NBER WORKING PAPER SERIES

INVERSE PROBABILITY TILTING AND MISSING DATA PROBLEMS

Daniel Egel
Bryan S. Graham
Cristine Campos de Xavier Pinto

Working Paper 13981
http://www.nber.org/papers/w13981

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2008

We would like to thank Stephen Cosslett, Jinyong Hahn, Guido Imbens, Michael Jansson, Richard Smith, Tom Rothenberg, and members of the Berkeley Econometrics Reading Group for helpful discussions. We are particularly grateful to Geert Ridder for detailed comments on an earlier draft. 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 and the University of California - Berkeley. All the usual disclaimers apply. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Inverse Probability Tilting and Missing Data Problems
Daniel Egel, Bryan S. Graham, and Cristine Campos de Xavier Pinto
NBER Working Paper No. 13981
April 2008
JEL No. C14,C21,C23

ABSTRACT

This paper outlines a new minimum empirical discrepancy (MD) estimator for missing data, sample combination and related problems: inverse probability tilting (IPT). Covered examples include estimation of the average treatment effect (ATE), the average treatment effect on the treated (ATT) and the two sample instrumental variables (TSIV) model. The proposed estimator attains the semiparametric efficiency bound under two auxiliary parametric restrictions (local efficiency), but is consistent so long as one or the other holds (double robustness). A novel feature of IPT is its 'exact balancing' property: after reweighting, sample moments of always-observed covariates in the complete-case subsample equal their corresponding (unweighted) full sample means. We also show how prior restrictions on the marginal distribution of always-observed covariates can be efficiently incorporated into our procedure. We use our methods, and compare them to several alternatives, in an evaluation of the National Supported Work (NSW) demonstration using 'non-experimental' comparison groups drawn from the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS) as in LaLonde (1986) and Dehejia and Wahba (1999). We explore the small sample properties of IPT in a Monte Carlo study. IPT performs well, relative to several alternative estimators, across a variety of data generating processes.

Daniel Egel
Department of Economics
University of California, Berkeley
Berkeley , CA 94720
egel@berkeley.edu

Cristine Campos de Xavier Pinto
Department of Economics
University of California Berkely
508-1 Evans Hall #3880
Berkeley, CA 94720

Bryan S. Graham
UC, Berkeley
Department of Economics
508-1 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
bgraham@econ.berkeley.edu

1 Introduction

Let $\{D, X', DY'\}_{i=1}^{\infty}$ be an independent and identically distributed random sequence drawn from the unknown distribution F with D a binary 'missingness' indicator. When $D = 1$ we observe both X and Y, when $D = 0$ we only observe X. The sampling process identifies $F(x, y|D = 1)$ and $F(d, x)$, but we seek to identify functionals of $F(y, x)$, a distribution not identified by this process alone. To ensure identification we assume that Y is 'missing-at-random' (MAR)[2] conditional on X:

$$F(y|x) = F(y|x, D = d), \quad d \in \{0, 1\}. \tag{1}$$

The only other prior restriction on F is that for some unique γ_0

$$\mathbb{E}[\psi(Z, \gamma_0)] = \int \psi(z, \gamma_0) f(y, x) \, \mathrm{d}m(y)\mathrm{d}m(x) = 0, \tag{2}$$

where $\psi(z, \gamma_0)$ is a known function of $Z = (X', Y')'$ indexed by γ. For simplicity we consider the exactly identified case when $\dim(\gamma) = \dim(\psi(z, \gamma))$.[3]

A large body of research explores identification and estimation of γ_0 under restrictions (1), (2) and additional support conditions. The semiparametric efficiency bound for this problem was calculated by Robins, Rotnitzky and Zhao (1994) and Hahn (1998). Several estimators attaining this bound have been proposed. The above set-up is widely used in the analysis of 'causal effects' (Rosenbaum and Rubin 1983, Heckman and Robb 1985, Imbens 2004), missing regressors (Robins, Rotnitzky and Zhao 1994) and non-classical measurement error (Robins, Hsieh and Newey 1995, Chen, Hong and Tamer 2005, Chen, Hong and Tarozzi 2004, 2008). Below we survey additional applications and propose new ones.

Existing approaches to efficient estimation of γ_0 exploit one of two alternative factorizations of $f(y, x)$ implied by the MAR restriction. The imputation approach substitutes $f(y|x, D = 1) f(x)$ for $f(y, x)$ in (2). Imputation estimators then replace $f(y|x, D = 1)$ with a nonparametric estimate and $f(x)$ with the empirical measure of the full sample. Hahn (1998), Chen, Hong and Tarozzi (2008) and Imbens, Newey and Ridder (2007) pursue variants of this approach.

The second approach, inverse probability weighting (IPW), uses the factorization $f(y, x) = \frac{f(y,x|D=1)Q_0}{p_0(x)}$, where $p_0(x) = \Pr(D = 1|X = x)$ is the 'propensity score' and $Q_0 = \Pr(D = 1)$ the marginal probability of complete observation. IPW estimators replace $p_0(x)$ with a nonparametric estimate and $f(y, x|D = 1)$ with the empirical measure of the $D = 1$ subsample. Hirano, Imbens and Ridder (2003) and Wooldridge (2007) develop this approach (cf., Heckman 1987, Rosenbaum 1987). Imputation and IPW are, under appropriate conditions on the estimates of $f(y|x, D = 1)$ and $p_0(x)$, semiparametrically efficient. Unfortunately $f(y|x, D = 1)$ and $p_0(x)$ can be difficult to estimate nonparametrically in moderate-sized samples (Wang, Linton and Härdle 2004, Ichimura and Linton 2005).

This paper proposes a new method of estimation. We choose $\widehat{\gamma}$ to solve (2) after replacing $F(y, x)$ with an estimate. Our estimate of $F(y, x)$ is a multinomial distribution whose support coincides with that of $D = 1$ selected subsample. The distribution of probability mass is chosen to

[2] The MAR assumption is controversial in certain settings. Little and Rubin (2002) and Manski (2003) provide discussions of the restriction from differing perspectives.

[3] The $m(\cdot)$ in (2) denotes a counting measure or a Lebesgue measure as is appropriate.

be as close as possible to the empirical measure of that subsample, while simultaneously requiring that it be consistent with restrictions on the marginal distribution of X implied by the *full* sample. In particular, we require that the means of a finite number of known functions of X calculated with respect to our estimated measure coincide with their corresponding full sample means. We refer to this latter property as 'exact balancing' (of moments). We show that exactly balancing moments in this way generates attractive efficiency and robustness properties.

Our procedure, which we call inverse probability tilting (IPT), belongs to the family of minimum empirical discrepancy (MD) estimators (Corcoran 1998). Conventional MD estimation focuses on efficient estimation of a distribution function in the absence of sample selection. Our focus is different since the subsample being 'tilted' will generally be inconsistent (in a large sample sense) with the imposed constraints (since $F(x|D=1) \neq F(x)$).[4] In this setting the choice of discrepancy function and imposed constraints together imply a corresponding selection model or propensity score (cf., Little and Wu 1991, Hirano, Imbens, Ridder and Rubin 2001). This suggests non-standard choices for discrepancy functions (cf., Nevo 2002).

When the discrepancy and imposed constraints correspond to a correctly specified propensity score the inverse probability 'tilt' of the selected subsample consistently estimates the true distribution function $F(y,x)$. When this is not the case the tilted distribution can still be used to consistently estimate γ_0 when certain auxiliary restrictions are satisfied. In the language of Bang and Robins (2005), the method is 'doubly robust'.

Specialized applications of inverse probability tilting have been proposed elsewhere. Little and Wu (1991) suggest the method for calibrating 2×2 contingency tables to known margins when selection is logistic. Nevo (2002, 2003) extends their approach to moment condition models. Hellerstein and Imbens (1999), although with a different motivation, develop related methods for regression models.

Our contribution is distinctive in a number of ways. First, our focus is on identification. Hellerstein and Imbens (1999), in contrast, focus on interpreting the probability limit of their weighted least squares estimator under general forms of misspecification.

Second, we show how to use IPT to solve a large number of econometric problems. We provide formal results for ignorable missing data and sample combination problems (cf., Robins, Rotnitzky and Zhao 1994, Hahn 1998, Chen, Hong and Tarozzi 2008, Ridder and Moffitt 2007, Graham 2007). These two families respectively cover the average treatment effect (ATE) and the average treatment effect on the treated (ATT) estimands. We make the useful, and apparently new, observation that the two sample instrumental variables (TSIV) model of Angrist and Krueger (1992) belongs to our missing data family. We show that our IPT estimator is more efficient than the one proposed by Angrist and Krueger (1992). Additionally, we extend their model to allow for 'incompatible' samples; IPT is the first consistent estimator available for this extended TSIV model. Finally, we sketch how IPT can be applied to estimate other classes of models such as the additive non-ignorable (AN) attrition model of Hirano, Imbens, Ridder and Rubin (2001).[5]

Third, and more innovatively, we show that IPT is locally semiparametric efficient and doubly robust. Under auxiliary assumptions on the form of the propensity score and the conditional

[4] This suggests that some of our methods may be useful for analyzing and estimating misspecified moment condition models.

[5] Nevo (2003) develops a related method for estimating the AN attrition model.

expectation function (CEF) of $\psi(Z, \gamma_0)$ given X, IPT is semiparametrically efficient. Consistency, however, requires only one or the other of these two parametric restrictions to hold. For our sample combination family of problems, which includes the ATT estimand, we are aware of no competing estimators with similar properties. For missing data problems Robins, Rotnitzky and Zhao (1994) proposed a locally efficient and doubly robust augmented inverse probability weighting (AIPW) estimator (cf., Tsiatis 2006). Advantages of IPT relative to AIPW include the need to estimate fewer nuisance parameters, its exact balancing property, and applicability to settings where a maximum likelihood estimate of the propensity score is unavailable (e.g., the case of non-ignorable attrition in panel data).

Fourth, for missing data problems, we show how prior restrictions on the distribution of X can be incorporated into our procedure. Such restrictions may be available from census cross-tabs or administrative data. A modification of our IPT estimator is locally efficient for missing data problems when the distribution of X is known. Although Chen, Hong and Tarozzi (2004) calculate the efficiency bound for this problem, we are aware of no estimator, other than our own, which attains this bound.

Fifth, we provide duality results which facilitate computation. Here we build on results drawn from the generalized empirical likelihood (GEL) literature (e.g., Imbens 1997, Newey and Smith 2004).[6] We show that IPT is a numerically feasible procedure by using it to revisit the National Supported Work (NSW) Demonstration evaluation (LaLonde 1986, Dehejia and Wahba 1999).

Sixth, we explore the small sample properties of IPT in a series of Monte Carlo experiments. These experiments directly compare the performance of IPT with the parametric IPW estimator of Wooldridge (2007), the nonparametric IPW estimator of Hirano, Imbens and Ridder (2003), the imputation estimators of both Imbens, Newey and Ridder (2007) and Chen, Hong and Tarozzi (2004, 2008) and the AIPW estimator of Robins, Rotnitzky and Zhao (1994). IPT performs well relative to each of these estimators across a variety of data generating processes.

Our focus on locally efficient estimation requires some justification. Globally efficient estimators for γ_0 have been developed by Hahn (1998), Hirano, Imbens and Ridder (2003), Chen, Hong and Tarozzi (2008) and Imbens, Newey and Ridder (2007), implementation of these estimators typically requires high-dimensional nonparametric smoothing. While this does not affect their first order asymptotic properties, it does affect small sample performance (cf., Wang, Linton and Härdle 2004, Ichimura and Linton 2005, Imbens, Newey and Ridder 2007). This motivates the 'flexible parametric' approach taken here. In practice, many implementations of missing data methods take such a form. Our distribution theory is concordant with applied practice. When our parametric assumptions are mirrored in the data, IPT is as efficient as methods based on nonparametric smoothing. The double robustness property of IPT provides some protection against incorrect parametric assumptions. Robins and Ritov (1997) and Robins, Rotnitzky and van der Laan (2000) provide additional justifications for adopting a flexible parametric approach.

Relative to other methods for estimating moment condition models with missing data an attractive feature of IPT is its exact balancing property. In program evaluations it is standard practice to report differences in covariate means across treatment and control units after first blocking on, or weighting by, the propensity score (e.g., Hirano and Imbens 2001, Table 2, p. 270). In contrast,

[6] Some of the algorithms we propose will interest those studying GEL estimation. In particular, generalizing an idea due to Owen, we develop an approach to dealing with 'restricted domain' of the IPT/GEL criterion function.

inverse probability tilting constructs a weighting scheme such that treatment and control covariate means will be *identically equal* to each another after weighting as well as identically equal to the (unweighted) covariate means taken across all units. In principle covariate variances, covariances and higher order sample moments can also be exactly balanced. This property of IPT, in addition to being aesthetically attractive, is the source of its local semiparametric efficiency and double robustness.[7]

Our work exploits insights of Little and Wu (1991), Nevo (2002, 2003) and Hirano, Imbens, Ridder and Rubin (2001). Little and Wu (1991, p. 89) appear to be the first to note the mapping between calibration discrepancies and selection probabilities. In the context of a creative proof of semiparametric just identification of their additive nonignorable (AN) attrition model, Hirano, Imbens, Ridder and Rubin (2001) fully generalize Little and Wu's (1991) observation. We exploit a closely related mapping when constructing the inverse probability tilt. Nevo (2002) provides an information theoretic interpretation of Little and Wu's (1991) application of calibration methods in the presence of logistic sample selection. His 'generalized' exponential tilting (GET) criterion function is a special case of the missing data family of criteria outlined below. Nevo (2003) develops an estimator for the Hirano, Imbens, Ridder and Rubin (2001) AN attrition model based on GET. None of these papers demonstrate double robustness or semiparametric efficiency. Less obviously our work exploits insights from logistic discrimination as in Anderson (1982) and nonparametric density estimation under constraints as in Efron and Tibshirani (1996).

In independent work Qin and Zhang (2007) have proposed an average treatment effect estimator which uses empirical likelihood methods to adjust Horvitz-Thompson parametric inverse probability weights so that, after reweighting, covariate means across the $D = 1$ subsample match their overall sample means. Their method, like ours, is locally efficient and doubly robust. Unlike our method their approach requires estimation of both the propensity score as well as a vector of empirical likelihood Lagrange multipliers. Our method combines these two steps into one and hence requires estimation of fewer nuisance parameters and also applies to data combination and certain nonignorable missing data problems.[8]

Section 2 provides a heuristic overview of IPT as applied to the simplest of problems: estimation of the mean of a outcome variable that is missing at random (MAR). In the context of this example we provide intuition for the efficiency and robustness properties of the IPT. This section also details the connection between IPT and minimum empirical discrepancy (MD) estimation of distribution functions (Corcoran 1998). Developing the relationship between MD and IPT provides certain insights into IPT's attractive theoretical properties. It also connects our work with that generalized empirical likelihood alternatives to GMM (e.g., Imbens 1997, 2002, Newey and Smith 2004, Kitamura 2007).

In Section 3 we formally outline the application of IPT to semiparametric missing data problems. We specifically discuss application of IPT to ATE and TSIV estimation in detail. Section 4 then

[7] An additional advantage of IPT, relative to other currently available methods, is its interpretability under misspecification, although we do not develop this point (cf., Hellerstein and Imbens 1999).

[8] Formally Qin and Zhang (2007) only discuss estimation of the marginal mean of a response variable when it is missing at random (MAR). However their results easily extend to cover moment condition models with data missing at random. It also appears possible to adapt their methods to data combination problems (although their efficiency and robustness properties in that case are unclear). It is not possible to apply their method to nonignorable missing data problems because in such problems the propensity score cannot be estimated by maximum likelihood.

outlines the application of IPT to a class of semiparametric data combination problems. Examples covered there include estimation of the ATT and our extension of Angrist and Krueger's (1992) TSIV model to allow for 'incompatible' samples.

Section 5 sketches the application of IPT to a few non-standard problems, such as the additive nonignorable attrition model of Hirano, Imbens, Ridder and Rubin (2001). Section 6 reports the results of a series of Monte Carlo experiments and presents an illustrative application: estimation of the ATT for National Supported Work (NSW) participants as in LaLonde (1986) and Dehejia and Wahba (1999). Section 7 summarizes and suggests areas for future research.

Appendix A provides details of computation, while Appendix B collects proofs. In Appendix C we detail other possible applications of IPT. Examples covered there include M-estimation under variable probability sampling (Wooldridge 1999, 2007), the construction of counterfactual wage distributions as in Dinardo, Fortin and Lemieux (1996) and Barsky, Bound, Charles and Lutpon (2002), binary choice models under choice-based sampling as in Cosslett (1981), and case-control studies with contaminated controls as in Lancaster and Imbens (1996) and Qin (1998).

2 A simple example

It is helpful to begin by considering the population and asymptotic sampling properties of inverse probability tilting in a very simple setting. Given the missing data structure outlined in the introduction, as well as the MAR assumption, we seek to efficiently estimate $\gamma_0 = \mathbb{E}[Y]$. Under (1) our estimand is the solution to (2) with $\psi(Z, \gamma) = Y - \gamma$ or, equivalently,

$$\gamma_0 = \int y f(y, x) \, \mathrm{d}m(y)\mathrm{d}m(x) = \int y \frac{Q_0}{p_0(x)} f(y, x|D = 1) \, \mathrm{d}m(y)\mathrm{d}m(x),$$

where the second equality follows from Baye's Law, Equation (1), and the assumption that $p_0(x)$ is bounded away from zero for all $x \in \mathcal{X}$. The second representation of $\mathbb{E}[Y]$ is exploited by the inverse probability tilting procedure.

Double robust identification The population analog of the IPT estimate of γ_0 is the mean of Y under the 'tilted' distribution $F_*(y, x)$:

$$\mathbb{E}_{F_*}[Y] = \int y f_*(y, x) \, \mathrm{d}m(y)\mathrm{d}m(x). \tag{3}$$

To describe the construction of the tilted distribution function let $h(X, \zeta_0) = h(X) - \zeta_0$ be an $M \times 1$ vector of mean zero functions of X under $F(x)$, the population distribution of X. A leading form for $h(X)$ is

$$h(X) = \left(X, X^2, \ldots, X^M\right)',$$

which implies that ζ_0 equals $X's$ first M uncentered moments.[9] Note that Q_0, $p_0(x)$ and ζ_0 are all asymptotically identified by the sampling process since $(D, X')'$ is observed for all units.

[9] Alternatively we could choose

$$h(X) = (\mathbf{1}(X < x_1), \mathbf{1}(x_1 \leq X < x_2), \ldots, \mathbf{1}(x_{M-2} \leq X < x_{M-1}), \mathbf{1}(x_{M-1} \leq X))',$$

in which case ζ_0 would give the probability mass associated with each of M intervals of the support of X. We thank Michael Jansson for this suggestion.

The *inverse probability tilt* (IPT) of $f(y,x|D=1)$ is given by

$$f_*(y,x) = \frac{f(y,x|D=1)\,Q_0}{G\left(\alpha_* + h(x)'\beta_*\right)},$$

where α_* and β_* are the solutions to the concave programming problem

$$\max_{\alpha,\beta} \left\{\alpha - \mathbb{E}\left[\varphi^+\left(\alpha + (h(X) - \zeta_0)'\beta; Q_0\right)|D=1\right]\right\}, \tag{4}$$

with $\varphi^+(v;Q)$ given by

$$\varphi^+(v;Q) = \left[\frac{v}{G(v)}Q + \int_{Q/G(v)}^{a} G^{-1}\left(\frac{Q}{t}\right)\mathrm{d}m(t)\right], \tag{5}$$

and $G(\cdot)$ an increasing, differentiable and continuous function mapping the real line onto the unit interval.

The $1+M$ first order conditions for (4) are

$$\begin{aligned}\int \frac{f(x|d=1)\,Q_0}{G\left(\alpha_* + h(x)'\beta_*\right)}\mathrm{d}m(x) &= 1 \\ \int h(x)\frac{f(x|d=1)\,Q_0}{G\left(\alpha_* + h(x)'\beta_*\right)}\mathrm{d}m(x) &= \zeta_0.\end{aligned} \tag{6}$$

Equation (6) show that the inverse probability tilt is chosen so that it integrates to one, is greater than or equal to zero and shares (at least) M moments with $F(x)$. This is the 'exact balancing' (of moments) property of IPT.

Our main identification result is to show that the expectation of Y under the tilted distribution is consistent for its population expectation (i.e, $\mathbb{E}_{F_*}[Y] = \mathbb{E}[Y] = \gamma_0$) if *at least one of* two auxiliary parametric restrictions holds: (1) $p_0(x) = G(\alpha_0 + h(x)'\beta_0)$ *or* (2) $\mathbb{E}[Y|X=x] = \varsigma_0 + \Pi_0 h(x)$. This is our double robustness result.

Equality of $\mathbb{E}_{F_*}[Y]$ and γ_0 under the first condition follows from global concavity of $\varphi^+(v;Q)$ and the consequent equalities of $\alpha_* = \alpha_0$ and $\beta_* = \beta_0$. Equality of $\mathbb{E}_{F_*}[Y]$ and γ_0 under the second condition is less obvious and an important quality of inverse probability tilting. By iterated expectations and (1) we have

$$\mathbb{E}_{F_*}[Y] = \int \mathbb{E}[Y|X=x]\frac{f(x|d=1)\,Q_0}{G\left(\alpha_* + h(x)'\beta_*\right)}\mathrm{d}m(x).$$

Substituting in $\varsigma_0 + \Pi_0 h(x)$ for $\mathbb{E}[Y|X=x]$ gives

$$\mathbb{E}_{F_*}[Y] = \varsigma_0 \int \frac{f(x|d=1)\,Q_0}{G\left(\alpha_* + h(x)'\beta_*\right)}\mathrm{d}m(x) + \Pi_0 \int h(x)\frac{f(x|d=1)\,Q_0}{G\left(\alpha_* + h(x)'\beta_*\right)}\mathrm{d}m(x) = \varsigma_0 + \Pi_0\zeta_0 = \mathbb{E}[Y]$$

where the second equality follows from the first order conditions for (α_*, β_*) given by (6) and the third from iterated expectations. The mean of Y with respect to the inverse probability tilted distribution will equal its population mean if the two distributions, while different, are sufficiently similar. For example, if X is a scalar and $\mathbb{E}[Y|X=x]$ is quadratic in x, then requiring that the

inverse probability tilt of $F(y,x|D=1)$ shares the same mean and variance for X as $F(x)$ ensures that $\mathbb{E}_{F_*}[Y]=\gamma_0$.

Local semiparametric efficiency By iterated expectations we can show that (6) is equivalent to the moment restriction

$$\mathbb{E}\left[\left\{\frac{D}{G\left(\alpha_*+h(X)'\beta_*\right)}-1\right\}\left\{\begin{array}{c}1\\ h(X)\end{array}\right\}\right]=0, \tag{7}$$

while (3) is equivalent to

$$\mathbb{E}\left[\frac{D(Y-\gamma_0)}{G\left(\alpha_*+h(X)'\beta_*\right)}\right]=0. \tag{8}$$

Our IPT estimator (in this example) is simply the sequential method-of-moments estimate of γ_0 based on these two restrictions. This representation simplifies asymptotic analysis and clarifies that the inverse probability tilt is constructed using a 'working model' for the propensity score of the form $G\left(\alpha_*+h(X)'\beta_*\right)$. If this working model is correctly specified, $p_0(x)=G(\alpha_0+h(x)'\beta_0)$, and $\mathbb{E}[Y|X=x]=\varsigma_0+\Pi_0 h(x)$, then our estimator attains the semiparametric efficiency bound for this problem. This is our local semiparametric efficiency result.

Efficiency is a consequence of the exact balancing property of the IPT.[10] Our estimate of γ_0 is given by the solution to

$$\frac{1}{N}\sum\nolimits_{i=1}^{N}\frac{D_i(Y_i-\widehat{\gamma})}{G(\widehat{\alpha}+h(X_i)'\widehat{\beta})}=0, \tag{9}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are first step estimates based on (7), which solve

$$\frac{1}{N}\sum\nolimits_{i=1}^{N}\left(\begin{array}{c}\frac{D_i}{G(\widehat{\alpha}+h(X_i)'\widehat{\beta})}\\ \frac{D_i h(X_i)}{G(\widehat{\alpha}+h(X_i)'\widehat{\beta})}\end{array}\right)=\left(\begin{array}{c}1\\ \widehat{\zeta}\end{array}\right), \tag{10}$$

where $\widehat{\zeta}=\sum_{i=1}^{N}h(X_i)/N$ is the full sample mean of $h(X_i)$. Equation (10) shows that $\widehat{\alpha}$ and $\widehat{\beta}$ are chosen so that the inverse probability weights, $D_i/G(\widehat{\alpha}+h(X_i)'\widehat{\beta})$, sum to one and the inverse probability weighted mean of $h(X_i)$ equals its full sample (unweighted) mean. These are the sample analogs of (6) above. An implication of (10) is that the inverse probability weighted mean of $\varsigma_0+\Pi_0 h(X_i)$ in the $D_i=1$ subsample equals is overall sample mean:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i(\varsigma_0+\Pi_0 h(X_i))}{G(\widehat{\alpha}+h(X_i)'\widehat{\beta})}=\frac{1}{N}\sum_{i=1}^{N}(\varsigma_0+\Pi_0 h(X_i)).$$

[10] Graham (2007) provides a method-of-moments framework for understanding semiparametric efficiency in missing data models. Efficiency of our estimator can also be shown using his results.

Using this exact balancing implication of (10) we can add and subtract terms to get

$$0 = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i\left(Y_i-\widehat{\gamma}\right)}{G(\widehat{\alpha}+h\left(X_i\right)'\widehat{\beta})} - \frac{1}{N}\sum_{i=1}^{N}\frac{D_i\left(\varsigma_0+\Pi_0 h(X_i)-\gamma_0\right)}{G(\widehat{\alpha}+h\left(X_i\right)'\widehat{\beta})} - \frac{1}{N}\sum_{i=1}^{N}\left(\varsigma_0+\Pi_0 h(X_i)-\gamma_0\right).$$

Exploiting the equality $\mathbb{E}\left[Y|X\right]=\varsigma_0+\Pi_0 h(X)$ and solving for $\sqrt{N}(\widehat{\gamma}-\gamma_0)$ then gives

$$\sqrt{N}(\widehat{\gamma}-\gamma_0) = \frac{1}{\sqrt{N}}\left\{\sum_{i=1}^{N}\frac{D_iY_i}{G(\widehat{\alpha}+h\left(X_i\right)'\widehat{\beta})}-\gamma_0-\frac{\mathbb{E}\left[Y|X_i\right]}{G(\widehat{\alpha}+h\left(X_i\right)'\widehat{\beta})}(D_i-G(\widehat{\alpha}+h\left(X_i\right)'\widehat{\beta}))\right\}.$$

Finally, a mean-value expansion in $\widehat{\alpha}$ and $\widehat{\beta}$ about α_0 and β_0 yields the asymptotically linear representation

$$\sqrt{N}(\widehat{\gamma}-\gamma_0)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\frac{D_iY_i}{p_0\left(X_i\right)}-\gamma_0-\frac{\mathbb{E}\left[Y|X_i\right]}{p_0\left(X_i\right)}\left(D_i-p_0\left(X_i\right)\right)\right\}+o_p\left(1\right),$$

where the term in $\{\cdot\}$ is the efficient score (cf., Hahn 1998). Local semiparametric efficiency of $\widehat{\gamma}$ follows directly.

Connection to minimum empirical discrepancy estimation of distribution functions As noted in the introduction, IPT is usefully viewed as a minimum empirical discrepancy (MD) estimator. This representation is valuable primarily for pedagogical purposes, and for that reason, we briefly develop it in this subsection. In particular the MD formulation highlights that IPT involves first estimating the marginal distribution of the missing variable, Y, and then applying standard M-estimation techniques to reweighted data. For asymptotic analysis as well as estimation a method-of-moments representation of IPT is more convenient and hence emphasized in subsequent sections.

Observe that γ_0 is a functional of $F\left(y,x\right)$. Unfortunately, due to sample selection, the empirical distribution function (EDF) of the $D=1$ subsample is not a consistent estimate of $F\left(y,x\right)$. However, as X is observed for all units, $F_N\left(x\right)=\sum_{i=1}^{N}\mathbf{1}\left(X_i\leq x\right)/N$ is a consistent estimate of $F\left(x\right)$. Since there are no restrictions on $X's$ marginal distribution, this estimate is also efficient.

The MD representation of IPT shows that it uses $F_N\left(x\right)$ in conjunction with the selected or 'complete-case' subsample to construct an estimate of $F\left(y,x\right)$. This estimate is chosen to be 'as close as possible' to the empirical measures of the $D=1$ subsample subject to the restriction that it share M moments of X with $F_N\left(x\right)$. Operationalizing 'as close as possible' requires choosing a discrepancy metric or distance function. As pointed out by Hirano, Imbens, Ridder and Rubin (2001) in a related setting, this choice is isomorphic to specifying a (working) model for the propensity score.

Exploiting this insight we can construct a class of discrepancy functions appropriate for missing data problems. This class includes the generalized exponential tilting (GET) discrepancy of Little

and Wu (1991) and Nevo (2002) as a special case. Consider the function

$$D\left(P,R\right)=\int\varphi\left(\frac{\mathrm{d}P}{\mathrm{d}R}\right)\mathrm{d}R,$$

which measures the divergence between the probability measures P and R. This function is convex, differentiable on its domain, and chosen such that $D\left(\cdot,R\right)$ is minimized at R (cf., Bickel, Klassen, Ritov and Wellner 1993, Chapter 7, Kitamura 2007). We work with contrast functions of the form

$$\varphi\left(v;\kappa\right)=\begin{cases}-\frac{v}{k(\kappa)}G^{-1}\left(\kappa\right)-\frac{1}{k(\kappa)}\int_{v}^{a}G^{-1}\left(\frac{\kappa}{t}\right)\mathrm{d}t & v>\kappa\\ +\infty & v\leq\kappa\end{cases}\tag{11}$$

for $\kappa\in\left(0,1\right)$ and $k\left(\kappa\right)=-\kappa/G_{1}\left(G^{-1}\left(\kappa\right)\right)$. The function $G\left(\cdot\right)$ is strictly increasing, differentiable and maps into the unit interval with $\lim_{v\rightarrow-\infty}G\left(v\right)=0$ and $\lim_{v\rightarrow\infty}G\left(v\right)=1$. We also require that, $G_{1}\left(v\right)=\partial G\left(v\right)/\partial v$, is symmetric about zero. Observe that (11) is convex, differentiable and attains its minimum at $v=1$.[11] The term $k\left(\kappa\right)$ is a normalizing constant; its presence facilitates asymptotic analysis as well as comparisons with generalized empirical likelihood (GEL) estimation.

Any suitably well-behaved CDF can be used to construct a contrast function of the form given by (11). In the special case where $G(v)=\exp\left(v\right)/\left[1+\exp\left(v\right)\right]$ our family implies (see Appendix A for details):

$$\varphi(v;\kappa)\propto\left(v-\kappa\right)\ln\left(v-\kappa\right)-v\ln\left(1-\kappa\right)-\left(v-\kappa\right),\tag{12}$$

which is the GET discrepancy of Nevo (2002).

When $G\left(v\right)$ is the CDF of a uniform random variable with support $\left[-1,1\right]$, which corresponds to a linear probability working model for the propensity score, we have

$$\varphi(v,Q)\propto v-\ln v,$$

which equals the (normalized) discrepancy associated with empirical likelihood (Imbens 1997, Newey and Smith 2004).[12]

Assume, without loss of generality, that the first N_1 observations correspond to $D=1$ units, with the remaining N_0 observations corresponding to $D=0$ units. The inverse probability tilt of the selected subsample is given by the solution to

$$\min_{\pi_{11},\ldots,\pi_{1N_1}}\quad\frac{1}{N_1}\sum_{i=1}^{N_1}\varphi(N_1\pi_{1i};\widehat{Q}),\quad s.t.\quad\sum_{i=1}^{N_1}\pi_{1i}=1,\quad\sum_{i=1}^{N_1}\pi_{1i}h(X_i,\widehat{\zeta})=0,\tag{13}$$

with $h(X,\zeta_0)$ and $\widehat{\zeta}$ as defined above. For what follows we define $\rho_0=\left(Q_0,\zeta_0'\right)'$.[13]

[11] A closely related family of discrepancy metrics is used by Hirano, Imbens, Ridder and Rubin (2001) (cf., Equation (15) p. 1656 of their paper).

[12] Note that the linear probability form does not, strictly speaking, satisfy the requirements needed to belong to our family of discrepancies.

[13] An important feature of $\varphi(v;\widehat{Q})$ is that it only finite for $v>\widehat{Q}$. This effectively constrains $N_1\widehat{\pi}_{1i}$ to be greater than $\widehat{Q}$ for all $i=1,\ldots,N_1$; a restriction implied by the structure of the missing data problem. By Baye's Law, the MAR restriction, and the propensity score being bounded away from zero, we have

$$\frac{f\left(y_1,x\right)}{f\left(y_1,x|D=1\right)}=\frac{Q_0}{p_0\left(x\right)}>Q_0\quad\forall\quad x\in\mathcal{X}.$$

Let $F_{\widehat{\pi}_1}(y,x) = \sum_{i=1}^{N_1} \widehat{\pi}_{1i} \mathbf{1}(Y_i \leq y, X_i \leq x)$ denote the inverse probability tilted distribution function. Inspection of (13) shows that this distribution function is chosen to be as close as possible to the empirical measure of the $D = 1$ subsample subject to the requirement that it satisfies the 'moment balancing constraint'

$$\int h(x)\, \mathrm{d}F_{\widehat{\pi}_1}(y_1, x) = \int h(x)\, \mathrm{d}F_N(x). \tag{14}$$

Since $F_N(x)$ is an efficient estimate of the distribution of X, it seems reasonable to require $F_{\widehat{\pi}_1}(y,x)$ to satisfy (14). As sketched above, imposing (14) leads to a semiparametrically efficient estimate of γ_0.

Let $\mathcal{L}\left(\pi_{11}, \ldots \pi_{1N_1}, \eta^1, \lambda^1; \widehat{\rho}\right)$ be the Lagrangian associated with (13), where η^1 is the scalar multiplier associated with the adding-up constraint and λ^1 the $M \times 1$ vector of multipliers associated with requirement that $h(X, \widehat{\zeta})$ be mean zero. The N_1 first order conditions for the probabilities imply that

$$\widehat{\pi}_i = \frac{\widehat{Q}}{N_1} \frac{1}{G(k(\widehat{Q})t(X_i, \widehat{\zeta})'\widehat{\delta}^1 + G^{-1}(\widehat{Q}))}, \qquad i = 1, \ldots, N_1, \tag{15}$$

where $t(X,\zeta) = \left(1, h(X,\zeta)'\right)'$ and $\delta^1 = \left(\eta^1, \lambda^{1\prime}\right)'$. The first order conditions for $\widehat{\delta}^1$, after substituting in (15), are then

$$\frac{\widehat{Q}}{N_1} \sum_{i=1}^{N_1} \frac{t(X_i, \widehat{\zeta})}{G(k(\widehat{Q})t(X_i, \widehat{\zeta})'\widehat{\delta}^1 + G^{-1}(\widehat{Q}))} - t_0 = 0, \tag{16}$$

where $t_0 = (1, \underline{0}')'$.

Let $\delta^1_* = (\eta^1_*, \lambda^{1\prime}_*)$ denote the probability limit of $\widehat{\delta}^1$. If we define the one-to-one mapping $\alpha_* = k(Q_0)\eta^1_* - \zeta_0'\lambda^1_* + G^{-1}(Q_0)$ and $\beta_* = k(Q_0)\lambda^1_*$ it is apparent that (16) is numerically identical to (10) above and hence that the Lagrange multipliers on the adding-up and moment constraint index a working model for the propensity score.

Once the weights are constructed the IPT estimate of γ_0 is given by

$$\widehat{\gamma} = \sum\nolimits_{i=1}^{N_1} \widehat{\pi}_{1i} Y_i.$$

As discussed above, the MD representation's primary value is to provide insight into some of the properties of our procedure. To connect the MD approach to the formulation given in (4) we use duality. Specifically, the Fenchel duality theorem (Rockafellar 1970, Borwein and Lewis 1991) implies that $\widehat{\delta}^1$ (when it exists) is also the solution to

$$\max_{\delta^1} \left\{ t_0'\delta^1 + \frac{1}{N_1} \sum_{i=1}^{N_1} \varphi^+(t(X_i, \widehat{\zeta})'\delta^1, \widehat{Q}) \right\}, \tag{17}$$

The left-hand-side of this expression is the population analog of $N_1\widehat{\pi}_{1i}$ and hence the restriction follows when $p_0(x)$ is strictly bounded above by one.

When $p_0(X_i)$ is close to zero $\widehat{\pi}_{1i}$ may become quite large and (13) will be difficult to solve. Both of these problems are manifestations of 'limited overlap' (cf., Imbens 2004). Appendix A discusses computation in detail, with particular emphasis on how to handle the restricted domain of $\varphi(v;\kappa)$ and computation when overlap is limited.

where $\varphi^{+}(v,Q)$ is the negative of the Fenchel conjugate of $\varphi(v,Q)$[14]:

$$\varphi^{+}(v,Q)=-\frac{1}{k\left(Q\right)}\left[\frac{k\left(Q\right)v+G^{-1}\left(Q\right)}{G(k\left(Q\right)v+G^{-1}\left(Q\right))}Q+\int_{Q/G(k(Q)v+G^{-1}(Q)}^{a}G^{-1}\left(\frac{Q}{t}\right)\mathrm{d}t\right]. \tag{18}$$

When $G\left(v\right)$ is logistic we have

$$\varphi^{+}(v,Q)=-vQ-Q\left(1-Q\right)\exp\left[\frac{v}{1-Q}-\ln\left(\frac{Q}{1-Q}\right)\right],$$

while for the linear probability case we have

$$\varphi^{+}(v,Q)\propto\ln\left(1-v\right),$$

which is the GEL criterion associated with the empirical likelihood (EL) estimator (Newey and Smith 2004).

After accounting for normalization (17) is simply the sample analog of (4). The MD probabilities can be recovered by

$$\widehat{\pi}_{1i}=-\varphi_{1}^{+}(t(X_i,\widehat{\zeta})'\widehat{\delta^{1}},\widehat{Q})/N_1,\quad i=1,\ldots,N_1, \tag{19}$$

with $\varphi_{1}^{+}(v,Q)=-Q/G\left(k\left(Q\right)v+G^{-1}\left(Q\right)\right)$.

Relationship to inverse probability weighting It is illuminating to compare IPT with inverse probability weighting (IPW) as in Hirano, Imbens and Ridder (2003) and Wooldridge (2007). Wooldridge's (2007) estimator also solves (9) except $\widehat{\alpha}$ and $\widehat{\beta}$ are replaced with their MLEs; the solutions to

$$\frac{1}{N}\sum\nolimits_{i=1}^{N}\left(\frac{D_i-G(\widehat{\alpha}+h\left(X_i\right)'\widehat{\beta})}{G(\widehat{\alpha}+h\left(X_i\right)'\widehat{\beta})\left[1-G(\widehat{\alpha}+h\left(X_i\right)'\widehat{\beta})\right]}\right)\left\{\begin{array}{c}1\\ h\left(X_i\right)\end{array}\right\}=0,$$

instead of (10). Hirano, Imbens and Ridder's (2003) estimator is the same except that they restrict $G\left(\cdot\right)$ to be the logistic CDF but allow the dimension of $h\left(X\right)$ to increase with the sample size. The resulting inverse probability weights do not sum to one, nor does the weighted mean of $h\left(X_i\right)$ equal to its full sample one (cf., Imbens 2004, pp. 16 - 17). Both estimators perform poorly relative to IPT in our Monte Carlo experiments.[15]

Fundamental to both our double robust identification and local semiparametric efficiency results is IPT's unique exact balancing property. Exact balancing creates the possibility of consistency despite misspecification of the propensity score as well as of efficiency when the propensity score is correctly specified.

---

[14]Note that our definition of $k\left(Q\right)$ ensures that $\varphi_{1}^{+}\left(0;Q\right)=\varphi_{2}^{+}\left(0;Q\right)=-1$ (with $\varphi_{j}^{+}\left(x\right)\overset{def}{\equiv}\varphi^{+j}\left(x\right)/\partial x^{j}$ for $j=1,2,\ldots$). These are the same normalized imposed on the GEL family of discrepancies studied by Newey and Smith (2004). This facilitates some comparisons and observations we make in the conclusion.

[15]Hirano, Imbens and Ridder's (2003) estimator, because they require the dimension of $h\left(X\right)$ to increase with N, is globally efficient.

3 IPT and semiparametric missing data problems

In this section we formally outline the application, and characterize the large sample properties, of IPT as applied to semiparametric missing data problems (cf., Robins, Rotnitzky and Zhao 1994, Chen, Hong and Tarozzi 2008). To describe this class of problems we let $Z = (Y_1', Y_0', X')'$ be a random vector, $\gamma_0 \in \mathcal{G} \subset \mathbb{R}^K$ an unknown parameter and assume that:

Assumption 3.1 (IDENTIFICATION) *For some known function* $\psi(z, \gamma) = \psi_1(y_1, x, \gamma) - \psi_0(y_0, x, \gamma)$

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

with $\mathbb{E}[\psi(Z, \gamma)] \neq 0$ *for all* $\gamma \neq \gamma_0$, $\gamma \in \mathcal{G} \subset \mathbb{R}^K$, $z \in \mathcal{Z} \subset \mathbb{R}^{\dim(Z)}$.

If a random sample of Z is available then estimation of γ_0 is entirely standard. Instead we consider the case where the analyst never observes Y_1 and Y_0 for the same unit. Let D be a binary indicator variable. When $D = 1$ we observe Y_1, while when $D = 0$ we observe Y_0; X and D are observed for all units. The semiparametric missing data model is defined by Assumption 3.1 as well as:

Assumption 3.2 (RANDOM SAMPLING) $\{D_i Y_{1i}, (1 - D_i) Y_{0i}, X_i, D_i\}_{i=1}^N$ *is an independently and identically distributed random sequence.*

Assumption 3.3 (MISSING AT RANDOM) $(Y_1, Y_0) \perp D|X$.

Assumption 3.4 (STRONG OVERLAP) *Let* $p_0(x) = \Pr(D = 1|X = x)$, *then* $0 < \kappa < p_0(x) < 1 - \kappa < 1$ *for some* $0 < \kappa < 1$ *and all* $x \in \mathcal{X} \subset \mathbb{R}^{\dim(X)}$.

In what follows we refer to the problem defined by Assumptions 3.1 to 3.4 as the semiparametric missing data (SMD) problem. That γ_0 is identified is well known, following from a straightforward application of the law of iterated expectations (e.g., Imbens 2004, Wooldridge 2007).

3.1 Estimation

The application of IPT to this problem involves three steps. First, we use the full sample to compute the marginal probability of the event $D = 1$ as well as the mean of $h(X)$. That is we solve

$$\frac{1}{N}\sum_{i=1}^{N} m_1(Z_i, \widehat{\rho}) = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} D_i - \widehat{Q} \\ h(X_i) - \widehat{\zeta} \end{pmatrix} = 0, \tag{20}$$

for $\widehat{\rho} = (\widehat{Q}, \widehat{\zeta}')'$.

Second, we reweight the $D = 1$ and $D = 0$ subsamples to match M full sample moments of X. That is we compute the $D = 1$ and $D = 0$ inverse probability tilts

$$F_{\widehat{\pi}_1}(y_1, x) = \sum\nolimits_{i=1}^{N_1} \widehat{\pi}_{1i} \mathbf{1}(Y_{1i} \leq y_1, X_i \leq x), \quad F_{\widehat{\pi}_0}(y_0, x) = \sum\nolimits_{i=N_1+1}^{N} \widehat{\pi}_{0i} \mathbf{1}(Y_{0i} \leq y_0, X_i \leq x)$$

with probability weights given by

$$\widehat{\pi}_{1i} = -\varphi_1^+(t(X_i, \widehat{\zeta})'\widehat{\delta}^1, \widehat{Q})/N_1, \quad \widehat{\pi}_{0i} = -\varphi_1^+(t(X_i, \widehat{\zeta})'\widehat{\delta}^0, 1 - \widehat{Q})/N_0,$$

for $i = 1, \ldots, N_1$ and $i = N_1 + 1, \ldots, N$ respectively and where $\widehat{\delta}^1$ and $\widehat{\delta}^0$ are the solutions to

$$\max_{\delta^1} \left\{ t_0' \delta^1 + \frac{1}{N_1} \sum_{i=1}^{N_1} \varphi^+(t(X_i, \widehat{\zeta})' \delta^1, \widehat{Q}) \right\}, \tag{21}$$

and

$$\max_{\delta^0} \left\{ t_0' \delta^0 + \frac{1}{N_0} \sum_{i=N_1+1}^{N} \varphi^+(t(X_i, \widehat{\zeta})' \delta^0, 1 - \widehat{Q}) \right\}. \tag{22}$$

Stacking the two first order conditions associated with (21) and (22) on top of one another implies that $\widehat{\delta} = (\widehat{\delta}^{1\prime}, \widehat{\delta}^{0\prime})$ is, after some manipulation, the solution to a second step moment equation of

$$\frac{1}{N} \sum_{i=1}^{N} m_2(Z_i, \widehat{\rho}, \widehat{\delta}) = \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} t_0 + \frac{D_i}{\widehat{Q}} \varphi_1^+(t(X_i, \widehat{\zeta})' \widehat{\delta}^1, \widehat{Q}) t(X_i, \widehat{\zeta}) \\ t_0 + \frac{1 - D_i}{1 - \widehat{Q}} \varphi_1^+(t(X_i, \widehat{\zeta})' \widehat{\delta}^0, 1 - \widehat{Q}) t(X_i, \widehat{\zeta}) \end{pmatrix}, \tag{23}$$

where $\widehat{\rho} = (\widehat{Q}, \widehat{\zeta}')'$ is fixed at its first step value.

In the third and final step $\widehat{\gamma}$ is given by the solution to

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} m_3(Z_i, \widehat{\rho}, \widehat{\delta}, \widehat{\gamma}) = -\frac{1}{N} \Bigg\{ & \sum_{i=1}^{N} \frac{D_i \varphi_1^+(t(X_i, \widehat{\zeta})' \widehat{\delta}^1, \widehat{Q})}{\widehat{Q}} \psi_1(Y_{1i}, X_i, \widehat{\gamma}) \\ & - \frac{(1 - D_i) \varphi_1^+(t(X_i, \widehat{\zeta})' \widehat{\delta}^0, 1 - \widehat{Q})}{1 - \widehat{Q}} \psi_0(Y_{0i}, X_i, \widehat{\gamma}) \Bigg\} = 0, \end{aligned} \tag{24}$$

with $\widehat{\rho}$ fixed at its first step, and $\widehat{\delta}$ at its second step, value.

Standard sequential GMM results can be used to derive the large sample properties of $\widehat{\gamma}$ (cf., Newey and McFadden 1994). From a numerical standpoint the second step is the most difficult; Appendix A details the algorithm we implement in the NSW application and our Monte Carlo experiments.

To connect our estimator to the more familiar inverse probability weighting (IPW) method of Hirano, Imbens and Ridder (2003) and Wooldridge (2007) it is helpful to define the one-to-one mappings $\widehat{\alpha}^1 = k(\widehat{Q})\widehat{\eta}^1 - \widehat{\zeta}'\widehat{\lambda}^1 + G^{-1}(\widehat{Q})$, $\widehat{\beta}^1 = k(\widehat{Q})\widehat{\lambda}^1$, $\widehat{\alpha}^0 = -k(1 - \widehat{Q})\widehat{\eta}^0 + \widehat{\zeta}'\widehat{\lambda}^0 - G^{-1}(1 - \widehat{Q})$ and $\widehat{\beta}^0 = -k(1 - \widehat{Q})\widehat{\lambda}^0$. We can then re-write (23) as

$$\frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \frac{D_i}{G(\widehat{\alpha}^1 + h(X_i)' \widehat{\beta}^1)} \\ \frac{D_i h(X_i)}{G(\widehat{\alpha}^1 + h(X_i)' \widehat{\beta}^1)} \\ \frac{1 - D_i}{1 - G(\widehat{\alpha}^0 + h(X_i)' \widehat{\beta}^0)} \\ \frac{(1 - D_i) h(X_i)}{1 - G(\widehat{\alpha}^0 + h(X_i)' \widehat{\beta}^0)} \end{pmatrix} = \iota_2 \otimes \begin{pmatrix} 1 \\ \widehat{\zeta} \end{pmatrix}, \tag{25}$$

and (24) as

$$\frac{1}{N} \sum_{i=1}^{N} \frac{D_i \psi_1(Y_{1i}, X_i, \widehat{\gamma})}{G(\widehat{\alpha}^1 + h(X_i)' \widehat{\beta}^1)} - \frac{(1 - D_i) \psi_0(Y_{0i}, X_i, \widehat{\gamma})}{1 - G(\widehat{\alpha}^0 + h(X_i)' \widehat{\beta}^0)} = 0. \tag{26}$$

Equality (26) is an inverse probability weighting of the identifying moment. The final step of our procedure is therefore similar to conventional IPW methods. However, it differs in that (i) the

$D = 1$ and $D = 0$ subsamples are weighted by different estimates of the propensity score and (ii) neither of these two estimates is the maximum likelihood estimate. Inspection of (25) reveals that the two sets of propensity score coefficients are chosen to ensure that each set of inverse probability weights sum to one and also that the weighted subsample means of $h(X)$ equal the corresponding (unweighted) full sample mean. If the propensity score is correctly specified IPW will also satisfy these conditions, but only asymptotically, not in finite samples as with IPT.

3.2 Large sample properties

In order to discuss the large sample properties of $\widehat{\gamma}$ it is helpful to introduce the following additional assumptions

Assumption 3.5 (Propensity Score) *There is a unique* $(\alpha_0, \beta_0) \in \mathcal{A} \times \mathcal{B} \subset \mathbb{R}^{1+M}$ *such that* $p_0(x) = G(\alpha_0 + h(x)'\beta_0)$ *for all* $x \in \mathcal{X}$.

Assumption 3.6 (Moment CEF) *Let* $q_j(x; \gamma_0) = \mathbb{E}[\psi_j(Y_j, X, \gamma_0) | X = x]$ *for* $j = 0, 1$ *and*

$$q_1(x; \gamma_0) = \varsigma_1 + \Pi_1 h(x), \quad q_0(x; \gamma_0) = \varsigma_0 + \Pi_0 h(x)$$

for some unique $(\varsigma_1, \Pi_1) \in \mathcal{S} \times \mathcal{P} \subset \mathbb{R}^K \times \mathbb{R}^{KM}$ *and* $(\varsigma_0, \Pi_0) \in \mathcal{S} \times \mathcal{P} \subset \mathbb{R}^K \times \mathbb{R}^{KM}$.

Assumption 3.5 simply states that the working model for the propensity score implicit in the minimum empirical discrepancy procedure is correctly specified. Assumption 3.6 is less standard. Its precise content depends on the form of $\psi_1(Y_1, X, \gamma_0)$ and $\psi_0(Y_0, X, \gamma_0)$. We discuss its precise interpretation in the examples section below.

We provide distribution theory for $\widehat{\gamma}$ appropriate for two cases: (i) either Assumption 3.5 or 3.6, but not both, is true and (ii) both are true. We could also consider the case where neither of these two assumptions hold. This would characterize the large sample properties of $\widehat{\gamma}$ under misspecification (cf., Hellerstein and Imbens 1999, Nevo 2003). While we do not pursue this case here, we note that the asymptotic variance estimator given in the Appendix can also be used to conduct valid inference about the probability limit of $\widehat{\gamma}$ under misspecification.

Our first result shows that if either Assumption 3.5 or 3.6 holds $\widehat{\gamma}$ is consistent for γ_0 and asymptotically normal.

Theorem 3.1 (Double Robustness) *Suppose Assumptions 3.1 to 3.4, either Assumption 3.5 or 3.6, and additional regularity conditions hold, then as* $N \to \infty$

$$\widehat{\gamma} \xrightarrow{p} \gamma_0$$

and

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \Gamma_0^{-1} \Upsilon_0 \Gamma_0^{-1\prime}),$$

where $\Gamma_0 = \mathbb{E}[\partial \psi(Z, \gamma_0 / \partial \gamma')]$ *and the form of* Υ_0 *depends on whether Assumption 3.5 or 3.6 holds (see Appendix B).*

Proof. See Appendix B. ■

Theorem 3.1 implies that the researcher has two opportunities to consistently estimate γ_0. More heuristically it suggests that IPT will perform well for moderately rich forms of $h(X)$ when either propensity score or the conditional expectations of $\psi_1(Y_1, X, \gamma_0)$ and $\psi_0(Y_0, X, \gamma_0)$ are smooth in X.

An immediate corollary of Theorem 3.1 is:

Corollary 3.1 (Asymptotic Normality with data MCAR) *When the assumptions of Theorem 3.1 hold and additionally the data are missing completely at random (MCAR) (i.e., $Q_0 = p_0(X)$ for all $X \in \mathcal{X}$) we have*

$$\begin{aligned}\Upsilon_0 = \Gamma_0^{-1} \Bigg(& \frac{\Omega_{\psi_1\psi_1} - \Omega_{\psi_1 h}\Omega_{hh}^{-1}\Omega_{\psi_1 h}'}{Q_0} + \frac{\Omega_{\psi_0\psi_0} - \Omega_{\psi_0 h}\Omega_{hh}^{-1}\Omega_{\psi_0 h}'}{1 - Q_0} \\ & + (\Omega_{\psi_1 h} - \Omega_{\psi_0 h})\,\Omega_{hh}^{-1}\,(\Omega_{\psi_1 h} - \Omega_{\psi_0 h})'\Bigg)\,\Gamma_0^{-1\prime},\end{aligned} \tag{27}$$

where $\Omega_{\psi_j\psi_j} = \mathbb{V}(\psi_j(Y_j, X, \gamma_0))$, $\Omega_{\psi_j h} = \mathbb{C}(\psi_j(Y_j, X, \gamma_0), h(X))$ *for* $j = 0, 1$, *and* $\Omega_{hh} = \mathbb{V}(h(X))$.

Our next result shows that IPT is locally efficient. By local efficiency we mean that IPT attains the semiparametric efficiency bound for the SMD model when Assumptions 3.5 and 3.6 happen to be true in the sampled population but are not part of the prior restriction (cf., Newey 1990 Robins, Rotnitzky and Zhao 1994).

The maximal asymptotic precision with which γ_0 can be estimated has been characterized by Robins, Rotnitzky and Zhao (1994) and is given by the inverse of

$$\mathcal{I}(\gamma_0) = \Gamma_0'\Lambda_0^{-1}\Gamma_0, \tag{28}$$

with

$$\Lambda_0 = \mathbb{E}\left[\frac{\Sigma_0(X;\gamma_0)}{1 - p_0(X)} + \frac{\Sigma_1(X;\gamma_0)}{p_0(X)} + [q_1(X;\gamma_0) - q_0(X;\gamma_0)]\,[q_1(X;\gamma_0) - q_0(X;\gamma_0)]'\right], \tag{29}$$

where $\Sigma_j(x;\gamma_0) = \mathbb{V}(\psi_j(Y_j, X, \beta) | X = x)$ for $j = 0, 1$. Our next Theorem shows when IPT attains the bound derived by Robins, Rotnitzky and Zhao (1994).

Theorem 3.2 (Local Semiparametric Efficiency) *Suppose Assumptions 3.1 to 3.6 and additional regularity conditions hold, then as $N \to \infty$*

$$\widehat{\gamma} \xrightarrow{p} \gamma_0$$

and

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \mathcal{I}(\gamma_0)^{-1}),$$

with $\mathcal{I}(\gamma_0)$ as defined by (28) and (29).

Proof. See Appendix B. ■

In Appendix B we provide a single variance-covariance estimator for $\widehat{\gamma}$ that is consistent for the cases covered by Theorems 3.1 and 3.2 as well as Corollary 3.1. From the practitioner's standpoint

this is a considerable advantage.[16]

3.3 Incorporation of prior restrictions on $F(x)$

In this subsection we discuss how prior restrictions on the marginal distribution of X can be efficiently incorporated into the IPT procedure. It is well know that such information, which is often available from census cross-tabulations or administrative data, increases the asymptotic precision with which γ_0 can be estimated. Chen, Hong and Tarozzi (2004) show that if, in addition to Assumptions 3.1 to 3.4, $F(x)$ and Q_0 are known, then the variance bound is given by the inverse of (28), with Λ_0 redefined to equal

$$\Lambda_0 = \mathbb{E}\left[\frac{\Sigma_0(x;\gamma_0)}{1-p_0(x)} + \frac{\Sigma_1(x;\gamma_0)}{p_0(x)}\right]. \tag{30}$$

Assume that $Q_0 = \mathbb{E}[D]$ and $\zeta_0 = \mathbb{E}[h(X)]$ are known. Our final result characterizes the large sample properties of the (efficient) GMM estimator based on the restriction

$$\mathbb{E}\begin{bmatrix} m_1(Z) \\ m_2(Z,\delta_0) \\ m_3(Z,\delta_0,\gamma_0) \end{bmatrix} = 0, \tag{31}$$

with $m_1(Z) = m_1(Z,\rho_0)$, $m_2(Z,\delta) = m_2(Z,\rho_0,\delta)$ and $m_3(Z,\delta,\gamma) = m_3(Z,\rho_0,\delta,\gamma)$ and the right-hand-sides of these equalities as defined by (20), (23) and (24) above. Note that here $m_1(Z)$ plays the role of an auxiliary moment (cf., Imbens and Lancaster 1994, Qian and Schmidt 1999, Imbens and Hellerstein 1999).

Corollary 3.2 (Prior Knowledge of F(x)) *Suppose Assumptions 3.1 to 3.6 and additional regularity conditions hold, let $\widehat{\gamma}$ be the efficient GMM estimator based on (31), then as $N \to \infty$*

$$\widehat{\gamma} \xrightarrow{p} \gamma_0$$

and

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \mathcal{I}(\gamma_0)^{-1}),$$

with $\mathcal{I}(\gamma_0)$ as defined by (28) and (30).

Proof. See Appendix B. ■

The appendix provides a consistent variance-covariance estimator for the situation covered by Corollary 3.2. While we do not provide a formal statement of the result, consistency of our known ζ_0 estimator requires only one of Assumptions 3.5 or 3.6 to hold. Our variance estimator is consistent in those cases as well. To the best of our knowledge our estimator is the first estimator to attain the bound for the missing data problem with $F(x)$ known.

[16] We also note that the covariance estimator generally associated with Robins, Rotnitzky and Zhao's (1994) AIPW estimator is consistent only under the assumptions of Theorem 3.2 (cf., the inference approach adopted by Lunceford and Davidian 2004 for example). Consequently, AIPW-based inference may be invalid even when AIPW point estimates are consistent.

3.4 Examples

In order to illustrate the application of each of our large sample results in specific settings we discuss two examples in detail. Appendix C provides additional examples.

Average Treatment Effects 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

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

which corresponds to setting $\psi_1\left(Y_1, X, \gamma\right) = Y_1$ and $\psi_0\left(Y_0, X, \gamma\right) = Y_0 + \gamma$. Since each unit can only be exposed to one intervention, either Y_1 or Y_0 is missing for all units. The conditional probability of assignment to the active intervention is given by $p_0\left(x\right)$.

Efficient estimation of the ATE by IPT involves choosing $h\left(X\right)$ such that the true propensity score is contained in the parametric family $G\left(\alpha + h\left(X\right)'\beta\right)$ and the true potential outcome CEFs are linear in $h\left(X\right)$. This will ensure satisfaction of Assumptions 3.5 and 3.6. Consider the case where the propensity score is known to be constant, as in a randomized experiment. In that case $h\left(X\right)$ should be chosen such that the population regressions of Y_1 and Y_0 on $h\left(X\right)$ accurately approximate $\mathbb{E}\left[Y_1 \middle| X\right]$ and $\mathbb{E}\left[Y_0 \middle| X\right]$. In this case efficient estimation requires overparameterizing the working model of the propensity score; a result analogous to that of Hirano, Imbens and Ridder (2003). If $\mathbb{E}\left[Y_1 \middle| X\right]$ and $\mathbb{E}\left[Y_0 \middle| X\right]$ are smooth relative to $p_0\left(X\right)$, then $h\left(X\right)$ should be chosen such that $G\left(\alpha + h\left(X\right)'\beta\right)$ can provide an accurate approximation of the propensity score (cf., Imbens, Newey and Ridder 2007, Graham 2007).

Corollary 3.2 has implications important for experimental design. Imagine an experiment involving a random sample of schools from a well-defined population (e.g., California public elementary schools). In such a situation administrative data may reveal the marginal distribution of some unit characteristics perfectly. Inverse probability tilting allows this information to be straightforwardly incorporated into estimation of ATEs.[17]

Two sample instrumental variables estimation with compatible samples Assume that $\dim\left(X\right) \geq \dim\left(Y_0\right)$ and consider the following instrumental variables model

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

This suggests a moment function with $\psi_1\left(Y_1, X, \gamma\right) = XY_1$ and $\psi_0\left(Y_0, X, \gamma\right) = XY_0'\gamma$. 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 $\lim_{N_1, N_0 \to \infty} N_1/(N_1 + N_0) = Q_0 > 0$. 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 (cf., Arellano and Meghir 1992).

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.

[17] We also note that our known ζ_0 covariance estimator (with $\widehat{\zeta}$ replacing ζ_0) is consistent for the asymptotic sampling variance of the sample average treatment effect (SATE) of Imbens (2004).

With probability Q_0 the i^{th} unit's values for Y_1 and X are recorded, while with probably $1-Q_0$ 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.

Efficient estimation of the TSIV model involves correctly modelling the first stage regression of Y_0 onto X. Assume that X is such that $\mathbb{E}[Y_0|X]$ is linear in X. In that case

$$\mathbb{E}[\psi_0(Y_0, X, \gamma)|X] = \mathbb{E}[\psi_1(Y_1, X, \gamma)|X] = XX_0'\gamma_{X_0} + X\left(X'\pi\right)\gamma_{Y_0},$$

so that IPT with $h(X)$ including X, the squares of its elements and all pairwise cross-products results in an efficient estimator. The variance bound for this estimator is given by Corollary 3.1, which also demonstrates that the TSIV estimate of Angrist and Krueger (1992, Lemma 1, p. 331) is inefficient. Their estimator has a large sample variance of

$$\Gamma_0^{-1}\left(\frac{\Omega_{\psi_1\psi_1}}{Q_0} + \frac{\Omega_{\psi_0\psi_0}}{1-Q_0}\right)\Gamma_0^{-1\prime}, \tag{32}$$

which is larger, in a matrix sense, than (27). If X is predictive of Y_1 and Y_0 – as is required for identification – the degree of inefficiency can be quite large.

4 IPT and semiparametric data combination problems

In this section we consider semiparametric data combination (SDM) problems (cf., Graham 2007). Such problems correspond to the 'verify-out-of-sample' class of problems considered by Chen, Hong and Tarozzi (2004). A leading estimand which falls in this family of problems is the Average Treatment Effect on the Treated (ATT). In this section we also show that the TSIV model, where the two samples are drawn from different populations, can be analyzed as a SDM problem. This observation substantially extends the range of situations in which TSIV methods can be applied.

Let $F_t(z)$ denote the distribution function for Z in some *target population*. Let $\mathbb{E}_t[\cdot]$ denote expectations taken with respect to this distribution. In the case of the ATT the target population would correspond to the treated population. We seek to estimate γ_0, which is identified by a prior restriction on the target population:

Assumption 4.1 (IDENTIFICATION) *For some known function* $\psi(z,\gamma) = \psi_1(y_1, x, \gamma) - \psi_0(y_0, x, \gamma)$

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

with $\mathbb{E}_t[\psi(Z,\gamma)] \neq 0$ *for all* $\gamma \neq \gamma_0$, $\gamma \in \mathcal{G} \subset \mathbb{R}^K$, $z \in \mathcal{Z} \subset \mathbb{R}^{\dim(Z)}$.

Available is a random sample of size N_t of (Y_1, X) from the target population; Y_0 is not observed for units in the *target sample.* Also available is an *auxiliary sample* of size N_a containing measurements of (Y_0, X). This sample is drawn from some *auxiliary population* with distribution function $F_a(z)$ (let $\mathbb{E}_a[\cdot]$ denote expectations taken with respect to this distribution). The relationship between the two distributions is captured by the following two assumptions.

Assumption 4.2 (Conditional Distributional Equality)

$$F_t\left(y_0, y_1 \middle| x\right) = F_a\left(y_0, y_1 \middle| x\right)$$

Assumption 4.3 (Weak Overlap) *Let* $S_j = \{x : f_j(x) > 0\}$ *for* $j = t, a$, *then*

$$S_t \subset S_a.$$

Assumption 4.2 states that the conditional distribution of (Y_0, Y_1) given X is the same in the two populations. Equivalently the two populations differ only in terms of the distribution of 'always observed' variables, X. Assumption 4.3 states that $f_a(x)$ is positive if $f_t(x)$ is positive. This ensures that in large samples, for each unit in the target sample there will be units with similar values of X in the auxiliary sample.

The semiparametric data combination model is closed by a sampling assumption:

Assumption 4.4 (Random Sampling) $\{(Y_{1i}, X_i)\}_{i=1}^{N_t}$ *and* $\{(Y_{0i}, X_i)\}_{i=1}^{N_a}$ *are random samples from, respectively,* F_t *and* F_a.

The *merged sample* is given by $\left\{\left(D_i, (1-D_i)\, Y_{0i}', D_i Y_{1i}', X_i'\right)'\right\}_{i=1}^{N}$, where $N = N_a + N_t$ and D indicates whether a given unit originates from the target or auxiliary sample. We can treat this sample as a random one from hypothetical *merged population* F_m (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 (cf., Hahn 1998, Imbens 2004, Chen, Hong and Tarozzi 2004, 2008). For example Assumption 4.1 corresponds to assuming that $\mathbb{E}[\psi(Z, \gamma_0) | D = 1] = 0$, Assumption 4.2 to the missing at random restriction of the previous section (i.e., Assumption 3.3 above) and Assumption 4.3 to requiring $p_0(x) \leq 1 - \kappa < 1$ for some $0 < \kappa < 1$ and $p_0(x) = \mathbb{E}[D | X = x]$. One can also replace the assumption of two independent random samples with a multinomial sampling assumption (cf., Graham 2007). We prefer the formulation given above because it emphasizes that the problem is fundamentally one of data combination. Properties of the hypothetical merged population are of incidental interest. The goal is to use the auxiliary data to learn more about the target population.

Assumptions 4.3 and Bayes' rule give the relationship

$$f_t(x) = f_a(x) \left\{ \frac{1 - Q_0}{Q_0} \frac{p_0(x)}{1 - p_0(x)} \right\},$$

where $Q_0 = \mathbb{E}[D]$. This result and Assumption 4.2 then give the equality

$$\mathbb{E}_t[\psi(Z, \gamma)] = \int \psi_1(y_1, x, \gamma)\, f_t(y_1, x)\, \mathrm{d}m(y_1)\mathrm{d}m(x) - \int \psi_0(y_0, x, \gamma) \frac{1 - Q_0}{Q_0} \frac{p_0(x)}{1 - p_0(x)} f_a(y_1, x)\, \mathrm{d}m(y_0)\mathrm{d}m(x).$$

The IPT estimator chooses $\widehat{\gamma}$ to set a sample analog of right-hand-side of the above expression equal to zero. We replace $f_t(y_1, x)$ in the first integral with the empirical measure of the target subsample and $\frac{1-Q_0}{Q_0} \frac{p_0(x)}{1-p_0(x)} f_a(y_1, x)$ with a particular minimum empirical discrepancy 'tilt' of the empirical

measure of the auxiliary sample. The structure of this tilt differs from the one described in Section 3 for missing data problems.

4.1 Estimation

For data combination problems only the $D = 0$ subsample need be tilted. However the structure of the required tilt differs from that needed for missing data problems. For data combination problems we work with discrepancies of the form

$$\varphi(v;\kappa) \overset{def}{\equiv} \begin{cases} -\frac{v}{k(\kappa)}G^{-1}(Q) - \frac{1}{k(\kappa)}\int_v^a G^{-1}\left(\frac{t}{t+(1-\kappa)/\kappa}\right)\mathrm{d}t & x > 0 \\ +\infty & x \leq 0 \end{cases}, \tag{33}$$

where $k(\kappa) = \kappa(1-\kappa)/G_1\left(G^{-1}(\kappa)\right)$. The inverse probability tilt of the auxiliary sample is given by the solution to

$$\min_{\pi_{aN_1+1},\ldots,\pi_{aN}} \quad \frac{1}{N_a}\sum_{i=N_t+1}^{N}\varphi(N_a\pi_{ai};\widehat{Q}), \quad s.t. \quad \sum_{i=N_t+1}^{N}\pi_{ai} = 1, \quad \sum_{i=N_t+1}^{N}\pi_{ai}h(X_i,\widehat{\zeta}^t) = 0, \tag{34}$$

with $\widehat{Q} = N_t/N$ and $\widehat{\zeta}^t = \frac{1}{N_t}\sum_{i=1}^{N_t}h(X_i)$. The latter mean is taken over the target sample, not the full sample. This is because we tilt the empirical measure of the auxiliary sample to match moments of the target sample, not of the full sample as in the missing data case. Let $\mathcal{L}(\pi_{a1},\ldots\pi_{aN_0},\eta^a,\lambda^a;\widehat{\rho})$ be the Lagrangian associated with (34). The probability weights take the form

$$\widehat{\pi}_{ai} = \frac{1}{N_a}\frac{1-\widehat{Q}}{\widehat{Q}}\frac{G(k(\widehat{Q})t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a + G^{-1}(\widehat{Q}))}{1-G(k(\widehat{Q})t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a + G^{-1}(\widehat{Q}))}, \qquad i = N_t+1,\ldots,N, \tag{35}$$

for $t(X,\zeta)$ as defined before and $\widehat{\delta}^a = (\widehat{\eta}^a,\widehat{\lambda}^{a\prime})'$. The vector of Lagrange multipliers solve

$$\frac{1}{N_a}\frac{1-\widehat{Q}}{\widehat{Q}}\sum_{i=N_t+1}^{N}\frac{G(k(\widehat{Q})t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a + G^{-1}(\widehat{Q}))}{1-G(k(\widehat{Q})t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a + G^{-1}(\widehat{Q}))}t(X_i,\widehat{\zeta}^t) - t_0 = 0. \tag{36}$$

Inspection of (36) indicates the inverse probability tilt, $F_{\widehat{\pi}_a}(y_0,x) = \sum_{i=1}^{N_a}\widehat{\pi}_{ai}\mathbf{1}(Y_{0i}\leq y_0, X_i\leq x)$, is chosen to be as close as possible to the empirical measure of auxiliary sample while simultaneously satisfying the balancing restrictions

$$\int h(x)\,\mathrm{d}F_{\widehat{\pi}_a}(y_0,x) = \int h(x)\,\mathrm{d}F_{N_t}(x).$$

Finally, the IPT estimate of γ_0 is given by the solution to

$$0 = \frac{1}{N_t}\sum_{i=1}^{N_t}\psi_1(Y_{1i},X_i,\widehat{\gamma}) - \sum_{i=N_t+1}^{N}\widehat{\pi}_{ai}\psi_0(Y_{0i},X_i,\widehat{\gamma}).$$

As in the missing data case, the MD representation's value is primarily pedagogical. For estimation, as well as to characterize large sample properties, a sequential method of moments formulation is more convenient.

In step one $\widehat{Q}$ and $\widehat{\zeta}^t$ are chosen to solve

$$\frac{1}{N}\sum_{i=1}^{N} m_1(Z_i,\widehat{\rho}) = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} D_i - \widehat{Q} \\ \frac{D_i}{\widehat{Q}}\left(h\left(X_i\right) - \widehat{\zeta}^t\right)\end{pmatrix} = 0. \tag{37}$$

In step two the IPT of the auxiliary sample is computed. As in the missing data case, we work with the dual problem. The Fenchel conjugate of (33) is

$$\varphi^+(v,Q) = -\frac{1}{k\left(Q\right)}\left\{\left[k\left(Q\right)v + G^{-1}\left(Q\right)\right]\frac{1-Q}{Q}\frac{G(k\left(Q\right)v + G^{-1}\left(Q\right))}{1-G(k\left(Q\right)v + G^{-1}\left(Q\right))}\right\}$$
$$+\int_{\frac{1-Q}{Q}\frac{G(k(Q)v+G^{-1}(Q))}{1-G(k(Q)v+G^{-1}(Q))}}^{a} G^{-1}\left(\frac{t}{t+\left(1-Q\right)/Q}\right)\mathrm{d}t\Bigg\}.$$

We therefore estimate $\widehat{\delta}^a$ by the solution to

$$\max_{\delta^a}\left\{t_0'\delta^a + \frac{1}{N_a}\sum_{i=N_t+1}^{N}\varphi^+(t(X_i,\widehat{\zeta}^t)'\delta^a,\widehat{Q})\right\}.$$

The first order condition to this problem implies that $\widehat{\delta}^a$ solves

$$\frac{1}{N}\sum_{i=1}^{N} m_2(Z_i,\widehat{\rho},\widehat{\delta}^a) = \frac{1}{N}\sum_{i=1}^{N}\left(t_0 + \frac{1-D_i}{1-\widehat{Q}}\varphi_1^+(t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a,\widehat{Q})t(X_i,\widehat{\zeta}^t)\right) = 0, \tag{38}$$

where

$$\varphi_1^+(t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a,\widehat{Q}) = -\frac{1-\widehat{Q}}{\widehat{Q}}\frac{G(k(\widehat{Q})t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a + G^{-1}(\widehat{Q}))}{1-G(k(\widehat{Q})t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a + G^{-1}(\widehat{Q}))}.$$

The probability weights can then be recovered by $\widehat{\pi}_{ai} = -\varphi_1^+(t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a/N_a$. Finally $\widehat{\gamma}$ solves

$$\frac{1}{N}\sum_{i=1}^{N} m_3(Z_i,\widehat{\rho},\widehat{\delta}^a,\widehat{\gamma}) = -\frac{1}{N}\left\{\sum_{i=1}^{N} -\frac{D_i}{\widehat{Q}}\psi_1\left(Y_{1i},X_i,\widehat{\gamma}\right)\right. \tag{39}$$
$$\left. -\frac{\left(1-D_i\right)\varphi_1^+(t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a,\widehat{Q})}{1-\widehat{Q}}\psi_0\left(Y_{0i},X_i,\widehat{\gamma}\right)\right\}$$

When $G\left(v\right) = \exp\left[v\right]/\left[1+\exp\left[v\right]\right]$ our data combination estimator is particularly simple to implement. Appendix A shows that for this case $\varphi^+(v,Q) = -\exp\left[v\right]$, which is the GEL criterion function associated with the exponential tilting (ET) estimator. The second step of IPT in this case therefore involves reweighting the auxiliary sample by ET to match moments of X calculated using the target sample. It is well-known that, among GEL estimators, ET is computationally attractive (cf., Imbens, Spady and Johnson 1998).

4.2 Large sample properties

The large sample properties of $\widehat{\gamma}$ in data combination problems parallel those given above for the missing data case. Our estimator is consistent as long as one of Assumptions 3.5 or 3.6 hold and efficient if both hold. We are aware of no other estimators for the data combination family of

problems with similar asymptotic properties.[18]

Theorem 4.1 (Double Robustness) *Suppose Assumptions 4.1 to 4.4, either Assumption 3.5 or 3.6, and additional regularity conditions hold, then as* $N \to \infty$

$$\widehat{\gamma} \xrightarrow{p} \gamma_0$$

and

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \Upsilon_0),$$

with the form of Υ_0 *depending on whether Assumption 3.5 or 3.6 holds (see Appendix B).*

Proof. See Appendix B. ■

As in the missing data case, double robust consistency is due to the balancing properties of this estimator (i.e., because the auxiliary sample is tilted to match moments of the target sample). Imposing balancing in this way ensures that $m_3(Z_i, \rho_0, \delta_0, \gamma)$ will be uniquely mean zero at $\gamma = \gamma_0$ as long as Assumption 3.6 holds; even if the propensity score is misspecified.

IPT is also locally efficient for the SDC problem. The maximal asymptotic precision with which γ_0 can be estimated under the SDC setup has been characterized by Hahn (1998) and Chen, Hong and Tarozzi (2004, 2008) and is equal to the inverse of

$$\mathcal{J}(\gamma_0) = \mathbb{E}\left[\frac{p_0(X)}{Q_0}\Gamma_0(X)\right]' \mathbb{E}\left[\Phi_0(X)\right]^{-1} \mathbb{E}\left[\frac{p_0(X)}{Q_0}\Gamma_0(X)\right], \tag{40}$$

with $\Gamma_0(x) = \mathbb{E}\left[\partial\psi(Z, \gamma_0)/\partial\gamma' \middle| X = x\right]$ and

$$\begin{aligned}\Phi_0(x) = \left\{\frac{p_0(x)}{Q_0}\right\}^2 \Bigg\{&\frac{\Sigma_0(x;\gamma_0)}{1 - p_0(x)} + \frac{\Sigma_1(x;\gamma_0)}{p_0(x)} \\ &+ \frac{1}{p_0(x)}\left[q_1(x;\gamma_0) - q_0(x;\gamma_0)\right]\left[q_1(x;\gamma_0) - q_0(x;\gamma_0)\right]'\Bigg\}.\end{aligned} \tag{41}$$

If both Assumptions 3.5 and 3.6 hold simultaneously then IPT attains the bound given by (40).

Theorem 4.2 (Local Semiparametric Efficiency) *Suppose Assumptions 4.1 to 4.4, both Assumptions 3.5 or 3.6, and additional regularity conditions hold, then as* $N \to \infty$

$$\widehat{\gamma} \xrightarrow{p} \gamma_0$$

and

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \mathcal{J}(\gamma_0)^{-1}),$$

with $\mathcal{J}(\gamma_0)$ *as defined by (40) and (41).*

Proof. See Appendix B. ■

In Appendix B we provide a single variance-covariance estimator for $\widehat{\gamma}$ that is consistent for the cases covered by Theorems 4.1 and 4.2.

[18] Globally efficient estimators, based on non-parametric smooths, have been developed by Hahn (1998), Hirano, Imbens and Ridder (2003) and Chen, Hong and Tarozzi (2004, 2008).

4.3 Examples

To illustrate the application of IPT to data combination problems we discuss estimation of the ATT and TSIV model (with incompatible samples) in detail. Appendix C provides additional examples.

Average Treatment Effects for the Treated 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 on the Treated is

$$\gamma_0 = \mathbb{E}\left[Y_1 - Y_0 | D = 1\right] = \mathbb{E}_t\left[Y_1 - Y_0\right].$$

which corresponds to setting $\psi_1(Y_1, X, \gamma) = Y_1$ and $\psi_0(Y_0, X, \gamma) = Y_0 + \gamma$.

In this case the target population corresponds to the treated population. An example is provided by LaLonde (1986) who studies the effect of job-training on NSW participants. He has available a sample of NSW participants with measures of prior earnings and demographic characteristics, X, and post-training earnings, Y_1. Also available is an auxiliary sample taken from the Panel Study of Income Dynamics (PSID). This sample is not drawn from the same population as the NSW one. In the auxiliary sample LaLonde (1986) also observes prior earnings and demographics, X, and post-training earnings for *non-participants*, Y_0.

As with the ATE, efficient estimation of the ATT requires choosing $h(X)$ such that the propensity score in the parametric family $G\left(\alpha + h(X)'\beta\right)$ and the conditional expectation of Y_1 and Y_0 are linear functions of $h(X)$.

Two sample instrumental variables estimation with incompatible samples Currie and Yelowitz (2000) consider a model relating, Y_1, an indicator for whether a school-aged child has repeated a grade, with an indicator for residence in public housing, Y_0. Additional variables in the model are $X = (X_0, X_1')'$ with X_0 equalling the number of male siblings in the household, and X_1 equalling the overall number of siblings in the child's household and other household characteristics. The moment function is as given in Assumption 4.1 with $\psi_1(Y_1, X, \gamma) = XY_1$ and $\psi_0(Y_0, X, \gamma) = X\left(X_1'\gamma_{X_1} + Y_0'\gamma_{Y_0}\right)$.[19] The number of male siblings serves as an excluded instrument for 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.[20]

A total of N_t units are drawn from the target population and their realizations of (Y_1, X) recorded; this forms the target sample. An auxiliary random sample of size N_a is taken from the auxiliary population with realizations of (Y_0, X) recorded. Currie and Yelowitz (2000) get information on (Y_0, X) from the US Current Population Survey (CPS), a stratified random sample, and information on (Y_1, X) from a random subsample of the US Census. In this example the population of interest is school-aged children living in the United States. The US Census subsample can be viewed as a pure random sample from this population.

Since the CPS is not a random sample of the US population, the TSIV estimator of Angrist and Krueger (1992) is inconsistent (cf., Ridder and Moffitt 2007).[21] Other available estimators for the

[19] This moment corresponds to the textbook linear-IV model.

[20] This is because children of opposites sexes are not allowed to share rooms under HUD guidelines.

[21] In principal CPS sampling weights could be incorporated into Angrist and Krueger's (1992) procedure to ensure

TSIV model also require that the two samples be random ones from a common population (Arellano and Meghir 1992). Unfortunately, in many empirical applications estimated moments for variables common to the two samples differ significantly. Currie and Yelowitz (2000), for example, find significant differences in the probability of a household being female-headed and ethnic classification in the CPS and Census samples. IPT could be used to re-weight the CPS to match moments of X calculated using the Census. Theorems 4.1 and 4.2 provide conditions under which such a procedure is consistent and efficient.

5 Application of IPT some additional non-standard problems

An attractive feature of IPT is its applicability to a wide range of problems. In particular, the IPT estimator is often available in settings where inverse probability weighting, based on maximum likelihood estimates of the propensity score, is not. In this section we sketch the application of IPT to a few problems that do not belong to either our missing data or data combination families. Each of these problems has sparked its own, independent literature. The application of IPT to each of them is conceptually straightforward.

Additive Nonignorable attrition Hirano, Imbens, Ridder and Rubin (2001) consider the following two-period panel data problem. In period one Z_1 is observed for a random sample of N units. In period two Z_2 is observed for a subset of originally sampled units ($N_2 < N$). In period two an independent refreshment sample is taken and Z_2 recorded for N_r units. Let $D = 1$ denote the event of not attriting from the initial panel sample. This sampling structure asymptotically reveals $F(Z_1, Z_2|\, D = 1)$ as well as $F(Z_1)$ and $F(Z_2)$. The probability of not attriting, $p_0(z_1, z_2) = \Pr(D = 1|\, Z_1 = z_1, Z_2 = z_2)$, is also identified under quasi-additivity restrictions. For example, Hirano, Imbens, Ridder and Rubin (2001) show that if

$$p_0(z_1, z_2) = G\left(\alpha_0 + h(z_1)'\beta_{10} + h(z_2)'\beta_{20}\right), \tag{42}$$

α_0, β_{10} and β_{20} are identified. To apply IPT to this model we using the missing data class of discrepancies to reweight the $D = 1$ complete panel in order to match moments of the marginal distributions of Z_1 and Z_2. The former set of moments are identified by the original panel and the latter by the refreshment sample. Nevo (2003) applies a method similar to IPT to this problem.

Case-control studies In this example interest centers on estimating the parameter indexing the propensity score. Assume that $D \in \{0, 1\}$ is the binary outcome of interest. The researcher has access to a random sample of X from the (selected) subpopulation of individuals with outcome $D = 1$ and a separate random sample of X from the $D = 0$ subpopulation. The marginal probability $Q_0 = \Pr(D = 1)$ is known. Such a sampling scheme is often attractive when $D = 1$ is a rare event. The researcher has prior knowledge that

$$\Pr(D = 1|X) = G\left(\alpha_0 + h(X)'\beta_0\right).$$

consistency.

The Lagrange multipliers associated with a tilt of the $D = 0$ sample to match the means of each element of $h(X)$ in the $D = 1$ sample identify α_0 and β_0. The data combination family of discrepancies should be used with Q_0. The variance-covariance estimator associated Theorem 4.1 can be used for inference.

Case-control studies with contaminated controls In this example interest also centers on estimating the parameter indexing the propensity score. As before the researcher has access to a random sample of X from the (selected) subpopulation of individuals with outcome $D = 1$. Also available is either (i) an auxiliary random sample from the overall population which contains values of X for each sampled unit but no information on D or (ii) prior knowledge of $\zeta_0 = \mathbb{E}[h(X)]$ (e.g., from census cross tabs).

For example D might indicate whether an individual born in the 1970s is incarcerated in 2006, with X being a vector of childhood demographic and socioeconomic characteristics. A random sample of X from the $D = 1$ population is available from the Survey of Inmates in State and Federal Correctional Facilities. The marginal distribution of X characteristics is identified by an auxiliary random sample (e.g., from the 1980 census). The researcher has prior knowledge that

$$\Pr(D = 1|X) = G\left(\alpha_0 + h(X)'\beta_0\right),$$

with $h(X)$ a known function. In this case the estimated Lagrange multipliers from the inverse probability tilt of the $D = 1$ sample identify α_0 and β_0 (using the missing data class of discrepancies with Q_0). When ζ_0 is known the variance-covariance estimator associated Corollary 3.2 can be used for inference. When ζ_0 is estimated from an independent 'contaminated sample' new distribution theory is required. Lancaster and Imbens (1996) and Qin (1998) discuss this model is detail and suggest alternative estimators.

6 Monte Carlo experiments and application

6.1 Monte Carlo experiments

In this section we compare the small sample performance of IPT with that of several alternative missing data estimators. Specifically we consider the parametric inverse probability weighting estimator described by Wooldridge (2007), henceforth IPW; the non-parametric IPW estimator of Hirano, Imbens and Ridder (2003), henceforth HIR; the imputation estimator of Imbens, Newey and Ridder (2007) (with their data-dependent choice of smoothing parameter), henceforth INR; the conditional expectation projection estimator of Chen, Hong and Tarozzi (2008), henceforth CHT; and the augmented inverse probability weighting estimator of Robins, Rotnitzky and Zhao (1994) as described by Tsiatis (2006), henceforth AIPW.[22]

We assume that Y, the outcome of interest, is generated according to

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 \Phi\left(\frac{X - a}{b}\right) + U, \quad U|\, X \sim \mathcal{N}(0, 1/64),$$

[22] Matlab replication files for our Monte Carlo experiments as well as a technical Appendix, describing in detail our implementation of each estimator, is available online.

Table 1: Parameter values for the four Monte Carlo experiments.

	Design 1: 'smooth-smooth'	Design 2: 'rough-smooth'	Design 3: 'smooth-rough'	Design 4: 'rough-rough'
α_0	0.00000	-0.12510	0.00000	-0.12510
α_1	-0.25000	-0.25000	-0.25000	-0.25000
α_2	0.00000	0.50000	0.00000	0.50000
a	0.50000	0.50000	0.50000	0.50000
b	0.20000	0.20000	0.20000	0.20000
β_0	0.00000	0.00000	2.00000	2.00000
β_1	2.19722	2.19722	4.19722	4.19722
β_2	0.00000	0.00000	-4.00000	-4.00000
c	0.15000	0.15000	0.15000	0.15000
$\sqrt{\mathcal{I}(\gamma_0)/N}$	0.00825	0.00752	0.00785	0.00708

NOTES: The square root of Hahn's (1998) variance bound for each design (divided by $N^{1/2} = \sqrt{1,000}$) is reported in the last row of the table.

where $\Phi(\cdot)$ is the CDF of a standard normal random variable and $X \sim \mathcal{U}(-1,1)$. Observations of Y are missing at random with Y observed if $D = 1$ where

$$D = \mathbf{1}(\beta_0 + \beta_1 X + \beta_2 \Phi(X/c) - V),$$

with $V|X,Y$ logistic.

We consider four different data generating processes (DGPs), the parameterizations of which are given in Table 1. In the first design both the outcome CEF and the propensity score are smooth in X. In the second, the outcome CEF is inhomogenous, while the propensity score remains smooth. In the third, the CEF is smooth, while the propensity score is now inhomogenous. In the fourth design both the CEF and propensity score are inhomogenous. The smooth and inhomogenous outcome CEFs and propensity scores are rendered in Figure 1.

Each design is calibrated such that the propensity score ranges from 0.1 to 0.9 with a marginal probability of missingness equal to one half. The target estimand is $\gamma_0 = \mathbb{E}[Y]$, which is identically equal to zero for each design. For each experiment the sample size is set equal to $N = 1,000$ with $5,000$ Monte Carlo replications. The IPW, AIPW and IPT estimators are based upon a logistic model for the propensity score with X entering the index linearly. The HIR estimator is also based on a logistic model for the propensity score but with X entering quadratically.[23] The INR estimator is based on a polynomial series estimate of $\mathbb{E}[Y|X]$ with the number of series terms chosen to minimize the estimated mean square error of $\widehat{\gamma}$ (see Imbens, Newey and Ridder 2007). The CHT estimator is based on a polynomial series estimate of $\mathbb{E}[Y|X]$ with the number of series terms fixed at 3 (i.e., a quadratic approximation).

Our designs are chosen to highlight the strengths and weakness of each estimator. In the first design we expect all estimators to perform well. In the second design we expect the IPW and HIR to perform acceptably and the imputation-based INR and CHT estimators to perform poorly. In design 3 we expect the opposite pattern. The AIPW and IPT estimators, due to their double-robustness

[23] Hirano, Imbens and Ridder's (2003) theoretical results suggest using a series logit estimator with $N^{1/9}$ polynomial terms. For our designs $1000^{1/9} \approx 2.15$ which we round up to 3, yielding the quadratic specification.

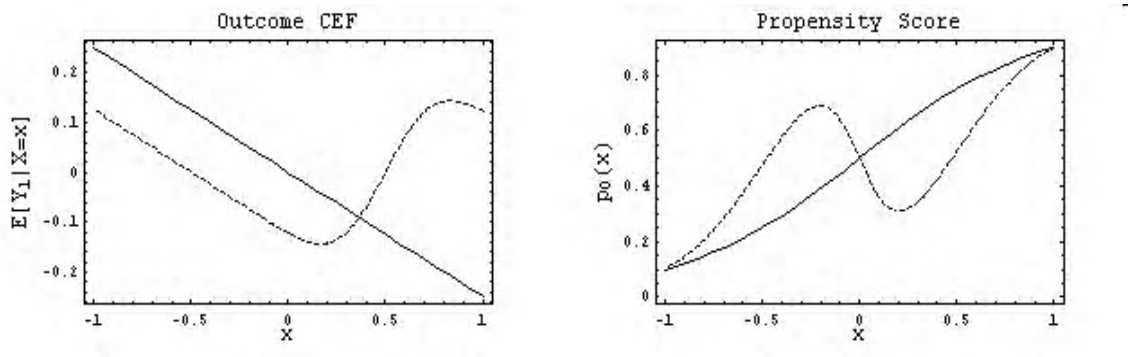

Figure 1: Conditional expectation function (CEF) of Y given X and the propensity score for the Monte Carlo experiments.

attribute, should perform well in both designs 2 and 3. In design 4 we expect all estimators to perform poorly.

Formally HIR, INR and CHT are consistent and attain Hahn's (1998) bound in all designs. IPW is never efficient but is consistent in designs 1 and 2. AIPW and IPT are consistent in designs 1, 2 and 3. In design 1 they attain Hahn's (1998) bound. In design 2 their large sample variance lies above Hahn's (1998) bound but below that of parametric IPW. In design 3 their large sample variance is actually smaller than Hahn's (1998) bound. Recall that Hahn's (1998) bound corresponds to the semiparametric missing data model defined by Assumptions 3.1 to 3.4. The imposition of Assumption 3.6 further lowers the bound. Therefore, under misspecification of the propensity score the large sample variance of IPT and AIPW lies between that of the efficient parametric imputation estimator and Hahn's (1998) bound.[24]

For purposes of comparison we also include an inverse probability weighting estimator based on the true propensity score. This estimator is consistent across all designs, but never efficient. We also include an infeasible oracle estimator: the mean of Y across all missing as well as non-missing observations. This estimator calibrates the cost of missingness in each of our designs.

Table 2 summarizes the results from the first design. As expected each estimator performs well, with small sample properties well-appoximated by asymptotic distribution theory. An approximate standard error for the Column 2 Monte Carlo scaled bias estimates is $\sqrt{(\pi/2)/5,000} \approx 0.0177$; differences in median bias are not significant across estimators for this design.

Table 3 reports results from the second design. In this design the imhomogeneity of $\mathbb{E}[Y|X]$ creates problems for both the INR and CHT imputation estimators, with each exhibiting significant median bias. Unsurprisingly, given the underlying smoothness of the propensity score, both the IPW and HIR estimators do well. In this design the HIR procedure involves an overfit of the propensity score. Consistent with the theoretical results of their paper, such overfitting results in improved precision (cf., Graham 2007 for a related discussion). The Monte Carlo sampling standard deviation of the HIR point estimates are about 10 percent lower than the corresponding

[24] Bang and Robins (2005, p. 966) comment that, when the propensity score is misspecified, but the outcome CEF is not, AIPW is often 'nearly' as efficient as parametric imputation. This assessment is based on their Monte Carlo experiments. In the designs whose results are reported in the lower half of Table 2 (p. 966) and the upper and lower portions of Table 4 (p. 970) of their paper, the AIPW estimator with a misspecified propensity score has a Monte Carlo sampling variance that lies below that of the AIPW estimator based on a correct model for the propensity score, but above that of the parametric imputation estimator. Our Design 3 results replicate this ordering.

Table 2: Monte Carlo results for Design 1

	(1) Asymptotic Bias	(2) Median Bias	(3) Asymptotic Std. Err.	(4) Median Std. Err.	(5) Standard Deviation	(6) Coverage of 95% CI
Oracle	0.0000	-0.0363	0.0060	0.0060	0.0060	0.952
True Weights	0.0000	-0.0304	0.0114	0.0113	0.0114	0.947
IPW	0.0000	-0.0217	0.0091	0.0090	0.0092	0.948
HIR	0.0000	-0.0359	0.0083	0.0084	0.0085	0.946
INR	0.0000	-0.0098	0.0083	0.0082	0.0082	0.950
CHT	0.0000	-0.0240	0.0083	0.0082	0.0083	0.949
AIPW	0.0000	-0.0162	0.0083	0.0082 [0.0082]	0.0083	0.949 [0.948]
IPT	0.0000	-0.0224	0.0083	0.0082	0.0083	0.947

NOTES: Each row corresponds to a specific estimator as described in the main text. Column 1 reports the scaled large sample bias of each estimator (i.e., its probability limit minus the true parameter divided by the square root of its large sample variance, $(AVar(\widehat{\gamma})/N)^{1/2}$). Column 2 reports the median Monte Carlo bias of each estimator scaled by its asymptotic standard error. Column 1 calibrates the scale of inconsistency for each estimator, while a comparison of Columns 1 and 2 allows for an assessment of whether an estimator's actual sampling distribution is centered at is probability limit. Column 3 gives the large sample standard error of each estimator (i.e., $(AVar(\widehat{\gamma})/N)^{1/2}$), Column 4 the median estimated standard error and column 5 the standard deviation of the point estimates across the 5,000 Monte Carlo replications. Column 6 reports the actual coverage of a 95 percent Wald-based confidence interval (the asymptotic variance estimators are described in a technical appendix available online). Two standard errors are provided for the AIPW estimator. The first assumes that the working model for both $E[Y|X]$and $p(X)$is correctly specified. This is the covariance estimator typically used in applied work (e.g., Lunceford and Davidian 2004). The second, given in the square brackets, treats the AIPW estimator as a sequential method-of-moments estimator with standard errors calculated accordingly. These standard errors are valid whenever AIPW is consistent. The mean number of series terms selected by the INR estimator is 2.25, while the median is 2.

Table 3: Monte Carlo results for Design 2

	(1) Asymptotic Bias	(2) Median Bias	(3) Asymptotic Std. Err.	(4) Median Std. Err.	(5) Standard Deviation	(6) Coverage of 95% CI
Oracle	0.0000	-0.0153	0.0050	0.0050	0.0049	0.954
True Weights	0.0000	-0.0207	0.0084	0.0083	0.0085	0.947
IPW	0.0000	-0.0149	0.0083	0.0082	0.0084	0.944
HIR	0.0000	-0.0031	0.0075	0.0075	0.0077	0.941
INR	0.0000	0.0849	0.0075	0.0080	0.0083	0.941
CHT	0.0000	0.0561	0.0075	0.0076	0.0076	0.948
AIPW	0.0000	-0.0150	0.0093	0.0105 [0.0092]	0.0095	0.968 [0.945]
IPT	0.0000	-0.0113	0.0080	0.0080	0.0081	0.945

NOTES: The mean number of series terms selected by the INR estimator is 3.1, while the median is 5. See notes to Table 2 for additional information.

Table 4: Monte Carlo results for Design 3

	(1) Asymptotic Bias	(2) Median Bias	(3) Asymptotic Std. Err.	(4) Median Std. Err.	(5) Standard Deviation	(6) Coverage of 95% CI
Oracle	0.0000	0.0217	0.0060	0.0060	0.0061	0.950
True Weights	0.0000	-0.0360	0.0104	0.0103	0.0105	0.943
IPW	-0.4042	-0.4181	0.0074	0.0073	0.0075	0.933
HIR	0.0000	-0.3977	0.0079	0.0073	0.0075	0.932
INR	0.0000	-0.0061	0.0079	0.0074	0.0078	0.936
CHT	0.0000	0.0014	0.0079	0.0074	0.0076	0.943
AIPW	0.0000	-0.0005	0.0074	0.0074 [0.0074]	0.0076	0.945 [0.942]
IPT	0.0000	0.0017	0.0074	0.0073	0.0076	0.942

NOTES: The mean number of series terms selected by the INR estimator is 3.06, while the median is 2. See notes to Table 2 for additional information.

IPW estimates (cf., Column 5).

Both AIPW and IPT estimators also perform well. When both $\mathbb{E}[Y|X]$ and $p(X)$ are correctly modelled the two estimators are first-order equivalent and, if the results from Design 1 are indicative, have similar small sample properties. However under partial misspecification they are no longer first-order equivalent. In design 2, while both are consistent, their asymptotic variances differ, with IPT being more precisely determined (Columns 1 and 3 of Table 3). The small sample properties of the two estimators mirror their large sample ones. Both are approximately median unbiased, but the sampling variability of the IPT estimator is about 10 percent lower than that of AIPW (Columns 2 and 5 of Table 3).

Table 4 reports results from the third design. In this case the IPW and HIR perform extremely poorly, reflecting their fragility vis-a-vis misspecification of the propensity score. The remaining estimators all perform well. Importantly, the AIPW and IPT estimators, while based on misspecified models for the propensity score, are nevertheless consistent. However, as in design 2, their asymptotic variances differ with IPT's again being smaller, albeit negligibly so. These large sample properties are reflected in small samples. In design 3 both AIPW and IPW are approximately median unbiased with similar sampling variances (Columns 2 and 5 of Table 4).

Table 5 reports results from the our fourth, and final, design. In this design the AIPW and IPT estimators are formally inconsistent as is IPW. Formally HIR, INR and CHT are consistent and efficient. In practice all estimators, with the exception of IPT, exhibit significant median bias.

The performance of IPT across designs 2 and 3 highlights the practical value of using a 'doubly robust' estimator. If either the propensity score or outcome CEF is well approximated by the implicit parametric models used, then IPT will perform well. While the HIR, INR and CHT estimators have attractive large sample properties irrespective of the form of the propensity score and outcome CEF, in practice their small performance is highly sensitive to the smoothness of one of these two functions. The HIR estimator performs well when the propensity score is smooth, but very poorly when it is not. The INR and CHT imputation estimators, in contrast, perform best when the outcome response is smooth.

Table 5: Monte Carlo results for Design 4

	(1) Asymptotic Bias	(2) Median Bias	(3) Asymptotic Std. Err.	(4) Median Std. Err.	(5) Standard Deviation	(6) Coverage of 95% CI
Oracle	0.0000	-0.0024	0.0050	0.0050	0.0051	0.946
True Weights	0.0000	-0.0159	0.0081	0.0080	0.0081	0.948
IPW	0.0691	0.0442	0.0071	0.0070	0.0071	0.946
HIR	0.0000	0.0513	0.0071	0.0066	0.0066	0.947
INR	0.0000	0.2631	0.0071	0.0065	0.0073	0.913
CHT	0.0000	0.9864	0.0071	0.0066	0.0067	0.803
AIPW	-0.0790	-0.1118	0.0073	0.0073 [0.0073]	0.0073	0.947 [0.949]
IPT	0.0536	0.0269	0.0071	0.0071	0.0071	0.946

NOTES: The mean number of series terms selected by the INR estimator is 5, while the median is 5. See notes to Table 2 for additional information.

The Monte Carlo experiments also highlight differences between the AIPW estimator of Robins, Rotnitzky and Zhao (1994) and our IPT procedure. Under misspecification of either $\mathbb{E}[Y|X]$ or $p(X)$, but not both simultaneously, they remain consistent but have different large sample variances. In the designs considered here IPT has a smaller asymptotic sampling variance than AIPW in such cases.

6.2 Application to National Supported Work (NSW) demonstration

LaLonde (1986), in an influential paper, compared a variety of non-experimental program evaluation estimators with an experimental benchmark using data from the National Supported Work (NSW) demonstration, the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS). Dehejia and Wahba (1999) revisited LaLonde's work using newer non-experimental program evaluation estimators. In this section we examine LaLonde's data once again. We use IPT to estimate the effect of NSW participation on post-treatment earnings. As our purposes are more illustrative than substantive, we refer the reader to LaLonde (1986) and Dehejia and Wahba (1999) for full details on the NSW demonstration and the particular data extracts employed here.[25]

The evaluation dataset consists of 297 NSW participants and 425 controls. Assignment to participation was random, hence a difference in post-intervention earnings across NSW participants and non-participants (in the evaluation dataset) can serve as a benchmark against which alternative, non-experimental, estimators can be compared. A sample of (presumably) non-participants drawn from the PSID and CPS is also available. These samples will form the basis of our non-experimental estimate of NSW participation on post-treatment earnings (i.e., the average treatment effect on the treated). In the language of Section 4 the subsample of NSW participants is the target sample. The samples of non-participants drawn from the PSID and the CPS serve as auxiliary samples.

Table 6 reports the means of a variety of demographic characteristics as well as pre- and post-intervention earnings amongst NSW participants, a subsample of NSW participants for which (pre-intervention) 1974 as well as 1975 earnings are available, and the PSID and CPS controls. There

[25] The data we used was accessed from `http://www.nber.org/%7Erdehejia/nswdata.html` in the fall of 2007.

Table 6: Means of pre-intervention covariates amongst NSW participants and PSID/CPS 'controls'

	(1) NSW Participants	(2) Dehejia-Wahba Sample	(3) PSID Controls	(4) CPS Controls
Age	24.63	25.82	34.85	33.23
Years-of-Schooling	10.38	10.35	12.12	12.03
Black	0.80	0.84	0.25	0.07
Hispanic	0.09	0.06	0.03	0.07
Married	0.17	0.19	0.87	0.71
No Degree	0.73	0.71	0.31	0.30
1974 Earnings	-	2,096	19,429	14,017
1975 Earnings	3,066	1,532	19,063	13,651
1978 Earnings (post intervention)	5,976	6,349	21,554	14,847
N	297	185	2,490	15,992

NOTES: The 'Dehejia-Wahba Sample' of Column (2) includes the 185 NSW participants with earnings information available for both 1974 and 1975. LaLonde (1986) and Dehejia and Wahba (1999) provide full details on each sample.

are substantial differences in the distribution of pre-treatment characteristics across the different samples. Following the suggestion of Dehejia and Wahba (1999) we use the Column 2 restricted subsample of NSW participants as our 'target sample' in the analysis which follows. Our target sample only includes the 185 NSW participants for which we observe two years of pre-treatment earnings.[26]

Using logistic regression we fitted a model for the propensity score using a merged sample consisting of NSW participants and either the PSID or CPS controls. In both cases the propensity score took a logit form with each of the pre-treatment characteristics listed in Table 6 entering linearly. For non-binary characteristics, specifically age, years-of-schooling and the logarithm of pre-treatment earnings in 1974 and 1975, we also included squared terms. While Dehejia and Wahba (1999) suggest using a somewhat richer model for the propensity score, we nevertheless maintain this specification throughout to keep our application as transparent and simple as possible.

Column 2 of Table 7 reports the difference in the means, and for non-binary variables, the variances[27] of pre-treatment characteristics in the target sample and the corresponding HIR inverse probability weighted means (based on the propensity score specification described above) of both the PSID (Panel A) and CPS (Panel B) controls. For comparison, Column 1 gives the raw difference in (unweighted) means. Inverse probability weighting does lead to considerably more balance in pre-treatment characteristics across NSW participants and controls drawn from both for the PSID and the CPS. This balance, however, is far from perfect. Inverse probability weighting has difficulty balancing the mean and variance of pre-intervention earnings (particularly when using the PSID as a control sample).

[26] Our benchmark estimates of the effect of NSW participation on earnings also use just those NSW controls for which we observe two years of pre-treatment earnings.

[27] Actually we report uncentered second moments.

Table 7: Difference between means of pre-intervention covariates amongst NSW participants and PSID/CPS 'controls' before and after reweighting

	Panel A: PSID Controls			Panel B: CPS Controls		
	(1)	(2)	(3)	(1)	(2)	(3)
	No Weights	HIR-IPW Weights	IPT Weights	No Weights	HIR-IPW Weights	IPT Weights
Age	9.03	-2.33	0	7.41	0.65	0
Age^2	606.14	-144.01	0	508.51	37.02	0
Years-of-Schooling	1.77	-0.06	0	1.68	-0.03	0
$(\text{Years-of-Schooling})^2$	45.26	0.10	0	41.84	-1.09	0
Black	-0.59	-0.06	0	-0.77	0.02	0
Hispanic	-0.03	0.11	0	0.01	-0.01	0
Married	0.68	-0.12	0	0.52	-0.02	0
No Degree	-0.40	-0.01	0	-0.41	0.01	0
log(1974 Earnings)	6.46	-1.04	0	5.77	-0.09	0
$\log(\text{1974 Earnings})^2$	66.56	-9.14	0	56.96	-0.85	0
log(1975 Earnings)	5.68	1.04	0	5.19	0.09	0
$\log(\text{1974 Earnings})^2$	61.69	3.38	0	53.71	-0.22	0

NOTES: The 'HIR-IPW Weight' attached to the i^{th} unit (with $D_i = 0$) equals $\frac{p(X_i,\pi)}{1-p(X_i,\pi)} / \sum_{i=1}^{N} (1 - D_i) \frac{p(X_i,\pi)}{1-p(X_i,\pi)}$ (see Hirano, Imbens and Ridder 2003). The IPT weights are as described in Section 4 of the text. The target sample includes the 185 NSW participants with two years of pre-intervention earnings data.

Table 8: Estimates of the ATT: NSW Demonstration

	(1) PSID Controls	(2) CPS Controls
Experimental	1,794	1,794
	(669)	(669)
HIR	−1,370	1,012
	(701)	(834)
CHT	3,240	1,308
	(5,415)	(812)
IPT	2,031	1,068
	(752)	(727)

NOTES: The 'Experimental' ATT estimate is a difference between the raw means of the 1978 earnings of NSW participants and non-participants in the NSW evaluation dataset. The included group of participants and non-participants are those for which earnings information in both 1974 and 1975 is available (see Dehejia and Wahba (1999) for additional details).

By construction, the IPT estimator, reweights the PSID and CPS data to exactly match the sample moments of the NSW participants target sample.[28] Table 7 verifies this claim, showing that after inverse probability tilting the differences in the two sets of sample moments are identically zero (Column 3).

In Table 8 we report estimates of average effect of NSW participation, amongst participants, on post-intervention earnings. (i.e., the ATT). The top row of the panel reports the 'experimental' ATT estimates based on the NSW evaluation dataset. The second two rows report estimates based on the Hirano, Imbens and Ridder (2003) and Chen, Hong and Tarozzi (2008) procedures.[29]

Both the HIR and CHT estimators dramatically fail to reproduce the experimental benchmark ATT estimate when using the PSID controls. When using the CPS controls they reproduce the experimental benchmark up to sampling error. In contrast, our IPT estimates, reproduce the experimental benchmark, again up to sampling error, for both sets of controls.

7 Summary

In this paper we have proposed a new estimator for semiparametric missing data and data combination problems. For both classes of problems our estimator is locally efficient and doubly robust. For the class of missing data problems our procedure is first-order equivalent to the AIPW estimator of Robins, Rotnitzky and Zhao (1994) when both the conditional mean of the identifying moment and the propensity score are correctly specified. Under partial misspecification, however, the two procedures, whilst both remaining consistent, have different large sample variances. For the class of data combination problems we are aware of no alternative to IPT with comparable properties.

The attractive theoretical properties of our inverse probability tilting procedure stem from its distinctive 'exact balancing' property. Our re-visit of LaLonde's (1986) NSW evaluation and a series of Monte Carlo experiments illustrate the good small sample performance of our procedure. While

[28] When 'overlap' is limited a valid set of balancing weights may not exist.

[29] Our implementation of Chen, Hong and Tarozzi (2008) is based on a least squares fit of NSW post-intervention earnings on the same terms entering our HIR propensity score model.

IPT performs very well relative to several alternative estimation procedures in the Monte Carlo designs we considered, before making strong prescriptions to practitioners more research is required (cf., Imbens 2004).

Our basic estimation strategy is also applicable, with minor modification, to other, less standard, missing data problems such as the additive nonignorable attrition model of Hirano, Imbens, Ridder and Rubin (1998) or case-control designs. A feature of these problems is that a maximum likelihood estimate of the propensity score is unavailable, rendering several other missing data methods unusable. Appendix C discusses a number of additional problems to which IPT might be fruitfully applied.

Developing connections between our approach and the recent literature on the estimation of moment condition models by generalized empirical likelihood (GEL) would be an interesting area for additional research. Our family of missing data discrepancies can be used, in a manner entirely analogous to the GEL estimators described by Newey and Smith (2004), to estimate moment condition models. In such cases κ in (11) may be chosen to calibrate the higher-order properties of the IPT estimates. In that case it is possible to show that the higher-order bias of IPT is given by Theorem 4.2 of Newey and Smith (2004, p. 228); with $\varphi_3^+(0,\kappa)$ replacing the ρ_3 term in their expression. This term is given by

$$\varphi_3^+(0,\kappa) = \frac{G_2\left(G^{-1}(\kappa)\right)}{G_1\left(G^{-1}(\kappa)\right)^2}\kappa - 2.$$

Therefore choosing κ such that $G_2\left(G^{-1}(\kappa)\right)\kappa/G_1\left(G^{-1}(\kappa)\right)^2 = 0$ results in an estimator with the same bias as the oracle GMM estimator which chooses the parameter by setting the optimal linear combination of sample moments equal to zero. For $G(\cdot)$ logistic $\varphi_3^+(0,\kappa) = -2$ for $\kappa = 1/2$. This is equivalent to IPT estimation using a dual contrast function of

$$\varphi^+(v,Q) \propto -\frac{v}{2} - \frac{1}{4}\exp[2v].$$

It would be interesting to compare the small sample performance of this estimator with that of various GEL estimators, for example, empirical likelihood (EL).

Returning to missing data problems, an open question is whether IPT has attractive higher-order properties relative to, say, AIPW. Our Monte Carlo experiments establish that the two estimators are distinct in practice as well as theory. For the case with data missing completely at random (MCAR) we conjecture that IPT does exhibit attractive higher-order bias properties. Verifying this conjecture would require taking stochastic expansions of each estimator. While this appears feasible it is beyond the scope of this paper.

Finally, it should also be possible to develop globally efficient IPT estimators by allowing the dimension of $h(X)$ to increase with the sample size at the appropriate rate.

Appendices

A Computation

In this appendix we outline our algorithm for computing IPT point estimates of γ_0. As noted in the main text, estimation consists of three steps. The first step is standard, involving only the computation of sample means. The second step is the calculation of the IPT weights. This step can be computationally challenging. The third step involves weighted M-estimation and is application specific. Here we provide details on the first and second step of estimation. Our proposed algorithm is a generalization of an idea developed by Owen (2001) for the computation of empirical likelihood confidence intervals.

A.1 Missing data problems

We being by computing the full sample means of D and $h(X)$:

$$\widehat{Q} = \frac{1}{N}\sum_{i=1}^{N} D_i, \qquad \widehat{\zeta} = \frac{1}{N}\sum_{i=1}^{N} h(X_i).$$

We then solve for the inverse probability tilt of the $D = 1$ subsample:

$$\max_{\delta^1 \in \widehat{\Delta}_{N_1}} \; l_{N_1}\left(\delta^1; \widehat{\rho}\right), \tag{43}$$

where $\widehat{\rho} = (\widehat{Q}, \widehat{\zeta}')'$, the set $\widehat{\Delta}_{N_1}$ is defined below and

$$l_{N_1}\left(\delta^1; \rho\right) = t_0'\delta^1 + \frac{1}{N_1}\sum_{i=1}^{N_1} \varphi^+(t(X_i,\zeta)'\delta^1, Q).$$

The form of $\varphi^+(v,Q)$ is given by Equation (18) of the main text. For future reference we also have the first and second derivatives

$$\varphi_1^+(v,Q) = -\frac{Q}{G\left(k(Q)\,v + G^{-1}(Q)\right)}, \quad \varphi_2^+(v,Q) = \frac{G_1\left(k(Q)\,v + G^{-1}(Q)\right)}{G\left(k(Q)\,v + G^{-1}(Q)\right)^2} k(Q)\,Q$$

with $k(Q) = -Q/G_1\left(G^{-1}(Q)\right) < 0$. It is straightforward to verify that $\varphi_1^+(0,Q) = \varphi_2^+(0,Q) = -1$, and hence that the Fenchel conjugate is appropriately normalized. Also recall that $\delta^1 = \left(\eta^1, \lambda^{1\prime}\right)'$ are the $1+M$ Lagrange multiplier(s) on the adding up and moment constraints in the primal problem (13), $t_0 = (1, \underline{0}')'$ and

$$t(X_i,\zeta) = \begin{pmatrix} 1 \\ h(Z_i,\zeta) \end{pmatrix}.$$

It is also helpful to establish the notational conventions $t_i(\zeta) = t(X_i,\zeta)$ and $t_i = t(X_i,\zeta_0)$.

As noted in the main text, inspection of the minimum discrepancy problem (13) shows that a valid solution requires each of the empirical probabilities to be bounded below by Q_0/N_1 and above by 1. The first restriction holds automatically since $G(\cdot)$ is increasing with $G(\infty) = 1$. The second restriction, however, implies a substantive domain restriction on $l_{N_1}\left(\delta^1; \rho\right)$. In particular, excluding improper probability weights requires that the maximization (13) occurs over the set

$$\widehat{\Delta}_{N_1} \overset{def}{\equiv} \left\{\delta^1 : Q_0/N_1 < -\varphi_1^+(t_i(\widehat{\zeta})'\delta^1, \widehat{Q})/N_1 < 1, \qquad i = 1,\ldots,N_1\right\}.$$

Instead of imposing these N_1 nonlinear constraints directly we extend an idea due to Owen (2001). To describe this idea it is helpful to first analyze the structure of the maximization problem in more detail. Differentiating $l_{N_1}\left(\delta^1; \rho\right)$ with respect to δ^1 gives a gradient vector of

$$\underset{1+M\times 1}{\nabla_{\delta^1} l_{N_1}\left(\delta^1; \rho\right)} = t_0 + \frac{1}{N_1}\sum_{i=1}^{N_1} \varphi_1^+(t_i(\zeta)'\delta^1, Q)\,t_i(\zeta). \tag{44}$$

Note that (44) will have an 'exploding denominator' when $t_i(\zeta)'\delta^1$ is large and positive. Differentiating again gives a Hessian matrix of

$$\underset{1+M\times 1+M}{\nabla_{\delta^1\delta^1} l_{N_1}\left(\delta^1; \rho\right)} = \frac{1}{N_1}\sum_{i=1}^{N_1} \varphi_2^+(t_i(\zeta)'\delta^1, Q)\,t_i(\zeta)t_i(\zeta)', \tag{45}$$

which is a negative semi-definite function of δ^1.[30] Solving for $\widehat{\delta}^1$ therefore involves maximizing a concave function over a convex domain. This follows from the outer-product structure of (45) and strict concavity of $\varphi^+(v, Q)$.

In order to avoid maximization over a restricted domain we redefine $l_{N_1}\left(\delta^1;\rho\right)$ so that it is concave in δ^1 over $\mathbb{R}^{1+M}$ without changing its value near the solution (cf., Owen 2001). At any valid solution the estimated probabilities must lie between Q_0/N_1 and 1. As noted previously the lower-bound will be satisfied automatically for any valid IPT contrast function. Let $v_i = t_i(\zeta)'\,\delta^1$, then the second inequality requires that

$$v_i < \frac{1}{k(Q_0)}\left[G^{-1}\left(\frac{Q_0}{N_1}\right) - G^{-1}(Q_0)\right], \qquad i = 1, \ldots, N_1$$

where the left-hand side of the above expression is a positive number. Let $v^*_{N_1} = \frac{1}{k(Q_0)}\left[G^{-1}\left(\frac{Q_0}{N_1}\right) - G^{-1}(Q_0)\right]$. Observe that $v^*_{N_1} \to \infty$ as $N_1 \to \infty$, suggesting that, in large enough samples, computation can occur without explicitly imposing the domain restriction.

Our modified estimator replaces $\varphi^+(v, Q_0)$ in (43) with the hybrid function

$$\varphi^\circ(v, Q_0) \stackrel{def}{\equiv} \begin{cases} \varphi^+(v, Q_0) & v < v^*_{N_1} \\ a_{N_1} + b_{N_1} v + \frac{1}{2} c_{N_1} v^2 & v \geq v^*_{N_1} \end{cases},$$

where a_{N_1}, b_{N_1} and c_{N_1} are the solutions to

$$\begin{aligned} c_{N_1} &= \varphi_2^+(v^*_{N_1}, Q_0) \\ b_{N_1} + c_{N_1} v^*_{N_1} &= \varphi_1^+(v^*_{N_1}, Q_0) \\ a_{N_1} + b_{N_1} v^*_{N_1} + \frac{c_{N_1}}{2}(v^*_{N_1})^2 &= \varphi^+(v^*_{N_1}, Q_0). \end{aligned}$$

This choice of coefficients ensures that $\varphi^\circ(v^*_{N_1}, Q_0)$ equals $\varphi^+(v^*_{N_1}, Q_0)$ and also equality of first and second derivatives at $v^*_{N_1}$.

Solving these equations yields

$$\begin{aligned} c_{N_1} &= \varphi_2^+(v^*_{N_1}, Q_0) \\ b_{N_1} &= \varphi_1^+(v^*_{N_1}, Q_0) - \varphi_2^+(v^*_{N_1}, Q_0) v^*_{N_1} \\ a_{N_1} &= \varphi^+(v^*_{N_1}, Q_0) - \varphi_1^+(v^*_{N_1}, Q_0) v^*_{N_1} + \frac{\varphi_2^+(v^*_{N_1}, Q_0)}{2}\left(v^*_{N_1}\right)^2. \end{aligned}$$

We then estimate δ^1 by solving

$$\max_{\delta^1 \in \mathbb{R}^{1+M}} l^\circ_{N_1}\left(\delta^1;\widehat{\rho}\right),$$

where

$$l^\circ_{N_1}\left(\delta^1;\rho\right) \stackrel{def}{\equiv} t_0'\delta^1 + \frac{1}{N_1}\sum_{i=1}^{N_1} \varphi^\circ(t_i(\zeta)'\delta^1, Q). \tag{46}$$

Since $l^\circ_{N_1}\left(\delta^1;\rho\right)$ coincides with $l_{N_1}\left(\delta^1;\rho\right)$ at valid solutions, solving the modified problem yields the same solution as the original one. In practice it is useful to check that the domain restrictions are satisfied at the candidate solution (i.e., the solution is a valid one). This can be done by checking that the probability weights sum to one and are all bounded below by Q_0/N_1 and above by 1. In addition to providing a simple way to avoid maximization over a restricted domain, using (46) in place of (43) avoids numerical problems caused by a 'exploding denominator' in (44).

We calculate $\widehat{\delta}^1$ by applying gradient-based procedures to maximize $l^\circ_{N_1}\left(\delta^1;\widehat{\rho}\right)$. This is a concave unconstrained maximization problem and hence straightforward to solve. The first and second derivatives are given by (44) and (45) above with $\varphi_1^\circ(v, Q)$ and $\varphi_2^\circ(v, Q)$ replacing $\varphi_1^+(v, Q)$ and $\varphi_2^+(v, Q)$ where

$$\begin{aligned} \varphi_1^\circ(v, Q) &= \mathbf{1}\left(v < v^*_{N_1}\right) \cdot \varphi_1^+(v, Q) + \mathbf{1}\left(v \geq v^*_{N_1}\right) \cdot (b_{N_1} + c_{N_1} v) \\ \varphi_2^\circ(v, Q) &= \mathbf{1}\left(v < v^*_{N_1}\right) \cdot \varphi_2^+(v, Q) + \mathbf{1}\left(v \geq v^*_{N_1}\right) \cdot c_{N_1}. \end{aligned}$$

To get starting values for this procedure we take a Taylor expansion of (44) around $\widehat{\delta}^1 = 0$ to get

$$0 \simeq t_0 - \frac{1}{N_1}\sum_{i=1}^{N_1} t_i(\widehat{\zeta}) - \left\{\frac{1}{N_1}\sum_{i=1}^{N_1} t_i(\widehat{\zeta}) t_i(\widehat{\zeta})'\right\}\widehat{\delta}^1,$$

[30] As long as the $t_i(\zeta)$ do not lie in a linear subspace of dimension less than $1 + M$, it is a negative definite function of δ^1.

and use the solution $\widehat{\delta}^1 = \left(\sum_{i=1}^{N_1} t_i(\widehat{\zeta}) t_i(\widehat{\zeta})'\right)^{-1} \left(N_1 t_0 - \sum_{i=1}^{N_1} t_i(\widehat{\zeta})\right)$ as starting values. Alternatively a $1+M$ vector of zeros can serve as starting values.

Calculation of $\widehat{\delta}^0$ is entirely parallel with N_0 and $1-\widehat{Q}$ replacing N_1 and $\widehat{Q}$ throughout and summation occurring from N_1+1 to N instead of 1 to N_1.

Computational details for inverse logistic tilting In this subsection we provide closed form expressions for $\varphi(v,Q)$, $\varphi^+(v,Q)$ and $\varphi^\circ(v,Q)$ when $G(v)$ takes the logistic form. We begin by noting that $G^{-1}(Q) = \ln\left(\frac{Q}{1-Q}\right)$ and hence that

$$k(Q) = -Q/G_1\left(G^{-1}(Q)\right) = -(1-Q)^{-1}.$$

To derive the closed form expressions for $\varphi(v,Q)$ and $\varphi^+(v,Q)$ given in the paper involves applications of 'integration by substitution':

$$\int_{t=a}^{t=b} f(t)\,\mathrm{d}t = \int_{u=h^{-1}(a)}^{u=h^{-1}(b)} f(h(u))\,h'(u)\,\mathrm{d}u, \tag{47}$$

where $h'(u) \neq 0$ for all $u \in [a,b]$.

We begin with the MD contrast function, $\varphi(v,Q)$. Let $u = \frac{Q}{t}/\left(1-\frac{Q}{t}\right) = \frac{Q}{t-Q}$ and hence $t = \left(\frac{1+u}{u}\right)Q = h(u)$, we then have

$$\begin{aligned}
\varphi(v,Q) &= -\frac{v}{k(Q)}G^{-1}(Q) - \frac{1}{k(Q)}\int_v^a G^{-1}(Q/t)\,\mathrm{d}t \\
&= v(1-Q)\ln\left(\frac{Q}{1-Q}\right) + (1-Q)\int_{\frac{Q}{v-Q}}^{\frac{Q}{a-Q}} \ln(u)\left(-\frac{Q}{u^2}\right)\mathrm{d}u \\
&\propto (v-Q)\ln(v-Q) - v\ln(1-Q) - (v-Q)
\end{aligned}$$

which is a normalized version of Nevo's (2002) generalized exponential tilting contrast function. The normalized criterion for the dual problem, $\varphi^+(v,Q)$, is

$$\begin{aligned}
\varphi^+(v,Q) &= -\frac{1}{k(Q)}\left[\frac{k(Q)v + G^{-1}(Q)}{G\left(k(Q)v + G^{-1}(Q)\right)}Q + \int_{Q/G\left(k(Q)v+G^{-1}(Q)\right)}^{a} G^{-1}\left(\frac{Q}{t}\right)\mathrm{d}t\right] \\
&= -\frac{1}{k(Q)}\frac{k(Q)v + G^{-1}(Q)}{G\left(k(Q)v + G^{-1}(Q)\right)}Q - \frac{1}{k(Q)}\int_{\frac{G\left(k(Q)v+G^{-1}(Q)\right)}{1-G\left(k(Q)v+G^{-1}(Q)\right)}}^{\frac{Q}{a-Q}} \ln(u)\left(-\frac{Q}{u^2}\right)\mathrm{d}u
\end{aligned}$$

where the second line follows from integration by substitution with, once again, $u = \frac{Q}{t}/\left(1-\frac{Q}{t}\right) = \frac{Q}{t-Q}$. Continuing we have

$$\begin{aligned}
\varphi^+(v,Q) &= -\frac{1}{k(Q)}\frac{k(Q)v + G^{-1}(Q)}{G\left(k(Q)v + G^{-1}(Q)\right)}Q + \frac{Q}{k(Q)}\left[\frac{u^{-1}}{-1}\ln(u) - u^{-1}\right]_{\frac{G\left(k(Q)v+G^{-1}(Q)\right)}{1-G\left(k(Q)v+G^{-1}(Q)\right)}}^{\frac{Q}{a-Q}} \\
&\propto -\frac{Q}{k(Q)}\frac{k(Q)v + G^{-1}(Q)}{G\left(k(Q)v + G^{-1}(Q)\right)} \\
&\quad - \frac{Q}{k(Q)}\left[-\frac{1-G\left(k(Q)v+G^{-1}(Q)\right)}{G\left(k(Q)v+G^{-1}(Q)\right)}\ln\left(\frac{G\left(k(Q)v+G^{-1}(Q)\right)}{1-G\left(k(Q)v+G^{-1}(Q)\right)}\right) - \frac{1-G\left(k(Q)v+G^{-1}(Q)\right)}{G\left(k(Q)v+G^{-1}(Q)\right)}\right].
\end{aligned}$$

Using the facts that

$$\frac{G(v)}{1-G(v)} = e^v, \qquad \frac{1}{G(v)} = 1 + e^{-v},$$

to simplify further gives

$$\varphi^+(v,Q) \propto -vQ - Q(1-Q)\exp\left[\frac{v}{1-Q} - \ln\left(\frac{Q}{1-Q}\right)\right].$$

For completeness the following are handy for computation

$$\begin{aligned}
\varphi_1^+(v,Q) &= -Q - Q\exp\left[\frac{v}{1-Q} - \ln\left(\frac{Q}{1-Q}\right)\right] \\
\varphi_2^+(v,Q) &= -\frac{Q}{1-Q}\exp\left[\frac{v}{1-Q} - \ln\left(\frac{Q}{1-Q}\right)\right].
\end{aligned}$$

Calculating the modified version of the criterion function, $\varphi^\circ(v,Q)$, is straightforward, but tedious. Using the

general expression for $v^*_{N_1}$ given above yields

$$v^*_{N_1} = (1-Q)\ln\left(\frac{N_1-Q}{1-Q}\right),$$

and therefore

$$\begin{aligned}
\varphi^+(v^*_{N_1}, Q) &= -Q(1-Q)\ln\left(\frac{N_1-Q}{1-Q}\right) - (1-Q)(N_1-Q) \\
\varphi_1^+(v^*_{N_1}, Q) &= -N_1 \\
\varphi_2^+(v^*_{N_1}, Q) &= -\frac{N_1-Q}{1-Q}.
\end{aligned}$$

Using these expressions we can solve for the coefficients in the quadratic portion of $\varphi^\circ(v, Q)$:

$$\begin{aligned}
c_{N_1} &= -\frac{N_1-Q}{1-Q} \\
b_{N_1} &= -N_1 + (N_1-Q)\ln\left(\frac{N_1-Q}{1-Q}\right) \\
a_{N_1} &= -(1-Q)(N_1-Q)\left\{\frac{1}{2}\left[\ln\left(\frac{N_1-Q}{1-Q}\right)\right]^2 - \ln\left(\frac{N_1-Q}{1-Q}\right) + 1\right\}.
\end{aligned}$$

This gives a modified criterion function of

$$\varphi^\circ(v,Q) = \begin{cases}
-xQ - Q(1-Q)\exp\left[\frac{v}{1-Q} - \ln\left(\frac{Q}{1-Q}\right)\right] & v < (1-Q)\ln\left(\frac{N_1-Q}{1-Q}\right) \\
= -(1-Q)(N_1-Q)\left\{\frac{1}{2}\left[\ln\left(\frac{N_1-Q}{1-Q}\right)\right]^2 - \ln\left(\frac{N_1-Q}{1-Q}\right) + 1\right\} & \\
\quad - \left[N_1 - (N_1-Q)\ln\left(\frac{N_1-Q}{1-Q}\right)\right]v & v \geq (1-Q)\ln\left(\frac{N_1-Q}{1-Q}\right) \\
\quad -\frac{1}{2}\frac{N_1-Q}{1-Q}v^2 &
\end{cases}.$$

Computational details for inverse linear probability tilting The CDF of a uniform random variable on $[-1,1]$ is given by $G(v) = \frac{1}{2}(1+v)$ and hence $G^{-1}(v) = 2v-1$. This gives $k(Q) = -2Q$ and hence an empirical discrepancy of

$$\begin{aligned}
\varphi(v,Q) &= -\frac{v}{k(Q)}G^{-1}(Q) - \frac{1}{k(Q)}\int_v^a G^{-1}(Q/t)\,\mathrm{d}t \\
&= \frac{v}{2Q}(2Q-1) + \frac{1}{2Q}\int_v^a \left(\frac{2Q}{t} - 1\right)\mathrm{d}t \\
&\propto v - \ln v.
\end{aligned}$$

The corresponding Fenchel conjugate is given by

$$\begin{aligned}
\varphi^+(v,Q) &= \frac{1}{2Q}\left[\frac{-2Qv + 2Q - 1}{1-v} + \int_{\frac{1}{1-v}}^a \left\{\frac{2Q}{t} - 1\right\}\mathrm{d}t\right] \\
&\propto \ln(1-v).
\end{aligned}$$

Calculating the modified dual criterion function is relatively straightforward. We have $\varphi_1^+(v, Q_0) = -(1-v)^{-1}$ and $\varphi_2^+(v, Q_0) = -(1-v)^{-2}$ and therefore $v^*_{N_1} = 1 - 1/N_1$. This gives

$$\begin{aligned}
c_{N_1} &= -N_1^2 \\
b_{N_1} &= -N_1 + N_1^2\left(1 - \frac{1}{N_1}\right) = N_1^2 - 2N_1 \\
a_{N_1} &= \ln(1/N_1) + N_1\left(1 - \frac{1}{N_1}\right) - \frac{N_1^2}{2}\left(1 - \frac{1}{N_1}\right)^2 = \ln(1/N_1) - \frac{N_1^2}{2} + 2N_1 - \frac{3}{2},
\end{aligned}$$

and hence, after some re-arranging,

$$\varphi^\circ(v, Q_0) = \begin{cases}
\ln(1-v) & v < 1 - 1/N_1 \\
\ln(1/N_1) - \frac{3}{2} + 2N_1(1-v) - \frac{N_1^2}{2}(1-v)^2 & v \geq 1 - 1/N_1
\end{cases}.$$

This is exactly the modification to the $\ln(1-v)$ function proposed by Owen (2001, p. 235).

A.2 Data combination problems

Computing the inverse probability tilt of the auxiliary sample in data combination problems follows the basis algorithm outlined above. We begin by computing

$$\widehat{Q} = \frac{1}{N}\sum_{i=1}^{N} D_i, \qquad \widehat{\zeta}^t = \frac{1}{N_t}\sum_{i=1}^{N_t} h(X_i).$$

We then solve the dual problem

$$\max_{\delta^a \in \widehat{\Delta}_{N_a}} \; l_{N_a}(\delta^a; \widehat{\rho}),$$

where $l_{N_a}(\delta^a;\rho) = t_0'\delta^a + \frac{1}{N_a}\sum_{i=N_t+1}^{N_a}\varphi^+(t(X_i,\zeta^t)'\delta^a, Q)$, $\widehat{\rho} = (\widehat{Q}, \widehat{\zeta}^{t\prime})'$, the set $\widehat{\Delta}_{N_a}$ is defined below, and the form of $\varphi^+(v,Q)$ is as given in Section 4 of the main text. The first and second derivatives of $\varphi^+(v,Q)$ are given by

$$\varphi_1^+(v,Q) = -\frac{1-Q}{Q}\frac{G(k(Q)v + G^{-1}(Q))}{1-G(k(Q)v+G^{-1}(Q))}, \quad \varphi_2^+(v,Q) = -\frac{1-Q}{Q}\frac{G_1(k(Q)v+G^{-1}(Q))}{[1-G(k(Q)v+G^{-1}(Q))]^2}k(Q).$$

The set $\widehat{\Delta}_{N_a}$ is given by $\left\{\delta^a : 0 \le -\varphi_1^+(t_i(\widehat{\zeta}^t)'\delta^a, \widehat{Q})/N_a < 1, \quad i = N_t+1,\ldots,N\right\}$ which ensures that the probabilities attached to each support point of the auxiliary sample are between zero and one. The lower-bound requirement will be satisfied automatically for any valid data combination contrast function. The upper bound is a substantive domain restriction. Let $v_i = t_i\left(\zeta^t\right)'\delta^0$; we require that for $i = N_t+1,\ldots,N$

$$v_i < \frac{1}{k(Q)}\left\{G^{-1}\left(\left[1+\frac{1}{N_a}\frac{1-Q}{Q}\right]^{-1}\right) - G^{-1}(Q)\right\}.$$

Let $v_{N_a}^* = \frac{1}{k(Q)}\left\{G^{-1}\left(\left[1+\frac{1}{N_a}\frac{1-Q}{Q}\right]^{-1}\right) - G^{-1}(Q)\right\}$ (note that $v_{N_a}^* \to \infty$ as $N_a \to \infty$). We define $\varphi^\circ(v,Q_0)$ as above with N_a and $v_{N_a}^*$ replacing N_1 and $v_{N_1}^*$ in the relevant expressions. The vector $\widehat{\delta}^a$ is then computed using a Newton-Raphson algorithm applied to the modified criterion function.

Computational details for inverse logistic tilting When the propensity score is logistic our algorithm is particularly straightforward to implement. We have $k(Q) = 1$, $v_{N_a}^* = \ln N_a$ and.

$$\varphi^+(v,Q_0) = \varphi_1^+(v,Q_0) = \varphi_2^+(v,Q_0) = -\exp[v].$$

Our modified criterion function is thus given by

$$\varphi^\circ(v,Q_0) = \begin{cases} -\exp[v] & v < \ln N_a \\ -N_a\left[1-\ln N_a + \frac{1}{2}(\ln N_a)^2 + (1-\ln N_a)v + \frac{1}{2}v^2\right] & v \ge \ln N_a \end{cases}.$$

We comment that this procedure is an apparently new algorithm for ET estimation.

B Proofs of theorems

B.1 Proof of Theorem 3.1

Since the IPT estimate of γ_0 can be represented as the solution to a three-step sequential GMM problem, Theorem 3.1 follows from a direct application of Theorem 6.1 of Newey and McFadden (1994, p. 2178). Here we demonstrate double robustness and derive expressions for the asymptotic variance of $\widehat{\gamma}$ under different assumptions.

The IPT procedure outlined in Section 3 consists of first estimating ρ_0 by the full sample analog estimates

$$\frac{1}{N}\sum_{i=1}^{N} m_1(Z_i,\widehat{\rho}) = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} D_i - \widehat{Q} \\ h(X_i) - \widehat{\zeta}\end{pmatrix} = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} D_i-\widehat{Q} & 0 \\ 0 & I_M\end{pmatrix} t(X_i,\widehat{\zeta}) = 0. \tag{48}$$

Second we solve (21) and (22) for $\widehat{\delta}^1$ and $\widehat{\delta}^0$. The corresponding first order conditions are

$$\frac{1}{N}\sum_{i=1}^{N} m_2(Z_i,\widehat{\rho},\widehat{\delta}) = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} t_0 + \frac{D_i}{\widehat{Q}}\varphi_1^+(t(X_i,\widehat{\zeta})'\widehat{\delta}^1, \widehat{Q})t(X_i,\widehat{\zeta}) \\ t_0 + \frac{1-D_i}{1-\widehat{Q}}\varphi_1^+(t(X_i,\widehat{\zeta})'\widehat{\delta}^0, 1-\widehat{Q})t(X_i,\widehat{\zeta})\end{pmatrix}, \tag{49}$$

where $\delta = \left(\delta^{1\prime}, \delta^{0\prime}\right)'$.

Consistency of the solution to (49) follows from standard M-estimator arguments. Since $\varphi^{+}(v, Q)$ is globally concave sufficient variation in $h(X)$ in each of the two subsamples will generally suffice for (48) to have a unique solution.

In the third step γ_0 is estimated by the solution to

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} m_3(Z_i, \widehat{\rho}, \widehat{\delta}, \widehat{\gamma}) &= \sum_{i=1}^{N_1} \widehat{\pi}_{1i}\psi_1\left(Y_{1i}, X_i, \widehat{\gamma}\right) - \sum_{i=N_1+1}^{N} \widehat{\pi}_{0i}\psi_0\left(Y_{0i}, X_i, \widehat{\gamma}\right) \\ &= -\frac{1}{N}\left\{\sum_{i=1}^{N} \frac{D_i\varphi_1^{+}(t(X_i,\widehat{\zeta})'\widehat{\delta}^1, \widehat{Q})}{\widehat{Q}}\psi_1\left(Y_{1i}, X_i, \widehat{\gamma}\right)\right. \\ &\quad \left. -\frac{(1-D_i)\,\varphi_1^{+}(t(X_i,\widehat{\zeta})'\widehat{\delta}^0, 1-\widehat{Q})}{1-\widehat{Q}}\psi_0\left(Y_{0i}, X_i, \widehat{\gamma}\right)\right\} = 0\end{aligned} \tag{50}$$

where $\widehat{\rho}$ is fixed at its first step, and $\widehat{\delta}$ at its second step, value.

Demonstration of 'double robustness' First consider the case where the propensity score is correctly modelled (i.e., Assumption 3.5 holds) but where $\mathbb{E}\left[\psi_1\left(Y_1, X, \gamma_0\right)|\, X\right]$ and $\mathbb{E}\left[\psi_0\left(Y_0, X, \gamma_0\right)|\, X\right]$ are not linear functions of $h(X)$ (i.e., Assumption 3.6 does not hold). By the one to one mapping between δ^1 and (α_0, β_0) and δ^0 and (α_0, β_0), and iterated expectations we then have

$$\begin{aligned}\mathbb{E}\left[m_3(Z, \rho_0, \delta_0, \gamma)\right] &= \mathbb{E}\left[\frac{D}{G\left(\alpha_0 + h(X)'\beta_0\right)}\psi_1\left(Y_1, X, \gamma\right) - \frac{1-D}{1-G\left(\alpha_0 + h(X)'\beta_0\right)}\psi_0\left(Y_0, X, \gamma\right)\right] \\ &= \mathbb{E}\left[\psi_1\left(Y_1, X, \gamma\right) - \psi_0\left(Y_0, X, \gamma\right)\right],\end{aligned}$$

which by Assumption 3.1 is uniquely zero at $\gamma = \gamma_0$.

Consistency of $\widehat{\gamma}$ for γ_0 then follows under additional (standard) regularity conditions (e.g., Newey and McFadden 1994, Section 2.5). These conditions, which include moment and continuity conditions on $\psi(Z, \gamma)$ and consistency of $\widehat{\delta}$ for δ_0, ensure that

$$\frac{1}{N}\sum_{i=1}^{N} m_3(Z_i, \widehat{\rho}, \widehat{\delta}, \gamma)$$

converges uniformly in $\gamma \in \mathcal{G} \subset \mathbb{R}^K$ to $\mathbb{E}\left[m_3(Z, \rho_0, \delta_0, \gamma)\right] = \mathbb{E}\left[\psi_1\left(Y_1, X, \gamma\right) - \psi_0\left(Y_0, X, \gamma\right)\right] = \mathbb{E}\left[\psi(Z, \gamma)\right]$.

Now consider the case where the propensity score is misspecified, but $\mathbb{E}\left[\psi_1\left(Y_1, X, \gamma_0\right)|\, X\right]$ and $\mathbb{E}\left[\psi_0\left(Y_0, X, \gamma_0\right)|\, X\right]$ are linear functions of $h(X)$. In this case denote the probability limits of $\widehat{\delta^1}$ and $\widehat{\delta^0}$ by δ_*^1 and δ_*^0. By Assumptions 3.1, 3.3 and 3.4 and iterated expectations we have

$$\begin{aligned}\mathbb{E}\left[m_3(Z, \rho_0, \delta_*, \gamma)\right] &= \mathbb{E}\left[\frac{D}{G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))}\psi_1\left(Y_1, X, \gamma\right)\right. \\ &\quad \left. -\frac{1-D}{G\left(k(1-Q_0)t(X,\zeta_0)'\delta_*^0 + G^{-1}(1-Q_0)\right)}\psi_0\left(Y_0, X, \gamma\right)\right] \\ &= \mathbb{E}\left[\frac{p_0(X)}{G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))}\mathbb{E}\left[\psi_1\left(Y_1, X, \gamma\right)|\, X\right]\right. \\ &\quad - \frac{1-p_0(X)}{G\left(k(1-Q_0)t(X,\zeta_0)'\delta_*^0 + G^{-1}(1-Q_0)\right)}\mathbb{E}\left[\psi_0\left(Y_0, X, \gamma\right)|\, X\right] \\ &\quad + \frac{D}{G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))}\Pi_1^* t(X,\zeta_0) \\ &\quad \left. -\frac{1-D}{G\left(k(1-Q_0)t(X,\zeta_0)'\delta_*^0 + G^{-1}(1-Q_0)\right)}\Pi_0^* t(X,\zeta_0)\right] \\ &= \mathbb{E}\left[\frac{p_0(X)}{G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))}\left\{\mathbb{E}\left[\psi_1\left(Y_1, X, \gamma\right)|\, X\right] - \mathbb{E}\left[\psi_1\left(Y_1, X, \gamma_0\right)|\, X\right]\right\}\right. \\ &\quad \left. -\frac{1-p_0(X)}{G\left(k(1-Q_0)t(X,\zeta_0)'\delta_*^0 + G^{-1}(1-Q_0)\right)}\left\{\mathbb{E}\left[\psi_0\left(Y_0, X, \gamma\right)|\, X\right] - \mathbb{E}\left[\psi_0\left(Y_0, X, \gamma_0\right)|\, X\right]\right\}\right]\end{aligned}$$

Clearly $\mathbb{E}\left[m_2(Z, \rho_0, \delta_*, \gamma)\right] = 0$ at $\gamma = \gamma_0$, however it need not be true that γ_0 is a unique solution. If $\psi_0\left(Y_0, X, \gamma\right)$ does not depend on γ then

$$\mathbb{E}\left[m_3(Z, \rho_0, \delta_*, \gamma)\right] = \mathbb{E}\left[\frac{p_0(X)}{G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))}\left\{\mathbb{E}\left[\psi_1\left(Y_1, X, \gamma\right)|\, X\right] - \mathbb{E}\left[\psi_1\left(Y_1, X, \gamma_0\right)|\, X\right]\right\}\right].$$

Since $p_0(x)/G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0)) > 0$ for all $x \in \mathcal{X}$ Assumption 3.1 implies that $\mathbb{E}[m_3(Z,\rho_0,\delta_*,\gamma)] = 0$ uniquely at $\gamma = \gamma_0$ for that case. If $\psi_1(Y_1,X,\gamma)$ does not depend on γ a parallel argument applies. One of these two conditions hold for many of our motivating examples. In other cases we must assume uniqueness of the solution to $\mathbb{E}[m_3(Z,\rho_0,\delta_*,\gamma_0)] = 0$ by hypothesis.

Asymptotic normality As long as at least one of Assumptions 3.5 and 3.6 it follows from standard GMM results that

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \to N(0, \Upsilon_0),$$

where Υ_0 is the lower $K \times K$ block of $M_0^{-1} V_0 M_0^{-1\prime}$ where M_0 and V_0 are given by

$$M_0 = \mathbb{E}\left[\frac{\partial m(Z,\rho_0,\delta_0,\gamma_0)}{\partial \theta'}\right] = \begin{pmatrix} M_{1\rho} & 0 & 0 \\ M_{2\rho} & M_{2\delta} & 0 \\ M_{3\rho} & M_{3\delta} & M_{3\gamma} \end{pmatrix}, \qquad V_0 = \mathbb{E}\left[m(Z,\rho_0,\delta_0,\gamma_0)m(Z,\rho_0,\delta_0,\gamma_0)'\right], \tag{51}$$

with $\theta = (\rho', \delta', \gamma')'$. Define

$$\begin{aligned} p_*^1(X) &= G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0)) \\ 1 - p_*^0(X) &= G(k(1-Q_0)t(X,\zeta_0)'\delta_*^0 + G^{-1}(1-Q_0)) \end{aligned}$$

and

$$\begin{aligned} \underset{1+M\times 1}{A_1} &= \begin{pmatrix} \frac{\partial}{\partial Q}\left\{\frac{1}{G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))}\right\} \\ \frac{\partial}{\partial \zeta}\left\{\frac{1}{G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))}\right\} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{G_1(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))}{G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))^2}\left[k_1(Q_0)t(X,\zeta_0)'\delta_*^1 + G_1\left(G^{-1}(Q_0)\right)^{-1}\right] \\ -\frac{G_1(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))}{G(k(Q_0)t(X,\zeta_0)'\delta_*^1 + G^{-1}(Q_0))^2}\left[k(Q_0)\begin{pmatrix} 0 & -I_M \end{pmatrix}'\delta_*^1\right] \end{pmatrix}, \end{aligned}$$

where the $*$ subscripts emphasize the possibility that the (implicitly defined) propensity score is incorrect.

The vector A_0 is defined analogously to that of A_1 with $1 - Q_0$ and δ_*^0 replacing Q_0 and δ_*^1 throughout. Using these definition we can write non-zero elements of M_0 as

$$\begin{aligned} \underset{1+M\times 1+M}{M_{1\rho}} &= -I_{1+M} \\ \underset{2(1+M)\times 1+M}{M_{2\rho}} &= \mathbb{E}\begin{bmatrix} Dt(X,\zeta_0)A_1' - \frac{D}{p_*^1(X)}\begin{pmatrix} 0 & 0 \\ 0 & -I_M \end{pmatrix} \\ (1-D)\,t(X,\zeta_0)A_0' - \frac{1-D}{1-p_*^0(X)}\begin{pmatrix} 0 & 0 \\ 0 & -I_M \end{pmatrix} \end{bmatrix} \\ \underset{K\times 1+M}{M_{3\rho}} &= \mathbb{E}\left[D\psi_1(Y_1,X,\gamma_0)A_1' - (1-D)\,\psi_0(Y_0,X,\gamma_0)A_0'\right] \\ \underset{2(1+M)\times 2(1+M)}{M_{2\delta}} &= \mathbb{E}\left[\begin{pmatrix} \frac{Dk(Q)G_1\left(G^{-1}\left(p_*^1(X)\right)\right)}{p_*^1(X)^2} & 0 \\ 0 & \frac{(1-D)k(1-Q)G_1\left(G^{-1}\left(1-p_*^0(X)\right)\right)}{\left(1-p_*^0(X)\right)^2} \end{pmatrix} \otimes t(X,\zeta_0)t(X,\zeta_0)'\right] \\ \underset{K\times 2(1+M)}{M_{3\delta}} &= -\mathbb{E}\left[\frac{Dk(Q)G_1\left(G^{-1}\left(p_*^1(X)\right)\right)}{p_*^1(X)^2}\psi_1(Y_1,X,\gamma_0)\,t(X,\zeta_0)',\right. \\ &\quad \left. -\frac{(1-D)\,k(1-Q)G_1\left(G^{-1}\left(1-p_*^0(X)\right)\right)}{(1-p_*^0(X))^2}\psi_0(Y_0,X,\gamma_0)\,t(X,\zeta_0)'\right] \\ \underset{K\times K}{M_{3\gamma}} &= \mathbb{E}\left[\frac{D}{p_*^1(X)}\frac{\partial\psi_1(Y_{1i},X_i,\gamma_0)}{\partial\gamma'} - \frac{1-D}{1-p_*^0(X)}\frac{\partial\psi_0(Y_{0i},X_i,\gamma_0)}{\partial\gamma'}\right]. \end{aligned}$$

These expressions can be used to construct an analog estimate of Υ_0 in the usual way (e.g., Newey and McFadden 1994, Section 4). This estimate will be consistent for the asymptotic sampling variance of $\widehat{\gamma}$ as long as one of Assumptions 3.5 or 3.6 holds; we therefore recommend this estimator in practice.

B.2 Proof of Theorem 3.2

We provide interpretable forms for the large sample variance of $\widehat{\gamma}$ under two cases: ρ_0 is (i) known and (ii) unknown. A third case, where some elements of ρ_0 are known and others are not, follows directly. In both cases we maintain Assumptions 3.5 and 3.6.

Form of Υ_0 when ρ_0 is known This case corresponds to the result given in Corollary 3.2. The estimator is equivalent to GMM using the moment function

$$m\left(Z,\delta,\gamma\right)=\left(\begin{array}{c} m_1(Z) \\ m_2(Z,\delta) \\ m_3(Z,\delta,\gamma) \end{array}\right),$$

with $m_1(Z)=m_1(Z,\rho_0)$, $m_2(Z,\delta)=m_2(Z,\rho_0,\delta)$ and $m_3(Z,\delta,\gamma)=m_3(Z,\rho_0,\delta,\gamma)$. In this case, since ρ_0 is known, $m_1(Z)$ plays the role of an auxiliary moment (cf., Hellerstein and Imbens 1999, Qian and Schmidt 1999).

The relevant Jacobian is now

$$M_0=\mathbb{E}\left[\begin{array}{cc} \frac{\partial m(Z,\delta,\gamma)}{\partial\delta'} & \frac{\partial m(Z,\delta,\gamma)}{\partial\gamma'} \end{array}\right]=\left(\begin{array}{cc} 0 & 0 \\ M_{2\delta} & 0 \\ M_{3\delta} & M_{3\gamma} \end{array}\right).$$

where the three non-zero elements of M_0 are given by

$$\begin{aligned}
\underset{2(1+M)\times 2(1+M)}{M_{2\delta}} &= \mathbb{E}\left[\left(\begin{array}{cc} \frac{k(Q)G_1\left(G^{-1}(p_0(X))\right)}{p_0(X)} & 0 \\ 0 & \frac{k(1-Q)G_1\left(G^{-1}(1-p_0(X))\right)}{1-p_0(X)} \end{array}\right)\otimes t(X,\zeta_0)t(X,\zeta_0)'\right] \\
\underset{K\times 2(1+M)}{M_{3\delta}} &= -\mathbb{E}\left[\frac{k(Q)G_1\left(G^{-1}\left(p_0\left(X\right)\right)\right)}{p_0\left(X\right)}\psi_1\left(Y_1,X,\gamma_0\right)t(X,\zeta_0)',\right. \\
&\quad \left.-\frac{k(1-Q)G_1\left(G^{-1}\left(1-p_0\left(X\right)\right)\right)}{1-p_0\left(X\right)}\psi_0\left(Y_0,X,\gamma_0\right)t(X,\zeta_0)'\right] \\
\underset{K\times K}{M_{3\gamma}} &= \mathbb{E}\left[\frac{\partial\psi\left(Z,\gamma_0\right)}{\partial\gamma'}\right]=\Gamma_0,
\end{aligned}$$

while those of V_0 are given by

$$\begin{aligned}
\underset{1+M\times 1+M}{V_{11}} &= \mathbb{E}\left[\begin{array}{cc} Q_0\left(1-Q_0\right) & \left(D-Q\right)h(X,\zeta_0)' \\ h(X,\zeta_0)\left(D-Q\right) & h(X,\zeta_0)h(X,\zeta_0)' \end{array}\right] \\
\underset{2(1+M)\times 1+M}{V_{21}} &= -\mathbb{E}\left[\begin{array}{cc} 1-Q_0 & 0 \\ 0 & h(X,\zeta_0)h(X,\zeta_0)' \\ -Q_0 & 0 \\ 0 & h(X,\zeta_0)h(X,\zeta_0)' \end{array}\right] \\
\underset{2(1+M)\times 2(1+M)}{V_{22}} &= \mathbb{E}\left[\begin{array}{cc} \frac{t(X,\zeta_0)t(X,\zeta_0)'}{p_0(X)}-t_0t_0' & -t_0t_0' \\ -t_0t_0' & \frac{t(X,\zeta_0)t(X,\zeta_0)'}{1-p_0(X)}-t_0t_0' \end{array}\right] \\
\underset{K\times 1+M}{V_{31}} &= \mathbb{E}\left[\left(\begin{array}{cc} \psi_1\left(Y_1,X,\gamma_0\right) & -\psi_0\left(Y_0,X,\gamma_0\right) \end{array}\right)\left(\begin{array}{cc} 1-Q_0 & h(X,\zeta_0)' \\ -Q_0 & h(X,\zeta_0)' \end{array}\right)\right] \\
\underset{K\times 2(1+M)}{V_{32}} &= \mathbb{E}\left[\left(\begin{array}{cc} -\frac{q_1(X;\gamma_0)}{p_0(X)} & \frac{q_0(X;\gamma_0)}{1-p_0(X)} \end{array}\right)t(X,\zeta_0)'\right] \\
\underset{K\times K}{V_{33}} &= \mathbb{E}\left[\frac{\Sigma_1\left(X;\gamma_0\right)}{p_0\left(X\right)}+\frac{1-p_0\left(X\right)}{p_0\left(X\right)}q_1\left(X;\gamma_0\right)q_1\left(X;\gamma_0\right)'+q_1\left(X;\gamma_0\right)q_1\left(X;\gamma_0\right)'\right. \\
&\quad \left.+\frac{\Sigma_0\left(X;\gamma_0\right)}{1-p_0\left(X\right)}+\frac{p_0\left(X\right)}{1-p_0\left(X\right)}q_0\left(X;\gamma_0\right)q_0\left(X;\gamma_0\right)'+q_0\left(X;\gamma_0\right)q_0\left(X;\gamma_0\right)'\right].
\end{aligned}$$

The forms of both M_0 and V_0 have been simplified by using Assumptions 3.5 or 3.6. The asymptotic variance-covariance matrix is given by $\left(M_0'V_0^{-1}M_0\right)^{-1}$. An analog estimate of this can be constructed in the usual way. In practice M_0 and V_0 should be estimated without exploiting the simplifications due to Assumptions 3.5 or 3.6.

Under Assumption 3.6 we have

$$\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,X\right]=\Pi_1^*t(X,\zeta_0),\qquad \mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)|\,X\right]=\Pi_0^*t(X,\zeta_0)$$

for $\Pi_0^*=\left(\varsigma_0+\Pi_0\zeta_0,\Pi_0\right)$ and $\Pi_1^*=\left(\varsigma_1+\Pi_1\zeta_0,\Pi_1\right)$.

Applying this result and iterated expectations to $M_{3\delta}$ yields

$$\begin{aligned}
M_{3\delta} &= -\mathbb{E}\left[\frac{k(Q)G_1\left(G^{-1}\left(p_0\left(X\right)\right)\right)}{p_0\left(X\right)}\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,X\right]t(X,\zeta_0)',\right.\\
&\left.-\frac{k(1-Q)G_1\left(G^{-1}\left(1-p_0\left(X\right)\right)\right)}{1-p_0\left(X\right)}\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)|\,X\right]t(X,\zeta_0)'\right]\\
&= -\mathbb{E}\left[\frac{k(Q)G_1\left(G^{-1}\left(p_0\left(X\right)\right)\right)}{p_0\left(X\right)}\Pi_1^* t(X,\zeta_0)t(X,\zeta_0)',\right.\\
&\left.-\frac{k(1-Q)G_1\left(G^{-1}\left(1-p_0\left(X\right)\right)\right)}{1-p_0\left(X\right)}\Pi_0^* t(X,\zeta_0)t(X,\zeta_0)'\right]\\
&= -\begin{pmatrix}\Pi_1^* & -\Pi_0^*\end{pmatrix}\mathbb{E}\left[\begin{pmatrix}\frac{k(Q)G_1\left(G^{-1}\left(p_0(X)\right)\right)}{p_0(X)} & 0\\ 0 & \frac{k(1-Q)G_1\left(G^{-1}\left(1-p_0(X)\right)\right)}{1-p_0(X)}\end{pmatrix}\otimes t(X,\zeta_0)t(X,\zeta_0)'\right]\\
&= -\begin{pmatrix}\Pi_1^* & -\Pi_0^*\end{pmatrix}M_{2\delta},
\end{aligned}$$

which then gives

$$\begin{aligned}
M_{3\delta}M_{2\delta}^{-1}V_{22} &= -\begin{pmatrix}\Pi_1^* & -\Pi_0^*\end{pmatrix}\mathbb{E}\begin{bmatrix}\frac{t(X,\zeta_0)t(X,\zeta_0)'}{p_0(X)}-t_0t_0' & -t_0t_0'\\ -t_0t_0' & \frac{t(X,\zeta_0)t(X,\zeta_0)'}{1-p_0(X)}-t_0t_0'\end{bmatrix}\\
&= -\mathbb{E}\begin{bmatrix}\frac{\Pi_1^*t(X,\zeta_0)t(X,\zeta_0)'}{p_0(X)}-\Pi_1^*t_0t_0'+\Pi_0^*t_0t_0' & -\Pi_1^*t_0t_0'-\frac{\Pi_0^*t(X,\zeta_0)t(X,\zeta_0)'}{1-p_0(X)}+\Pi_0^*t_0t_0'\end{bmatrix}\\
&= \mathbb{E}\left[\begin{pmatrix}-\frac{q_1(X;\gamma_0)}{p_0(X)} & \frac{q_0(X;\gamma_0)}{1-p_0(X)}\end{pmatrix}t(X,\zeta_0)'\right]\\
&= V_{32}
\end{aligned}$$

where the second to last line uses the fact that $\Pi_1^*t_0t_0'-\Pi_0^*t_0t_0'=\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)-\psi_0\left(Y_0,X,\gamma_0\right)\right]t_0'=0$.

This result and some straightforward algebra then give

$$\left(M_0'V_0^{-1}M_0\right)^{-1}=\begin{pmatrix}M_{2\delta}^{-1}V_{22}M_{2\delta}^{-1\prime} & 0\\ 0 & \Gamma_0^{-1}\left(V_{33}-M_{3\delta}M_{2\delta}^{-1}V_{22}M_{2\delta}^{-1\prime}M_{3\delta}'\right)\Gamma_0^{-1\prime}\end{pmatrix}.$$

Now observe that

$$\begin{aligned}
M_{3\delta}M_{2\delta}^{-1}V_{22}M_{2\delta}^{-1\prime}M_{3\delta}' &= \begin{pmatrix}\Pi_1^* & -\Pi_0^*\end{pmatrix}\mathbb{E}\begin{bmatrix}\frac{t(X,\zeta_0)t(X,\zeta_0)'}{p_0(X)}-t_0t_0' & -t_0t_0'\\ -t_0t_0' & \frac{t(X,\zeta_0)t(X,\zeta_0)'}{1-p_0(X)}-t_0t_0'\end{bmatrix}\begin{pmatrix}\Pi_1^*\\ -\Pi_0^*\end{pmatrix}\\
&= \begin{pmatrix}\frac{\Pi_1^*t(X,\zeta_0)t(X,\zeta_0)'}{p_0(X)}-\Pi_1^*t_0t_0'+\Pi_0^*t_0t_0' & -\Pi_1^*t_0t_0'-\frac{\Pi_0^*t(X,\zeta_0)t(X,\zeta_0)'}{1-p_0(X)}+\Pi_0^*t_0t_0'\end{pmatrix}\\
&\times\begin{pmatrix}\Pi_1^*\\ -\Pi_0^*\end{pmatrix}\\
&= \begin{pmatrix}\frac{\Pi_1^*t(X,\zeta_0)t(X,\zeta_0)'}{p_0(X)} & -\frac{\Pi_0^*t(X,\zeta_0)t(X,\zeta_0)'}{1-p_0(X)}\end{pmatrix}\begin{pmatrix}\Pi_1^*\\ -\Pi_0^*\end{pmatrix}\\
&= \frac{q_1\left(X;\gamma_0\right)q_1\left(X;\gamma_0\right)'}{p_0\left(X\right)}+\frac{q_0\left(X;\gamma_0\right)q_0\left(X;\gamma_0\right)'}{1-p_0\left(X\right)}.
\end{aligned}$$

Therefore the large sample variance of $\widehat{\gamma}$ when ρ_0 is known is given by

$$\Gamma_0^{-1}\mathbb{E}\left[\frac{\Sigma_0\left(X;\gamma_0\right)}{1-p_0\left(X\right)}+\frac{\Sigma_1\left(X;\gamma_0\right)}{p_0\left(X\right)}\right]\Gamma_0^{-1\prime},$$

as claimed.

Form of Υ_0 when ρ_0 is unknown This case corresponds to the result given in Theorem 3.2. IPT is equivalent to GMM using the moment function

$$m\left(Z,\rho,\delta,\gamma\right)=\begin{pmatrix}m_1(Z,\rho)\\ m_2(Z,\rho,\delta)\\ m_3(Z,\rho,\delta,\gamma)\end{pmatrix}.$$

The variance of $m(Z, \rho_0, \delta_0, \gamma_0)$ is the same as in the known ρ_0 case, the Jacobian matrix partitioned as in (51) with

$$\begin{aligned}
\underset{1+M\times 1+M}{M_{1\rho}} &= -I_{1+M} \\
\underset{2(1+M)\times 1+M}{M_{2\rho}} &= \mathbb{E}\left[\begin{array}{c} Dt(X,\zeta_0)A_1' - \frac{D}{p_0(X)}\begin{pmatrix} 0 & 0 \\ 0 & -I_M \end{pmatrix} \\ (1-D)\,t(X,\zeta_0)A_0' - \frac{1-D}{1-p_0(X)}\begin{pmatrix} 0 & 0 \\ 0 & -I_M \end{pmatrix} \end{array}\right] \\
\underset{K\times 1+M}{M_{3\rho}} &= \mathbb{E}\left[D\psi_1(Y_1, X, \gamma_0)A_1' - (1-D)\,\psi_0(Y_0, X, \gamma_0)A_0'\right]
\end{aligned}$$

and all other elements are as defined in the known ρ_0 case.

For this case note that

$$\begin{aligned}
M_{3\delta}M_{2\delta}^{-1}V_{21} &= \begin{pmatrix} \Pi_1^* & -\Pi_0^* \end{pmatrix}\mathbb{E}\left[\begin{array}{cc} 1-Q_0 & 0 \\ 0 & h(X,\zeta_0)h(X,\zeta_0)' \\ -Q_0 & 0 \\ 0 & h(X,\zeta_0)h(X,\zeta_0)' \end{array}\right] \\
&= \begin{pmatrix} \varsigma_1 + \Pi_1\zeta_0 & \Pi_1 & -\varsigma_0 - \Pi_0\zeta_0 & -\Pi_0 \end{pmatrix}\mathbb{E}\left[\begin{array}{cc} 1-Q_0 & 0 \\ 0 & h(X,\zeta_0)h(X,\zeta_0)' \\ -Q_0 & 0 \\ 0 & h(X,\zeta_0)h(X,\zeta_0)' \end{array}\right] \\
&= \mathbb{E}\left[\begin{pmatrix} \psi_1(Y_1, X, \gamma_0) & -\psi_0(Y_0, X, \gamma_0) \end{pmatrix}\begin{pmatrix} 1-Q_0 & h(X,\zeta_0)' \\ -Q_0 & h(X,\zeta_0)' \end{pmatrix}\right] \\
&= V_{31},
\end{aligned}$$

which, along with the equality $M_{3\delta}M_{2\delta}^{-1}V_{22} = V_{32}$, and some tedious but straightforward algebra implies that

$$\begin{aligned}
\left\{M_0^{-1}V_0^{-1}M_0^{-1\prime}\right\}_{33} = \Gamma_0^{-1}V_{33}\Gamma_0^{-1\prime} &- \Gamma_0^{-1}M_{3\delta}M_{2\delta}^{-1}V_{22}M_{2\delta}^{-1\prime}M_{3\delta}'\Gamma_0^{-1\prime} \\
&+ \Gamma_0^{-1}\left(M_{3\rho} - M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)M_{1\rho}^{-1}V_{11}M_{1\rho}^{-1\prime}\left(M_{3\rho} - M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)'\Gamma_0^{-1\prime},
\end{aligned}$$

where we can also show that

$$M_{3\rho} - M_{3\delta}M_{2\delta}^{-1}M_{2\rho} = \begin{pmatrix} 0 & \Pi_1 - \Pi_0 \end{pmatrix}.$$

This gives a large sample variance of $\widehat{\gamma}$ when ρ_0 is unknown of

$$\Gamma_0^{-1}\mathbb{E}\left[\frac{\Sigma_0(X;\gamma_0)}{1-p_0(X)} + \frac{\Sigma_1(X;\gamma_0)}{p_0(X)} + \left[q_1(X;\gamma_0) - q_0(X;\gamma_0)\right]\left[q_1(X;\gamma_0) - q_0(X;\gamma_0)\right]'\right]\Gamma_0^{-1\prime}$$

as claimed.

B.3 Proof of Corollary 3.1

When the data are MCAR we have $\widehat{\delta}^1 \xrightarrow{p} 0$ and $\widehat{\delta}^0 \xrightarrow{p} 0$. Recalling that $k(Q) = -Q/G_1\left(G^{-1}(Q)\right)$ then gives

$$M_{2\delta} = -I_2 \otimes \begin{pmatrix} 1 & 0 \\ 0 & \Omega_{hh} \end{pmatrix}, \quad M_{3\delta} = \begin{pmatrix} \mathbb{E}\left[\psi_1(Y_1, X, \gamma_0)\right] & \Omega_{\psi_1 h} & -\mathbb{E}\left[\psi_0(Y_0, X, \gamma_0)\right] & -\Omega_{\psi_0 h} \end{pmatrix}, \quad M_{3\gamma} = \Gamma_0$$

as well as $M_{3\rho} - M_{3\delta}M_{2\delta}^{-1}M_{2\rho} = \begin{pmatrix} 0 & \left(\Omega_{\psi_1 h} - \Omega_{\psi_0 h}\right)\Omega_{hh}^{-1} \end{pmatrix}$ and

$$\begin{aligned}
V_{22} &= \begin{pmatrix} \frac{1}{Q_0} - 1 & 0 & -1 & 0 \\ 0 & \frac{\Omega_{hh}}{Q_0} & 0 & 0 \\ -1 & 0 & \frac{1}{1-Q_0} - 1 & 0 \\ 0 & 0 & 0 & \frac{\Omega_{hh}}{1-Q_0} \end{pmatrix} \\
V_{33} &= \frac{\Omega_{\psi_1\psi_1} + \mathbb{E}\left[\psi_1(Y_1, X, \gamma_0)\right]\mathbb{E}\left[\psi_1(Y_1, X, \gamma_0)\right]'}{Q_0} + \frac{\Omega_{\psi_0\psi_0} + \mathbb{E}\left[\psi_0(Y_0, X, \gamma_0)\right]\mathbb{E}\left[\psi_0(Y_0, X, \gamma_0)\right]'}{1-Q_0}.
\end{aligned}$$

This gives $M_{3\delta}M_{2\delta}^{-1} = -\left(\begin{array}{cccc} \mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)\right] & \Omega_{\psi_1 h}\Omega_{hh}^{-1} & -\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)\right] & -\Omega_{\psi_0 h}\Omega_{hh}^{-1}\end{array}\right)$ and hence

$$\begin{aligned}
M_{3\delta}M_{2\delta}^{-1}V_{22}M_{2\delta}^{-1\prime}M_{3\delta}' &= \left(\begin{array}{cccc} \mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)\right] & \Omega_{\psi_1 h}\Omega_{hh}^{-1} & -\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)\right] & -\Omega_{\psi_0 h}\Omega_{hh}^{-1}\end{array}\right) \\
&\times \left(\begin{array}{cccc} \frac{1}{Q_0}-1 & 0 & -1 & 0 \\ 0 & \frac{\Omega_{hh}}{Q_0} & 0 & 0 \\ -1 & 0 & \frac{1}{1-Q_0}-1 & 0 \\ 0 & 0 & 0 & \frac{\Omega_{hh}}{1-Q_0}\end{array}\right)\left(\begin{array}{c} \mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)\right]' \\ \Omega_{hh}^{-1}\Omega_{\psi_1 h}' \\ -\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)\right]' \\ -\Omega_{hh}^{-1}\Omega_{\psi_0 h}'\end{array}\right) \\
&= \left(\begin{array}{cccc} \frac{\mathbb{E}[\psi_1(Y_1,X,\gamma_0)]}{Q_0} & \frac{\Omega_{\psi_1 h}}{Q_0} & -\frac{\mathbb{E}[\psi_0(Y_0,X,\gamma_0)]}{1-Q_0} & -\frac{\Omega_{\psi_0 h}}{1-Q_0}\end{array}\right)\left(\begin{array}{c} \mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)\right]' \\ \Omega_{hh}^{-1}\Omega_{\psi_1 h}' \\ -\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)\right]' \\ -\Omega_{hh}^{-1}\Omega_{\psi_0 h}'\end{array}\right) \\
&= \frac{\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)\right]\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)\right]'}{Q_0} + \frac{\Omega_{\psi_1 h}\Omega_{hh}^{-1}\Omega_{\psi_1 h}'}{Q_0} \\
&+ \frac{\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)\right]\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)\right]'}{1-Q_0} + \frac{\Omega_{\psi_0 h}\Omega_{hh}^{-1}\Omega_{\psi_0 h}'}{1-Q_0}.
\end{aligned}$$

Similarly

$$\begin{aligned}
\left(M_{3\rho}-M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)M_{1\rho}^{-1}V_{11}M_{1\rho}^{-1\prime}\left(M_{3\rho}-M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)' &= \left(\Omega_{\psi_1 h}-\Omega_{\psi_0 h}\right)\Omega_{hh}^{-1}\left(\Omega_{\psi_1 h}-\Omega_{\psi_0 h}\right)' \\
&= \Omega_{\psi h}\Omega_{hh}^{-1}\Omega_{\psi_0 h}.
\end{aligned}$$

The asymptotic variance of $\widehat{\gamma}$ is thus given by

$$\Gamma_0^{-1}\left(\frac{\Omega_{\psi_1\psi_1}-\Omega_{\psi_1 h}\Omega_{hh}^{-1}\Omega_{\psi_1 h}'}{Q_0}+\frac{\Omega_{\psi_0\psi_0}-\Omega_{\psi_0 h}\Omega_{hh}^{-1}\Omega_{\psi_0 h}'}{1-Q_0}+\Omega_{\psi h}\Omega_{hh}^{-1}\Omega_{\psi_0 h}\right)\Gamma_0^{-1\prime},$$

which, when Assumption 3.6 holds, can be shown to be identical to the variance bound.

B.4 Proof of Theorem 4.1

As in the missing data case, the IPT estimate of γ_0 in the data combination case can be represented as a solution to a three-step sequential GMM problem. First, we estimate $\rho_0 = \left(Q_0, \zeta_0^{t\prime}\right)'$ using the target sample analog estimates

$$\frac{1}{N}\sum_{i=1}^{N} m_1(Z_i,\widehat{\rho}) = \frac{1}{N}\sum_{i=1}^{N}\left(\begin{array}{c} \frac{D_i}{\widehat{Q}}-1 \\ \frac{D_i}{\widehat{Q}}\left(h\left(X_i\right)-\widehat{\zeta}^t\right)\end{array}\right) = 0.$$

Note that $\zeta_0^t = \mathbb{E}\left[h\left(X\right)|\, D=1\right] = \mathbb{E}_t\left[h\left(X\right)\right]$.

Second, we find $\widehat{\delta}^a$ by solving

$$\frac{1}{N}\sum_{i=1}^{N} m_2(Z_i,\widehat{\rho},\widehat{\delta}^a) = \frac{1}{N}\sum_{i=1}^{N}\left[t_0+\frac{1-D_i}{1-\widehat{Q}}\varphi_1^+(t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a,\widehat{Q})t(X_i,\widehat{\zeta}^t)\right] = 0,$$

where

$$\varphi_1^+(t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a,\widehat{Q}) = -\frac{1-\widehat{Q}}{\widehat{Q}}\frac{G(k(\widehat{Q})t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a+G^{-1}(\widehat{Q}))}{1-G(k(\widehat{Q})t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a+G^{-1}(\widehat{Q}))}.$$

In the third, and final, step, we estimate γ_0 by solving

$$0 = \frac{1}{N}\sum_{i=1}^{N} m_3(Z_i,\widehat{\rho},\widehat{\delta}^a,\widehat{\gamma}) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i}{\widehat{Q}}\psi_1\left(Y_{1i},X_i,\widehat{\gamma}\right)+\frac{1-D_i}{1-\widehat{Q}}\varphi_1^+(t(X_i,\widehat{\zeta}^t)'\widehat{\delta}^a,\widehat{Q})\psi_0\left(Y_{0i},X_i,\widehat{\gamma}\right)\right\}.$$

Demonstration of 'double robustness' We first consider the case where the propensity score is correctly specified, but the conditional expectation functions, $\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\, X\right]$ and $\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)|\, X\right]$, are not linear in $h\left(X\right)$. We have, by iterated expectations,

$$\begin{aligned}
\mathbb{E}\left[m_3\left(Z_i,\rho_0,\delta_0^a,\gamma\right)\right] &= \mathbb{E}\left[\psi_1\left(Y_1,X,\gamma\right)|\, D=1\right] - \mathbb{E}\left[\frac{p_0\left(X\right)}{Q_0}\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma\right)|\, X,D=1\right]\right] \\
&= \mathbb{E}\left[\psi_1\left(Y_1,X,\gamma\right)|\, D=1\right] - \mathbb{E}\left[\psi_0\left(Y_1,X,\gamma\right)|\, D=1\right] \\
&= \mathbb{E}\left[\psi\left(Z,\gamma\right)|\, D=1\right]
\end{aligned}$$

which is uniquely zero at $\gamma = \gamma_0$. Consistency of $\widehat{\gamma}$ for γ_0 then follows under regularity conditions.

Now consider the case where the propensity score is misspecified, but Assumption 3.6 holds. Denote the probability limit of $\widehat{\delta}^a$ by δ_*^a and define $p_*^a(X) = G(k(Q_0)\, t\left(X_i, \zeta_0^t\right)' \delta_*^a + G^{-1}(Q_0))$ and $\Pi_0^* = \left(\varsigma_0 + \Pi_0 \zeta_0^t, \Pi_0\right)$. Observe that the second step moment restriction implies the population equality

$$0 = \mathbb{E}\left[t_0 - \frac{1-D}{1-Q_0}\frac{1-Q_0}{Q_0}\frac{p_*^a(X)}{1-p_*^a(X)} t(X, \zeta_0^t)\right].$$

Multiplying through by Π_0^* and rearranging gives, using Assumption 3.6,

$$\mathbb{E}\left[\psi_0(Y_0, X, \gamma)|\, D=1\right] = \mathbb{E}\left[\frac{1-p_0(X)}{Q_0}\frac{p_*^a(X)}{1-p_*^a(X)}\mathbb{E}\left[\psi_0(Y_0, X, \gamma)|\, X\right]\right]$$

since $\Pi_0^* t_0 = \varsigma_0 + \Pi_0 \zeta_0^t = \mathbb{E}\left[\psi_0(Y_0, X, \gamma)|\, D=1\right]$. This equality is a consequence of linearity of $\mathbb{E}\left[\psi_0(Y_0, X, \gamma)|\, X\right]$ in $h(X)$ and the imposition of the moment balancing constraints.

Using this result, Assumptions 4.1 to 4.4, and iterated expectations we have

$$\begin{aligned}
\mathbb{E}\left[m_3(Z, \rho_0, \delta_*^a, \gamma)\right] &= \mathbb{E}\left[\psi_1(Y_1, X, \gamma)|\, D=1\right] - \mathbb{E}\left[\frac{1-p_0(X)}{Q_0}\frac{p_*^a(X)}{1-p_*^a(X)}\mathbb{E}\left[\psi_0(Y_0, X, \gamma)|\, X\right]\right] \\
&= \mathbb{E}\left[\psi_1(Y_1, X, \gamma)|\, D=1\right] - \mathbb{E}\left[\psi_0(Y_0, X, \gamma)|\, D=1\right] \\
&= \mathbb{E}\left[\psi(Z, \gamma)|\, D=1\right]
\end{aligned}$$

which is uniquely zero at $\gamma = \gamma_0$ by Assumption 4.1.

Asymptotic normality If Assumptions 4.1 to 4.4 hold as well as either Assumption 3.5 or 3.6, we can use the standard GMM results to show that

$$\sqrt{N}(\widehat{\gamma} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, \Upsilon_0),$$

where Υ_0 is the lower $K \times K$ block of $M_0^{-1} V_0 M_0^{-1\prime}$ with M_0^{-1} and V_0 given by

$$M_0^{-1} = \begin{pmatrix} M_{1\rho}^{-1} & 0 & 0 \\ -M_{2\delta}^{-1} M_{2\rho} M_{1\rho}^{-1} & M_{2\delta}^{-1} & 0 \\ M_{3\gamma}^{-1}\left(M_{3\rho} - M_{3\delta} M_{2\delta}^{-1} M_{2\rho}\right) M_{1\rho}^{-1} & M_{3\gamma}^{-1} M_{3\delta} M_{2\delta}^{-1} & M_{3\gamma}^{-1} \end{pmatrix}, \quad V_0 = \mathbb{E}\left[m(Z, \rho_0, \delta_*^a, \gamma_0)\, m(Z, \rho_0, \delta_*^a, \gamma_0)'\right].$$

After defining

$$\begin{aligned}
\underset{(M+1\times 1)}{A_a} &= \begin{pmatrix} \frac{\partial}{\partial Q}\left\{\frac{G(k(Q_0)t(X,\zeta_0^t)'\delta_*^a + G^{-1}(Q_0))}{1-G(k(Q_0)t(X,\zeta_0^t)'\delta_*^a + G^{-1}(Q_0))}\right\} \\ \frac{\partial}{\partial \zeta_0}\left\{\frac{G(k(Q_0)t(X,\zeta_0^t)'\delta_*^a + G^{-1}(Q_0))}{1-G(k(Q_0)t(X,\zeta_0^t)'\delta_*^a + G^{-1}(Q_0))}\right\} \end{pmatrix} \\
&= \frac{G_1\left(k(Q_0)\, t(X,\zeta_0^t)'\delta_*^a + G^{-1}(Q_0)\right)}{1-G(k(Q_0)t(X,\zeta_0^t)'\delta_*^a + G^{-1}(Q_0))} \begin{pmatrix} k_1(Q_0)\, t(X,\zeta_0^t)'\delta_*^a + G_1\left(G^{-1}(Q_0)\right)^{-1} \\ k(Q_0)\begin{pmatrix} 0 & -I_M \end{pmatrix}' \delta_*^a \end{pmatrix},
\end{aligned}$$

we can write the non-zero elements of the Jacobian matrix as

$$\begin{aligned}
\underset{1+M\times 1+M}{M_{1\rho}} &= -\begin{pmatrix} 1/Q_0 & 0 \\ 0 & I_M \end{pmatrix} \\
\underset{M+1 x M+1}{M_{2\rho}} &= \mathbb{E}\left[\frac{1-D}{Q_0^2}\frac{p_*^a(X)}{1-p_*^a(X)} t\left(X, \zeta_0^t\right) t_0' - \frac{1-D}{Q_0}\frac{1}{1-p_*^a(X)} t\left(X, \zeta_0^t\right) A_a' - \frac{1-D}{Q_0}\frac{p_*^a(X)}{1-p_*^a(X)}\begin{pmatrix} 0 & 0 \\ 0 & -I_M \end{pmatrix}\right] \\
\underset{K\times 1+M}{M_{3\rho}} &= \mathbb{E}\left[-\frac{D}{Q_0^2}\psi_1(Y_1, X, \gamma_0)\, t_0' + \psi_0(Y_0, X, \gamma_0)\left\{\frac{1-D}{Q_0^2}\frac{p_*^a(X)}{1-p_*^a(X)} t_0' - \frac{1-D}{Q_0}\frac{1}{1-p_*^a(X)} A_a'\right\}\right] \\
\underset{1+M\times 1+M}{M_{2\delta}} &= -\mathbb{E}\left[\frac{1-D}{Q_0}\frac{k(Q_0)}{\left(1-p_*^a(X)\right)^2} G_1(k(Q_0)\, t\left(X, \zeta_0^t\right)'\delta_*^a + G^{-1}(Q_0)) t\left(X, \zeta_0^t\right) t\left(X, \zeta_0^t\right)'\right] \\
\underset{K\times 1+M}{M_{3\delta}} &= -\mathbb{E}\left[\frac{1-D}{Q_0}\frac{k(Q_0)}{\left(1-p_*^a(X)\right)^2} G_1(k(Q_0)\, t\left(X, \zeta_0^t\right)'\delta_*^a + G^{-1}(Q_0))\psi_0(Y_0, X, \gamma_0)\, t\left(X, \zeta_0^t\right)'\right] \\
\underset{K\times K}{M_{3\gamma}} &= \mathbb{E}\left[\frac{D}{Q_0}\frac{\partial\psi_1(Y_1, X, \gamma)}{\partial\gamma} - \frac{1-D}{Q_0}\frac{p_*^a(X)}{1-p_*^a(X)}\frac{\partial\psi_0(Y_0, X, \gamma)}{\partial\gamma}\right]
\end{aligned}$$

As in the missing data case, these expressions can be used to construct an analog estimate of Υ_0.

B.5 Proof of Theorem 4.2

Here we derive the large sample variance of $\widehat{\gamma}$ when both Assumptions 3.5 and 3.6 hold.

Form of Υ_0 when ρ_0 is unknown

$$
\begin{aligned}
m_1(Z,\rho_0) &= \begin{pmatrix} D-Q_0 \\ \frac{D}{Q_0} h\left(X,\zeta_0^t\right) \end{pmatrix} \\
m_2(Z,\rho_0,\delta_0^a) &= t_0 - \frac{1-D}{Q_0}\frac{p_0(X)}{1-p_0(X)} t(X,\zeta_0^t) \\
m_2(Z,\rho_0,\delta_0^a,\gamma_0) &= \frac{D}{Q_0}\psi_1(Y_1,X,\gamma_0) - \frac{1-D}{Q_0}\frac{p_0(X)}{1-p_0(X)}\psi_0(Y_0,X,\gamma_0)
\end{aligned}
$$

Under Assumptions 3.5 and 3.6 we have

$$
\underset{(M+1\times 1)}{A_a} = \frac{G_1\left(k(Q_0)\, t(X,\zeta_0^t)'\delta_0^a + G^{-1}(Q_0)\right)}{1-p_0(X)} \begin{pmatrix} k_1(Q_0)\, t(X,\zeta_0^t)'\delta_0^a + G_1\left(G^{-1}(Q_0)\right)^{-1} \\ k(Q_0)\begin{pmatrix} 0 & -I_M \end{pmatrix}' \delta_0^a \end{pmatrix},
$$

and hence the elements of the Jacobian matrix equal to

$$
\begin{aligned}
\underset{1+M\times 1+M}{M_{1\rho}} &= -I_{1+M} \\
\underset{M+1xM+1}{M_{2\rho}} &= \mathbb{E}\left[\frac{p_0(X)}{Q_0^2} t\left(X,\zeta_0^t\right) t_0' - \frac{1}{Q_0} t\left(X,\zeta_0^t\right) A_a' - \frac{p_0(X)}{Q_0}\begin{pmatrix} 0 & 0 \\ 0 & -I_M \end{pmatrix}\right] \\
\underset{K\times 1+M}{M_{3\rho}} &= \mathbb{E}\left[-\frac{p_0(X)}{Q_0^2} q_1(X;\gamma_0)\, t_0' + q_0(X,\gamma_0)\left\{\frac{p_0(X)}{Q_0^2} t_0' - \frac{1}{Q_0} A_a'\right\}\right] \\
\underset{1+M\times 1+M}{M_{2\delta}} &= -\mathbb{E}\left[\frac{1}{Q_0}\frac{k(Q_0)}{1-p_0(X)} G_1(k(Q_0)\, t\left(X,\zeta_0^t\right)'\delta_0^a + G^{-1}(Q_0)) t\left(X,\zeta_0^t\right) t\left(X,\zeta_0^t\right)'\right] \\
\underset{K\times 1+M}{M_{3\delta}} &= -\mathbb{E}\left[\frac{1}{Q_0}\frac{k(Q_0)}{1-p_0(X)} G_1(k(Q_0)\, t\left(X,\zeta_0^t\right)'\delta_0^a + G^{-1}(Q_0)) q_0(X;\gamma_0)\, t\left(X,\zeta_0^t\right)'\right] \\
\underset{K\times K}{M_{3\gamma}} &= \mathbb{E}\left[\frac{p_0(X)}{Q_0}\Gamma_0(X)\right].
\end{aligned}
$$

Notice that $M_{3\delta} = \Pi_0^* M_{2\delta}$.

Partitioning V_0 we have

$$
\begin{aligned}
\underset{1+M\times 1+M}{V_{11}} &= \begin{bmatrix} Q_0(1-Q_0) & 0 \\ 0 & \frac{p_0(X)}{Q_0^2} h\left(X,\zeta_0^t\right) h\left(X,\zeta_0^t\right)' \end{bmatrix} \\
\underset{1+M\times 1+M}{V_{21}} &= \mathbb{E}\begin{bmatrix} p_0(X)\, t\left(X,\zeta_0^t\right) & \mathbf{0}_{1+M\times M} \end{bmatrix} \\
\underset{K\times 1+M}{V_{31}} &= \mathbb{E}\begin{bmatrix} p_0(X)\frac{1-Q_0}{Q_0} q_1(X;\gamma_0) + p_0(X)\, q_0(X;\gamma_0) & \frac{p_0(X)}{Q_0^2} q_1(X;\gamma_0)\, h\left(X,\zeta_0^t\right)' \end{bmatrix} \\
\underset{1+M\times 1+M}{V_{22}} &= t_0 t_0' - \mathbb{E}\left[\frac{p_0(X)}{Q_0} t_0 t(X,\zeta_0^t)' + \frac{p_0(X)}{Q_0} t(X,\zeta_0^t) t_0' - \left\{\frac{p_0(X)}{Q_0}\right\}^2 \frac{t(X,\zeta_0^t) t(X,\zeta_0^t)'}{1-p_0(X)}\right] \\
\underset{K\times 1+M}{V_{32}} &= \mathbb{E}\left[\left\{\frac{p_0(X)}{Q_0}\right\}^2 \frac{q_0(X;\gamma_0)\, t(X,\zeta_0^t)'}{1-p_0(X)}\right] \\
\underset{K\times K}{V_{33}} &= \mathbb{E}\left[\left\{\frac{p_0(X)}{Q_0}\right\}^2 \left\{\frac{\mathbb{E}\left[\psi_1(Y_1,X,\gamma_0)\,\psi_1(Y_1,X,\gamma_0)'\middle| X\right]}{p_0(X)} + \frac{\mathbb{E}\left[\psi_0(Y_0,X,\gamma_0)\,\psi_0(Y_0,X,\gamma_0)'\middle| X\right]}{1-p_0(X)}\right\}\right].
\end{aligned}
$$

Some tedious algebra gives

$$\begin{aligned}
\Upsilon_0 &= \left\{M_0^{-1}V_0^{-1}M_0^{-1\prime}\right\}_{33} \\
&= M_{3\gamma}^{-1}\left[V_{33} + M_{3\delta}M_{2\delta}^{-1}V_{22}M_{2\delta}^{-1\prime}M_{3\delta}'\right. \\
&+ \left(M_{3\rho} - M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)M_{1\rho}^{-1}V_{11}M_{1\rho}^{-1\prime}\left(M_{3\rho} - M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)' \\
&+ \left(M_{3\delta}M_{2\delta}^{-1}V_{12}' - V_{13}'\right)M_{1\rho}^{-1\prime}\left(M_{3\rho} - M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)' \\
&+ \left(M_{3\rho} - M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)M_{1\rho}^{-1}\left(M_{3\delta}M_{2\delta}^{-1}V_{12}' - V_{13}'\right)' \\
&\left.-V_{23}'M_{2\delta}^{-1\prime}M_{3\delta}' - M_{3\delta}M_{2\delta}^{-1}V_{23}\right]M_{3\gamma}^{-1\prime}
\end{aligned}$$

Using the fact that $M_{3\delta}M_{2\delta}^{-1} = \Pi_0^*$, $\Pi_0^* t(X,\zeta_0^t) = q_0(X;\gamma_0)$ and $\Pi_0^* t_0 = \varsigma_0 + \Pi_0\zeta_0^t = \mathbb{E}\left[\psi_0(Y_0,X,\gamma)|\,D=1\right]$ we can show that

$$\begin{aligned}
M_{3\delta}M_{2\delta}^{-1}V_{22}M_{2\delta}^{-1\prime}M_{3\delta}' &= \Pi_0^* t_0 t_0'\Pi_0^* - \mathbb{E}\left[\frac{p_0(X)}{Q_0}\Pi_0^* t_0 t(X,\zeta_0^t)'\Pi_0^* + \frac{p_0(X)}{Q_0}\Pi_0^* t(X,\zeta_0^t)t_0'\Pi_0^* - \left\{\frac{p_0(X)}{Q_0}\right\}^2\frac{\Pi_0^* t(X,\zeta_0^t)t(X,\zeta_0^t)'\Pi_0^*}{1-p_0(X)}\right] \\
&= \mathbb{E}\left[\left\{\frac{p_0(X)}{Q_0}\right\}^2\frac{q_0(X;\gamma_0)\,q_0(X;\gamma_0)}{1-p_0(X)}\right] - \mathbb{E}\left[\psi_0(Y_0,X,\gamma_0)|\,D=1\right]\mathbb{E}\left[\psi_0(Y_0,X,\gamma_0)|\,D=1\right]',
\end{aligned}$$

and similarly that

$$M_{3\delta}M_{2\delta}^{-1}V_{23} = \mathbb{E}\left[\left\{\frac{p_0(X)}{Q_0}\right\}^2\frac{q_0(X;\gamma_0)\,q_0(X;\gamma_0)'}{1-p_0(X)}\right].$$

We further have

$$\begin{aligned}
M_{3\rho} - M_{3\delta}M_{2\delta}^{-1}M_{2\rho} &= \mathbb{E}\left[-\frac{p_0(X)}{Q_0^2}q_1(X;\gamma_0)\,t_0' + q_0(X;\gamma_0)\left\{\frac{p_0(X)}{Q_0^2}t_0' - \frac{1}{Q_0}A_a'\right\}\right] \\
&- \mathbb{E}\left[\frac{p_0(X)}{Q_0^2}\Pi_0^* t\left(X,\zeta_0^t\right)t_0' - \frac{1}{Q_0}\Pi_0^* t\left(X,\zeta_0^t\right)A_a' - \frac{p_0(X)}{Q_0}\Pi_0^*\begin{pmatrix}0 & 0\\ 0 & -I_M\end{pmatrix}\right] \\
&= \mathbb{E}\left[-\frac{p_0(X)}{Q_0^2}q_1(X;\gamma_0)\,t_0' + q_0(X;\gamma_0)\left\{\frac{p_0(X)}{Q_0^2}t_0' - \frac{1}{Q_0}A_a'\right\}\right] \\
&- \mathbb{E}\left[\frac{p_0(X)}{Q_0^2}q_0(X;\gamma_0)\,t_0' - \frac{1}{Q_0}q_0(X;\gamma_0)\,A_a' + \frac{p_0(X)}{Q_0}\begin{pmatrix}0 & \Pi_0\end{pmatrix}\right] \\
&= \mathbb{E}\left[-\frac{p_0(X)}{Q_0^2}q_1(X;\gamma_0)\,t_0' - \frac{p_0(X)}{Q_0}\begin{pmatrix}0 & \Pi_0\end{pmatrix}\right] \\
&= -\begin{pmatrix}\frac{\mathbb{E}[\psi_1(Y_0,X,\gamma)|D=1]}{Q_0} & \Pi_0\end{pmatrix},
\end{aligned}$$

and

$$\begin{aligned}
\underset{1+M\times 1+M}{V_{11}} &= \begin{bmatrix} Q_0(1-Q_0) & 0 \\ 0 & \frac{p_0(X)}{Q_0^2}h\left(X,\zeta_0^t\right)h\left(X,\zeta_0^t\right)'\end{bmatrix} \\
\underset{1+M\times 1+M}{V_{21}} &= \mathbb{E}\begin{bmatrix} p_0(X)\,t\left(X,\zeta_0^t\right) & \mathbf{0}_{1+M\times M}\end{bmatrix} \\
\underset{K\times 1+M}{V_{31}} &= \mathbb{E}\begin{bmatrix} p_0(X)\frac{1-Q_0}{Q_0}q_1(X;\gamma_0) + p_0(X)\,q_0(X;\gamma_0) & \frac{p_0(X)}{Q_0^2}q_1(X;\gamma_0)\,h\left(X,\zeta_0^t\right)'\end{bmatrix}.
\end{aligned}$$

This gives

$$\begin{aligned}
M_{3\delta}M_{2\delta}^{-1}V_{12}' - V_{13}' &= M_{3\delta}M_{2\delta}^{-1}V_{21} - V_{31} \\
&= \Pi_0^* V_{21} - V_{31} \\
&= \mathbb{E}\begin{bmatrix} p_0(X)\,q_0(X;\gamma_0) & \mathbf{0}_{K\times M}\end{bmatrix} \\
&- \mathbb{E}\begin{bmatrix} p_0(X)\frac{1-Q_0}{Q_0}q_1(X;\gamma_0) + p_0(X)\,q_0(X;\gamma_0) & \frac{p_0(X)}{Q_0^2}q_1(X;\gamma_0)\,h\left(X,\zeta_0^t\right)'\end{bmatrix} \\
&= -\mathbb{E}\begin{bmatrix} p_0(X)\frac{1-Q_0}{Q_0}q_1(X;\gamma_0) & \frac{p_0(X)}{Q_0^2}q_1(X;\gamma_0)\,h\left(X,\zeta_0^t\right)'\end{bmatrix} \\
&= -\begin{pmatrix}(1-Q_0)\,\mathbb{E}\left[\psi_1(Y_1,X,\gamma_0)|\,D=1\right] & \mathbb{E}\left[\frac{p_0(X)}{Q_0^2}q_1(X;\gamma_0)\,h\left(X,\zeta_0^t\right)'\right]\end{pmatrix}.
\end{aligned}$$

Combining the previous results gives

$$\begin{aligned}
\left(M_{3\delta}M_{2\delta}^{-1}V_{12}'-V_{13}'\right)M_{1\rho}^{-1\prime}\left(M_{3\rho}-M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)' &= -\begin{pmatrix} (1-Q_0)\,\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,D=1\right] & \mathbb{E}\left[\frac{p_0(X)}{Q_0^2}q_1\left(X;\gamma_0\right)h\left(X,\zeta_0^t\right)'\right] \end{pmatrix} \\
&\times\begin{pmatrix} \frac{\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|D=1\right]'}{Q_0} \\ \Pi_0' \end{pmatrix} \\
&= -\frac{1-Q_0}{Q_0}\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,D=1\right]\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,D=1\right]' \\
&-\mathbb{E}\left[\frac{p_0\left(X\right)}{Q_0^2}q_1\left(X;\gamma_0\right)h\left(X,\zeta_0^t\right)'\Pi_0'\right] \\
&= -\frac{1-Q_0}{Q_0}\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,D=1\right]\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,D=1\right]' \\
&-\mathbb{E}\left[\frac{p_0\left(X\right)}{Q_0^2}q_1\left(X;\gamma_0\right)\left[\left(\varsigma_0+\Pi_0h\left(X\right)\right)'-\left(\varsigma_0+\Pi_0\zeta_0^t\right)'\right]\right] \\
&= -\frac{1-Q_0}{Q_0}\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,D=1\right]\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,D=1\right]' \\
&-\mathbb{E}\left[\frac{p_0\left(X\right)}{Q_0^2}q_1\left(X;\gamma_0\right)q_0\left(X;\gamma_0\right)'\right] \\
&+\frac{1}{Q_0}\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,D=1\right]\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)|\,D=1\right]'
\end{aligned}$$

and

$$\begin{aligned}
&= \left(M_{3\rho}-M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)M_{1\rho}^{-1}V_{11}M_{1\rho}^{-1}\left(M_{3\rho}-M_{3\delta}M_{2\delta}^{-1}M_{2\rho}\right)' \\
&= \begin{pmatrix} \frac{\mathbb{E}\left[\psi_1(Y_1,X,\gamma)|D=1\right]}{Q_0} & \Pi_0 \end{pmatrix}\mathbb{E}\begin{pmatrix} Q_0\left(1-Q_0\right) & 0 \\ 0 & \frac{p_0(X)}{Q_0^2}h\left(X,\zeta_0^t\right)h\left(X,\zeta_0^t\right)' \end{pmatrix}\begin{pmatrix} \frac{\mathbb{E}\left[\psi_1(Y_1,X,\gamma)|D=1\right]'}{Q_0} \\ \Pi_0' \end{pmatrix} \\
&\begin{pmatrix} (1-Q_0)\,\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma\right)|\,D=1\right] & \mathbb{E}\left[\frac{p_0(X)}{Q_0^2}\Pi_0h\left(X,\zeta_0^t\right)h\left(X,\zeta_0^t\right)'\right] \end{pmatrix}\begin{pmatrix} \frac{\mathbb{E}\left[\psi_1(Y_1,X,\gamma)|D=1\right]'}{Q_0} \\ \Pi_0' \end{pmatrix} \\
&= \frac{1-Q_0}{Q_0}\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma\right)|\,D=1\right]\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma\right)|\,D=1\right]'+\mathbb{E}\left[\frac{p_0\left(X\right)}{Q_0^2}\Pi_0h\left(X,\zeta_0^t\right)h\left(X,\zeta_0^t\right)'\Pi_0'\right] \\
&= \frac{1-Q_0}{Q_0}\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma\right)|\,D=1\right]\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma\right)|\,D=1\right]' \\
&+\mathbb{E}\left[\frac{p_0\left(X\right)}{Q_0^2}\left[\left(\varsigma_0+h\left(X\right)\Pi_0\right)-\left(\varsigma_0+\Pi_0\zeta_0^t\right)\right]\left[\left(\varsigma_0+h\left(X\right)\Pi_0\right)-\left(\varsigma_0+\Pi_0\zeta_0^t\right)\right]'\right] \\
&= \frac{1-Q_0}{Q_0}\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma\right)|\,D=1\right]\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma\right)|\,D=1\right]' \\
&+\mathbb{E}\left[\frac{p_0\left(X\right)}{Q_0^2}q_0\left(X;\gamma_0\right)q_0\left(X;\gamma_0\right)'\right] \\
&-\frac{1}{Q_0}\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)|\,D=1\right]\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)|\,D=1\right]'
\end{aligned}$$

Using these expressions and the fact that $\mathbb{E}\left[\psi_1\left(Y_1,X,\gamma_0\right)|\,D=1\right]-\mathbb{E}\left[\psi_0\left(Y_0,X,\gamma_0\right)|\,D=1\right]=\mathbb{E}_t\left[\psi\left(Z,\gamma_0\right)\right]=0$, to evaluate Υ_0 gives

$$\begin{aligned}
\Upsilon_0 &= \mathbb{E}\left[\frac{p_0\left(X\right)}{Q_0}\Gamma_0\left(X\right)\right]^{-1} \\
&\times\mathbb{E}\left[\left\{\frac{p_0\left(X\right)}{Q_0}\right\}^2\left\{\frac{\Sigma_1\left(X;\gamma_0\right)}{p_0\left(X\right)}+\frac{\Sigma_0\left(X;\gamma_0\right)}{1-p_0\left(X\right)}+\frac{1}{p_0\left(X\right)}\left[q_1\left(X;\gamma_0\right)-q_0\left(X;\gamma_0\right)\right]\left[q_1\left(X;\gamma_0\right)-q_0\left(X;\gamma_0\right)\right]'\right\}\right] \\
&\times\mathbb{E}\left[\frac{p_0\left(X\right)}{Q_0}\Gamma_0\left(X\right)\right]^{-1\prime}
\end{aligned}$$

as claimed.

C Additional examples

In this appendix we briefly outline the application of IPT to some additional missing data and data combination problems.

C.1 Missing data examples

Variable probability sampling Assume that $\mathcal{Y}_1 \subset \mathbb{R}^{\dim(Y_1)}$ is partitioned into $1+M$ exhaustive and mutually exclusive strata

$$\mathcal{Y}_{10}, \mathcal{Y}_{11}, \ldots, \mathcal{Y}_{1M},$$

and let X be an $M \times 1$ vector of corresponding strata indicator variables,

$$X = (\mathbf{1}(Y_1 \in \mathcal{Y}_{11}), \ldots, \mathbf{1}(Y_1 \in \mathcal{Y}_{1M}))',$$

where the 0^{th} strata is omitted. A total of N draws are taken from some target population. Each draw is retained with a known probability depending on its strata, p_m, $m = 0, \ldots, M$. A total of $N_1 < N$ units are retained and their realizations of Y_1 recorded. No information on Y_1 is available for non-retained units. Either strata membership, X, for all sampled units is available or the population frequencies of each strata, $\mathbb{E}[X]$, are available. The moment function is as in Assumption 3.1 with $\psi_0(Y_0, X, \gamma)$ a vector of zeros.

In this example Assumptions 3.5 and 3.6 place no additional restrictions on the model. When the population strata frequencies, ζ_0, are known, Corollary 3.2 gives a variance bound of $\Gamma_0^{-1}\Lambda_0\Gamma_0^{-1\prime}$ with

$$\Lambda_0 = \sum\nolimits_{m=0}^{M} \zeta_{0,m} \left\{ \frac{\mathbb{V}(\psi_1(Y_1, X, \gamma_0) | Y_1 \in \mathcal{Y}_{1m})}{p_m} \right\},$$

which is the variance of Wooldridge's (1999) estimator (cf., Theorem 7.1, pp. 1399 - 1400). When ζ_0 is unknown but the strata of discarded units are available, Theorem 3.2 gives a bound of $\Gamma_0^{-1}\Lambda_0\Gamma_0^{-1\prime}$ with

$$\begin{aligned}&\sum\nolimits_{m=0}^{M} \zeta_{0,m} \left\{ \frac{\mathbb{V}(\psi_1(Y_1, X, \gamma_0) | Y_1 \in \mathcal{Y}_{1m})}{p_m} \right. \\ &\left. + \mathbb{E}[\psi_1(Y_1, X, \gamma_0) | Y_1 \in \mathcal{Y}_{1m}] \mathbb{E}[\psi_1(Y_1, X, \gamma_0) | Y_1 \in \mathcal{Y}_{1m}]' \right\}.\end{aligned}$$

This implies that Wooldridge's (1999) estimator for ζ_0 unknown is inefficient (in fairness he does not require knowledge of the strata of discarded units as we do here). In later work, Wooldridge (2007) proposes an efficient estimator for the unknown ζ_0 case. Some simple algebra shows that the IPT estimator is numerically identical to his.

C.2 Data combination examples

Small area estimation Let Y_0 be an indicator for household poverty and X a vector of household characteristics. We seek to estimate the poverty rate in a specific target municipality. Available is a random sample of N_t observations of X from this municipality. Also available is a random sample of size N_a of both Y_0 and X from the entire country. Our estimand is

$$\gamma_0 = \mathbb{E}_t[Y_0]$$

which corresponds to setting $\psi_1(Y_1, X, \gamma) = 0$ and $\psi_0(Y_0, X, \gamma) = Y_0 - \gamma$. In this example Assumption 4.2 establishes that the conditional distribution of Y_0 is the same in the entire country as it is in the specific municipality of interest. Under this assumption Tarozzi and Deaton (2007) identify γ_0 by

$$\gamma_0 = \int \mathbb{E}_a[Y_0 | X = x] \cdot f_t(x) \, \mathrm{d}m(x)$$

They then suggest applying the non-parametric estimator of Chen, Hong and Tarozzi (2008) to this problem. See Tarozzi (2007) for related applications. The application of IPT to this problem is straightforward.

Earnings decompositions Let Y_1 denote earnings in 1979 and Y_0 earnings in 1992. Let X be a vector of worker characteristics (e.g., age, race, education, union coverage). Available are two random samples of workers which record characteristics and earnings in 1979 and 1992. We seek to decompose changes in specific quantiles of the earnings distribution across the two periods into portions due to changes in the distribution of worker characteristics and changes in the mapping from characteristics to earnings.

Such a decomposition requires a characterization of the distribution of 1992 earnings that would prevail under the 1979 distribution of worker characteristics. The α^{th} quantile of this counterfactual distribution, $\gamma^{\alpha}_{92|79}$, is identified by

$$\mathbb{E}_t\left[\mathbf{1}(Y_0 \leq \gamma^{\alpha}_{92|79}) - \alpha\right] = 0,$$

which corresponds to setting $\psi_0(Y_0, X, \gamma) = \alpha - \mathbf{1}(Y_0 \leq \gamma^{\alpha}_{92|79})$ and $\psi_1(Y_1, X, \gamma)$ to a vector of zeros. Here 1979 and 1992 workers respectively play the role of the target and auxiliary populations.

The α^{th} quantiles of the actual 1979 and 1992 earnings distributions are denoted by $\gamma^{\alpha}_{79|79}$ and $\gamma^{\alpha}_{92|92}$. A decomposition into compositional and wage structure effects is then given by

$$\gamma^{\alpha}_{92|92} - \gamma^{\alpha}_{79|79} = \left(\gamma^{\alpha}_{92|92} - \gamma^{\alpha}_{92|79}\right) - \left(\gamma^{\alpha}_{79|79} - \gamma^{\alpha}_{92|79}\right).$$

Dinardo, Fortin and Lemieux (1996) and Barsky, Bound, Charles and Lupton (2002) develop alternative methods for earnings decompositions. Firpo, Fortin and Lemieux (2007) suggest a method of finer decomposition into the contributions from changes in each individual worker characteristic and its price.

M-estimation with endogenously stratified random samples Assume that $\mathcal{Y}_0 \subset \mathbb{R}^{\dim(Y_0)}$ is partitioned into $1+M$ exhaustive and mutually exclusive strata

$$\mathcal{Y}_{00}, \mathcal{Y}_{01}, \ldots, \mathcal{Y}_{0M}$$

and let X be an $M \times 1$ vector of corresponding strata indicator variables,

$$X = (\mathbf{1}(Y_0 \in \mathcal{Y}_{01}), \ldots, \mathbf{1}(Y_0 \in \mathcal{Y}_{0M}))',$$

with the 0^{th} strata omitted. A total of N_a draws of Y_0 are taken via multinomial sampling: with probability H_m the researcher draws (at random) from the m^{th} strata. The moment function is as in Assumption 4.1 with $\psi_1(Y_1, X, \gamma)$ a vector of zeros. A random sample of size N_t from the target population identifies the aggregate population strata shares $\zeta_0^t = \mathbb{E}_t[X]$ (alternatively these may be known a priori).

The distribution of (Y_0, X) in the stratified sample is connected to its target population counterpart by the equality

$$f_t(y_0, x) = \frac{1-Q_0}{Q_0} \frac{G\left(\alpha_0 + h(x)'\beta_0\right)}{1 - G\left(\alpha_0 + h(x)'\beta_0\right)} f_a(y_0, x),$$

where $h(x) = x$ and

$$\alpha_0 = G^{-1}\left(\left[1 + \frac{1-Q_0}{Q_0}\frac{H_0}{1-\sum_{m=1}^M \zeta_{0m}^t}\right]^{-1}\right)$$

and

$$\beta_0 = \begin{pmatrix} G^{-1}\left(\left[1 + \frac{1-Q_0}{Q_0}\frac{H_1}{\zeta_{01}^t}\right]^{-1}\right) - G^{-1}\left(\left[1 + \frac{1-Q_0}{Q_0}\frac{H_0}{1-\sum_{m=1}^M \zeta_{0m}^t}\right]^{-1}\right) \\ \vdots \\ G^{-1}\left(\left[1 + \frac{1-Q_0}{Q_0}\frac{H_M}{\zeta_{0M}^t}\right]^{-1}\right) - G^{-1}\left(\left[1 + \frac{1-Q_0}{Q_0}\frac{H_0}{1-\sum_{m=1}^M \zeta_{0m}^t}\right]^{-1}\right) \end{pmatrix}$$

with $H = (H, \ldots, H_m)'$ (cf., Imbens and Lancaster 1996, Equation (5), p. 294). This demonstrates that this model is contained within our data combination set-up as long as the probability of sampling each stratum is positive. Wooldridge (2001) and Tripathi (2007) also considers M-estimation with stratified random samples. Cosslett (1981) and Imbens and Lancaster (1996) consider the case where γ_0 indexes a conditional density. That case allows for more efficient estimation, but does not fall within our basic set-up.[31] Our estimator is semiparametrically efficient in the M-estimator case, but not the conditional likelihood case.

Non-classical measurement error Partition $X = (X_0, X_1', X_2')'$ with X_0 a noisy measure of Y_0, some regressor of interest, X_1 an outcome variable, and X_2 additional controls (measured without error). Available is a random sample of X from the target population and a separate 'validation sample' of Y_0 and X from some different, auxiliary, population. The moment function is as in Assumption 4.1 with $\psi_1(Y_1, X, \gamma)$ a vector of zeros. For example X_0 might be self-reported earnings and Y_0 earnings as recorded by the Social Security Administration. Robins, Hsieh and Newey (1995), Chen, Hong and Tamer (2005) and Chen, Hong and Tarozzi (2004, 2008) develop estimators for this problem. Ridder and Moffitt (2007) survey the literature.

References

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

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.

Arellano, Manuel and Costas Meghir. (1992). "Female labour supply and on-the-job search: an empirical model estimated using complementary data sets," *Review of Economic Studies* 59 (3): 537 - 559.

[31] Partition $Y_0 = (Y_{00}', Y_{01}')'$ and let γ index the parametric conditional density $f(y_{01}|y_{00};\gamma)$ with the marginal density $f(y_{00})$ left nonparametric; set $\psi_0(Y_0, X, \gamma) = -\frac{1}{f(y_{01}|y_{00};\gamma)}\frac{\partial f(y_{01}|y_{00};\gamma)}{\partial \gamma}$. Observe that $\mathbb{E}[\psi_0(Y_0, X, \gamma)|Y_{00}] = 0$ is conditionally mean zero. This extra information is not exploited by our estimator.

Bang, Heejung and James M. Robins. (2005). "Doubly robust estimation in missing data and causal inference models," *Biometrics* 61 (4): 962 - 972.

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.

Bickel, Peter J., Chris A.J. Klaassen, Ya'acov Ritov and Jon A. Wellner. (1993). *Efficient and adaptive estimation for semiparametric models.* New York: Springer-Verlag, Inc.

Borwein, J. M. and A.S. Lewis. (1991). "Duality relationships for entropy-like minimization problems," *Siam Journal of Control and Optimization* 29 (2): 325 - 338.

Chen, Xiaohong, Han Hong and Elie Tamer. (2005). "Measurement error models with auxiliary data," *Review of Economic Studies* 72 (2): 343 - 366.

Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2004). "Semiparametric efficiency in GMM models of nonclassical measurement errors, missing data and treatment effects, *Mimeo.*

Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2008). "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics* 36 (2): 808 - 843.

Corcoran, Stephen A. (1998). "Bartlett adjustment for empirical discrepancy statistics," *Biometrika* 85 (4): 967 - 972.

Cosslett, Stephen R. (1981). "Maximum likelihood estimator for choice-based samples," *Econometrica* 49 (5): 1289 - 1316.

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

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.

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.

Efron, Bradley and Robert Tibshirani. (1996). "Using special designed exponential families for density estimation," *Annals of Statistics* 24 (6): 2431 - 2461.

Firpo, Sergio, Nicole Fortin and Thomas Lemieux. (2007). "Decomposing wage distributions using recentered influence functions," *Mimeo.*

Graham, Bryan S. (2007). "GMM 'equivalence' for semiparametric missing data models ," *Mimeo.*

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

Heckman, James J. (1987). "Selection bias and self-selection," *The New Palgrave: A Dictionary of Economics*: 287 - 297 (J. Eatwell et al. Eds.). London: Macmillan Press Ltd.

Heckman, James J. Richard Robb, Jr. (1985). "Alternative methods for evaluating the impact of interventions," *Longitudinal Analysis of Labor Market Data*: 156 - 245 (J.J. Heckman & B. Singer, Eds.). Cambridge: Cambridge University Press.

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.

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.

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.

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.

Ichimura, Hidehiko and Oliver Linton. (2005). "Asymptotic expansions for some semiparametric program evaluation estimators," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: 149 -170 (D.W.K Andrews & J.H. Stock, Eds). Cambridge: Cambridge University Press.

Imbens, Guido. W. (1997). "One-step estimators of over-identified generalized method of moments models," *Review of Economic Studies* 64 (3): 359 - 383.

Imbens, Guido. W. (2002). "Generalized method of moments and empirical likelihood," *Journal of Business and Economic Statistics* 20 (4): 493 - 506.

Imbens, Guido W. (2004). "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics* 86 (1): 4 - 29.

Imbens, Guido W. and Tony Lancaster. (1994). "Combining micro and macro data in microeconometric models," *Review of Economic Studies* 61 (4): 655 - 680.

Imbens, Guido W. and Tony Lancaster. (1996). "Efficient estimation and stratified sampling," *Journal of Econometrics* 74 (2): 289 - 318.

Imbens, Guido W., Whitney K. Newey and Geert Ridder (2007). "Mean-square-error calculations for average treatment effects," *Mimeo.*

Imbens, Guido W, Richard Spady and Phillip Johnson (1998). "Information theoretic approaches to inference in moment condition models," *Econometrica* 66 (2): 333 - 357.

Kitamura, Yuichi. (2007). "Empirical likelihood methods in econometrics: theory and practice," *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress* 3: 174 - 237 (R. Blundell, W. Newey & T. Persson, Eds.). Cambridge: Cambridge University Press.

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

Lancaster, Tony and Guido W. Imbens. (1996). "Case-control studies with contaminated controls," *Journal of Econometrics* 71 (1-2): 145 - 160.

Little, Roderick J.A. and Donald B. Rubin. (2002). *Statistical analysis with missing data, 2nd Ed.* Hoboken, NJ: Jon Wiley & Sons, Inc.

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.

Lunceford, Jared K. and Marie Davidian. (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in Medicine* 23 (19): 2937 - 2960.

Manski, Charles F. (2003). *Partial Identification of Probability Distributions.* New York: Springer-Verlag.

Nevo, Aviv. (2002). "Sample selection and information-theoretic alternatives to GMM," *Journal of Econometrics* 107 (1-2): 149 - 157.

Nevo, Aviv. (2003). "Using weights to adjust for sample selection when auxiliary information is available," *Journal of Business and Economic Statistics* 21 (1): 43 - 52.

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

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). Amsterdam: North Holland.

Newey, Whitney K. and Richard J. Smith. (2004). "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica* 72 (1): 219 - 255.

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

Qian, Hailong and Peter Schmidt. (1999). "Improved instrumental variables and generalized method of moments estimators," *Journal of Econometrics* 91 (1): 145 - 169.

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

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 B* 69 (1): 101 - 122.

Ridder, Geert and Robert Moffitt. (2007). "The econometrics of data combination," *Handbook of Econometrics* 6B: 5469 - 5547 (J.J. Heckman & E Leamer, Eds.). New York: North-Holland.

Robins, James M., Fushing Hsieh and Whitney Newey. (1995). "Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates," *Journal of the Royal Statistical Society B* 57 (2): 409 - 424.

Robins, James M. and Ya'acov Ritov. (1997). "Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models," *Statistics in Medicine* 16 (3): 285 - 319.

Robins, James M., Andrea Rotnitzky and Mark ban der Laan. (2000). "On profile likelihood: a comment," *Journal of the American Statistical Association* 95 (450): 477 - 482.

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.

Rockafellar, R. Tyrell. (1970). *Convex Analysis.* Princeton: Princeton University Press.

Rosenbaum,Paul R. (1987). "Model-based direct adjustment," *Journal of the American Statistical Association* 82 (398): 387 - 394.

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.

Tarozzi, Alessandro. (2007). "Calculating comparable statistics from incomparable surveys with an application to poverty in India," *Journal of Business and Economic Statistics* 25 (3): 314 - 336.

Tarozzi, Alessandro and Angus Deaton. (2007). "Using census and survey data to estimate poverty and inequality for small areas," *Mimeo.*

Tripathi, Gautam. (2007). "Moment based inference with stratified data," *Econometric Theory*, forthcoming.

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

Wang, Qihua, Oliver Linton and Wolfgang Härdle. (2004). "Semiparametric regression analysis with missing response at random," *Journal of the American Statistical Association* 99 (466): 334 - 345.

Wooldridge, Jeffrey M. (1999). "Asymptotic properties of weighted M-estimators for variable probability samples," *Econometrica* 67 (6): 1385 - 1406.

Wooldridge, Jeffrey M. (2001). "Asymptotic properties of weighted M-estimators for standard stratified samples," *Econometric Theory* 17 (2): 451 - 470.

Wooldridge, Jeffrey M. (2007). "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics* 141 (2): 1281 - 1301.