Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST)[1]

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

Initial Draft: July 2006

This Draft: March 18, 2015

[1]We would like to thank David Card, Stephen Cosslett, Jinyong Hahn, Michael Jansson, Patrick Kline, Richard Smith, Tom Rothenberg, and members of the Berkeley Econometrics Reading Group for helpful discussions. We are particularly grateful to Gary Chamberlain, Guido Imbens, Justin McCrary, Geert Ridder, Enrique Sentana and Leonard Stefanski for detailed comments on earlier drafts. This draft has benefited from comments by the co-editor, associate editor and three anonymous referees. We thank Jing Qin and Biao Zhang for assistance in replicating the Monte Carlo designs in Qin and Zhang (2008). 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, University of California - Berkeley, CEMFI (Madrid), Pontifícia Universidade Católica do Rio de Janeiro and the 2013 North American Summer Meeting of the Econometric Society. Preliminary portions of the current paper previously appeared in Section 4 of an early draft of the NBER Working Paper 'Inverse probability tilting and missing data problems'. The published version of that paper excludes the material reported here. A supplemental appendix with proofs and additional details regarding computation may be found on the first author's web page. All the usual disclaimers apply.

[◇]Department of Economics, 530 Evans Hall, University of California - Berkeley, Berkeley, CA 94720-3880 and NBER. E-Mail: bgraham@econ.berkeley.edu, Web: https://emlab.berkeley.edu/~bgraham/

[+]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. E-mail: Daniel_Egel@rand.org.

Abstract

We propose a locally efficient estimator for a class of semiparametric data combination problems. A leading estimand in this class is the Average Treatment Effect on the Treated (ATT). Data combination problems are related to, but distinct from, the class of missing data problems analyzed by Robins, Rotnitzky and Zhao (1994) (of which the Average Treatment Effect (ATE) estimand is a special case). Our estimator also possesses a double robustness property. Our procedure may be used to efficiently estimate, among other objects, the ATT, the two-sample instrumental variables model (TSIV), counterfactual distributions, poverty maps, and semiparametric difference-in-differences. In an empirical application we use our procedure to characterize residual Black-White wage inequality after flexibly controlling for 'pre-market' differences in measured cognitive achievement as in Neal and Johnson (1996).

JEL Classification: C14, C21, C23

Key Words: Data Combination, Two-Sample Instrumental Variables (TSIV), Average Treatment Effect on the Treated (ATT), Poverty Maps, Earnings decompositions, Direct Standardization, Semiparametric Difference-in-Differences, Semiparametric Efficiency, Black-White Gap, Double Robustness, Propensity Score

1 Introduction

Let $Z = (W', X', Y')'$ denote a random vector drawn from some *study population* of interest with distribution function F_s. For some unique γ_0, and known function $\psi(z, \gamma)$ of the same dimension, we assume that

$$\mathbb{E}_s[\psi(Z, \gamma_0)] = 0, \tag{1}$$

where $\mathbb{E}_s[\cdot]$ denotes expectations taken with respect to the study population. If a random sample of Z is available, then consistent estimation of γ_0 (under regularity conditions) is straightforward (e.g., Newey and McFadden, 1994). Many statistical models of interest can be represented in terms of moment restrictions like (1); see Wooldridge (2002) for a textbook exposition.

In this paper we consider estimation of γ_0 when a random sample of Z is unavailable. Instead two separate samples are available. The first is drawn from the study population and contains N_s measurements of (Y, W). The second is drawn from an *auxiliary population* (with distribution function F_a; $\mathbb{E}_a[\cdot]$ denotes expectations taken with respect to this distribution) and contains N_a measurements of (X, W). While the variable W is common to the two samples, X and Y are not. Hahn (1998) and Chen, Hong and Tarozzi (2008) show that identification of γ_0 follows if (i) the conditional distributions of X given W in the two populations coincide (although their marginal distributions for W may differ), (ii) the support of W in the auxiliary population is at least as large as that in the study population and (iii) $\psi(z, \gamma_0)$ is separable in the components depending on the 'non-common' variables Y and X

$$\psi(Z, \gamma_0) = \psi_s(Y, W, \gamma_0) - \psi_a(X, W, \gamma_0). \tag{2}$$

Examples of statistical problems to which the above setup applies include the two sample instrumental variables (TSIV) model of Angrist and Krueger (1992) and Ridder and Moffitt (2007), the average treatment effect on the treated (ATT) estimand from the program evaluation literature (e.g., Heckman and Robb, 1985; Imbens, 2004), counterfactual earn-

ings/wealth decompositions as in Dinardo, Fortin and Lemieux (1996) and Barsky, Bound, Charles and Lupton (2002), poverty mapping as in Elbers, Lanjouw and Lanjouw (2003) and Tarozzi and Deaton (2009), direct standardization methods used in demography (e.g., Kitagawa, 1964), and models with mismeasured regressors and validation samples (e.g., Carroll and Wand, 1991).

To help fix ideas consider the ATT example. Here Y denotes an individual's potential outcome under active treatment, say earnings given participation in a job training program, X denotes her outcome under control (earnings in the absence of training) and W is a vector of baseline covariates. Available is a random sample of (Y, W) from the population assigned active treatment (i.e., 'the treated'). A separate sample of measurements of (X, W) is drawn from a population of controls. The ATT, $\gamma_0 = \mathbb{E}_s[Y - X]$, is given by the solution to (1) with $\psi_s(Y, W, \gamma_0) = Y$ and $\psi_a(X, W, \gamma_0) = X + \gamma_0$.

Dehejia and Wahba (1999), revisiting earlier work by LaLonde (1986), combine two distinct samples to estimate the effect of the National Supported Work (NSW) demonstration, a labor training program, on the post-intervention earnings of trainees. Their study sample consists of 185 NSW participants, while their auxiliary sample includes 2,490 non-participants drawn from the Panel Study of Income Dynamics (PSID). These two samples consist of random draws from distinct, non-overlapping, populations. The two sample feature of their analysis distinguishes it from one seeking to estimate a population average treatment effect (ATE). In that case the researcher generally bases her analysis on a random sample from the population of interest, where some units happen to be treated, and others not (e.g., Rosenbaum and Rubin, 1983). There the inferential problem is usefully conceptualized as one of missing data and the general theory of Robins, Rotnitzky and Zhao (1994) directly applies.

Relationship between data combination and missing data problems One perspective is that data combination problems are nothing more than a particular class of "missing

data" problems in which the auxiliary sample is collected independently, and from a different population than that, of the study sample. Our use of the term missing data is more technical, referring, in particular, to the family of problems analyzed by Robins, Rotnitzky and Zhao (1994, Section 8.1). In this family both the study and auxiliary samples are random ones from the population of interest. It turns out that this difference has statistical content with, as we emphasize here (and others have before us), implications for estimator formulation and properties. In an important paper Hahn (1998) showed that while prior restrictions on the form of the propensity score do not lower the semiparametric variance bound for the ATE, they do lower the corresponding bound for the ATT. Chen, Hong and Tarozzi (2008) generalize this result, showing that, unlike in the missing data context (their 'verify-in-sample' case), knowledge of the form of the propensity score is asymptotically valuable in data combination problems (their 'verify-out-of-sample' case).

Our contribution is to develop a flexible parametric estimator for general data combination problems with good efficiency and robustness properties. Similar to the augmented inverse probability weighting (AIPW) estimator for missing data problems due to Robins, Rotnitzky and Zhao (1994), our data combination procedure is locally efficient and possesses a double robustness property. This latter property, given the non-ancillarity of the propensity score in the data combination problem, is surprising.

To our knowledge we are the first to propose a locally efficient estimator in the data combination context. Chen, Hong and Tarozzi (2008) propose a globally efficient estimator, but their procedure requires nonparametric modelling as opposed to the flexible parametric approach adopted here. Our methods provide a practical alternative to theirs when W is high dimensional (cf., Firpo and Rothe, 2013). Abadie (2005) develops a parametric propensity score reweighting (PSR) estimate of the ATT. Qin and Zhang (2008) show that Abadie's estimator can have low efficiency in some settings and propose an alternative that uses empirical likelihood ideas. Qin and Zhang (2008) do not characterize the semiparametric efficiency or robustness properties of their ATT estimator, nor show how to extend it to

the wider class of problems considered here. Hirano and Imbens (2001) also propose a type of propensity score reweighting estimator for the ATT. Their estimator exhibits a double robustness property, but they do not consider issues of semiparametric efficiency nor general data combination problems as we do. Besides its robustness and efficiency properties, our estimator is simple to compute and is suitable for many applied problems, like the estimation of the ATT, two sample instrumental variables and others cited above.

In Section 2 we define the semiparametric data combination model. Modestly extending the work of Chen, Hong and Tarozzi (2008) we calculate the semiparametric efficiency bound for our model. We relate our efficiency bound analysis to prior work on distribution function estimation based on a random sample from the population of interest and a second, biased, sample from the same population (e.g., Qin, 1998; Gilbert, Lele, Vardi, 1999). This discussion motivates the form of our AST estimator, which we introduce in Section 3, where we also formally characterize its large sample properties. Our key results are Theorems 3.1 to 3.3 below. Section 4 provides an illustrative empirical application and reports on the results of several Monte Carlo experiments. Proofs of our main results are contained in the Appendix. The Supplemental Web Appendix contains additional proof details, extra examples of data combination problems, and additional Monte Carlo results. An algorithm for computing our estimator, that we have found to work well in practice, is also described in the Supplemental Web Appendix.

2 Semiparametric data combination model

A formal definition of the data combination model is given by Assumption 2.1 below. Let $A \Subset \mathbb{R}^P$denote a compact subset of $\mathbb{R}^P$.

Assumption 2.1 ***Semiparametric Data Combination Model***

(i) (IDENTIFICATION) *For some* $\psi(z,\gamma) = \psi_s(y,w,\gamma) - \psi_a(x,w,\gamma)$*, equation (1) holds with* $\mathbb{E}_s[\psi(Z,\gamma)] \neq 0$ *for all* $\gamma \neq \gamma_0$, $\gamma \in \mathcal{G} \Subset \mathbb{R}^K$, $z \in \mathcal{Z} \Subset \mathbb{R}^{\dim(Z)}$;

(ii) (Conditional Distributional Equality) $F_s(x|w) = F_a(x|w)$ *and* $F_s(y|w) = F_a(y|w)$ *for all* $w \in \mathcal{W} \Subset \mathbb{R}^{\dim(W)}$, $x \in \mathcal{X} \Subset \mathbb{R}^{\dim(X)}$ *and* $y \in \mathcal{Y} \Subset \mathbb{R}^{\dim(Y)}$;

(iii) (Weak Overlap) *Let* $S_j = \{w : f_j(w) > 0\}$ *for* $j = s, a$, *then* $S_s \subset S_a$;

(iv) (Multinomial Sampling) *With probability* $Q_0 \in (\xi, 1-\xi)$ *for* $0 < \xi < 1$ *we draw a unit at random from* F_s *and record its realizations of* Y *and* W, *otherwise we draw a unit at random from* F_a *and record its realizations of* X *and* W. *Let* $D_i = 1$ *if the* i^{th} *draw* $(i = 1, \ldots, N)$ *corresponds to a study population unit and* $D_i = 0$ *otherwise;*

(v) (Propensity score model) *There is a unique* $\delta_0 \in \mathcal{D} \Subset \mathbb{R}^{\dim(\delta)}$, *known vector* $r(W)$ *of linearly independent functions of* W *with a constant in the first row, and known function* $G(\cdot)$ *such that (a)* $G(\cdot)$ *is strictly increasing, differentiable and maps into the unit interval with* $\lim_{v \to -\infty} G(v) = 0$ *and* $\lim_{v \to \infty} G(v) = 1$, *(b)* $\frac{f_s(w)}{f_a(w)} = \frac{1-Q_0}{Q_0} \frac{G(r(w)'\delta_0)}{1-G(r(w)'\delta_0)}$ *for all* $w \in \mathcal{W}$, *and (c)* $0 < G(r(w)'\delta) \leq \kappa < 1$ *for all* $\delta \in \mathcal{D}$ *and* $w \in \mathcal{W}$.

The first part of Assumption 2.1 implies global identifiability of the complete data model. The second part implies that the distributions of (Y, W) and (X, W) in the two populations differ only in terms of their marginal distributions for the always measured variable, W. The third part ensures that, in large samples, for each unit in the study sample there will be matching units with similar values of W in the auxiliary sample. The fourth part of Assumption 2.1 allows us to treat the *merged sample*

$$\left\{(D_i, W_i,' (1 - D_i) X_i', D_i Y_i')'\right\}_{i=1}^{N},$$

'as if' it were a random one from a pseudo *merged population* with distribution function F (let $\mathbb{E}[\cdot]$ denote expectations taken with respect to this distribution). The semiparametric data combination model is typically defined by specifying properties of the merged population (e.g., Hahn, 1998; Chen, Hong and Tarozzi, 2008). We prefer the formulation given above because it (i) emphasizes that the problem is fundamentally one of combining two datasets and (ii) in many applications the merged population does not correspond a real

world population. Neither (i) or (ii) are features of standard missing data problems (i.e., Robins, Rotnitzky and Zhao (1994)). We also note that formulating a model by imposing restrictions on a pseudo-population is somewhat awkward (cf., the discussion in Abadie and Imbens (2006, p. 239)).

The sampling distribution induced by the multinomial scheme, F, has density

$$f(z,d) = Q_0^d (1-Q_0)^{1-d} f_s(z)^d f_a(z)^{1-d},$$

such that $f(z|\, d=1) = f_s(z)$ and $f(z|\, d=0) = f_a(z)$. Now consider the conditional probability given $W = w$ that a unit in the merged sample corresponds to a draw from the study population. Let $\mathbb{E}[D|\, W=w] = p_0(w)$ denote this 'propensity score', by Bayes' Law we can define a relationship between the study and auxiliary densities of W in terms of $p_0(w)$

$$f_s(w) = f_a(w) \left\{ \frac{1-Q_0}{Q_0} \frac{p_0(w)}{1-p_0(w)} \right\}. \tag{3}$$

Under the merged population formulation of the problem it is clear that part (i) of Assumption 2.1 corresponds to requiring that $\mathbb{E}[\psi(Z,\gamma_0)|\, D=1] = 0$, part (ii) to conditional independence restrictions on the merged population distribution function of $F(y|\, w, d=1) = F(y|\, w, d=0)$ and $F(x|\, w, d=1) = F(x|\, w, d=0)$, and parts (iii) and (iv) to assuming that $p_0(w)$ is bounded away from one. Part (v) implies that the density ratio $f_s(w)/f_a(w)$ takes a parametric form or, equivalently, that the propensity score is known up to a finite dimensional parameter.

Identification of γ_0 follows from, using parts (ii) and (iii) of Assumption 2.1 and Equation (3), the equality

$$\mathbb{E}_s[\psi(Z,\gamma)] = \mathbb{E}\left[\frac{D}{Q_0}\psi_s(Y,W,\gamma)\right] - \mathbb{E}\left[\frac{1-D}{Q_0}\frac{p_0(W)}{1-p_0(W)}\psi_a(X,W,\gamma)\right], \tag{4}$$

which is, by part (i) of Assumption 2.1, uniquely zero at $\gamma = \gamma_0$. See Lemma 3.1 of Abadie (2005) for a formal proof.

2.1 Example: Two sample instrumental variables (TSIV)

To give some idea of the range of problems to which our methods apply, we elaborate on one common data combination problem in detail: the two sample instrumental variable model (TSIV). This model is widely used by empirical researchers in economics (cf., Inoue and Atsushi, 2010). Our observation that TSIV is a special case of the model defined by Assumption 2.1 is a new one, with empirically relevant implications. In particular, the Auxiliary-to-Study (AST) estimator we propose below is both (i) more efficient and (ii) consistent under a wider, and empirically relevant, set of assumptions, than, for example, the estimators of Angrist and Krueger (1992) and Ridder and Moffitt (2007).

Additional examples of data combination problems are outlined in the Supplemental Web Appendix. Chen, Hong and Tarozzi (2008), Ridder and Moffitt (2007) and Abadie (2005) provide further examples.

Following Ridder and Moffitt (2007), consider two sample instrumental variables (TSIV) models of the form

$$\mathbb{E}_s \left[\{ f(Y;\gamma) - g(X, W_1;\gamma) \} e(W) \right] = 0,$$

with $W = (W_0', W_1')'$. The first sample consists of measurements of (Y, W) and the second of (X, W). They assume that both samples are random ones from the study population (i.e., the samples are 'compatible'). This corresponds to augmenting Assumption 2.1 with the additional requirement that $F_s(w) = F_a(w)$. The TSIV model is of the form required by (2) with $\psi_s(y, w, \gamma) = f(Y;\gamma) e(W)$ and $\psi_a(x, w, \gamma) = g(X, W_1;\gamma) e(W)$. When $e(W) = W$, $f(Y;\gamma) = Y$ and $g(X, W_1;\gamma) = X'\alpha + W_1'\beta$ with $\gamma_0 = (\alpha_0, \beta_0')'$ we have the linear model analyzed by Angrist and Krueger (1992). Ridder and Moffitt (2007) show how one may estimate the Mixed Proportional Hazard (MPH) model under this setup, while Ichimura

and Martinez-Sanchis (2004) discuss binary choice models.

A concrete example of a TSIV problem is provided by the work of Currie and Yelowitz (2000), who consider the model $\mathbb{E}_s\left[W\left(Y - X'\alpha_0 - W_1'\beta_0\right)\right] = 0$, where Y is an indicator for whether a school-aged child has repeated a grade, X an indicator for residence in public housing, W_0 equals the number of male siblings in the household, and W_1 equals the overall number of siblings and also contains other household characteristics; $W = (W_0', W_1')'$. Their interest centers on the causal effect of residence in public housing on human capital acquisition. The number of male siblings changes the probability of residence in public housing since, conditional on the overall number of siblings, families with a mixture of boys and girls qualify for larger units and hence higher (implicit) housing subsidies. Currie and Yelowitz (2000) additionally argue that, conditional on the total number of one's siblings, their gender mix should not influence schooling independently of any effect mediated by exposure to public housing. Hence W_0 may serve as an instrumental variable for X.

Currie and Yelowitz (2000) observe Y and W for a random subsample of children drawn from the US Census. The Census, however, does not collect information on residence in public housing, X. This information is available in the US Current Population Survey (CPS), which also includes measurements of W (but not Y). They treat both the Census and CPS samples as random ones from their study population (school-aged children living in the United States) and use a variant of Angrist and Krueger's (1992) method to estimate $\gamma_0 = (\alpha_0, \beta_0')'$.

In applications of the TSIV model, like Currie and Yelowitz's (2000), it is often found that the sample moments of the common variables W differ significantly across the two datasets being combined (see also Björkland and Jäntti, 1997). This suggests that full compatibility may fail in practice (i.e., $F_s(w) \neq F_a(w)$). The estimator presented below does not require full compatibility and is generally more efficient than the one proposed by Angrist and Krueger (1992) (compare Theorems 3.1 and 3.2 below with Angrist and Krueger (1992, p. 331) or Ridder and Moffitt (2007, p. 5505)).

2.2 Efficiency bound

Hahn (1998, Theorem 1) calculated the semiparametric variance bound for the special case where γ_0 is the ATT and part (v) of Assumption 2.1 is not part of the prior restriction. Chen, Hong and Tarozzi (2008, Theorem 3) include part (v) in their prior, but assume that $\psi_s(Y, W, \gamma) = 0$. The following result generalizes that of Chen, Hong and Tarozzi (2008) to the case where the moment function is of the form given in (2). To present this result we require some additional notation. Let $\mathbb{E}^*[Y|X]$ denote the mean squared error minimizing linear predictor of Y given X and define

$$\Gamma_0(w) = \mathbb{E}\left[\frac{\partial\psi(Z,\gamma_0)}{\partial\gamma'}\middle| W = w\right], \quad \Gamma_0 = \mathbb{E}[\Gamma_0(W)], \quad p_0(w) = G(t(w)'\delta_0)$$

$$q_s(w) = \mathbb{E}[\psi_s(Y, W, \gamma_0)|W = w], \quad q_a(w) = \mathbb{E}[\psi_a(X, W, \gamma_0)|W = w]$$

$$\Sigma_s(w;\gamma_0) = \mathbb{V}(\psi_s(Y, W, \gamma_0)|W = w), \quad \Sigma_a(w;\gamma_0) = \mathbb{V}(\psi_a(X, W, \gamma_0)|W = w)$$

$$\mathbb{S}_\delta = \frac{D - G(r(W)'\delta_0)}{G(r(W)'\delta_0)[1 - G(r(W)'\delta_0)]} G_1(r(W)'\delta_0)\, r(W)$$

with $G_1(v) = \partial G(v)/\partial v$ and

$$\begin{aligned}\Lambda(W) \;=\; & \left(\frac{p_0(W)}{Q_0}\right)^2 \left\{\frac{\Sigma_s(W;\gamma_0)}{p_0(W)} + \frac{\Sigma_a(W;\gamma_0)}{1 - p_0(W)}\right. \\ & \left. + [q_s(W) - q_a(W)][q_s(W) - q_a(W)]'\right\} \\ & + \mathbb{E}\left[\left(\frac{D}{p_0(W)} - 1\right)\frac{p_0(W)\{q_s(W) - q_a(W)\}}{Q_0}\mathbb{S}_\delta'\right] \\ & \times \mathbb{E}[\mathbb{S}_\delta\mathbb{S}_\delta']^{-1}\,\mathbb{E}\left[\left(\frac{D}{p_0(W)} - 1\right)\frac{p_0(W)\{q_s(W) - q_a(W)\}}{Q_0}\mathbb{S}_\delta'\right]'.\end{aligned} \tag{5}$$

Theorem 2.1 (Semiparametric Variance Bound) *Under Assumption 2.1 (i) the maximal asymptotic precision with which γ_0 may be regularly estimated is given by the inverse*

of $\mathcal{I}(\gamma_0) = \Gamma_0' \mathbb{E}[\Lambda(W)]^{-1} \Gamma_0$ *and (ii) the efficient influence function is*

$$\begin{aligned}
\phi^{\text{eff}}(Z, \gamma_0) &= -\Gamma_0^{-1} \times \left[\frac{D}{Q_0} \{\psi_s(Y, W, \gamma_0) - q_s(W)\} \right. \\
&\quad - \frac{1-D}{Q_0} \frac{p_0(W)}{1-p_0(W)} \{\psi_a(X, W, \gamma_0) - q_a(W)\} \\
&\quad + \frac{p_0(W)}{Q_0} \{q_s(W) - q_a(W)\} \\
&\quad \left. + \frac{1}{Q_0} \mathbb{E}^* \left[\left(\frac{D}{p_0(W)} - 1 \right) p_0(W) \{q_s(W) - q_a(W)\} \middle| \mathbb{S}_\delta \right] \right]. \qquad (6)
\end{aligned}$$

Proof. The proof, which involves a modest extension of the analysis of Chen, Hong and Tarozzi (2008, Theorem 3), is in the Supplemental Web Appendix. ■

It is easy to show that the information bound for γ_0 is smaller in the model which leaves $p_0(W)$ nonparametric (i.e., where part (v) of Assumption 2.1 is not part of the prior). Knowledge of the parametric form of the propensity score increases the large sample precision with which γ_0 may be estimated. In contrast, in semiparametric missing data problems it is well-known that parametric restrictions on the propensity score do not shift the efficiency bound (e.g., Robins, Rotnitzky and Zhao, 1994; Hahn, 1998). The value of prior restrictions on the propensity score distinguishes the data combination problem from the missing data one.

To understand this difference, we use the well known result that a biased sample may be combined with a random one to form a more efficient distribution function estimate as long as the biasing function is known or parametrically specified. Parts (v) of Assumption 2.1 implies that we can view the auxiliary sample as a biased sampled from the study population of interest where the biasing function is known up to a finite dimensional parameter (cf., Qin, 1998; Gilbert, Lele and Vardi, 1999; Ridder and Moffitt, 2007).

Here, and in what follows, we assume without loss of generality that the merged sample is arranged such that its first N_s units correspond to study population draws, and its remaining N_a units to auxiliary sample draws. Let $G(r(w)'\widehat{\delta}_{ML})$ denote the conditional maximum

likelihood estimate of the propensity score (based on the merged sample), then the estimate

$$\widehat{F}_s^{\text{eff}}(w) = \sum_{i=1}^{N} \widehat{\pi}_i^{\text{eff}} \mathbf{1}(W_i \leq w), \quad \widehat{\pi}_i^{\text{eff}} = \frac{G(r(W_i)'\widehat{\delta}_{ML})}{\sum_{i=1}^{N} G(r(W_i)'\widehat{\delta}_{ML})} \tag{7}$$

efficiently uses the information in both the study and auxiliary samples to estimate $F_s(w)$. To understand (7) note that Bayes' law gives $f_s(W_i) = f(W_i | D_i = 1) = p_0(W_i) f(W_i) / Q_0$; replacing $p_0(W_i)$ and Q_0 with their maximum likelihood estimates, and $f(W_i)$ with the empirical measure of the merged sample, $1/N$, gives $\widehat{f}_s(W_i) = \widehat{\pi}_i^{\text{eff}}$, for $\widehat{\pi}_i^{\text{eff}}$ defined in (7). Equation (7) uses *both* study and auxiliary units – linked via a parametric form for the propensity score – to efficiently estimate $F_s(w)$.

In contrast, in missing data problems the population of interest corresponds to what we have termed the merged population. The most efficient estimate of the merged population distribution function of W is the merged sample empirical distribution function. This is true irrespective of the form of the propensity score. This provides one intuition for why prior knowledge of the form of the propensity score is not valuable in the missing data context (cf., Graham, 2011).

3 Auxiliary-to-Study Tilting

In this section we present our Auxiliary-to-Study Tilting (AST) estimator and characterize its large sample properties under different sets of assumptions. Since the parameter of interest, γ_0, involves integration over the study population distributions of (Y, W) and (X, W), these two distribution functions must be (implicitly) estimated in order to estimate γ_0. The AST estimator utilizes distribution function estimates that share a finite number of moments of W in common with $\widehat{F}_s^{\text{eff}}(w)$. That is we calibrate our estimates of the study population distributions of (Y, W) and (X, W) to features of (7) (which is a semiparametrically efficient estimate of $F_s(w)$ when the propensity score takes a parametric form). This, as we explain below, is the source of the efficiency gains associated with our procedure.

The idea of calibrating a distribution function estimate to information garnered from auxiliary sources arises in other contexts. Little and Wu (1991) discuss contingency table calibration to known margins and provide historical references (cf., Hellerstein and Imbens, 1999). Bickel, Ya'Acov and Wellner (1991) study estimation of linear functionals of probability measures with known marginals. Hirano, Imbens, Ridder and Rubin (2001) show how calibration to marginal information from refreshment samples may be used to correct for certain types of nonignorable attrition in panel data. In the context of average treatment effect estimation, Tan (2006) calibrates estimates of the two potential outcome distributions to features of the empirical distribution of always observed variables (cf., Qin and Zhang, 2007; Graham, Pinto and Egel, 2012). Recently Cheng, Small, Tan, and Ten Have (2009) apply related ideas to an instrumental variables model.

3.1 Outline of the AST estimator

Our estimator for γ_0, which we call the auxiliary-to-study tilting (AST) estimator, is a sequential method of moments estimator. In the first step we estimate the propensity score parameter δ by conditional maximum likelihood:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i - G\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)}{G\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)\left[1-G\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)\right]}G_1\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)r\left(W_i\right)=0. \tag{8}$$

In the second step we compute a reweighting of both the study and auxiliary samples. Let $t\left(W\right)$ be vector of known linearly independent functions of W with a constant 1 in the first row and λ_a and λ_s be 'tilting' parameters of the same dimension. We allow for $r\left(W\right)$ and $t\left(W\right)$ to include common elements or even coincide. Fixing δ at $\widehat{\delta}_{ML}$ and Q at $\widehat{Q}_{ML}$ we choose $\widehat{\lambda}_a$ to solve:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1-D_i}{1-G\left(r\left(W_i\right)'\widehat{\delta}_{ML}+t\left(W_i\right)'\widehat{\lambda}_a\right)}-1\right)\frac{G\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)}{\widehat{Q}_{ML}}t\left(W_i\right)=0. \tag{9}$$

To understand this method of choosing $\widehat{\lambda}_a$ its helpful to rearrange (9) to get

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}\frac{1-D_i}{\widehat{Q}_{ML}}\frac{G\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)t\left(W_i\right)}{1-G\left(r\left(W_i\right)'\widehat{\delta}_{ML}+t\left(W_i\right)'\widehat{\lambda}_a\right)} &= \frac{1}{N}\sum_{i=1}^{N}\frac{G\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)t\left(W_i\right)}{\widehat{Q}_{ML}} \quad (10) \\ \sum_{i=N_s+1}^{N}\widehat{\pi}_i^a t\left(W_i\right) &= \sum_{i=1}^{N}\widehat{\pi}_i^{\text{eff}} t\left(W_i\right), \end{aligned}$$

for

$$\widehat{\pi}_i^a = \frac{G\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)}{\sum_{i=1}^{N}G(r(W_i)'\widehat{\delta}_{ML})}\frac{1}{1-G\left(r\left(W_i\right)'\widehat{\delta}_{ML}+t\left(W_i\right)'\widehat{\lambda}_a\right)}, \quad i=N_s+1,\ldots,N,$$

and where the second line of (10) is equivalent to the first. The term to the right of the equality in (10) is an estimate of $\mathbb{E}_s\left[t\left(W_i\right)\right]$ – the study population mean of $t\left(W_i\right)$ – based on the *efficient* distribution function estimate (7). It is consequently an efficient estimate of $\mathbb{E}_s\left[t\left(W_i\right)\right]$. The solution to (9) – our estimate of λ_a – is chosen to form a reweighting of the auxiliary sample such that $\sum_{i=1}^{N_s}\widehat{\pi}_i^a t\left(W_i\right)$ is *numerically identical* to the efficient estimate of $\mathbb{E}_s\left[t\left(W_i\right)\right]$ based on $\widehat{F}_s^{\text{eff}}\left(w\right)$.

To better understand (10) recall that, as shown by Abadie (2005) and others, the propensity score reweighting type estimator

$$\widehat{F}_s^{\text{PSR}}\left(x,w\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{1-D_i}{\widehat{Q}_{ML}}\frac{G\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)}{1-G\left(r\left(W_i\right)'\widehat{\delta}_{ML}\right)}\mathbf{1}\left(X_i\leq x, W_i\leq w\right),$$

is consistent for the study population distribution function of (X,W). Our AST estimator replaces $\widehat{F}_s^{\text{PSR}}\left(x,w\right)$ with the more efficient tilted version

$$\widehat{F}_s^{\text{AST}}\left(x,w\right) = \sum_{i=N_s+1}^{N}\widehat{\pi}_i^a\mathbf{1}\left(X_i\leq x, W_i\leq w\right).$$

This tilted distribution estimate, unlike $\widehat{F}_s^{\text{PSR}}\left(x,w\right)$, is guaranteed to integrate to one and

shares a finite number of moment in common with $\widehat{F}_s^{\text{eff}}(w)$.

We also compute an analogous tilt of the study sample

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i}{G\left(r(W_i)'\widehat{\delta}_{ML}+t(W_i)'\widehat{\lambda}_s\right)}-1\right)\frac{G\left(r(W_i)'\widehat{\delta}_{ML}\right)}{\widehat{Q}_{ML}}t(W_i)=0, \tag{11}$$

so that

$$\sum_{i=1}^{N_s}\widehat{\pi}_i^s t(W_i)=\sum_{i=1}^{N}\widehat{\pi}_i^{\text{eff}}t(W_i), \tag{12}$$

for

$$\widehat{\pi}_i^s=\frac{G\left(r(W_i)'\widehat{\delta}_{ML}\right)}{\sum_{i=1}^{N}G(r(W_i)'\widehat{\delta}_{ML})}\frac{1}{G\left(r(W_i)'\widehat{\delta}_{ML}+t(W_i)'\widehat{\lambda}_s\right)},\quad i=1,\ldots,N_s.$$

With the auxiliary and study sample tilts in hand we then choose $\widehat{\gamma}_{AST}$ to solve, holding $\widehat{\lambda}_a$ and $\widehat{\lambda}_s$ fixed at their second step values,

$$\sum_{i=1}^{N_s}\widehat{\pi}_i^s\psi_s(Y_i,W_i,\widehat{\gamma}_{AST})-\sum_{i=N_s+1}^{N}\widehat{\pi}_i^a\psi_a(X_i,W_i,\widehat{\gamma}_{AST})=0. \tag{13}$$

Inspection of (13) indicates that our estimate of γ_0 is based on two separate estimates of the study population distribution function. The first, corresponding to the *study tilt* $\{\widehat{\pi}_i^s\}_{i=1}^{N_s}$ is an estimate of the study population distribution of (Y_i,W_i), the second, corresponding to the *auxiliary tilt*, $\{\widehat{\pi}_i^a\}_{i=N_s+1}^{N}$, is an estimate of the study population distribution of the (X_i,W_i). Neither of these two estimates coincide with the efficient estimate of the study population distribution of W_i alone (i.e, with (7)), but they do share important features with it. Specifically they are constructed so that the means of $t(W_i)$, computed using the two tilts, coincide with the efficient estimate.

3.2 Large sample properties

Our next three results provide formal descriptions of the asymptotic sampling properties of $\widehat{\gamma}_{AST}$ under different combinations of assumptions. We begin with a characterization of the

sampling properties of $\sqrt{N}\left(\widehat{\gamma}_{AST}-\gamma_0\right)$ under our baseline model (i.e., Assumption 2.1). We then outline our local semiparametric efficiency and double robustness results.

To state our first result we require some additional notation. Let

$$q_a^*\left(W\right)=\Pi_a^* t\left(W\right),\quad q_s^*\left(W\right)=\Pi_s^* t\left(W\right),$$

be weighted projections of $\psi_s\left(Y,W,\gamma_0\right)$ and $\psi_a\left(X,W,\gamma_0\right)$ onto the space spanned by $t\left(W\right)$, with projection coefficients of

$$\begin{aligned}
\Pi_s^* &= \mathbb{E}\left[G_1\left(r\left(W\right)'\delta_0\right)q_s\left(W\right)t\left(W\right)'\right] \\
&\quad\times\mathbb{E}\left[G_1\left(r\left(W\right)'\delta_0\right)t\left(W\right)t\left(W\right)'\right]^{-1}, \\
\Pi_a^* &= \mathbb{E}\left[\frac{p_0\left(W\right)}{1-p_0\left(W\right)}G_1\left(r\left(W\right)'\delta_0\right)q_a\left(W\right)t\left(W\right)'\right] \\
&\quad\times\mathbb{E}\left[\frac{p_0\left(W\right)}{1-p_0\left(W\right)}G_1\left(r\left(W\right)'\delta_0\right)t\left(W\right)t\left(W\right)'\right]^{-1}. \qquad (14)
\end{aligned}$$

Also define $R\left(D,W\right)=R_s\left(D,W\right)-R_a\left(D,W\right)$ with

$$\begin{aligned}
R_s\left(D,W\right) &= \frac{1}{Q_0}\left\{\left(\frac{D}{p_0\left(W\right)}-1\right)p_0\left(W\right)\left\{q_s^*\left(W\right)-q_s\left(W\right)\right\}\right. \qquad (15) \\
&\quad\left.-\mathbb{E}^*\left[\left(\frac{D}{p_0\left(W\right)}-1\right)p_0\left(W\right)\left\{q_s^*\left(W\right)-q_s\left(W\right)\right\}\middle|\mathbb{S}_\delta\right]\right\} \\
R_a\left(D,W\right) &= \frac{1}{Q_0}\left\{\left(\frac{1-D}{1-p_0\left(W\right)}-1\right)p_0\left(W\right)\left\{q_a^*\left(W\right)-q_a\left(W\right)\right\}\right. \qquad (16) \\
&\quad\left.-\mathbb{E}^*\left[\left(\frac{1-D}{1-p_0\left(W\right)}-1\right)p_0\left(W\right)\left\{q_a^*\left(W\right)-q_a\left(W\right)\right\}\middle|\mathbb{S}_\delta\right]\right\}.
\end{aligned}$$

Theorem 3.1 (Asymptotic Distribution) *Suppose that Assumption 2.1 and additional regularity conditions hold, then (i)* $\widehat{\gamma}_{AST}\xrightarrow{p}\gamma_0$, *(ii)*

$$\sqrt{N}\left(\widehat{\gamma}_{AST}-\gamma_0\right)\xrightarrow{D}\mathcal{N}\left(0,I\left(\gamma_0\right)^{-1}+\Sigma_{RR}\left(\gamma_0\right)\right),$$

with

$$\Sigma_{RR}(\gamma_0) = \Gamma_0^{-1}\mathbb{E}\left[R(D,W)\,R(D,W)'\right]\Gamma_0^{-1\prime}$$

and (iii) the asymptotic efficiency of $\sqrt{N}a'(\widehat{\gamma}_{AST} - \gamma_0)$*, for any vector of constants* a*, is bounded below by*

$$\text{ae}\left(\sqrt{N}a'(\widehat{\gamma}_{AST} - \gamma_0)\right) \geq \frac{a'I(\gamma_0)^{-1}a}{a'I(\gamma_0)^{-1}a + \frac{\epsilon^2}{Q_0^2}\mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)}\right]a'\left(\Gamma_0^{-1}\iota\right)\left(\Gamma_0^{-1}\iota\right)'a} \tag{17}$$

where $\epsilon = \max(\epsilon_s, \epsilon_a)$ *with*

$$\epsilon_s = \sup_{w\in\mathcal{W}}\left\|q_s(w) - \Pi_s^* t(w)\right\|_\infty, \quad \epsilon_a = \sup_{w\in\mathcal{W}}\left\|q_a(w) - \Pi_a^* t(w)\right\|_\infty,$$

and $\|\cdot\|_\infty$ *denoting the maximum absolute row sum norm.*

Theorem 3.1 indicates that under Assumption 2.1 our AST estimator is consistent and asymptotically normal, but inefficient relative to a semiparametrically efficient estimator (cf., Theorem 2.1 above). Some insight into the degree of AST's inefficiency is provided by the bound (17). First, the term, ϵ^2, indicates that the AST estimator performs better when $q_s(w)$ and $q_a(w)$ are well-approximated by a linear combination of the elements of $t(w)$. We discuss the nature of this approximation further below. Second, the performance of the AST estimator will, in general, be sensitive to the degree of overlap. If the expected value of the propensity score weight, $p_0(W)/(1-p_0(W))$, used to reweight auxiliary units is large, as may be true if κ, the upper bound on $p_0(W)$, is close to one (cf. part (v) of Assumption 2.1), then the performance of the AST estimator may be poor (cf., Khan and Tamer, 2010).

More generally the form of $R(D,W)$ indicates that the relative efficiency of $\widehat{\gamma}_{AST}$ depends on the quality of the linear approximations $q_s(W) \simeq \Pi_s^* t(W)$ and $q_a(W) \simeq \Pi_a^* t(W)$. This is easiest to see in the special case where $r(W) \subset t(W)$, in which case (see (31) in Appendix A), defining $U_s^* = q_s(W) - \Pi_s^* t(W)$, $U_a^* = q_a(W) - \Pi_a^* t(W)$ and $U^* = (1-p_0(W))U_s^* +$

$p_0(W)U_a^*$:

$$\Sigma_{RR}(\gamma_0) = \frac{1}{Q_0^2}\Gamma_0^{-1}\mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)}U^*U^{*\prime}\right]\Gamma_0^{-1\prime},$$

so that the degree of inefficiency depends on a (weighted) expectation of the squares and cross products of a linear combination of the approximation errors. The form of $\Sigma_{RR}(\gamma_0)$ indicates that $\widehat{\gamma}_{AST}$ will have high relative efficiency whenever $q_s(W)$ and $q_a(W)$ are well approximated by a linear combination of the elements of $t(W)$. This will be particularly true when overlap is good such that the weight $\frac{p_0(W)}{1-p_0(W)}$ does not take on extreme values.

Our next result, which characterizes when $\widehat{\gamma}_{AST}$ will be efficient, is anticipated by the discussion above. Consider the assumption:

Assumption 3.1 (MOMENT CEF) *For some unique pair of matrices Π_s, Π_a and vector of linear independent functions $t(W)$ with a constant in the first row, we have*

$$\mathbb{E}[\psi_s(Y,W,\gamma_0)|W] = \Pi_s t(W), \quad \mathbb{E}[\psi_a(X,W,\gamma_0)|W] = \Pi_a t(W).$$

Assumption 3.1 posits a working model for the conditional expectation functions (CEFs) of $\psi_s(Y,W,\gamma_0)$ and $\psi_a(X,W,\gamma_0)$ given W. The substantive content of this assumption is, of course, model and application specific. The ATT example discussed in the introduction provides a simple illustration. In that case Assumption 3.1 implies that the CEFs of the potential outcomes given active and control treatment, Y and X, are linear in $t(W)$. Thus, if the object of interest is the ATT, the analyst should pick the elements of $t(W)$ so as to provide a good approximation to these two CEFs. For the two sample instrumental variables (TSIV) model it is possible to show that the correct $t(W)$ is an implication of the structure of the first stage relationship between the endogenous right hand side variable, X, and the instrument vector, W.

If both Assumptions 2.1 and 3.1 hold the Appendix shows that $\widehat{\gamma}_{AST}$ is asymptotically

linear with representation

$$\sqrt{N}\left(\widehat{\gamma}_{AST}-\gamma_0\right)=\frac{1}{\sqrt{N}}\sum\nolimits_{i=1}^{N}\phi^{\text{eff}}\left(Z_i,\gamma_0\right)+o_p\left(1\right)$$

from which our next Theorem directly follows.

Theorem 3.2 (Local Semiparametric Efficiency) *Suppose that Assumption 2.1 and additional regularity conditions hold, then for $\widehat{\gamma}_{AST}$ the solution to (13), $\widehat{\gamma}_{AST}$ is locally efficient at Assumption 3.1 such that $\sqrt{N}\left(\widehat{\gamma}_{AST}-\gamma_0\right)\xrightarrow{D}\mathcal{N}\left(0,\mathcal{I}\left(\gamma_0\right)^{-1}\right)$ with $\mathcal{I}\left(\gamma_0\right)$ as defined in Theorem 2.1.*

Proof. See Appendix A. ■

Our efficiency bound calculation, Theorem 2.1, gives the information bound for γ_0 without imposing the additional auxiliary Assumption 3.1. 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 may be possible to estimate γ_0 more precisely. Our estimator is not efficient with respect to this augmented model. Rather it attains the bound provided by Theorem 2.1 if Assumption 3.1 *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 $\widehat{\gamma}_{AST}$ is locally efficient at Assumption 3.1. The form of the variance bound when semiparametric, or parametric (as in Assumption 3.1), restrictions on $q_s\left(w\right)$ and $q_a\left(w\right)$ are maintained as part of the prior restriction is unknown. Graham (2011) studies such restrictions in the missing data context.

Next we give our double robustness result. Here our result is slightly less general than similar results in the missing data literature, but nevertheless may be useful in practice.

Theorem 3.3 (Double Robustness) *Under parts (i) to (iv) of Assumption 2.1, $\widehat{\gamma}_{AST}\xrightarrow{p}$*

γ_0 with a limiting normal distribution if ***either*** *(a) part (v) of Assumption 2.1 also holds* ***or*** *(b) the analyst chooses* $G(v) = \frac{\exp(v)}{1+\exp(v)}$ *and Assumption 3.1 holds.*

Proof. See Appendix A. ■

Theorem 3.3 indicates that the advantage of choosing $t(W)$ with Assumption 3.1 in mind is twofold. Under the baseline model defined by Assumption 2.1, Theorem 3.2 implies that $\widehat{\gamma}_{AST}$ will have low sampling variation if $q_s(w) = \mathbb{E}[\psi_s(Y, W, \gamma_0)|\, W = w]$ and $q_a(w) = \mathbb{E}[\psi_a(X, W, \gamma_0)|\, W = w]$ are approximately linear in $t(w)$ (see also part (iii) of Theorem 3.1). This is the case covered by part (a) of the Theorem. Now consider the case where the analyst misspecifies the propensity score model, but Assumption 3.1 holds, part (b) of Theorem 3.3 indicates that $\widehat{\gamma}_{AST}$ will remain consistent for γ_0 in this case if the analyst chooses $G(v)$ to take the logit form. We emphasize that the true propensity score model may or may not be of the logit form.

The peculiar feature of Theorem 3.3, relative to analogous results in the missing data literature (e.g., Tsiatis, 2006), is the requirement that the assumed propensity score take the logit form. To understand this requirement note that, in general, (7) will be an inconsistent estimate of the study population distribution of W when the propensity score is misspecified. Calibrating the study and auxiliary tilts to moments of this distribution will therefore typically produce an inconsistent estimate of γ_0. However when condition (b) of Theorem 3.3 holds we have, from the estimating equations for the propensity score parameter,

$$\frac{1}{N}\sum_{i=1}^{N}\left(D_i - G\left(t(W_i)'\widehat{\delta}_{ML}\right)\right)t(W_i) = 0. \tag{18}$$

Now consider the mean of $t(W_i)$ with respect to $\widehat{F}_s^{\text{eff}}(w)$. Using (18), and the fact that $t(W_i)$ contains a constant so that $\sum_{i=1}^{N} G(t(W_i)'\widehat{\delta}_{ML}) = \sum_{i=1}^{N} D_i$, we have the equalities

$$\sum_{i=1}^{N}\widehat{\pi}_i^{\text{eff}}t(W_i) = \sum_{i=1}^{N}\frac{G(t(W_i)'\widehat{\delta}_{ML})}{\sum_{j=1}^{N}G(t(W_j)'\widehat{\delta}_{ML})}t(W_i) = \frac{\sum_{i=1}^{N}D_i t(W_i)}{\sum_{i=1}^{N}D_i}.$$

Therefore, under the conditions of part (b) of Theorem 3.3, $\sum_{i=1}^{N} \widehat{\pi}_i^{\text{eff}} t(W_i) \xrightarrow{p} \mathbb{E}_s[t(W)]$ irrespective of whether the propensity score is correctly model. This implies that the study and auxiliary tilts will be correctly calibrated such that, when Assumption 3.1 holds, $\widehat{\gamma}_{AST}$ will remain consistent for γ_0. Note that this estimate of $\mathbb{E}_s[t(W)]$ will not be efficient when the propensity score is misspecified.

Although the propensity score is not ancillary in the data combination problem, our estimator remains consistent in the presence of propensity score misspecification when $G(v)$ takes the logit form. It is an open question where there exist a locally efficient *and* doubly robust estimator under non-logit parametric forms for the propensity score.

The alternative estimator, which replaces maximum likelihood (ML) propensity score fit computed in the first step of our procedure with the method of moments (MM) one

$$\frac{1}{N}\sum_{i=1}^{N}\left(D_i - G\left(t(W_i)'\widehat{\delta}_{MM}\right)\right)t(W_i) = 0,$$

will be double robust but not locally efficient (unless a logit form for $G(v)$ is maintained as part of Assumption 2.1, in which case the ML and MM fits coincide). More generally there is a tension between efficiency, which requires using the MLE of the propensity score for reweighting, and robustness to propensity score misspecification.

Implications for practitioners Collectively Theorems 3.1 to 3.3 suggest several useful guidelines for empirical researchers. First, when overlap is good, or equivalently the propensity score weights $p_0(W)/(1-p_0(W))$ do not take very large values, Theorems 3.1 to 3.3 provide a very strong theoretical case for using AST in practice. If Assumption 3.1 happens to be true in the sampled populations, then AST will be more efficient than the propensity score reweighting approach of Abadie (2005). This result is analogous to the enhanced efficiency of the Augmented Inverse Probability Weighting (AIPW) estimator of Robins, Rotnitzky and Zhao (1994) relative to conventional Inverse Probability Weighting (IPW)

in the missing data context. In practice high levels of precision will be observed whenever $q_s(w)$ and $q_a(w)$ are reasonably well approximated by a linear combination of the elements of $t(w)$. A further advantage of the AST procedure is that, if the propensity score is inadvertently misspecified, AST will nevertheless remain consistent for γ_0 if Assumption 3.1 holds (and the analyst works with a logit form for $G(v)$).

In settings with poor overlap, the AST estimator may be highly variable and, in extreme cases, may not even exist. To understand this last observation consider the case where $G(v)$ takes the logit form. In that case the computation of the auxiliary tilt requires that the study sample mean of $t(W)$ lie within the convex hull of the auxiliary sample. If the study and auxiliary distributions of W are very different from one another, this convex hull condition may fail in practice even if Assumption 2.1 holds in the population. We do not view this as a weakness of our procedure, rather such situations alert the researcher to the fragility of identification when overlap is poor (cf., Khan and Tamer, 2010). When overlap is poor direct imputation approach may be preferable (e.g., Kline, 2011; Chen, Hong and Tarozzi, 2008). However imputation will be very sensitive to violations of Assumption 3.1; this limitation is illustrated by our Monte Carlo experiments below.

The computational algorithm detailed in the Supplemental Web Appendix is designed to work well in situations where the convex hull condition is "nearly" violated and we recommend its routine use. For covariance matrix estimation we recommend use the textbook formulae for the GMM estimator based on the moment vector implied by (8), (9), (11) and (13) above and explicitly defined in the Appendix.

4 Application and Monte Carlo experiments

Empirical application Neal and Johnson (1996) study the role of 'pre-market' (i.e., acquired prior to age 18) differences in cognitive achievement in explaining differences in earnings between young Black and White men. Using a sample of employed Black and White

males drawn from the National Longitudinal Survey of Youth 1979 (NLSY79), Neal and Johnson (1996) compute the least squares fit of the logarithm of hourly wages on a constant, a black dummy, age, and Armed Forces Qualification Test (AFQT) percentile score measured at age 16 to 18. They find that the coefficient on the black dummy variable drops by two thirds to three quarters when AFQT score is included as a covariate. On the basis of this finding they argue that differences in the rate of cognitive skill acquisition across Blacks and White prior to age 18, due to differences in family background, school quality and neighborhood characteristics, explains a substantial portion of subsequent Black-White wage inequality. We do not provide an assessment of this interpretation here, rather our goal is to illustrate the use of AST in a familiar setting.

Let Y denote real average wages from 1990 to 1993 for a random draw from the population of Black men aged 16 to 18 in 1979 and residing in the United States. This population corresponds to our study population of interest. Let X denote real wages for a random draw from the population of White men aged 16 to 18 in 1979 and residing in the United States. This corresponds to our auxiliary population. Let W be a vector including year of birth and AFQT score (We transform the percentile scores used by Neal and Johnson (1996) onto the real line using the inverse standard normal CDF). We compare features of the observed distribution of Black wages with those of a hypothetical White population whose age and AFQT distribution *coincides* with that of the Blacks (i.e., with study population's). These types of hypothetical comparisons underlie Oaxaca decompositions, as used in labor and health economics, and similar exercises undertaken in demography (e.g., Kitagawa, 1964). Barsky, Bound, Charles and Lupton (2002) and Fortin, Lemieux and Firpo (2010) survey the application of decomposition methods in economics.

Our sample closely resembles that used in Johnson and Neal (1998). It includes 1,371 measurements of real wages, race, age and AFQT scores drawn from the NLSY79. Throughout we replace the empirical measure of our sample with the NLSY79 base year sampling weights (although this adjustment has little effect on our results). The age distributions

for Blacks and Whites in the merged sample are, as would be expected, quite similar. The distribution of AFQT scores across the two groups are quite different. The mean Black score is approximately one standard deviation lower than the mean White score. The two distributions also substantially differ in their second, third and fourth moments (not reported).

Panel A of Table 1 reports estimates of mean log Wages for Blacks (Column 1), as well as the Black-White average difference (Column 2). On average, Blacks earn almost 28 percent less per hour than Whites in our sample. Panel A also reports estimates of the CDF of the Black wage distribution at selected points, and the corresponding Black-White CDF differences. For example, while over 45 percent of Blacks earn less than \$7.50 per hour in our sample, fewer than 30 percent of Whites do (Table 1, Row 3). Inspection of the CDF differences indicates that, while the distributions are most different at the lower wage levels, differences exist across the entire support of wages.

Panel B of Table 1 reports average wage differences between Blacks and a hypothetical population of Whites whose distribution of age and AFQT score coincides with the Black distribution. This allows for a comparison between Black and White wages that flexibly controls for differences between the two populations in age and AFQT score.

In Column 1 of Panel B we report age- and AFQT-adjusted differences in mean wages and wage CDFs based on the conditional expectation projection (CEP) estimator of Chen, Hong, and Tarozzi (2008). Our implementation of their procedure models the conditional expectation functions (CEFs) of Y and X given W as a separable functions of a constant, two year of birth dummies, a quadratic polynomial in transformed AFQT score, and twelve dummy variables for the transformed AFQT score lying respectively below $-2, -1.75, \ldots, 0.25, 0.5$. Let $t(W)$ be the vector containing all these functions of W. In principle, if the dimension of the approximating model is allowed to grow with the sample size, the Chen, Hong, and Tarozzi (2008) estimator is consistent for, and efficient under, all data generating processes satisfying parts (i) to (iv) of Assumption 2.1. In small samples the performance of the estimator is heavily dependent on the quality of the two CEF approximations.

Table 1: Raw and adjusted differences in Black versus White hourly wages

	Panel A		**Panel B**		
	(1)	(2)	(1)	(2)	(3)
	Black	**B − W**	**CEP**	**PSR**	**AST**
Average (log(Wage))	6.749	−0.279	−0.1108	−0.1072	−0.1052
	(0.021)	(0.026)	(0.0348)	(0.0303)	(0.0298)
Pr (Wage ≤ \$5.00)	0.0801	0.0566	0.0243	0.0246	0.0278
	(0.0125)	(0.0135)	(0.0216)	(0.0193)	(0.0187)
Pr (Wage ≤ \$7.50)	0.4505	0.2948	0.1780	0.1737	0.1757
	(0.0244)	(0.0275)	(0.0391)	(0.0355)	(0.0350)
Pr (Wage ≤ \$10.00)	0.6590	0.2691	0.0987	0.0964	0.0903
	(0.0244)	(0.0300)	(0.0406)	(0.0358)	(0.0353)
Pr (Wage ≤ \$12.50)	0.8020	0.2001	0.0417	0.0386	0.0348
	(0.0198)	(0.0265)	(0.0328)	(0.0288)	(0.0284)
Pr (Wage ≤ \$15.00)	0.8896	0.1426	0.0176	0.0129	0.0109
	(0.0153)	(0.0219)	(0.0238)	(0.0203)	(0.0202)

NOTES: Results based on an extract of 1,371 Black and White men ages 16 to 18 in 1979 from the NLSY79. Estimated standard errors, which account for within-household dependence in outcomes across siblings, are reported in parentheses.

Column 2 of Panel B implements the propensity score reweighting (PSR) estimator of Hirano and Imbens (2001) and Abadie (2005). We model the propensity score as a logit function with an index linear in $t(W)$ as defined above for the CEP estimator. The PSR estimates are very close in magnitude and precision to the CEP estimates.

Column 3 of Panel B implements our AST procedure using the same choice of $t(W)$ and $r(W) = t(W)$. This choice ensures that the study and auxiliary sample tilts share the following features with the efficient distribution function estimate of W: (i) the marginal year of birth distributions coincide, (ii) the means and variances of the transformed AFQT score coincide, (iii) the probability masses assigned to the intervals defined by the $-2, -1.75, \ldots, 0.25, 0.5$ grid of AFQT score intervals coincide. Figure 1 plots undersmoothed kernel density estimates of the actual Black and White AFQT score densities; the two distributions are very different from one another. The figure also plots a density estimate based on the auxiliary sample tilt. This corresponds to the AFQT score density in the hypothetical comparison population of Whites. As is evident from the figure, our choice of $t(W)$ is rich

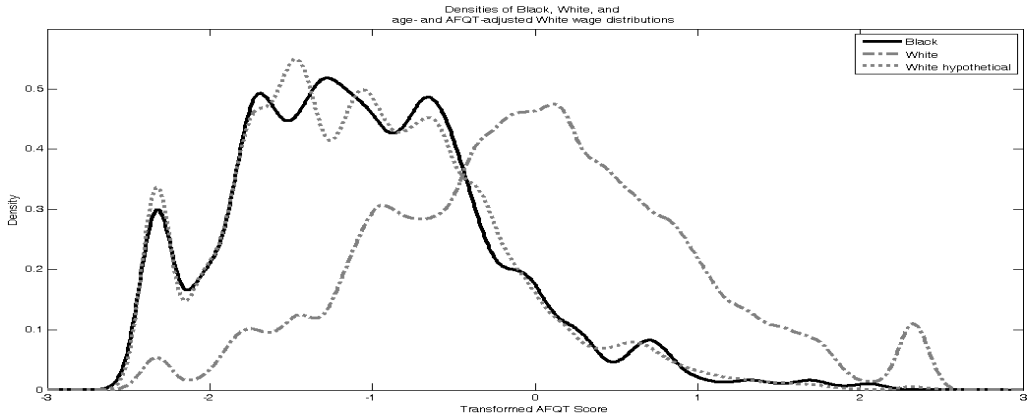

Figure 1: AFQT Densities
NOTES: The figure plots kernel density estimates of the actual Black and White AFQT score distributions as well as an estimate based on the auxiliary sample tilt. A Gaussian kernel is used with a bandwidth equal to 1/2 of Silverman's 'rule-of-thumb' choice. Undersmoothing highlights the ability of the auxiliary tilt to match local features of the Black AQFT density.

enough to closely match this density with its target Black one.

After adjusting for age and AFQT differences we find that, while a Black-White residual log wage CDF gap remains at middle parts of the wage distribution, it disappears at the low and high ends of this distribution. The average log wage gaps falls, after adjusting for age and AFQT differences, from -0.279 to -0.111.

While the AST point estimates are similar to the corresponding CEP and PSR ones, their estimated sampling precision is uniformly superior (as Theorem 3.2 would suggest). The close correspondence between the CEP, PSR and AST point estimates in our application likely reflects a combination of two factors. First, while the AFQT distributions across Blacks and Whites differ dramatically, the support of the Black distribution is clearly contained within that of the White distribution. Hence part (iii) of Assumption 2.1 is well satisfied. Second the approximating models underlying each of the estimators are quite flexible. In settings where overlap is weaker, and/or the approximating models more parsimonious (as would be required when the dimension of W is large), we would expect the three estimators to more often yield different point estimates depending on the true data generating process.

Table 2: Parameter values for Monte Carlo experiments

Design	(1)	(2)	(3)	(4)
σ_a^2	1	2/3	1	2/3
σ_Y^2	3.4823	2.6590	1.7496	0.9253
α_2	0	0	-1	-1

Monte Carlo We now report on a number of Monte Carlo experiments we conducted to verify the theoretical properties described in Theorems 3.1 to 3.3. In particular we wish to assess the relevance of our theoretical robustness and efficiency results. To do this we consider a stylized program evaluation setting. The analyst wishes to estimate the average treatment effect on the treated (ATT).

In each of our first set of experiments we assume that W is distributed according to a truncated normal distribution, with support $[-c, c]$, in both the study (treated) and auxiliary (control) populations. The location and scale parameters of these two distributions, respectively (μ_s, σ_s^2) and (μ_a, σ_a^2), may differ. We assume a multinomial sampling scheme: with probability $Q_0 = 1/2$ a draw of (Y, W) is taken at random from the study (treated) population, otherwise a draw of (X, W) is taken from the auxiliary (control) population. Finally we assume that Y and X, which play the roles of the outcome under treatment and control, are generated according to

$$\begin{aligned} Y \,|\, W, D &\sim \mathcal{N}\left(0, \sigma_Y^2\right) \\ X \,|\, W, D &\sim \mathcal{N}\left(\alpha_0 + \alpha_1 \left(W - \mu_{W|D=1}\right) + \alpha_2 \left[\left(W - \mu_{W|D=1}\right)^2 - \sigma_{W|D=1}^2\right], \sigma_X^2\right), \end{aligned}$$

where $\mu_{W|D=1}$ and $\sigma_{W|D=1}^2$ are the study population mean and variance of W (which differ from μ_s and σ_s^2 due to truncation).

The target parameter is $\gamma_0 = \mathbb{E}_s[Y - X] = \alpha_0$. The propensity score induced by these designs is of the logit form with an index quadratic in W :

$$p_0(w) = \left[1 + \exp\left(-\beta_0 - \beta_1 W - \beta_2 W^2\right)\right]^{-1},$$

where β_0, β_1 and β_2 are functions of (μ_s, σ_s^2) and (μ_a, σ_a^2) (cf., Anderson, 1982). When the study and auxiliary population distributions of W have different means, but a common variance, the logit index will be linear in W. When both the means and variances differ, then the index will generally be nontrivially quadratic in W .

Across all designs we assume a sample size of $N = 1,000$ and set $\mu_s = 0$, $\sigma_S^2 = 1$, $\mu_a = -1/2$, $\alpha_0 = 0$, $\alpha_1 = 1/2$, $\sigma_X^2 = 1$ and $c = 3$. We vary σ_A^2 and α_2 across designs to, respectively, induce nonlinearity in the (index of) the propensity score and $\mathbb{E}\left[\psi_a\left(X, W, \gamma_0\right) | W\right] = q_a\left(W\right)$. We vary σ_Y^2 across designs to keep the variance bound fixed. Across each of our designs an efficient estimator (under Assumption 2.1) will have an asymptotic standard error of $\sqrt{\mathcal{I}\left(\gamma_0\right)^{-1}/1000} = 1/10$.

Table 2 gives the parameter configurations for each of four Monte Carlo designs. In the first design both the propensity score, $p_0\left(w\right)$, and $q_a\left(w\right)$ are 'linear' in w (for $p_0\left(w\right)$ 'linear' means linear in the logit index). In the second design the propensity score is quadratic in w, while $q_a\left(w\right)$ remains linear. In Design three the reverse is true, while in Design four both objects are 'quadratic'. Across each design we implement the AST estimator with $G\left(\cdot\right)$ being the logit function and $r\left(W\right) = t\left(W\right) = \left(1, W\right)'$. For the conditional expectation projection (CEP) estimator we proceed 'as if' $\mathbb{E}\left[X | W\right]$ were linear in W, while our implementation of propensity score reweighting (PSR) uses a logit propensity score with a linear index.

Our AST estimator is consistent for γ_0 in designs 1 through 3. CEP is consistent in designs 1 and 2, but inconsistent in design 3. The PSR estimator is consistent in designs 1 and 3, but inconsistent in design 2. All estimators are inconsistent in design 4 due to the nonlinearity of both $p_0\left(w\right)$ and $q_a\left(w\right)$. Table 3 reports the results of our experiments. Column 1 lists a 'pencil and paper' asymptotic bias calculation, while Column 2 gives the median bias across 5,000 Monte Carlo replications (in both cases bias is scaled by the 'pencil and paper' asymptotic standard error reported in Column 3). As predicted, AST is median unbiased (up to simulation error) in designs 1 through 3. In contrast, PSR is severely biased in design 2 and CEP in design 3. As expected, all estimators perform poorly in design

Table 3: Monte Carlo results

	(1) Asym. Bias	(2) Med. Bias	(3) Asym. SE.	(4) Median SE.	(5) Std. Dev.	(6) Cov. of 95% CI	(7) RMSE
Design 1: $p_0(w)$ linear, $q_a(w)$ linear							
CEP	0.0000	0.0097	0.0997	0.0996	0.0986	0.9526	0.0986
PSR	0.0000	0.0164	0.1007	0.1006	0.1005	0.9506	0.1005
AST	0.0000	0.0055	0.0100	0.0998	0.0998	0.9540	0.0997
Design 2: $p_0(w)$ quadratic, $q_a(w)$ linear							
CEP	0.0000	0.0137	0.0925	0.0924	0.0947	0.9480	0.0947
PSR	0.5053	0.5437	0.0905	0.0911	0.0912	0.9126	0.1039
AST	0.0000	0.0169	0.0941	0.0931	0.0941	0.9470	0.0942
Design 3: $p_0(w)$ linear, $q_a(w)$ quadratic							
CEP	-1.6125	-2.0082	0.1309	0.1296	0.1627	0.6204	0.3111
PSR	0.0000	-0.0137	0.1063	0.1037	0.1068	0.9420	0.1068
AST	0.0000	-0.0266	0.1076	0.1054	0.1081	0.9416	0.1081
Design 4: $p_0(w)$ quadratic, $q_a(w)$ quadratic							
CEP	-4.6038	-6.7095	0.1192	0.1157	0.1728	0.0010	0.8196
PSR	-3.0049	-3.1031	0.0847	0.0821	0.0858	0.1694	0.2670
AST	-2.8789	-2.9313	0.0941	0.0873	0.0953	0.1726	0.2908

4. These bias properties are reflected in the coverage of standard, Wald-based, 95 percent confidence intervals for γ_0 (Column 6). By comparing columns 1 and 2 and columns 3 and 5, we see that – for the designs considered here – the finite sample distributions of all of the estimators are very well approximated by their asymptotic counterparts.

5 Summary

When the propensity score is parametrically specified information in both the study and auxiliary samples may be used to form an efficient estimate of W, the variable common to both datasets. An intuition for this insight follows from recognizing that, under part (v) of Assumption 2.1, the auxiliary sample is equivalent to a biased sample from the study population with the biasing function known up to a finite dimensional parameter. Using this efficient distribution function estimate we tilt the propensity score reweighting (study

population) distribution function estimates of (Y, W) and (X, W) so that they share certain moments in common. By choosing these moments carefully (i.e., with reference to Assumption 3.1) we can produce a locally efficient estimate of γ_0. Even if the parametric relationship between the study and auxiliary populations, as embodied in the propensity score model, is misspecified, AST remains consistent for γ_0 if Assumption 3.1 holds.

To our knowledge we are the first to propose a locally efficient estimator for the class of data combination problems defined by Assumption 2.1. Our procedure also has a double robustness property. Our results provide a useful complement to the work of Robins, Rotnitzky and Zhao (1994), Tan (2006) and others for missing data problems. Relative to Chen, Hong and Tarozzi (2008), who do provide explicit results for data combination problems (their so called 'verify-out-of-sample' case), our approach may be useful when W is high dimensional such that their method, which requires nonparametric estimation of $q_s(w)$ and $q_a(w)$, is impractical.

In future work it would be useful to study data dependent methods for choosing $t(W)$. Similarly it would be interesting to construct a locally efficient estimator with minimal variance across all estimators based on the linear approximating models $q_s(W) \simeq \Pi_s t(W)$ and $q_a(W) \simeq \Pi_a t(W)$. In the missing data context such estimators are called "improved locally efficient" (e.g., Tan (2010)).

A Proofs

Proof of Theorem 3.1: The AST procedure coincides with a just identified GMM estimator based on the $\dim(r(W)) + 2\dim(t(W)) + \dim(\gamma_0)$ vector of moment functions $m(Z_i, \theta_0)$

with $\theta = (\delta', \lambda_a', \lambda_s', \gamma')'$. This vector of moment functions is composed of the subvectors:

$$\begin{aligned}
\underset{\dim(r(W))\times 1}{m_1(Z,\delta_0)} &= \frac{D - G\left(r(W)'\delta_0\right)}{G\left(r(W)'\delta_0\right)\left(1 - G\left(r(W)'\delta_0\right)\right)} G_1\left(r(W)'\delta_0\right) r(W) \\
\underset{\dim(t(W))\times 1}{m_2(Z,\delta_0,\lambda_{a0})} &= \frac{1}{Q_0}\left(\frac{1-D}{1 - G\left(r(W)'\delta_0 + t(W)'\lambda_{a0}\right)} - 1\right) G\left(r(W)'\delta_0\right) t(W) \\
\underset{\dim(t(W))\times 1}{m_3(Z,\delta_0,\lambda_{s0})} &= \frac{1}{Q_0}\left(\frac{D}{G\left(r(W)'\delta_0 + t(W)'\lambda_{s0}\right)} - 1\right) G\left(r(W)'\delta_0\right) t(W) \\
\underset{\dim(\gamma_0)\times 1}{m_4(Z,\delta_0,\lambda_{a0},\lambda_{s0},\gamma_0)} &= \frac{D}{Q_0}\frac{G\left(r(W)'\delta_0\right)}{G\left(r(W)'\delta_0 + t(W)'\lambda_{s0}\right)}\psi_s(Y,W,\gamma_0) \\
&\quad -\frac{1-D}{Q_0}\frac{G\left(r(W)'\delta_0\right)}{1 - G\left(r(W)'\delta_0 + t(W)'\lambda_{a0}\right)}\psi_a(X,W,\gamma_0).
\end{aligned}$$

Let $M = \mathbb{E}\left[\partial m(Z,\theta_0)/\partial\theta'\right]$; a standard argument (e.g., Newey and McFadden, 1994) gives, under regularity conditions, the asymptotically linear representation

$$\sqrt{N}\left(\widehat{\theta} - \theta_0\right) = -M^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m(Z_i,\theta_0)\right) + o_p(1). \tag{19}$$

The influence function for $\widehat{\gamma}_{AST}$ corresponds to the last K elements of (19). By tedious, but straightforward, calculation we can show that this subvector equals

$$\begin{aligned}
\sqrt{N}\left(\widehat{\gamma} - \gamma_0\right) &= \frac{-M_{44}^{-1}}{\sqrt{N}}\sum_{i=1}^{N}\Big\{m_4(Z_i,\delta_0,\lambda_{a0},\lambda_{s0},\gamma_0) - M_{41}M_{11}^{-1}m_1(Z_i,\delta_0) \\
&\quad + M_{42}M_{22}^{-1}\left(M_{21}M_{11}^{-1}m_1(Z_i,\delta_0) - m_2(Z_i,\delta_0,\lambda_{a0})\right) \\
&\quad + M_{43}M_{33}^{-1}\left(M_{31}M_{11}^{-1}m_1(Z_i,\delta_0) - m_3(Z_i,\delta_0,\lambda_{s0})\right)\Big\} + o_p(1).
\end{aligned} \tag{20}$$

where M_{kl} equals the expected value of the derivative of the k^{th} subvector of $m(Z,\theta)$ with respect to the l^{th} subvector of θ evaluated at $\theta = \theta_0$.

Under part (v) of Assumption 2.1 the Information Matrix equality gives $M_{11} = -\mathbb{E}\left[\mathbb{S}_\delta\mathbb{S}_\delta'\right]$.

Calculation gives M_{41} equal to

$$M_{41} = \frac{1}{Q_0}\mathbb{E}\left[\frac{1-D}{1-p_0(W)}\psi_a(X,W,\gamma_0)\,\mathbb{S}'_\delta\right]. \tag{21}$$

Using this result, iterated expectations and part (ii) of Assumption 2.1 we then get

$$-M_{41}M_{11}^{-1}m_1(Z,\delta_0) = \frac{1}{Q_0}\mathbb{E}^*\left[\frac{1-D}{1-p(W)}q_a(W)\middle|\mathbb{S}_\delta\right].$$

Evaluating M_{21} yields, after some manipulation,

$$M_{21} = -\frac{1}{Q_0}\mathbb{E}\left[\left(\frac{1-D}{1-p_0(W)}-1\right)p_0(W)\,t(W)\,\mathbb{S}'_\delta\right], \tag{22}$$

where $p_0(W) = G\left(r(W)'\delta_0\right) = G\left(r(W)'\delta_0 + t(W)'\lambda_{a0}\right)$. These results imply that

$$M_{21}M_{11}^{-1}m_1(Z,\delta) = \frac{1}{Q_0}\mathbb{E}^*\left[\left(\frac{1-D}{1-p_0(W)}-1\right)p_0(W)\,t(W)\middle|\mathbb{S}_\delta\right].$$

Similar calculations give

$$M_{31} = -\frac{1}{Q_0}\mathbb{E}\left[\left(\frac{D}{p_0(W)}-1\right)p_0(W)\,t(W)\,\mathbb{S}'_\delta\right], \tag{23}$$

yielding

$$M_{31}M_{11}^{-1}m_1(Z,\delta_0) = \frac{1}{Q_0}\mathbb{E}^*\left[\left(\frac{D}{p_0(W)}-1\right)p_0(W)\,t(W)\middle|\mathbb{S}_\delta\right].$$

Evaluating M_{22} and M_{42} yields

$$M_{22} = \frac{1}{Q_0}\mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)}G_1\left(r(W)'\delta_0\right)t(W)\,t(W)'\right] \tag{24}$$

$$M_{42} = -\frac{1}{Q_0}\mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)}G_1\left(r(W)'\delta_0\right)\psi_a(X,W,\gamma_0)\,t(W)'\right]. \tag{25}$$

Using (24) and (25) and iterated expectations we get

$$\begin{aligned} M_{42}M_{22}^{-1} &= -\mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)}G_1\left(r(W)'\delta_0\right)q_a(W)\,t(W)'\right] \\ &\quad\times\mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)}G_1\left(r(W)'\delta_0\right)t(W)\,t(W)'\right]^{-1} \\ &\quad\stackrel{def}{=} -\Pi_a^*, \end{aligned}$$

as defined in (14) of the main text.

Now consider M_{33} and M_{43}; we have

$$M_{33} = -\frac{1}{Q_0}\mathbb{E}\left[G_1\left(r(W)'\delta_0\right)t(W)\,t(W)'\right] \tag{26}$$

$$M_{43} = -\frac{1}{Q_0}\mathbb{E}\left[G_1\left(r(W)'\delta_0\right)\psi_s(X,W,\gamma_0)\,t(W)'\right]. \tag{27}$$

Using (26) and (27) and iterated expectations we get

$$\begin{aligned} M_{43}M_{33}^{-1} &= \mathbb{E}\left[G_1\left(r(W)'\delta_0\right)q_s(W)\,t(W)'\right] \\ &\quad\times\mathbb{E}\left[G_1\left(r(W)'\delta_0\right)t(W)\,t(W)'\right]^{-1} \stackrel{def}{=} \Pi_s^*, \end{aligned}$$

also as defined in (14) of the main text.

Recalling the definitions, $q_a^*(W) = \Pi_a^* t(W)$ and $q_s^*(W) = \Pi_s^* t(W)$, substituting the expressions derived immediately above into (20), and rearranging, yields the form of the influence function stated in the theorem.

Now recall the definitions of $R_s(D,W)$ and $R_a(D,W)$ given in (15) and (16) of the main text. By the definition of the LP operator we have that $\mathbb{E}[R_a(D,W)\,\mathbb{S}_\delta'] = \mathbb{E}[R_s(D,W)\,\mathbb{S}_\delta'] = 0$. This follows since $R_a(D,W)$ and $R_s(D,W)$ are linear prediction errors, with $\mathbb{S}_\delta$ the vector of predictors. The conditional mean zero property of the score function also yields the restrictions $\mathbb{E}[R_a(D,W)|\,W] = \mathbb{E}[R_s(D,W)|\,W] = 0$. From these properties, and direct

calculation, we have that

$$\mathbb{E}\left[\phi^{\text{eff}}\left(Z,\gamma_0\right)\left\{R_s\left(D,W\right)-R_a\left(D,W\right)\right\}\right]=0$$

from which the claimed form of the asymptotic variance of γ_0 follows.

Let a be a vector of constants. By linearity of the LP operator, the Cauchy-Schwartz inequality, and recalling that $R\left(D,W\right)=R_s\left(D,W\right)-R_a\left(D,W\right)$, we have that

$$\begin{aligned}
&a'\mathbb{V}\left(R\left(D,W\right)\right)a \\
&\qquad\leq a'\mathbb{V}\left(\frac{1}{Q_0}\left(\frac{1-D}{1-p_0\left(W\right)}-1\right)p_0\left(W\right)\left\{q_a^*\left(W\right)-q_a\left(W\right)\right\}\right. \\
&\qquad\qquad\left.+\frac{1}{Q_0}\left(\frac{D}{p_0\left(W\right)}-1\right)p_0\left(W\right)\left\{q_s^*\left(W\right)-q_s\left(W\right)\right\}\right)a.
\end{aligned}$$

This bound will hold with equality if $r\left(W\right)\subset t\left(W\right)$ since, by the definitions of Π_s^* and Π_a^*, we will have (in that case) the zero covariance results

$$\begin{aligned}
\mathbb{E}\left[\left(q_a\left(W\right)-q_a^*\left(W\right)\right)\frac{p_0\left(W\right)}{1-p_0\left(W\right)}G_1\left(r\left(W\right)'\delta_0\right)r\left(W\right)'\right] &= 0 \qquad (28)\\
\mathbb{E}\left[\left\{q_a^*\left(W\right)-q_a\left(W\right)\right\}G_1\left(r\left(W\right)'\delta_0\right)r\left(W\right)'\right] &= 0.
\end{aligned}$$

Iterated expectations and (28) then give

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{D}{p_0\left(W\right)}-1\right)p_0\left(W\right)\left\{q_a^*\left(W\right)-q_a\left(W\right)\right\}\mathbb{S}_\delta'\right] \\
=\;&\mathbb{E}\left[\left\{q_a^*\left(W\right)-q_a\left(W\right)\right\}G_1\left(r\left(W\right)'\delta_0\right)r\left(W\right)'\right]=0 \qquad (29)
\end{aligned}$$

and also

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1-D}{1-p_0\left(W\right)}-1\right)p_0\left(W\right)\left\{q_a^*\left(W\right)-q_a\left(W\right)\right\}\mathbb{S}_\delta'\right] \\
=\;&-\mathbb{E}\left[\left\{q_a^*\left(W\right)-q_a\left(W\right)\right\}\frac{\left(D-p_0\left(W\right)\right)^2}{\left(1-p_0\left(W\right)\right)^2}G_1\left(r\left(W\right)'\delta_0\right)r\left(W\right)'\right]=0. \qquad (30)
\end{aligned}$$

Equations (29) and (30) imply that, if $r(W) \subset t(W)$, then, after manipulation,

$$\mathbb{V}(R(D,W)) = \frac{1}{Q_0^2}\mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)}U^*U^{*\prime}\right], \tag{31}$$

with the right-hand-side of (31) an upper bound otherwise. Recalling the definition of ϵ given in the statement of the Theorem, and making use of the various compact support conditions embedded in Assumption 2.1, we get the bound

$$a'\mathbb{V}(R(D,W))\,a \leq \frac{\epsilon^2}{Q_0^2}\mathbb{E}\left[\frac{p_0(W)}{1-p_0(W)}\right]a'\iota\iota'a,$$

from which (17) follows directly.

Proof of Theorem 3.2: Under Assumption 3.1, we have that $\Pi_s^* = \Pi_s$ and $\Pi_a^* = \Pi_a$. This implies that $R_s(D,W)$ and $R_a(D,W)$ are identically equal to zero. The result then follows directly from Theorem 2.1.

Proof of Theorem 3.3: Asymptotic normality follows from standard results. Consistency under part (a) is a consequence of Equation (4) in the main text. Showing consistency under part (b) is more complicated. Denote the probability limits of $\widehat{\delta}$, $\widehat{\lambda}_a$,and $\widehat{\lambda}_s$ when part (v) of Assumption 2.1 fails to hold by, respectively δ_*, λ_{a*}, and λ_{s*}. Let $p_*(W) = G\left(r(W)'\delta_*\right)$ and $p_j(W) = G\left(r(W)'\delta_* + t(W)'\lambda_{j*}\right)$ for $j = s, a$. If $G(\cdot)$ takes the logit form, then $p_*(W)$ will satisfy the population restriction $\mathbb{E}[m_1(Z,\delta_*)] = \mathbb{E}[(D - p_*(W))\,t(W)] = 0$ so that, using iterated expectations and rearranging, we have the equality.

$$\mathbb{E}[t(W)|\,D=1] = \mathbb{E}\left[\frac{p_*(W)}{Q_0}t(W)\right]. \tag{32}$$

We also have $\mathbb{E}\left[m_2\left(Z,\delta_*,\lambda_{a*}\right)\right]=\mathbb{E}\left[m_3\left(Z,\delta_*,\lambda_{s*}\right)\right]=0$, which, respectively multiplying by Π_a and Π_s (using Assumption 3.1), gives the additional equalities:

$$\mathbb{E}\left[\frac{1-D}{1-p_a\left(W\right)}p_*\left(W\right)q_a\left(W\right)\right] = \mathbb{E}\left[p_*\left(W\right)q_a\left(W\right)\right] \tag{33}$$

$$\mathbb{E}\left[\frac{D}{p_s\left(W\right)}p_*\left(W\right)q_s\left(W\right)\right] = \mathbb{E}\left[p_*\left(W\right)q_s\left(W\right)\right]. \tag{34}$$

Using (32), (33), (34), Assumption 3.1, iterated expectations, and part (ii) of Assumption 2.1 yields

$$\begin{aligned}
\mathbb{E}\left[m_4\left(Z,\delta_*,\lambda_{a*},\lambda_{s*},\gamma\right)\right] &= \mathbb{E}\left[\frac{p_*\left(W\right)}{Q_0}\left\{q_s\left(W\right)-q_a\left(W\right)\right\}\right] \\
&= \left(\Pi_s-\Pi_a\right)\mathbb{E}\left[\frac{p_*\left(W\right)}{Q_0}t\left(W\right)\right] \\
&= \mathbb{E}\left[q_s\left(W\right)-q_a\left(W\right)\middle|\,D=1\right] \\
&= \mathbb{E}\left[\psi\left(Z,\gamma\right)\middle|\,D=1\right],
\end{aligned}$$

which by part (i) of Assumption 2.1 is uniquely zero at $\gamma=\gamma_0$.

References

[1] Abadie, Alberto. (2005). "Semiparametric difference-in-differences," *Review of Economic Studies* 72 (1): 1 - 19.

[2] Abadie, Alberto and Guido W. Imbens. (2006). "Large sample properties of matching estimators for average treatment effects," *Econometrica* 74 (1): 235 - 267.

[3] Anderson, J.A. (1982). "Logistic discrimination," *Handbook of Statistics* 2: 169 - 191 (P.R. Krishnaiah & L.N. Kanal, Eds.). Amsterdam: North-Holland.

[4] 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.

[5] Barsky, Robert, John Bound, Kerwin Ko' Charles and Joseph P. Lupton. (2002). "Accounting for the black-white wealth gap: a nonparametric approach," *Journal of the American Statistical Association* 97 (459): 663 - 673.

[6] Bickel, Peter J., Ya'Acov Ritov and Jon A. Wellner. (1991). "Efficient estimation of linear functionals of a probability measure P with known marginal distributions," *Annals of Statistics* 19 (3): 1316 - 1346.

[7] Björklund, Anders and Markus Jäntti. (1997). "Intergenerational income mobility in Sweden compared to the United States," *American Economic Review* 87 (5): 1009 - 1018.

[8] Carroll, R. J. and M. P. Wand. (1991). "Semiparametric estimation in logistic measurement error models," *Journal of the Royal Statistical Society B* 53 (3): 573 - 585.

[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] Cheng, Jing, Dylan S. Small, Zhiqiang Tan, and Thomas R. Ten Have. (2009). "Efficient nonparametric estimation of causal effects in randomized trials with noncompliance," *Biometrika* 96 (1): 19 - 36.

[11] Currie, Janet and Aaron Yelowitz. (2000). "Are public housing projects good for kids?" *Journal of Public Economics* 75 (1): 99 - 124

[12] Dehejia, Rajeev H. and Sadek Wahba. (1999). "Causal effects in nonexperimental studies: reevaluating the evaluation of training programs," *Journal of the American Statistical Association* 94 (448): 1053 - 1062.

[13] Dinardo, John, Nicole M. Fortin, Thomas Lemieux. (1996). "Labor market institutions and the distribution of wages, 1973 - 1992: a semiparametric approach," *Econometrica* 64 (5): 1001 - 1044.

[14] Elbers, Chris, Jean O. Lanjouw and Peter Lanjouw. (2003). "Micro–level estimation of poverty and inequality," *Econometrica* 71 (1): 355 - 364.

[15] Firpo, Sergio and Cristoph Rothe. (2013). "Semiparametric estimation and inference using doubly robust moment conditions," *Mimeo.*

[16] Fortin, Nicole, and Thomas Lemieux, and Sergio Firpo. (2011). "Decomposition methods in economics," *Handbook of Labor Economics 4A*: 1 - 102 (O. Ashenfelter & D. Card, Eds.). Amsterdam: North-Holland.

[17] Gilbert, Peter B., Subhash R. Lele and Yehuda Vardi. (1999). "Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials," *Biometrika* 86 (1): 27 - 43.

[18] Graham, Bryan S. (2011). "Efficiency bounds for missing data models with semiparametric restrictions," *Econometrica* 79 (2): 437 - 452.

[19] Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel. (2012). "Inverse probability tilting for moment condition models with missing data," *Review of Economic Studies* 79 (3): 1053 - 1079.

[20] Hahn, Jinyong. (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica* 66 (2): 315 - 331.

[21] Heckman, James J. and R. Robb. (1985). "Alternative Methods for Evaluating the Impact of Interventions". In *Longitudinal Analysis of Labor Market Data*, ed.., J. Heckman and B Singer Cambridge: Cambridge University Press.

[22] Hellerstein, Judith K. and Guido W. Imbens. (1999). "Imposing moment restrictions from auxiliary data by weighting," *Review of Economics and Statistics* 81 (1): 1 - 14.

[23] 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.

[24] 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.

[25] Hirano, Keisuke, Guido W. Imbens, Geert Ridder, Donald B. Rubin. (2001). "Combining panel data sets with attrition and refreshment samples," *Econometrica* 69 (6): 1645 - 1659.

[26] Ichimura, Hidehiko and Elena Martinez-Sanchis. (2004). "Identification and estimation of GMM models by a combination of two data sets," *Mimeo.*

[27] Imbens, Guido W. (2004). "Nonparametric Estimation of Average Treatment Effect under Exogeneity: A Review," *Review of Economic and Statistics* 86(1): 4-29.

[28] Inoue, Atsushi and Gary Solon. (2010). "Two-sample instrumental variables estimators," *Review of Economics and Statistics* 92 (3): 557 - 561.

[29] 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.

[30] Khan, Shakeeb and Elie Tamer (2010). "Irregular identification, support conditions, and inverse weight estimation," *Econometrica* 78 (6): 2021 - 2042.

[31] Kitagawa, Evelyn M. (1964). "Standardized comparisons in population research," *Demography* 1 (1): 296 - 315.

[32] Kline, Patrick. (2011). "Oaxaca-Blinder as a reweighting estimator," *American Economic Review Papers & Proceedings*, forthcoming.

[33] Lalonde, Robert J. (1986). "Evaluating the econometric evaluations of training programs," *American Economic Review* 76 (4): 604 - 620.

[34] Little, Roderick J.A. and Mei-Miau Wu. (1991). "Models for contingency tables with known margins when target and sampled populations differ," *Journal of the American Statistical Association* 86 (413): 87 - 95.

[35] 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.

[36] Newey, Whitney K. (1990). "Semiparametric efficiency bounds," *Journal of Applied Econometrics* 5 (2): 99 - 135.

[37] Newey, Whitney K. and Daniel McFadden. (1994). "Large sample estimation and hypothesis testing," *Handbook of Econometrics* 4: 2111 - 2245 (R.F. Engle & D.L. McFadden, Eds.). Amsterdam: North-Holland.

[38] Qin, Jing. (1998). "Inferences for case-control and semiparametric two-sample density ratio models," *Biometrika* 85 (3): 619 - 630.

[39] Qin, Jing, and Biao Zhang. (2007). "Empirical-likelihood-based inference in missing response problems and its application in observational studies," *Journal of the Royal Statistical Society: Series B* 69 (1): 101 – 122.

[40] Qin, Jing, and Biao Zhang. (2008). "Empirical-likelihood-based difference-in-differences estimators," *Journal of the Royal Statistical Society B* 70 (2): 329 - 349.

[41] Ridder, Geert and Robert Moffitt. (2007). "The Econometrics of Data Combination," *Handbook of Econometrics* 6B: 5469 - 5547 (J.J. Heckman & E.E. Leamer, Ed.). Amsterdam: North-Holland.

[42] 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.

[43] Rosenbaum, Paul R. and Donald B. Rubin. (1983). "The central role of the propensity score in observational studies for causal effects," *Biometrika* 70 (1): 41 - 55.

[44] Tan, Zhiqiang. (2006). "A distributional approach for causal inference using propensity scores," *Journal of the American Statistical Association* 101 (476): 1619 - 1637.

[45] Tan, Zhiqiang. (2010). "Bounded, efficient and doubly robust estimation with inverse weighting", *Biometrika*, 97(3): 661-682.

[46] Tarozzi, Alessandro and Angus Deaton. (2009). "Using census and survey data to estimate poverty and inequality for small areas," *Review of Economics and Statistics* 91 (4): 773 - 792.

[47] Tsiatis, Anastasios A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer

[48] Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.