Inverse Probability Tilting and Missing Data Problems[1]

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

Initial Draft: July 2006

This Draft: April 24, 2007

Very Preliminary and Incomplete

(Please contact corresponding author for latest draft)

To be completed....

JEL Classification: C14, C21, C23

Key Words: Missing Data, Minimum Empirical Discrepancy, Generalized Empirical Likelihood, Inverse Probability Weighting, Double Robustness, Sample Selection, Attrition, Treatment Effects, Stratified Sampling, Semiparametric Efficiency

[1]We would like to thank Stephen Cosslett, Jinyong Hahn, Guido Imbens, Michael Jansson, Geert Ridder, Richard Smith, and members of the Berkeley Econometrics Reading Group for helpful discussions. We also acknowledge feedback and suggestions from participants in seminars at the University of Pittsburgh, Ohio State University and the University of Southern California. All the usual disclaimers apply.

[+]Department of Economics, University of California - Berkeley, 549 Evans Hall #3880, Berkeley, CA 94720 E-mail: egel@econ.berkeley.edu. Web: https://webfiles.berkeley.edu/~egel/

[†]*Corresponding Author.* Department of Economics, University of California - Berkeley, 549 Evans Hall #3880, Berkeley, CA 94720 and National Bureau of Economic Research. E-mail: bgraham@econ.berkeley.edu. Web: http://www.econ.berkeley.edu/~bgraham/.

[◇]Department of Economics, University of California - Berkeley, 549 Evans Hall #3880, Berkeley, CA 94720. E-mail: cristine@econ.berkeley.edu.

1 Introduction

Let $\{D, X', DY_1'\}_{i=1}^{\infty}$ be an independent and identically distributed random sequence drawn from the unknown distribution F_0 with D a binary 'missingness' indicator. When $D = 1$ we observe both $X \in \mathcal{X} \subset R^{\dim(X)}$ and $Y_1 \in \mathcal{Y}_1 \subset R^{\dim(Y_1)}$, when $D = 0$ we only observe X. The sampling process identifies $F_0(y_1, x|D = 1)$ and $F_0(d, x)$, but we seek to identify functionals of $F_0(y_1, x)$, a distribution not identified by this process alone. To ensure identification we assume that Y_1 is 'missing-at-random' (MAR)[2] conditional on $X = x$:

$$F_0(y_1|x) = F_0(y_1|x, D = 1) = F_0(y_1|x, D = 0). \tag{1}$$

The only other prior restriction on F_0 is that for some unique $\gamma_0 \in \mathcal{G} \subset R^{\dim(\gamma)}$

$$\mathbb{E}_{F_0}[m(Y_1, X, \gamma_0)] = \int\int m(y_1, x, \gamma_0) f_0(y_1, x) \, \mathrm{d}y_1 \mathrm{d}x = 0, \tag{2}$$

where $m(y_1, x, \gamma)$ is an $L \times 1$ known function of y_1 and x indexed by γ. For simplicity we consider the exactly identified case when $\dim(\gamma) = L$.

A large body of research explores identification and estimation of γ_0 under restrictions (1) and (2). 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, Imbens 2004), missing regressors (Robins, Rotnitzky and Zhao 1994) and non-classical measurement error (Chen, Hong and Tarozzi 2004). Below we survey additional applications.

Existing approaches to efficient estimation of γ_0 exploit one of two alternative factorizations of $f_0(y_1, x)$ implied by the MAR restriction. The imputation approach substitutes $f_0(y_1|x, D = 1) f_0(x)$ for $f_0(y_1, x)$ in (2). Imputation estimators then replace $f_0(y_1|x, D = 1)$ with a nonparametric estimate and $f_0(x)$ with the empirical measure of the complete sample. Hahn (1998), Chen, Hong and Tarozzi (2004) and Imbens, Newey and Ridder (2005) pursue variants of this approach.

Inverse probability weighting (IPW) uses the factorization $f_0(y_1, x) = f_0(y_1, x|D = 1) Q_0/p_0(x)$, where $p_0(x) = \mathbb{E}_{F_0}[D|X = x]$ is the propensity score and $Q_0 = \mathbb{E}_{F_0}[D]$ the marginal probability of complete observation. IPW estimators replace $p_0(x)$ with a nonparametric estimate and $f_0(y_1, x|D = 1)$ with the empirical measure of the $D = 1$ subsample. Hirano, Imbens and Ridder (2003), Wooldridge (2007) and Chen, Hong and Tarozzi (2004) pursue this approach.

Imputation and IPW are, under appropriate conditions on the estimates of $f_0(y_1|x, D = 1)$ and $p_0(x)$, globally efficient. Unfortunately $f_0(y_1|x, D = 1)$ and $p_0(x)$ can be difficult to estimate nonparametrically in moderate-sized samples.

Robins, Rotnizky and Zhao (1994) proposed an augmented inverse probability weighting (AIPW) estimator. AIPW is an M-estimator based on a parametric estimate of the efficient score function (cf., Graham 2007). It requires auxiliary parametric models for $f_0(y_1|x, D = 1)$ and $p_0(x)$. When the true data generating process satisfies these auxiliary restrictions AIPW is locally efficient (cf., Newey 1990). Consistency of the estimator, however, only requires the restrictions on $f_0(y_1|x, D = 1)$ or $p_0(x)$ to be

[2] Little and Rubin (2002) and Manski (2003) provide comprehensive discussions of restriction (1).

true. In the language of Scharfstein, Rotnitzky and Robins (1999) AIPW is 'doubly robust'. Graham (2007) provides a method-of-moments interpretation of double robustness.

This paper proposes a new approach to estimation: inverse probability tilting (IPT). Like AIPW, the proposed method requires making auxiliary parametric assumptions about $f_0(y_1|x, D=1)$ and $p_0(x)$, is efficient when these assumptions are true, and robust to departures from one or the other of them. Unlike AIPW, only the nuisance parameter indexing the propensity need be estimated (i.e., an estimate of that indexing $f_0(y_1|x, D=1)$ is not required).

We begin by approximating $F_0(y_1, x)$ with a multinomial distribution with support coinciding with that of the $D=1$ selected subsample. The distribution of probability mass is chosen to be as close as possible to the empirical measure of the selected subsample, subject to the restriction that various moments of X, calculated using the estimated distribution, are identical to the corresponding sample moments calculated using all observations of X. We refer to this latter property as 'exact balancing' (of moments). Making the 'close as possible' terminology precise requires choosing a discrepancy metric, a choice that is equivalent to choosing a model for the propensity score (Little and Wu 1991, Hirano, Imbens, Ridder and Rubin 2001).

When the propensity score is correctly modelled, and hence the discrepancy metric appropriately chosen, the inverse probability 'tilt' of the empirical measure of the selected subsample provides a consistent estimate of the true distribution function $F_0(y_1, x)$. When this is not the case, however, the tilted distribution can still be used to consistently estimate γ_0 when certain auxiliary restrictions are satisfied (i.e., the method is 'doubly robust').

Specific applications of inverse probability tilting have been proposed elsewhere. Little and Wu (1991) suggest the method for calibration of 2×2 contingency tables to known margins when selection is logistic. Nevo (2002, 2003) extends their results to moment condition models. Hirano, Imbens, Ridder and Rubin (2001), in the context of a creative proof of just identification of their additive non-ignorable attrition model, formalize the mapping between discrepancy metrics and propensity scores.

Distinctive contributions of our work include: (1) complete generalization of the method, (2) application to a far wider range of models, (3) demonstration of local efficiency and double robustness and (4) providing duality results which facilitate both computation and interpretation. We apply the method to a wide range of missing data and data combination problems. Local efficiency and double robustness hold in all cases. Examples not covered by the work cited above include estimation of the average treatment effect (ATE), M-estimation with variably probability samples or missing regressors, estimation of the average treatment effect on the treated (ATT), two sample instrumental variables estimation when the two samples are not random draws from the same population, small area estimation and non-classical measurement error with validation samples. We are also able to substantially extend Hellerstein and Imbens (1999) results on M-estimation when the sampled and target populations don't coincide.

Semiparametric efficiency and double robustness are also properties of AIPW. Advantages of IPT, relative to AIPW, include applicability to a wider-range of estimands, the need to estimate fewer nuisance parameters, and the exact balancing property of the IPT estimate of $F_0(y_1, x)$. Inverse probability tilting could be generalized to achieve global efficiency. Such a generalization would involve increasing the number of moments balanced as the sample size grows. In practice applications will involve flexible

parametric specifications and hence we develop theory appropriate for this case. In this context IPT is particularly attractive relative to competing approaches: parametric imputation is not robust to misspecification of $f_0(y_1|x, D=1)$ (Imbens 2004) and parametric inverse probability weighting is generally neither robust nor efficient (Chen, Hong and Tarozzi 2004, Graham 2007).

Papers to discuss:

Anderson (1982)

Qin and Zhang (2007)

Efron and Ibshirani (1996)

2 The population problem in a nutshell

It is helpful to begin by informally considering the 'population' properties of inverse probability tilting. Recall that the assumed sampling process asymptotically reveals $F_0(y_1, x|D=1)$ and $F_0(d, x)$. Let $t_0 = (1, \underline{0}')'$ and

$$t(x, \zeta_0) = \begin{pmatrix} 1 \\ h(x) - \zeta_0 \end{pmatrix},$$

with $h(x)$ an $M \times 1$ vector of known functions of x with mean ζ_0 under F_0 (i.e., $\mathbb{E}_{F_0}[h(X)] = \zeta_0$).[3] We assume that $\mathbb{E}_{F_0}\left[(h(X) - \zeta_0)(h(X) - \zeta_0)'|D=1\right]$ is of full rank.

Now consider the properties of the inverse probability tilted distribution F_* with density

$$f_*(y_1, x) = \frac{f_0(y_1, x|d=1) Q_0}{G\left(k(Q_0) t(x, \zeta_0)' \delta_* + G^{-1}(Q_0)\right)},$$

where δ_* is the solution to

$$\max_{\delta} \left\{ t_0'\delta - \mathbb{E}_{F_0}\left[\varphi^+\left(t(X, \zeta_0)'\delta; Q_0\right)|D=1\right]\right\}, \tag{3}$$

with $\varphi_j^+(v;\cdot) = \partial^j \varphi^+(v;\cdot)/\partial v^j$ for $j = 0, 1, 2, \ldots$ and $\varphi^+(x; Q)$ given by

$$\varphi^+(x; Q) \overset{def}{\equiv} \frac{1}{k(Q)} \left[\frac{k(Q) x + G^{-1}(Q)}{G(k(Q) x + G^{-1}(Q))} Q + \int_{Q/G(k(Q)x + G^{-1}(Q))}^{a} G^{-1}\left(\frac{Q}{t}\right) \mathrm{d}t \right],$$

for $k(Q) \overset{def}{\equiv} -Q/G_1\left(G^{-1}(Q)\right)$ and $G(\cdot)$ an increasing, differentiable and continuous function mapping

[3] Note that both ζ_0 and Q_0 are asymptotically identified since realizations of D and X are recorded for all units.

the real line onto the unit interval.[4] The $1+M$ first order conditions for (3) are

$$\begin{aligned}\int\int \frac{f_0\left(y_1,x|d=1\right)Q_0}{G\left(k\left(Q_0\right)t\left(x,\zeta_0\right)'\delta_*+G^{-1}\left(Q_0\right)\right)}\mathrm{d}y_1\mathrm{d}x &= \int\int f_*\left(y_1,x\right)\mathrm{d}y_1\mathrm{d}x=1 \\ \int\int h(x)\frac{f_0\left(y_1,x|d=1\right)Q_0}{G\left(k\left(Q_0\right)t\left(x,\zeta_0\right)'\delta_*+G^{-1}\left(Q_0\right)\right)}\mathrm{d}y_1\mathrm{d}x &= \int\int h(x)f_*\left(y_1,x\right)\mathrm{d}y_1\mathrm{d}x=\zeta_0.\end{aligned} \tag{4}$$

The inverse probability tilted density $f_*\left(y_1,x\right)$ is a proper density and shares (at least) M moments with F_0.

Our main identification result concerns equality of γ_0 and γ_*, where γ_* is the root to the tilted analog of (2)

$$\mathbb{E}_{F_*}\left[m\left(Y_1,X,\gamma_*\right)\right]=\int\int m\left(y_1,x,\gamma_*\right)f_*\left(y_1,x\right)\mathrm{d}y_1\mathrm{d}x=0. \tag{5}$$

We can show that $\gamma_*=\gamma_0$ if at least one of two conditions hold: (1) $p_0\left(x\right)=G\left(\alpha_0+h\left(x\right)'\beta_0\right)$ for all $x\in\mathcal{X}$ or (2) $\mathbb{E}_{F_0}\left[m\left(Y_1,X,\gamma_0\right)|X=x\right]=\Upsilon_0 w\left(x\right)$ for all $x\in\mathcal{X}$, some $L\times 1+M$ matrix Υ_0 and $w\left(x\right)=\left(1,h\left(x\right)'\right)'$.

Equality of $\gamma_*=\gamma_0$ under the first condition follows from the fact that after making the substitutions $\zeta_0=\mathbb{E}_{F_0}\left[h(X)\right]$ and

$$\alpha_0=k\left(Q_0\right)\eta^*+G^{-1}\left(Q_0\right),\ \beta_0=k\left(Q_0\right)\lambda^*,$$

where $\delta=\left(\eta,\lambda'\right)'$ and using iterated expectations we can show equivalence of (4) and (5) with the pair of moment restrictions

$$\mathbb{E}_{F_0}\left[\left\{\frac{D}{G\left(\alpha_0+h\left(X\right)'\beta_0\right)}-1\right\}w\left(X\right)\right] = 0 \tag{6}$$

$$\mathbb{E}_{F_0}\left[\frac{Dm\left(Y_1,X,\gamma_0\right)}{G\left(\alpha_0+h\left(X\right)'\beta_0\right)}\right] = 0, \tag{7}$$

which identify α_0, β_0 and γ_0 if $p_0\left(x\right)=G\left(\alpha_0+h\left(x\right)'\beta_0\right)$ and the MAR restriction holds.

Equality of $\gamma_*=\gamma_0$ under the second condition is less obvious. In this case we allow the propensity score to be misspecified (i.e., there are no α_0 and β_0 such that $p_0\left(x\right)=G\left(\alpha_0+h\left(x\right)'\beta_0\right)$ for all $x\in\mathcal{X}$. Let α_* and β_* denote the solution to (6) under misspecification. Exploiting the well-known equivalence of (6) and (7) with the pair which replaces (7) with the population residual associated with the (mean squared error minimizing) linear predictor of (7) given (6) (e.g., Qian and Schmidt 1999), yields the

[4] Under these restrictions on $G\left(\cdot\right)$, $\varphi^+\left(x;\cdot\right)$ is concave and increasing in x.That $\varphi^+(x;Q)$ is increasing in x follows from the fact that

$$\varphi_1^+(x;Q)=\frac{Q}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)},$$

is bounded below by zero. Concavity from the fact that

$$\varphi_2^+(x;Q)=\frac{G_1\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^2}k\left(Q\right)Q$$

is negative for all $x\in\mathbb{R}^1$.

equivalent system

$$\begin{aligned}\mathbb{E}_{F_0}\left[\left\{\frac{D}{G\left(\alpha_* + h\left(X\right)'\beta_*\right)} - 1\right\} w\left(X\right)\right] &= 0 \\ \mathbb{E}_{F_0}\left[\frac{Dm\left(Y_1, X, \gamma_0\right)}{G(\alpha_* + h\left(X\right)'\beta_*)} - \frac{\Upsilon_0 w\left(X\right)}{G(\alpha_* + h\left(X\right)'\beta_*)}\left(D - G(\alpha_* + h\left(X\right)'\beta_*)\right)\right] &= 0.\end{aligned}$$

Applying iterated expectations, the MAR asumption and the second condition $\mathbb{E}_{F_0}\left[m\left(Y_1, X, \gamma_0\right)|X = x\right] = \Upsilon_0 w\left(x\right)$ to the residualized moment gives

$$\mathbb{E}_{F_0}\left[\Upsilon_0 w\left(X\right)\right] = \mathbb{E}_{F_0}\left[\mathbb{E}_{F_0}\left[m\left(Y_1, X, \gamma_0\right)|X\right]\right] = \mathbb{E}_{F_0}\left[m\left(Y_1, X, \gamma_0\right)\right] = 0.$$

The restrictions (6) and (7) therefore identify γ_0 when the second condition holds even under misspecification of the propensity score.

3 Sampling structure and estimands

We begin by describing our basic sampling framework in detail. Available are two samples. The study or treatment sample consists of N_1 records of Y_1, X. This sample is a random sample from the *study population.* The auxiliary or control sample consists of N_0 records of Y_0, X. It is a random sample from the *auxiliary population.* The merged sample consists of $N = N_1 + N_1$ records of D, X, DY_1 and $(1 - D)\,Y_0$, where D is a binary variable taking a value of one if a unit originates from the study sample and value of zero if it originates from the auxiliary sample. That is, D indicates whether a unit is member of the study or auxiliary populations. Often this distinction reflects some measurable attribute of a unit. In program evaluation problems D indicates exposure to a policy ($D = 1$) or not ($D = 0$). Without loss of generality we assume that the first N_1 observations in the merged sample corresponds to study units and the remaining N_0 to auxiliary units.

It is helpful, both conceptually and for the development large sample theory, to think of the merged sample as arising from a simple, albeit fictional, multinomial sampling scheme: with probability Q_0 a unit is randomly drawn from the study population and its values of Y_1 and X recorded, with probability $1 - Q_0$ a unit is randomly drawn from the auxiliary population and its values of Y_0 and X recorded. This is repeated N times such that as $N \to \infty$ we have $N_1/N \xrightarrow{p} Q_0$. Realizations of Y_0 are not recorded for units drawn from the study population, while realizations of Y_1 are not recorded for units drawn from the auxiliary population.

Each of our motivating examples is based on some variant of the above sampling scheme. Some of our more complicated examples, particularly those with 'non-ignorable' missing data, require augmentations of the above scheme. We defer a discussion of this until Section 7. It is helpful to organize our main examples into two classes of problems. We refer to the first class as *missing data problems* and the second as *sample combination problems.* While this terminology is imperfect, we feel it best captures the fundamental identification challenge of each set of examples.

In both cases the study and auxiliary populations are distinct subpopulations which together form some larger *target population.*[5] In missing data problems the multinomial sampling probabilities at-

[5]The term 'target population' is borrowed from Little and Wu (1991). However, our use of the term differs from theirs.

tached to the study and auxiliary populations equal their respective shares in the target population. That is $Q_0 = P_0$, with P_0 denoting the probability that a random draw from the target population is a member of the study subpopulation. In this case the merged sample is (equivalent to) a random sample of $Z = (Y_1', Y_0', X')'$ from the target population where, with some probability, possibly depending on Z, some elements of Z not recorded for some sampled units. The goal is to identify some feature of the target population. Let F be the distribution function for this population. Our starting point is the following prior restriction on F.

Assumption 3.1 (Global Identification in the Target Population) *For some known function* $\psi(z,\gamma) = m(y_1, x, \gamma) + l(y_0, x, \gamma)$

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

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

The notation $\mathbb{E}_F[\cdot]$ denotes expectations taken with respect to F.

In data combination problems the goal is to identify some feature of the study population. In these problems Q_0 typically differs from P_0; that is the study and auxiliary samples are typically collected independently. LaLonde (1986) works with a study sample of National Supported Work (NSW) Demonstration participants and an auxiliary sample on non-participants from the Current Population Survey (CPS). Let H denote the distribution of (Z, D) induced by the multinomial sampling scheme. Its associated density is

$$h(z,d) = [Q_0 f(z|d=1)]^d [(1-Q_0) f(z|d=0)]^{1-d}. \tag{8}$$

This density characterizes a hypothetical *combined population.* Observe that the distribution of Z in the study population equals the conditional distribution of Z given $D = 1$ in the combined population. For data combination problems our starting point is following prior restriction on the study/combined population.

Assumption 3.2 (Global Identification in the Study Population) *For some known function* $\psi(z,\gamma) = m(y_1, x, \gamma) + l(y_0, x, \gamma)$

$$\mathbb{E}_H[\psi(Z,\gamma_0)|D=1] = \int \psi(z,\gamma_0)\, dF(z|d=1) = 0,$$

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

Since $\mathbb{E}_H[\cdot|D=1] = \mathbb{E}_F[\cdot|D=1]$ we will often omit subscripts from the expectations operator in what follows. The distinction between the target and the combined population is important only for determining whether the available merged sampled is suitable for identifying γ_0 as defined by Assumption 3.1 or 3.2. A good discussion of this issue in the context of evaluating the NSW Demonstration is found in Dehejia and Wahba (1999, p. 1057). When discussing estimation of γ_0 as defined by Assumption 3.1 we implicitly assume that $Q_0 = P_0$.

Assumptions 3.1 and 3.2 are closely related to the distinction between the 'average treatment effect' (ATE) and the 'average treatment effect on the treated' (ATT) estimands of the program evaluation literature (e.g., Hahn 1998, Hirano, Imbens and Ridder 2003, Imbens 2004) and the 'verify-in-sample' and 'verify-out-of-sample' distinction made in the context of estimating nonclassical measurement error models with validation samples as in Chen, Hong and Tarozzi (2005). Assumptions 3.1 and 3.2 are sufficiently flexible to allow us to apply inverse probability tilting to most of the missing data and biased sampling problems of which we are aware as well as to provide results for several new problems.

4 Examples

To illustrate the range of problems to which inverse probability tilting can be applied we introduce, and subsequently develop in some detail, the following concrete examples.

4.1 Missing data problems

Our first class of problems are explicit missing data problems. The leading case of such problems are those with data 'missing at random' (MAR), in the sense defined by Little and Rubin (2002). Efficient estimation under the MAR set-up has been widely studied since the pioneering work of Robins, Rotnitzky and Zhao (1994). We focus on this case initially. In Section 7, we show how inverse probability tilting can also be applied to non-ignorable missing data problems. A leading example is provided by the two-period panel data problem with second period attrition depending on unobservables studied by Hirano, Imbens, Ridder and Rubin (2001). Such examples require generalizing our sampling framework to include additional auxiliary samples.

Variable probability sampling with known population strata frequencies (Wooldridge 1999) Assume that $(\mathcal{Y}_1, \mathcal{X}) \subset \mathbb{R}^{\dim(Y_1)} \times \mathbb{R}^{\dim(X)}$ is partitioned into $M+1$ exhaustive and mutually exclusive strata $(\mathcal{Y}_{10}, \mathcal{X}_0), (\mathcal{Y}_{11}, \mathcal{X}_1), \ldots (\mathcal{Y}_{1M}, \mathcal{X}_M)$. A total of N draws are taken from the target population. Each draw is retained with known probabilities $p_0, \ldots, p_M$, which vary across strata. A total of $N_1 < N$ units are retained and their realizations of Y_1 and X recorded. No information on Y_1 or X is available for non-retained units. Either knowledge of the initial sample size, N, or $P_0 = \Pr(D = 1)$ is available. The moment vector is as given by Assumption 3.1 with $l(y_0, x, \gamma)$ a vector of zeros. Table 1 summarizes the structure of the moment function for this example and each of those that follow.

Missing regressors or expensive covariates (Robins, Rotnitzky and Zhao 1994) We partition $X = (X_0, X_1')'$ with X_0 the outcome of interest and X_1 controls. We assume Y_0 is a noisy measure of, or surrogate for, Y_1, the regressor of interest. The moment function corresponds to setting

$$m(y_1, x; \gamma) = \frac{\partial e(y_1, x_1, \gamma)}{\partial \gamma} (x_0 - e(y_1, x_1, \gamma)),$$

with $e(y_1, x_1, \gamma)$ a known (conditional expectation) function indexed by γ, and $l(y_0, x; \gamma)$ a vector of zeros. For example Y_1 might be calorie consumption measured using a long questionnaire which is expensive to administer and/or difficult to get sampled individuals to complete and Y_0 might be a calorie consumption measure associated with a shorter questionnaire that is inexpensive to administer. The

outcome variable, X_0, might be some aspect of child health or early childhood cognitive development and X_1 various demographic controls. This example could easily be modified to incorporate endogenous regressors.

Average Treatment Effects (ATE) (Rosenbaum and Rubin 1983, Imbens 2004) Let $D = 1$ denote assignment to active treatment and $D = 0$ assignment to the status quo or control treatment; Y_1 and Y_0 are (potential) outcomes under the active treatment and status quo respectively; X is a vector of pre-treatment controls. The estimand of interest is the average treatment effect (ATE)

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

which is equivalent to γ_0 defined by Assumption 3.1 with $m\left(y_1, x, \gamma\right) = y_1$, $l\left(y_0, x, \gamma\right) = -y_0 - \gamma$.

4.2 Sample combination problems

The second family of problems to which inverse probability tilting can be applied are sample combination problems.[6] A leading example is combining a stratified random sample from the target population with a pure random sample from that population.

Two sample instrumental variables (TSIV) (Angrist and Krueger 1992, Arellano and Meghir 1992) Currie and Yelowitz (2000) consider a model where Y_1 indicates whether a school-aged child has repeated a grade, Y_0 indicates residence in public housing, and $X = \left(X_0, X_1'\right)'$ 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 3.2 with $m\left(y_1, x, \gamma\right) = g\left(x\right) d\left(y_1, \gamma\right)$ and $l\left(y_0, x, \gamma\right) = -g\left(x\right) e\left(y_0, x_1, \gamma\right)$. That is, by

$$\psi\left(z, \gamma\right) = g\left(x\right)\left(d\left(y_1, \gamma\right) - e\left(y_0, x_1, \gamma\right)\right),$$

where $d\left(\cdot, \gamma\right)$, $g\left(\cdot\right)$ and $e\left(\cdot, \cdot, \gamma\right)$ are known functions; Ridder and Moffitt (2007, Section 2.4) provide a number of specific examples. 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.[7]

A total of N_1 units are drawn from the study population and their realizations of (Y_1, X) recorded; this forms the study sample. An auxiliary random sample of size N_0 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 Current Population Survey (a stratified random sample) and information on (Y_1, X) from a random sub-sample of the Census. In this example the population of interest is school-aged children living in the United States. The Census sub-sample can be viewed as a pure random sample from this population.

Available results for TSIV require that the two samples be random samples from the same population (Angrist and Krueger 1992, Arellano and Meghir 1992, Ridder and Moffitt 2007). Unfortunately, in

[6] Ridder and Moffitt (2007) provide a recent survey on research on data combination.

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

many empirical applications estimated moments from the two samples for the common variables, X, differ significantly. Currie and Yelowitz (2000), for example, find significant differences in the probability of a household being female-headed and ethnic classification. As we show below IPT can be used to combine datasets that are not random draws from the same population.

M-estimation when the target and sampled populations differ (Little and Wu 1991, Hellerstein and Imbens 1999, Tarozzi 2005) In this example the auxiliary sample is not one for which Assumption 3.2 holds. Under certain restrictions, however, the study samples provide sufficient information to identify the study distribution function. Conventional M-estimation proceeds using an estimate of the study distribution function in place of the empirical distribution function.

Hellerstein and Imbens (1999) estimate an Mincerian log-earnings regression with an ability measure. Their auxiliary sample, the NLS, is not a random sample from their study population (non-self-employed working white men between the ages of 28 and 38 in 1980). In their example earnings and basic demographic controls are included in X, Y_0 is an IQ score and Y_1 is empty. The NLS includes information on X and Y_0, while a sample drawn from the 1980 Census identifies the marginal distribution of X in the study population. The moment vector is as given in Assumption 3.2 with $m(y_1, x, \gamma)$ equalling a vector of zeros.

In related work Little and Wu (1991) study the calibration of contingency tables to known margins when the target population differs from the sampled population. Tarozzi (2005) is interested in computing Indian poverty rates in a consistent way over time using samples with changing expenditure questionnaires.

Average Treatment Effects on the Treated (ATT) (Hahn 1998, Imbens 2004) LaLonde (1986) and Dehejia and Wahba (1999) combine earnings outcomes from participants in the National Supported Work (NSW) Demonstration with earnings from nonexperimental comparison samples drawn from either the Panel Study of Income Dynamics (PSID) or the CPS to estimate the ATT associated with the NSW Demonstration. Their target population is NSW participants. The study sample, which consists of the PSID (or CPS) controls, contains information on Y_0, earnings given nonparticipation, and X, pre-intervention controls. Their auxiliary sample consists of a random sample of NSW participants and contains information on Y_1, earnings given participation, and X.

The object of interest is

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

which is equivalent to γ_0 defined by Assumption 3.2 with $m(y_1, x, \gamma) = y_1$ and $l(y_0, x, \gamma) = -y_0 - \gamma$.

Nonclassical measurement error with validation samples (Chen, Hong and Tamer 2005, Chen, Hong and Tarozzi 2005) In this example the auxiliary sample is a validation sample consisting of Y_0, a measurement of, say, income free of error, as well as other variables X. Included in X is a noisy measure of Y_0 or other surrogates. The study sample is a random sample from the population of interest and consists of records of X (Y_1 is left empty); γ_0 is defined by Assumption 3.2 with $m(y_1, x, \gamma)$ equal to a vector of zeros.

Small error estimation (Elbers, Lanjouw and Lanjouw 2003) An auxiliary sample consists of a measurements of Y_0, say household consumption, and demographic controls, X from a country-wide household survey. Precise information on the distribution of X in a small area of interest, say a specific municipality, is provided by a study sample (typically extracted from a Census). The vector Y_1 is empty. Interest centers are estimation of

$$\gamma_0 = \mathbb{E}\left[Y_0 | D = 1\right],$$

where expectations are taken with respect to the population in the municipality of interest. This estimand is equivalent to γ_0 defined by Assumption 3.2 with $m\left(y_1, x, \gamma\right)$ set equal to zero and $l\left(y_0, x, \gamma\right) = y_0 - \gamma$.

5 Inverse probability tilting

In this section we show how inverse probability tilting (IPT) can be used to consistently estimate γ_0 as defined by either Assumption 3.1 or 3.2 under the sampling scheme described above and the additional restrictions:

Assumption 5.1 (Random Sampling) *$\{Y_1, X\}_{i=1}^{N_1}$ is an independently and identically distributed random sequence from the study population and $\{Y_0, X\}_{i=1}^{N_0}$ is an independently and identically distributed random sequence from the auxiliary population.*

Assumption 5.2 (Relative Sample Size Growth) *As $N_1 \to \infty$ and $N_0 \to \infty$, $N_1/N_0 \to Q_0/\left(1 - Q_0\right)$.*

Assumption 5.3 (Missing at Random) $\left(Y_1, Y_0\right) \perp D | X$

Assumption 5.4 (Overlap) *Let $p_0\left(x\right) = \Pr\left(D = 1 | X = x\right)$, then $0 < \kappa < p_0\left(x\right) < 1 - \kappa < 1$ for all $x \in \mathcal{X} \subset \mathbb{R}^{\dim(X)}$.*

Assumption 5.5 (Parametric Propensity Score) *Let $G\left(w\left(X\right)'\alpha\right)$ be a parametric model for $p\left(x\right)$ with (i) $w\left(x\right) = \left(1, h^*\left(x\right)'\right)'$, $h^*\left(x\right)$ a $M \times 1$ vector of known functions of X, $\zeta_0 = \mathbb{E}\left[h^*\left(X\right)\right]$ or $\zeta_0 = \mathbb{E}\left[h^*\left(X\right) | D = 0\right]$, $h\left(x, \zeta_0\right) = h^*\left(x\right) - \zeta_0$ and $\mathbb{E}\left[h\left(X, \zeta_0\right) h\left(X, \zeta_0\right)' | D = 1\right]$ or $\mathbb{E}\left[h\left(X, \zeta_0\right) h\left(X, \zeta_0\right)' | D = 0\right]$ of full rank, (ii) a unique $\alpha_0 \in \mathcal{A} \subset \mathbb{R}^{1+M}$ such that $p_0\left(x\right) = G(w\left(x\right)'\alpha_0)$, and (iii) $G\left(\cdot\right)$ a continuous, differentiable and increasing function with*

$$\lim_{v \to -\infty} G\left(v\right) = 0, \qquad \lim_{v \to +\infty} G\left(v\right) = 1.$$

Assumptions 5.1 and 5.2 formalize the sampling scheme outlined in Section 3. Assumptions 5.3 and 5.4 ensure identification of γ_0. Assumption 5.3 implies the distribution of (Y_1, Y_0) conditional on X is the same in the study and auxiliary populations. This restriction and the common support restriction implied by Assumption 5.4 identify γ_0.

Consider γ_0 as defined by Assumption 3.1. We have

$$\begin{aligned}
\mathbb{E}\left[\psi\left(Z, \gamma\right)\right] &= \mathbb{E}\left[m\left(Y_1, X, \gamma\right)\right] + \mathbb{E}\left[l\left(Y_0, X, \gamma\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[m\left(Y_1, X, \gamma\right) | X\right]\right] + \mathbb{E}\left[\mathbb{E}\left[l\left(Y_0, X, \gamma\right) | X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[m\left(Y_1, X, \gamma\right) | X, D = 1\right]\right] + \mathbb{E}\left[\mathbb{E}\left[l\left(Y_0, X, \gamma\right) | X, D = 0\right]\right],
\end{aligned}$$

where line two follows by iterated expectations and line three by Assumption 5.3. The study and auxiliary samples identify $\mathbb{E}\left[m\left(Y_1, X, \gamma\right)|X, D=1\right]$ and $\mathbb{E}\left[l\left(Y_0, X, \gamma\right)|X, D=0\right]$. By Assumption 5.4 they do so at all points of support in the target distribution of X and hence $\mathbb{E}\left[\psi\left(Z, \gamma\right)\right]$ is identified. Identification of γ_0 as defined by Assumption 3.2 follows from a similar argument.

While Assumptions 5.1 to 5.4 ensure point identification of γ_0 we additionally impose Assumption 5.5 for convenience. To understand this restriction note that it implies, by Baye's Law and Assumptions 5.3 and 5.4 that

$$\begin{aligned} G^{-1}\left(\frac{Q_0 f\left(z|d=1\right)}{f\left(z\right)}\right) &= w\left(x\right)'\alpha_0 \\ &= \alpha_{10} + h^*\left(x\right)'\alpha_{20}, \end{aligned} \tag{9}$$

where $G^{-1}\left(\cdot\right)$ is the inverse of $G\left(\cdot\right)$ and we partition $\alpha_0 = \left(\alpha_{10}, \alpha_{20}'\right)'$. Equation (9) shows that Assumptions 5.3 to 5.5 restrict some known transformation of the ratio of the density of Z in the study and target populations to be linear in $h^*\left(x\right)$. Such restrictions are similar to those made in discriminant analysis. To make this connection explicit note that if $G\left(t\right) = \exp\left(t\right)/\left[1+\exp\left(t\right)\right]$ is the logistic function we can, using Assumption 5.3, show, after some manipulation, that (9) is equivalent to the restriction

$$\ln\left[\frac{f\left(x|d=1\right)}{f\left(x|d=0\right)}\right] = \alpha_{10}^* + h^*\left(x\right)'\alpha_{20}, \tag{10}$$

with $\alpha_{10}^* = \alpha_{10} + \ln\left(1-Q_0\right) - \ln\left(Q_0\right)$.[8]

It is well-known that (10) holds for $h^*\left(x\right)$ including x, x^2 and all cross-product terms when f$(x|d=1)$ and $f\left(x|d=0\right)$ are multivariate normal (with different mean vectors and covariance matrices). When x is discrete and $h^*\left(x\right)$ is a vector of dummy variables for each support point, then (10) obviously holds. Restriction (10) can also accommodate joint distributions of continuous and discrete random variables with dependence between the continuous and discrete random variables (Anderson 1982). Barron and Sheu (1991) show that for X a continuous random variable with bounded support, the log density ratio can be well-approximated by allowing the dimension of $h^*\left(x\right)$ to grow with the sample size. We focus on 'flexible parametric' specifications in this paper.[9]

[8]This follows from:

$$\begin{aligned} G^{-1}\left(\frac{Q_0 f\left(z|d=1\right)}{f\left(z\right)}\right) &= G^{-1}\left(\frac{Q_0 f\left(y_0, y_1|x, d=1\right) f\left(x|d=1\right)}{f\left(y_0, y_1|x\right) f\left(x\right)}\right) \\ &= G^{-1}\left(\frac{Q_0 f\left(x|d=1\right)}{Q_0 f\left(x|d=1\right) + \left(1-Q_0\right) f\left(x|d=0\right)}\right) \\ &= \ln\left(\frac{f\left(x|d=1\right)}{f\left(x|d=0\right)}\right) - \ln\left(\frac{1-Q_0}{Q_0}\right) \\ &= w\left(x\right)'\alpha_0 \end{aligned}$$

[9]For modest numbers of covariates an feasible implementation of IPT will necessarily be of the flexible parametric variety. When $f\left(x|d=1\right)$ and $f\left(x|d=1\right)$ are multivariate normal densities satisfying (10) requires $h^*\left(x\right)$ to be of dimension $2p + \frac{1}{2}p\left(p-1\right)$, with $p = \dim\left(X\right)$. Dehejia and Wahba's (1999) evaluation of the NSW Demonstration includes 8 covariates and hence $2p + \frac{1}{2}p\left(p-1\right) = 44$.

5.1 IPT as a calibration estimator

The inverse probability tilting estimation of γ_0 can be informally derived by calibration arguments (e.g., Little and Wu 1991).

5.1.1 IPT estimation of γ_0 defined by Assumption 3.1

By Bayes Law and Assumptions 5.3, 5.4, and 5.5 we have

$$\int \psi(z,\gamma)\,\mathrm{d}F(z) = \int m(y_1,x,\gamma)\,\frac{Q_0}{G\left(w(x)'\alpha_0\right)}\mathrm{d}F(z|d=1) + \int l(y_0,x,\gamma)\,\frac{1-Q_0}{1-G\left(w(x)'\alpha_0\right)}\mathrm{d}F(z|d=1).$$

Observe that the target density is a 'inverse probability tilt' of the study and auxiliary densities. The method of inverse probability tilting is built on this insight. The procedure has two steps. First, the empirical measures of the study and auxiliary samples are inverse probability tilted toward a multinomial approximation of target measure. Second γ_0 is estimated by method-of-moments with the study and auxiliary empirical measures replaced with their inverse probability tilts.

Replacing $f(z|d=1)$ with the empirical measure of the study sample and assuming $f(z)$ to be multinomial with coinciding support gives

$$G^{-1}\left(\frac{Q_0}{N_1\pi_i}\right) = w(X_i)'\alpha_0, \quad i=1,\ldots,N_1,$$

where $\pi_1,\ldots,\pi_{N_1}$ are the multinomial probabilities characterizing F.

Let

$$\widehat{Q} = \frac{1}{N}\sum\nolimits_{i=1}^{N} D_i, \qquad \widehat{\zeta} = \frac{1}{N}\sum\nolimits_{i=1}^{N} h^*(X_i), \tag{11}$$

be analog estimates of $\rho_0 = \left(Q_0, \zeta_0'\right)'$. The inverse probability tilt of the study sample is given by

$$\widehat{\pi}_i(\widehat{\rho}) = \frac{\widehat{Q}}{N_1}\frac{1}{G(w(X_i)'\,\widehat{\alpha}^S)}, \quad i=1,\ldots,N_1, \tag{12}$$

with $\widehat{\alpha}^S$ uniquely determined by imposing the adding-up constraint, $\sum_{i=1}^{N_1}\widehat{\pi}_i = 1$, and the requirement that $h(X,\widehat{\zeta})$ be mean zero in the target population, $\sum_{i=1}^{N_1}\widehat{\pi}_i h(X,\widehat{\zeta}) = 0$:

$$\frac{\widehat{Q}}{N_1}\sum_{i=1}^{N_1}\frac{1}{G(w(X_i)'\,\widehat{\alpha}^S)} - 1 = 0 \tag{13}$$

$$\frac{\widehat{Q}}{N_1}\sum_{i=1}^{N_1}\frac{h(X_i,\widehat{\zeta})}{G(w(X_i)'\,\widehat{\alpha}^S)} = 0. \tag{14}$$

We also have

$$G^{-1}\left(\frac{1-Q_0}{N_1\pi_i}\right) = w(X_i)'\alpha_0, \quad i=N_1+1,\ldots,N,$$

and hence the inverse probability tilt of the auxiliary sample is given by

$$\widehat{\pi}_i(\widehat{\rho}) = \frac{1-\widehat{Q}}{N_0}\frac{1}{1-G(w\left(X_i\right)'\widehat{\alpha}^A)}, \quad i = N_1+1,\ldots,N,$$

with

$$\begin{aligned}\frac{1-\widehat{Q}}{N_0}\sum_{i=N_1+1}^{N}\frac{1}{1-G(w\left(X_i\right)'\widehat{\alpha}^A)} - 1 &= 0\\ \frac{1-\widehat{Q}}{N_0}\sum_{i=N_1+1}^{N}\frac{h(X_i,\widehat{\zeta})}{1-G(w\left(X_i\right)'\widehat{\alpha}^A)} &= 0.\end{aligned}$$

The 'S' and 'A' superscripts denote which sample was used to estimate the propensity score coefficients.

The inverse probability tilting estimate of γ_0 in then given by the solution to

$$\begin{aligned}0 &= \sum_{i=1}^{N_1}\widehat{\pi}_i m\left(Y_{1i},X_i,\widehat{\gamma}\right) + \sum_{i=N_1+1}^{N}\widehat{\pi}_i l\left(Y_{0i},X_i,\widehat{\gamma}\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{D_i m\left(Y_{1i},X_i,\widehat{\gamma}\right)}{G(w\left(X_i\right)'\widehat{\alpha}^S)} + \frac{\left(1-D_i\right)l\left(Y_{0i},X_i,\widehat{\gamma}\right)}{1-G(w\left(X_i\right)'\widehat{\alpha}^A)}\right\},\end{aligned}$$

where for simplicity we have assumed that $\dim\left(\psi\left(z,\gamma\right)\right) = \dim\left(\gamma\right)$.

Some important properties to discuss:

1. Like inverse probability weighting (Hirano, Imbens and Ridder 2003, Wooldridge 2007), but $\widehat{\alpha}^S$ and $\widehat{\alpha}^A$ not estimated by Sieve/Parametric MLE.
2. While $\widehat{\alpha}^S$ and $\widehat{\alpha}^A$ are both consistent for α_0, they are numerically different.
3. $\sum_{i=1}^{N} D_i\widehat{\pi}_i$ and $\sum_{i=1}^{N}\left(1-D_i\right)\widehat{\pi}_i$ sum to one by construction – no ex post Hájek normalization required.
4. $\sum_{i=1}^{N} D_i\widehat{\pi}_i h^*(X_i) = \sum_{i=1}^{N}\left(1-D_i\right)\widehat{\pi}_i h^*(X_i) = \sum_{i=1}^{N} h^*(X_i)/N$ – the inverse probability tilted moments in the study and auxiliary samples exactly match the sample moments of $h^*(X_i)$ in the merged sample. IPT imposes 'balance' by construction.
5. For IPW these last two properties are asymptotic only.

5.1.2 IPT estimation of γ_0 defined by Assumption 3.2

By Bayes Law and Assumptions 5.3, 5.4, and 5.5 we have

$$\begin{aligned}&\int\psi\left(z,\gamma\right)\mathrm{d}F\left(z|d=1\right)\\ &\qquad = \int m\left(y_1,x,\gamma\right)\mathrm{d}F\left(z|d=1\right) + \int l\left(y_0,x,\gamma\right)\frac{1-Q_0}{Q_0}\frac{G\left(w\left(x\right)'\alpha_0\right)}{1-G\left(w\left(x\right)'\alpha_0\right)}\mathrm{d}F\left(z|d=0\right).\end{aligned}$$

To estimate γ_0 defined by Assumption 3.2 we tilt the auxiliary sample toward a multinomial approximation of the study population measure.

Manipulating (8) and using Bayes Law and Assumptions 5.3, 5.4, and 5.5 gives

$$\frac{f\left(z|d=1\right)}{f\left(z|d=0\right)}=\frac{1-Q_0}{Q_0}\frac{G\left(w\left(x\right)'\alpha_0\right)}{1-G\left(w\left(x\right)'\alpha_0\right)}$$

and hence

$$G^{-1}\left(\frac{Q_0}{Q_0+\left(1-Q_0\right)\frac{f(z|d=0)}{f(z|d=1)}}\right)=w\left(x\right)'\alpha_0$$

Replacing $f\left(z|d=0\right)$ with the empirical measure of the auxiliary sample and assuming $f\left(z|d=1\right)$ to be multinomial with coinciding support gives

$$G^{-1}\left(\frac{N_0\pi_i}{N_0\pi_i+\left(1-Q_0\right)/Q_0}\right)=w\left(X_i\right)'\alpha_0,\quad i=N_1+1,\ldots,N,$$

where $\pi_1,\ldots,\pi_{N_0}$ are the multinomial probabilities characterizing $F_{Z|D=1}$.

Let

$$\widehat{Q}=\frac{1}{N}\sum\nolimits_{i=1}^{N}D_i,\qquad \widehat{\zeta}=\frac{1}{N_1}\sum\nolimits_{i=1}^{N_1}h^*\left(X_i\right), \tag{15}$$

be analog estimates of $\rho_0=\left(Q_0,\zeta_0'\right)'$. Note that $\widehat{\zeta}$ is the mean of $h^*\left(X_i\right)$ in the study sample, not the merged sample as before. The inverse probability tilt of the auxiliary sample is given by

$$\widehat{\pi}_i(\widehat{\rho})=\frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\frac{G(w\left(X_i\right)'\widehat{\alpha}^A)}{1-G(w\left(X_i\right)'\widehat{\alpha}^A)},\quad i=N_1+1,\ldots,N,$$

with $\widehat{\alpha}$ uniquely determined by imposing the adding-up constraint, $\sum_{i=N_1+1}^{N}\widehat{\pi}_i=1$, and the requirement that $h(X,\widehat{\zeta})$ be mean zero in the study population, $\sum_{i=N_1+1}^{N}\widehat{\pi}_i h(X,\widehat{\zeta})=0$:

$$\begin{aligned}\frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\sum_{i=N_1+1}^{N}\frac{G(w\left(X_i\right)'\widehat{\alpha}^A)}{1-G(w\left(X_i\right)'\widehat{\alpha}^A)}-1 &= 0\\ \frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\sum_{i=N_1+1}^{N}\frac{G(w\left(X_i\right)'\widehat{\alpha}^A)}{1-G(w\left(X_i\right)'\widehat{\alpha}^A)}h(X,\widehat{\zeta}) &= 0.\end{aligned}$$

The IPT estimate of γ_0 is given by the solution to

$$\begin{aligned}0 &= \frac{1}{N_1}\sum_{i=1}^{N_1}m\left(Y_{1i},X_i,\widehat{\gamma}\right)+\sum_{i=N_1+1}^{N}\widehat{\pi}_i l\left(Y_{0i},X_i,\widehat{\gamma}\right)\\ &= \frac{1}{N_1}\sum_{i=1}^{N_1}m\left(Y_{1i},X_i,\widehat{\gamma}\right)+\frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\sum_{i=N_1+1}^{N}\frac{G(w\left(X_i\right)'\widehat{\alpha}^A)}{1-G(w\left(X_i\right)'\widehat{\alpha}^A)}l\left(Y_{0i},X_i,\widehat{\gamma}\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{G(w\left(X_i\right)'\widehat{\alpha}^A)}{\widehat{Q}}\left\{\frac{D_i m\left(Y_{1i},X_i,\widehat{\gamma}\right)}{G(w\left(X_i\right)'\widehat{\alpha}^A)}+\frac{\left(1-D_i\right)l\left(Y_{0i},X_i,\widehat{\gamma}\right)}{1-G(w\left(X_i\right)'\widehat{\alpha}^A)}\right\}.\end{aligned}$$

Some important properties to discuss:

1. Like inverse probability weighting (Hirano, Imbens and Ridder 2003), but $\widehat{\alpha}^A$ not estimated by Sieve/Parametric MLE.

2. $\sum_{i=1}^{N}(1-D_i)\widehat{\pi}_i$ sums to one and $\sum_{i=1}^{N}(1-D_i)\widehat{\pi}_i h^*(X_i) = \sum_{i=1}^{N_1} h^*(X_i)/N_1$ – the inverse probability tilted moments in the auxiliary samples exactly match the sample moments of $h^*(X_i)$ in the study sample. IPT imposes 'balance' by construction.

5.2 IPT as a minimum empirical discrepancy estimator

For the rest of the paper we assume that, when γ_0 is as defined by Assumption 3.1, that

$$\psi(z,\gamma) = m(y_1, x, \gamma),$$

and hence only the study sample need be tilted. Similarly, when γ_0 is as defined by Assumption 3.2, we assume that

$$\psi(z,\gamma) = l(y_0, x, \gamma).$$

This simplifies exposition.

Consider the function

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

which measures the divergence between the probability measures P and Q. The contrast function $\varphi(x)$ is convex, differentiable on its domain, and chosen such that $D(\cdot, Q)$ is minimized at Q (cf., Bickel, Klassen, Ritov and Wellner 1993, Chapter 7, Kitamura 2006). For what follows it will be useful to establish the notation $f_j(x) \overset{def}{\equiv} \partial^j f(x)/\partial x^j$ for any function $f(\cdot)$.

Inverse probability tilting estimates of γ_0 can be viewed as minimum empirical discrepancy estimators.

5.2.1 Minimum empirical discrepancy representation and duality for estimation under Assumption 3.1

The inverse probability tilt for the empirical measure of the study sample is given by the solution to

$$\min_{\pi_1,\ldots,\pi_{N_1}} \quad \frac{1}{N_1}\sum_{i=1}^{N_1}\varphi(N_1\pi_i; Q_0), \quad s.t. \quad \sum_{i=1}^{N_1}\pi_i = 1, \quad \sum_{i=1}^{N_1}\pi_i h(Z_i, \zeta_0) = 0, \tag{16}$$

with $\rho_0 = \left(Q_0, \zeta_0'\right)'$ replaced by the consistent estimates given in Equation (11) above and

$$\varphi(x;Q) \overset{def}{\equiv} \begin{cases} -\frac{x}{k(Q)}G^{-1}(Q) - \frac{1}{k(Q)}\int_x^a G^{-1}\left(\frac{Q}{t}\right)\mathrm{d}t & x > Q \\ +\infty & x \leq Q \end{cases} \tag{17}$$

for $Q \in (0,1)$ and $k(Q) \stackrel{def}{\equiv} -Q/G_1\left(G^{-1}(Q)\right)$. Observe that (17) is convex, differentiable and attains its minimum at $x = 1$.[10]

Note that since $\varphi(x; Q_0)$ is only finite for $x > Q_0$, it constrains the estimated probabilities to be greater than Q_0/N_1. This restriction is implied by the structure of the problem since

$$\frac{f(z)}{f(z|d=1)} = \frac{Q_0}{p_0(x)} > Q_0.$$

Replacing $F(z)$ with its the multinomial approximation and $F(z|d=1)$ with the empirical measure of the study sample gives $\pi_i > Q_0/N_1$.

Let $\mathcal{L}\left(\pi_1, \ldots \pi_{N_1}, \eta^S, \lambda^S; \rho_0\right)$ be the Lagrangian associated with (16), where η^S is the scalar multiplier associated with the adding-up constraint and λ^S the $M \times 1$ vector of multipliers associated with requirement that $h(Z, \zeta_0)$ be mean zero. The N_1 first order conditions for the probabilities are

$$-\frac{G^{-1}(\widehat{Q})}{k(\widehat{Q})} + \frac{1}{k(\widehat{Q})} G^{-1}\left(\frac{\widehat{Q}}{N_1 \widehat{\pi}_i}\right) - \widehat{\eta}^S - h(Z, \widehat{\zeta})' \widehat{\lambda}^S = 0, \qquad i = 1, \ldots, N_1.$$

Solving for $\widehat{\pi}_i$ gives

$$\widehat{\pi}_i(\widehat{\rho}) = \frac{\widehat{Q}}{N_1} \frac{1}{G(k(\widehat{Q}) t(Z, \widehat{\zeta})' \widehat{\delta}^S + G^{-1}(\widehat{Q}))}, \qquad i = 1, \ldots, N_1, \tag{18}$$

where

$$t(Z, \zeta) = \begin{pmatrix} 1 \\ h(Z, \zeta) \end{pmatrix}, \qquad \delta^S = \begin{pmatrix} \eta^S \\ \lambda^S \end{pmatrix}.$$

Note the identities

$$\alpha_{10} = k(Q_0)\eta_0 - \zeta_0' \lambda_0 + G^{-1}(Q_0), \quad \alpha_{20} = k(Q_0)\lambda_0.$$

The first order condition for $\widehat{\delta}^S$, after substituting in (18), is

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

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

The negative of the Fenchel conjugate of $\varphi(x, Q)$ is

$$\varphi^+(x, Q) \stackrel{def}{\equiv} -\max_y \left\{ xy + \frac{y}{k(Q)} G^{-1}(Q) + \frac{1}{k(Q)} \int_y^a G^{-1}\left(\frac{Q}{t}\right) \mathrm{d}t \right\}.$$

[10] These claims follow from the first order condition

$$\varphi_1(x^*; Q) = -\frac{1}{k(Q)} \left[G^{-1}(Q) - G^{-1}\left(\frac{Q}{x^*}\right) \right] = 0$$

which implies an extreme point at $x^* = 1$. That this point is a global minimum follows from convexity of $\varphi(x; Q)$, i.e.,

$$\varphi_2(x; Q) = -\frac{1}{k(Q)} \frac{1}{G_1\left(G^{-1}\left(\frac{Q}{x}\right)\right)} \frac{Q}{x^2} > 0.$$

The first order condition for this problem is given by

$$x+\frac{G^{-1}(Q)}{k(Q)}-\frac{G^{-1}(Q/y^*)}{k(Q)}=0,\Rightarrow y^*=\frac{Q}{G(k(Q)x+G^{-1}(Q))},$$

so we have[11]

$$\varphi^+(x,Q)\overset{def}{\equiv}-\frac{1}{k(Q)}\left[\frac{k(Q)x+G^{-1}(Q)}{G(k(Q)x+G^{-1}(Q))}Q+\int_{Q/G(k(Q)x+G^{-1}(Q)}^{a}G^{-1}\left(\frac{Q}{t}\right)\mathrm{d}t\right]. \tag{19}$$

The Fenchel duality theorem (Rockafellar 1970, Borwein and Lewis 1991) implies that $\widehat{\delta}^S$ is also the solution to

$$\max_{\delta}\left\{t_0'\delta+\frac{1}{N_1}\sum_{i=1}^{N_1}\varphi^+(t(Z,\widehat{\zeta})'\delta,\widehat{Q})\right\}. \tag{20}$$

The dual problem simplifies computation and asymptotic analysis. The probabilities can be recovered by

$$\widehat{\pi}_i(\widehat{\rho})=-\frac{\varphi_1^+(t(Z,\widehat{\zeta})'\widehat{\delta}^S,\widehat{Q})}{N_1},\qquad i=1,\ldots,N_1.$$

5.2.2 Minimum empirical discrepancy representation and duality for estimation under Assumption 3.2

To do for this section:

1. Normalize the discrepancy function.

For estimation of γ_0 as defined by Assumption 3.2 the appropriate contrast function[12] is

$$\varphi(x;Q)\overset{def}{\equiv}\begin{cases}-xG^{-1}(Q)-\int_x^a G^{-1}\left(\frac{t}{t+(1-Q)/Q}\right)\mathrm{d}t & x>0\\ +\infty & x\leq 0\end{cases}. \tag{21}$$

[11] This follows since

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

[12]

$$\begin{aligned}\varphi_1(x;Q) &= -G^{-1}(Q)+G^{-1}\left(\frac{x}{x+(1-Q)/Q}\right)\\ \varphi_2(x;Q) &= \frac{1}{G_1\left(G^{-1}\left(\frac{x}{x+(1-Q)/Q}\right)\right)}\left[\frac{1}{x+(1-Q)/Q}-\frac{x}{(x+(1-Q)/Q)^2}\right]\\ &= \frac{1}{G_1\left(G^{-1}\left(\frac{x}{x+(1-Q)/Q}\right)\right)}\left[\frac{(1-Q)/Q}{(x+(1-Q)/Q)^2}\right]>0\end{aligned}$$

The inverse probability tilt for the empirical measure of the auxiliary sample is given by the solution to

$$\min_{\pi_1,\ldots,\pi_{N_0}} \quad \frac{1}{N_0}\sum_{i=N_1+1}^{N}\varphi\left(N_1\pi_i;Q_0\right), \quad s.t. \quad \sum_{i=1}^{N_1}\pi_i=1, \quad \sum_{i=1}^{N_1}\pi_i h\left(Z_i,\zeta_0\right)=0, \tag{22}$$

with $\rho_0=\left(Q_0,\zeta_0'\right)'$ replaced by the consistent estimates given in Equation (15) above.

Let $\mathcal{L}\left(\pi_1,\ldots\pi_{N_0},\eta^A,\lambda^A;\rho_0\right)$ be the Lagrangian associated with (22). The first order conditions for the probabilities are

$$-G^{-1}(\widehat{Q})+G^{-1}\left(\frac{N_0\widehat{\pi}_i}{N_0\widehat{\pi}_i+(1-\widehat{Q})/\widehat{Q}}\right)-\widehat{\eta}^A-h(X,\widehat{\zeta})'\widehat{\lambda}^A=0, \qquad i=N_1+1,\ldots,N,$$

which gives

$$\widehat{\pi}_i(\widehat{\rho})=\frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\frac{G(\widehat{\eta}^A+h(X,\widehat{\zeta})'\widehat{\lambda}^A+G^{-1}(\widehat{Q}))}{1-G(\widehat{\eta}^A+h(X,\widehat{\zeta})'\widehat{\lambda}^A+G^{-1}(\widehat{Q}))}, \qquad i=N_1+1,\ldots,N,$$

where $\widehat{\delta}^A$ is the solution to

$$\frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\sum_{i=N_1+1}^{N}\frac{G(\widehat{\eta}^A+h(X,\widehat{\zeta})'\widehat{\lambda}^A+G^{-1}(\widehat{Q}))}{1-G(\widehat{\eta}^A+h(X,\widehat{\zeta})'\widehat{\lambda}^A+G^{-1}(\widehat{Q}))}t(Z,\widehat{\zeta})-t_0=0.$$

The Fechel conjugate is

$$\varphi^+(x,Q)\overset{def}{\equiv}-\max_y\left\{xy+yG^{-1}\left(Q\right)+\int_y^a G^{-1}\left(\frac{t}{t+(1-Q)/Q}\right)\mathrm{d}t\right\}.$$

The first order condition for this problem is given by

$$x+G^{-1}\left(Q\right)-G^{-1}\left(\frac{y^*}{y^*+(1-Q)/Q}\right)=0,\Rightarrow y^*=\frac{1-Q}{Q}\frac{G(x+G^{-1}\left(Q\right))}{1-G(x+G^{-1}\left(Q\right))},$$

and hence

$$\begin{aligned}\varphi^+(x,Q) &= -\left\{\begin{array}{c} x\frac{1-Q}{Q}\frac{G(x+G^{-1}(Q))}{1-G(x+G^{-1}(Q))}+G^{-1}\left(Q\right)\frac{1-Q}{Q}\frac{G(x+G^{-1}(Q))}{1-G(x+G^{-1}(Q))} \\ +\int_{\frac{1-Q}{Q}\frac{G(x+G^{-1}(Q))}{1-G(x+G^{-1}(Q))}}^{a}G^{-1}\left(\frac{t}{t+(1-Q)/Q}\right)\mathrm{d}t\end{array}\right\} \\ &= -\left\{\left[x+G^{-1}\left(Q\right)\right]\frac{1-Q}{Q}\frac{G(x+G^{-1}\left(Q\right))}{1-G(x+G^{-1}\left(Q\right))}+\int_{\frac{1-Q}{Q}\frac{G(x+G^{-1}(Q))}{1-G(x+G^{-1}(Q))}}^{a}G^{-1}\left(\frac{t}{t+(1-Q)/Q}\right)\mathrm{d}t\right\}.\end{aligned}$$

Inverse logistic tilting Little and Wu (1991) suggested calibration of contingency table cell frequencies to known marginals using (12), (13) and (14). In their example X consists of two binary variables with $h^*\left(X\right)=X$ and $G(t)$ equal to the logistic function

$$G(t)=\frac{e^t}{1+e^t}.$$

For this model $\varphi(x,Q)$ and $\varphi^{+}(x,Q)$ can be expressed in closed form. For the case where interest centers on γ_0 defined by Assumption 3.1 we have

$$\varphi(x,Q) \propto (x-Q)\ln(x-Q) - x\ln(1-Q) - (x-Q),$$

which is a normalized version of Nevo's (2002) 'generalized' exponential tilting distance metric. The Fenchel conjugate is given by

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

For the case where interest centers on γ_0 defined by Assumption 3.2 we have

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

which yields the exponential tilting discrepancy with a Fenchel conjugate of

$$\varphi^{+}(x,Q) = -\exp(x),$$

which is the associated GEL criterion function.

Inverse linear probability tilting

$$G(t) = \begin{cases} 0 & t < -1/2 \\ \frac{1}{2} + v & -1/2 \leq t \leq 1/2 \\ 1 & t > -1/2 \end{cases}$$

For Assumption 3.1 we get

$$\begin{aligned} \varphi(x,Q) &\propto x - \ln x \\ \varphi^{+}(x,Q) &\propto \ln(1-x). \end{aligned}$$

For Assumption 3.2 we get....

Discuss connection to Hellerstein and Imbens (1999).

6 Large sample properties

6.1 Large sample results for estimation under Assumption 3.1

Recall that, for simplicity, we continue to work with the special case of $\psi(z,\gamma) = m(y_1,x,\gamma)$.

In this section we provide conditions for local efficiency and double robustness of inverse probability tilting estimates of γ_0 under Assumption 3.1. The maximal asymptotic precision with which γ_0 can be estimated under the MAR setup has been characterized by Robins, Rotnitzky and Zhao (1994). For future reference, the information bound for this problem is given by

$$\mathcal{I}(\gamma_0) = \mathbb{E}[\Gamma_0(X)]'\,\mathbb{E}[\Sigma_0(X)]^{-1}\mathbb{E}[\Gamma_0(X)] \tag{23}$$

where

$$\Gamma_0(x) = \mathbb{E}\left[\partial\psi(Z,\gamma_0)/\partial\gamma | X = x\right], \quad \Sigma_0(x) = Var\left(\psi(Z,\gamma_0)|X=x\right)/p_0(x) + q_0(x,\gamma_0)\,q_0(x,\gamma_0)'.$$

The corresponding efficient score is

$$s(Z,\gamma_0,p_0,q_0) = \frac{D}{p_0(X)}\psi(Z,\gamma_0) - \frac{q_0(X,\gamma_0)}{p_0(X)}\left(D - p_0(X)\right), \tag{24}$$

where $q_0(x,\gamma_0) = \mathbb{E}\left[\psi(Z,\gamma_0)|X=x,D\right] = \mathbb{E}\left[\psi(Z,\gamma_0)|X=x\right]$ (with the second equality following from Assumption 5.3).

Globally efficient estimates of γ_0 have been proposed by Hahn (1998), Hirano, Imbens and Ridder (2003), Chen, Hong and Tarozzi (2004) and Imbens, Newey and Ridder (2005). While global efficiency is a desirable large sample property, such estimators' asymptotic sampling distributions may only poorly approximate actual ones in moderated-sized samples due to their dependence on infinite-dimensional nuisance parameters. For this reason estimators which are (locally) efficient under auxiliary parametric assumptions, but (partially) robust to violations of these assumptions are attractive (Newey 1990, Robins, Rotnitzky and van der Laan 2000).

One candidate approach to constructing such estimators would be to directly incorporate parametric restrictions into estimators known to be globally efficient. Consider the imputation approach, as in Chen, Hong and Tarozzi (2004) and Imbens, Newey and Ridder (2005), which estimates γ_0 by the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{q}(X_i,\widehat{\gamma}) = 0,$$

where $\widehat{q}(x,\gamma)$ is a nonparametric estimate of $q_0(x,\gamma)$. If the conditional density of Y_1 given X belongs to a known parametric family, indexed by an unknown finite-dimensional parameter (i.e., $f_0(y_1|x) = f(y_1|x;\tau_0)$), then a parametric imputation estimator of γ_0 is given by the solution to

$$\frac{1}{N}\sum_{i=1}^{N} q(X_i,\widehat{\gamma};\widehat{\tau}) = 0,$$

where $\widehat{\tau}$ is the MLE of τ_0, as computed using the $D=1$ sub-sample, and $q(x,\gamma;\tau) = \int \psi(z,\gamma)\, f(y_1|x;\tau)\,\mathrm{d}x$.[13] This estimator, while locally efficient, is inconsistent under departures from the auxiliary parametric assumption. For that reason it is not particularly attractive (cf., Imbens 2004, pp. 12 - 13).

Next consider modifying the globally efficient inverse probability weighting (IPW) estimator of Hirano, Imbens and Ridder (2003). This estimate is given by the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i\psi(Z_i,\widehat{\gamma})}{\widehat{p}(X_i)} = 0,$$

with $\widehat{p}(x)$ a particular nonparametric estimate of the propensity score. Unfortunately, the IPW estima-

[13] Use of the $D=1$ subsample to estimate the conditional density of Y_1 given X is appropriate due to conditional independence of D and Y_1 given X.

tor, which replaces $p(x)$ with a parametric model for propensity score, $p(x;\upsilon_0)$, where υ_0 is an unknown finite-dimensional parameter, is typically inefficient and is generally inconsistent under departures from the parametric selection model (Chen, Hong and Tarozzi 2004, Graham 2007). While parametric IPW is easy to implement its intrinsic inefficiency and lack of robustness are considerable shortcomings.[14]

A final candidate approach to locally efficient estimation was proposed by Robins, Rotnitzky and Zhao (1994).[15] They suggested an augmented inverse probability weighting (AIPW) estimator where $\widehat{\gamma}$ is given by the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{p(X_i;\widehat{\upsilon})}\psi(Z_i,\widehat{\gamma})-\frac{q(X_i,\widehat{\gamma};\widehat{\tau})}{p(X_i;\widehat{\upsilon})}(D_i-p(X_i;\widehat{\upsilon}))=0,$$

with $\widehat{\tau}$ and $\widehat{\upsilon}$ being the first-step MLEs used to implement parametric imputation and IPW respectively. Robins, Rotnitzky and Zhao (1994) show local efficiency of AIPW, while Scharfstein, Rotnitzky and Robins (1999) show that it is 'doubly robust' in the sense that it remains consistent if either the parametric model for the conditional density of Y_1 given X or the propensity score is misspecified, but not both simultaneously. Graham (2007) provides a GMM-characterization of the doubly robust property and proposes a globally efficient AIPW estimator.

Bang and Robins (2005) argue that double robustness is a 'highly desirable property' (p. 962), giving researchers two chances to impose correct parametric restrictions. Local efficiency implies that, when the auxiliary parametric models are true, the AIPW estimator also exploits all the information about γ_0 contained in the data. While AIPW is relatively straightforward to compute, it does require calculating two auxiliary MLEs in addition to the parameter of interest.

We now show that IPT, like the AIPW estimator of Robins, Rotnitzky and Zhao (1994), is locally efficient and doubly robust. Unlike AIPW, the IPT estimator only requires computation of the parameters indexing the (parametric) propensity score. Those indexing the conditional density of Y_1 given X need not be estimated. We assume:

Assumption 6.1 (Parametric Heterogeneity) *Let* $\mathbb{E}^*[\psi(Z,\gamma_0)|h^*(X)]=\mathbb{E}[\psi(Z,\gamma_0)|X]$ *where*

$$\mathbb{E}^*[\psi(Z,\gamma_0)|h^*(X)]=\Upsilon_0 w(X),\qquad \Upsilon_0=\mathbb{E}\left[\psi(Z,\gamma_0)\,w(X)'\right]\mathbb{E}\left[w(X)\,w(X)'\right]^{-1}.$$

Assumption 5.5 specifies the presumed parametric family for the propensity score and provides conditions ensuring consistency of the (implicit) IPT estimates of α_0. Assumption 6.1 does not fully specify the conditional distribution of Y_1 given X but does restrict it in certain ways. When $\psi(Z,\gamma_0)=Y_1-\gamma_0$, a moment appropriate for estimation of the marginal mean of Y_1 (as required in, say, estimation of the ATE), Assumption 6.1 is equivalent to assuming that the conditional mean of Y_1 given X is linear in $h^*(X)$.

Recall that we estimate γ_0 by the solution to

$$\sum_{i=1}^{N_1}\widetilde{\pi}_i(\widehat{\rho})\psi(Z_i,\widehat{\gamma})=0,$$

[14] Wooldridge (2007, Section 4), in a very elegant paper, does provide some results on robustness of IPW to misspecification of the propensity score.

[15] See Graham (2007) for further discussion and additional references.

with $\widetilde{\pi}_i(\rho)$ as defined previously.

The following theorem summarizes the large sample properties of the IPT estimate.

Theorem 6.1 (Double Robustness and Local Efficiency of $\widehat{\gamma}_{IPT}$) *Suppose Assumption 3.1, Assumptions 5.1 to 5.4, additional regularity conditions, and at least one of Assumption 5.5 or 6.1 hold, then as* $N \to \infty$

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

and

$$\sqrt{N}(\widehat{\gamma}_{IPT} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, B_0^{-1} A_0 B_0^{-1}),$$

with A_0 *and* B_0 *defined in Appendix B. If both Assumption 5.5 or 6.1 hold, then*

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

with $\mathcal{I}(\gamma_0)$ *equalling the information bound given in (23).*

Proof. See Appendix B. ■

One intuition for the local efficiency result of Theorem 6.1 views it as an consequence of the exact balancing property of the IPT estimate of the target distribution function. That is, since $\sum_{i=1}^{N_1} \widetilde{\pi}_i(\widehat{\rho}) = 1$ and $\sum_{i=1}^{N_1} \widetilde{\pi}_i(\widehat{\rho}) h^*(X_i) = \sum_{i=1}^{N} h^*(X_i)/N$ we have, *exactly*,

$$\sum_{i=1}^{N_1} \widetilde{\pi}_i(\widehat{\rho}) \psi(Z_i, \gamma) = \sum_{i=1}^{N_1} \widetilde{\pi}_i(\widehat{\rho}) [\psi(Z_i, \gamma) - \Upsilon_0 w(X_i)] + \frac{1}{N} \sum_{i=1}^{N} \Upsilon_0 w(X_i).$$

As shown in Section 5 there is a one-to-one mapping between the IPT tilting parameter, δ, and the coefficients indexing the propensity score, α. Partitioning, we have, for $\delta = (\eta, \lambda')'$ and $\alpha = (\alpha_1, \alpha_2')'$

$$\eta_0 = k(Q_0)^{-1} \left(\alpha_{10} + \zeta_0' \lambda_0 - G^{-1}(Q_0)\right), \quad \lambda_0 = k(Q_0)^{-1} \alpha_{20}.$$

Using the definition of $\widetilde{\pi}_i(\widehat{\rho})$ and this one-to-one mapping we have

$$\sum_{i=1}^{N_1} \widetilde{\pi}_i(\widehat{\rho}) \psi(Z_i, \gamma) = \frac{1}{N_1} \sum_{i=1}^{N_1} \frac{\widehat{Q}}{G(w(X_i)'\widehat{\alpha})} [\psi(Z_i, \gamma) - \Upsilon_0 w(X_i)] + \frac{1}{N} \sum_{i=1}^{N} \Upsilon_0 w(X_i),$$

which after using the definition of $\widehat{Q}$ and Assumption 6.1 gives

$$\sum_{i=1}^{N_1} \widetilde{\pi}_i(\widehat{\rho}) \psi(Z_i, \gamma) = \frac{1}{N} \sum_{i=1}^{N} \frac{D_i \psi(Z_i, \gamma)}{G(w(X_i)'\widehat{\alpha})} - \frac{\Upsilon_0 w(X_i)}{G(w(X_i)'\widehat{\alpha})}(D_i - G(w(X_i)'\widehat{\alpha})),$$

which is of the form of the efficient score (24).[16]

Another, complementary, intuition for local efficiency follows from the GMM equivalence results of Graham (2007). Appendix B shows that the first order conditions for the sequential IPT estimator

[16] Observe that only the first term of the right-hand side of this expression varies with γ. The second term is always evaluated (implicitly) at γ_0. The IPT estimate is therefore, unlike the AIPW estimate, not exactly an M-estimate based on a parametric estimate of the efficient score vector. Nonetheless, as shown in Appendix B, the two estimators share a common influence function and hence are asymptotically equivalent.

described above are equivalent to the sample analogs of the unconditional population moments

$$\mathbb{E}\left[\left\{\frac{D}{G(w(X)'\alpha_0)}-1\right\}w(X)\right] = 0 \tag{25}$$

$$\mathbb{E}\left[\frac{D}{G(w(X)'\alpha_0)}\psi(Z,\gamma_0)\right] = 0. \tag{26}$$

We can therefore evaluate the efficiency properties of IPT using standard results on efficient GMM estimation. A particularly illuminating method for doing so in the present context is to exploit the well-known asymptotic equivalence between the joint GMM estimate of γ_0 based on (25) and (26) and the one based upon the population residual associated with the (mean squared error minimizing) linear predictor of (25) given (26) (e.g., Qian and Schmidt 1999).[17]

The linear predictor in question is given by

$$\begin{aligned}\mathbb{E}^*\left[\frac{D\psi(Z,\gamma_0)}{G(w(X)'\alpha_0)}\middle|\left\{\frac{D}{G(w(X)'\alpha_0)}-1\right\}w(X)\right]&\\ =\mathbb{E}\left[\frac{D\psi(Z,\gamma_0)}{G(w(X)'\alpha_0)}\left\{\frac{D}{G(w(X)'\alpha_0)}-1\right\}w(X)'\right]&\\ \times\mathbb{E}\left[\left\{\frac{D}{G(w(X)'\alpha_0)}-1\right\}^2 w(X)\,w(X)'\right]^{-1}\left\{\frac{D}{G(w(X)'\alpha_0)}-1\right\}w(X),&\end{aligned}$$

which, after using iterated expectations and Assumptions 5.3, 5.4, and 5.5, gives

$$\begin{aligned}\mathbb{E}\left[\left\{\frac{1}{G(w(X)'\alpha_0)}-1\right\}\mathbb{E}\left[\psi(Z,\gamma_0)\,|X\right]w(X)'\right]&\\ \times\mathbb{E}\left[\left\{\frac{1}{G(w(X)'\alpha_0)}-1\right\}w(X)\,w(X)'\right]^{-1}\left\{\frac{D}{G(w(X)'\alpha_0)}-1\right\}w(X),&\end{aligned}$$

which, after substituting for $\mathbb{E}\left[\psi(Z,\gamma_0)\,|X\right]$ using Assumption 6.1, factoring, and rearranging, finally gives

$$\frac{\Upsilon_0 w(X)}{G(w(X)'\alpha_0)}\left(D-G(w(X)'\alpha_0)\right).$$

Therefore IPT is asymptotically equivalent to an efficient GMM estimator based upon the 'residualized' moment

$$\mathbb{E}\left[\frac{D}{G(w(X)'\alpha_0)}\psi(Z,\gamma_0)-\frac{\Upsilon_0 w(X)}{G(w(X)'\alpha_0)}\left(D-G(w(X)'\alpha_0)\right)\right]=0,$$

which also equals the efficient score under the maintained auxiliary parametric Assumptions 5.5 and 6.1. Double robustness of IPT follows from the asymptotic equivalence to AIPW.

Several features of Theorem 6.1 merit emphasis. As is apparent from inspection of (25) and (26), IPT is equivalent to parametric IPW estimation where the propensity score is estimated in a very special way by method-of-moments. In particular, the first-step propensity score estimator is constructed to impose balance, that is to ensure that the IPW mean of $h^*(X)$ in the selected sample equals its unweighted mean in the full sample. The weights are also constructed to sum to one (no ex-post Hajek-type

[17] In the present case, such 'residualization' also implicitly accounts for the impact of estimation error in $\widehat{\alpha}$ on the asymptotic sampling distribution of $\widehat{\gamma}$.

normalization is required, cf., Imbens 2004, pp. 16 - 17). When Assumptions 5.5 and 6.1 hold IPT is locally efficient. Although IPT only imposes a finite subset of the unconditional moments implied by the missing-at-random assumption, the chosen subset is sufficient to both identify the propensity score and to exhaust the all information content of the MAR setup.

Hirano, Imbens and Ridder (2003) show global efficiency of IPW when $p(x)$ is estimated nonparametrically in a particular way. Wooldridge (2007) analyzes IPW when the selection probability is estimated parametrically by maximum likelihood. Wooldridge's approach is generally inefficient, while the Hirano, Imbens and Ridder (2003) approach provides no natural mechanism to simultaneously maintain semiparametric efficiency and incorporate prior knowledge about the degree of response heterogeneity. Inverse probability tilting is attractive relative to both approaches for these reasons. The efficiency result of Theorem 6.1, however, is local, not global like that of Hirano, Imbens and Ridder (2003). Inverse probability tilting can be made globally efficient by allowing the dimension of $h^*(X)$ to grow with the sample size at a particular rate. Flexible parametric modelling will likely be appropriate for many practical applications.

Another attractive feature of IPT with data missing at random is that achieving local efficiency, while necessarily requiring two auxiliary parametric assumptions, only involves estimating the nuisance parameters indexing the propensity score. The AIPW approach, in contrast, requires estimation of both those parameters indexing the propensity score as well as the conditional density of Y_1 given X. This difference, while leaving the first order asymptotics unchanged, may have consequences for finite sample performance which we investigate through a series of Monte Carlo experiments in Section 8 below.

6.2 Large sample results for estimation under Assumption 3.2

Recall that, for simplicity, we continue to work with the special case of $\psi(z,\gamma) = l(y_0, x, \gamma)$.

For future reference, the information bound for this problem is given by (cf., Chen, Hong and Tarozzi 2004)

$$\mathcal{I}(\gamma_0) = \mathbb{E}[\Gamma_0(X)|D=1]' \mathbb{E}[\Sigma_0(X)]^{-1} \mathbb{E}[\Gamma_0(X)|D=1] \tag{27}$$

where

$$\Gamma_0(x) = \mathbb{E}[\partial\psi(Z,\gamma_0)/\partial\gamma | X=x, D=1]$$
$$\Sigma_0(x) = \left[\frac{p_0(x)}{Q_0}\right]^2 \left\{\frac{Var(\psi(Z,\gamma_0)|X=x)}{1-p_0(x)} + q_0(x,\gamma_0)\, q_0(x,\gamma_0)'\right\}$$

The corresponding efficient score is

$$s(Z,\gamma_0,p_0,q_0) = (1-D)\frac{1}{Q_0}\frac{p_0(X)}{1-p_0(X)}\psi(Z,\gamma_0) - \frac{1}{Q_0}\frac{p_0(X)}{1-p_0(X)} q_0(X,\gamma_0)(D-p_0(X)).$$

Theorem 6.2 (Double Robustness and Local Efficiency of $\widehat{\gamma}_{IPT}$) *Suppose Assumption 3.2, Assumptions 5.1 to 5.4, additional regularity conditions, and at least one of Assumption 5.5 or 6.1 hold, then as* $N \to \infty$

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

and

$$\sqrt{N}(\widehat{\gamma}_{IPT} - \gamma_0) \xrightarrow{D} \mathcal{N}(0, B_0^{-1} A_0 B_0^{-1}),$$

with A_0 and B_0 defined in Appendix B. If both Assumption 5.5 or 6.1 hold, then

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

with $\mathcal{I}(\gamma_0)$ equalling the information bound given in (27).

Proof. See Appendix B. ■

7 Extensions

Non-ignorable attrition in panel data (Hirano, Imbens, Ridder and Rubin 2001, Nevo 2003)

Case-control studies with contaminated controls (Lancaster and Imbens 1996, Qin 1998)

8 Monte Carlo

Appendices

A Computation

NOTE: Some disjunct between this appendix and main text that needs to be fixed.

In this appendix we outline our algorithm for computing IPT point estimates of γ_0 and asymptotic standard errors for $\widehat{\gamma}$. Our algorithm builds on that of Graham (2005), itself an extension of work by Owen (2001), Mittelhammer, Judge and Schoenberg (2005) and Imbens (2002). We develop a modification of the saddle-point criterion function that deals with restricted domain issues without changing the solution. Our modification specializes to a proposal of Owen (2001) for the case of empirical likelihood, but is new for other GEL as well as IPT estimators. Finally we describe how to solve the (modified) saddle-point problem. When the sampled and target populations coincide we are able to apply Graham's (2005) hybrid Newton-Raphson/Fisher-Scoring algorithm to IPT estimation. All of our suggested algorithms are gradient-based. This appendix considers computation of $\widehat{\gamma}$ for the case where γ_0 is overidentified (i.e., $\dim(\gamma_0) = K < L = \dim(\psi(Z, \gamma_0))$). The results given for the exactly identified case in the main text are special cases of those given below.

The dual (saddle-point) problem is

$$\min_{\gamma \in \Gamma} \max_{\delta \in \widehat{\Delta}_{N_1}(\gamma)} l_{N_1}(\delta, \gamma; \rho_0), \tag{28}$$

where $l_{N_1}(\delta, \gamma; \rho_0)$ is the sample average of the saddle-point criterion function

$$l_{N_1}(\delta, \gamma; \rho_0) \overset{def}{\equiv} t_0'\delta + \frac{1}{N_1}\sum_{i=1}^{N_1} \varphi^+(t(Z_i, \gamma; \zeta_0)'\delta, Q_0), \tag{29}$$

with $\varphi^+(x, Q)$ given by () in the main text, $\delta = (\eta, \lambda', \xi')'$ the $1 + M + L$ Lagrange multiplier(s) on the three sets of constraints in the primal problem (**??**), $t_0 = (1, \underline{0}', \underline{0}')'$ and

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

The set $\widehat{\Delta}_{N_1}(\gamma)$ is defined below. It is also helpful to establish the conventions $t_i(\gamma; \zeta) = t(Z_i, \gamma; \zeta)$, $t_i(\gamma) = t(Z_i, \gamma; \zeta_0)$ and $t_i = t(Z_i, \gamma_0; \zeta_0)$. Let

$$\mathbb{L}_{N_1}(\gamma; \rho_0) \overset{def}{\equiv} \max_{\delta \in \widehat{\Delta}_{N_1}(\gamma)} l_{N_1}(\delta, \gamma; \rho_0)$$

be the concentrated saddle-point criterion function. Let $\widetilde{\delta}(\gamma;\rho_0)$ denote the concentrated value of the tilting parameter.

As noted in the main text, inspection of the minimum discrepancy problem (**??**) 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 $l_{N_1}(\delta,\gamma;\rho_0)$. In particular, excluding improper probability weights requires that the inner maximization (28) occurs over the set

$$\widehat{\Delta}_{N_1}(\gamma) \overset{def}{\equiv} \left\{\delta : Q_0/N_1 < -\varphi_1^+(t_i(\gamma;\zeta_0)'\widetilde{\delta}(\gamma;\rho_0),Q_0)/N_1 < 1, \qquad i=1,\ldots,N_1\right\}.$$

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 inner maximization problem in more detail. Differentiating $l_{N_1}(\delta,\gamma;\rho_0)$ with respect to δ gives a gradient vector of

$$\underset{1+M+L\times 1}{\nabla_\delta l_{N_1}(\delta,\gamma;\rho)} = t_0 + \frac{1}{N_1}\sum_{i=1}^{N_1}\varphi_1^+(t_i(\gamma;\zeta)'\delta,Q)t_i(\gamma;\zeta). \tag{30}$$

Note that (30) will have an 'exploding denominator' when $t_i(\gamma;\zeta)'\delta$ is large and positive. Differentiating again gives a Hessian matrix of

$$\underset{1+M+L\times 1+M+L}{\nabla_{\delta\delta} l_{N_1}(\delta,\gamma;\rho)} = \frac{1}{N_1}\sum_{i=1}^{N_1}\varphi_2^+(t_i(\gamma;\zeta)'\delta,Q)t_i(\gamma;\zeta)t_i(\gamma;\zeta)'. \tag{31}$$

For a fixed value of γ (and ρ) (31) is a negative semi-definite function of δ.[18] Solving for $\widetilde{\delta}(\gamma;\rho_0)$ therefore involves maximizing a concave function over a convex domain. This follows from the outer-product structure of (31) and strict concavity of $\varphi^+(x,Q)$.

In order to avoid maximization over a restricted domain we redefine $l_{N_1}(\delta,\gamma;\rho_0)$ so that it is concave in δ over $\mathbb{R}^{1+M+L}$ 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 $x_i = t_i(\gamma;\zeta)'\delta$, then the second inequality requires that

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

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

Our modified estimator replaces $\varphi^+(x,Q_0)$ in (29) with the hybrid function

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

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

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

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

Solving these equations yields

$$\begin{aligned} c_{N_1} &\overset{def}{\equiv} \varphi_2^+(x^*_{N_1},Q_0) \\ b_{N_1} &\overset{def}{\equiv} \varphi_1^+(x^*_{N_1},Q_0) - \varphi_2^+(x^*_{N_1},Q_0)x^*_{N_1} \\ a_{N_1} &\overset{def}{\equiv} \varphi^+(x^*_{N_1},Q_0) - \varphi_1^+(x^*_{N_1},Q_0)x^*_{N_1} + \frac{\varphi_2^+(x^*_{N_1},Q_0)}{2}(x^*_{N_1})^2. \end{aligned}$$

We then estimate γ_0 by solving

$$\min_{\gamma\in\Gamma}\mathbb{L}^\circ_{N_1}(\gamma;\rho_0),$$

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

where

$$\mathbb{L}_{N_1}^{\circ}(\gamma;\rho_0) \stackrel{def}{\equiv} \max_{\delta\in\mathbb{R}^{1+M+L}} l_{N_1}^{\circ}(\delta,\gamma;\rho_0),$$

is the concentrated modified saddle-point criterion function with

$$l_{N_1}^{\circ}(\delta,\gamma;\rho) \stackrel{def}{\equiv} t_0'\delta + \frac{1}{N_1}\sum_{i=1}^{N_1}\varphi^{\circ}(t_i(\gamma;\zeta)'\delta, Q). \tag{32}$$

Since $\mathbb{L}_{N_1}^{\circ}(\gamma;\rho_0)$ coincides $\mathbb{L}_{N_1}(\gamma;\rho_0)$ at valid solutions, solving the modified saddle-point 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 (32) in place of (29) avoids numerical problems caused by a 'exploding denominator' in (30).

Modified saddle-point criterion for empirical likelihood Calculating the modified saddle-point criterion function for empirical likelihood (EL) is relatively straightforward. We have $k(Q_0) = -Q$, $G^{-1}(x) = -1/2 + x$, $\varphi^{+}(x,Q_0) = \ln(1-x)$, $\varphi_1^{+}(x,Q_0) = -(1-x)^{-1}$ and $\varphi_2^{+}(x,Q_0) = -(1-x)^{-2}$, and therefore $x_{N_1}^{*} = 1 - 1/N_1$. 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,

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

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

Modified saddle-point criterion for inverse logistic tilting Calculating $\varphi^{\circ}(x,Q_0)$ for inverse logistic tilting (ILT) is also straightforward, albeit a bit more tedious than for EL. For this case we have $k(Q_0) = -(1-Q_0)^{-1}$, $G^{-1}(x) = \ln(x/(1-x))$, $\varphi^{+}(x,Q_0) = -xQ_0 + \frac{Q_0}{k(Q_0)}\exp\left[-k(Q_0)\,x - G^{-1}(Q_0)\right]$, $\varphi_1^{+}(x,Q_0) = -Q_0 - Q_0\exp\left[-k(Q_0)\,x - G^{-1}(Q_0)\right]$ and $\varphi_2^{+}(x,Q_0) = k(Q_0)\,Q_0\exp\left[-k(Q_0)\,x - G^{-1}(Q_0)\right]$, and therefore

$$x_{N_1}^{*} = (1-Q_0)\ln\left(\frac{N_1-Q_0}{1-Q_0}\right),$$

and

$$\begin{aligned}
\varphi^{+}(x_{N_1}^{*},Q_0) &= -Q_0(1-Q_0)\ln\left(\frac{N_1-Q_0}{1-Q_0}\right) - (1-Q_0)(N_1-Q_0) \\
\varphi_1^{+}(x_{N_1}^{*},Q_0) &= -N_1 \\
\varphi_2^{+}(x_{N_1}^{*},Q_0) &= -\frac{N_1-Q_0}{1-Q_0}.
\end{aligned}$$

Using these expressions we can solve for the coefficients in the quadratic as

$$\begin{aligned}
c_{N_1} &= -\frac{N_1-Q_0}{1-Q_0} \\
b_{N_1} &= -N_1 + (N_1-Q_0)\ln\left(\frac{N_1-Q_0}{1-Q_0}\right) \\
a_{N_1} &= -(1-Q_0)(N_1-Q_0)\left\{\frac{1}{2}\left[\ln\left(\frac{N_1-Q_0}{1-Q_0}\right)\right]^2 - \ln\left(\frac{N_1-Q_0}{1-Q_0}\right) + 1\right\}.
\end{aligned}$$

This gives a modified ILT criterion function of

$$\varphi^{\circ}(x,Q_0)=\begin{cases} -xQ_0-Q_0\left(1-Q_0\right)\exp\left[\frac{x}{1-Q_0}-\ln\left(\frac{Q_0}{1-Q_0}\right)\right] & x<\left(1-Q_0\right)\ln\left(\frac{N_1-Q_0}{1-Q_0}\right) \\ =-\left(1-Q_0\right)\left(N_1-Q_0\right)\left\{\frac{1}{2}\left[\ln\left(\frac{N_1-Q_0}{1-Q_0}\right)\right]^2-\ln\left(\frac{N_1-Q_0}{1-Q_0}\right)+1\right\} & \\ -\left[N_1-\left(N_1-Q_0\right)\ln\left(\frac{N_1-Q_0}{1-Q_0}\right)\right]x & x\geq\left(1-Q_0\right)\ln\left(\frac{N_1-Q_0}{1-Q_0}\right) \\ -\frac{1}{2}\frac{N_1-Q_0}{1-Q_0}x^2 & \end{cases}.$$

Solving the modified saddle-point problem We calculate $\widehat{\gamma}$ by applying gradient-based procedures to the profiled saddle-point criterion function, $\mathbb{L}^{\circ}_{N_1}(\gamma;\rho_0)$. In order to describe these procedures we require expressions for the first and second derivations of the unprofiled saddle-point criterion function. Differentiating $l^{\circ}_{N_1}(\delta,\gamma;\rho_0)$ with respect to δ and γ yields

$$\underset{1+M+L\times 1}{\nabla_{\delta}l^{\circ}_{N_1}(\delta,\gamma;\rho_0)} = t_0+\frac{1}{N_1}\sum_{i=1}^{N_1}\varphi_1^{\circ}(t_i(\gamma)'\delta,Q_0)t_i(\gamma) \tag{33}$$

$$\underset{K\times 1}{\nabla_{\gamma}l^{\circ}_{N_1}(\delta,\gamma;\rho_0)} = \frac{1}{N_1}\sum_{i=1}^{N_1}\varphi_1^{\circ}(t_i(\gamma)'\delta,Q_0)\left[\frac{\partial t_i(\gamma)}{\partial\gamma'}\right]'\delta. \tag{34}$$

Note that when $K=L$ (i.e., when γ_0 is just identified in the target population), if we set the last L elements of $\widehat{\delta}$ equal to zero and its first $1+M$ as given in (X) and set $\widehat{\gamma}$ as given in (X), then we have

$$\begin{pmatrix} \nabla_{\delta}l^{\circ}_{N_1}(\widehat{\delta},\widehat{\gamma};\rho_0) \\ \nabla_{\delta}l^{\circ}_{N_1}(\widehat{\delta},\widehat{\gamma};\rho_0) \end{pmatrix}=0,$$

and hence the just identified case is indeed a special case of the saddle-point estimation procedure considered here.

The second derivatives and cross-partial are given by

$$\underset{1+M+L\times 1+M+L}{\nabla_{\delta\delta}l^{\circ}_{N_1}(\delta,\gamma;\rho_0)} = \frac{1}{N_1}\sum_{i=1}^{N_1}\varphi_2^{\circ}(t_i(\gamma)'\delta,Q_0)t_i(\gamma)\,t_i(\gamma)' \tag{35}$$

$$\begin{aligned}\underset{K\times K}{\nabla_{\gamma\gamma}l^{\circ}_{N_1}(\delta,\gamma;\rho_0)} &= \frac{1}{N_1}\sum_{i=1}^{N_1}\varphi_1^{\circ}(t_i(\gamma)'\delta,Q_0)\left[\sum_{j=1}^{1+M+L}\left[\frac{\partial^2 t_i^{(j)}(\gamma)}{\partial\gamma\partial\gamma'}\right]'\delta_j\right] \\ &\quad+\frac{1}{N_1}\sum_{i=1}^{N_1}\varphi_2^{\circ}(t_i(\gamma)'\delta,Q_0)\left[\frac{\partial t_i(\gamma)}{\partial\gamma'}\right]'\delta\delta'\left[\frac{\partial t_i(\gamma)}{\partial\gamma'}\right]\end{aligned} \tag{36}$$

$$\begin{aligned}\underset{K\times 1+M+L}{\nabla_{\gamma\delta}l^{\circ}_{N_1}(\delta,\gamma;\rho_0)} &= \frac{1}{N_1}\sum_{i=1}^{N_1}\varphi_1^{\circ}(t_i(\gamma)'\delta,Q_0)\left[\frac{\partial t_i(\gamma)}{\partial\gamma'}\right]' \\ &\quad+\frac{1}{N_1}\sum_{i=1}^{N_1}\varphi_2^{\circ}(t_i(\gamma)'\delta,Q_0)\left[\frac{\partial t_i(\gamma)}{\partial\gamma'}\right]'\delta t_i(\gamma)'.\end{aligned} \tag{37}$$

We also have $\varphi_1^{\circ}(x,Q)$ and $\varphi_2^{\circ}(x,Q)$ given by

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

with

$$\begin{aligned}\varphi_1^{+}(x,Q) &= -\frac{1}{k(Q)}\frac{k(Q)}{G\left(k(Q)x+G^{-1}(Q)\right)}Q+\frac{Q}{k(Q)}\frac{k(Q)x+G^{-1}(Q)}{G\left(k(Q)x+G^{-1}(Q)\right)^2}G_1\left(k(Q)x+G^{-1}(Q)\right)k(Q) \\ &\quad-\frac{Q}{k(Q)}\frac{k(Q)x+G^{-1}(Q)}{G(k(Q)x+G^{-1}(Q))^2}G_1\left(k(Q)x+G^{-1}(Q)\right)k(Q) \\ &= -\frac{Q}{G\left(k(Q)x+G^{-1}(Q)\right)} \\ \varphi_2^{+}(x,Q) &= \frac{G_1\left(k(Q)x+G^{-1}(Q)\right)}{G\left(k(Q)x+G^{-1}(Q)\right)^2}k(Q)\,Q.\end{aligned}$$

It is straightforward to verify that $\varphi_1^+(0,Q) = \varphi_2^+(0,Q) = -1$, and hence that the Fenchel conjugate is appropriately normalized.

By the chain and product rules the score of the profiled saddle point criterion function is

$$\begin{aligned} s_{N_1}^p(\gamma;\rho_0) &= \nabla_\gamma \mathbb{L}_{N_1}^\circ(\gamma;\rho_0) = \left[\frac{\partial\widetilde{\delta}(\gamma;\rho_0)}{\partial\gamma'}\right]' \nabla_\delta l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right) + \nabla_\gamma l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right) \\ &= \nabla_\gamma l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right), \end{aligned} \tag{38}$$

where the second equality follows from the Envelope Theorem (i.e., since $\nabla_\delta l_{N_1}^\circ(\delta,\gamma;\rho_0) = 0$ at the maximizer $\widetilde{\delta}(\gamma;\rho_0)$). The i^{th}'s contribution to the score is

$$s^p(Z_i,\gamma;\rho_0) = \left[\frac{\partial\widetilde{\delta}(\gamma;\rho_0)}{\partial\gamma'}\right]' (t_0 + \varphi_1^\circ(t_i(\gamma)'\widetilde{\delta}(\gamma;\rho_0),Q_0)t_i(\gamma)) + \varphi_1^\circ(t_i(\gamma)'\widetilde{\delta}(\gamma;\rho_0),Q_0)\left[\frac{\partial t_i(\gamma)}{\partial\gamma'}\right]'\widetilde{\delta}(\gamma;\rho_0).$$

The Hessian of the profile-saddle point criterion is given by

$$\left[\frac{\partial\widetilde{\delta}(\gamma;\rho_0)}{\partial\gamma'}\right]' \nabla_{\gamma\gamma} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right)\left[\frac{\partial\widetilde{\delta}(\gamma;\rho_0)}{\partial\gamma'}\right] + 2\left[\frac{\partial\widetilde{\delta}(\gamma;\rho_0)}{\partial\gamma'}\right]' \nabla'_{\gamma\delta} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right) + \nabla_{\delta\delta} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right).$$

However we can use the equality

$$\left[\frac{\partial\widetilde{\delta}(\gamma;\rho_0)}{\partial\gamma'}\right] = -\left[\nabla_{\delta\delta} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right)\right]^{-1} \nabla_{\gamma\delta} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right)' \tag{39}$$

to write the Hessian as

$$H_{N_1}^p(\gamma;\rho_0) = \nabla_{\gamma\gamma} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right) - \nabla_{\gamma\delta} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right)\left[\nabla_{\delta\delta} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right)\right]^{-1} \nabla_{\gamma\delta} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right)'. \tag{40}$$

We can derive equality (39) by total differentiating $\nabla_\delta l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right)$ with respect to γ, yielding

$$0 = \frac{1}{N_1}\sum\nolimits_{i=1}^{N_1} \varphi_1^\circ(t_i(\gamma)'\widetilde{\delta}(\gamma;\rho_0),Q_0)\left[\frac{\partial t_i(\gamma)}{\partial\gamma'}\right] + \frac{1}{N_1}\sum\nolimits_{i=1}^{N_1} \varphi_2^\circ(t_i(\gamma)'\widetilde{\delta}(\gamma;\rho_0),Q_0)t_i(\gamma)\left[\widetilde{\delta}(\gamma;\rho_0)'\frac{\partial t_i(\gamma)}{\partial\gamma'} + t_i(\gamma)'\frac{\partial\widetilde{\delta}(\gamma;\rho_0)}{\partial\gamma'}\right].$$

Solving for $\partial\widetilde{\delta}(\gamma;\rho_0)/\partial\gamma'$ then gives

$$\begin{aligned} \frac{\partial\widetilde{\delta}(\gamma;\rho_0)}{\partial\gamma'} &= -\left[\frac{1}{N_1}\sum\nolimits_{i=1}^{N_1} \varphi_2^\circ(t_i(\gamma)'\widetilde{\delta}(\gamma;\rho_0),Q_0)t_i(\gamma)t_i(\gamma)'\right]^{-1} \\ &\quad\times\left[\frac{1}{N_1}\sum\nolimits_{i=1}^{N_1} \varphi_1^\circ(t_i(\gamma)'\widetilde{\delta}(\gamma;\rho_0),Q_0)\left[\frac{\partial t_i(\gamma)}{\partial\gamma'}\right] + \varphi_2^\circ(t_i(\gamma)'\widetilde{\delta}(\gamma;\rho_0),Q_0)t_i(\gamma)\widetilde{\delta}(\gamma;\rho_0)'\frac{\partial t_i(\gamma)}{\partial\gamma'}\right] \\ &= -\left[\nabla_{\delta\delta} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right)\right]^{-1} \nabla_{\gamma\delta} l_{N_1}^\circ\left(\widetilde{\delta}(\gamma;\rho_0),\gamma;\rho_0\right)', \end{aligned}$$

as claimed.

Expressions (38) and (40) can be used to apply a gradient-based optimization procedure to the profiled IPT saddle-point criterion function. Doing so requires numerically calculating the concentrated criterion function at each iteration. Fortunately, given the modification to $\varphi_1^+(x,Q)$ suggested above (to deal with restricted domain issues), this is relatively straightforward to do accurately and quickly.

For a given value of γ, we compute the $1 + M + L \times 1$ vector of tilting parameters $\widetilde{\delta}(\gamma;\rho_0)$ and form $\mathbb{L}_{N_1}^\circ(\gamma;\rho_0)$, the profiled IPT criterion function. We solve for $\widetilde{\delta}(\gamma;\rho_0)$ numerically by using the first and second derivatives of $l_{N_1}(\delta,\gamma;\rho_0)$ with respect to δ – given by (33) and (35) above – to implement a Newton-Raphson procedure. To get starting values for this procedure we take a Taylor expansion of (33) around $\widetilde{\delta}(\gamma;\rho_0) = 0$ to get

$$0 \simeq t_0 - \frac{1}{N_1}\sum_{i=1}^{N_1} t_i(\gamma) - \left\{\frac{1}{N_1}\sum_{i=1}^{N_1} t_i(\gamma)t_i(\gamma)'\right\}\delta^*(\gamma;\rho_0),$$

and use the solution $\delta^*(\gamma;\rho_0) = \left(\sum_{i=1}^{N_1} t_i(\gamma)t_i(\gamma)'\right)^{-1}\left(N_1 t_0 - \sum_{i=1}^{N_1} t_i(\gamma)\right)$ as starting values.

The profiled IPT criterion function, $\mathbb{L}^{\circ}_{N_1}(\gamma;\rho_0)$, can be combined with the expression for its score vector given by (38) to implement any of a number of gradient-based optimization procedures.[19] We have found that the BFGS quasi-Newton method, as implemented by MATLAB 6.0's *fminunc()* function, works well in practice and this is the method used in the Monte Carlo experiments undertaken for the paper.

One can also use the expression given for the Hessian, equation (40) above, to implement a Newton-Raphson procedure where the $j+1$ update to γ equals

$$\gamma^{(j+1)} = \gamma^{(j)} - [H^p_N(\gamma^{(j)};\rho_0)]^{-1} s^p_N(\gamma^{(j)};\rho_0),$$

with iteration continuing until $\gamma^{(j+1)} \simeq \gamma^{(j)}$ or some other standard stopping criterion is achieved.

Details on inverse logistic tilting Let $G(v)$ equal the logistic function:

$$G(x) = \frac{\exp(x)}{1+\exp(x)}, \qquad G^{-1}(x) = \ln\left(\frac{x}{1-x}\right).$$

Case 1: Estimation under Assumption 3.1 We also have the simplification

$$k(Q) = -\frac{Q}{G_1(G^{-1}(Q))} = -Q\frac{\left[1+\exp\left[G^{-1}(Q)\right]\right]^2}{\exp\left[G^{-1}(Q)\right]} = -\kappa\frac{\left[1+\exp\ln\left(\frac{Q}{1-Q}\right)\right]^2}{\exp\ln\left(\frac{Q}{1-Q}\right)}$$
$$= -Q\frac{\left[\frac{1}{1-Q}\right]^2}{\frac{Q}{1-Q}} = -\frac{1}{1-Q}.$$

The key to getting the closed form expression for $\varphi(x,Q)$ and $\varphi^+(x,Q)$ given in the paper involves an application of 'integration by substitution':

$$\int_{x=a}^{x=b} f(x)\,\mathrm{d}x = \int_{u=h^{-1}(a)}^{u=h^{-1}(b)} f(h(u))\,h'(u)\,\mathrm{d}u, \tag{41}$$

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

We begin with the MD contrast function. Let $u = \frac{Q}{t}/\left(1-\frac{Q}{t}\right) = \frac{Q}{t-Q}$; this implies that $t = \left(\frac{1+u}{u}\right)Q = h(u)$, we then have

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

which is a normalized version of Nevo's (2002) 'generalized exponential tilting' contrast function. The step from the third to the forth line follows from

$$\int u^q \ln(u)\,\mathrm{d}u = \frac{u^{q+1}}{q+1}\ln u - \frac{u^{q+1}}{(q+1)^2} + C.$$

[19] We have found that the version of $s^p_{N_1}(\gamma;\rho_0)$ that does note take advantage of the simplification due to the Envelope Theorem, given in the first line of (38), works best in practice.

The normalized criterion for the dual problem is

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

where the final 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

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

Now use the facts that

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

to simplify further. We have

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

We can now verify that the first order conditions calculated from the closed-form criterion match those calculated from the general criterion function.

For logistic tilting we have

$$\begin{aligned}
\varphi_{1}^{+}(x,Q) &= -\frac{Q}{G\left(k(Q)x+G^{-1}(Q)\right)} \\
&= -Q\left(1+\exp\left[-k(Q)x-G^{-1}(Q)\right]\right) \\
&= -Q-Q\exp\left[-k(Q)x-G^{-1}(Q)\right].
\end{aligned}$$

This is exactly what one gets by differentiation of the closed-form expression for the criterion. For completeness the

following are handy for computation

$$\begin{aligned}
\varphi_1^+(x,Q) &= -Q - Q\exp\left[-k(Q)\,x - G^{-1}(Q)\right] \\
\varphi_2^+(x,Q) &= k(Q)\,Q\exp\left[-k(Q)\,x - G^{-1}(Q)\right] \\
\varphi_3^+(x,Q) &= -k(Q)^2\,Q\exp\left[-k(Q)\,x - G^{-1}(Q)\right].
\end{aligned}$$

Evaluating the above at $x = 0$ verifies that our normalizations are working as desired:

$$\begin{aligned}
\varphi_1^+(0,Q) &= -Q - Q\exp\left[-\ln\left(\frac{Q}{1-Q}\right)\right] = -Q + \frac{-Q(1-Q)}{Q} = -1 \\
\varphi_2^+(0,Q) &= -\frac{Q^2}{\frac{\exp\left[G^{-1}(Q)\right]}{\left[1+\exp\left[G^{-1}(Q)\right]\right]^2}}\exp\left[-G^{-1}(Q)\right] = -\frac{Q^2}{G\left(G^{-1}(Q)\right)^2} = -1 \\
\varphi_3^+(0,Q) &= -\frac{Q^3}{\frac{\exp\left[G^{-1}(Q)\right]}{\left[1+\exp\left[G^{-1}(Q)\right]\right]^2}\frac{\exp\left[G^{-1}(Q)\right]}{\left[1+\exp\left[G^{-1}(Q)\right]\right]^2}}\exp\left[-G^{-1}(Q)\right] = -\frac{Q^3}{G\left(G^{-1}(Q)\right)^2}\frac{1}{G_1\left(G^{-1}(\kappa)\right)} = k(Q) = -\frac{1}{1-Q}.
\end{aligned}$$

Case 2: Estimation under Assumption 3.2 We have

$$\begin{aligned}
\varphi(x;Q) &= -xG^{-1}(Q) - \int_x^a G^{-1}\left(\frac{t}{t+(1-Q)/Q}\right)\mathrm{d}t \\
&= -x\ln\left(\frac{Q}{1-Q}\right) - \int_x^a \ln\left(\frac{tQ}{1-Q}\right) \\
&= -x\ln\left(\frac{Q}{1-Q}\right) - \frac{1-Q}{Q}\int_x^a \frac{Q}{1-Q}\ln\left(\frac{tQ}{1-Q}\right) \\
&= -x\ln\left(\frac{Q}{1-Q}\right) - \frac{1-Q}{Q}\left[\frac{tQ}{1-Q}\ln\left(\frac{tQ}{1-Q}\right) - \frac{tQ}{1-Q}\right]_x^a \\
&= -x\ln\left(\frac{Q}{1-Q}\right) + \frac{1-Q}{Q}\left[\frac{xQ}{1-Q}\ln\left(\frac{xQ}{1-Q}\right) - \frac{xQ}{1-Q}\right] + c(a) \\
&= -x\ln\left(\frac{Q}{1-Q}\right) + x\ln(x) + x\ln\left(\frac{Q}{1-Q}\right) - x \\
&= x\ln(x) - x,
\end{aligned}$$

and

$$\begin{aligned}
\varphi^+(x,Q) &= -\left\{\left[x + G^{-1}(Q)\right]\frac{1-Q}{Q}\frac{G(x+G^{-1}(Q))}{1-G(x+G^{-1}(Q))} + \int_{\frac{1-Q}{Q}\frac{G(x+G^{-1}(Q))}{1-G(x+G^{-1}(Q))}}^a G^{-1}\left(\frac{t}{t+(1-Q)/Q}\right)\mathrm{d}t\right\} \\
&= -\left\{\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right] + \int_{\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]}^a \ln\left(\frac{tQ}{1-Q}\right)\right\} \\
&= -\left\{\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right] + \frac{1-Q}{Q}\int_{\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]}^a \frac{Q}{1-Q}\ln\left(\frac{tQ}{1-Q}\right)\right\} \\
&= -\left\{\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right] + \frac{1-Q}{Q}\left[\frac{tQ}{1-Q}\ln\left(\frac{tQ}{1-Q}\right) - \frac{tQ}{1-Q}\right]_{\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]}^a\right\} \\
&= -\left\{\begin{array}{c}\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right] \\ -\frac{1-Q}{Q}\left[\begin{array}{c}\frac{Q}{1-Q}\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\ln\left(\frac{Q}{1-Q}\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\right) \\ -\frac{Q}{1-Q}\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\end{array}\right] + c(a)\end{array}\right\} \\
&= -\left\{\begin{array}{c}\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right] \\ -\frac{1-Q}{Q}\left[\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\left[x+\ln\left(\frac{Q}{1-Q}\right)\right] - \exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right]\right] + c(a)\end{array}\right\} \\
&= -\frac{1-Q}{Q}\exp\left[x+\ln\left(\frac{Q}{1-Q}\right)\right] - c(a) \\
&\propto -\exp[x].
\end{aligned}$$

Details on inverse linear probability tilting

Case 2: Estimation under Assumption 3.2

$$\begin{aligned}
\varphi(x,Q) &= -xG^{-1}(Q) - \int_x^a G^{-1}\left(\frac{t}{t+(1-Q)/Q}\right)\mathrm{d}t \\
&= -x\left[Q-\frac{1}{2}\right] - \int_x^a \left[\frac{t}{t+(1-Q)/Q} - \frac{1}{2}\right]\mathrm{d}t \\
&= -x\left[Q-\frac{1}{2}\right] - \left[\frac{t}{2} - \frac{1-Q}{Q}\ln\left[t+\frac{1-Q}{Q}\right]\right]_x^a \\
&= -xQ + \frac{x}{2} + \frac{x}{2} - \frac{1-Q}{Q}\ln\left[x+\frac{1-Q}{Q}\right] + c(a) \\
&\propto (1-Q)\,x - \frac{1-Q}{Q}\ln\left[x+\frac{1-Q}{Q}\right]
\end{aligned}$$

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

B Proofs

Uniform weak law of large numbers (UWLLN)

Central limit theorem (CLT)

B.1 Proof of Theorem 6.1

Since the IPT estimate of γ_0 can be represented as the solution to a two-step sequential GMM problem, Theorem 6.1 follows from standard results on sequential GMM estimators. The 'regularity conditions' referred to in the statement of Theorem 6.1 are those referenced in Theorem 6.1 of Newey and McFadden (1994, p. 2178). This particular application of Newey and McFadden's theorem is similar to Theorems 3.1 and 4.1 of Wooldridge (2002b).

The first step of the proof involves demonstrating equivalence of the IPT estimator to a particular sequential GMM estimator. The procedure outlined in Section 6 consists of first estimating ρ_0 by the full sample analog estimates

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

Second IPT, with ρ_0 replaced by $\widehat{\rho}$, is applied to the selected sample (i.e., the $D=1$ sub-sample). The first order condition

for IPT are given by

$$\begin{aligned} 0 &= t_0 + \frac{1}{N_1}\sum\nolimits_{i=1}^{N_1} \varphi_1^+(t(Z,\widehat{\zeta})'\widehat{\delta},\widehat{Q})t(Z,\widehat{\zeta}) \\ &= \frac{1}{N}\sum\nolimits_{i=1}^{N}\left\{\frac{D_i}{G(k(\widehat{Q})t(Z,\widehat{\zeta})'\widehat{\delta}+G^{-1}(\widehat{Q}))}-1\right\}\left\{\begin{array}{c} 1 \\ h^*(X_i) \end{array}\right\}. \end{aligned}$$

Let $\alpha=(\alpha_1,\alpha_2')'$ with $\alpha_1 = k(Q)\delta_1 + G^{-1}(Q) - \zeta'\delta_2$ and $\alpha_2=\delta_2$. Since there is a one-to-one mapping between α and δ we rewrite

$$\frac{1}{N}\sum_{i=1}^{N} m_1(Z_i,\widehat{\alpha}) = 0, \qquad m_1(Z,\alpha) = \left\{\frac{D}{G(w(X)'\alpha)}-1\right\}w(X), \tag{42}$$

with $w(x)=(1,h^*(x)')'$. Consistency of the solution to (42) for α_0 requires that $\mathbb{E}[h(X,\zeta_0)h(X,\zeta_0)'|D=1]$ is of full rank as well as the monotonicity and continuity conditions imposed on $G(\cdot)$ by Assumption 5.5 (cf., Nevo 2003, Proposition 1, p. 45).

In the second and final step γ_0 is estimated by the solution to

$$\frac{1}{N}\sum_{i=1}^{N} m_2(Z_i,\widehat{\alpha},\widehat{\gamma}) = 0, \tag{43}$$

where $\widehat{\alpha}$ is fixed at its first step value and

$$m_2(Z_i,\alpha,\gamma) = \frac{D\psi(Z,\gamma)}{G(w(X)'\alpha)} - \frac{\Upsilon_0 w(X)}{G(w(X)'\alpha)}\left(D - G(w(X)'\alpha_0)\right).$$

Note that γ only enters the first term, with the second term (implicitly) evaluated at γ_0.

To see that the solution to (43) is equivalent to the IPT estimate of γ_0 given by the solution to

$$\sum_{i=1}^{N_1}\widetilde{\pi}_i(\widehat{\rho})\psi(Z_i,\widehat{\gamma}) = 0, \tag{44}$$

manipulate (44) as follows:

$$\begin{aligned} \sum_{i=1}^{N_1}\widetilde{\pi}_i(\widehat{\rho})\psi(Z_i,\widehat{\gamma}) &= \sum_{i=1}^{N_1}\widetilde{\pi}_i(\widehat{\rho})[\psi(Z_i,\widehat{\gamma})-\Upsilon_0 w(X_i)] + \frac{1}{N}\sum_{i=1}^{N}\Upsilon_0 w(X_i) \\ &= \frac{1}{N_1}\sum_{i=1}^{N_1}\frac{\widehat{Q}}{G(k(\widehat{Q})t(Z_i,\widehat{\zeta})'\widehat{\delta}+G^{-1}(\widehat{Q}))}[\psi(Z_i,\widehat{\gamma})-\Upsilon_0 w(X_i)] + \frac{1}{N}\sum_{i=1}^{N}\Upsilon_0 w(X_i) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{D_i\psi(Z_i,\widehat{\gamma})}{G(w(X_i)'\widehat{\alpha})} - \frac{\Upsilon_0 w(X)}{G(w(X_i)'\widehat{\alpha})}(D_i - G(w(X_i)'\widehat{\alpha})) \\ &= \frac{1}{N}\sum_{i=1}^{N} m_2(Z_i,\widehat{\alpha},\widehat{\gamma}) = 0. \end{aligned}$$

Line one follows from the fact that $\sum_{i=1}^{N_1}\widetilde{\pi}_i(\widehat{\rho})=1$ and $\sum_{i=1}^{N_1}\widetilde{\pi}_i(\widehat{\rho})h(X_i,\widehat{\zeta})=0$, line two from the definition of $\widetilde{\pi}_i(\widehat{\rho})$, and line three from the definitions of $\widehat{Q}$ and $\widehat{\delta}$ and rearrangement.

Second we verify identification and consistency; using Assumptions 5.3 to 5.5 and iterated expectations we have

$$\begin{aligned} \mathbb{E}[m_2(Z,\alpha_0,\gamma)] &= \mathbb{E}\left[\frac{D\psi(Z_i,\gamma)}{G(w(X)'\alpha_0)} - \frac{\Upsilon_0 w(X)}{G(w(X)'\alpha_0)}(D_i - G(w(X)'\alpha_0))\right] \\ &= \mathbb{E}\left[\frac{D\psi(Z,\gamma)}{p_0(X)} - \frac{\mathbb{E}[\psi(Z,\gamma_0)|X]}{p_0(X)}(D-p_0(X))\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{D\psi(Z,\gamma)}{p_0(X)} - \frac{\mathbb{E}[\psi(Z,\gamma_0)|X]}{p_0(X)}(D-p_0(X))\middle| X,D\right]\right] \\ &= \mathbb{E}[\psi(Z,\gamma)], \end{aligned}$$

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

Consistency of $\widehat{\gamma}$ for γ_0 follows from (global) identification and additional (standard) regularity conditions (e.g., Newey

and McFadden 1994, Section 2.5, Wooldridge 2002a, Chapter 12). These conditions, which include moment and continuity conditions on $\psi(Z,\gamma)$ and consistency of $\widehat{\alpha}$ for α_0, ensure that

$$\frac{1}{N}\sum_{i=1}^{N} m_2(Z_i,\widehat{\alpha},\gamma)$$

converges uniformly in $\gamma \in \mathcal{G} \subset \mathbb{R}^K$ to $\mathbb{E}[m_2(Z,\alpha_0,\gamma)]$.

Finally we show asymptotic normality. Expanding each row of (44) with respect to γ around γ_0 in a mean value expansion and using positive definiteness of $\mathbb{E}[\Gamma_0(X)]$ as well as the UWLLN gives

$$\sqrt{N}(\widehat{\gamma}-\gamma_0) = \mathbb{E}[\Gamma_0(X)]^{-1\prime}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_2(Z_i,\widehat{\alpha},\gamma_0)\right] + o_p(1),$$

a second mean value expansion of the term in $[\cdot]$ gives

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_2(Z_i,\widehat{\alpha},\gamma_0) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_2(Z_i,\alpha_0,\gamma_0) + \mathbb{E}\left[\frac{\partial m_2(Z_i,\alpha_0,\gamma_0)}{\partial \alpha'}\right](\widehat{\alpha}-\alpha_0) + o_p(1)\\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_2(Z_i,\alpha_0,\gamma_0) + o_p(1),\end{aligned}$$

which, after substitution, yields the asymptotic linear representation

$$\sqrt{N}(\widehat{\gamma}-\gamma_0) = \mathbb{E}[\Gamma_0(X)]^{-1\prime}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_2(Z_i,\alpha_0,\gamma_0)\right] + o_p(1).$$

Note that no adjustment needs to be made for first and second step estimation. Since

$$\mathbb{E}[\Sigma_0(X)] = \mathbb{E}\left[m_2(Z,\alpha_0,\gamma_0)m_2(Z,\alpha_0,\gamma_0)'\right]$$

the result follows immediately from an application of the CLT, Slutsky's Theorem, and a comparison of the resulting asymptotic variance expression with (23).

Now consider double robustness. First assume that $\Upsilon_0 w(X) \neq \mathbb{E}[\psi(Z,\gamma_0)|X]$, $m_2(Z,\alpha_0,\gamma_0)$ continues to be a valid moment function (cf., Robins, Rotnitzky and Zhao 1994). Now assume that there is no α such that $G(w(x)'\alpha) = p_0(x)$. Let α_* be the probability limit of the first step estimate and observe that since

$$\begin{aligned}\mathbb{E}[m_2(Z,\alpha_*,\gamma_0)] &= \mathbb{E}\left[\mathbb{E}\left[\frac{p_0(X)\psi(Z,\gamma_0)}{G(w(X)'\alpha_*)} - \frac{\Upsilon_0 w(X)}{G(w(X)'\alpha_*)}(p_0(X)-G(w(X)'\alpha_*))\middle| Z\right]\right]\\ &= \mathbb{E}\left[\frac{p_0(X)\mathbb{E}[\psi(Z,\gamma_0)|X]}{G(w(X)'\alpha_*)} - \frac{\Upsilon_0 w(X)}{G(w(X)'\alpha_*)}(p_0(X)-G(w(X)'\alpha_*))\right]\\ &= \mathbb{E}\left[\frac{p_0(X)}{G(w(X)'\alpha_*)}\{\mathbb{E}[\psi(Z,\gamma_0)|X] - \Upsilon_0 w(X)\}\right]\\ &= 0,\end{aligned}$$

$m_2(Z,\alpha_*,\gamma_0)$ continues to be a valid moment under propensity score misspecification.

B.2 Proof of Theorem ??

The structure of the proof is analogous to that for Theorem 6.1 and is only sketched.

The first-step moment is

$$m_1(Z,\alpha) = \frac{G(w(X)'\alpha_0)}{1-G(w(X)'\alpha_0)}\left\{\frac{D}{G(w(X)'\alpha_0)} - 1\right\} w(X) = 0$$

since

$$\begin{aligned}
0 &= \frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\sum_{i=N_1+1}^{N}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}-1=0\\
0 &= \frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\sum_{i=N_1+1}^{N}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\left[h^*(X)-\frac{1}{N_1}\sum_{i=1}^{N_1}h^*(X)\right]\\
&= \frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\sum_{i=N_1+1}^{N}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}h^*(X)-\frac{1}{N_1}\sum_{i=1}^{N_1}h^*(X),
\end{aligned}$$

therefore

$$\begin{aligned}
0 &= \frac{1}{N_0}\frac{1-\widehat{Q}}{\widehat{Q}}\sum_{i=N_1+1}^{N}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}w(X)-\frac{1}{N_1}\sum_{i=1}^{N_1}w(X)=0\\
&= \frac{1}{N}\frac{1}{\widehat{Q}}\sum_{i=1}^{N}(1-D_i)\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}w(X)-\frac{1}{N}\frac{1}{\widehat{Q}}\sum_{i=1}^{N}D_iw(X)\\
&= \frac{1}{N}\frac{1}{\widehat{Q}}\sum_{i=1}^{N}\left\{(1-D_i)\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}-D_i\frac{1-G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\right\}w(X)\\
&= \frac{1}{N}\frac{1}{\widehat{Q}}\sum_{i=1}^{N}\left\{\frac{D_i-G(w(X_i)'\widehat{\alpha}^A}{1-G(w(X_i)'\widehat{\alpha}^A)}\right\}w(X)\\
&= \frac{1}{N}\frac{1}{\widehat{Q}}\sum_{i=1}^{N}\frac{G(w(X_i)'\widehat{\alpha}^A}{1-G(w(X_i)'\widehat{\alpha}^A)}\left\{\frac{D_i}{G(w(X_i)'\widehat{\alpha}^A}-1\right\}w(X).
\end{aligned}$$

The second-step moment is

$$m_2(Z,\alpha,\gamma)=(1-D)\frac{1}{Q}\frac{G(w(X)'\alpha)}{1-G(w(X)'\alpha)}\psi(Z,\gamma)-\frac{\Upsilon_0 w(X)}{Q}\frac{G(w(X)'\alpha)}{1-G(w(X)'\alpha)}\left[D-G(w(X)'\alpha)\right],$$

since

$$\begin{aligned}
0 &= \sum_{i=N_1+1}^{N}\pi_i(\widehat{\rho})\psi(Z_i,\widehat{\gamma})\\
&= \frac{1}{N_0}\sum_{i=1}^{N}(1-D_i)\frac{1-\widehat{Q}}{\widehat{Q}}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\psi(Z_i,\widehat{\gamma})\\
&= \frac{1}{N}\sum_{i=1}^{N}(1-D_i)\frac{1}{\widehat{Q}}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\psi(Z_i,\widehat{\gamma})+\frac{1}{N}\sum_{i=1}^{N}(1-D_i)\frac{1}{\widehat{Q}}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\Upsilon_0 w(X_i)-\frac{1}{N_1}\sum_{i=1}^{N_1}\Upsilon_0 w(X_i)\\
&= \frac{1}{N}\sum_{i=1}^{N}(1-D_i)\frac{1}{\widehat{Q}}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\psi(Z_i,\widehat{\gamma})+\frac{1}{N}\sum_{i=1}^{N}(1-D_i)\frac{1}{\widehat{Q}}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\Upsilon_0 w(X_i)-\frac{1}{N}\sum_{i=1}^{N}D_i\frac{1}{\widehat{Q}}\Upsilon_0 w(X_i)\\
&= \frac{1}{N}\sum_{i=1}^{N}(1-D_i)\frac{1}{\widehat{Q}}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\psi(Z_i,\widehat{\gamma})+\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\widehat{Q}}\left[(1-D_i)\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}-D_i\right]\Upsilon_0 w(X_i)\\
&= \frac{1}{N}\sum_{i=1}^{N}(1-D_i)\frac{1}{\widehat{Q}}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\psi(Z_i,\widehat{\gamma})+\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\widehat{Q}}\left[(1-D_i)\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}-D_i\frac{1-G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\right]\Upsilon_0 w(X_i)\\
&= \frac{1}{N}\sum_{i=1}^{N}(1-D_i)\frac{1}{\widehat{Q}}\frac{G(w(X_i)'\widehat{\alpha}^A)}{1-G(w(X_i)'\widehat{\alpha}^A)}\psi(Z_i,\widehat{\gamma})-\frac{1}{\widehat{Q}}\frac{\Upsilon_0 w(X_i)}{1-G(w(X_i)'\widehat{\alpha}^A)}\left[D_i-G(w(X_i)'\widehat{\alpha}^A)\right].
\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\frac{\partial m_2\left(Z,\alpha_0,\gamma_0\right)}{\partial\gamma'}\right] &= \mathbb{E}\left[\left(1-D\right)\frac{1}{Q_0}\frac{G(w\left(X\right)'\alpha_0)}{1-G(w\left(X\right)'\alpha_0)}\frac{\partial\psi\left(Z,\gamma\right)}{\partial\gamma'}-\frac{\Upsilon_0 w\left(X\right)}{Q}\frac{G(w\left(X\right)'\alpha_0)}{1-G(w\left(X\right)'\alpha_0)}\left[D-G(w\left(X\right)'\alpha_0)\right]\right]\\ &= \mathbb{E}\left[\frac{G(w\left(X\right)'\alpha_0)}{Q_0}\frac{\partial\psi\left(Z,\gamma\right)}{\partial\gamma'}\right]\\ &= \mathbb{E}\left[\left.\frac{\partial\psi\left(Z,\gamma\right)}{\partial\gamma'}\right| D=1\right].\end{aligned}$$

The linear predictor in question is given by

$$\begin{aligned}&\mathbb{E}^*_{F_0}\left[\left.\frac{Dm\left(Y_1,X,\gamma_0\right)}{G(\alpha_0+h\left(X\right)'\beta_0)}\right|\left\{\frac{D}{G(\alpha_0+h\left(X\right)'\beta_0)}-1\right\}w\left(X\right)\right]\\ &\qquad=\mathbb{E}_{F_0}\left[\frac{Dm\left(Y_1,X,\gamma_0\right)}{G(\alpha_0+h\left(X\right)'\beta_0)}\left\{\frac{D}{G(\alpha_0+h\left(X\right)'\beta_0)}-1\right\}w\left(X\right)'\right]\\ &\qquad\quad\times\mathbb{E}_{F_0}\left[\left\{\frac{D}{G(\alpha_0+h\left(X\right)'\beta_0)}-1\right\}^2 w\left(X\right)w\left(X\right)'\right]^{-1}\left\{\frac{D}{G(\alpha_0+h\left(X\right)'\beta_0)}-1\right\}w\left(X\right),\end{aligned}$$

which, after using iterated expectations and the missing-at-random assumption, gives

$$\begin{aligned}&\mathbb{E}_{F_0}\left[\left\{\frac{p_0\left(X\right)}{G(\alpha_0+h\left(X\right)'\beta_0)^2}-\frac{p_0\left(X\right)}{G(\alpha_0+h\left(X\right)'\beta_0)}\right\}\mathbb{E}_{F_0}\left[m\left(Y_1,X,\gamma_0\right)|X\right]w\left(X\right)'\right]\\ &\qquad\times\mathbb{E}_{F_0}\left[\left\{\frac{p_0\left(X\right)}{G(\alpha_0+h\left(X\right)'\beta_0)^2}-\frac{2p_0\left(X\right)}{G(\alpha_0+h\left(X\right)'\beta_0)}+1\right\}w\left(X\right)w\left(X\right)'\right]^{-1}\left\{\frac{D}{G(\alpha_0+h\left(X\right)'\beta_0)}-1\right\}w\left(X\right).\end{aligned}$$

Substituting $\Upsilon_0 w\left(X\right)=\mathbb{E}_{F_0}\left[m\left(Y_1,X,\gamma_0\right)|X\right]$, factoring and rearranging, finally gives

$$\frac{\Upsilon_0 w\left(X\right)}{G(\alpha_0+h\left(X\right)'\beta_0)}\left(D-G(\alpha_0+h\left(X\right)'\beta_0)\right)+\Upsilon_0\mathbb{E}^*_{F_0}\left[w\left(X\right)|\left\{\frac{D}{G(\alpha_0+h\left(X\right)'\beta_0)}-1\right\}w\left(X\right)\right].$$

Therefore the equivalent system is given by

where the second line uses the linear predictor property

$$\mathbb{E}\left[\mathbb{E}^*\left[w\left(X\right)|\left\{\frac{D}{G(\alpha_0+h\left(X\right)'\beta_0)}-1\right\}w\left(X\right)\right]\right]=\mathbb{E}\left[w\left(X\right)\right]=\begin{pmatrix}1\\ \zeta_0\end{pmatrix},$$

and that $\Upsilon_0\left[1,\zeta_0'\right]=\mathbb{E}_{F_0}\left[m\left(Y_1,X,\gamma_0\right)\right]=0$.

which also equals the efficient score under the maintained auxiliary parametric Assumptions 5.5 and 6.1. Double robustness of IPT follows from the asymptotic equivalence to AIPW.

the second condition is a direct consequence of (**??**) and (**??**). Iterated expectations, (1) and the second condition imply that

$$\begin{aligned}\int\int m\left(y_1,x,\gamma_0\right)f_*\left(y_1,x\right)\mathrm{d}y_1\mathrm{d}x &= \int\mathbb{E}_{F_0}\left[m\left(Y_1,X,\gamma_0\right)|X=x\right]f_*\left(x\right)\mathrm{d}x\\ &= \Upsilon_0\int w\left(x\right)f_*\left(x\right)\mathrm{d}x\\ &= \Upsilon_0[1,\zeta_0']'=0\end{aligned}$$

with $f_*\left(x\right)=f_0\left(x|d=1\right)\varphi_1^+\left(t\left(x,\zeta_0\right)'\delta_*;Q_0\right)$ and the final line due to (**??**) and (**??**).

B.3 Stochastic expansion

The first, second, third and fourth derivatives of $\varphi^{+}(x,Q)$ are given by

$$\begin{aligned}
\varphi_3^{+}(x,Q) &= \frac{G_2\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^2}k\left(Q\right)^2Q-2\frac{G_1\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^2}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^3}k\left(Q\right)^2Q\\
\varphi_4^{+}(x,Q) &= \frac{G_3\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^2}k\left(Q\right)^3Q\\
&\quad -2\frac{G_1\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)G_2\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^3}k\left(Q\right)^3Q\\
&\quad -4\frac{G_1\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)G_2\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^3}k\left(Q\right)^3Q\\
&\quad +6\frac{G_1\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^3}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^4}k\left(Q\right)^3Q\\
&= \frac{G_3\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^2}k\left(Q\right)^3Q\\
&\quad -6\frac{G_1\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)G_2\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^3}k\left(Q\right)^3Q\\
&\quad +6\frac{G_1\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^3}{G\left(k\left(Q\right)x+G^{-1}\left(Q\right)\right)^4}k\left(Q\right)^3Q,
\end{aligned}$$

with

$$\begin{aligned}
\varphi_3^{+}(0,Q) &= \frac{G_2\left(G^{-1}\left(Q\right)\right)}{Q}k\left(Q\right)^2-2\frac{G_1\left(G^{-1}\left(Q\right)\right)^2}{Q^2}k\left(Q\right)^2\\
&= \frac{G_2\left(G^{-1}\left(Q\right)\right)}{Q}\left[\frac{-Q}{G_1\left(G^{-1}\left(Q\right)\right)}\right]^2-2\frac{G_1\left(G^{-1}\left(Q\right)\right)^2}{Q^2}\left[\frac{-Q}{G_1\left(G^{-1}\left(Q\right)\right)}\right]^2\\
&= \frac{G_2\left(G^{-1}\left(Q\right)\right)}{G_1\left(G^{-1}\left(Q\right)\right)}\frac{Q}{G_1\left(G^{-1}\left(Q\right)\right)}-2.\\
&= \frac{G_2\left(G^{-1}\left(Q\right)\right)}{G_1\left(G^{-1}\left(Q\right)\right)^2}Q-2\\
&= -\frac{G_2\left(G^{-1}\left(Q\right)\right)}{G_1\left(G^{-1}\left(Q\right)\right)}k\left(Q\right)-2,
\end{aligned}$$

and

$$\begin{aligned}
\varphi_4^{+}(0,Q) &= \frac{G_3\left(G^{-1}\left(Q\right)\right)}{Q}k\left(Q\right)^3\\
&\quad -6\frac{G_1\left(G^{-1}\left(Q\right)\right)G_2\left(G^{-1}\left(Q\right)\right)}{Q^2}k\left(Q\right)^3\\
&\quad +6\frac{G_1\left(G^{-1}\left(Q\right)\right)^3}{Q^3}k\left(Q\right)^3\\
&= \frac{G_3\left(G^{-1}\left(Q\right)\right)}{Q}\frac{-Q^3}{G_1\left(G^{-1}\left(Q\right)\right)^3}\\
&\quad -6\frac{G_1\left(G^{-1}\left(Q\right)\right)G_2\left(G^{-1}\left(Q\right)\right)}{Q^2}\frac{-Q^3}{G_1\left(G^{-1}\left(Q\right)\right)^3}\\
&\quad +6\frac{G_1\left(G^{-1}\left(Q\right)\right)^3}{Q^3}\frac{-Q^3}{G_1\left(G^{-1}\left(Q\right)\right)^3}\\
&= -\frac{G_3\left(G^{-1}\left(Q\right)\right)}{G_1\left(G^{-1}\left(Q\right)\right)^3}Q^2+6\frac{G_2\left(G^{-1}\left(Q\right)\right)}{G_1\left(G^{-1}\left(Q\right)\right)^2}Q-6\\
&= -\frac{G_3\left(G^{-1}\left(Q\right)\right)}{G_1\left(G^{-1}\left(Q\right)\right)}k\left(Q\right)^2-6\frac{G_2\left(G^{-1}\left(Q\right)\right)}{G_1\left(G^{-1}\left(Q\right)\right)}k\left(Q\right)-6.
\end{aligned}$$

Inspection of $\varphi_3^+(0,Q)$ and $\varphi_4^+(0,Q)$ reveals that third order equivalence with EL requires that

$$\begin{aligned} -\frac{G_2\left(G^{-1}(Q)\right)}{G_1\left(G^{-1}(Q)\right)}k(Q) &= 0 \\ -\frac{G_3\left(G^{-1}(Q)\right)}{G_1\left(G^{-1}(Q)\right)}k(Q)^2 - 6\frac{G_2\left(G^{-1}(Q)\right)}{G_1\left(G^{-1}(Q)\right)}k(Q) &= 0. \end{aligned}$$

Further manipulation reveals that these two conditions are equivalent to the requirement that

$$G_1\left(G^{-1}(Q)\right) \neq 0, \qquad G_2\left(G^{-1}(Q)\right) = 0, \qquad G_3\left(G^{-1}(Q)\right) = 0.$$

References

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

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

Barron, Andrew R. and Chyong-Hwa Sheu. (1991). "Approximation of density functions by sequences of exponential families," *Annals of Statistics* 19 (3): 1347 - 1369.

Bhattacharya, Debopam. (2005). "Inference in panel data models under attrition caused by unobservables," *Mimeo.*

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.

Chamberlain, Gary. (1987). "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics* 34 (1): 305 - 334.

Chamberlain, Gary. (1992). "Efficiency bounds for semiparametric regression," *Econometrica* 60 (3): 567 - 596.

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 non-classical measurement errors, missing data and treatment effects, *Mimeo.*

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

Cosslett, Stephen R. (1993). "Estimation from endogenously stratified samples," *Handbook of Statistics 11*: 1 - 44 (G.S. Maddala et al., Eds.). Amsterdam: Elsevier.

Cosslett, Stephen R. (1997). "Nonparametric maximum likelihood methods," *Handbook of Statistics 15*: 385 - 404 (G.S. Maddala & C.R. Rao, Eds.). Amsterdam: Elsevier.

Crump, Richard K. et al. (2006). "Nonparametric tests of treatment effect heterogeneity," *Mimeo.*

Dehejia, Rajeev H. and Sadek Wahba. (1999). "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and Statistics* 84 (1): 151 - 161.

Dehejia, Rajeev H. and Sadek Wahba. (2002). "Causal effects in nonexperimental studies: reevaluating the evaluation of training programs," *Journal of the American Statistical Association* 94 (448): 1053 - 1062.

Donald, Stephen G., Guido W. Imbens and Whitney K. Newey. (2003). "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics* 117 (1): 55 - 93.

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

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

Deming, W. Edwards and Frederick F. Stephan. (1940). "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *Annals of Mathematical Statistics* 11 (4): 427 - 444.

Donald, Stephen G., Guido W. Imbens and Whitney K. Newey. (2003). "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Journal of Econometrics* 117 (1): 55 - 93.

Fitzgerald, John, Peter Gottschalk and Robert Moffitt. (1998). "An analysis of sample attrition in panel data: the Michigan panel study of income dynamics," *Journal of Human Resources* 33 (2): 251 - 299.

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

Gilbert, Peter B. (2000). "Large sample theory of maximum likelihood estimations in semiparametric biased sampling models," *Annals of Statistics* 28 (1): 151 - 194.

Gill, Richard D., Yehuda Vardi and Jon A. Wellner. (1988). "Large sample theory of empirical distributions in biased samples," *Annals of Statistics* 16 (3): 1069 - 1112.

Graham, Bryan S. (2007). "A note on semiparametric efficiency in moment condition models with missing data," *Mimeo.*

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

Hansen, Lars Peter, John Heaton and Amir Yaron. (1996). "Finite-sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics* 14 (3): 262 - 280.

Hausman, Jerry A. and David A. Wise. (1979). "Attrition bias in experimental and panel data: the Gary income maintenance experiment," *Econometrica* 47 (2): 455 - 474.

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, 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.

Horvitz, D.G. and D. J. Thompson. (1952). "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association* 47 (260): 663 - 685.

Imbens, Guido W. (1992). "An efficient method of moments estimator for discrete choice models with choice-based sampling," *Econometrica* 60 (5): 1187 - 1214.

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

Kitamura, Yuichi and Michael Stutzer. (1997). "An information-theoretic alternative to generalized method of moments estimation," *Econometrica* 65 (4): 861 - 874.

Kitamura, Yuichi. (2006). "Empirical likelihood methods in econometrics: theory and practice," *Mimeo.*

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.

Lawless, J.F., J.D. Kalbfleisch and C.J. Wild. (1999). "Semiparametric methods for response-selective and missing data problems in regression," *Journal of the Royal Statistical Society B* 61 (2): 413 - 438.

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.

Manski, Charles F. (2005). *Social Choice with Partial Knowledge of Treatment Response.* Princeton: Princeton University Press.

Mittelhammer, Ron C., George G. Judge and Ron Schoenberg. (2005). "Empirical evidence concerning the finite sample performance of EL-type structural equation estimation and inference methods," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: 292 - 305 (D.W.K. Andrews & J.H. Stock, Eds.). Cambridge: Cambridge University Press.

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.

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

Prokhorov, Artem and Peter Schmidt. (2005). "GMM redundancy results for general missing data problems," *Mimeo.*

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

Qin, Jin and Jerry Lawless. (1994). "Empirical likelihood and general estimating equations," *Annals of Statistics* 22 (1): 300 - 325.

Qin, Jing. (1993). "Empirical likelihood in biased sample problems," *Annals of Statistics* 21 (3): 1182 - 1196.

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

Qin, Jing, Denis Leung and Jun Shao. (2002). "Estimation with survey data under nonignorable nonresponse or informative sampling," *Journal of the American Statistical Association* 97 (457): 193 - 200.

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.

Robins, James M., Miguel Ángel Hernán and Babette Brumback. (2000). "Marginal structural models and causal inference in epidemiology," *Epidemiology* 11 (5): 550 - 560.

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., Steven D. Mark and Whitney Newey. (1992). "Estimating exposure effects by modelling the expectation of exposure conditional on confounders," *Biometrics* 48 (2): 479 - 495.

Robins, James M. and Andrea Rotnitzky. (1995). "Semiparametric efficiency in multivariate regression models," *Journal of the American Statistical Association* 90 (429): 122 - 129.

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

Särndal, Carl Erik and Sixten Lundström. (2005). *Estimation in Surveys with Nonresponse.* West Sussex: John Wiley and Sons Ltd.

Scharfstein, Daniel O., Andrea Rotnitzky and James M. Robins. (1999). "Adjusting for nonignorable drop-out using semiparametric nonresponse models," *Journal of the American Statistical Association* 94 (448): 1096 - 1120.

Vardi,Yehuda. (1985). "Empirical distributions in selection bias models," *Annals of Statistics* 13 (1): 178 - 203.

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

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

Wooldridge, Jeffrey M. (2002b). "Inverse probability weighted M-estimators for sample selection, attrition and stratification," *Portuguese Economic Journal* 1 (2): 117 - 139.

Wooldridge, Jeffrey M. (2007). "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics,* forthcoming.

	$m(y_1, x, , \gamma)$	$l(y_0, x, \gamma)$	Comments
Panel A: Missing data			
VP Sampling	$m(y_1, x; \gamma)$	$\underline{0}$	Y_0 empty
Missing regressors	$\frac{\partial e(y_1, x_1, \gamma)}{\partial \gamma}(x_0 - e(y_1, x_1, \gamma))$	$\underline{0}$	$X = (X_0, X_1')'$, $e(\cdot, \cdot, \gamma)$ a CEF
ATE	y_1	$-y_0 - \gamma$	
Panel B: Sample combination			
TSIV	$g(x) d(y_1, \gamma)$	$-g(x) e(y_0, x_1, \gamma)$	$X = (X_0, X_1')'$, $d(\cdot, \gamma)$, $e(\cdot, \cdot, \gamma)$, $g(\cdot)$ known
M-Estimation	$\underline{0}$	$l(y_0, x; \gamma)$	Y_1 empty
ATT	y_1	$-y_0 - \gamma$	
Measurement error	$\underline{0}$	$l(y_0, x, \gamma)$	
Small area	$\underline{0}$	$y_0 - \gamma$	Y_1 empty

Table 1: Moment function corresponding to motivating examples