GMM 'equivalence' for semiparametric missing data models[1]

Bryan S. Graham[†]

Initial Draft: August 2006
This Draft: August 23, 2007

[1]I would like to thank Gary Chamberlain, Jinyong Hahn, Guido Imbens, Michael Jansson and Whitney Newey for comments on earlier draft. Helpful discussions with Oliver Linton, Cristine Pintos, Jim Powell, Geert Ridder as well as participants in the Berkeley Econometrics Reading Group are gratefully acknowledged. All the usual disclaimers apply. This is a heavily revised version of a paper which previously circulated under the title "A note on semiparametric efficiency in moment condition models with missing data".

[†]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/.

Abstract

This paper shows that the semiparametric efficiency bound for a parameter identified by an unconditional moment restriction with data missing at random (MAR), coincides with that of a particular augmented moment condition problem. The augmented system consists of the inverse probability of observation weighted (IPW) original moment restriction and an additional conditional moment restriction which exhausts all other implications of the MAR assumption. While efficiency bounds for these types of problems are widely-known, the general equivalence result is apparently new. Demonstrating equivalence provides fresh intuitions for several apparent 'paradoxes' in the missing data literature, including the well-known results that smoothness and exclusion priors on the propensity score do not increase the efficiency bound for the parameter of interest and that weighting by a nonparametric estimate of the propensity score results in an efficient estimator while weighting by the true propensity score does not. The 'equivalent' GMM problem also suggests new efficient estimators.

This paper also analyzes the effect of imposing additional semiparametric restrictions on the conditional expectation function (CEF) of the original moment function given always-observed covariates on the variance bound. In the program evaluation context such restrictions are generated by semiparametric models for the CEFs of the two potential outcomes given covariates. By exploiting the insight that these restrictions simply add conditional moments to the 'equivalent' augmented system I apply Chamberlain's (1992a) methods to calculate the corresponding variance bound.

Some related results and intuitions are provided for a family of data combination problems. This family of problems includes the average treatment effect on the treated (ATT) estimand.

JEL Classification: C14, C21, C31

Key Words: Missing Data, Semiparametric Efficiency, Propensity Score, (Augmented) Inverse Probability Weighting, Double Robustness, Empirical Likelihood, Average Treatment Effects, Causal Inference

1 Introduction

Let $Z = (Y_1', Y_0', X)'$ be vector of modelling variables, $\{Z_i\}_{i=1}^{\infty}$ be an independent and identically distributed random sequence drawn from the unknown distribution F_0, β a $K \times 1$ unknown parameter vector and $\psi(Z, \beta) = \psi_1(Y_1, X, \beta) - \psi_0(Y_0, X, \beta)$ a known vector-valued function of the same dimension.[2] The only prior restriction on F_0 is that for some $\beta_0 \in \mathcal{B} \subset \mathbb{R}^K$

$$\mathbb{E}[\psi(Z, \beta_0)] = 0. \tag{1}$$

Chamberlain (1987) showed that the maximal asymptotic precision with which β_0 can be estimated under (1) (subject to identification and regularity conditions) is given by the inverse of

$$\mathbb{E}[\Gamma_0(X)]' \mathbb{E}[\Omega_0(X)]^{-1} \mathbb{E}[\Gamma_0(X)], \tag{2}$$

with

$$\Gamma_0(x) = \mathbb{E}\left[\partial\psi(Z, \beta_0)/\partial\beta' \middle| x\right]$$
$$\Omega_0(x) = \mathbb{V}(\psi(Z, \beta_0)|x) + [q_1(x; \beta_0) - q_0(x; \beta_0)][q_1(x; \beta_0) - q_0(x; \beta_0)]',$$

where $\mathbb{E}[A|c] = \mathbb{E}[A|C = c]$, $\mathbb{V}(A|c) = Var(A|C = c)$ and $q_j(x; \beta) = \mathbb{E}\left[\psi_j(Y_j, X, \beta)|x\right]$ for $j = 0, 1$.

Now consider the case where a random sequence from F_0 is unavailable. Instead only a 'selected' sequence of samples is available. Let D be a binary selection indicator. When $D = 1$ we observe Y_1 and X, when $D = 0$ we observe Y_0 and X; Y_1 and Y_0 are never simultaneously observed for a single unit. This paper considers estimation of β_0 under restriction (1) and the following additional assumptions.

Assumption 1.1 (Random Sampling) *$\{Z_i, D_i\}_{i=1}^{\infty}$ is an independent and identically distributed random sequence from F_0.*

Assumption 1.2 (Observed Data) *For each unit we observe X, D and $Y = (1 - D)Y_0 + DY_1$.*

Assumption 1.3 (Conditional Independence) $(Y_1, Y_0) \perp D|X$.

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

[2] Extending what follows to the overidentified case is straightforward.

Restriction (1) and Assumptions 1.1 to 1.4 constitute a semiparametric model for the data. Henceforth I refer to this model as the semiparametric missing data model or the missing at random (MAR) setup. The 'MAR setup' has been applied to a number of important econometric and statistical problems, including program evaluation as surveyed by Imbens (2004), non-classical measurement error (e.g., Robins, Hsieh and Newey 1995, Chen, Hong and Tamer 2005), missing regressors (Robins, Rotnitzky and Zhao 1994) and attrition in panel data (e.g., Wooldridge 2002). Chen, Hong and Tarozzi (2004) and Wooldridge (2007) discuss several other applications.

The maximal asymptotic precision with which β_0 can be estimated under the MAR setup has been characterized by Robins, Rotnitzky and Zhao (1994) and is given by the inverse of

$$\mathcal{I}(\beta_0) = \mathbb{E}[\Gamma_0(X)]' \mathbb{E}[\Lambda_0(X)]^{-1} \mathbb{E}[\Gamma_0(X)], \tag{3}$$

with

$$\Lambda_0(x) = \frac{\Sigma_0(x;\beta_0)}{1-p_0(x)} + \frac{\Sigma_1(x;\beta_0)}{p_0(x)} + [q_1(x;\beta_0) - q_0(x;\beta_0)][q_1(x;\beta_0) - q_0(x;\beta_0)]',$$

where $\Sigma_j(x;\beta_0) = \mathbb{V}(\psi_j(Y_j, X, \beta_0) | X = x)$ for $j = 0, 1$.

The associated efficient influence function (Bickel, Klassen, Ritov and Wellner 1993, Newey 1990) is given by

$$\begin{aligned} \phi(z,\theta_0) &= \mathbb{E}[\Gamma_0(X)]^{-1} \times \left\{ \frac{d}{p_0(x)}\psi_1(y_1,x,\beta_0) - \frac{1-d}{1-p_0(x)}\psi_0(y_0,x,\beta_0) \right. \\ &\quad \left. - \left\{ \frac{q_1(x;\beta_0)}{p_0(x)} + \frac{q_0(x;\beta_0)}{1-p_0(x)} \right\} (d - p_0(x)) \right\}. \end{aligned} \tag{4}$$

for $\theta = (p, q_0', q_1', \beta')'$.

The calculation of (3) is a now standard.[3] Knowledge of (3) is useful because it quantifies the cost – in terms of asymptotic precision – of the missing data[4] and because it can be used to verify whether a specific estimator for β_0 is efficient. To simplify what follows I will explicitly assume that $\mathcal{I}(\beta_0)$ is well-defined (i.e., that all its component expectations exist and are finite, and that all its component matrices are nonsingular).

Several globally efficient estimators for β_0 are available. The inverse probability weighting (IPW)

[3] An accessible derivation of this result can be found in Chen, Hong and Tarozzi (2004, Theorem 8).

[4] The difference $\Lambda_0(x) - \Omega_0(x)$ equals

$$\Sigma_0(x;\beta_0)\left\{\frac{p_0(x)}{1-p_0(x)}\right\} + \Sigma_1(x;\beta_0)\left\{\frac{1-p_0(x)}{p_0(x)}\right\} + 2\Sigma_{12}(x;\beta_0)$$

where $\Sigma_{01}(x;\beta_0) = \mathbb{C}(\psi_0(Y_0,X,\beta_0), \psi_1(Y_1,X,\beta_0) | X = x)$ with $\mathbb{C}(A, B|c) = Cov(A, B|C = c)$. This sum is bounded below by zero.

estimator with a nonparametrically estimated selection probability attains the bound (Hirano, Imbens and Ridder 2003). Cheng (1994), Hahn (1998), Chen, Hong and Tarozzi (2004, 2007) and Imbens, Newey and Ridder (2005) propose efficient imputation estimators. Locally efficient, in the sense defined by Newey (1990, p. 120), augmented inverse probability weighting (AIPW) estimators, which combine the weighting and imputation approaches along with auxiliary parametric assumptions, are common in the statistics literature (cf., Robins, Rotnitzky and Zhao 1994, Scharfstein, Rotnitzky and Robins 1999, Bang and Robins 2005, Tsiatis 2006). Wooldridge (2007) studies IPW estimation with parametrically estimated selection probabilities.

This paper shows that the semiparametric efficiency bound for β_0 under the MAR setup, coincides with the bound for a particular augmented moment condition problem. The augmented system consists of the inverse probability of observation weighted (IPW) original moment restriction (1) and an additional conditional moment restriction which exhausts all other implications of the MAR setup (useful for estimating β_0).[5] This general equivalence result is apparently new. Demonstrating equivalence provides fresh intuitions for several apparent 'paradoxes' in the missing data literature, including the well-known results that smoothness and exclusion priors on the propensity score do not increase the precision with which β_0 can be estimated (Robins, Hsieh and Newey 1995, Robins and Rotnitzky 1995, Hahn 1998, 2004) and that weighting by a nonparametric estimate of the propensity score results in an efficient estimator while weighting by the true propensity score does not (Hirano, Imbens and Ridder 2003, cf., Wooldridge 2007).

Equivalence also allows for a simple GMM characterization of the double robustness property of the class of locally efficient augmented inverse probability weighted (AIPW) estimators introduced by Robins, Rotnitzky and Zhao (1994). AIPW estimators require the analyst to make two auxiliary parametric assumptions, one about the form of the propensity score and the other about (certain features of) the conditional distribution of (Y_0, Y_1) given X. The AIPW estimate is semiparametrically efficient if the maintained auxiliary restrictions happen to hold in the population being sampled from but remains consistent if one or the other of the two parametric restrictions are violated (i.e., it is 'doubly robust'). Using the 'equivalent' GMM problem I show that this property follows from standard results on robustness of sequential GMM estimators to first step misspecification (e.g., Theorem 6.2 of Newey and McFadden (1994, p. 2180)).

Augmented inverse probability weighting is only efficient when the auxiliary restrictions imposed at the estimation stage are not viewed as part of the prior restriction. When they are viewed

[5] Similar 'equivalence' results hold in other well-known statistical models. I am grateful to Jinyong Hahn for providing the following simple but illuminating example. Consider the classical regression model where $Y = X'\beta_0 + \varepsilon$ with $\varepsilon | X \sim \mathcal{N}\left(0, \sigma_0^2\right)$; in that model the variance bound is given by the Cramer-Rao lower bound. Since the variance bound for the unconditional moment problem $\mathbb{E}\left[X\varepsilon\right] = 0$ coincides with that Cramer-Rao lower bound (when the homoscedasticity restriction is used to simplify the variance expression), we can conclude that the weaker moment restrictions 'exhausts' all the information content of the full model.

as part of the prior restriction they alter the efficiency bound. As is well-known the first set of auxiliary restrictions, those on the propensity score do not alter the bound, however those on the conditional distribution of (Y_0, Y_1) given X do so. This also paper analyzes the effect of imposing additional semiparametric restrictions on the conditional expectation functions (CEFs) $q_0(x;\beta) = \mathbb{E}[\psi_0(Y_0, X, \beta)|X = x]$ and $q_1(x;\beta) = \mathbb{E}[\psi_1(Y_1, X, \beta)|X = x]$. In the program evaluation context such restrictions are generated by semiparametric models for the CEFs of the two potential outcomes given covariates.

In an innovative paper, Wang, Linton and Härdle (2004) consider this problem with $\psi_1(Y_1, X, \beta) = Y_1 - \beta$ and $\psi_0(Y_0, X, \beta) = 0$ (i.e., they seek to estimate the marginal mean of an outcome which is missing at random). They impose a partial linear structure, as in Engle et al (1986), on $\mathbb{E}[Y_1|X]$. In making their variance bound calculation they assume that the conditional distribution of Y_1 given X is normal with a variance that does not depend on X. They do not provide a bound for the general case but conjecture that it is "very complicated" (p. 338).

By exploiting the insight that semiparametric restrictions on the forms of $q_0(x;\beta)$ and $q_1(x;\beta)$ simply add conditional moments to the 'equivalent' GMM problem I am able to apply Chamberlain's (1992a) methods to calculate the variance bound given the extra restrictions. Formally I derive the variance bound for the semiparametric problem defined by (1), Assumptions 1.1 to 1.4 and

Assumption 1.5 (Functional Restrictions) *Partition* $X = (X_1', X_2')'$, *then*

$$\begin{aligned}\mathbb{E}[\psi_0(Y_0, X, \beta_0)|X = x] &= q_0(x, \delta_{00}, h_{00}(x_2); \beta_0)\\ \mathbb{E}[\psi_1(Y_1, X, \beta_0)|X = x] &= q_1(x, \delta_{10}, h_{10}(x_2); \beta_0),\end{aligned}$$

where $q_0(x, \delta_0, h_0(x_2); \beta)$ *and* $q_1(x, \delta_1, h_1(x_2); \beta)$ *are known functions,* δ_0 *and* δ_1 *are* $J \times 1$ *finite dimensional unknown parameters, and* $h_0(\cdot)$ *and* $h_1(\cdot)$ *are unknown functions of* X_2.

To the best of my knowledge, the variance bound for this problem, the MAR setup with 'functional' restrictions, has not been previously calculated. The most relevant research is that of Wang, Linton and Härdle (2004). I am also aware of two additional related results. Robins, Mark and Newey (1992) consider the model with $\psi_1(Y_1, X, \beta) = Y_1$ and $\psi_0(Y_0, X, \beta) = Y_0 + \beta$ and the functional restrictions $\mathbb{E}[Y_0|X = x] = g_0(x)$ and $\mathbb{E}[Y_1|X = x] = \beta_0 + g_0(x)$ for $g_0(x)$ some smooth unknown function of X.[6] These additional restrictions are implied by a constant additive treatment effect (CATE) assumption. Under the CATE assumption they note that $\mathbb{E}[Y|X, D] = \beta_0 D + g_0(X)$ for $Y = DY_1 + (1 - D)Y_0$. This gives the partially linear model for which Chamberlain (1992a) derived the semiparametric efficiency bound.

[6] When evaluating the variance bound they also assume homoscedasticity, but this is not part of the prior restriction.

Hahn (2004), working with the same $\psi(Z, \beta)$ function partitions $X = (X_1', X_2')'$ and assumes that the two potential outcomes are independent of X_2 given X_1. This implies that Assumption 1.3 holds conditional on X_1 alone. The variance bound is therefore given by (3) with X_1 replacing X. The bounds calculated by Wang, Linton and Härdle (2004) and Robins, Mark and Newey (1992) are special cases of the one given in Theorem 5.1 below. Hahn's (2004) results, as they involve independence, as opposed to only mean independence, assumptions do not fit into my setup.

While it is relatively straightforward to construct globally efficient estimators for β_0 in the model which also imposes Assumption 1.5, it does not appear generally possible to construct estimators that are efficient while remaining robust its violation. The benefits of the extra semiparametric restrictions imposed by Assumption 1.5 must be weighed against the risk of misspecification.

Section 2 reports the first result of the paper: an equivalence between the 'MAR setup' and a particular method-of-moments problem. Specific examples of equivalence are discussed by Newey (1994a) and Hirano, Imbens and Ridder (2003). I discuss the connection between their results and the general result provided below. Section 3 uses the equivalent method-of-moments problem to develop alternative intuitions for the various 'puzzles' mentioned above. This section demonstrates the pedagogical value of the equivalence result. It provides a simple explanation for Hahn's (1998) finding that knowledge of the propensity score does not lower the information bound for the MAR problem, why the estimator of Hirano, Imbens and Ridder (2003) attains this bound and for the double robustness property of AIPW estimators. I also show that Wooldridge's (2007) three-step average treatment effect (ATE) estimator (cf., Hirano and Imbens 2002, Robins and Rotnitzky 1995) is asymptotically equivalent to Robins, Rotnitzky and Zhao's (1994) AIPW estimator.

Section 4 explores the implications of GMM equivalence for estimation. First, building on work by Robins, Rotnitzky and Zhao (1994) and Newey (1994a, Section 5.3), I propose a new globally efficient estimator for β_0. An advantage of the proposed estimator is that it allows for straightforward incorporation of smoothness and exclusion priors on the propensity score without sacrificing asymptotic efficiency; something that is not straightforward with other available efficient estimators. Second, for the known propensity score case, I show how the equivalent GMM problem suggests a simple way to modify the Horvitz-Thompson estimator so that it is semiparametrically efficient.

Section 5 calculates the variance bound for β_0 when the MAR setup is augmented by Assumption 1.5. It also (informally) discusses estimators which exploit Assumption 1.5. Requiring double robustness apparently precludes fully efficient estimation. Section 6 briefly discusses a related GMM equivalence result for a family of data combination problems. This family includes the average treatment effect on the treated (ATT) estimand studied by Hahn (1998). Section 7 summarizes and concludes with a discussion of open questions.

2 Equivalence result

Under Assumptions 1.1 to 1.4 the inverse probability of observation weighted moment condition

$$\mathbb{E}\left[\frac{D}{p_0(X)}\psi_1(Y_1, X, \beta_0) - \frac{1-D}{1-p_0(X)}\psi_0(Y_0, X, \beta_0)\right] = 0, \tag{5}$$

is valid (e.g., Hirano, Imbens and Ridder 2004, Wooldridge 2007). Under Assumptions 1.1 to 1.4 the conditional moment restriction

$$\mathbb{E}\left[\frac{D}{p_0(X)} - 1 \middle| X\right] = 0 \quad \forall \quad X \in \mathcal{X}, \tag{6}$$

also holds and nonparametrically identifies $p_0(x)$. While the terminology is inexact, in what follows I call (5) the *identifying moment* and (6) the *auxiliary moment.*

That the MAR setup is equivalent, with respect to the information it provides about β_0, to an augmented moment problem defined by restrictions (5) and (6) is implied by the following Theorem.

Theorem 2.1 (GMM Equivalence) *Suppose that (i) the distribution of Z has a known, finite support, (ii) there is some $\beta_0 \in \mathcal{B} \subset \mathbb{R}^K$ and $\rho_0 = (\rho_1, \ldots, \rho_L)'$ where $\rho_l = p_0(x_l) \in \mathcal{P} \subset [\kappa, 1-\kappa]$ for each $l = 1, \ldots, L$ and some $0 < \kappa < 1$ (with $\mathcal{X} = \{x_1, \ldots, x_L\}$ the known support of X) such that restrictions (5) and (6) hold, (iii) $\mathbb{E}[\Lambda_0(X)]$ and $\mathcal{I}(\beta_0) = \mathbb{E}[\Gamma_0(X)]'\mathbb{E}[\Lambda_0(X)]^{-1}\mathbb{E}[\Gamma_0(X)]$ are nonsingular with probability one and (iv) other regularity conditions hold (cf., Chamberlain 1992b, Section 2), then $\mathcal{I}(\beta_0)$ is the Fisher information bound for β_0.*

Proof. See Appendix A. ■

The proof of Theorem 2.1 involves only some tedious algebra and a straightforward application of Lemma 2 of Chamberlain (1987). Assuming that Z has known, finite support makes the problem fully parametric. The unknown parameters are the probabilities associated with each possible realization of Z, the values of the propensity score at each of the L mass points of the distribution of X, $\rho_0 = (\rho_1, \ldots, \rho_L)'$, and the parameter of interest, β_0.

The multinomial assumption is not apparent in the form of $\mathcal{I}(\beta_0)$, which involves only conditional expectations of certain functions of the data. This suggests that the bound holds in general since any F_0 which satisfies (5) and (6) can be arbitrarily well-approximated by a multinomial distribution also satisfying the restrictions. Chamberlain (1992a, Theorem 1) demonstrates that this is indeed the case. Therefore $\mathcal{I}(\beta_0)^{-1}$ is the maximal asymptotic precision with which β_0 can be estimated when the only prior restrictions on F_0 are (5) and (6). Since this variance bound coincides with (3) I conclude that (5) and (6) 'exhaust' all of the prior restrictions implied by the MAR setup (that are helpful for estimating β_0).

The connection between semiparametrically efficient estimation of moment condition models with missing data and 'augmented' systems of moment restrictions has been noted previously for the special case of data missing completely at random (MCAR) with $\psi(Z,\beta) = \psi_1(Y_1,X,\beta)$. In that case Assumptions 1.1 to 1.4 hold with $p_0(X)$ equal to a (sometimes known) constant. Newey (1994a) shows that an efficient estimate of β_0 can be based on the pair of moment restrictions

$$\mathbb{E}\left[D\psi_1(Y_1,X,\beta_0)\right], \qquad Cov\left(D, q_1(X;\beta_0)\right) = 0,$$

with $q_1(X;\beta)$ as defined above. Hirano, Imbens and Ridder (2003) discuss a related example with X binary and the data also MCAR. In their example efficient estimation is possible with only a finite number of unconditional moment restrictions (a construction that is also used in the proof of Theorem 2.1). Theorem 2.1 provides a formal generalization of the Newey (1994a) and Hirano, Imbens and Ridder (2003) examples to the missing at random (MAR) case.

Hirano, Imbens and Ridder (2003) emphasize connections between semiparametric efficiency and empirical likelihood estimation and it is illuminating to discuss these connections here as well. When Z is multinomial, as is assumed in Theorem 2.1, the results of Chamberlain (1987) imply that the GMM estimate of θ_0 is equivalent to the (constrained) maximum likelihood estimate, the Fisher information bound of which is provided by Theorem 2.1. Imbens (1997, Section 3) shows an equivalence of the constrained MLE with the empirical likelihood estimate when Z is multinomial. Therefore, for multinomial Z empirical likelihood estimation of β_0 is semiparametrically efficient. For continuous X efficient estimation is less straightforward due to the presence of the unknown function $p_0(x)$, but the equivalent method-of-moments problem nonetheless suggests a number of 'natural' estimators some of which are discussed below.

3 Some method-of-moments intuitions

The method-of-moments formulation provides a useful framework for understanding several apparent paradoxes found in the missing data literature. It clarifies why smoothness and exclusion priors on the propensity score do not lower the asymptotic variance bound for β_0 (Robins, Hsieh and Newey 1995, Hahn 1998, 2004) and why weighting by an estimate of the propensity score is typically more efficient than weighting by the true propensity score (Hirano, Imbens and Ridder 2003, Wooldridge 2007). The GMM structure also provides a simple intuition for the 'doubly robust' property of the class of AIPW estimators introduced by Robins, Rotnitzky and Zhao (1994).

Consider first the absence of any efficiency gain associated with imposing (valid) prior restrictions on the propensity score. An extreme example of such restrictions is complete knowledge of the propensity score. One might expect that such information, by eliminating sampling uncertainty in

the propensity score, would increase the precision with which it is possible to estimate β_0. That such information is not helpful in this way is somewhat puzzling. This result has proved particularly perplexing to practitioners as it apparently suggests that incorporation of prior knowledge regarding features of the selection probability is incompatible with efficient estimation (cf., Robins, Rotnitzky and Zhao 1994, pp. 854 - 855, Hahn 1998, pp. 324 - 325, Imbens 2004, pp. 16 - 17).

Under the conditions of Theorem 2.1 calculations provided in Appendix A imply that the GMM estimates of β_0 and ρ_0 (recall that ρ_0 contains the values for the propensity score at each of the mass points of the distribution of X) have an asymptotic sampling distribution of

$$\sqrt{N}\left(\begin{bmatrix}\widehat{\rho}\\ \widehat{\beta}\end{bmatrix}-\begin{bmatrix}\rho_0\\ \beta_0\end{bmatrix}\right)\xrightarrow{D}\mathcal{N}\left(\begin{bmatrix}0\\ 0\end{bmatrix},\begin{bmatrix}\mathcal{I}_M(\rho_0)^{-1} & 0\\ 0 & \mathcal{I}_M(\beta_0)^{-1}\end{bmatrix}\right),$$

with $\mathcal{I}(\beta_0)$ as defined in (3) and $\mathcal{I}(\rho_0)$ as defined in Appendix A.

That knowledge of various features of the propensity score does not alter the efficiency bound is thus a consequence of information matrix block diagonality between β_0 and ρ_0. As is well-known, under block diagonality sampling error in $\widehat{\rho}$ does not affect, at least to first order, the asymptotic sampling properties of $\widehat{\beta}$. While block diagonality is formally only a feature of the multinomial approximation to the true data generating process, the result nonetheless provides a useful intuition for understanding why prior knowledge of the propensity score is not valuable asymptotically.

Analogous results hold for other important GMM problems. For example, in unconditional moment problems, the infeasible estimator which sets an optimal linear combination of the sample moments equal to zero is first-order equivalent to the feasible one based on setting a noisy estimate of that linear combination equal to zero. It is well-known, however, that this largely reflects the limits of the usual asymptotic approximation: weight matrix estimation does affect the higher-order properties of GMM estimates (Newey and Smith 2004). Ichimura and Linton (2005) show that the effects of propensity score estimation do show-up in the 'second-order' asymptotics of the Hirano, Imbens and Ridder (2003) estimate of β_0. This suggests that the ability to incorporate smoothness and exclusion priors on the propensity score, while at the same time maintaining (first order) efficiency, is likely to result in an estimator with superior small-sample performance.

A related puzzle is the finding of Hirano, Imbens and Ridder (2003) and Wooldridge (2007) that weighting by the true propensity score is typically inefficient.[7] The true weights estimator is the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{p_0(X_i)}\psi_1(Y_{1i},X_i,\widehat{\beta})-\frac{1-D_i}{1-p_0(X_i)}\psi_0(Y_{0i},X_i,\widehat{\beta})=0. \tag{7}$$

[7] Versions of this puzzle are actually quite old. For example it shows up in the choice-based sampling literature (cf., Cosslett 1981). Imbens (1992) provides method of moments intuition for that case (cf., Wooldridge 1999a).

Inefficiency of the true weights estimator stems from its failure to impose the additional restrictions given by (6). These additional restrictions are valuable because they can be used to reduce sampling variance in (5). One approach to efficient estimation in this case would be to consider a joint GMM estimator which chooses $\widehat{\beta}$ such that sample analogs of (5) and (6) hold simultaneously. An alternative approach, which is practically and pedagogically useful, is to, following Newey (1994a), modify (6) so that it is conditionally uncorrelated with $D/p_0(X)-1$ given X. This is accomplished by choosing $\widehat{\beta}$ to set the sample mean of the (population) residuals associated with the conditional linear predictor of $D\psi(Z,\beta_0)/p_0(X)$ given $D/p(X)-1$ conditional on X equal to zero. Such a $\widehat{\beta}$ solves

$$\frac{1}{N}\sum_{i=1}^{N} s(Z_i, p_0(X_i), \widehat{q}_0(X_i;\widehat{\beta}), \widehat{q}_1(X_i;\widehat{\beta}), \widehat{\beta}) = 0, \tag{8}$$

where $\widehat{q}_0(X_i;\beta)$ and $\widehat{q}_1(X_i;\beta)$ are some nonparametric estimates and $s(Z,p,q_0,q_1,\beta)$ is given by

$$\begin{aligned} s(Z,p,q_0,q_1,\beta) &= \frac{D}{p(X)}\psi(Z,\beta) \\ &\quad -\mathbb{E}^*\left[\frac{D}{p(X)}\psi_1(Y_1,X,\beta) - \frac{1-D}{1-p(X)}\psi_0(Y_0,X,\beta)\middle|\frac{D}{p(X)}-1;X\right] \\ &= \frac{D}{p(X)}\psi_1(Y_1,X,\beta) - \frac{1-D}{1-p(X)}\psi_0(Y_0,X,\beta) \\ &\quad -\left[\frac{q_1(X;\beta)}{p(X)} + \frac{q_0(X;\beta)}{1-p(X)}\right](D-p_0(X)), \end{aligned}$$

which is identical to the efficient score function for β_0 (the notation $\mathbb{E}^*[Y|X;Z]$ denotes the (mean squared error minimizing) linear predictor of Y given X within a subpopulation homogenous in Z).[8] Efficiency of this estimator follows from the fact that it incorporates all the information about β_0 contained in the auxiliary conditional moment restriction by using it to reduce the sampling variance in the identifying moment (since $s(Z,p,q_0,q_1,\beta_0)$ and $D/p(X)-1$ are conditionally uncorrelated by construction). Inefficiency of (7) follows from its failure to exploit this information.

The GMM formulation also provides a simple intuition for why replacing $p_0(X_i)$ in (7) with a nonparametric estimate of the propensity score is efficient. Hirano, Imbens and Ridder (2003) propose approximating the propensity score by $p\left(R^M(X_i)'\pi_M\right)$ where $p(\cdot)$ is the logistic CDF and $R^M(X_i)$ is a vector of series terms, the length of which is indexed by M. They show that the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{p\left(R^M(X_i)'\widehat{\pi}_M\right)}\psi_1(Y_{1i},X_i,\widehat{\beta}) - \frac{1-D_i}{1-p\left(R^M(X_i)'\widehat{\pi}_M\right)}\psi_0(Y_{0i},X_i,\widehat{\beta}) = 0,$$

[8] Wooldridge (1999b, Section 4) collects some useful results on conditional linear predictors.

attains the semiparametric efficiency bound when M grows with the N at a certain rate, $\widehat{\pi}_M$ is estimated by a logit procedure and other technical conditions hold. One intuition for their efficiency result follows from the fact that any given estimator in their sequence is numerically identical to the unconditional joint GMM estimator which chooses $\widehat{\beta}$ and $\widehat{\pi}_M$ to solve

$$\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} m_1(Z_i,\widehat{\pi}_M) \\ m_2(Z_i,\widehat{\beta},\widehat{\pi}_M) \end{pmatrix} = 0,$$

where

$$m_1(Z,\pi_M) = \left(\frac{D}{p\left(R^M(X)'\pi_M\right)} - 1\right)\frac{R^M(X)}{p\left(R^M(X)'\pi_M\right)}$$
$$m_2(Z,\beta,\pi_M) = \frac{D}{p\left(R^M(X)'\pi_M\right)}\psi_1(Y_1,X,\beta) - \frac{D}{1-p\left(R^M(X)'\pi_M\right)}\psi_0(Y_0,X,\beta).$$

Replacing $p_0(X_i)$ with the sequence of estimated propensity scores $p\left(R^M(X_i)'\widehat{\pi}_M\right)$ is therefore numerically equivalent to solving a joint method-of-moments problem that imposes both the identifying moment (5) and a sequence of unconditional moment restrictions that, in large enough samples, are equivalent to the conditional restriction given in (6) above. Note that weighting by a parametric estimate of the propensity score, as in Wooldridge (2007), is generally inefficient: the MLE first order conditions do not (asymptotically) impose all the restrictions implied by (6).[9]

The equivalent GMM problem also provides a simple explanation for the 'doubly robust' property of the class of AIPW estimations introduced by Robins, Rotnitzky and Zhao (1994). These estimators are motivated by the structure of the efficient score. A common method of constructing globally efficient estimators is to form a M-estimator based on a nonparametric estimate of the efficient score (cf., Newey 1990). Robins, Rotnitzky and Zhao (1994, cf., Equations 9 and 15) instead propose a M-estimator based on a parametric estimate of the efficient score. Their method requires making auxiliary assumptions about the form of the propensity score as well as the conditional distributions of Y_0 and Y_1 given X. Under these maintained extra assumptions their estimator attains the semiparametric efficiency bound. Scharfstein, Rotnitzky and Robins (1999) show that AIPW remains consistent if either, but not both, of the auxiliary parametric models are misspecified (although it is not efficient in such cases). Tsiatis (2006) provides a survey of research on double robustness in missing data models.

Assume, possibly incorrectly, that both the propensity score and the conditional densities of Y_0 and Y_1 given X belong to known parametric families, $p_0(X) = p(X,\gamma_0)$, $f_0(Y_0|X) = f_0(Y_0|X;\delta_{00})$, and $f_1(Y_1|X) = f_1(Y_1|X;\delta_{10})$ indexed by the unknown finite dimensional parameters γ and δ_0 and

[9] Wooldridge's (2007) estimator is generally more efficient than the one which uses the true propensity score.

δ_1. Let $\widehat{\gamma}$, $\widehat{\delta}_0$ and $\widehat{\delta}_1$ denote MLEs (the latter two computed using the $D=0$ and $D=1$ subsamples respectively) and define

$$q_j(x, \widehat{\delta}_j; \beta) = \int \psi_j(y_j, x, \beta) f_j(y_j|x; \widehat{\delta}_j) \mathrm{d}y_j, \qquad j = 0, 1.$$

The 'doubly robust' AIPW estimate of β_0 solves

$$\frac{1}{N} \sum_{i=1}^{N} s(Z_i, \widehat{\beta}, p\left(X_i, \widehat{\gamma}\right), q_0(X_i, \widehat{\delta}_0; \widehat{\beta}), q_1(X_i, \widehat{\delta}_1; \widehat{\beta})) = 0. \tag{9}$$

An important application of (9) arises in the causal inference literature when $\psi_1(Y_1, X, \beta) = Y_1$ and $\psi_0(Y_0, X, \beta) = Y_0 + \beta$. In that case calculating $q_1(x, \widehat{\delta}_1; \beta) = \int y_1 f(y_1|x; \widehat{\delta}_1) \mathrm{d}y_1$ amounts to specifying and estimating a parametric model for $\mathbb{E}\left[Y_1|X=x\right]$ and similarly for $q_0(x, \widehat{\delta}_0; \beta)$ (Lunceford and Davidian 2004, Bang and Robins 2005).

To understand the robustness properties of AIPW estimators it is helpful to first consider an M-estimator for β_0 based on a nonparametric estimate of the efficient score. That is, choose $\widehat{\beta}$ to solve

$$\frac{1}{N} \sum_{i=1}^{N} s(Z_i, \widehat{\beta}, \widehat{p}\left(X_i\right), \widehat{q}_0(X_i; \widehat{\beta}), \widehat{q}_1(X_i; \widehat{\beta})) = 0, \tag{10}$$

where $\widehat{p}\left(X_i\right)$, $\widehat{q}_0(X_i; \beta)$ and $\widehat{q}_1(X_i; \beta)$ are nonparametric estimates.

From Newey (1994b, Proposition 3, p. 1360), it follows that nonparametric estimation of the propensity score does not affect the asymptotic variance of $\widehat{\beta}$ since

$$\mathbb{E}\left[\left. \frac{\partial s\left(Z, \beta_0, p_0\left(X\right), q_0\left(X; \beta_0\right), q_1\left(X; \beta_0\right)\right)}{\partial p_0} \right| X \right] = 0. \tag{11}$$

Similarly, nonparametric estimation of $q_j(X_i; \beta)$ for $j = 0, 1$ does not effect the asymptotic variance of $\widehat{\beta}$ since

$$\mathbb{E}\left[\left. \frac{\partial s\left(Z, \beta_0, p_0\left(X\right), q_0\left(X; \beta_0\right), q_1\left(X; \beta_0\right)\right)}{\partial q_j} \right| X \right] = 0, \qquad j = 0, 1. \tag{12}$$

In the parametric GMM context zeroness of the expectation of the derivative of a moment with respect to an estimated nuisance parameter is associated with robustness to 'first step' mis-specification. A similar property holds here. Consider the solution to (10) where the consistent nonparametric estimate $\widehat{p}\left(X_i\right)$ is replaced with the inconsistent parametric estimate $p\left(X_i, \widehat{\gamma}\right)$. Let γ_* be the limiting value of $\widehat{\gamma}$; although $p\left(X, \gamma_*\right) \neq p_0\left(X\right)$ this estimator remains consistent for β_0

since, by iterated expectations,

$$\begin{aligned}\mathbb{E}\left[s\left(Z,\beta_0,p\left(X,\gamma_*\right),q_0\left(X;\beta_0\right),q_1\left(X;\beta_0\right)\right)\right] &= \mathbb{E}\left[q_1\left(X;\beta_0\right)-q_0\left(X;\beta_0\right)\right] \\ &= \mathbb{E}\left[\psi(Z,\beta_0)\right]=0.\end{aligned}$$

Now replace $\widehat{q}_j(X_i;\beta)$ in (10) with $q_j(X_i,\widehat{\delta}_j;\beta)$ and assume that $f_j\left(Y_j|X\right)$ does not belong to the presumed parametric family. Analogously to the misspecified propensity score case we have $q_j(x,\delta_{j*};\beta)\neq q_j\left(x;\beta\right)$ for $j=0,1$. Nonetheless $\widehat{\beta}$ remains consistent for β_0 since

$$\mathbb{E}\left[s\left(Z,\beta_0,p_0\left(X\right),q_0\left(X,\delta_{0*};\beta_0\right),q_1\left(X,\delta_{1*};\beta_0\right)\right)\right]=0$$

for any $q_j(x,\delta_{j*};\beta)$. The above arguments demonstrates that the 'doubly robust' property of AIPW estimators follows from standard results on robustness of sequential GMM estimators to first-stage misspecification (e.g., Newey and McFadden 1994, Theorem 6.2).

Wooldridge (2007) has recently suggested a doubly robust alternative to AIPW estimation of the average treatment effect (cf., Hirano and Imbens 2001, Robins and Rotnitzky 1995). Wooldridge (2007) does not provide distribution theory for his estimator, here I note that his estimator is locally efficient.

Define $\theta=\left(\gamma_0',\delta_0',\delta_1',\beta\right)'$, with β_0 equal to the ATE; Wooldridge's sequential estimator is equivalent to joint GMM applied to the moment function

$$\psi\left(Z,\theta\right)=\left\{\begin{array}{c}\frac{D-p(X,\gamma)}{p(X,\gamma)[1-p(X,\gamma)]}\frac{\partial p(X,\gamma)}{\partial\gamma}\\ \frac{1-D}{1-p(X,\gamma)}X(Y_0-X'\delta_0)\\ \frac{D}{p(X,\gamma)}X(Y_1-X'\delta_1)\\ \beta-X'\left(\delta_1-\delta_0\right)\end{array}\right\}.$$

The maintained assumptions are that $p_0\left(X\right)=p\left(X,\gamma_0\right)$, $\mathbb{E}\left[Y_0|\,X\right]=X'\delta_{00}$ and $\mathbb{E}\left[Y_1|\,X\right]=X'\delta_{10}$. Under these assumptions it is straightforward to show that the GMM estimate of $\widehat{\beta}$ has an asymptotic sampling variance of

$$\mathbb{E}\left[\frac{\sigma_0^2\left(X\right)}{1-p\left(X,\gamma_0\right)}\right]+\mathbb{E}\left[\frac{\sigma_1^2\left(X\right)}{p\left(X,\gamma_0\right)}\right]+\left(\delta_1-\delta_0\right)'Var\left(X\right)\left(\delta_1-\delta_0\right),$$

where $\sigma_j^2\left(X\right)=\mathbb{V}\left(Y_j|\,X\right)$ for $j=0,1$. Wooldridge's (2007) estimator is thus locally efficient. It is attractive relative to AIPW since it can be computed in a small number of steps using standard software.

4 New approaches to estimation

This section proposes two new estimators for β_0 (under restriction (1) and Assumptions 1.1 to 1.4) suggested by the equivalent GMM problem. The first proposal builds on work by Newey (1994a). Its large sample properties can be rigorously derived using his results. For this reason I do not specify the detailed regularity conditions needed for a formal demonstration of its consistency and asymptotic normality. Second, for the known propensity score case, I suggest a simple and efficient modification of the Horvitz-Thompson estimator.[10]

The first proposed estimate of β_0 is an M-estimate based on a nonparametric estimate of the efficient score (10), where $\widehat{p}(X_i)$, $\widehat{q}_0(X_i;\beta)$ and $\widehat{q}_1(X_i;\beta)$ are nonparametric series estimates. Let $R^{M_k}(X) = (r_1(X), \ldots, r_{M_k}(X))'$ be the first M_k terms in a sequence of approximating functions (e.g., a power series in a one-to-one bounded transformation of X). Let $M = M_0 + M_1 + \ldots + M_{2K}$ and define

$$\underset{M\times(1+2K)}{R_i} = \begin{pmatrix} R^{\widehat{M}_0}(X_i) & 0 & \cdots & 0 \\ 0 & R^{\widehat{M}_1}(X_i) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & R^{\widehat{M}_{2K}}(X_i) \end{pmatrix}$$

with $\widehat{M}_k$ for $k = 0, \ldots, 2K$ denoting the (possibly) data-dependent number of approximating terms in each column of R_i. Define the $1 + 2K$ row vector

$$W_i(\beta) = \left(D_i, (1-D_i)\psi_0(Y_{0i}, X_i, \beta)', D_i\psi_1(Y_{1i}, X_i, \beta)'\right).$$

A series estimate of $h(X_i,\beta) = (h_1(X_i), h_2(X_i,\beta), h_3(X_i,\beta)) = \mathbb{E}[W_i(\beta)|X_i]$ is given by

$$\widehat{h}(X_i,\beta) = (\tau_N(\widetilde{h}_1(X_i)), \widetilde{h}_2(X_i,\beta), \widetilde{h}_3(X_i,\beta))$$

where $\tau_N(\cdot)$ is a trimming function which ensures that $\widehat{h}_1(X_i)$ lies in the $[\kappa, 1-\kappa]$ interval and

$$\widetilde{h}(X_i,\beta_i) = \widehat{\Pi}R_i, \qquad \widehat{\Pi} = \left(\frac{1}{N}\sum_{i=1}^{N} W_i(\beta) R_i'\right)\left(\frac{1}{N}\sum_{i=1}^{N} R_i R_i'\right)^{-},$$

with $(\cdot)^{-}$ denoting a generalized inverse.

The nonparametric series estimates of the propensity score and the conditional means of the two

[10] I thank Guido Imbens for suggesting that I attempt to develop a simple modification of the Horvitz-Thompson estimator that is efficient in the known propensity score case.

parts of the moment function are then given by

$$\widehat{p}(X_i) = \widehat{h}_1(X_i), \qquad \widehat{q}_0(X_i;\beta) = \frac{\widehat{h}_2(X_i,\beta)'}{\widehat{h}_1(X_i)}, \qquad \widehat{q}_1(X_i;\beta) = \frac{\widehat{h}_3(X_i,\beta)'}{\widehat{h}_1(X_i)},$$

with $\widehat{\beta}$ found by solving (10).[11]

A consistent estimate of the large sample variance of $\widehat{\beta}$ is given by $(\widehat{\Gamma}'\widehat{\Sigma}^{-1}\widehat{\Gamma})^{-1}/N$ with

$$\widehat{\Gamma} = \frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\widehat{p}(X_i)} \frac{\partial \psi(Z_i,\widehat{\beta})}{\partial \beta'}, \qquad \widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \widehat{s}_i \widehat{s}_i',$$

where $\widehat{s}_i = s(Z_i, \widehat{\beta}, \widehat{p}(X_i), \widehat{q}_0(X_i;\widehat{\beta}), \widehat{q}_1(X_i;\widehat{\beta}))$.

A simpler estimator would use the same number of series terms to approximate each of the $1+2K$ elements of $h(X_i,\beta)$. To understand why the increased generality might be useful consider the case where $\psi_0(Y_0, X, \beta) = Y_0 + \beta$ and $\psi_1(Y_1, X, \beta) = Y_1$. In that case $h_1(X) = p_0(X)$, $h_2(X,\beta) = (1 - p_0(X))\{\mathbb{E}[Y_0|X] + \beta\}$ and $h_3(X,\beta) = p_0(X)\,\mathbb{E}[Y_1|X]$. The finite sample properties of $\widehat{\beta}$ might improve if the number of series terms used to approximate $h_1(X_i)$, $h_2(X_i,\beta)$ and $h_2(X_i,\beta)$ respectively reflect the amount of smoothness in the propensity score and the degree of 'response heterogeneity' in the control and active treatments (as opposed to using an equally rich approximation for all three CEFs). In contrast, the estimator of Hirano, Imbens and Ridder (2003) effectively requires overfitting of the propensity score if $q_0(X;\beta)$ and/or $q_1(X;\beta)$ vary sharply in X, while the imputation estimators of Chen, Hong and Tarozzi (2004) or Imbens, Newey and Ridder (2005) provide no mechanisms for incorporating a correctly smoothed estimate of the propensity score.[12]

In some cases prior knowledge of the selection probability might be particularly sharp. If β_0 is the ATE and the data are generated by a randomized experiment then the propensity score is known. Another case where the propensity score is known is M-estimation under variable probability sampling with known retention frequencies as in Wooldridge (1999a, 2007). In other situations we may be willing to assume that $p_0(X) = p(X_1, \gamma_0)$ for some known function $p(\cdot,\cdot)$, unknown finite dimensional parameter γ_0 and X_1 a subvector of X. Let $\widehat{\gamma}$ be the MLE of γ_0 and $\widehat{q}_0(X_i;\beta)$ and $\widehat{q}_1(X_i;\beta)$ as estimated above (in particular we require that the denominators of $\widehat{q}_0(X_i;\beta)$ and

[11] An alternative estimator would replace $\widehat{h}_1(X_i)$ with the series logit estimator (SLE) of Hirano, Imbens and Ridder (2003).

[12] Imbens, Newey and Ridders' (2005) results do imply that such information would help determine the optimal (in a MSE sense) number of approximating terms to use when estimating $q_0(X;\beta)$ and $q_1(X;\beta)$.

$\widehat{q}_1(X_i;\beta)$ use the nonparametric series estimate of the propensity score), then

$$\frac{1}{N}\sum_{i=1}^{N} s(Z_i, \widehat{\beta}, p\left(X_{1i}, \widehat{\gamma}\right), \widehat{q}_0(X_i;\widehat{\beta}), \widehat{q}_1(X_i;\widehat{\beta})) = 0, \tag{13}$$

is locally efficient for β_0 and consistent irrespective of whether or not $p_0\left(X\right) = p\left(X_1, \gamma_0\right)$ for some γ_0 (for the known propensity score case we simply replace $p\left(X_i, \widehat{\gamma}\right)$ with the true propensity score in (13)). Like the 'doubly robust' estimator of Robins, Rotnitzky and Zhao (1994) semiparametric efficiency of $\widehat{\beta}$ requires that the imposed restrictions on the propensity score are correct. However, unlike their estimator, semiparametric efficiency and consistency holds without having to correctly specify parametric forms for $q_0\left(X;\beta\right)$ and $q_1\left(X;\beta\right)$.

My second proposal is a simple modification of the Horvitz-Thompson estimator. This estimator replaces the empirical measure used by the Horvitz-Thompson estimator with an estimated measure which satisfies a sequence of unconditional restrictions implied by (6). Let $a_{Mi} = \left(D_i/p_0\left(X_i\right) - 1\right) R^M\left(X_i\right)$, $\overline{a}_M = \sum_{i=1}^{N} a_{Mi}/N$, $B_M = \sum_{i=1}^{N} a_{Mi}a_{Mi}'/N$ and define the (closed-form) weights

$$\omega_{Mi} = \frac{1 - \overline{a}_M' B_M^{-1} a_{Mi}}{\sum_{i=1}^{N} 1 - \overline{a}_M' B_M^{-1} a_{Mi}}, \qquad i = 1, \ldots, N,$$

then let $\widehat{\beta}$ be the solution to

$$\sum_{i=1}^{N} \omega_{Mi} \left\{ \frac{D_i}{p_0\left(X_i\right)} \psi_1(Y_{1i}, X_i, \widehat{\beta}) - \frac{1 - D_i}{1 - p_0\left(X_i\right)} \psi_0(Y_{0i}, X_i, \widehat{\beta}) \right\} = 0.$$

If M grows with N at the appropriate rate then $\widehat{\beta}$ will attain the semiparametric efficiency bound (under appropriate regularity conditions). The intuition for this claim is that $\widehat{\beta}$ is equivalent to the GMM estimate where the 'known weights' identifying moment (5) is augmented by the additional M unconditional moment restrictions (cf., Brown and Newey 1998)

$$\mathbb{E}\left[\left(\frac{D}{p_0\left(X\right)} - 1\right) R^M\left(X\right)\right] = 0.$$

If M grows with N then this sequence of unconditional auxiliary moment restrictions will be equivalent to the conditional moment restriction (6) in large samples and $\widehat{\beta}$ should be efficient. Any GMM software that accepts user weights can be used to implement this estimator. If $\psi_1\left(Y_1, X, \beta\right) = Y_1 - \beta$ and $\psi_0\left(Y_0, X, \beta\right) = 0$, then $\widehat{\beta}$ exists in closed-form as

$$\widehat{\beta} = \sum\nolimits_{i=1}^{N} \omega_{Mi} \frac{D_i}{p_0\left(X_i\right)} Y_{1i}, \tag{14}$$

which is a weighted version of the Horvitz-Thompson estimator of the marginal mean of Y_1. Observe that (14) replaces the empirical distribution function with the $\widehat{F}_{\omega_M}(z) = \sum_{i=1}^{N} \omega_{Mi} \mathbf{1}(Z_i \leq z)$. An important feature of $\widehat{F}_{\omega_M}(z)$ is that

$$\int \frac{d}{p_0(x)} \mathrm{d}\widehat{F}_{\omega_M}(z) = 1, \qquad \int \frac{d}{p_0(x)} R^M(x) \, \mathrm{d}\widehat{F}_{\omega_M}(z) = \int R^M(x) \, \mathrm{d}\widehat{F}_{\omega_M}(z).$$

Replacing the empirical distribution function with $\widehat{F}_{\omega_M}(z)$ ensures that the inverse probability weights, $D_i/p_0(X_i)$, sum to one and also that the inverse probability weighted mean of $R^M(X_i)$ in the $D_i = 1$ subsample equals its mean over the entire sample. Intuitively the 'balancing' property of the propensity score holds in the reweighted sample. For example if $R^M(X_i) = \left(1, X_i, \ldots, X_i^{M-1}\right)$, then using $\widehat{F}_{\omega_M}(z)$ ensures that the first $M-1$ inverse probability weighted sample moments of X_i in the $D_i = 1$ subsample equal their overall sample moments. Closely related estimators, including ones appropriate for the unknown propensity score case, are developed fully in Egel, Graham and Pintos (2007).[13] They also develop connections with empirical likelihood (e.g., Imbens 1997). The main point I want to emphasize here is how the equivalent GMM problem suggests new and conceptually simple approaches to efficient estimation of β_0 in the semiparametric missing data problem.

5 Semiparametric functional restrictions

Consider the MAR setup augmented by Assumption 1.5. To the best of my knowledge, the maximal asymptotic precision with which β_0 can be estimated in this model has not been previously characterized. In this section I exploit the fact that Assumption 1.5 simply adds the two conditional moment restrictions

$$\begin{aligned} \mathbb{E}\left[\psi_0(Y_0, X, \beta_0) - q_0(X, \delta_{00}, h_{00}(X_2); \beta_0) \,|X\right] &= 0 \\ \mathbb{E}\left[\psi_1(Y_1, X, \beta_0) - q_1(X, \delta_{10}, h_{10}(X_2); \beta_0) \,|X\right] &= 0, \end{aligned} \tag{15}$$

to the equivalent GMM problem (defined by (5) and (6)). I then apply Chamberlain's (1992a) approach to the new problem defined by (5), (6) and (15) to calculate the variance bound for β_0.

[13] In fact, replacing $p_0(X_i)$ with a MLE in the above procedure is semiparametrically efficient.

For $j = 0, 1$ let $q_j(X) = q_j(X, \delta_{j0}, h_{j0}(X_2); \beta_0)$ and $\Sigma_j(X) = \mathbb{V}(\psi_j(Y_j, X, \beta_0) \big| X)$ define

$$\underset{K\times 1}{\Delta_{h_j}(X_2)} = \mathbb{E}\left[\left.\frac{\partial q_j(X)}{\partial h_j'}\right| X_2\right], \qquad \underset{K\times J}{\Delta_{\delta_j}(X_2)} = \mathbb{E}\left[\left.\frac{\partial q_j(X)}{\partial \delta_j'}\right| X_2\right]$$

$$\underset{1\times 1}{\Upsilon_{h_j}(X_2)} = \mathbb{E}\left[\left(\frac{\partial q_j(X)}{\partial h_j'}\right)'\left[\frac{\Sigma_j(X)}{(1-j)(1-p(X))+jp(X)}\right]^{-1}\left(\frac{\partial q_j(X)}{\partial h_j'}\right)\middle| X_2\right]$$

$$\underset{J\times J}{\Upsilon_{\delta_j}(X_2)} = \mathbb{E}\left[\left(\frac{\partial q_j(X)}{\partial \delta_j'}\right)'\left[\frac{\Sigma_j(X)}{(1-j)(1-p(X))+jp(X)}\right]^{-1}\left(\frac{\partial q_j(X)}{\partial \delta_j'}\right)\middle| X_2\right]$$

$$\underset{1\times J}{\Upsilon_{h_j\delta_j}(X_2)} = \mathbb{E}\left[\left(\frac{\partial q_j(X)}{\partial h_j'}\right)'\left[\frac{\Sigma_j(X;\beta_0)}{(1-j)(1-p(X))+jp(X)}\right]^{-1}\left(\frac{\partial q_j(X)}{\partial \delta_j'}\right)\middle| X_2\right],$$

and

$$A_j(X_2) = \Delta_{h_j}(X_2)\,\Upsilon_{h_j}(X_2)^{-1}\,\Delta_{h_j}(X_2)'$$
$$B_j(X_2) = \Delta_{\delta_j}(X_2) - \Delta_{h_j}(X_2)\,\Upsilon_{h_j}(X_2)^{-1}\,\Upsilon_{h_j\delta_j}(X_2)$$
$$C_j(X_2) = \Upsilon_{\delta_j}(X_2) - \Upsilon_{h_j\delta_j}(X_2)'\,\Upsilon_{h_j}(X_2)^{-1}\,\Upsilon_{h_j\delta_j}(X_2)$$

and

$$\begin{aligned}\Xi = \mathbb{E}\left[(q_1(X) - q_0(X))(q_1(X) - q_0(X))'\right] &\\ + \mathbb{E}\left[A_0(X_2)\right] + \mathbb{E}\left[B_0(X_2)\right]\mathbb{E}\left[C_0(X_2)\right]^{-1}\mathbb{E}\left[B_0(X_2)\right]' &\\ + \mathbb{E}\left[A_1(X_2)\right] + \mathbb{E}\left[B_1(X_2)\right]\mathbb{E}\left[C_1(X_2)\right]^{-1}\mathbb{E}\left[B_1(X_2)\right]'.&\end{aligned}$$

The variance bound for β_0 is given in the following theorem.

Theorem 5.1 (Efficiency with Semiparametric Functional Restrictions) *Suppose that (i) the distribution of Z has a known, finite support, (ii) there is some $\beta_0 \in \mathcal{B} \subset \mathbb{R}^K$, $\rho_0 = (\rho_1, \ldots, \rho_L)'$ where $\rho_l = p_0(x_l) \in \mathcal{P} \subset [\kappa, 1-\kappa]$ for each $l = 1, \ldots, L$ and some $0 < \kappa < 1$ (with $\mathcal{X} = \{x_1, \ldots, x_L\}$ the known support of X), $\delta_{j0} \in \mathcal{D}_j \subset \mathbb{R}^J$ for $j = 0, 1$ and $h_{j0}(x_{2,m}) = \lambda_{0,m}$ $\in \mathcal{L} \subset \mathbb{R}^1$ for $j = 0, 1$ and each $m = 1, \ldots, M$ (with $\mathcal{X}_2 = \{x_{2,1}, \ldots, x_{2,M}\}$ the known support of X_2) such that restrictions (5), (6) and (15) hold, (iii) Ξ_0 and*

$$\mathcal{I}^f(\beta_0) = \mathbb{E}\left[\Gamma_0(X)\right]'\Xi_0^{-1}\mathbb{E}\left[\Gamma_0(X)\right]$$

are nonsingular with probability one and (iv) other regularity conditions hold (cf., Chamberlain 1992b, Section 2), then $\mathcal{I}^f(\beta_0)$ is the Fisher information bound for β_0.

Proof. See Appendix A. ■

The following Corollary gives the bound when $q_0(X;\beta_0)$ and $q_1(X;\beta_0)$ are parametrically specified.

Corollary 5.1 (Efficiency with Parametric Functional Restrictions) *When $q_j(X;\beta_0) = q_0(X,\delta_{j0};\beta_0)$ for some $\delta_{j0} \in \mathcal{D}_j \subset \mathbb{R}^J$, $j = 0,1$ and the other assumptions of Theorem 5.1 are satisfied then*

$$\begin{aligned}\mathcal{I}^f(\beta) = \mathbb{E}[\Gamma_0(X)]' \Big[\mathbb{E}\left[(q_1(X)-q_0(X))(q_1(X)-q_0(X))'\right] \\ + \mathbb{E}[\Delta_{\delta_1}(X)]\,\mathbb{E}[\Upsilon_{\delta_1}(X)]^{-1}\,\mathbb{E}[\Delta_{\delta_1}(X)]' \\ + \mathbb{E}[\Delta_{\delta_0}(X)]\,\mathbb{E}[\Upsilon_{\delta_0}(X)]^{-1}\,\mathbb{E}[\Delta_{\delta_0}(X)]'\Big]^{-1}\mathbb{E}[\Gamma_0(X)]\end{aligned}$$

is the Fisher information bound for β_0.

The general expression for $\mathcal{I}^f(\beta_0)$ is admittedly unwieldy, however it is interpretable for some important special cases. A leading example is when β_0 equals the ATE so that $\psi_1(Y_1,X,\beta) = Y_1$ and $\psi_0(Y_0,X,\beta) = Y_1+\beta$ and the CEFs of Y_0 and Y_1 take a partially linear structure. If the variance of Y_0 and Y_1 are both constant in X (but homoscedasticity is not part of the prior restriction) and we define $e_0(X_2) = \Pr(D=1|X_2)$, then evaluating the bound gives

$$\begin{aligned}\mathcal{I}^f(\beta_0) = Var\left(X_1'(\delta_1-\delta_0)+h_1(X_2)-h_0(X_2)\right) + \sigma_1^2\left\{\mathbb{E}\left[\frac{1}{e_0(X_2)}\right]\right. \\ \left. +\mathbb{E}\left[\frac{\mathbb{C}(D,X_1'|X_2)}{e_0(X_2)}\right]\mathbb{E}[e_0(X_2)\,\mathbb{V}(X_1|X_2,D=1)]^{-1}\,\mathbb{E}\left[\frac{\mathbb{C}(D,X_1'|X_2)}{e_0(X_2)}\right]'\right\} \\ + \sigma_0^2\left\{\mathbb{E}\left[\frac{1}{1-e_0(X_2)}\right]\right. \\ \left. +\mathbb{E}\left[\frac{\mathbb{C}(D,X_1'|X_2)}{1-e_0(X_2)}\right]\mathbb{E}[(1-e_0(X_2))\,\mathbb{V}(X_1|X_2,D=0)]^{-1}\,\mathbb{E}\left[\frac{\mathbb{C}(D,X_1'|X_2)}{1-e_0(X_2)}\right]'\right\},\end{aligned}$$

where $\mathbb{C}(A,B|c) = Cov(A,B|C=c)$. After some manipulation, this expression agrees with the one obtained by Wang, Linton and Härdle (2004, Theorem 3.4) (who did assume normality and homoscedasticity as part of the prior restriction).[14]

When is imposing the partial linear structure on $\mathbb{E}[Y_0|X]$ and $\mathbb{E}[Y_1|X]$ likely to be valuable? If X_1 is highly predictive for treatment in subpopulations homogenous in X_2 (i.e., $\mathbb{C}(D,X_1|X_2) \neq 0$), then within such subpopulations the distribution of X_1 will differ across treatment and controls. In

[14] To be precise Wang, Linton and Härdle (2004) only provide the bound for the marginal mean of Y_1, but their result easily generalizes to the ATE estimand.

such situations imposing the restriction that X_1 enters $\mathbb{E}[Y_0|X]$ and $\mathbb{E}[Y_1|X]$ linearly facilitates extrapolation. This may substantially improve the precision with which β_0 can be estimated.

Efficient estimation under Assumption 1.5 is conceptually straightforward (Appendix A provides expressions for the efficient influence function). For example, the semiparametric regression imputation estimator which solves

$$\sum_{i=1}^{N} D_i \left\{ \psi_{1i}(Y_{1i}, X_i, \widehat{\beta}) - q_0(X_i, \widehat{\delta}_0, \widehat{h}_0(X_{2_i}); \widehat{\beta}) \right\} - (1 - D_i) \left\{ \psi_0(Y_{0i}, X_i, \widehat{\beta}) - q_1(X_i, \widehat{\delta}_1, \widehat{h}_1(X_{2_i}); \widehat{\beta}) \right\} = 0,$$

should be efficient as long as the estimates of $\widehat{\delta}_0$ and $\widehat{\delta}_1$ are semiparametrically efficient. Wang, Linton and Härdle (2004), however, argue that such an estimator is unlikely to be useful in practice (p. 338). For example, in the partially linear model semiparametric efficient estimation of $\widehat{\delta}_0$ requires nonparametric estimation of the conditional variance of Y_0 given X; a difficult problem when the dimension of X is high.

When $q_0(X;\beta)$ and $q_1(X;\beta)$ are parametrically specified, the imputation estimator is generally considered unattractive due to its sensitivity to misspecification (e.g., Imbens 2004, p. 24). Similar concerns arise in the semiparametric case: partial linear imputation is an efficient way to 'deal with' limited (conditional on X_2) overlap of X_1 across treatment and controls but it may be very biased if the partial linear structure is false. In practice, the efficiency gains promised by Theorem 5.1, must be weighed against the risk of bias due to violations of Assumption 1.5.

6 Equivalent GMM problem for semiparametric data combination

Consider the following multinomial sampling procedure. With probability Q_0 the analyst randomly draws a unit from a population with distribution function G_0 and records its realizations of Y_1 and X; with probability $1 - Q_0$ the analyst randomly draws a unit from a different population (with distribution function H_0) and records its realizations of Y_0 and X_0. Let $D = 1$ if a unit so sampled is from the first population and zero otherwise. The sampling distribution induced by the multinomial scheme, F_0, has density

$$f_0(z,d) = Q_0^d (1 - Q_0)^{1-d} g_0(z)^d h_0(z)^{1-d},$$

where $g_0(z)$ and $h_0(z)$ are the densities of G_0 and H_0. The parameter of interest, β_0, is defined by the restriction

$$\mathbb{E}_{F_0}[\psi(Z,\beta_0)|D=1] = \mathbb{E}_{G_0}[\psi(Z,\beta_0)] = 0. \tag{16}$$

A familiar example helps to fix ideas. Deheija and Wahba (1999) combine two distinct samples to estimate the effect of the National Supported Work (NSW) demonstration, a labor training program, on post-intervention earnings. Their merged sample consists of 185 NSW participants from an evaluation sample and 2,490 non-participants drawn from the Panel Study of Income Dynamics (PSID). Let Y_1 and Y_0 denote the potential post-intervention earnings associated with assignment to NSW training ($D = 1$) and non-training ($D = 0$) respectively. This merged sampled cannot be conceptualized as a random one from a meaningful population. Deheija and Wahba (1999) therefore focus on estimating a feature of the 'study' population (NSW participants). In my notation, they set $\psi_1(Y_1, X, \beta) = Y_1$ and $\psi_0(Y_0, X, \beta) = Y_0 + \beta$ and define β_0 as the solution to (16) or

$$\beta_0 = \mathbb{E}_{F_0}[Y_1 - Y_0 | D = 1],$$

which gives the average treatment effect on the treated (ATT).

I call the problem defined by restriction (16) and Assumptions 1.1 to 1.4 the semiparametric data combination problem.[15] Chen, Hong and Tarozzi (2004), Tarozzi and Deaton (2007) and Egel, Graham and Pintos (2007) provide several examples of semiparametric data combination problems. The bound for these problems was first calculated by Hahn (1998) for the special case where β_0 is the ATT and the propensity score is either completely known or unknown. Recently Chen, Hong and Tarozzi (2004, 2007) have extended these results to the general moment condition case and have also considered the bound when the propensity score is known to belong to a parametric family (indexed by an unknown parameter).

When the propensity score is unknown the information bound is (Chen, Hong and Tarozzi 2004, 2007)

$$\mathcal{J}(\beta_0) = \mathbb{E}\left[\frac{p_0(X)}{Q_0}\Gamma_0(X)\right]' \mathbb{E}[\Phi_0(X)]^{-1} \mathbb{E}\left[\frac{p_0(X)}{Q_0}\Gamma_0(X)\right], \tag{17}$$

with

$$\begin{aligned}\Phi_0(x) = \left\{\frac{p_0(x)}{Q_0}\right\}^2 \Bigg\{&\frac{\Sigma_0(x;\beta_0)}{1-p_0(x)} + \frac{\Sigma_1(x;\beta_0)}{p_0(x)} \\ &+ \frac{1}{p_0(x)}[q_1(x;\beta_0) - q_0(x;\beta_0)][q_1(x;\beta_0) - q_0(x;\beta_0)]'\Bigg\}.\end{aligned} \tag{18}$$

When $p_0(X) = p(X, \gamma_0)$ for some known function $p(\cdot,\cdot)$ and unknown parameter γ_0 the bound is

[15] In fact for data combination problems Assumption 1.4 can be weakened. The propensity score can be zero for some values of X but still needs to be strictly below one for all values of X (i.e., 'weak overlap').

given by (17) with

$$\begin{aligned}\Phi_0(x) = &\left\{\frac{p_0(x)}{Q_0}\right\}^2 \left\{\frac{\Sigma_0(x;\beta_0)}{1-p_0(x)} + \frac{\Sigma_1(x;\beta_0)}{p_0(x)}\right. \\ &\left. + [q_1(x;\beta_0) - q_0(x;\beta_0)][q_1(x;\beta_0) - q_0(x;\beta_0)]'\right\} \\ &+ \mathbb{E}\left[\frac{q_1(X;\beta_0) - q_0(X;\beta_0)}{Q_0}\frac{\partial p(X,\gamma_0)}{\partial \gamma}'\right] \mathbb{E}\left[S_\gamma S_\gamma'\right]^{-1} \mathbb{E}\left[\frac{\partial p(X,\gamma_0)}{\partial \gamma}\frac{q_1(X;\beta_0) - q_0(X;\beta_0)'}{Q_0}\right],\end{aligned} \tag{19}$$

where $S_\gamma = \frac{D-p_0(X)}{p_0(X)[1-p_0(X)]}\frac{\partial p(X,\gamma_0)}{\partial \gamma}$. Finally, when $p_0(X)$ is known the bound is given by (17) with

$$\begin{aligned}\Phi_0(x) = &\left\{\frac{p_0(x)}{Q_0}\right\}^2 \left\{\frac{\Sigma_0(x;\beta_0)}{1-p_0(x)} + \frac{\Sigma_1(x;\beta_0)}{p_0(x)}\right. \\ &\left. + [q_1(x;\beta_0) - q_0(x;\beta_0)][q_1(x;\beta_0) - q_0(x;\beta_0)]'\right\}.\end{aligned} \tag{20}$$

Here I provide an analog of Theorem 2.1 for this problem and briefly discuss some of its implications for efficient estimation. I work with an identifying moment of (cf., Hirano, Imbens and Ridder 2003)

$$\mathbb{E}_{F_0}\left[\frac{p_0(X)}{Q_0}\left\{\frac{D}{p_0(X)}\psi_1(Y_1,X,\beta_0) - \frac{1-D}{1-p_0(X)}\psi_0(Y_0,X,\beta_0)\right\}\right] = 0, \tag{21}$$

and the auxiliary moment given by (6).

Theorem 6.1 (GMM Equivalence for Data Combination Problems with Unknown Propensity Score) *Suppose that (i) the distribution of Z has a known, finite support, (ii) there is some unknown $\beta_0 \in \mathcal{B} \subset \mathbb{R}^K$ and $\rho_0 = (\rho_1, \ldots, \rho_L)'$ where $\rho_l = p_0(x_l) \in \mathcal{P} \subset [0, 1-\kappa]$ for each $l = 1, \ldots, L$ and some $0 < \kappa < 1$ (with $\mathcal{X} = \{x_1, \ldots, x_L\}$ the known support of X) such that restrictions (21) and (6) hold, (iii) $\mathbb{E}[\Phi_0(X)]$ and $\mathcal{J}(\beta_0) = \mathbb{E}[\Gamma_0(X)]'\mathbb{E}[\Phi_0(X)]^{-1}\mathbb{E}[\Gamma_0(X)]$ are nonsingular with probability one for $\Phi_0(x)$ as defined by (18) and (iv) other regularity conditions hold (cf., Chamberlain 1992b, Section 2), then $\mathcal{J}(\beta_0)$ is the Fisher information bound for β_0.*

Proof. See Appendix A. ■

The bounds for the cases where the propensity score belongs to a parametric family and is known are given by the following Corollaries.

Corollary 6.1 (GMM Equivalence for Data Combination Problems with Parametric Propensity Score) *Suppose the conditions of Theorem 6.1 hold and $p_0(x) = p(x,\gamma_0) \in \mathcal{P} \subset [0, 1-\kappa]$ for some unknown $\gamma_0 \in \mathcal{G} \subset \mathbb{R}^{\dim(\gamma)}$, $0 < \kappa < 1$ and all $x \in \mathcal{X}$, then $\mathcal{J}(\beta_0)$ is the Fisher information bound for β_0 with $\Phi_0(x)$ as given in (19).*

Corollary 6.2 (GMM Equivalence for Data Combination Problems with Known Propensity Score) *Suppose the conditions of Theorem 6.1 hold with $p_0(x)$ known, then $\mathcal{J}(\beta_0)$ is the Fisher information bound for β_0 with $\Phi_0(x)$ as given in (20).*

Since their respective bounds coincide, I conclude that restrictions (21) and (6) exhaust the information content of the semiparametric data combination model. The equivalent GMM problem can be used to provide fresh intuitions for various features of the data combination problem. These extensions are relatively straightforward. Here I confine myself to a brief comment about efficient estimation.

Define the moment function

$$\begin{aligned} m\left(Z, p_0(X), q_0(X;\beta_0), q_1(X;\beta_0), \beta_0\right) &= \frac{p_0(X)}{Q_0}\left\{\frac{D}{p_0(X)}\psi_1(Y_1, X, \beta) - \frac{1-D}{1-p_0(X)}\psi_0(Y_0, X, \beta)\right\} \\ &\quad -\mathbb{E}^*\left[\frac{p_0(X)}{Q_0}\left\{\frac{D}{p_0(X)}\psi_1(Y_1, X, \beta_0) - \right.\right. \\ &\quad \left.\left.\frac{1-D}{1-p_0(X)}\psi_0(Y_0, X, \beta_0)\right\}\middle| \frac{D}{p_0(X)} - 1; X\right] \\ &= \frac{p_0(X)}{Q_0}\left\{\frac{D}{p_0(X)}\psi_1(Y_1, X, \beta) - \frac{1-D}{1-p_0(X)}\psi_0(Y_0, X, \beta)\right. \\ &\quad \left. - \left[\frac{q_1(X;\beta_0)}{p_0(X)} + \frac{q_0(X;\beta_0)}{1-p_0(X)}\right](D - p_0(X))\right\}. \end{aligned}$$

Now consider the estimator which chooses $\widehat{\beta}$ to solve

$$\frac{1}{N}\sum\nolimits_{i=1}^{N} m(Z_i, \widehat{p}_0(X_i), \widehat{q}_0(X_i;\widehat{\beta}), \widehat{q}_1(X_i;\widehat{\beta}), \widehat{\beta}),$$

with $\widehat{q}_0(X_i;\widehat{\beta})$ and $\widehat{q}_1(X_i;\widehat{\beta})$ nonparametric series estimates as described in Section 4 and $\widehat{p}_0(X_i)$ a series estimate, the parametric MLE, or the known propensity score (as is appropriate). Using standard results on semiparametric- or parametric- two-step estimation (e.g., Newey 1994a, 1994b, Newey and McFadden 1994) it is straightforward to show that $\widehat{\beta}$ is semiparametrically efficient (under regularity conditions). An advantage of this approach to estimation is that the same moment condition is used for all levels of prior knowledge of the propensity score. In contrast both the IPW and imputation estimators of, respectively, Hirano, Imbens and Ridder (2003) and Chen, Hong and Tarozzi (2004, 2007) must be modified to efficiently incorporate prior knowledge about the selection process.

7 Conclusion

This paper has shown that the semiparametric efficiency bound associated with what I have termed the 'MAR setup' is equivalent to the bound associated with a particular augmented set of moment restrictions. The demonstration of equivalence improves our understanding of this class of estimation problems. Several 'puzzles' in the missing data literature can be understood as special cases of standard GMM results (e.g., Newey and McFadden 1994). In particular, I have shown that redundancy of knowledge of the propensity score follows from information matrix block diagonality, that IPW with a nonparametric estimate of the propensity score is equivalent to solving a sequence of unconditional augmented moment problems (explaining the efficiency gains found by Hirano, Imbens and Ridder 2003), and, finally, that double robustness of AIPW follows from standard results on sequential GMM estimators (I also show that Wooldridge's (2007) three-step estimator for the average treatment effect is locally efficient as well as doubly robust).

Equivalence is also constructive, suggesting new estimators with desirable features not displayed by other currently available efficient estimators. For example, the first estimator outlined in Section 4 allows a researcher to incorporate smoothness and exclusion priors on the propensity score while maintaining asymptotic efficiency and robustness. As noted earlier, the 'second order' expansions of Ichimura and Linton (2005) suggest that imposing such restrictions may result in appreciably better small sample performance. I am aware of no other efficient estimation strategy with similar properties.

I also derive the efficiency bound for β_0 when semiparametric restrictions on the CEFs of $\psi_1(Y_1, X, \beta)$ and $\psi_0(Y_0, X, \beta)$ given X are imposed. My result fully generalizes the work of Wang, Linton and Härdle (2004) for the partial linear case under normality and homoscedasticity.

Finally I provide an equivalent GMM problem for a class of semiparametric data combination models which covers the ATT estimand. Inspired by the equivalent GMM problem I suggest an approach to efficient estimation that automatically makes optimal use of smoothness and exclusion priors on the propensity score. Again I am aware of no other efficient estimators with this property.

This paper suggests several open questions that may merit further research. First, a rigorous development of the various estimators suggested here is required, as is an evaluation of their small sample properties. Second, GMM equivalence may facilitate the derivation of new results on the 'higher order' properties of various estimates of β_0 as in Ichimura and Linton (2005) and Imbens, Newey and Ridder (2005). Consider the case where the propensity score is parametrically specified. Let $\widehat{\gamma}$ be the MLEs of the parameters indexing the parametric family and $\widehat{\beta}$ the solution to augmented

problem

$$\sum_{i=1}^{N}\left(\begin{array}{c}\left(\frac{D_i}{p(X_i,\widehat{\gamma})}-1\right)R^M(X_i)\\ \frac{D_i}{p(X_i,\widehat{\gamma})}\psi_1\left(Y_{1i},X_i,\widehat{\beta}\right)-\frac{1-D_i}{1-p(X_i,\widehat{\gamma})}\psi_0\left(Y_{0i},X_i,\widehat{\beta}\right)\end{array}\right)=0.$$

If M grows with N at the appropriate rate this estimator will attain the efficiency bound. However, in finite samples it is not known how to optimally choose M. However the similarity between this problem and that of choosing the number of moments in conditional moment problems, as in Donald, Imbens and Newey (2002), suggests a natural starting point.

Finally, the modified Horvitz-Thompson estimator discussed in Section 4 hints at connections with empirical likelihood. Some of these connections are developed more fully in Egel, Graham and Pintos (2007).

A Proofs and derivations

A.1 Proof of Theorem 2.1

To simplify notation in the Appendices let β denote the true parameter value β_0 unless explicitly stated otherwise (similarly the '0' subscript is removed from other objects, such as the propensity score, when doing so does not cause confusion).

The proof closely follows that of Theorem 1 in Chamberlain (1992b) and consists of three steps.

Step 1: Demonstration of equivalence with unconditional GMM problem The first step is to show that restrictions (5) and (6) are, in the multinomial case, equivalent to a finite set of unconditional moment restrictions. Under the multinomial assumption we have $X\in\{x_1,\ldots,x_L\}$ for some L. Let the $L\times 1$ vector B have a 1 in the l^{th} row if $X=x_l$ and zeros elsewhere and $\tau_l=\Pr(X=x_l)$ (observe that $\sum_{l=1}^{L}\tau_l=1$). Denote the value of the selection probability at $X=x_l$ by ρ_l and define $\rho=\{\rho_1,\ldots,\rho_L\}'$; this vector gives the values of $p(\cdot)$ at each of the mass points of X. Using this notation we can write $p(X)=B'\rho$.

Under the multinomial assumption restrictions (5) and (6) are equivalent to the $L+K\times 1$ vector of unconditional moment restrictions

$$\mathbb{E}\left[m(Z,\beta,\rho)\right]=\mathbb{E}\left[\begin{array}{c}m_1(Z,\rho)\\ m_2(Z,\beta,\rho)\end{array}\right]=0,$$

where

$$m_1(Z,\rho)=B\left(\frac{D}{B'\rho}-1\right),\qquad m_2(Z,\beta,\rho)=\frac{D}{B'\rho}\psi_1(Y_1,X,\beta)-\frac{1-D}{1-B'\rho}\psi_0(Y_0,X,\beta).$$

To verify that this is the case note that by iterated expectations

$$\mathbb{E}\left[m_1(Z,\rho)\right]=\left(\begin{array}{c}\tau_1\mathbb{E}\left[\left(\frac{D}{p(X)}-1\right)\middle|X=x_1\right]\\ \vdots\\ \tau_L\mathbb{E}\left[\left(\frac{D}{p(X)}-1\right)\middle|X=x_L\right]\end{array}\right),$$

and hence $\mathbb{E}\left[m_1(Z,\beta,\rho)\right]=0$ if and only if $\mathbb{E}\left[\frac{D}{p(X)}-1\middle|X\right]=0$ for all $X\in\{x_1,\ldots,x_L\}$. We also have

$$\mathbb{E}\left[m_2(Z,\beta,\rho)\right]=\mathbb{E}\left[\frac{D}{p(X)}\psi_1(Y_1,X,\beta)-\frac{1-D}{1-p(X)}\psi_0(Y_0,X,\beta)\right]=0,$$

so $\mathbb{E}\left[m\left(Z,\beta,\rho\right)\right]=0$ if and only if (5) and (6) are satisfied as claimed.

Step 2: Application of Lemma 2 of Chamberlain (1987) Chamberlain (1987, Lemma 2) shows that for Z a multinomial random variable the variance bound for β under the sole restriction that $\mathbb{E}\left[m\left(Z,\beta,\rho\right)\right]=0$ is

$$\left\{\left(M'V^{-1}M\right)^{-1}\right\}_{22}$$

where $\left\{\left(M'V^{-1}M\right)^{-1}\right\}_{22}$ is the lower-right $K\times K$ block of $\left(M'V^{-1}M\right)^{-1}$ with

$$V\overset{def}{\equiv}\mathbb{E}\left[m\left(Z,\beta,\rho\right)m\left(Z,\beta,\rho\right)'\right],\quad M\overset{def}{\equiv}\mathbb{E}\left[\frac{\partial m\left(Z,\beta,\rho\right)}{\partial\rho'},\frac{\partial m\left(Z,\beta,\rho\right)}{\partial\beta'}\right].$$

The application of Chamberlain's result requires that M has full column rank and that V is non-singular. The calculations made in Step 3 below demonstrate that these conditions are implied by the assumption that $\mathbb{E}\left[\Gamma(X)\right]$ has full column rank, $p\left(X\right)$ is bounded away from zero and one and non-singularity of $\mathbb{E}\left[\Omega\left(X\right)\right]$.

Step 3: Calculation of the bound The final step is to solve for an explicit expression for $\left\{\left(M'V^{-1}M\right)^{-1}\right\}_{22}$. This requires some simple, albeit tedious, algebra. Partitioning V_0

$$\underset{L+K\times L+K}{V}=\left(\begin{array}{cc}V_{11} & V_{12}\\ V_{12}' & V_{22}\end{array}\right)$$

we have the lower right-hand block, letting $\psi_j=\psi_j\left(Y_j,X,\beta\right)$ and $q_j\left(X\right)=\mathbb{E}\left[\psi_j\middle|X\right]$ for $j=0,1$, given by

$$\begin{aligned}V_{22} &= \mathbb{E}\left[m_2\left(Z,\beta,\rho\right)m_2\left(Z,\beta,\rho\right)'\right] \qquad (22)\\ &= \mathbb{E}\left[\left\{\frac{D\psi_1}{p\left(X\right)}-\frac{\left(1-D\right)\psi_0}{1-p\left(X\right)}\right\}\left\{\frac{D\psi_1}{p\left(X\right)}-\frac{\left(1-D\right)\psi_0}{1-p\left(X\right)}\right\}'\right]\\ &= \mathbb{E}\left[\frac{\mathbb{E}\left[\psi_1\psi_1'\middle|X\right]}{p\left(X\right)}+\frac{\mathbb{E}\left[\psi_0\psi_0'\middle|X\right]}{1-p\left(X\right)}\right]\\ &= \mathbb{E}\left[\frac{Var\left(\psi_1|X\right)}{p\left(X\right)}+\frac{1-p\left(X\right)}{p\left(X\right)}q_1\left(X\right)q_1\left(X\right)'+q_1\left(X\right)q_1\left(X\right)'\right.\\ &\quad \left.\frac{Var\left(\psi_0|X\right)}{1-p\left(X\right)}+\frac{p\left(X\right)}{1-p\left(X\right)}q_0\left(X\right)q_0\left(X\right)'+q_0\left(X\right)q_0\left(X\right)'\right]\\ &= \sum\nolimits_{l=1}^{L}\tau_l\left[\frac{\Sigma_{1,l}}{\rho_l}+\frac{1-\rho_l}{\rho_l}q_{1,l}q_{1,l}'+q_{1,l}q_{1,l}'\right.\\ &\quad \left.+\frac{\Sigma_{0,l}}{1-\rho_l}+\frac{\rho_l}{1-\rho_l}q_{0,l}q_{0,l}'+q_{0,l}q_{0,l}'\right],\end{aligned}$$

where

$$q_{j,l}=\mathbb{E}\left[\psi_j\left(Y_j,X,\beta\right)|X=x_l\right],\quad \Sigma_{j,l}=Var\left(\psi_j\left(Y_j,X,\beta\right)|X=x_l\right),\quad j=0,1.$$

The upper right-hand block is similarly derived as

$$\begin{aligned} V_{12} &= \mathbb{E}\left[m_1(Z,\beta)\, m_2(Z,\beta,\rho)'\right] \\ &= \mathbb{E}\left[B\left(\frac{D}{B'\rho}-1\right)\left\{\frac{D\psi_1(Y_1,X,\beta)}{B'\rho}-\frac{(1-D)\,\psi_0(Y_0,X,\beta)}{1-B'\rho}\right\}'\right] \\ &= \mathbb{E}\left[B\left(\frac{1-p(X)}{p(X)}q_1(X)'+q_0(X)'\right)\right] \\ &= \begin{pmatrix} \tau_1\frac{1-\rho_1}{\rho_1}q'_{1,1}+\tau_1 q'_{0,1} \\ \vdots \\ \tau_L\frac{1-\rho_L}{\rho_L}q'_{1,L}+\tau_L q'_{0,L} \end{pmatrix}. \end{aligned} \tag{23}$$

Finally the upper left-hand block is given by

$$\begin{aligned} V_{11} &= \mathbb{E}\left[B\left(\frac{D}{B'\rho}-1\right)\left(\frac{D}{B'\rho}-1\right)B'\right] \\ &= \mathbb{E}\left[BB'\left(\frac{D}{p(X)}-1\right)\left(\frac{D}{p(X)}-1\right)\right] \\ &= \mathbb{E}\left[BB'\left(\frac{1-p(X)}{p(X)}\right)\right] \\ &= diag\left\{\begin{array}{ccc} \tau_1\frac{1-\rho_1}{\rho_1} & \cdots & \tau_L\frac{1-\rho_L}{\rho_L} \end{array}\right\}. \end{aligned} \tag{24}$$

Partition M

$$\underset{L+K\times L+K}{M} = \begin{pmatrix} M_{1\rho} & 0 \\ M_{2\rho} & M_{2\beta} \end{pmatrix},$$

where, from similar calculations to those made above, we have

$$\begin{aligned} M_{1\rho} &= -diag\left\{\begin{array}{ccc} \frac{\tau_1}{\rho_1} & \cdots & \frac{\tau_L}{\rho_L} \end{array}\right\} \\ M_{2\rho} &= -\left(\begin{array}{ccc} \tau_1\frac{q_{1,1}}{\rho_1}+\tau_1\frac{q_{0,1}}{1-\rho_1} & \cdots & \tau_L\frac{q_{1,L}}{\rho_L}+\tau_L\frac{q_{0,L}}{1-\rho_L} \end{array}\right), \quad M_{2\beta} = \mathbb{E}\left[\Gamma(X)\right]. \end{aligned} \tag{25}$$

Applying standard results on partitioned inverses then yields

$$M^{-1} = \begin{pmatrix} M_{1\rho}^{-1} & 0 \\ -M_{2\beta}^{-1}M_{2\rho}M_{1\rho}^{-1} & M_{2\beta}^{-1} \end{pmatrix},$$

Note that the existence of $M_{1\rho}^{-1}$ and $M_{2\beta}^{-1}$ follows from the assumptions that $p(X)$ is bounded away from zero and one and the assumption that $\mathbb{E}\left[\Gamma(X)\right]$ has full column rank.

Redundancy of knowledge of the propensity score suggests that $M^{-1}VM^{-1\prime}$ will be block diagonal. A sufficient condition for this is that (cf., Prokhorov and Schmidt 2006)

$$V'_{12} = M_{2\rho}M_{1\rho}^{-1}V_{11}. \tag{26}$$

To verify that this condition holds use (24) and (25) to show that

$$M_{2\rho}M_{1\rho}^{-1}V_{11} = \left(\begin{array}{ccc} \tau_1\frac{1-\rho_1}{\rho_1}q_{1,1}+\tau_1 q_{0,1} & \cdots & \tau_L\frac{1-\rho_L}{\rho_L}q_{1,L}+\tau_L q_{0,L} \end{array}\right),$$

which equals V_{12}' as required. Exploiting the resulting simplifications yields

$$M^{-1}VM^{-1\prime} = \begin{pmatrix} M_{1\rho}^{-1}V_{11}M_{1\rho}^{-1} & 0 \\ 0 & M_{2\beta}^{-1}\left(V_{22} - V_{12}'V_{11}^{-1}V_{12}\right)M_{2\beta}^{-1\prime} \end{pmatrix}$$

and hence

$$\left(M^{-1}VM^{-1\prime}\right)_{22} = M_{2\beta}^{-1}\left(V_{22} - V_{12}'V_{11}^{-1}V_{12}\right)M_{2\beta}^{-1\prime}.$$

By $M_{2\rho}M_{1\rho}^{-1} = \left(q_{1,1} + \frac{\rho_1}{1-\rho_1}q_{0,1}, \ldots, q_{1,L} + \frac{\rho_L}{1-\rho_L}q_{0,L}\right)$ and (26) we have $V_{12}'V_{11}^{-1}V_{12}$ equal to

$$\begin{aligned} V_{12}^{'}V_{11}^{-1}V_{12} &= M_{2\rho}M_{1\rho}^{-1}V_{11}M_{1\rho}^{-1\prime}M_{2\rho}' \\ &= \sum_{l=1}^{L}\tau_l\frac{1-\rho_l}{\rho_l}q_{1,l}q_{1,l}' + \tau_l\frac{\rho_l}{1-\rho_l}q_{0,l}q_{0,l}' + \tau_l q_{1,l}q_{0,l}' + \tau_l q_{0,l}q_{1,l}' \\ &= \mathbb{E}\left[\frac{1-p(X)}{p(X)}q_1(X)q_1(X)' + \frac{p(X)}{1-p(X)}q_0(X)q_0(X)'\right. \\ &\quad \left. + q_1(X)q_0(X)' + q_0(X)q_1(X)'\right], \end{aligned}$$

and hence, using (22),

$$\begin{aligned} V_{22} - V_{12}'V_{11}^{-1}V_{12} &= \mathbb{E}\left[\frac{Var(\psi_1|X)}{p(X)} + \frac{Var(\psi_0|X)}{1-p(X)}\right. \\ &\quad \left. + (q_1(X) - q_0(X))(q_1(X) - q_0(X))'\right] \\ &= \mathbb{E}[\Lambda(X)]. \end{aligned}$$

Using this result and taking the partitioned determinant gives

$$\begin{aligned} \det(V) &= \det(V_{11})\det\left(V_{22} - V_{12}'V_{11}^{-1}V_{12}\right) \\ &= \mathbb{E}\left[\frac{1-p(X)}{p(X)}\right]\det\{\mathbb{E}[\Lambda(X)]\}, \end{aligned}$$

and hence V is non-singular under strong overlap (Assumption 1.4) and non-singularity of $\mathbb{E}[\Omega(X)]$.

Since $M_{2\beta} = \mathbb{E}[\Gamma(X)]$ we have

$$\mathcal{I}(\beta_0) = \mathbb{E}[\Gamma(X)]'\mathbb{E}[\Lambda(X)]^{-1}\mathbb{E}[\Gamma(X)],$$

as claimed.

For completeness the upper left-hand portion of the full variance covariance matrix is given by

$$M_{11}^{-1}V_{11}M_{11}^{-1\prime} = \mathcal{I}^{-1}(\rho_0) = \begin{pmatrix} \frac{1}{f(x_1)}p(x_1)(1-p(x_1)) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{f(x_L)}p(x_L)(1-p(x_L)) \end{pmatrix}$$

where $f(x) = \sum_{l=1}^{L}\tau_l \times \mathbf{1}(x = x_l)$.

A.2 Proof of Theorem 5.1

The first two steps of the proof of Theorem 5.1 are analogous to those of Theorem 2.1 and therefore omitted. The actual calculation of the bound, while conceptually straightforward, is considerably more tedious. Details of this step are provided here.

Assume that the marginal distributions of X_1 and X_2 have I and M points of support with probabilities $\pi_1, \ldots, \pi_I$ and $\varsigma_1, \ldots, \varsigma_M$. Let $L = I \times M$ and τ_{im} denote the joint probability $\Pr\left(X_1 = x_{1,i}, X_2 = x_{2,m}\right)$. Let $\lambda_j = \left(\lambda_{j,1}, \ldots, \lambda_{j,M}\right)'$ be the values of $h_j(\cdot)$ at each of the mass points of X_2. Let C be a $M \times 1$ vector with a 1 in the m^{th} row if $X_2 = x_{2,m}$ and zeros elsewhere. Finally it is convenient to use the shorthand $\Psi = \left[q_1(X) - q_0(X)\right]\left[q_1(X) - q_0(X)\right]'$. In what follows I use both the single and double subscript notation to denote a point on the support of X as is convenient. We can map between the two notations by observing that $x_{im} = x_l$ for $l = (i-1)M + m$.

For the multinomial case the conditional moment problem defined by (5), (6) and (15) is equivalent to the unconditional problem

$$\mathbb{E}\left[m(Z,\theta)\right] = \mathbb{E}\left[\begin{array}{c} m_1(Z,\rho) \\ m_2(Z,\rho,\lambda_0,\delta_0,\beta) \\ m_3(Z,\rho,\lambda_1,\delta_1,\beta) \\ m_4(Z,\rho,\beta) \end{array}\right] = 0,$$

with $\theta = \left(\rho', \lambda_0', \delta_0', \lambda_1', \delta_1', \beta'\right)'$ and

$$\underset{L\times 1}{m_1(Z,\rho)} = B\left(\frac{D}{B'\rho} - 1\right), \quad \underset{LK\times 1}{m_2(Z,\rho,\lambda_0,\delta_0,\beta)} = -\left(B \otimes I_K\right)\left(\frac{1-D}{1-B'\rho}\left(\psi(Z,\beta) - q_0(X,\delta_0,C'\lambda_0;\beta)\right)\right),$$

$$\underset{LK\times 1}{m_3(Z,\rho,\lambda_1,\delta_1,\beta)} = \left(B \otimes I_K\right)\left(\frac{D}{B'\rho}\left(\psi_1(Y_1,X,\beta) - q_1(X,\delta_1,C'\lambda_1;\beta)\right)\right)$$

$$\underset{K\times 1}{m_3(Z,\rho,\beta)} = \frac{D}{B'\rho}\psi_1(Y_1,X,\beta) - \frac{1-D}{1-B'\rho}\psi_0(Y_0,X,\beta),$$

where, for $j = 0, 1$, $q_j(X, \delta_j, h_j(X_2); \beta)$ is a known function, δ_j an unknown finite-dimensional parameter and $h_j(X_2)$ an unknown function of X_2 with $X = \left(X_1', X_2'\right)'$.

Partition $V = \mathbb{E}\left[m(Z,\theta)\, m(Z,\theta)'\right]$ as

$$\underset{L+2KL+K\times L+2KL+K}{V} = \left(\begin{array}{cccc} V_{11} & & & \\ V_{21} & V_{22} & & \\ V_{31} & V_{32} & V_{33} & \\ V_{41} & V_{42} & V_{43} & V_{44} \end{array}\right),$$

where, using calculations similar to those given in the proof of Theorem 2.1, we have

$$\begin{aligned}
\underset{L\times L}{V_{11}} &= diag\left\{\tau_1\frac{1-\rho_1}{\rho_1},\ldots,\tau_L\frac{1-\rho_L}{\rho_L}\right\}, \quad \underset{L\times KL}{V_{12}} = (\underline{0},\ldots,\underline{0}), \quad \underset{L\times KL}{V_{13}} = (\underline{0},\ldots,\underline{0}),\\
\underset{KL\times KL}{V_{0,22}} &= diag\left\{\tau_1\frac{\Sigma_{0,1}}{1-\rho_1},\ldots,\tau_L\frac{\Sigma_{0,L}}{1-\rho_L}\right\}, \quad \underset{KL\times KL}{V_{23}} = 0,\\
\underset{KL\times KL}{V_{33}} &= diag\left\{\tau_1\frac{\Sigma_{1,1}}{\rho_1},\ldots,\tau_L\frac{\Sigma_{1,L}}{\rho_L}\right\}\\
\underset{K\times L}{V_{41}} &= \left(\tau_1\left[\frac{1-\rho_1}{\rho_1}q_{1,1}+q_{0,1}\right],\ldots,\tau_L\left[\frac{1-\rho_L}{\rho_L}q_{1,L}+q_{0,L}\right]\right)\\
\underset{K\times KL}{V_{42}} &= \left(\tau_1\frac{\Sigma_{0,1}}{1-\rho_1},\ldots,\tau_L\frac{\Sigma_{0,L}}{1-\rho_L}\right), \quad \underset{K\times KL}{V_{43}} = \left(\tau_1\frac{\Sigma_{1,1}}{\rho_1},\ldots,\tau_L\frac{\Sigma_{1,L}}{\rho_L}\right)\\
\underset{K\times K}{V_{44}} &= \sum\nolimits_{l=1}^{L}\tau_l\left[\frac{\Sigma_{1,l}}{\rho_l}+\frac{1-\rho_l}{\rho_l}q_{1,l}q_{1,l}'+q_{1,l}q_{1,l}'+\frac{\Sigma_{0,l}}{1-\rho_l}+\frac{\rho_l}{1-\rho_l}q_{0,l}q_{0,l}'+q_{0,l}q_{0,l}'\right],
\end{aligned}$$

and hence

$$V=\begin{pmatrix} V_{11} & 0 & 0 & V_{41}'\\ 0 & V_{22} & 0 & V_{42}'\\ 0 & 0 & V_{33} & V_{43}'\\ V_{41} & V_{42} & V_{43} & V_{44}\end{pmatrix}$$

We can partition the Jacobian matrix

$$\underset{L+2KL+K\times L+2M+2J+K}{M}=\begin{pmatrix} M_{1\rho} & 0 & 0 & 0 & 0 & 0\\ 0 & M_{2\lambda_0} & M_{2\delta_0} & 0 & 0 & 0\\ 0 & 0 & 0 & M_{3\lambda_1} & M_{3\delta_1} & 0\\ M_{4\rho} & 0 & 0 & 0 & 0 & M_{4\beta}\end{pmatrix},$$

where

$$\begin{aligned}
\underset{L\times L}{M_{1\rho}} &= -diag\left\{\frac{\tau_1}{\rho_1},\ldots,\frac{\tau_L}{\rho_L}\right\}\\
\underset{KL\times M}{M_{2\lambda_0}} &= \left(H_{0,1}',\ldots,H_{0,I}'\right)', \quad \underset{KL\times J}{M_{2\delta_0}} = \begin{pmatrix}\tau_1\nabla_{\delta_0}q_{0,1}\\ \vdots\\ \tau_L\nabla_{\delta_0}q_{0,L}\end{pmatrix}\\
\underset{KL\times M}{M_{3\lambda_1}} &= -\left(H_{1,1}',\ldots,H_{1,I}'\right)', \quad \underset{KL\times J}{M_{3\delta_1}} = -\begin{pmatrix}\tau_1\nabla_{\delta_1}q_{1,1}\\ \vdots\\ \tau_L\nabla_{\delta_1}q_{1,L}\end{pmatrix}\\
M_{4\rho} &= -\begin{pmatrix}\tau_1\left(\frac{q_{1,1}}{\rho_1}+\frac{q_{0,1}}{1-\rho_1}\right) & \cdots & \tau_L\left(\frac{q_{1,L}}{\rho_L}+\frac{q_{0,L}}{1-\rho_L}\right)\end{pmatrix}, \quad M_{4\beta}=\mathbb{E}\left[\Gamma\left(X\right)\right].
\end{aligned}$$

where $H_{j,i}=diag\left\{\tau_{i1}\nabla_{h_j}q_{j,i1},\ldots,\tau_{iM}\nabla_{h_j}q_{j,iM}\right\}$ for $i=1,\ldots,I$ and $j=0,1$ with $q_{j,im}=q_j(x_{im},\delta_j,h_j\left(x_{2,m}\right);\beta)$.

The variance bound for β is given by the lower right-hand $K\times K$ block of $\left(M'V^{-1}M\right)^{-1}$. We begin by calculating V^{-1}. Partition V

$$V=\begin{pmatrix} B_{11} & B_{12}\\ B_{12}' & B_{22}\end{pmatrix},$$

with

$$\underset{L+2KL\times L+2KL}{B_{11}} = diag\left\{ \begin{array}{cccc} V_{11} & V_{22} & V_{33} \end{array} \right\}, \quad \underset{L+2KL\times K}{B_{12}} = \left(\begin{array}{ccc} V_{41} & V_{42} & V_{43} \end{array} \right)', \quad B_{22} = V_{44}.$$

Now partition V^{-1} as

$$V_0^{-1} = \left(\begin{array}{cc} C_{11} & C_{12} \\ C_{12}' & C_{22} \end{array} \right), \tag{27}$$

where the partitioned inverse formula gives

$$\underset{L+2KL\times L+2KL}{C_{11}} = diag\left\{ \begin{array}{cccc} V_{11}^{-1} & V_{22}^{-1} & V_{33}^{-1} \end{array} \right\} + D'\mathbb{E}\left[\Psi\right]^{-1} D, \quad \underset{K\times L+2KL}{C_{12}'} = -\mathbb{E}\left[\Psi\right]^{-1} D, \quad \underset{K\times K}{C_{22}} = \mathbb{E}\left[\Psi\right]^{-1}$$

with $D = \left(\begin{array}{ccc} A' & (\iota_L \otimes I_K)' & (\iota_L \otimes I_K)' \end{array} \right) = B_{12}'B_{11}^{-1}$ and $A = \left(\begin{array}{ccc} q_{1,1} + \frac{\rho_1}{1-\rho_1} q_{0,1} & \cdots & q_{1,L} + \frac{\rho_L}{1-\rho_L} q_{0,L} \end{array} \right)'$ a $L\times K$ matrix.

Expression (27) follows since

$$\begin{aligned}
C_{22} &= \left(B_{22} - B_{12}'B_{11}^{-1}B_{12}\right)^{-1} \\
&= \left\{ \sum_{l=1}^{L} \tau_l \left[\frac{\Sigma_{1,l}}{\rho_l} + \frac{1-\rho_l}{\rho_l} q_{1,l}q_{1,l}' + q_{1,l}q_{1,l}' + \frac{\Sigma_{0,l}}{1-\rho_l} + \frac{\rho_l}{1-\rho_l} q_{0,l}q_{0,l}' + q_{0,l}q_{0,l}' \right] \right. \\
&\quad - \left(\tau_1 \left[\frac{1-\rho_1}{\rho_1} q_{1,1} + q_{0,1} \right], \ldots, \tau_L \left[\frac{1-\rho_L}{\rho_L} q_{1,L} + q_{0,L} \right], \tau_1 \frac{\Sigma_{0,1}}{1-\rho_1}, \ldots, \tau_L \frac{\Sigma_{0,L}}{1-\rho_L}, \tau_1 \frac{\Sigma_{1,1}}{\rho_1}, \ldots, \tau_L \frac{\Sigma_{1,L}}{\rho_L} \right) \\
&\quad \times diag\left\{ \tau_1 \frac{1-\rho_1}{\rho_1}, \ldots, \tau_L \frac{1-\rho_L}{\rho_L}, \tau_1 \frac{\Sigma_{0,1}}{1-\rho_1}, \ldots, \tau_L \frac{\Sigma_{0,L}}{1-\rho_L}, \tau_1 \frac{\Sigma_{1,1}}{\rho_1}, \ldots, \tau_L \frac{\Sigma_{1,L}}{\rho_L} \right\}^{-1} \\
&\quad \left. \times \left(\tau_1 \left[\frac{1-\rho_1}{\rho_1} q_{1,1} + q_{0,1} \right], \ldots, \tau_L \left[\frac{1-\rho_L}{\rho_L} q_{1,L} + q_{0,L} \right], \tau_1 \frac{\Sigma_{0,1}}{1-\rho_1}, \ldots, \tau_L \frac{\Sigma_{0,L}}{1-\rho_L}, \tau_1 \frac{\Sigma_{1,1}}{\rho_1}, \ldots, \tau_L \frac{\Sigma_{1,L}}{\rho_L} \right)' \right\}^{-1} \\
&= \left\{ \sum_{l=1}^{L} \tau_l \left[\frac{\Sigma_{1,l}}{\rho_l} + \frac{1-\rho_l}{\rho_l} q_{1,l}q_{1,l}' + q_{1,l}q_{1,l}' + \frac{\Sigma_{0,l}}{1-\rho_l} + \frac{\rho_l}{1-\rho_l} q_{0,l}q_{0,l}' + q_{0,l}q_{0,l}' \right] \right. \\
&\quad \left. - \sum_{l=1}^{L} \tau_l \left[\frac{1-\rho_l}{\rho_l} q_{1,l}q_{1,l}' + \frac{\rho_l}{1-\rho_l} q_{0,l}q_{0,l}' + q_{1,l}q_{0,l}' + q_{0,l}q_{0,1}' + \frac{\Sigma_{0,l}}{1-\rho_l} + \frac{\Sigma_{1,l}}{\rho_l} \right] \right\} \\
&= \left\{ \sum_{l=1}^{L} \tau_l \left[q_{1,l} - q_{0,l} \right] \left[q_{1,l} - q_{0,l} \right]' \right\}^{-1} \\
&= \mathbb{E}\left[\Psi\right]^{-1}.
\end{aligned}$$

We also have $C_{12}' = -C_{22}B_{12}'B_{11}^{-1} = -\mathbb{E}\left[\Psi\right]^{-1} D$ and

$$C_{11} = B_{11}^{-1} + B_{11}^{-1}B_{12}C_{22}B_{12}'B_{11}^{-1} = diag\left\{ \begin{array}{cccc} V_{11}^{-1} & V_{22}^{-1} & V_{33}^{-1} \end{array} \right\} + D'\mathbb{E}\left[\Psi\right]^{-1} D.$$

We now evaluate $\mathcal{I}^f(\theta) = M'V^{-1}M$ to

$$\left(\begin{array}{cc}
M_{1\rho}'V_{11}^{-1}M_{1\rho} & 0 \\
0 & M_{2\lambda_0}'\left[V_{22}^{-1}+(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'\right]M_{2\lambda_0} \\
0 & M_{2\delta_0}'\left[V_{22}^{-1}+(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'\right]M_{2\lambda_0} \\
0 & M_{3\lambda_1}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'M_{2\lambda_0} \\
0 & M_{3\delta_1}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'M_{2\lambda_0} \\
0 & -M_{4\beta}'\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'M_{2\lambda_0}
\end{array}\right.$$

$$\left.\begin{array}{cc}
0 & 0 \\
M_{2\lambda_0}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'M_{2\delta_0} & M_{2\lambda_0}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'M_{3\delta_1} \\
M_{2\delta_0}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'M_{2\delta_0} & M_{2\delta_0}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'M_{3\delta_1} \\
M_{3\lambda_1}'\left[V_{33}^{-1}+(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'\right]M_{2\delta_0} & M_{3\lambda_1}'\left[V_{33}^{-1}+(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'\right]M_{3\delta_1} \\
M_{3\delta_1}'\left[V_{33}^{-1}+(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'\right]M_{2\delta_0} & M_{3\delta_1}'\left[V_{33}^{-1}+(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'\right]M_{3\delta_1} \\
-M_{4\beta}'\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'M_{3\lambda_1} & -M_{4\beta}'\mathbb{E}\left[\Psi\right]^{-1}(\iota_L\otimes I_K)'M_{3\delta_1}
\end{array}\right.$$

$$\left.\begin{array}{c}
0 \\
-M_{2\lambda_0}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}M_{4\beta} \\
-M_{2\delta_0}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}M_{4\beta} \\
-M_{3\lambda_1}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}M_{4\beta} \\
-M_{3\delta_1}'(\iota_L\otimes I_K)\,\mathbb{E}\left[\Psi\right]^{-1}M_{4\beta} \\
M_{4\beta}'\mathbb{E}\left[\Psi\right]^{-1}M_{4\beta}
\end{array}\right)$$

where I have made use of the equality $M_{1\rho}'A = M_{4\rho}'$.

Observe that, as in the standard semiparametric missing data model, $\mathcal{I}^f(\theta)$ satisfies Stein's condition for redundancy of knowledge of the propensity score for β. However the structure of the bound does indicate that knowledge of the finite dimensional parameters and nonparametric portions of the CEFs of $\psi_1(Y_1, X, \beta)$ and $\psi_0(Y_1, X, \beta)$ given X does increase the precision with which β can be estimated.

The variance bound for β_0 is given by the lower right-hand $K \times K$ block of the inverse of this matrix. Because of block diagonality we only need to consider the lower right-hand block. Partition this block as

$$\left(\begin{array}{cc} B_{11} & B_{12} \\ B_{12}' & B_{22} \end{array}\right)$$

where B_{11}, B_{12} and B_{22} are redefined to equal

$$\begin{aligned}
B_{11} &= \begin{pmatrix} M'_{2\lambda_0}V_{22}^{-1}M_{2\lambda_0} & M'_{2\lambda_0}V_{22}^{-1}M_{2\delta_0} & 0 & 0 \\ M'_{2\delta_0}V_{22}^{-1}M_{2\lambda_0} & M'_{2\delta_0}V_{22}^{-1}M_{2\delta_0} & 0 & 0 \\ 0 & 0 & M'_{3\lambda_1}V_{33}^{-1}M_{3\lambda_1} & M'_{3\lambda_1}V_{33}^{-1}M_{3\delta_1} \\ 0 & 0 & M'_{3\delta_1}V_{33}^{-1}M_{3\lambda_1} & M'_{3\delta_1}V_{33}^{-1}M_{3\delta_1} \end{pmatrix} \\
&\quad + \begin{pmatrix} M'_{2\lambda_0}(\iota_L \otimes I_K) \\ M'_{2\delta_0}(\iota_L \otimes I_K) \\ M'_{3\lambda_1}(\iota_L \otimes I_K) \\ M'_{3\delta_1}(\iota_L \otimes I_K) \end{pmatrix} \mathbb{E}[\Psi]^{-1} \\
&\quad \times \begin{pmatrix} (\iota_L \otimes I_K)' M_{0,2\lambda_0} & (\iota_L \otimes I_K)' M_{0,2\delta_0} & (\iota_L \otimes I_K)' M_{0,3\lambda_1} & (\iota_L \otimes I_K)' M_{0,3\delta_1} \end{pmatrix} \\
B_{12} &= \begin{pmatrix} M'_{2\lambda_0}(\iota_L \otimes I_K) \\ M'_{2\delta_0}(\iota_L \otimes I_K) \\ M'_{3\lambda_1}(\iota_L \otimes I_K) \\ M'_{3\delta_1}(\iota_L \otimes I_K) \end{pmatrix} \mathbb{E}[\Psi]^{-1} M_{4\beta} \\
B_{33} &= M'_{4\beta}\mathbb{E}[\Psi]^{-1} M_{4\beta}.
\end{aligned}$$

The information bound is therefore given by

$$\begin{aligned}
\mathcal{I}^f(\beta_0) &= B_{22} - B'_{12}B_{11}^{-1}B_{12} \\
&= M'_{4\beta}\mathbb{E}[\Psi]^{-1} M_{4\beta} - M'_{4\beta}\mathbb{E}[\Psi_0]^{-1} \\
&\quad \times \begin{pmatrix} (\iota_L \otimes I_K)' M_{0,2\lambda_0} & (\iota_L \otimes I_K)' M_{0,2\delta_0} & (\iota_L \otimes I_K)' M_{0,3\lambda_1} & (\iota_L \otimes I_K)' M_{0,3\delta_1} \end{pmatrix} \\
&\quad \times \left\{ \begin{pmatrix} M'_{2\lambda_0}V_{22}^{-1}M_{2\lambda_0} & M'_{2\lambda_0}V_{22}^{-1}M_{2\delta_0} & 0 & 0 \\ M'_{2\delta_0}V_{22}^{-1}M_{2\lambda_0} & M'_{2\delta_0}V_{22}^{-1}M_{2\delta_0} & 0 & 0 \\ 0 & 0 & M'_{3\lambda_1}V_{33}^{-1}M_{3\lambda_1} & M'_{3\lambda_1}V_{33}^{-1}M_{3\delta_1} \\ 0 & 0 & M'_{3\delta_1}V_{33}^{-1}M_{3\lambda_1} & M'_{3\delta_1}V_{33}^{-1}M_{3\delta_1} \end{pmatrix} \right. \\
&\quad + \begin{pmatrix} M'_{2\lambda_0}(\iota_L \otimes I_K) \\ M'_{2\delta_0}(\iota_L \otimes I_K) \\ M'_{3\lambda_1}(\iota_L \otimes I_K) \\ M'_{3\delta_1}(\iota_L \otimes I_K) \end{pmatrix} \mathbb{E}[\Psi]^{-1} \\
&\quad \left. \times \begin{pmatrix} (\iota_L \otimes I_K)' M_{0,2\lambda_0} & (\iota_L \otimes I_K)' M_{0,2\delta_0} & (\iota_L \otimes I_K)' M_{0,3\lambda_1} & (\iota_L \otimes I_K)' M_{0,3\delta_1} \end{pmatrix} \right\}^{-1} \\
&\quad \times \begin{pmatrix} M'_{2\lambda_0}(\iota_L \otimes I_K) \\ M'_{2\delta_0}(\iota_L \otimes I_K) \\ M'_{3\lambda_1}(\iota_L \otimes I_K) \\ M'_{3\delta_1}(\iota_L \otimes I_K) \end{pmatrix} \mathbb{E}[\Psi]^{-1} M_{4\beta} \\
&= M'_{4\beta} \left[\mathbb{E}[\Psi] + \begin{pmatrix} (\iota_L \otimes I_K)' M_{2\lambda_0} & (\iota_L \otimes I_K)' M_{2\delta_0} \end{pmatrix} \right. \\
&\quad \times \begin{pmatrix} M'_{2\lambda_0}V_{22}^{-1}M_{2\lambda_0} & M'_{2\lambda_0}V_{22}^{-1}M_{2\delta_0} \\ M'_{2\delta_0}V_{22}^{-1}M_{2\lambda_0} & M'_{2\delta_0}V_{22}^{-1}M_{2\delta_0} \end{pmatrix}^{-1} \begin{pmatrix} M'_{2\lambda_0}(\iota_L \otimes I_K) \\ M'_{2\delta_0}(\iota_L \otimes I_K) \end{pmatrix} \\
&\quad + \begin{pmatrix} (\iota_L \otimes I_K)' M_{3\lambda_1} & (\iota_L \otimes I_K)' M_{3\delta_1} \end{pmatrix} \\
&\quad \left. \times \begin{pmatrix} M'_{3\lambda_1}V_{33}^{-1}M_{3\lambda_1} & M'_{3\lambda_1}V_{33}^{-1}M_{3\delta_1} \\ M'_{3\delta_1}V_{33}^{-1}M_{3\lambda_1} & M'_{3\delta_1}V_{33}^{-1}M_{3\delta_1} \end{pmatrix}^{-1} \times \begin{pmatrix} M'_{3\lambda_1}(\iota_L \otimes I_K) \\ M'_{3\delta_1}(\iota_L \otimes I_K) \end{pmatrix} \right]^{-1} M_{4\beta},
\end{aligned}$$

where I have used the identity $A^{-1} - A^{-1}U(B^{-1} + U'A^{-1}U)^{-1}U'A^{-1} = \left(A + UBU'\right)^{-1}$.

Using the partitioned inverse formula and multiplying out the expression in [·] above then gives

$$\begin{aligned}
\mathcal{I}^f(\beta_0) &= M'_{4\beta} \times [\mathbb{E}[\Psi] + (\iota_L \otimes I_K)' \Big[M_{2\lambda_0} \left(M'_{2\lambda_0} V_{22}^{-1} M_{2\lambda_0} \right) M'_{2\lambda_0} \\
&+ \left(M_{2\delta_0} - M_{2\lambda_0} \left(M'_{2\lambda_0} V_{22}^{-1} M_{2\lambda_0} \right)^{-1} M'_{2\lambda_0} V_{22}^{-1} M_{2\delta_0} \right) \\
&\times \left(M'_{2\delta_0} V_{22}^{-1} M_{2\delta_0} - M'_{2\delta_0} V_{22}^{-1} M_{2\lambda_0} \left(M'_{2\lambda_0} V_{22}^{-1} M_{2\lambda_0} \right)^{-1} M'_{2\lambda_0} V_{22}^{-1} M_{2\delta_0} \right)^{-1} \\
&\times \left(M_{2\delta_0} - M_{2\lambda_0} \left(M'_{2\lambda_0} V_{22}^{-1} M_{2\lambda_0} \right)^{-1} M'_{2\lambda_0} V_{22}^{-1} M_{2\delta_0} \right)' \Big] \\
&+ (\iota_L \otimes I_K)' \Big[M_{3\lambda_1} \left(M'_{3\lambda_1} V_{33}^{-1} M_{3\lambda_1} \right) M'_{3\lambda_1} \\
&+ \left(M_{3\delta_1} - M_{3\lambda_1} \left(M'_{3\lambda_1} V_{33}^{-1} M_{3\lambda_1} \right)^{-1} M'_{3\lambda_1} V_{33}^{-1} M_{3\delta_1} \right) \\
&\times \left(M'_{3\delta_1} V_{33}^{-1} M_{3\delta_1} - M'_{3\delta_1} V_{33}^{-1} M_{3\lambda_1} \left(M'_{3\lambda_1} V_{33}^{-1} M_{3\lambda_1} \right)^{-1} M'_{3\lambda_1} V_{33}^{-1} M_{3\delta_1} \right)^{-1} \\
&\times \left(M_{3\delta_1} - M_{3\lambda_1} \left(M'_{3\lambda_1} V_{33}^{-1} M_{3\lambda_1} \right)^{-1} M'_{3\lambda_1} V_{33}^{-1} M_{3\delta_1} \right)' \Big] (\iota_L \otimes I_K) \Big]^{-1} \times M_{4\beta}.
\end{aligned}$$

We can now use the explicit expressions for V_0 and M_0 give above to generate an interpretable bound. The required calculations are tedious but straightforward (details are available in a supplemental Web Appendix), they give an information bound of:

$$\begin{aligned}
\mathcal{I}^f(\beta_0) &= \mathbb{E}[\Gamma(X)]' \times \big[\mathbb{E}\left[(q_1(X) - q_0(X))(q_1(X) - q_0(X))' \right] \\
&\quad + \mathbb{E}[A_0(X_2)] \\
&\quad + \mathbb{E}[B_0(X_2)] \mathbb{E}[C_0(X_2)]^{-1} \mathbb{E}[B_0(X_2)]' \\
&\quad + \mathbb{E}[A_1(X_2)] \\
&\quad + \mathbb{E}[B_1(X_2)] \mathbb{E}[C_1(X_2)]^{-1} \mathbb{E}[B_1(X_2)]' \big]^{-1} \times \mathbb{E}[\Gamma_0]
\end{aligned}$$

where

$$\begin{aligned}
A_j(X_2) &= \Delta_{h_j}(X_2) \Upsilon_{h_j}(X_2)^{-1} \Delta_{h_j}(X_2)' \\
B_j(X_2) &= \Delta_{\delta_j}(X_2) - \Delta_{h_j}(X_2) \Upsilon_{h_j}(X_2)^{-1} \Upsilon_{h_j \delta_j}(X_2) \\
C_j(X_2) &= \Upsilon_{\delta_j}(X_2) - \Upsilon_{h_j \delta_j}(X_2)' \Upsilon_{h_j}(X_2)^{-1} \Upsilon_{h_j \delta_j}(X_2)
\end{aligned}$$

for $j = 0, 1$ with

$$\begin{aligned}
\underset{K\times 1}{\Delta_{h_j}(X_2)} &= \mathbb{E}\left[\left.\frac{\partial q_j(X)}{\partial h_j'}\right| X_2\right] \\
\underset{K\times J}{\Delta_{\delta_j}(X_2)} &= \mathbb{E}\left[\left.\frac{\partial q_j(X)}{\partial \delta_j'}\right| X_2\right] \\
\underset{1\times 1}{\Upsilon_{h_j}(X_2)} &= \mathbb{E}\left[\left.\left(\frac{\partial q_j(X)}{\partial h_j'}\right)'\left[\frac{\Sigma_j(X)}{(1-j)(1-p(X))+jp(X)}\right]^{-1}\left(\frac{\partial q_j(X)}{\partial h_j'}\right)\right| X_2\right] \\
\underset{J\times J}{\Upsilon_{\delta_j}(X_2)} &= \mathbb{E}\left[\left.\left(\frac{\partial q_j(X)}{\partial \delta_j'}\right)'\left[\frac{\Sigma_j(X)}{(1-j)(1-p(X))+jp(X)}\right]^{-1}\left(\frac{\partial q_j(X)}{\partial \delta_j'}\right)\right| X_2\right] \\
\underset{1\times J}{\Upsilon_{h_j\delta_j}(X_2)} &= \mathbb{E}\left[\left.\left(\frac{\partial q_j(X)}{\partial h_j'}\right)'\left[\frac{\Sigma_j(X)}{(1-j)(1-p(X))+jp(X)}\right]^{-1}\left(\frac{\partial q_j(X)}{\partial \delta_j'}\right)\right| X_2\right].
\end{aligned}$$

Let $U_j = \psi_j\left(Y_j, X, \beta\right) - q_j(X)$, the the efficient influence function is evidently

$$\begin{aligned}
\phi(Z,\theta) &= \mathbb{E}[\Gamma(X)]^{-1}\left[D\Delta_{h_1}(X_2)\Upsilon_{h_1}(X_2)^{-1}\left(\frac{\partial q_1(X)}{\partial h_1}\right)\Sigma_1(X)^{-1}U_1\right. \\
&\quad + D\mathbb{E}[B_1(X_2)]\mathbb{E}[C_1(X_2)]^{-1}\left(\frac{\partial q_1(X)}{\partial \delta_1}\right)\Sigma_1(X)^{-1}U_1 \\
&\quad - D\mathbb{E}[B_1(X_2)]\mathbb{E}[C_1(X_2)]^{-1}\Upsilon_{h_1\delta_1}(X_2)\Upsilon_{h_1}(X_2)^{-1}\left(\frac{\partial q_1(X)}{\partial h_1}\right)\Sigma_1(X)^{-1}U_1 \\
&\quad - (1-D)\Delta_{h_0}(X_2)\Upsilon_{h_0}(X_2)^{-1}\left(\frac{\partial q_0(X)}{\partial h_0}\right)\Sigma_0(X)^{-1}U_0 \\
&\quad - (1-D)\mathbb{E}[B_0(X_2)]\mathbb{E}[C_0(X_2)]^{-1}\left(\frac{\partial q_0(X)}{\partial \delta_0}\right)\Sigma_0(X)^{-1}U_0 \\
&\quad \left. + (1-D)\mathbb{E}[B_0(X_2)]\mathbb{E}[C_0(X_2)]^{-1}\Upsilon_{h_0\delta_0}(X_2)\Upsilon_{h_0}(X_2)^{-1}\left(\frac{\partial q_0(X)}{\partial h_0}\right)\Sigma_0(X)^{-1}U_0\right].
\end{aligned}$$

The bound for the parametric case is easily seen to be

$$\begin{aligned}
\mathcal{I}^f(\beta_0) &= \mathbb{E}[\Gamma(X)]'\left[\mathbb{E}\left[(q_1(X)-q_0(X))(q_1(X)-q_0(X))'\right]\right. \\
&\quad + \mathbb{E}\left[\Delta_{\delta_1}(X)\right]\mathbb{E}\left[\Upsilon_{\delta_1}(X)\right]^{-1}\mathbb{E}\left[\Delta_{\delta_1}(X)\right]' \\
&\quad \left. + \mathbb{E}\left[\Delta_{\delta_0}(X)\right]\mathbb{E}\left[\Upsilon_{\delta_0}(X)\right]^{-1}\mathbb{E}\left[\Delta_{\delta_0}(X)\right]'\right]^{-1}\mathbb{E}[\Gamma(X)],
\end{aligned}$$

with an efficient influence function evidently given by

$$\begin{aligned}
\phi(Z;\theta) &= D\mathbb{E}[\Gamma(X)]^{-1}\mathbb{E}\left[\Delta_{\delta_1}(X)\right]\left\{\frac{\partial q_1(X)}{\partial \delta_1'}\right\}'\Sigma_1(X)^{-1}U_1 \\
&\quad - (1-D)\mathbb{E}[\Gamma(X)]^{-1}\mathbb{E}\left[\Delta_{\delta_0}(X)\right]\left\{\frac{\partial q_0(X)}{\partial \delta_0'}\right\}'\Sigma_0(X)^{-1}U_0 \\
&\quad + \mathbb{E}[\Gamma(X)]^{-1}(q_1(X)-q_0(X)).
\end{aligned}$$

A.3 Proof of Theorem 6.1.

The first two parts of the proof are analogous to those of Theorem 2.1 above and are omitted. Here I sketch the calculation of the bound. We let

$$m_1(Z,\rho) = B\left(\frac{D}{B'\rho} - 1\right), \qquad m_2(Z,\beta,\rho) = \frac{B'\rho}{Q}\left\{\frac{D}{B'\rho}\psi_1(Y_1,X,\beta) - \frac{1-D}{1-B'\rho}\psi_0(Y_0,X,\beta)\right\}$$

and define V and M as in the proof of Theorem 2.1. We again partition V and evaluate V_{22} as

$$\begin{aligned} V_{22} &= \mathbb{E}\left[m_2(Z,\beta,\rho)\, m_2(Z,\beta,\rho)'\right] \\ &= \frac{1}{Q^2}\mathbb{E}\left[\left\{D\psi_1 - \frac{p(X)}{1-p(X)}(1-D)\psi_0\right\}\left\{D\psi_1 - \frac{p(X)}{1-p(X)}(1-D)\psi_0\right\}'\right] \\ &= \frac{1}{Q^2}\mathbb{E}\left[p(X)\,\mathbb{E}\left[\psi_1\psi_1'|\,X\right] + \frac{p(X)^2}{1-p(X)}\mathbb{E}\left[\psi_0\psi_0'|\,X\right]\right] \\ &= \mathbb{E}\left[\frac{p(X)^2}{Q^2}\left\{\frac{\mathbb{E}\left[\psi_1\psi_1'|\,X\right]}{p(X)} + \frac{\mathbb{E}\left[\psi_0\psi_0'|\,X\right]}{1-p(X)}\right\}\right] \\ &= \sum_{l=1}^{L}\tau_l\frac{\rho_l^2}{Q_0^2}\left[\frac{\Sigma_{1,l}}{\rho_l} + \frac{1-\rho_l}{\rho_l}q_{1,l}q_{1,l}' + q_{1,l}q_{1,l}'\right. \\ &\quad \left. + \frac{\Sigma_{0,l}}{1-\rho_l} + \frac{\rho_l}{1-\rho_l}q_{0,l}q_{0,l}' + q_{0,l}q_{0,l}'\right], \end{aligned} \tag{28}$$

and V_{12} as

$$\begin{aligned} V_{0,12} &= \mathbb{E}\left[m_1(Z,\beta)\, m_2(Z,\beta,\rho)'\right] \\ &= \mathbb{E}\left[B\left(\frac{D}{B'\rho} - 1\right)\frac{B'\rho}{Q}\left\{\frac{D\psi_1}{B'\rho} - \frac{(1-D)\psi_0}{1-B'\rho}\right\}'\right] \\ &= \frac{1}{Q}\mathbb{E}\left[B\left(\frac{D}{B'\rho} - 1\right)\left\{D\psi_1 - \frac{B'\rho}{1-B'\rho}(1-D)\psi_0\right\}'\right] \\ &= \frac{1}{Q}\mathbb{E}\left[B\left\{[1-p(X)]\,q_1(X)' + p_0(X)\,q_0(X)'\right\}\right] \\ &= \frac{1}{Q_0}\begin{pmatrix} \tau_1(1-\rho_1)\,q_{1,1}' + \tau_1\rho_1 q_{0,1}' \\ \vdots \\ \tau_L(1-\rho_L)\,q_{1,L}' + \tau_L\rho_L q_{0,L}' \end{pmatrix}, \end{aligned} \tag{29}$$

with V_{11} remaining as defined in the proof of Theorem 2.1 above. The components of the Jacobian M are given by

$$\begin{aligned} M_{1\rho} &= -diag\begin{Bmatrix} \frac{\tau_1}{\rho_1} & \cdots & \frac{\tau_L}{\rho_L} \end{Bmatrix} \\ M_{2\rho} = -\frac{1}{Q_0}\begin{pmatrix} \tau_1\frac{q_{0,1}}{1-\rho_1} & \cdots & \tau_L\frac{q_{0,L}}{1-\rho_L} \end{pmatrix}, &\quad M_{2\beta} = \mathbb{E}\left[\frac{p(X)}{Q}\Gamma(X)\right]. \end{aligned} \tag{30}$$

Straightforward algebra gives a variance bounds for β of

$$\begin{aligned} \left\{M^{-1}VM^{-1\prime}\right\}_{22} &= M_{2\beta}^{-1}\left(M_{2\rho}M_{1\rho}^{-1}V_{11}M_{1\rho}^{-1\prime}M_{2\rho}' - V_{12}'M_{1\rho}^{-1\prime}M_{2\rho}' - M_{2\rho}M_{1\rho}^{-1}V_{12} + V_{22}\right)M_{2\beta}^{-1\prime} \\ &= \mathbb{E}\left[\frac{p(X)}{Q}\Gamma(X)\right]^{-1}\mathbb{E}\left[\Phi(X)\right]\mathbb{E}\left[\frac{p(X)}{Q}\Gamma(X)\right]^{-1\prime}, \end{aligned}$$

with $\Phi(x)$ as defined in (10).

If the propensity score is known, then the Jacobian simplifies to $M_0 = (0, M_{2\beta}')'$. The variance-bounds is then

$$\begin{aligned} M'V^{-1}M &= M_{2\beta}'\left(V_{22} - V_{12}'V_{11}^{-1}V_{12}\right)^{-1} M_{2\beta} \\ &= \mathbb{E}\left[\frac{p(X)}{Q}\Gamma(X)\right]' \times \mathbb{E}\left[\Phi(X)\right]^{-1} \\ &\quad \times \mathbb{E}\left[\frac{p(X)}{Q}\Gamma(X)\right]', \end{aligned}$$

with $\Phi(x)$ as defined in (20). The bound given for the case where the propensity score is parametrically specified follows from similar calculations.

References

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

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

Brown, Bryan W. and Whitney K. Newey. (1998). "Efficient semiparametric estimation of expectations," *Econometrica* 66 (2): 453 - 464.

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

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

Chamberlain, Gary. (1992b). "Comment: sequential moment restrictions in panel data," *Journal of Business and Economic Statistics* 10 (1): 20 - 26.

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

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

Chen, Xiaohong, Han Hong and Alessandro Tarozzi. (2007). "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics*, forthcoming.

Cheng, Philip E. (1994). "Nonparametric estimation of mean functionals with data missing at random," *Journal of the American Statistical Association* 89 (425): 81 - 87.

Cosslett, Stephen R. (1981). "Efficient estimation of discrete-choice models," *Structural Analysis of Discrete Data with Econometric Applications*: 51 - 111 (C.F. Manski, D. McFadden, Eds.). Cambridge, MA: The MIT Press.

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

Donald, Stephen G., Guido W. Imbens and Whitney K. Newey (2002). "Choosing the number of moments in conditional moment restriction models," *Mimeo.*

Engle, Robert F., C. W. J. Granger, John Rice and Andrew Weiss. (1986). "Semiparametric estimates of the relation between weather and electricity sales," Journal of the American Statistical Association 81 (394): 310 - 320.

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

Hahn, Jinyong. (2004). "Functional restriction and efficiency in causal inference," *Review of Economics and Statistics* 86 (1): 73 - 76.

Hirano, Keisuke and Guido W. Imbens. (2001). "Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization," *Health Services and Outcomes Research* 2 (3-4): 259 -278.

Hirano, Keisuke, Guido W. Imbens and Geert Ridder. (2003). "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica* 71 (4): 1161 - 1189.

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

Imbens, Guido W. (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 for over-identified generalized method of moment models," *Review of Economic Studies* 64 (3): 359 - 383.

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

Imbens, Guido W., Whitney K. Newey and Geert Ridder (2005). "Mean-square-error calculations for average treatment effects," *IEPR Working Paper 05.34.*

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

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

Newey, Whitney K. (1994a). "Series estimation of regression functionals," *Econometric Theory* 10 (1): 1 - 28.

Newey, Whitney K. (1994b). "The asymptotic variance of semiparametric estimators," *Econometrica* 62 (6): 1349 - 1382.

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

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

Prokhorov, Artem and Peter J, Schmidt. (2006). "GMM redundancy results for general missing data problems," *Mimeo.*

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

Robins, James M., Steven D. Mark and Whitney K. 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 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.

Scharfstein, Daniel O., Andrea Rotnitzky and James M. Robins. (1999). "Rejoinder," *Journal of the American Statistical Association* 94 (448): 1135- 1146.

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

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

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

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

Wooldridge, Jeffrey M. (1999b). "Distribution-free estimation of some nonlinear panel data models," *Journal of Econometrics* 90 (1): 77 - 97.

Wooldridge, Jeffrey M. (2002). "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.