsg1/public/>.

⁴Given the severity of the missing data problem in our sample one may reasonably question the plausibility of the missing at random assumption. We emphasize that the goal of our empirical application is illustrative.

1 A semiparametric missing data model

Here we describe a general moment condition model with data missing at random (MAR). Our set-up is as in Wooldridge (2007) except that our parameter is the solution to a moment condition, as opposed to a population optimization, problem. Let $Z = (Y_1', X')'$ be a random vector, γ_0 an unknown parameter, and assume that:

Assumption 1.1 (IDENTIFICATION) *For some known $K \times 1$ vector of functions $\psi(z, \gamma)$*

$$\mathbb{E}[\psi(Z, \gamma_0)] = 0,$$

with (i) $\mathbb{E}[\psi(Z, \gamma)] \neq 0$ for all $\gamma \neq \gamma_0$, $\gamma \in \mathcal{G} \subset \mathbb{R}^K$ and \mathcal{G} compact with $\gamma_0 \in \text{int}(\mathcal{G})$, (ii) $|\psi(z, \gamma)| \leq b(z)$ for all $z \in \mathcal{Z}$ with $b(z)$ a non-negative function on \mathcal{Z} and $\mathbb{E}[b(Z)] < \infty$, (iii) $\psi(z, \gamma)$ is continuous on \mathcal{G} for each $z \in \mathcal{Z}$ and continuously differentiable in a neighborhood of γ_0 , (iv) $\mathbb{E}[\|\psi(Z, \gamma_0)\|^2] < \infty$, and (v) $\mathbb{E}[\sup_{\gamma \in \mathcal{G}} \|\nabla_{\gamma} \psi(Z, \gamma)\|] < \infty$.

Assumption 1.1 provides a standard set of conditions under which the full sample method-of-moments estimate of γ_0 , the solution to $\sum_{i=1}^N \psi(Z_i, \hat{\gamma})/N = 0$, will be consistent and asymptotically normal (cf., Newey and McFadden 1994, Theorems 2.6 and 3.4). Our interest is in identification and estimation when Y_1 is not observed for all units. Let D be a binary indicator variable. When $D = 1$ we observe Y_1 and X , while when $D = 0$ we observe only X . Our benchmark model is defined by Assumption 1.1 as well as:

Assumption 1.2 (RANDOM SAMPLING) *$\{D_i, X_i, Y_{1i}\}_{i=1}^N$ is an independently and identically distributed random sequence. We observe D , X and $Y = DY_1$ for each sampled unit.*

Assumption 1.3 (MISSING AT RANDOM) $\Pr(D = 1 | X, Y_1) = \Pr(D = 1 | X)$

Assumption 1.4 (STRONG OVERLAP) *Let $p_0(x) = \Pr(D = 1 | X = x)$, then $0 < \kappa \leq p_0(x) \leq 1$ for some $0 < \kappa < 1$ and all $x \in \mathcal{X} \subset \mathbb{R}^{\dim(X)}$.*

Assumption 1.5 (PROPENSITY SCORE MODEL) *There is a unique $\delta_0^* \in \text{int}(\mathcal{D}^*)$ with $\mathcal{D}^* \subset \mathbb{R}^{\dim(\delta^*)}$ and compact, known vector $r(X)$ of linearly independent functions of X , and known function $G(\cdot)$ such that (i) $G(\cdot)$ is strictly increasing, continuously*

differentiable and maps into the unit interval with $\lim_{v \rightarrow -\infty} G(v) = 0$ and $\lim_{v \rightarrow \infty} G(v) = 1$, (ii) $p_0(x) = G(r(x)'\delta_0^*)$ for all $x \in \mathcal{X}$, and (iii) $0 < \kappa \leq G(r(x)'\delta^*) \leq 1$ for all $\delta^* \in \mathcal{D}^*$ and $x \in \mathcal{X}$.

We refer to the model defined by Assumptions 1.1 to 1.5 as the semiparametric missing data model. Chen, Hong and Tarozzi (2008) study this model without maintaining Assumption 1.5, that is, with the propensity score left nonparametric. As is well-known, removing Assumption 1.5 from the prior restriction does not affect the asymptotic precision with which γ_0 may be estimated (Hahn, 1998). We nevertheless maintain it when deriving our local efficiency result (Theorem 2.1). Doing so is important for establishing regularity of our estimator. We also assess the properties of IPT when Assumption 1.5 fails (Theorem 2.2).

To get a sense of the range of problems to which our methods may be applied it is helpful to consider a few specific examples.

Example 1.1 (MEAN OF A VARIABLE MISSING AT RANDOM) *Let Y_1 be a binary indicator for an individual's HIV status, let $D = 1$ if an individual is tested and zero otherwise; Y_1 is logically observable only when $D = 1$. We would like to estimate the population prevalence of HIV: $\gamma_0 = \mathbb{E}[Y_1]$. This corresponds to setting $\psi(Z, \gamma) = Y_1 - \gamma$. Assumption 1.3 says that the testing decision is independent of HIV status in subpopulations homogenous in X . This may be plausible if X includes measures of risk-taking behavior and other background characteristics so that it closely approximates an individual's own information set regarding their status. Assumption 1.4 requires that at least some individuals in every subpopulation defined in terms of $X = x$ get tested. Assumption 1.5 presumes the availability of a parametric model for the testing decision. This example is closely related to that of average treatment effect (ATE) estimation under exogenous treatment assignment (see Section 5 below).*

Example 1.2 (REGRESSION FUNCTION ESTIMATION WITH MISSING REGRESSORS) *Let X_1 be a vector of demographic characteristics, X_2 log earnings, Y_1 armed forces qualification test (AFQT) score, and X_3 a vector of always observed surrogates or proxies for Y_1 (e.g., scores on subcomponents of the test, on earlier tests, etc.). Let $D = 1$ if a unit's test score is available and zero otherwise. Let $X = (X_1', X_2', X_3')'$, $\gamma = (\gamma_1', \gamma_2')'$ and $\psi(Z, \gamma) = (X_1', Y_1')'(X_2 - X_1'\gamma_1 - Y_1'\gamma_2)$. Here γ corresponds to the coefficient vector indexing the linear predictor of log earnings given demographics and*

AFQT score as in Johnson and Neal (1998). This corresponds to a linear regression model where the covariate of interest is subject to item non-response. Assumption 1.3 requires that across individuals with identical earnings (X_2), demographics (X_1), and test proxies (X_3) the probability of observing the AFQT score is independent of its value.

Other examples of the semiparametric missing data model defined by Assumptions 1.1 to 1.5 include panel data models with attrition, certain forms of censored durations and M-estimation under variable probability sampling. Chen, Hong and Tarozzi (2004) and Wooldridge (2007) survey additional examples. See also Section 5 below.

2 Inverse probability tilting

Our first result shows that standard IPW, where the propensity score is estimated by CMLE, is typically inefficient under Assumptions 1.1 to 1.5. This motivates our search for an efficient variant of IPW. The maximal asymptotic precision with which γ_0 can be estimated under these assumptions was characterized by Robins, Rotnitzky and Zhao (1994) and is given by the inverse of

$$\mathcal{I}(\gamma_0) = \Gamma_0' \Lambda_0^{-1} \Gamma_0, \quad (1)$$

with

$$\Gamma_0 = \mathbb{E} \left[\frac{\partial \psi(Z, \gamma_0)}{\partial \gamma'} \right], \quad \Lambda_0 = \mathbb{E} \left[\frac{\Sigma(X; \gamma_0)}{p_0(X)} + q(X; \gamma_0) q(X; \gamma_0)' \right], \quad (2)$$

where $\Sigma(x; \gamma) = \mathbb{V}(\psi(Z, \gamma) | X = x)$ and $q(X; \gamma) = \mathbb{E}[\psi(Z, \gamma) | X = x]$. We seek an estimator which attains this bound.

To describe the textbook IPW estimator we require some notation. Let $r_i = r(X_i)$, $\psi_i(\gamma) = \psi(Z_i, \gamma)$ and $\psi_i = \psi(Z_i, \gamma_0)$. Similarly let $G_i(\delta^*) = G(r_i' \delta^*)$ and $G_i = G(r_i' \delta_0^*)$. Denote a random unit's contribution to the score of the propensity score log-likelihood evaluated at $\delta^* = \delta_0^*$ by

$$S_{\delta^*} = \frac{D - G}{G(1 - G)} G_1 r,$$

with $G_s(v) = \partial^s G(v) / \partial v^s$ for $s = 1, 2$.⁵ Finally let $q(X_i; \gamma) = \mathbb{E}[\psi(Z_i, \gamma) | X_i]$ and

⁵To economize on notation we often omit an argument of a function when it is being evaluated

$q_i = q(X_i; \gamma_0)$. The inverse probability weighted estimate of γ_0 is given by the solution to

$$\frac{1}{N} \sum_{i=1}^N \frac{D_i \psi(Z_i, \hat{\gamma}_{IPW})}{G(r(X_i)' \hat{\delta}_{ML}^*)} = 0, \quad (3)$$

with $\hat{\delta}_{ML}^*$ the CMLE estimate of δ_0^* . Proposition 2.1 summarizes the first order asymptotic properties of $\hat{\gamma}_{IPW}$.

Proposition 2.1 (ASYMPTOTIC SAMPLING DISTRIBUTION OF $\hat{\gamma}_{IPW}$) *Suppose Assumptions 1.1 to 1.5 hold, then (i) $\sqrt{N}(\hat{\gamma}_{IPW} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \text{AVar}(\hat{\gamma}_{IPW}))$ with*

$$\begin{aligned} \text{AVar}(\hat{\gamma}_{IPW}) &= \mathcal{I}(\gamma_0)^{-1} \\ &+ \Gamma^{-1} \mathbb{E} \left[\left(\left(\frac{D}{G} - 1 \right) q - \Pi_S S_{\delta^*} \right) \left(\left(\frac{D}{G} - 1 \right) q - \Pi_S S_{\delta^*} \right)' \right] \Gamma^{-1}, \end{aligned} \quad (4)$$

for $\Pi_S = \mathbb{E} \left[\frac{D}{G} \psi S'_{\delta^*} \right] \mathbb{E} [S_{\delta^*} S'_{\delta^*}]^{-1}$ and (ii) $k' [\text{AVar}(\hat{\gamma}_{IPW}) - \mathcal{I}(\gamma_0)^{-1}] k \geq 0$ for any vector of constants k .

Proof. See the supplemental web appendix. ■

While the inefficiency of IPW, part (ii) of Proposition 2.1, is well known, the asymptotic variance expression (4) provides new insight into its large sample properties. Observe that $\Pi_S S_{\delta^*}$ equals the best (i.e., mean squared error minimizing) linear predictor of $\left(\frac{D}{G} - 1\right) q$ given S_{δ^*} .⁶ If S_{δ^*} happens to be a good predictor of $\left(\frac{D}{G} - 1\right) q$, then IPW will be nearly efficient. Consider the case where the propensity score takes a logit form so that $G(v) = \exp(v) / [1 + \exp(v)]$. Some basic calculations give $S_{\delta^*} = \left(\frac{D}{G} - 1\right) G \cdot r$; therefore if it so happens that q can be written as a linear function of $G \cdot r$, then the asymptotic variance of IPW will coincide with that of an efficient estimator. An interpretation of Hirano, Imbens and Ridder (2003) is that if the dimension of r is allowed to grow with the sample size, then q will eventually be arbitrarily well-approximated by a linear function of $G \cdot r$, so that this coincidence holds generally (i.e., for *all* data generating processes (DGPs)). Wooldridge (2007) makes a related point: (4) cannot increase if the dimension of r increases.

at the ‘truth’. For example $G_1 = G_1(r(X)' \delta_0^*) = \partial G(r(X)' \delta_0^*) / \partial v$.

⁶Note that by the conditional mean zero property of the score function and Assumption 1.3

$$\mathbb{E} \left[\left(\frac{D}{G} - 1 \right) q S'_{\delta^*} \right] = \mathbb{E} \left[\frac{D}{G} q S'_{\delta^*} \right] = \mathbb{E} \left[\frac{D}{G} \psi S'_{\delta^*} \right].$$

In practice the researcher is only able to fit a finite dimensional model for the propensity score. Proposition 2.1 indicates that, except under very special circumstances, the resulting IPW estimate of γ_0 is inefficient under Assumptions 1.1 to 1.5. Expression (4) indicates this inefficiency is most acute when $(\frac{D}{G} - 1)q$ is poorly approximated by a linear combination of S_{δ^*} , the vector of estimating equations for the propensity score parameter δ^* . This suggests that changing the estimating equations for δ^* , such that a linear combination of them closely approximates $(\frac{D}{G} - 1)q$, might improve estimator precision. This conjecture turns out to be correct. To show this result we begin by positing a working model for the conditional mean of $\psi(Z, \gamma_0)$ given X .

Assumption 2.1 (MOMENT CEF MODEL) *For some unique matrix Π_0^* and vector of linear independent functions $t^*(X)$ with a constant in the first row, we have*

$$\mathbb{E}[\psi(Z, \gamma_0)|X] = \Pi_0^* t^*(X).$$

The precise content of Assumption 2.1 depends on the form of $\psi(Z, \gamma)$. If $\psi(Z, \gamma) = Y_1 - \gamma$, as in Example 1.1, then it is equivalent to assuming that the conditional mean of Y_1 is a linear function of $t^*(X)$. Example 1.2 provides a more complicated illustration. In that case

$$\mathbb{E}[\psi(Z, \gamma_0)|X] = \begin{pmatrix} X_1 X_2 - X_1 X_1' \gamma_1 - X_1 \mathbb{E}[Y_1|X]' \gamma_2 \\ \mathbb{E}[Y_1|X] X_2 - \mathbb{E}[Y_1|X] X_1' \gamma_1 - \mathbb{E}[Y_1 Y_1'|X] \gamma_2 \end{pmatrix},$$

so that selecting $t^*(X)$ requires formulating models for the first and second conditional moments of Y_1 .⁷

When $\psi(Z, \gamma)$ is nonlinear in γ choosing $t^*(X)$ such that Assumption 2.1 holds is more difficult. In this case one can think of $t^*(X)$ as a vector of approximating functions as in the literature on nonparametric sieve estimation (e.g., Chen, 2007; see also Section 5 below). We emphasize that any approach to missing data which involves imputation also requires formulating a model for $\mathbb{E}[\psi(Z, \gamma_0)|X]$.

Let $t(X)$ denote the union of all linearly independent elements in $t^*(X)$ and $r(X)$ (recall that $r(X)$ are the functions of X entering the propensity score model

⁷To be explicit assume that $\mathbb{E}[Y_1|X] = h_1(X)' \pi_1$ and $vech(\mathbb{E}[Y_1 Y_1'|X]) = h_2(X)' \pi_2$. Let $h_3(X)$ consist of $h_1(X)$ and all non-redundant interactions between its elements and those of X_1 and X_2 , then setting $t^*(X) = (h_2(X)', h_3(X)')'$ with any redundant entries removed is sufficient for Assumption 2.1 to hold.

in Assumption 1.5). Let $1 + M$ equal the dimension of $t(X)$; this vector will include a constant and M known functions of X . Note that $t(X) = (r(X)', r^*(X)')'$ where $r^*(X)$ is the relative complement of $r(X)$ in $t^*(X)$. Letting $\delta_0 = (\delta_0^*, \eta_0)'$, where $\eta_0 = 0$, we have under Assumptions 1.1 to 1.5 the following *just-identified* unconditional moment problem

$$\mathbb{E} \left[\frac{D}{G(t(X)' \delta_0)} \psi(Z, \gamma_0) \right] = 0 \quad (5)$$

$$\mathbb{E} \left[\left(\frac{D}{G(t(X)' \delta_0)} - 1 \right) t(X) \right] = 0. \quad (6)$$

Our proposed estimator chooses $\hat{\beta}_{IPT} = (\hat{\gamma}'_{IPT}, \hat{\delta}'_{IPT})'$ to set the sample analog of (5) and (6) equal to zero:

$$\frac{1}{N} \sum_{i=1}^N \frac{D_i}{G(t(X_i)' \hat{\delta}_{IPT})} \psi(Z_i, \hat{\gamma}_{IPT}) = 0 \quad (7)$$

$$\frac{1}{N} \sum_{i=1}^N \left(\frac{D_i}{G(t(X_i)' \hat{\delta}_{IPT})} - 1 \right) t(X_i) = 0. \quad (8)$$

Several features of this estimator merit comment. First, as with the standard IPW estimator, $\hat{\gamma}_{IPT}$ is the solution to an inverse probability weighted method of moments problem (compare (7) with (3)). However, the fitted propensity score values used to construct the weights are not conditional maximum likelihood estimates. Instead $\hat{\delta}_{IPT}$ is the solution to a method of moments problem.⁸ Importantly, under Assumption 2.1, a linear combination of the estimating equations for $\hat{\delta}_{IPT}$ equals $(\frac{D}{G} - 1) q$, which Proposition 2.1 suggests might be important for efficiency.⁹

Second, if $r(X)$ is not contained within $t^*(X)$, then we add moments to the propensity score estimating equation, replacing $t^*(X)$ with $t(X)$. These additional moments do not improve the precision of $\hat{\gamma}_{IPT}$, but they do ensure that (6) contains a sufficient number of moment restrictions to pin down the propensity score parameter. Third, in the opposite case where $t^*(X)$ is not contained within $r(X)$, we enrich the propensity score model, replacing $r(X)' \delta_0^*$ with $t(X)' \delta_0$ in $G(\cdot)$. The effect of this

⁸Consequently $\hat{\delta}_{IPT}$ is an inefficient estimate of $\delta_0 = (\delta_0^*, \underline{0})'$.

⁹An earlier version of this paper derived (6) as the solution to an optimal instrumental variables problem based on the conditional moment formulation of the semiparametric missing data model studied by Graham (2011). For brevity this derivation is omitted here.

replacement is to eliminate any overidentifying restrictions. To see this note that

$$t(X)' \delta_0 = r(X)' \delta_0^* + r^*(X)' \eta_0,$$

where, by Assumption 1.5, $\eta_0 = \underline{0}$. Nevertheless including $r^*(X)$ in the propensity score model ensures that the combined dimension of (5) and (6) coincides with $\dim(\gamma_0) + \dim(\delta_0) = K + 1 + M$ so that $\beta_0 = (\gamma_0', \delta_0')'$ is just-identified. This approach to overidentification appears to be novel.¹⁰ Theorem 3.1 below shows that it results in attractive higher order properties.

An example helps to fix ideas. Let $\psi(Z, \gamma) = Y_1 - \gamma$ as in Example 1.1 with X scalar. We assume that Assumption 1.5 holds with $r(X) = (1, X)'$ so that the propensity score is, for example, logit with an index linear in X . In choosing $t^*(X)$ such that Assumption 2.1 holds we are concerned about possible nonlinearities in $\mathbb{E}[Y_1 | X = x]$, therefore we set $t^*(X) = (1, X, X^2)'$. This gives $t(X) = t^*(X)$ and $r^*(X) = X^2$. In this case we fit a propensity score model with an index that is quadratic in X despite the fact that Assumption 1.5 says that a linear one will suffice. We fit this model not by CMLE but by choosing $\widehat{\delta}_{IPT}$ to solve (8). Once we have fitted our propensity score we compute $\widehat{\gamma}_{IPT}$ by choosing it to solve (7).

Now consider the case where the analyst believes that the propensity score might vary sharply with X so that Assumption 1.5 requires $r(X) = (1, X, X^2)'$, but that $\mathbb{E}[Y_1 | X = x]$ is linear in X so that Assumption 2.1 requires only $t^*(X) = (1, X)'$. In this case $t(X) = r(X)$ and $r^*(X)$ is empty. Here the added moment serves only to tie down the propensity score parameter; it does not increase the precision of $\widehat{\gamma}_{IPT}$. There is no need to overfit the propensity score in this case.

The main difference between IPW and IPT is that the latter approach (i) overfits the propensity score if Assumption 2.1 requires us to do so and (ii) we do not use CMLE to fit the propensity score. In Appendix A we show that the first step of our procedure requires solving a globally concave programming problem with unrestricted domain. In theory this is no harder than computing the CMLE associated with a binary choice logit model and in practice we have found this step to be straightforward. The second step of our procedure, as with the standard IPW one, can be completed by any M-estimation program that is able to accept user-specified weights.

¹⁰It is similar in spirit to the introduction of ‘tilting’ parameters in the context of generalized empirical likelihood (GEL) estimation of overidentified moment condition models (e.g., Imbens, 1997). This observation is the source of inverse probability tilting’s name.

The next two theorems characterize the first order asymptotic properties of $\hat{\gamma}_{IPT}$. The first result shows that when Assumptions 1.1 to 1.5, *and* Assumption 2.1 hold, the asymptotic variance of $\hat{\gamma}_{IPT}$ equals $\mathcal{I}(\gamma_0)^{-1}$. More precisely $\hat{\gamma}_{IPT}$ is locally efficient for γ_0 in the semiparametric model defined by Assumptions 1.1 to 1.5 *at* DGPs which also satisfy Assumption 2.1.

Equation (1) is the information bound for γ_0 without imposing the additional auxiliary Assumption 2.1. This assumption imposes restrictions on the joint distribution of the data not implied by the baseline model. If these restrictions are added to the prior used to calculate the efficiency bound, then it is generally possible to estimate γ_0 more precisely. We emphasize that our estimator is not efficient with respect to this augmented model. Rather it attains the bound defined by (1) if Assumption 2.1 *happens to be true* in the population being sampled from, but *is not part of the prior restriction* used to calculate the bound. Newey (1990, p. 114), Robins, Rotnitzky and Zhao (1994, p. 852 - 3) and Tsiatis (2006) discuss the concept of local efficiency in detail. In what follows we will, for brevity, say $\hat{\gamma}_{IPT}$ is locally efficient at Assumption 2.1.

Theorem 2.1 (LOCAL SEMIPARAMETRIC EFFICIENCY OF $\hat{\gamma}_{IPT}$) *Consider the semiparametric missing data model defined by Assumptions 1.1 to 1.5, then for $\hat{\gamma}_{IPT}$ the solution to (7), (i) $\hat{\gamma}_{IPT}$ is regular and (ii) locally efficient at Assumption 2.1 with $\sqrt{N}(\hat{\gamma}_{IPT} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \mathcal{I}(\gamma_0)^{-1})$.*

Proof. See Appendix A.

Theorem 2.1 indicates that $\hat{\gamma}_{IPT}$ has good efficiency properties. By choosing the estimating equation for the propensity score with the properties of $\mathbb{E}[\psi(Z, \gamma_0) | X]$ in mind, efficiency improvements over the standard IPW estimator are possible.¹¹ ■

Our next Theorem shows that IPW has a double robustness property (cf., Bang and Robins, 2005; Tsiatis, 2006; Wooldridge, 2007). Restrictions (5) and (6) were derived under the baseline missing data model defined by Assumptions 1.1 to 1.5.

¹¹We comment that the standard IPW estimator is also locally efficient. However this occurs not at DGPs which satisfy Assumption 2.1, but rather at ones where $\mathbb{E}[\psi(Z, \gamma_0) | X]$ is linear in $r(X) \cdot G(r(X)' \delta_0^*)$. We find this condition a bit awkward from a modelling standpoint, however it does help to explain why IPW is often nearly efficient in Monte Carlo experiments where the outcome equation is a direct function of the propensity score (e.g., Busso, DiNardo, and McCrary, 2009). If the data are missing completely at random (MCAR) such that $p_0(x) = \Pr(D = 1) = Q_0$ for all $x \in \mathcal{X}$, then IPW and IPT will be locally efficient at the same DGPs as long as $r(X) = t^*(X)$.

Consequently *regardless* of whether Assumption 2.1 also holds $\widehat{\gamma}_{IPT}$ will be consistent for γ_0 and asymptotically normal.¹² This is the first part of double robustness.

Now consider a DGP where Assumptions 1.1 to 1.4 and 2.1, but not 1.5, hold. That is, a situation where the propensity score is misspecified but the implicit moment CEF model is not. In this case $\widehat{\delta} \xrightarrow{p} \delta_*$ where δ_* is the pseudo-true value which solves (6). This pseudo-true value has an interesting property. Rearranging (6) we get

$$\mathbb{E} \left[\frac{D}{G(t(X)' \delta_*)} t(X) \right] = \mathbb{E} [t(X)]. \quad (9)$$

The inverse probability weighted mean of $t(X)$ in the $D = 1$ *subpopulation* coincides with its full population mean, $\mathbb{E} [t(X)]$. This property holds *regardless* of whether the true propensity score is of the form $G(t(X)' \delta)$ for some $\delta = \delta_0$.

In the sample, rearranging (8), we get

$$\sum_{i=1}^N \widehat{\pi}_{IPT,i} t(X_i) = \frac{1}{N} \sum_{i=1}^N t(X_i), \quad \widehat{\pi}_{IPT,i} = \frac{1}{N} \frac{D_i}{G(t(X_i)' \widehat{\delta}_{IPT})}, \quad (10)$$

so that the inverse probability weighted mean of $t(X)$ in the $D = 1$ complete case *subsample* coincides with its full sample mean. By choosing the propensity score parameter to solve (8) we ensure that the estimated inverse probability weights satisfy an *exact balancing* property. For example, if $t(X_i) = (1, X, X^2)'$ with X scalar, then, after reweighting the complete case sample with $\widehat{\pi}_{IPT,i}$, the mean and variance of X will coincide with their full sample counterparts. Since the first element of $t(X_i)$ is a constant, the $\widehat{\pi}_{IPT,i}$ weights will also sum to 1.¹³

Let $F(x, y_1)$ be the joint distribution of X, Y_1 , then

$$\widehat{F}_{IPT}(x, y_1) = \sum_{i=1}^N \widehat{\pi}_{IPT,i} \mathbf{1}(X_i \leq x) \mathbf{1}(Y_{1i} \leq y_1), \quad (11)$$

is the estimate for the joint distribution of X and Y_1 implied by the IPT estimator (cf., Back and Brown, 1993; Imbens, 1997). By (10) this distribution function satisfies

¹²Its asymptotic variance, however, will lie above $\mathcal{I}(\gamma_0)^{-1}$, in the matrix sense, unless Assumption 2.1 also holds.

¹³Equation (10) highlights that the existence of $\widehat{\delta}_{IPT}$ requires that $\sum_{i=1}^N t(X_i) / N$ lie within the convex hull of the complete case subsample (a condition that is easy to check). Under Assumption 1.4 this will be true in large enough samples, but may not be in small samples; particularly when overlap is poor.

the exact balancing condition

$$\int t(x) d\widehat{F}_{IPT}(x, y_1) = \int t(x) dF_N(x), \quad (12)$$

where $F_N(x)$ is the full sample empirical distribution function of X . Since $F_N(x)$ is an efficient estimate of the distribution of X , it is reassuring that $\widehat{F}_{IPT}(x, y_1)$ satisfies (12). We discuss the properties of $\widehat{F}_{IPT}(x, y_1)$ further in Section 3.

The exact balancing property of $\widehat{F}_{IPT}(x, y_1)$ implies that $\widehat{\gamma}_{IPT}$ may be consistent for γ_0 , even if the maintained propensity score model is incorrect. Let $\Pi_0 = (\Pi_0^*, \underline{0})$, under Assumption 2.1 we have $\Pi_0 \mathbb{E}[t(X)] = \mathbb{E}[\Pi_0^* t^*(X)] = \mathbb{E}[\psi(Z, \gamma_0)]$. Using this equality, Assumption 1.3, and exact balancing (9) we get

$$\begin{aligned} \mathbb{E} \left[\frac{D\psi(Z, \gamma)}{G(t(X)' \delta_*)} \right] &= \mathbb{E} \left[\frac{p_0(X) \psi(Z, \gamma)}{G(t(X)' \delta_*)} \right] - \mathbb{E}[\psi(Z, \gamma_0)] \\ &= \mathbb{E} \left[\frac{p_0(X) \psi(Z, \gamma)}{G(t(X)' \delta_*)} \right] - \Pi_0 \mathbb{E}[t(X)] \\ &= \mathbb{E} \left[\frac{p_0(X)}{G(t(X)' \delta_*)} \psi(Z, \gamma) \right] - \Pi_0 \mathbb{E} \left[\frac{p_0(X)}{G(t(X)' \delta_*)} t(X) \right] \\ &= \mathbb{E} \left[\frac{p_0(X)}{G(t(X)' \delta_*)} \{ \mathbb{E}[\psi(Z, \gamma) | X] - \mathbb{E}[\psi(Z, \gamma_0) | X] \} \right] = 0. \quad (13) \end{aligned}$$

Therefore $\gamma = \gamma_0$ is a solution to the inverse probability weighted population moment even if there is no δ_0 such that $G(t(x)' \delta_0) = p_0(x)$ for all $x \in \mathcal{X}$. This is the second part of double robustness.

If $\psi(Z, \gamma)$ is linear in γ , as in Examples 1.1 and 1.2 above, then $\gamma = \gamma_0$ uniquely solves (13). In the general nonlinear case ensuring uniqueness of the solution to (13) may require the imposition of additional conditions, depending on the form of $\psi(Z, \gamma)$. As such conditions are model-specific we do not formulate them here, but note that doing so is facilitated by the fact that Assumption 1.4 and part (iv) of Assumption 1.5 ensure that $p_0(x)/G(t(x)' \delta_*)$ is bounded below by some positive constant.¹⁴ Proceeding under the assumption that $\gamma = \gamma_0$ uniquely solves (13) we get our second result.

¹⁴Wooldridge (2001, pp. 458 - 459) develops conditions for consistency of unweighted M-estimators when the underlying sample is a stratified random one. His argument could be adapted to the current setting for cases where $E[\psi(Z, \gamma_0)] = 0$ corresponds to the first order condition of a population optimization problem.

Theorem 2.2 (DOUBLE ROBUSTNESS OF $\hat{\gamma}_{IPT}$) *Suppose Assumptions 1.1 to 1.4, either Assumption 1.5 or 2.1, and $\gamma = \gamma_0$ uniquely solves (13), then $\sqrt{N}(\hat{\gamma}_{IPT} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \Psi_0)$, where the form of Ψ_0 depends on whether Assumption 1.5 or 2.1 holds (see Appendix A).*

Proof. See Appendix A. ■

Our formulation of the IPT estimator was undertaken with efficiency considerations at the forefront. This led to an approach where the propensity score was parameterized with two concerns in mind. First, the parametric propensity score family needs to be rich enough to contain the true score. Second, it needs to be rich enough to balance those functions of X which enter the CEF of $\psi(Z, \gamma_0)$. Theorem 2.2 shows that the dividend to this approach extends beyond local efficiency. Even if the propensity score is misspecified, IPT will remain consistent if $\mathbb{E}[\psi(Z, \gamma_0)|X]$ is linear in $t(X)$. More heuristically Theorem 2.2 suggests that IPT will perform well for moderately rich forms of $t(X)$ when *either* the propensity score or the conditional expectation of $\psi(Z, \gamma_0)$ is smooth in X . Researchers should choose $t(X)$ to be rich enough so that it accurately approximates whichever function, either $p_0(x)$ or $q_0(x) = \mathbb{E}[\psi(Z, \gamma_0)|X = x]$, is believed to be the least smooth. The double robustness properties of IPT are illustrated via a series of Monte Carlo experiments, summarized in the Supplemental Web Appendix.

3 Other alternatives to IPW and higher order comparisons

Theorems 2.1 and 2.2 provide one argument for routine use of IPT: it is (i) more robust than either standard IPW or parametric imputation and (ii) locally efficient at Assumption 2.1. Computationally it is no harder than standard IPW (see Appendix A). Finally the exact balancing property is likely to be attractive to applied researchers. It is consistent with the intuition that reweighting makes the complete case subsample more like the full sample. Tables which assess balance after IPW are commonly featured in applied work (e.g., Hirano and Imbens, 2001; see also Table 14 in the Supplemental Web Appendix).

While the argument privileging IPT over IPW appears to be straightforward, other alternatives to IPW exist. One such alternative is the class of augmented

inverse probability weighting (AIPW) estimators introduced by Robins, Rotnitzky, and Zhao (1994). Like IPT, AIPW is locally efficient at Assumption 2.1. It is also doubly robust. In this section we present two theoretical arguments for privileging our IPT method over AIPW ones. First we show that the implicit estimate of the joint distribution of X and Y_1 associated with IPT is attractive relative to the ones associated with AIPW. Second we compare the higher order bias of the two types of estimators.

3.1 A class of iterated AIPW estimators

Several versions of AIPW are now available (see Tan (2010) for a recent survey). Here we describe a general set-up which captures many of them. Let $\omega_i(\delta) = \omega(X_i, \delta)$ and $\nu_i(\delta) = \nu(D_i, X_i, \delta)$ be known, scalar-valued, nonnegative weight functions. We require that $\nu(D_i, X_i, \delta)$ is such that $\mathbb{E}[\nu(D_i, X_i, \delta) | X] = 1$. Our family of AIPW estimators will be indexed by these two weight functions. Let $\hat{\gamma}_{(\nu, \omega)}$ be an AIPW estimate in the family, which is defined as the solution to

$$\frac{1}{N} \sum_{i=1}^N \left\{ \frac{D_i}{G_i(\hat{\delta}_{ML})} \psi(Z_i, \hat{\gamma}_{(\nu, \omega)}) - \frac{\hat{q}_{(\nu, \omega)}(X_i; \hat{\gamma}_{(\nu, \omega)})}{G_i(\hat{\delta}_{ML})} (D_i - G_i(\hat{\delta}_{ML})) \right\} = 0, \quad (14)$$

with $\hat{\delta}_{ML}$ the CMLE of the propensity score parameter and

$$\hat{q}_{(\nu, \omega)}(x; \gamma) = \left[\frac{1}{N} \sum_{i=1}^N \frac{D_i}{\hat{G}_i} \hat{\omega}_i \psi_i(\gamma) t'_i \right] \times \left[\frac{1}{N} \sum_{i=1}^N \hat{\nu}_i \hat{\omega}_i t_i t'_i \right]^{-1} t(x),$$

with $\hat{G}_i = G_i(\hat{\delta}_{ML})$, $\hat{\nu}_i = \nu_i(\hat{\delta}_{ML})$ and $\hat{\omega}_i = \omega_i(\hat{\delta}_{ML})$. Note that $\hat{q}_{(\nu, \omega)}(x; \gamma)$ is the fitted value associated with a weighted least squares fit of $\psi_i(\gamma)$ onto t_i .

Setting $\nu_i(\delta) = D_i/G_i(\delta)$ and $\omega_i(\delta) = G_i(\delta)$ we get the original AIPW estimator of Robins, Rotnitzky and Zhao (1994); $\nu_i(\delta) = 1$ and $\omega_i(\delta) = 1$ yields Newey's (1994, Section 5.3) estimator, while $\nu_i(\delta) = D_i/G_i(\delta)$ and $\omega_i(\delta) = (1 - G_i(\delta))/G_i(\delta)$ gives the estimator suggested by Cao, Tsiatis and Davidian (2009) (see Table 1).¹⁵

Hirano and Imbens (2001) and Wooldridge (2007) propose a doubly robust esti-

¹⁵Many of the estimators listed in Table 1 were originally proposed in the context of a specific form for $\psi(Z, \gamma)$. We adapt to the general case as necessary. Newey (1994) derives the large sample properties of his estimator where the dimension of $t(X)$ grows with N . Here we consider his estimator with the dimension of $t(X)$ held fixed.

mator for the average treatment effect under exogenous treatment assignment.¹⁶ It turns out that setting $\nu_i(\delta) = D_i/G_i(\delta)$ and $\omega_i(\delta) = 1$ gives their estimator. In the general moment model case their approach chooses $\hat{\gamma}_{HIW}$ to solve

$$\frac{1}{N} \sum_{i=1}^N \hat{q}_{HIW}(X_i; \hat{\gamma}_{HIW}) = 0 \quad (15)$$

where $\hat{q}_{HIW}(x; \gamma)$ is the weighted least squares fit

$$\hat{q}_{HIW}(x; \gamma) = \left[\frac{1}{N} \sum_{i=1}^N \frac{D_i}{\widehat{G}_i} \psi_i(\gamma) t'_i \right] \times \left[\frac{1}{N} \sum_{i=1}^N \frac{D_i}{\widehat{G}_i} t_i t'_i \right]^{-1} t(x). \quad (16)$$

The following Proposition shows that (15) is also a member of our class of AIPW estimators.

Proposition 3.1 *The solution to (15) is numerically identical to $\hat{\gamma}_{(\nu, \omega)}$ with $\nu_i(\delta) = D_i/G_i(\delta)$ and $\omega_i(\delta) = 1$.*

Proof. Since the first element of t_i is a constant we have, by the first order condition associated with (16),

$$\frac{1}{N} \sum_{i=1}^N \frac{D_i}{\widehat{G}_i} \{\psi(Z_i, \hat{\gamma}) - \hat{q}_{HIW}(x; \gamma)\} = 0. \quad (17)$$

Adding the left-hand side of (17) to (15) and re-arranging gives the result. ■

3.2 Implicit distribution function estimates

A useful way to understand the properties of first order equivalent estimators is in terms of their implicit distribution function estimates. After some simple algebra we can show that the solution to (14) coincides with that to

$$\sum_{i=1}^N \hat{\pi}_{(v, \omega), i} \psi(Z_i, \hat{\gamma}_{(v, \omega)}) = 0,$$

where

$$\hat{\pi}_{(v, \omega), i} = \frac{1}{N} \frac{D_i}{\widehat{G}_i} \hat{\zeta}_{(v, \omega), i} \quad (18)$$

¹⁶Wooldridge's (2007) estimator is actually slightly more general than the one described here in that $\hat{q}_{HIW}(x; \gamma)$ need not correspond to a least squares fit.

with

$$\widehat{\zeta}_{(v,\omega),i} = \left\{ 1 - \left[\frac{1}{N} \sum_{i=1}^N \left(\frac{D_i}{\widehat{G}_i} - 1 \right) t'_i \right] \times \left[\frac{1}{N} \sum_{i=1}^N \widehat{\nu}_i \widehat{\omega}_i t_i t'_i \right]^{-1} \times \widehat{\omega}_i t_i \right\}, \quad (19)$$

for $i = 1, \dots, N$. This implies that the estimate of the joint distribution associated with $\widehat{\gamma}_{(v,\omega)}$ is

$$\widehat{F}_{(v,\omega)}(x, y_1) = \sum_{i=1}^N \widehat{\pi}_{(v,\omega),i} \mathbf{1}(X_i \leq x) \mathbf{1}(Y_{1i} \leq y_1), \quad (20)$$

(see Back and Brown, 1993, Proposition 1).

This distribution function has several interesting properties. First if $\nu_i = D_i/G_i(\delta)$, which is true for all the estimators listed in Table 1 except Newey's (1994), then

$$\int t(x) d\widehat{F}_{(v,\omega)}(x, y_1) = \int t(x) dF_N(x).$$

The re-weighted mean $t(X)$ in the complete case ($D = 1$) subsample coincides with its unweighted full sample mean. Since the unweighted full sample mean of $t(X)$ is an efficient estimate of its population analog, then so is the re-weighted complete case sample mean. In this sense the $\widehat{F}_{(v,\omega)}(x, y_1)$ inherits some of the efficiency properties of $F_N(x)$. Since the first element of $t(X_i)$ is 1 the AIPW distribution function estimate also integrates to 1 (i.e., $\int d\widehat{F}_{(v,\omega)}(x, y_1) = 1$).

As noted in the previous section the IPT distribution function estimate (11) also exactly balances the mean of $t(X_i)$ and integrates to one. However, it differs from $\widehat{F}_{(v,\omega)}(x, y_1)$ in that it is guaranteed to be non-decreasing, whereas $\widehat{F}_{(v,\omega)}(x, y_1)$ may be decreasing in x and/or y_1 over some ranges. Put differently some of the $\widehat{\pi}_{(v,\omega),i}$ weights may be negative, while the $\widehat{\pi}_{IPT,i}$ weights are positive by construction.

To gain further insight into this problem consider the distribution function estimator associated with standard IPW (e.g., Imbens, 2004):

$$\widehat{F}_{IPW}(x, y_1) = \sum_{i=1}^N \widehat{\pi}_{IPW,i} \mathbf{1}(X_i \leq x) \mathbf{1}(Y_{1i} \leq y_1), \quad \widehat{\pi}_{IPW,i} = \frac{1}{N} \frac{D_i}{\widehat{G}_i}. \quad (21)$$

Now consider a random sample where

$$\frac{1}{N} \sum_{i=1}^N \left(\frac{D_i}{\widehat{G}_i} - 1 \right) t(X_i) > 0 \Leftrightarrow \sum_{i=1}^N \widehat{\pi}_{IPW,i} t(X_i) > \frac{1}{N} \sum_{i=1}^N t(X_i). \quad (22)$$

In this case the IPW estimate of the mean of $t(X_i)$ exceeds its unweighted full sample counterpart. The fact that the latter mean is efficient, implies that former is not. The AIPW distribution function estimator corrects this inefficiency by adjusting the IPW weights as follows

$$\widehat{\pi}_{(v,\omega),i} = \widehat{\pi}_{IPW,i} \times \widehat{\zeta}_{(v,\omega),i},$$

with $\widehat{\zeta}_{(v,\omega),i}$ as defined in (19). Under (22) large realizations of $t(X)$ are ‘too frequent’ in the complete case subsample (even after reweighting by the inverse of the estimated propensity score). In such a situation $\widehat{\zeta}_{(v,\omega),i}$ will be less than one for $D = 1$ units with large values of $t(X)$ and greater than one for units with small values. In extreme cases the resulting $\widehat{\pi}_{(v,\omega),i}$ may be negative or exceed one. Condition (22) is especially likely to occur when the propensity score model is misspecified. In that case \widehat{G}_i corresponds to a quasi-MLE propensity score estimate and hence $\frac{1}{N} \sum_{i=1}^N \left(D_i / \widehat{G}_i - 1 \right) t(X_i)$ may differ from zero even in large samples.

In practice the IPW and AIPW distribution functions can generate nonsensical estimates. Let $\psi(Z, \gamma) = Y_1 - \gamma$. Neither $\widehat{\gamma}_{IPW}$ and $\widehat{\gamma}_{(v,\omega)}$ are guaranteed to lie within the convex hull of the data. If $Y_1 \in \{0, 1\}$, for example, this means it is possible for $\widehat{\gamma}_{IPW}$ and $\widehat{\gamma}_{(v,\omega)}$ to exceed one. In contrast $\widehat{\gamma}_{IPT}$ will lie in the convex hull of the data by construction. In our view an estimator which sets a weighted mean of $\psi(Z, \gamma)$ equal to zero, where these weights need not lie on the unit interval is *a priori* unattractive (cf., Robins, Sued, Lei-Gomez and Rotnitzky, 2007; Tan, 2010).

3.3 Higher order bias

Another way IPT and AIPW can be compared is in terms of their higher order bias. In this section we present higher order bias expressions for both IPT and AIPW when Assumptions 1.1 to 1.5 and Assumption 2.1 hold. Bias comparisons are interesting in this case because IPT and AIPW are first order equivalent. Theorem 3.1, which is based on an application of Lemma A.4 of Newey and Smith (2004), gives the result.

Theorem 3.1 (HIGHER ORDER BIAS) *Suppose Assumptions 1.1 to 1.5, Assumption 2.1, and additional regularity conditions hold, then as $N \rightarrow \infty$*

$$\widehat{\gamma}_{(v,\omega)} = \gamma_0 + \frac{C_O}{N} + \frac{C_V(v,\omega)}{N} + O(N^{-2}) \quad (23)$$

$$\widehat{\gamma}_{IPT} = \gamma_0 + \frac{C_O}{N} + O(N^{-2}) \quad (24)$$

where

$$\begin{aligned} C_O &= -\frac{1}{2} \sum_{k=1}^K \Gamma^{-1} \mathbb{E} \left[\frac{\partial^2 \psi}{\partial \gamma_k \partial \gamma'} \right] \mathcal{I}(\gamma_0)^{-1} e_k \\ &\quad + \Gamma^{-1} \mathbb{E} \left[\frac{\partial \psi}{\partial \gamma'} \Gamma^{-1} \frac{1}{p} \{\psi - q\} \right] + \frac{1}{N} \Gamma^{-1} \mathbb{E} \left[\frac{\partial \psi}{\partial \gamma'} \Gamma^{-1} q \right] \\ C_V(v,\omega) &= -\Gamma^{-1} \mathbb{E} \left[\frac{D}{p^2} \Sigma(X) \Lambda^{-1} \Pi_S S_\delta \right] \\ &\quad + \Gamma^{-1} \mathbb{E} \left[\left\{ \frac{D}{p} \left(2\omega - \frac{1}{p} \right) - \omega v \right\} qq' \Lambda^{-1} \Pi_S S_\delta \right] \\ &\quad + \Gamma^{-1} \mathbb{E} \left[\omega \left(\frac{D}{p} - \nu \right) \left(\frac{D}{p} - 1 \right) qt' F_0^{-1} t \right], \end{aligned}$$

with e_k denoting a $K \times 1$ vector with a 1 in the k^{th} row and zeros elsewhere, $p = G(t(X)' \delta_0)$, and $F_0 = \mathbb{E} \left[\frac{1-p}{p} tt' \right]$.

Proof. See Appendix A and the Supplemental Web Appendix. ■

To understand Theorem 3.1 it is helpful to consider the asymptotic properties of an infeasible ‘oracle’ estimator. This estimator chooses $\widehat{\gamma}$ to set the optimal (i.e., asymptotic variance minimizing) linear combination of the sample mean of

$$\psi^I(Z, \gamma_0) = \left\{ \begin{array}{l} \frac{D}{p_0(X)} \psi(Z, \gamma_0) \\ \left(\frac{D}{p_0(X)} - 1 \right) q_0(X) \end{array} \right\} \quad (25)$$

equal to zero. This estimator is infeasible because (i) $p_0(X)$ and $q_0(X)$ are unknown and (ii) the optimal linear combination is also unknown. An implication of Graham (2011, Theorem 2.1) is that the efficient GMM estimator based on (25) is also semi-parametrically efficient for the missing data problem defined by Assumptions 1.1 to 1.5.

A direct application of Theorem 4.1 of Newey and Smith (2004) to (25) gives an asymptotic bias for this estimator of C_O . This bias coincides with that of $\hat{\gamma}_{IPT}$, despite the fact that the oracle estimator is based on the true propensity score, $p_0(X)$, conditional mean moment vector, $q_0(X)$, and optimal GMM weight matrix. In contrast, the bias expression for the AIPW estimate $\hat{\gamma}_{(v,\omega)}$ contains additional terms. The additional terms arise from AIPW's separation of the tasks of propensity score estimation and imposition of the optimal set of balancing restrictions implied by Assumption 2.1. The first task generates no gains in terms of asymptotic precision, while at the same time introducing sampling error into the vector of estimating equations for $\hat{\gamma}_{(v,\omega)}$. The second task results in an overidentified system of moment equations. The finite sample properties of $\hat{\gamma}_{(v,\omega)}$ may degrade as a result. It is straightforward to construct stylized examples where the bias of $\hat{\gamma}_{(v,\omega)}$ increases with M , the dimension of $t(X)$, while that of $\hat{\gamma}_{IPT}$ does not. This will be especially true if the distribution of $t(X)$ is skewed and/or that of $\psi(Z, \gamma_0)$ is heteroscedastic (see the Supplemental Web Appendix for Monte Carlo examples).

The contrast between the higher order bias of $\hat{\gamma}_{IPT}$ and $\hat{\gamma}_{(v,\omega)}$ in some ways parallels that between empirical likelihood (EL) and two-step GMM for general moment condition models (Newey and Smith, 2004). The empirical likelihood estimator transforms an overidentified moment condition problem into a just-identified one by introducing a vector of tilting parameters (cf., Imbens, 1997). Our approach to overidentification, in contrast, involves overparameterizing the propensity score. The idea of overfitting a nuisance function to eliminate overidentification appears to be novel.

An alternative to IPT would be to apply GEL directly to the set of moment conditions underlying the AIPW estimator (cf., Qin, Zhang and Leung, 2009). Let $L_r = \dim(r(X))$ and $L_{t^*} = \dim(t^*(X))$. Such an approach would apply GEL to the $K + L_{t^*} + L_r$ system of moments

$$\mathbb{E} \left[\begin{array}{c} \frac{D}{G(r(X)' \delta_0^*)} \psi(Z, \gamma_0) \\ \left(\frac{D}{G(r(X)' \delta_0^*)} - 1 \right) t^*(X) \\ \left(\frac{D - G(r(X)' \delta_0^*)}{G(r(X)' \delta_0^*) [1 - G(r(X)' \delta_0^*)]} \right) G_1(r(X)' \delta_0^*) r(X) \end{array} \right] = 0.$$

Computation of $\hat{\gamma}_{GEL}$ would involve solving a saddle point problem with $2(K + L_r) + L_{t^*}$ parameters (Newey and Smith, 2004; Section 2). In contrast computing $\hat{\gamma}_{IPT}$ requires solving a $1 + M \leq L_{t^*} + L_r$ dimensional globally concave problem and a

just-identified moment condition problem with K parameters. Our approach involves a smaller parameter and sidesteps the need to solve a saddle point problem.

4 Basic skills and the Black-White wage gap

In an important pair of papers Neal and Johnson (1996) and Johnson and Neal (1998) document that Black-White skill differences present *prior* to labor market entry (i.e., by age 18) can account for a substantial portion of the corresponding gap in adult hourly earnings. In particular they find that three fifths of the raw 28 percent Black-White gap in average hourly earnings can be predicted by differences in Armed Forces Qualification Test (AFQT) scores, a measure of basic skills used by the military.

Here we repeat Johnson and Neal 's (1998) analysis after replacing AFQT scores with measures of cognitive skills acquired *prior* to adolescence. The idea is to measure how much of Black-White differences in hourly earnings can be accounted for by differences in skills across the two groups already manifest prior to adolescence. If a substantial portion of the wage gap can be so accounted for, then educational interventions which aim to ameliorate racial inequality might be more appropriately targeted toward younger children.¹⁷

We reconstruct the National Longitudinal Survey of Youth 1979 (NLSY79) extract analyzed by Johnson and Neal (1998). This sample is a stratified random sample of young men from the United States born between 1962 and 1964. Measurements of average hourly wages over the 1990 to 1993 period, race, as well as AFQT scores are available for each individual. The NLSY79 also collected data from respondents' school records. In some cases these records included (nationally normed) percentile scores on IQ tests taken at various ages. We use those scores corresponding to tests taken between the ages of 7 and 12 as measures of cognitive skills acquired prior to adolescence. Unfortunately these scores are missing for almost 90 percent of individuals. An unweighted analysis based on those individuals with complete information would be problematic for two reasons: (i) there are few complete cases making precise inference difficult and (ii) the complete cases are not representative of the full sample in terms of always-observed characteristics. Our IPT estimator is designed to address

¹⁷Interpreting any predictive relationship between early childhood test scores and subsequent labor market earnings causally involves a number of subtleties. As our purposes are primarily illustrative, we do not dwell on this issue here. See Neal and Johnson (1996) for a discussion of some of the issues involved.

both of these problems.

Columns 1 and 2 of Table 2 replicate Columns 1 and 2 of Table 14-1 in Johnson and Neal (1998, p. 483) (with the exception that we exclude Hispanics from our analysis).¹⁸ The first column reports the least squares fit of LOGWAGE onto a constant, YEAROFBIRTH, and BLACK. The estimated wage gap between Blacks and Whites of the same age is 28 percent. Column 2 adds AQFT to the set of explanatory variables. The wage gap between Blacks and Whites of the same age with the same pre-market AFQT score is only 11 percent. Seventeen percentage points of the unconditional Black-White hourly wage gap can be accounted for by average differences in pre-market AFQT scores across the two groups. That a substantial portion of racial differences in hourly wages can be accounted for by differences in skills acquired prior to entry into the labor market is Neal and Johnson's (1996) central result.

Columns 3 and 4 of Table 2 replicate Columns 1 and 2 after replacing AFQT with our preadolescence test score (EARLYTEST). This is an unweighted analysis based on the 144 complete cases. Conditioning on age alone, racial wage gaps in the complete case subsample are very similar to those computed using the full sample (Column 3). The wage gap conditional on the pre-adolescent test score is substantially lower (Column 4). Unfortunately these wage gap estimates are very imprecise; their standard errors are almost four times those of their Columns 1 and 2 counterparts. A second problem with this analysis is that those individuals with early test scores differ systematically from those without them (See the Table 11 in the Supplemental Web Appendix).

To address bias due to non-randomness in the missingness process as well as to improve precision we re-estimated the Table 2, Column 4 model using our IPT procedure. To appropriately use IPT we require that EARLYTEST is missing at random (Assumption 1.3). That is, conditional on YEAROFBIRTH, BLACK, LOGWAGE and AFQT, we require that the probability of observing EARLYTEST is independent of its value. Given the severity of missingness in our dataset this assumption is potentially problematic. We nevertheless maintain it in order to illustrate the practical application of IPT.

The joint support of YEAROFBIRTH and BLACK contains $3 \times 2 = 6$ points. We included in $t(X)$ five non-redundant dummy variables for YEAROFBIRTH-by-BLACK cell (Whites born in 1962 are the excluded group). This resulted in full

¹⁸See also Columns 1 and 3 of Table 1 in Neal and Johnson (1996, p. 875).

distributional balance for the discretely-valued components of X . We also balanced the means, variances and covariance of AFQT and LOGWAGE conditional on race alone, and age alone, but not their interaction.¹⁹ That is $t(X)$ also included AFQT, LOGWAGE, AFQT², LOGWAGE² and AFQT×LOGWAGE as well as the interactions of these variables with BLACK and the two year of birth dummies (1962 being the excluded cohort). This led to a specification of $t(X)$ with 26 elements.

Our choice of $t(X)$ was informed by two considerations. First, we wanted $t(X)$ to be rich enough to allow for complex forms of selection into missingness (see Assumption 1.5) as well as for the conditional mean and variance of EARLYTEST (see Assumption 2.1 and Example 1.2). Second, we wanted to reweight the 144 complete cases such that an analyst with access to these data alone would *numerically exactly reproduce* the results of Johnson and Neal (1998) (i.e., the point estimates in Columns 1 and 2 of Table 2).²⁰

Column 2 of Table 3 reports IPT estimates of the best linear predictor of LOGWAGE given, YEAROFBIRTH, BLACK, and EARLYTEST. For comparison the unweighted complete case estimates are reproduced in Column 1 of the table, while the standard inverse probability weighted (IPW) estimates are given in Column 3. Relative to the unweighted complete case one, the IPT estimate of the Black-White wage gap, conditional on skills acquired prior to adolescence (EARLYTEST), is larger in absolute value with a standard error almost two thirds smaller. Recall that the wage gap conditional on age alone was 28 percent (Table 2, Column 1). Conditioning on skills acquired prior to adolescence this gap falls to 18 percent. This is larger than the 11 percent gap present after conditioning on the later AFQT score, but substantially smaller than the unconditional gap. Put differently roughly 40 percent of the raw Black-White wage gap can be accounted for by differences in average skill levels across the two groups manifest prior to adolescence. This represents about two-thirds of the pre-market effect found by Neal and Johnson (1996).

Column 3 of Table 3 reports IPW estimates of the same model. The IPW estimate of the Black-White wage gap is imprecisely determined with a standard error over

¹⁹Given the near normal distribution of AFQT and LOGWAGE in our sample focusing on the first two moments of these variables seemed appropriate.

²⁰Our choice of $t(X)$ ensures that all those moments used to compute the full sample least squares fit of LOGWAGE onto a constant, YEAROFBIRTH, BLACK and AFQT are exactly balanced. Consequently the corresponding IPT-weighted least squares fit based on the 144 complete cases alone will be numerically identical to the unweighted full sample fit.

twice as large as the IPT one. This provides a concrete example of the efficiency gains IPT can provide relative to IPW (see Proposition 2.1 and Theorem 2.1). Columns 4 through 7 report estimates based on the four implementations of AIPW described in Section 3. The AIPW point estimates, with the exception of Newey’s (1994), are very similar to their IPT counterpart, albeit with standard errors about 10 percent larger.²¹

5 Summary and extensions

The IPT procedure proposed in this paper is a promising alternative to standard IPW- and AIPW-based approaches to missing data. We end by outlining some possible extensions to IPT that might merit further research.

Program evaluation and related problems Thus far we have focused on problems where Z is completely observed if $D = 1$. Now consider the case where $Z = (X', Y_0', Y_1')'$ with D , X and $Y = (1 - D)Y_0 + DY_1$ observed. That is we observe Y_0 if $D = 0$ and Y_1 if $D = 1$. Let the moment function take the separable form

$$\psi(z, \gamma) = \psi_1(y_1, x, \gamma) - \psi_0(y_0, x, \gamma).$$

Many problems fall into this basic set-up.

Example 5.1 (AVERAGE TREATMENT EFFECTS (ATEs)) *Let $D = 1$ and $D = 0$ respectively denote assignment to an active and control program or intervention and Y_1 and Y_0 the corresponding potential outcomes. The Average Treatment Effect (ATE) is*

$$\gamma_0 = \mathbb{E}[Y_1 - Y_0],$$

which corresponds to setting $\psi_1(Y_1, X, \gamma) = Y_1$ and $\psi_0(Y_0, X, \gamma) = Y_0 + \gamma$. Since each unit can only be exposed to one intervention, either Y_1 or Y_0 is missing for all units. Graham, Pinto and Egel (2011) discuss the application of IPT to this problem in detail and outline an implementation in STATA.

²¹In this particular example the implicit AIPW distribution function estimates are reasonably similar to the IPT one; AIPW does not give inordinate weight to any particular respondent and negative weight is attached to only a handful of units. The exception is Newey’s (1994) variant of AIPW. Theorem 3.1 suggests this variant of AIPW is more biased than the others, consistent with our empirical results.

Example 5.2 (TWO SAMPLE INSTRUMENTAL VARIABLES (TSIV) ESTIMATION WITH COMPATIBLE SAMPLES) *Assume that $\dim(X) \geq \dim(Y_0)$ and consider the following instrumental variables model*

$$Y_1 = Y_0' \gamma_0 + U, \quad \mathbb{E}[UX] = 0.$$

This suggests a moment function with $\psi_1(Y_1, X, \gamma) = XY_1$ and $\psi_0(Y_0, X, \gamma) = XY_0' \gamma$. Two independent random samples of size N_1 and N_0 from the same population are available. In the first sample N_1 values of Y_1 and X are recorded, while in the second N_0 values of Y_0 and X are recorded. For asymptotic analysis we assume that $\lim_{N_1, N_0 \rightarrow \infty} N_1/(N_1 + N_0) = Q_0 > 0$. This is the two-sample instrumental variables (TSIV) model analyzed by Angrist and Krueger (1992). Ridder and Moffitt (2007) provide a technical and historical overview. This model is equivalent to a special case of the semiparametric missing data model, an observation that is apparently new. Assume N units are randomly drawn from some target population. With probability Q_0 the i^{th} unit's values for Y_1 and X are recorded, while with probability $1 - Q_0$ its values of Y_0 and X are recorded. The indicator variable D denotes which set of variables are measured. The only difference between this sampling scheme and that of Angrist and Krueger (1992) is that in the latter N_1 and N_0 are fixed by the researcher, whilst in the missing data formulation they are random variables. An adaptation of the argument given by Imbens and Lancaster (1996, Sections 2.1-2.2) shows that this difference does not affect inference.

To apply IPT to these problems we find the $\hat{\delta}_{IPT}^0$, $\hat{\delta}_{IPT}^1$ and $\hat{\gamma}_{IPT}$ which solve

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \left\{ \frac{D_i \psi_1(Y_{1i}, X_i, \hat{\gamma}_{IPT})}{G(t(X_i)' \hat{\delta}_{IPT}^1)} - \frac{(1 - D_i) \psi_0(Y_{0i}, X_i, \hat{\gamma}_{IPT})}{1 - G(t(X_i)' \hat{\delta}_{IPT}^0)} \right\} &= 0 \\ \frac{1}{N} \sum_{i=1}^N \left(\frac{1 - D_i}{1 - G(t(X_i)' \hat{\delta}_{IPT}^0)} - 1 \right) t(X_i) &= 0 \\ \frac{1}{N} \sum_{i=1}^N \left(\frac{D_i}{G(t(X_i)' \hat{\delta}_{IPT}^1)} - 1 \right) t(X_i) &= 0. \end{aligned}$$

Note that this involves computing two propensity score parameter estimates. One which balances the mean of $t(X)$ in the $D = 1$ subsample with its full sample mean ($\hat{\delta}_{IPT}^1$) and one which balances the mean of $t(X)$ in the $D = 0$ subsample with its

full sample mean ($\widehat{\delta}_{IPT}^0$). Each of these propensity score estimates may be computed using the algorithm described in Appendix A. The second step of estimation involves solving a just-identified moment condition problem.

It is straightforward to extend the arguments given above to show that the above estimator is locally efficient and doubly robust. As before $t(X)$ should be rich enough to adequately model the propensity score. Local efficiency requires that $\mathbb{E}[\psi_0(Y_0, X, \gamma)|X]$ and $\mathbb{E}[\psi_1(Y_1, X, \gamma)|X]$ be linear in $t(X)$ (this is also the condition for double robustness). As in the examples outlined above the form of $t(X)$ is often suggested by the structure of the problem. Consider efficient estimation of the ATE by IPT. This requires choosing $t(X)$ such that the true propensity score is contained in the parametric family $G(t(X)'\delta)$ and the true potential outcome CEFs are linear in $t(X)$. Consistency requires only one of these two restrictions to be true.

$\mathbb{E}[\psi(Z, \gamma_0)|X]$ **nonlinear** If there is no $t(X)$ such that $\mathbb{E}[\psi(Z, \gamma_0)|X]$ is linear in $t(X)$ then neither our local efficiency or double robustness result can exactly hold (although our procedure, like IPW, will still be consistent if the propensity score is correctly specified). Although, in practice, $\mathbb{E}[\psi(Z, \gamma_0)|X]$ may be well-approximated by a function linear in $t(X)$, it is of interest to allow $\mathbb{E}[\psi(Z, \gamma_0)|X]$ to be intrinsically nonlinear. As a concrete example assume that we seek to estimate the marginal mean of the binary-valued Y_1 . We posit the working model $\Pr(Y_1 = 1|X) = F(X'\eta)$ and choose $\widehat{\eta}$ to maximize the log-likelihood

$$\sum_{i=1}^N D_i \{Y_{1i} \log F(X_i'\eta) + (1 - Y_{1i}) \log (1 - F(X_i'\eta))\}.$$

Note we use only the complete cases ($D = 1$ units) for this computation.

Observe that if $t(X)$ included $F(X'\eta_0)$ as an element, then Assumption 2.1 would hold by construction. We approximate this ideal by including the estimate $F(X'\widehat{\eta})$ as an element of $t(X)$ (along with the elements of $r(X)$ and possibly other known functions of X). Denote $t(X)$, so defined, by $t(X; \widehat{\eta}_{ML})$. Using this vector of balancing functions we estimate the propensity score parameter by solving

$$\frac{1}{N} \sum_{i=1}^N \left(\frac{D_i}{G(t(X_i; \widehat{\eta}_{ML})'\widehat{\delta}_{IPT})} - 1 \right) t(X_i; \widehat{\eta}_{ML}) = 0.$$

The IPT estimate of γ_0 is solved for as before. The main difference between the IPT

procedure introduced in Section 2 and the one sketched above is the inclusion of a generated regressor in the propensity score model. It is possible that sampling error in $\hat{\eta}_{ML}$ could affect the asymptotic properties of $\hat{\gamma}_{IPT}$. We conjecture that, appropriately restated, Theorems 2.1 and 2.2 would remain valid, but that our higher-order bias calculations would be affected.

Data dependent choice of $t(X)$ when $\mathbb{E}[\psi(Z, \gamma_0) | X]$ is nonparametric Assume the propensity score is known, but that prior knowledge on the form of $\mathbb{E}[\psi(Z, \gamma_0) | X]$ is unavailable (i.e., it is nonparametric). If the first element of $t(X)$ is $G^{-1}(p_0(X))$, then $\hat{\gamma}_{IPT}$ will be consistent. The choice of what other functions of X to include in $t(X)$ has implications for efficiency alone (and perhaps finite sample bias). In this special case, the problem of choosing $t(X)$ is closely related to that of moment selection in conditional moment problems (e.g., Donald, Imbens and Newey, 2008). Hirano, Imbens and Ridder (2003) also suggest incorporating a known propensity score in a similar fashion, but do not make the connection between overparameterization of the propensity score and moment selection. This connection is made, by construction, explicit in the IPT setting. When the propensity score is also nonparametric, choosing $t(X)$ is no longer analogous to a pure moment selection problem since $t(X)$ also determines the quality with which the propensity score is approximated. It would be interesting to explore automated, data dependent, procedures for choosing the components and dimension of $t(X)$ in the above settings.

Estimation of overidentified moment condition models If $\dim(\psi(Z, \gamma)) > \dim(\gamma)$ the procedure outlined above is not directly applicable. One approach to overidentification would be to estimate the inverse probability tilt as described above. In step two the analyst could then apply two-step GMM (or GEL) using the IPT-reweighted data. We conjecture that this procedure would be locally efficient and doubly robust. It would be interesting to construct a one step estimator for overidentified models.

A Appendix

This appendix outlines the proofs of the results given in the main text. Throughout the Appendix we assume that $t(X) = t^*(X) = r(X)$ so that $\Pi_0 = \Pi_0^*$ and $\delta_0 = \delta_0^*$.

This is done only to simplify the notation and is without loss of generality. We also drop ‘0’ subscripts, used to denote (true) population values, when doing so causes no confusion. A supplemental web appendix, available at <https://files.nyu.edu/bsg1/public/>, contains additional proofs, Monte Carlo results, and empirical application details.

Local efficiency and double robustness of $\widehat{\gamma}_{IPT}$ (Theorems 2.2 and 2.1)

Consistency and double robustness When Assumptions 1.1 to 1.5 hold consistency follows from arguments analogous to those of Wooldridge (2007) for IPW. If Assumptions 1.1 to 1.4 and 2.1 hold, but not 1.5 (we *do* assume that the $G(\cdot)$ function satisfies the stated regularity conditions; in particular that $G(t(x)'\delta) > 0$ for all $x \in \mathcal{X}$ and $\delta \in \mathcal{D}$) we have $\widehat{\delta} \xrightarrow{P} \delta_*$ where δ_* is the pseudo-true value which solves $\mathbb{E}[(D/G(t(X)'\delta_*) - 1)t(X)] = 0$. This gives $\mathbb{E}[p_0(X)t(X)/G(t(X)'\delta_*)] = \mathbb{E}[t(X)]$ so that under Assumption 1.3 and 2.1 we have equation (13) of the main text. Therefore $\gamma = \gamma_0$ is a solution to the IPW population moment. If $\psi(Z, \gamma)$ is linear in γ , then this solution is also unique. Otherwise uniqueness follows by hypothesis.

Asymptotic normality Asymptotic normality of $\widehat{\gamma}_{IPT}$ follows from Theorem 6.1 of Newey and McFadden (1994). Let $\beta = (\gamma', \delta)'$. The $K + 1 + M \times 1$ moment vector and derivative matrix equal

$$m_i(\beta) = \begin{pmatrix} \frac{D_i}{G_i(\delta)} \psi_i(\gamma) \\ \left(\frac{D_i}{G_i(\delta)} - 1\right) t_i \end{pmatrix}, \quad M_i(\beta) = \begin{bmatrix} \frac{D_i}{G_i(\delta)} \frac{\partial \psi_i(\gamma)}{\partial \gamma'} & -\frac{D_i}{G_i(\delta)} \frac{G_{1i}(\delta)}{G_i(\delta)} \psi_i(\gamma) t_i' \\ 0 & -\frac{D_i}{G_i(\delta)} \frac{G_{1i}(\delta)}{G_i(\delta)} t_i t_i' \end{bmatrix}. \quad (26)$$

First consider the case where Assumptions 1.1 to 1.5 hold. Let $M = \mathbb{E}[M_i(\beta_0)]$ and $\Omega = \mathbb{E}[m_i(\beta_0)m_i(\beta_0)']$, then $\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \Psi_0)$ for $\Psi_0 = \left\{ (M'\Omega^{-1}M)^{-1} \right\}_{1:K, 1:K}$ (where $A_{1:K, 1:K}$ is the upper left hand $K \times K$ block of A). The covariance of $m_i = m_i(\beta_0)$ equals

$$\Omega = \begin{pmatrix} \mathbb{E}\left[\frac{\psi\psi'}{G}\right] & E_0 \\ E_0' & F_0 \end{pmatrix}, \quad (27)$$

with

$$E_0 = \mathbb{E}\left[\frac{1-G}{G}\psi t'\right], \quad F_0 = \mathbb{E}\left[\frac{1-G}{G}t t'\right]. \quad (28)$$

The the population mean of $M_i = M_i(\beta_0)$ equals

$$M = \begin{pmatrix} \Gamma & -\mathbb{E}\left[\frac{G_1}{G}\psi t'\right] \\ 0 & -\mathbb{E}\left[\frac{G_1}{G}tt'\right] \end{pmatrix}. \quad (29)$$

Using (27) and (29) we get a limiting sampling variance for $\sqrt{N}(\widehat{\beta} - \beta_0)$ equal to

$$M^{-1}\Omega M^{-1\nu} = \begin{pmatrix} \Gamma^{-1}\left(\mathbb{E}\left[\frac{\psi\psi'}{G}\right] - E_0F_0^{-1}E_0'\right)\Gamma^{-1\nu} + \Gamma^{-1}\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\Delta_0'F_0\Delta_0\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\Gamma^{-1\nu} \\ -\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}F_0\left\{E_0F_0^{-1} - \mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\right\}'\Gamma^{-1\nu} \\ -\Gamma^{-1}\left\{E_0F_0^{-1} - \mathbb{E}\left[\frac{G_1}{G}\psi t'\right]\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}\right\}F_0\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1} \end{pmatrix}, \quad (30)$$

$$V_{MM}(\delta)$$

where

$$\Delta_0 = \mathbb{E}\left[\frac{D}{G}(\psi - E_0F_0^{-1}t)S_\delta'\right], \quad V_{MM}(\delta_0) = \mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}F_0\mathbb{E}\left[\frac{G_1}{G}tt'\right]^{-1}. \quad (31)$$

Now consider the case where Assumptions 1.1 to 1.4 and 2.1 hold, but not 1.5. Let $\beta_* = (\gamma'_0, \delta'_*)'$, with δ_* the pseudo-true propensity score parameter. Let $G_* = G(t(X)'\delta_*)$ etc. Under this set of assumptions we have

$$\Omega_* = \begin{pmatrix} \mathbb{E}\left[\frac{p_0(X)}{G_*^2}\psi\psi'\right] & \mathbb{E}\left[\frac{p_0(X)}{G_*}\left(\frac{1-G_*}{G_*}\right)\psi t'\right] \\ \mathbb{E}\left[\frac{p_0(X)}{G_*}\left(\frac{1-G_*}{G_*}\right)t\psi'\right] & \mathbb{E}\left[\left(\frac{p_0(X)}{G_*^2} - 2\frac{p_0(X)}{G_*} + 1\right)tt'\right] \end{pmatrix},$$

and

$$M_* = \begin{pmatrix} \mathbb{E}\left[\frac{p_0(X)}{G_*}\frac{\partial\psi}{\partial\gamma'}\right] & -\mathbb{E}\left[\frac{p_0(X)}{G_*}\frac{G_{1*}}{G_*}\psi t'\right] \\ 0 & -\mathbb{E}\left[\frac{p_0(X)}{G_*}\frac{G_{1*}}{G_*}tt'\right] \end{pmatrix},$$

so that $\Psi_0 = \left\{(M_*'\Omega_*^{-1}M_*)^{-1}\right\}_{1:K,1:K}$.

Local efficiency If Assumption 2.1 also holds we have $\mathbb{E}[\psi|X] = \Pi_0 t = q$ so that $E_0F_0^{-1} = \Pi_0$ and hence

$$E_0F_0^{-1}E_0' = \mathbb{E}\left[\frac{1-G}{G}\Pi_0tt'\Pi_0'\right] = \mathbb{E}\left[\frac{1-G}{G}qq'\right], \quad (32)$$

which gives the equality $\Gamma^{-1} \left(\mathbb{E} \left[\frac{\psi \psi'}{G} \right] - E_0 F_0^{-1} E_0' \right) \Gamma^{-1'} = \mathcal{I}(\gamma_0)^{-1}$. In that case we also have $\Delta_0 = 0$ since $\mathbb{E}[\psi | D, X] = E_0 F_0^{-1} t$. Under these conditions (30) simplifies to

$$M^{-1} \Omega M^{-1'} = \text{diag} \left(\mathcal{I}(\gamma_0)^{-1}, V_{MM}(\delta_0) \right). \quad (33)$$

Local efficiency at Assumption 2.1 follows if we can show that IPT is regular under Assumptions 1.1 to 1.5. The score function for a parametric submodel of the semiparametric missing data model is (e.g., Chen, Hong and Tarozzi, 2008)

$$\begin{aligned} s_\eta(y, x, d; \eta) &= ds_\eta(y_1 | x; \eta) \\ &+ \frac{d - G(t(x)' \delta)}{G(t(x)' \delta) [1 - G(t(x)' \delta)]} G_1(t(x)' \delta) t(x) \times \left(\frac{\partial \delta}{\partial \eta} \right) + r_\eta(x; \eta). \end{aligned}$$

Under Assumption 1.1 we have, differentiating under the integral and using iterated expectations,

$$\begin{aligned} \frac{\partial \gamma(\eta_0)}{\partial \eta} &= -\Gamma^{-1} \mathbb{E} \left[\psi(Z, \gamma_0) \frac{\partial \log f(Y_1, X; \eta_0)}{\partial \eta} \right] \\ &= -\Gamma^{-1} \left\{ \mathbb{E}[\psi(Z, \gamma_0) s_\eta(Y_1 | X; \eta_0)] + \mathbb{E}[q(X; \gamma) r_\eta(X; \eta_0)] \right\}. \end{aligned}$$

Under Assumptions 1.1 to 1.5 standard calculations yield an asymptotically linear representation of $\hat{\gamma}$ equal to:

$$\hat{\gamma} = \gamma_0 - \frac{1}{N} \sum_{i=1}^N \Gamma^{-1} \left\{ \frac{D_i \psi(Z_i, \gamma_0)}{G(t(X_i)' \delta_0)} - M_{12} M_{22}^{-1} \left(\frac{D_i}{G(t(X_i)' \delta_0)} - 1 \right) t(X_i) \right\} + o_p(N^{-1/2}),$$

where $-\Gamma^{-1}$ times the term in $\{\cdot\}$ is the influence function and M_{12} and M_{22} denote the upper right-hand $K \times 1 + M$ and lower right-hand $1 + M \times 1 + M$ blocks of M as given in (29) above. Let ϕ denote this influence function, by Theorem 2.2 of Newey (1990), regularity of $\hat{\gamma}$ follows if

$$\frac{\partial \gamma(\eta_0)}{\partial \eta} = \mathbb{E}[\phi s_\eta(Y, X | \eta_0)] = -\Gamma^{-1} \left\{ \mathbb{E}[\psi(Z, \gamma_0) s_\eta(Y_1 | X; \eta_0)] + \mathbb{E}[q(X; \gamma_0) r_\eta(X; \eta_0)] \right\}.$$

We have, using the conditional mean zero property of scores, the MAR assumption,

and the fact that $p_0(X) = G(t(X)'\delta_0)$

$$\begin{aligned}
\mathbb{E}[\phi s_\eta(Y, X | \eta_0)] &= -\Gamma^{-1} \mathbb{E} \left[\begin{aligned} &\left\{ \frac{D_i \psi(Z, \gamma_0)}{G(t(X)'\delta_0)} - M_{12} M_{22}^{-1} \left(\frac{D_i}{G(t(X)'\delta_0)} - 1 \right) t(X_i) \right\} \\ &\times \{s_\eta(Y_1 | X; \eta_0) + r_\eta(X; \eta_0)\} \end{aligned} \right] \\
&= -\Gamma^{-1} \mathbb{E} \left[\frac{D_i \psi(Z, \gamma_0)}{G(t(X)'\delta_0)} \{s_\eta(Y_1 | X; \eta_0) + r_\eta(X; \eta_0)\} \right] \\
&= -\Gamma^{-1} \mathbb{E} [\psi(Z, \gamma_0) \{s_\eta(Y_1 | X; \eta_0) + r_\eta(X; \eta_0)\}] \\
&= -\Gamma^{-1} \{ \mathbb{E} [\psi(Z, \gamma_0) s_\eta(Y_1 | X; \eta_0)] + \mathbb{E} [q(X; \gamma_0) r_\eta(X; \eta_0)] \},
\end{aligned}$$

as required.

Consistent variance-covariance matrix estimation If Assumptions 1.1 to 1.4 and either 1.5 or 2.1 or both hold (as well as additional regularity conditions), then the asymptotic variance of $\hat{\gamma}$ may be consistently estimated by

$$\hat{\Psi} = \left\{ \left(\widehat{M}' \widehat{\Omega}^{-1} \widehat{M} \right)^{-1} \right\}_{1:K, 1:K}, \quad (34)$$

with $\widehat{M} = \sum_{i=1}^N M_i(\widehat{\beta}) / N$ and $\widehat{\Omega} = \sum_{i=1}^N m_i(\widehat{\beta}) m_i(\widehat{\beta})' / N$.

Derivation of the higher order bias of IPT (Theorem 3.1) Here we outline the derivation of the $O(N^{-1})$ bias expressions for $\widehat{\gamma}_{IPT}$ (i.e., equation (24) in the main text). The derivation of the corresponding bias expression for the class of AIPW estimators discussed in the main text can be found in the supplement. Newey and Smith (2004, Lemma A.4, pp. 241 - 242) provide a general formula for the $O(N^{-1})$ bias of M-estimators. As IPT and AIPW have M-estimator representations we use their general result in our calculations. We maintain Assumption 2.1 throughout in what follows (in addition to Assumptions 1.1 to 1.5).

Let $\widehat{\theta}$ be the solution to the $T = \dim(\theta)$ equations

$$\bar{b}(\widehat{\theta}) = \sum_{i=1}^N b_i(\widehat{\theta}) = 0. \quad (35)$$

Under regularity conditions (see below) Newey and Smith (2004, Lemma A.4) show

that the asymptotic bias of $\widehat{\theta}$ is given by

$$\text{Bias}(\widehat{\theta}) = \frac{-B^{-1}}{N} \left(\mathbb{E}[A_i \phi_i] + \frac{1}{2} \mathbb{E} \left[\sum_{q=1}^T \phi_{q,i} B_q \phi_i \right] \right), \quad (36)$$

where e_q is a $T \times 1$ column vector with a one in the q^{th} row and zeros elsewhere and

$$B = \mathbb{E} \left[\frac{\partial b_i(\theta)}{\partial \theta'} \right], \quad \phi_i = -B^{-1} b_i(\theta), \quad A_i = \frac{\partial b_i(\theta)}{\partial \theta'} - B, \quad B_q = \mathbb{E} \left[\frac{\partial^2 b_i(\theta)}{\partial \theta_q \partial \theta'} \right]. \quad (37)$$

The IPT estimator of $\theta = (\gamma', \delta')'$ is given by the solution to (35) with

$$b_i(\theta) = \begin{pmatrix} \frac{D_i}{G_i(\delta)} \psi_i(\gamma) \\ \left(\frac{D_i}{G_i(\delta)} - 1 \right) t_i \end{pmatrix}.$$

To apply (36) to IPT we require that the parameter space of θ is compact with θ_0 in its interior, continuity of $b_i(\theta)$ in θ and continuous differentiability in a neighborhood of θ_0 , and $\text{rank}(B) = \dim(\theta)$. These conditions are implied by Assumptions 1.1 and 1.5. Additionally we require a Lipschitz continuity condition on the third derivative of $b_i(\theta)$ and the existence of certain higher order moments. Specifically we assume that (i) for some $d(Z)$ with $\mathbb{E}[d(Z)] < \infty$

$$\left\| \frac{\partial^3 b_i(\theta)}{\partial \theta_j \partial \theta_k \partial \theta_l} - \frac{\partial^3 b_i(\theta_0)}{\partial \theta_j \partial \theta_k \partial \theta_l} \right\| \leq d(Z_i) \|\theta - \theta_0\|$$

and (ii) $\mathbb{E}[\|b_i(\theta_0)\|^6]$, $\mathbb{E}\left[\left\|\frac{\partial b_i(\theta_0)}{\partial \theta'}\right\|^6\right]$, $\mathbb{E}\left[\left\|\frac{\partial^2 b_i(\theta_0)}{\partial \theta_j \partial \theta_l}\right\|^6\right]$, and $\mathbb{E}\left[\left\|\frac{\partial^3 b_i(\theta_0)}{\partial \theta_j \partial \theta_k \partial \theta_l}\right\|^2\right]$ are finite for $j, k, l = 1, \dots, K + 1 + M$ (see Newey, 2002). These conditions will hold if $G(\cdot)$ and $\psi(z, \gamma)$ are both three times continuously differentiable with bounded derivatives and enough moments of $t(X)$ exist (e.g., if a component of $t(X)$ is a Cauchy random variable then (36) will not hold).

Objects, $\frac{\partial b_i(\theta_0)}{\partial \theta'}$, B and A_i of (37) above specialize to

$$\begin{aligned} \frac{\partial b_i(\theta_0)}{\partial \theta'} &= \begin{bmatrix} \frac{D_i}{G_i} \frac{\partial \psi_i}{\partial \gamma'} & -\frac{D_i}{G_i} \frac{G_{1i}}{G_i} \psi_i t'_i \\ 0 & -\frac{D_i}{G_i} \frac{G_{1i}}{G_i} t_i t'_i \end{bmatrix}, \quad B = \begin{bmatrix} \Gamma & -\mathbb{E}\left[\frac{G_1}{G} \psi t'\right] \\ 0 & -\mathbb{E}\left[\frac{G_1}{G} t t'\right] \end{bmatrix} \\ A_i &= \begin{bmatrix} \frac{D_i}{G_i} \frac{\partial \psi_i}{\partial \gamma'} - \Gamma & -\frac{D_i}{G_i} \frac{G_{1i}}{G_i} \psi_i t'_i + \mathbb{E}\left[\frac{G_1}{G} \psi t'\right] \\ 0 & -\frac{D_i}{G_i} \frac{G_{1i}}{G_i} t_i t'_i + \mathbb{E}\left[\frac{G_1}{G} t t'\right] \end{bmatrix}. \end{aligned}$$

Using the partitioned inverse formula we have

$$B^{-1} = \begin{bmatrix} \Gamma^{-1} & -\Gamma^{-1} \mathbb{E} \left[\frac{G_1}{G} \psi t' \right] \mathbb{E} \left[\frac{G_1}{G} t t' \right]^{-1} \\ 0 & -\mathbb{E} \left[\frac{G_1}{G} t t' \right]^{-1} \end{bmatrix}. \quad (38)$$

Combining the above expressions then gives

$$\begin{aligned} & \mathbb{E} [A_i \phi_i] \quad (39) \\ &= - \begin{bmatrix} \mathbb{E} \left[\frac{\partial \psi}{\partial \gamma'} \Gamma^{-1} \frac{1}{G} \psi \right] - \mathbb{E} \left[\frac{1-G}{G} \frac{\partial \psi}{\partial \gamma'} \Gamma^{-1} \mathbb{E} \left[\frac{G_1}{G} \psi t' \right] \mathbb{E} \left[\frac{G_1}{G} t t' \right]^{-1} t \right] + \mathbb{E} \left[\frac{1-G}{G} \frac{G_1}{G} \psi t' \mathbb{E} \left[\frac{G_1}{G} t t' \right]^{-1} t \right] \\ \mathbb{E} \left[\frac{1-G}{G} \frac{G_1}{G} t t' \mathbb{E} \left[\frac{G_1}{G} t t' \right]^{-1} t \right] \end{bmatrix}. \end{aligned}$$

Let $\Pi_* \stackrel{def}{=} \mathbb{E} \left[\frac{G_1}{G} \psi t' \right] \mathbb{E} \left[\frac{G_1}{G} t t' \right]^{-1}$; using (39) we have the first K rows of $-B^{-1} \mathbb{E} [A_i \phi_i]$ equal to

$$\begin{aligned} & \Gamma^{-1} \mathbb{E} \left[\frac{\partial \psi}{\partial \gamma'} \Gamma^{-1} \frac{1}{G} \{\psi - \Pi_* t\} \right] + \Gamma^{-1} \mathbb{E} \left[\frac{\partial \psi}{\partial \gamma'} \Gamma^{-1} \Pi_* t \right] \quad (40) \\ & + \Gamma^{-1} \mathbb{E} \left[\frac{1-G}{G} \frac{G_1}{G} \{\psi - \Pi_* t\} t' \mathbb{E} \left[\frac{G_1}{G} t t' \right]^{-1} t \right]. \end{aligned}$$

Assumption 2.1 gives $q = \Pi_0 t$ so that $\Pi_* = \Pi_0$; therefore, applying the law of iterated expectations, gives the last term in the expression above identically equal to zero.

Now consider the second component of the bias expression (36). Evaluating $\mathbb{E} [\phi_i \phi_i']$ yields

$$\mathbb{E} [\phi_i \phi_i'] = \begin{bmatrix} \mathcal{I}(\gamma_0)^{-1} & 0 \\ 0 & V_{MM}(\delta) \end{bmatrix}. \quad (41)$$

For $q = 1, \dots, K$, using the expression for $\partial b_i(\theta_0) / \partial \theta'$, we have

$$B_q = \mathbb{E} \begin{bmatrix} \frac{\partial^2 \psi}{\partial \gamma_q \partial \gamma'} & -\frac{G_1}{G} \frac{\partial \psi}{\partial \gamma_q} t' \\ 0 & 0 \end{bmatrix}, \quad (42)$$

for B_q as defined in (37) above. For $q = K+1, \dots, K+1+M (= T)$ we have instead

$$B_q = \mathbb{E} \begin{bmatrix} -\frac{G_1}{G} t_{q-K} \frac{\partial \psi}{\partial \gamma'} & \left(\frac{2G_1^2}{G^2} - \frac{G_2}{G} \right) t_{q-K} \psi t' \\ 0 & \left(\frac{2G_1^2}{G^2} - \frac{G_2}{G} \right) t_{q-K} t t' \end{bmatrix}. \quad (43)$$

Using (41), (42) and (43) the first K rows of $\frac{-B^{-1}}{2N} \mathbb{E} \left[\sum_{q=1}^T \phi_{q,i} B_q \phi_i \right]$ can be shown to equal

$$\left\{ \frac{-B^{-1}}{2N} \mathbb{E} \left[\sum_{q=1}^T \phi_{q,i} B_q \phi_i \right] \right\}_{1:K,:} = -\frac{1}{2N} \sum_{k=1}^K \Gamma^{-1} \mathbb{E} \left[\frac{\partial^2 \psi}{\partial \gamma_k \partial \gamma'} \right] \mathcal{I}(\gamma_0)^{-1} e_k. \quad (44)$$

Combining (40) and (44) yields C_O as given in the statement of the Theorem.

Computation Computation of $\widehat{\gamma}_{IPT}$ consists of two steps. In the first step, which is nonstandard and detailed here, $\widehat{\delta}$ is computed as the solution to (8). Here we outline an approach to solving (8) which we have found to be computationally convenient and very reliable in practice. This involves defining $\widehat{\delta}$ to be the solution to a globally concave programming problem with unrestricted domain. In the second $\widehat{\gamma}$ is computed as the solution to (7).²²

Consider the following function

$$\varphi(v) = \frac{v}{G(v)} + \int_{1/G(v)}^a G^{-1} \left(\frac{1}{t} \right) dt, \quad (45)$$

with $G(\cdot)$ as defined in Assumption 1.5. When the propensity score takes the logit $G(v) = (1 + \exp(-v))^{-1}$ form (45) exists in closed form (see below). We implement the logit specification in the empirical application and expect that most users will do likewise. If a different propensity score model is assumed, then (45) is can be evaluated numerically.

The first and second derivatives of $\varphi(v)$ are

$$\varphi_1(v) = \frac{1}{G(v)}, \quad \varphi_2(v) = -\frac{G_1(v)}{G(v)^2}, \quad (46)$$

so that (45) is strictly concave.

We compute $\widehat{\delta}$ by solving the following optimization problem

$$\max_{\delta} l_N(\delta), \quad l_N(\delta) = \frac{1}{N} \sum_{i=1}^N D_i \varphi(t(X_i)' \delta) - \frac{1}{N} \sum_{i=1}^N t(X_i)' \delta. \quad (47)$$

²²The second step is identical to that associated with standard inverse probability weighting (IPW). As the second step is both application specific, and typically straightforward to compute using standard software (that accepts user-specified weights), we do not detail it here.

Differentiating $l_N(\delta)$ with respect to δ gives an $1 + M \times 1$ gradient vector of

$$\nabla_{\delta} l_N(\delta) = \frac{1}{N} \sum_{i=1}^N D_i \varphi_1(t(X_i)' \delta) t(X_i) - \frac{1}{N} \sum_{i=1}^N t(X_i), \quad (48)$$

which coincides with (8) as required. The $1 + M \times 1 + M$ Hessian matrix is

$$\nabla_{\delta\delta} l_N(\delta) = \frac{1}{N} \sum_{i=1}^N D_i \varphi_2(t(X_i)' \delta) t(X_i) t(X_i)'. \quad (49)$$

This is a negative definite function of δ ; the problem (47) is consequently concave with a unique solution (if one exists). Existence of a solution requires that $\sum_{i=1}^N t(X_i)/N$ lie within the convex hull of the complete case subsample (this will be true in large samples under Assumption 1.4, but should nevertheless be checked prior to computation).²³

In practice (48) will have an ‘exploding denominator’ when $t(X_i)' \delta$ is a large negative number. This can lead to numerical instabilities by causing the Hessian to be ill-conditioned. We address this problem by noting that at a valid solution $\sum_{i=1}^N D_i/G(t(X_i)' \hat{\delta})/N = 1$. Since Assumption 1.5 implies that $G(v)$ is bounded below by zero, this means that $D_i/G(t(X_i)' \hat{\delta}) < N$ for all $i = 1, \dots, N$. Letting $v_i = t(X_i)' \delta$ this inequality corresponds to requiring that

$$G^{-1}(D_i/N) < v_i, \quad i = 1, \dots, N \quad (50)$$

at $\delta = \hat{\delta}$. Let $v_N^* = G^{-1}(1/N)$; note that $v_N^* \rightarrow -\infty$ as $N \rightarrow \infty$ suggesting that (50) will be satisfied for most values of δ in large enough samples. In small samples (50) may be violated for some i at some iterations of the maximization procedure (although not at a valid solution). Our approach to estimation involves replacing $\varphi(v)$ with a quadratic function when $v \leq v_N^*$; this ensures that the denominator in (48) is bounded. This will improve the condition of the Hessian with respect to δ without changing the solution. Owen (2001, Chapter 12) proposes a similar procedure in the context of empirical likelihood estimation of moment condition models.

²³Convex hull conditions also arise in research on empirical likelihood (e.g., Owen, 2001; pp. 85 - 87).

Specifically we replace $\varphi(v)$ in (47), (48) and (49) with

$$\varphi_N^\circ(v) = \begin{cases} \varphi(v) & v > v_N^* \\ a_N + b_N v_N^* + \frac{c_N}{2} (v_N^*)^2 & v \leq v_N^* \end{cases}, \quad (51)$$

where a_N , b_N and c_N are the solutions to

$$\begin{aligned} c_N &= \varphi_2(v_N^*) \\ b_N + c_N v_N^* &= \varphi_1(v_N^*) \\ a_N + b_N v_N^* + \frac{c_N}{2} (v_N^*)^2 &= \varphi_0(v_N^*). \end{aligned}$$

This choice of coefficients ensures that $\varphi_N^\circ(v)$ equals $\varphi(v)$, as well as equality of first and second derivatives, at $v = v_N^*$.

When $G(v)$ is logit our algorithm is particularly simple to implement. For $G(v) = \exp(v) / [1 + \exp(v)]$ we have

$$\varphi(v) \propto v - \exp(-v).$$

Differentiating with respect to v then gives $\varphi_1(v) = 1 + \exp(-v)$ and $\varphi_2(v) = -\exp(-v)$.

We also have $v_N^* = G^{-1}(1/N) = \ln\left(\frac{1/N}{1-1/N}\right) = \ln\left(\frac{1}{N-1}\right)$ so that solving for a_N , b_N and c_N yields

$$\begin{aligned} a_N &= -(N-1) \left[1 + \ln\left(\frac{1}{N-1}\right) + \frac{1}{2} \left[\ln\left(\frac{1}{N-1}\right) \right]^2 \right], \\ b_N &= N + (N-1) \ln\left(\frac{1}{N-1}\right), \quad c_N = -(N-1). \end{aligned}$$

References

- [1] Abowd, John M., Bruno Crépon, and Francis Kramarz. (2001). "Moment estimation with attrition: an application to economic models," *Journal of the American Statistical Association* 96 (456): 1223 - 1231.
- [2] Angrist, Joshua D. and Alan B. Krueger. (1992). "The effect of age at school entry on educational attainment: an application of instrumental variables with

- moments from two samples,” *Journal of the American Statistical Association* 87 (418): 328 - 336.
- [3] Back, Kerry and David P. Brown. (1993). “Implied probabilities in GMM estimators,” *Econometrica* 61 (4): 971 - 975.
- [4] Bang, Heejung and James M. Robins. (2005). “Doubly robust estimation in missing data and causal inference models,” *Biometrics* 61 (4): 962 - 972.
- [5] Busso, Matias, John DiNardo and Justin McCrary. (2009). “Finite sample properties of semiparametric estimators of average treatment effects,” *Mimeo*.
- [6] Cao, Weihua, Anastasios A. Tsiatis and Marie Davidian. (2009). “Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data,” *Biometrika* 96 (3): 723 - 734.
- [7] Chen, Xiaohong. (2007). “Large sample sieve estimation of semi-nonparametric models,” *Handbook of Econometrics* 6 (B): 5549 - 5632. (J.J. Heckman & E.E. Leamer, Eds.). Amsterdam: North-Holland.
- [8] Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2004). “Semiparametric efficiency in GMM models of nonclassical measurement errors, missing data and treatment effects, *Mimeo*.
- [9] Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2008). “Semiparametric efficiency in GMM models with auxiliary data,” *Annals of Statistics* 36 (2): 808 - 843.
- [10] Donald, Stephen G., Guido W. Imbens and Whitney K. Newey. (2008). “Choosing the number of moments in conditional moment restriction models,” *Mimeo*.
- [11] Graham, Bryan S. (2011). “Efficiency bounds for missing data models with semi-parametric restrictions,” *Econometrica* 79 (2): 437 - 452.
- [12] Graham, Bryan S., Cristine Campos de Xavier Pinto and Daniel Egel. (2011). “Inverse probability tilting estimation of Average Treatment Effects in Stata,” *Mimeo*.

- [13] Hahn, Jinyong. (1998). "On the role of the propensity score in efficient semi-parametric estimation of average treatment effects," *Econometrica* 66 (2): 315 - 331.
- [14] Hirano, Keisuke and Guido W. Imbens. (2001). "Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization," *Health Services and Outcomes Research* 2 (3-4): 259 -278.
- [15] Hirano, Keisuke, Guido W. Imbens and Geert Ridder. (2003). "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica* 71 (4): 1161 - 1189.
- [16] Imbens, Guido. W. (1997). "One-step estimators of over-identified generalized method of moments models," *Review of Economic Studies* 64 (3): 359 - 383.
- [17] Imbens, Guido W. (2004). "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics* 86 (1): 4 - 29.
- [18] Imbens, Guido W. and Tony Lancaster. (1996). "Efficient estimation and stratified sampling," *Journal of Econometrics* 74 (2): 289 - 318.
- [19] Johnson, William R. and Derek A. Neal (1998). "Basic skills and the black-white earnings gap," *The Black-White Test Score Gap*: 480 - 500. (C. Jencks & M. Phillips, Eds.). Washington, D.C.: The Brookings Institution.
- [20] Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*. Hoboken, N.J.: John Wiley & Sons, Inc.
- [21] Neal, Derek A. and William R. Johnson. (1996). "The role of premarket factors in black-white wage differences," *Journal of Political Economy* 104 (5): 869 - 895.
- [22] Newey, Whitney K. (1990). "Semiparametric efficiency bounds," *Journal of Applied Econometrics* 5 (2): 99 - 135.
- [23] Newey, Whitney K. (1994). "Series estimation of regression functionals," *Econometric Theory* 10 (1): 1 - 28.

- [24] Newey, Whitney K. (2002). “Stochastic Expansion for M-Estimator,” *Lecture Note*.
- [25] Newey, Whitney K. and Daniel McFadden. (1994). “Large sample estimation and hypothesis testing,” *Handbook of Econometrics 4*: 2111 - 2245 (R.F. Engle & D.F. McFadden, Eds.). Amsterdam: North-Holland.
- [26] Newey, Whitney K. and Richard J. Smith. (2004). “Higher order properties of GMM and generalized empirical likelihood estimators,” *Econometrica* 72 (1): 219 - 255.
- [27] Owen, Art. B. (2001). *Empirical Likelihood*. New York: Chapman & Hall/CRC.
- [28] Qin, Jing, Biao Zhang, and Denis H. Y. Leung. (2009). “Empirical likelihood in missing data problems,” *Journal of the American Statistical Association* 104 (488): 1492 - 1503.
- [29] Ridder, Geert and Robert Moffitt. (2007). “The econometrics of data combination,” *Handbook of Econometrics* 6 (2): 5469 - 5547 (J.J. Heckman & E Leamer, Eds.). New York: North-Holland.
- [30] Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. (1994). “Estimation of regression coefficients when some regressors are not always observed,” *Journal of the American Statistical Association* 89 (427): 846 - 866.
- [31] Robins, James, Mariela Sued, Quanhong Lei-Gomez and Andrea Rotnitzky. (2007). “Comment: performance of double-robust estimators when “inverse probability” weights are highly variable,” *Statistical Science* 22 (4): 544 - 559.
- [32] Tan, Zhiqiang. (2010). “Bounded, efficient and doubly robust estimation with inverse weighting,” *Biometrika* 97 (3): 661 - 682.
- [33] Tsiatis, Anastasios A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
- [34] Wooldridge, Jeffrey M. (2001). “Asymptotic properties of weighted M-estimators for standard stratified samples,” *Econometric Theory* 17 (2): 451 - 470.
- [35] Wooldridge, Jeffrey M. (2007). “Inverse probability weighted estimation for general missing data problems,” *Journal of Econometrics* 141 (2): 1281 - 1301.

Table 1: Weight functions for different AIPW estimators

AIPW Estimator	$\omega_i(\delta)$	$\nu_i(\delta)$	Locally Efficient?	Doubly Robust?
Robins, Rotnitzky, and Zhao (1994)	$G_i(\delta)$	$\frac{D_i}{G_i(\delta)}$	Yes	Yes
Newey (1994)	1	1	Yes	No
Cao, Tsiatis and Davidian (2009)	$\frac{1-G_i(\delta)}{G_i(\delta)}$	$\frac{D_i}{G_i(\delta)}$	Yes	Yes
Hirano and Imbens (2001) / Wooldridge (2007)	1	$\frac{D_i}{G_i(\delta)}$	Yes	Yes

Table 2: Replication of Table 14-1 of Johnson and Neal (1998) and unweighted complete case analysis with pre-adolescent test score

	(1)	(2)	(3)	(4)
	<i>OLS</i>	<i>OLS</i>	<i>CC - OLS</i>	<i>CC - OLS</i>
YEAROFBIRTH	-0.0458 (0.0151)**	-0.0466 (0.0147)**	-0.0947 (0.0464)*	-0.0940 (0.0470)*
BLACK	-0.2776 (0.0261)**	-0.1079 (0.0284)**	-0.2708 (0.0833)**	-0.1606 (0.0900) ⁺
AFQT	-	0.1645 (0.0146)**	-	-
EARLYTEST	-	-	-	0.1011 (0.0540) ⁺
R^2	0.062	0.183	0.068	0.11
N	1,371	1,371	144	144

NOTES: Estimation samples are as described in the main text. The 1979 baseline sampling weights are used in place of the empirical measure when computing all estimates. A ‘***’, ‘**’ and ‘+’ denotes that a point estimate is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of heteroscedasticity and dependence across units residing in the same household at baseline.

Table 3: IPT, IPW and AIPW estimates of the Black-White wage gap conditional on preadolescent skills

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	<i>CC - OLS</i>	<i>IPT</i>	<i>IPW</i>	<i>AIPW_{RRZ}</i>	<i>AIPW_{NEWWEY}</i>	<i>AIPW_{HIW}</i>	<i>AIPW_{CTD}</i>
YEAROFBIRTH	-0.0940 (0.0470)*	-0.0537 (0.0162)**	-0.0970 (0.0303)**	-0.0533 (0.0165)**	-0.0339 (0.0541)	-0.0535 (0.0166)**	-0.0543 (0.0167)**
BLACK	-0.1606 (0.0900)+	-0.1837 (0.0356)**	-0.2136 (0.0795)*	-0.1834 (0.0419)**	-0.0745 (0.157)	-0.1871 (0.0392)**	-0.1837 (0.0390)**
EARLYTEST	0.1011 (0.0540)+	0.1112 (0.0296)**	0.1221 (0.0362)**	0.1050 (0.0358)**	0.0951 (0.0361)	0.1072 (0.0346)**	0.1144 (0.0344)**

NOTES: Samples are as described in the main text. The 1979 baseline sampling weights are used when computing all estimates. A **, *, + and + denotes that a coefficient is significantly different from zero at the 1, 5 and 10 percent levels. Standard errors (in parentheses) allow for arbitrary patterns of heteroscedasticity and dependence across units residing in the same household at baseline.