Submitted to the Annals of Statistics

SEMIPARAMETRIC EFFICIENCY IN GMM MODELS WITH AUXILIARY DATA

By Xiaohong Chen, Han Hong and Alessandro Tarozzi

New York University and Duke University

We study semiparametric efficiency bounds and efficient estimation of parameters defined through general moment restrictions with missing data. Identification relies on auxiliary data containing information about the distribution of the missing variables conditional on proxy variables that are observed in both the primary and the auxiliary database, when such distribution is common to the two data sets. The auxiliary sample can be independent of the primary sample, or can be a subset of it. For both cases, we derive bounds when the probability of missing data given the proxy variables is unknown, or known, or belongs to a correctly specified parametric family. We find that the conditional probability is not ancillary when the two samples are independent. For all cases, we discuss efficient semiparametric estimators. An estimator based on a conditional expectation projection is shown to require milder regularity conditions than one based on inverse probability weighting.

1. Introduction. Many empirical studies are complicated by the presence of missing data. In such circumstances, identifying assumptions become necessary to overcome the lack of identification that results from the missing information. One solution to this identification problem is based on the assumption that information on the true value of the variables in the data set of interest (the *primary* data set) can be recovered using *auxiliary* data sources under a conditional independence assumption. The key element of this identification strategy is that the distribution of the variables of interest is assumed to be independent of whether they belong to the primary or the auxiliary sample, conditional on a set of proxy variables, which are observed in both samples.

The first goal of this paper is to study semiparametric efficiency bounds of parameters defined through general nonlinear and over-identified moment conditions for missing data models under a conditional independence assumption. We provide semiparametric efficiency bounds for the cases when the propensity score is unknown, or is known, or belongs to a correctly specified parametric family. In our context, the propensity score is defined as the probability that one observation belongs to the subsample where only the proxy variables are observed. The auxiliary sample can be either a subset of the primary sample ("verify-in-sample" case) or independent of the primary sample ("verify-out-of-sample"). The former case is a special case of the MAR or CAR missing

AMS 2000 subject classifications: Primary 62H12, 62D05; secondary 62F12, 62G20.

Keywords and phrases: Semiparametric Efficiency Bounds, GMM, Measurement Error, Missing Data, Auxiliary data, Sieve Estimation

data structure where the missing variables are common to all subjects. Semiparametric efficiency bounds for this case are closely related to the results in [26] when there is a single hierarchy in the case of monotone missing data patterns for a fixed set of instrument functions. (See also [25] and [7].) We provide new results on semiparametric efficiency bounds for the "verify-out-of-sample" case. We find that while more information on the propensity score will not affect the asymptotic efficiency bounds for parameters defined in the verify-in-sample case (as shown in, e.g. [26], [7] and [16]), it will improve the asymptotic efficiency for parameters defined in the verify-out-of-sample case. Our new efficiency bound results for the case when the parametric propensity is correctly specified should be useful in applied work because such an assumption is frequently adopted by empirical researchers.

The second goal of paper is to develop two classes of sieve-based, Generalized Method of Moments (GMM) estimators that achieve the efficiency bounds for parameters defined under either the "verify-out-of-sample" or the "verify-in-sample" framework. Each estimator relies only on one nonparametric estimate; a conditional expectation projection based GMM (hereafter CEP-GMM) estimator only requires the nonparametric estimation of a conditional expectation, while an inverse probability weighting based GMM (hereafter IPW-GMM) estimator only needs a nonparametric estimate of the propensity score. We establish asymptotic normality and efficiency properties of both estimators under weaker regularity conditions than the existing ones in the literature. In particular, we allow for nonlinear and non-smooth moment restrictions and for unbounded support of conditioning (or proxy) variables. The CEP-GMM estimator presents some advantages over the IPW-GMM estimator. First, its root-n asymptotic normality and efficiency can be derived without the strong assumption that the unknown propensity score is uniformly bounded away from zero and one. Second, the CEP-GMM estimator is characterized by a simple common format that achieves the relevant efficiency bound for all the cases we consider, regardless of whether the propensity score is unknown, or known or parametrically specified. Instead, the IPW-GMM estimator will be generally inefficient when the propensity score is known, or is parametrically estimated using a *correctly* specified parametric model; in such instances, different combinations of nonparametric and parametric estimates of the propensity score have to be specifically derived to achieve the efficiency bounds.

Our results can also be applied to the estimation of parametric nonlinear models with nonclassical measurement errors with validation data, a topic that has been studied in [6], [29], [5], [20], [8] among others.

Section 2 describes the model and presents the semiparametric efficiency bounds. Semiparametrically efficient CEP-GMM and IPW-GMM estimators are developed in sections 3 and 4 respectively. In Section 5 we illustrate empirically the performance of the different estimators in the estimation of the distribution of private consumption in rural India in the presence of missing data. Section 6 concludes. All proofs are given in the appendices.

2. Semiparametric Efficiency Bounds. Let $(X_i, Y_i, D_i)_{i=1}^n$ be an i.i.d. sample from (X, Y, D), and denote $Z_i = (Y_i, X_i)$ where Y_i is only observed when $D_i = 0$. We are interested in the estimation of parameters $\beta \in R^{d_\beta}$ defined implicitly in terms of general nonlinear moment conditions. In the first (verify-out-of-sample) case such conditions are described by

$$E\left[m(Z;\beta) \mid D = 1\right] = 0 \quad \text{if and only if} \quad \beta = \beta_0, \tag{1}$$

while in the second (verify-in-sample) case the condition is

$$E\left[m(Z;\beta)\right] = 0 \quad \text{if and only if} \quad \beta = \beta_0, \tag{2}$$

where $Z = (Y, X)$ and $m(\cdot;\beta)$ is a set of functions with dimension $d_m \geq d_\beta$.

In other words, under case (1) Y is always missing in the **primary** data set ($D = 1$), which is a random sample from the population of interest, while an independent **auxiliary** sample (where $D = 0$) will serve the purpose of ensuring the identification of parameters that would not be identified by the primary data set alone. Under case (2), the auxiliary sample is instead a subset of the entire primary sample.

In this section we present the semiparametric efficiency bound for the estimation of β implicitly defined by either moment conditions (1) or (2). To state the efficiency bounds we introduce some notations. Let $p = Pr(D = 1)$ and $p(X) = Pr(D = 1|X)$. In this paper β is typically used to denote an arbitrary value in the parameter space, but to save notation in this section β is also used as the true parameter value β_0. Define

$$\mathcal{E}(X;\beta) = E\left[m(Z;\beta)|X\right]$$

to be the conditional expectation of the moment conditions given X, and define

$$V(m(Z;\beta)|X) = E\left[m(Z;\beta)m(Z;\beta)'|X\right] - \mathcal{E}(X;\beta)\mathcal{E}(X;\beta)'$$

to be the conditional variance of the moment conditions given X. In addition, define

$$\mathcal{J}_\beta^1 = \frac{\partial}{\partial\beta} E\left[m(Z;\beta) \mid D = 1\right] \quad \text{and} \quad \mathcal{J}_\beta^2 = \frac{\partial}{\partial\beta} E\left[m(Z;\beta)\right].$$

Assumption 1. *(i) Both $\mathcal{J}_\beta^1$ and $\mathcal{J}_\beta^2$ have full column rank equal to d_β; (ii) The data $(X_i, Y_i, D_i)_{i=1}^n$ is an i.i.d. sample from (X, Y, D); (iii) $p = Pr(D = 1) \in (0, 1)$.*

Notice that in both case (1) and (2) the moment conditions are assumed to hold in the *primary* sample in which some information is missing. Identification is possible because of the access to an auxiliary data set ($D = 0$) which contains both $Y$and a set of proxy variables X that are also potentially of interest, if the following fundamental conditional independence assumption holds:

Assumption 2. $Y \perp D \mid X$.

Conditional independence assumptions have been used extensively in econometrics and statistics to achieve identification with missing data. Examples include inference in models with attrition or nonresponse (e.g. [21], [25], [27], [34], [35]), the estimation of treatment effects (see e.g. the references surveyed in [17]), the recovery of comparability over time of statistics calculated using data collected with different methodology (e.g. Clogg *et al.* (1991)[11], [28], [32]).

Under case (2), Assumption 1 would be satisfied if, for instance, the probability of validating a given observation only depends on X. In case (1), Assumption 2 requires that the sampling scheme used to create the auxiliary sample depends only on X. If a simple random subset of the primary data is validated, $p(X)$ is a constant and the auxiliary data set is characterized by the same distribution of (Y, X) as the primary data set, and Assumption 2 is easily seen satisfied. In this case, which is common in the statistics literature, the auxiliary data set is usually called a *validation* data set. A stratified sample satisfying Assumption 2 in model (2) can also be produced through a two-stage sampling design using a finite number of strata (see e.g. [4] and [3]), in which case the only variable X that is observed for all sampled observations is a discrete stratum indicator.

THEOREM 1. *Under Assumptions 1 and 2, the asymptotic variance lower bound for* $\sqrt{n}(\hat{\beta} - \beta)$ *for any regular estimator* $\hat{\beta}$ *is given by*

$$(\mathcal{J}_\beta' \Omega_\beta^{-1} \mathcal{J}_\beta)^{-1}.$$

When the moment condition case (1) holds, $\mathcal{J}_\beta \equiv \mathcal{J}_\beta^1$ *and* $\Omega_\beta = \Omega_\beta^1$ *where*

$$\Omega_\beta^1 = E\left[\frac{p(X)^2}{p^2(1-p(X))} V\left(m(Z;\beta) \mid X\right) + \frac{p(X)}{p^2} \mathcal{E}(X;\beta)\, \mathcal{E}(X;\beta)'\right].$$

When the moment condition case (2) holds, $\mathcal{J}_\beta \equiv \mathcal{J}_\beta^2$ *and* $\Omega_\beta = \Omega_\beta^2$ *where*

$$\Omega_\beta^2 = E\left[\frac{1}{1-p(X)} V\left[m(Z;\beta) \mid X\right] + \mathcal{E}(X;\beta)\, \mathcal{E}(X;\beta)'\right].$$

2.1. *Information content of the propensity score.* It is interesting to analyze whether the knowledge of the propensity score $p(X)$ decreases the semiparametric efficiency bounds for the parameters β.

THEOREM 2. *Under Assumptions 1 and 2, if* $p(X)$ *is known, then the semiparametric efficiency variance bound for estimating* β *is* $(\mathcal{J}_\beta' \tilde{\Omega}_\beta^{-1} \mathcal{J}_\beta)^{-1}$. *When the moment condition case (1) holds,* $\mathcal{J}_\beta \equiv \mathcal{J}_\beta^1$ *and* $\tilde{\Omega}_\beta = \tilde{\Omega}_\beta^1$ *where*

$$\tilde{\Omega}_\beta^1 = E\left[\frac{p(X)^2}{p^2(1-p(X))} V\left(m(Z;\beta) \mid X\right) + \frac{p(X)^2}{p^2} \mathcal{E}(X;\beta) \mathcal{E}(X;\beta)'\right].$$

When the moment condition (2) holds, $\mathcal{J}_\beta \equiv \mathcal{J}_\beta^2$ *and* $\tilde{\Omega}_\beta = \Omega_\beta^2$ *given in Theorem 1.*

In other words, knowledge of $p(X)$ reduces the semiparametric efficient variance bound for β under the "verify-out-of-sample" case, but it does not under the "verify-in-sample" case. The following argument provides an intuition for this result. When (2) holds, β is defined through the relation

$$\int\int m(y,x;\beta)f(y \mid x)dyf(x)dx = 0.$$

The propensity score $p(X)$ does not enter the definition of β, therefore its knowledge should not affect the variance bound for β. However, the relation that identifies β when (1) holds clearly depends on $p(X)$:

$$\int\int m(y,x;\beta)p(x)f(y \mid x)dyf(x)dx = 0.$$

Remark 1: A special case of Theorem 2 is when $p(X)$ is a constant p. In this case, the auxiliary sample is also called a validation sample and is drawn randomly from the same population as the primary sample, e.g. $Y, X \perp D$ ([6], [29], [20]). In such case it is then easy to see that the two efficiency bounds given in Theorem 2 become identical.

Another interesting question is what is the efficiency bound for the estimation of β defined by moment condition (1) if the propensity score is unknown but is assumed to belong to a correctly specified parametric family, so that $p(X) = p(X;\gamma)$. Let $p_\gamma(X) = \partial p(X;\gamma)/\partial\gamma$, and define the score function for γ as $S_\gamma = S_\gamma(D,X) = \frac{D-p(X;\gamma)}{p(X;\gamma)(1-p(X;\gamma))}p_\gamma(X)$.

THEOREM 3. *Under Assumptions 1 and 2, if $p(X) = p(X;\gamma)$ and $E[S_\gamma(D,X)S_\gamma(D,X)']$ is positive definite, then the efficiency variance bound for estimating β defined by moment condition (1) is given by $(\mathcal{J}_\beta'\tilde{\Omega}_\beta^{-1}\mathcal{J}_\beta)^{-1}$ where $\mathcal{J}_\beta = \mathcal{J}_\beta^1$ and*

$$\tilde{\Omega}_\beta = \tilde{\Omega}_\beta^1 + (E\frac{\mathcal{E}(X;\beta)p_\gamma(X)'}{p})\left[ES_\gamma S_\gamma'\right]^{-1}(E\frac{p_\gamma(X)\mathcal{E}(X;\beta)'}{p}).$$

This variance bound is clearly larger than $\tilde{\Omega}_\beta^1$ stated in Theorem 2, but it is smaller than the bound in Theorem 1. This latter result can be verified noting first that the bound in Theorem 3 corresponds to the variance of the following influence function:

$$\frac{(1-D)p(X)}{p(1-p(X))}(m(Z;\beta) - \mathcal{E}(X;\beta)) + \text{Proj}(\frac{\mathcal{E}(X;\beta)}{p}(D-p(X))\Big|S_\gamma(D,X)) + \frac{p(X)\mathcal{E}(X;\beta)}{p},$$

where we use $\text{Proj}(Z_1|Z_2)$ to denote the population least squares projection of a random variable Z_1 onto the linear space spanned by Z_2. The conclusion follows noting that the variance bound stated in Theorem 1 for moment condition (1) is instead the variance of the following influence function

$$\frac{1}{p}D\mathcal{E}(X;\beta) + \frac{(1-D)p(X)}{p(1-p(X))}\left[m(Z;\beta) - \mathcal{E}(X;\beta)\right],$$

whose corresponding variance is larger.

Our results for GMM models complement and extend the finding in the program evaluation literature that knowing the propensity score decreases the efficient variance bound for the estimation of the average effect of treatment on the treated, while the propensity score is ancillary for the average treatment effect parameter ([16]).

3. CEP-GMM Estimation. In this section, we consider a first class of semi-parametrically efficient estimators based on a conditional expectation projection (CEP) method. If Assumption 2 holds, identification follows by noting that, under case (1)

$$E\left[m(Z;\beta) \mid D=1\right] = \int E\left[m\left(Z;\beta\right) \mid x, D=0\right] f\left(x \mid D=1\right) dx,$$

while under case (2),

$$E\left[m\left(Z;\beta\right)\right] = \int E\left[m\left(Z;\beta\right) \mid x, D=0\right] f\left(x\right) dx.$$

Therefore, $E\left[m(Z;\beta)|x, D=0\right]$ can be recovered using observations where $D=0$, and it can then be integrated against either $f(x \mid D=1)$ or $f(x)$ to recover the parameters of interest. In other words, the projection method first estimates

$$\mathcal{E}(X;\beta) \equiv E\left[m(Z;\beta)|X\right] = E\left[m(Z;\beta)|X, D=0\right]$$

nonparametrically from the auxiliary sample, and then averages this nonparametric estimator over the primary sample.

3.1. *Efficient estimation with unknown propensity score.* In the following, we use subscripts p and a to refer to observations belonging to the primary sample and to the auxiliary sample respectively. Let n_p be the size of the primary sample and n_a be the size of the auxiliary sample. Observations in the primary sample are indexed by $i = 1, \ldots, n_p$. Observations in the auxiliary sample are indexed by $j = 1, \ldots, n_a$. Under moment condition (1) (verify-out-of-sample case), $n = n_p + n_a$. Under moment condition (2) (verify-in-sample case), $n = n_p$. Let $\hat{\mathcal{E}}(X;\beta)$ denote a nonparametric estimate of $\mathcal{E}(X;\beta)$ using the auxiliary sample. [8] (hereafter CHT) used a sieve based method for this nonparametric estimation. Let $\{q_l\left(X\right), l = 1, 2, ...\}$ denote a sequence of known basis functions that can approximate any square-measurable function of X arbitrarily well. Also let

$$q^{k(n_a)}\left(X\right) = \left(q_1\left(X\right), ..., q_{k(n_a)}\left(X\right)\right)' \quad \text{and} \quad Q_a = \left(q^{k(n_a)}\left(X_{a1}\right), ..., q^{k(n_a)}\left(X_{an_a}\right)\right)'$$

for some integer $k(n_a)$, with $k(n_a) \to \infty$ and $k(n_a)/n \to 0$ when $n \to \infty$. Then for each given β, the first step nonparametric estimation can be defined as,

$$\hat{\mathcal{E}}\left(X;\beta\right) = \sum_{j=1}^{n_a} m\left(Z_{aj};\beta\right) q^{k(n_a)}\left(X_{aj}\right) \left(Q_a' Q_a\right)^{-1} q^{k(n_a)}\left(X\right).$$

A generalized method of moment estimator for β_0 can then be defined as

$$\hat{\beta} = \arg\min_{\beta \in B} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \hat{\mathcal{E}}\left(X_{pi}; \beta\right) \right)' \hat{W} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \hat{\mathcal{E}}\left(X_{pi}; \beta\right) \right). \tag{3}$$

The $\sqrt{n}$-consistency and asymptotic normality of this CEP-GMM estimator have been established in CHT. Following the proof of their claim (A.2), we have the following asymptotic representation:

$$\begin{aligned} \frac{\sqrt{n}}{n_p} \sum_{i=1}^{n_p} \hat{\mathcal{E}}\left(X_{pi}; \beta_0\right) &= \frac{\sqrt{n}}{n_p} \sum_{i=1}^{n_p} \mathcal{E}(X_{pi}; \beta_0) \\ &+ \frac{\sqrt{n}}{n_a} \sum_{j=1}^{n_a} \frac{f_{X_p}(X_{aj})}{f(X_{aj}|D=0)} \left[m(Z_{aj}; \beta_0) - \mathcal{E}(X_{aj}; \beta_0) \right] + o_p(1), \end{aligned}$$

where we use $f_{X_p}(X)$ to denote the density of X in the primary data set, and $o_p(1)$ represents a term that converges to 0 in probability.

When moment condition (1) holds, $n = n_p + n_a$, $f_{X_p}(X) = f(X \mid D = 1)$ and

$$\frac{f_{X_p}(X)}{f(X \mid D=0)} = \frac{(1-p)p(X)}{p(1-p(X))}.$$

In this case we can also write the influence function for $\frac{\sqrt{n}}{n_p} \sum_{i=1}^{n_p} \hat{\mathcal{E}}\left(X_{pi}; \beta_0\right)$ as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{1}{p} D_i \mathcal{E}(X_i; \beta_0) + (1 - D_i) \frac{p(X_i)}{p(1-p(X_i))} \left[m(Z_i; \beta_0) - \mathcal{E}(X_i; \beta_0) \right] \right\} + o_p(1). \tag{4}$$

The proof of Theorem 1 shows that the two terms in the influence function correspond to the two components of the efficient influence function that contain information about $f(X|D=1)$ and $f(Y \mid X)$ respectively. These two terms are orthogonal to each other, so that

$$Avar\left(\frac{\sqrt{n}}{n_p} \sum_{i=1}^{n_p} \hat{\mathcal{E}}\left(X_{pi}; \beta_0\right)\right) = \Omega^1_\beta,$$

where Ω^1_β is given in Theorem 1.

When moment condition (2) holds, $f_{X_p}(X) = f(X)$, $n_p = n$ and

$$\frac{f_{X_p}(X)}{f(X \mid D=0)} = \frac{1-p}{1-p(X)}.$$

The influence function for $\frac{\sqrt{n}}{n_p} \sum_{i=1}^{n_p} \hat{\mathcal{E}}\left(X_{pi}; \beta_0\right)$ can then be written as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \mathcal{E}(X_i; \beta_0) + (1 - D_i) \frac{1}{1-p(X_i)} \left[m(Z_i; \beta_0) - \mathcal{E}(X_i; \beta_0) \right] \right\} + o_p(1). \tag{5}$$

The two terms in the influence function correspond to the two components of the projected efficiency influence function that contain information about $f(X)$ and $f(Y \mid X)$ in the proof of Theorem 1. The orthogonality between these two terms implies that

$$Avar(\frac{\sqrt{n}}{n_p}\sum_{i=1}^{n_p}\hat{\mathcal{E}}\left(X_{pi};\beta_0\right)) = \Omega_\beta^2,$$

where Ω_β^2 is given in Theorem 1. The semiparametric efficiency bounds given in Theorem 1 are then achieved by an optimally weighted GMM estimator $\hat{\beta}$ for β_0 that uses a weighting matrix $\hat{W} = \Omega_\beta^{-1} + o_p(1)$.

Before we formally present the semiparametric efficiency property of the CEP-GMM estimator, we need to introduce some notations and assumptions. Let the support of X be $\mathcal{X} = \mathcal{R}^{d_x}$. We could use more complicated notations and let $\mathcal{X} = \mathcal{X}_c \times \mathcal{X}_{dc}$, with $\mathcal{X}_c$ being the support of the continuous variables and $\mathcal{X}_{dc}$ the support of the finitely many discrete variables. Further we could decompose $\mathcal{X}_c = \mathcal{X}_{c1} \times \mathcal{X}_{c2}$ with $\mathcal{X}_{c1} = \mathcal{R}^{d_{x,1}}$ and $\mathcal{X}_{c2}$ being a compact and connected subset of $\mathcal{R}^{d_{x,2}}$. Then, under simple and usual modifications of the assumptions, the large sample results stated below would remain valid. To avoid tedious notation yet to allow for some unbounded support elements of X, we assume $\mathcal{X} = \mathcal{X}_c = \mathcal{R}^{d_x}$. For any $1 \times d_x$ vector $\mathbf{a} = (a_1, \ldots, a_{d_x})$ of non-negative integers, we write $|\mathbf{a}| = \sum_{k=1}^{d_x} a_k$, and for any $x = (x_1, ..., x_{d_x})' \in \mathcal{X}$, we denote the $|\mathbf{a}|$-th derivative of a function $h : \mathcal{X} \to \mathcal{R}$ as:

$$\nabla^{\mathbf{a}} h(x) = \frac{\partial^{|\mathbf{a}|}}{\partial x_1^{a_1} \ldots \partial x_{d_x}^{a_{d_x}}} h(x).$$

For some $\gamma > 0$, let $\underline{\gamma}$ be the largest integer smaller than γ, and let $\Lambda^\gamma(\mathcal{X})$ denote a Hölder space with smoothness γ, i.e., a space of functions $h : \mathcal{X} \to \mathcal{R}$ which have up to $\underline{\gamma}$ continuous derivatives, and the highest ($\underline{\gamma}$-th) derivatives are Hölder continuous with the Hölder exponent $\gamma - \underline{\gamma} \in (0, 1]$. The Hölder space becomes a Banach space when endowed with the Hölder norm:

$$||h||_{\Lambda^\gamma} = \sup_x |h(x)| + \max_{|\mathbf{a}|=\underline{\gamma}} \sup_{x \neq \overline{x}} \frac{|\nabla^{\mathbf{a}} h(x) - \nabla^{\mathbf{a}} h(\overline{x})|}{\sqrt{(x-\overline{x})'(x-\overline{x})}^{\gamma-\underline{\gamma}}} < \infty.$$

Let $\Lambda^\gamma(\mathcal{X}, \omega_1)$ denote a weighted Hölder space of functions $h : \mathcal{X} \to \mathcal{R}$ such that $h(\cdot)[1 + |\cdot|^2]^{-\omega_1/2}$ is in $\Lambda^\gamma(\mathcal{X})$. We call $\Lambda_c^\gamma(\mathcal{X}, \omega_1) \equiv \{h \in \Lambda^\gamma(\mathcal{X}, \omega_1) : ||h(\cdot)[1 + |\cdot|^2]^{-\omega_1/2}||_{\Lambda^\gamma} \leq c < \infty\}$ a weighted Hölder ball (with radius c).

The sieve estimator $\widehat{\mathcal{E}}(X;\beta)$ needs to converge to $\mathcal{E}(X;\beta)$ in some metric. We allow supports of the proxy variables to be unbounded, and use a weighted sup-norm metric defined as

$$||g||_{\infty,\omega} \equiv \sup_{x \in \mathcal{X}, \beta \in B} \left| g(x,\beta)[1 + |x|^2]^{-\omega/2} \right|$$

for some $\omega > 0$. Also we let $\Pi_{\infty n} g$ denote the projection of g onto the closed linear span of $q^{k(n_a)}(x) = (q_1(x), ..., q_{k(n_a)}(x))'$ under the norm $||\cdot||_{\infty,\omega}$. Let $f_{X_a}(x) = f_{X|D=0}(x)$ and $f_{X_p}(x) = f_{X|D=1}(x)$.

The following assumption is sufficient to ensure that $\widehat{\mathcal{E}}(\cdot;\beta)$ converges to $\mathcal{E}(\cdot;\beta)$ under the supremum norm $||\cdot||_{\infty,\omega}$.

ASSUMPTION 3. *Let $\widehat{W} - W = o_p(1)$ for a positive semidefinite matrix W, and the following hold: (1) for all $\beta \in B$, $\mathcal{E}(\cdot;\beta)$ belongs to a weighted Hölder ball $\Lambda_c^\gamma(\mathcal{X},\omega_1)$ for some $\gamma > 0$ and $\omega_1 \geq 0$; (2) $\int(1+|x|^2)^\omega f_{X_p}(x)dx < \infty$, $\int(1+|x|^2)^\omega f_{X_a}(x)dx < \infty$ for some $\omega > \omega_1 \geq 0$; (3) For each fixed x, $\mathcal{E}(x;\beta)$ is continuous at β for all $\beta \in B$; (4) $Var[m(Z_i;\beta) \mid X_i = x, D_i = 0]$ is bounded uniformly over x and β; (5) For any $\mathcal{E}(\cdot;\beta) \in \Lambda_c^\gamma(\mathcal{X},\omega_1)$, there is a sequence $\Pi_{\infty n}\mathcal{E}$ in the sieve space $\mathcal{G}_n = \{g(\cdot;\beta) \in \Lambda_c^\gamma(\mathcal{X},\omega_1) : g(x;\beta) = q^{k(n_a)}(x)'\pi(\beta)\}$ such that $||\mathcal{E}(\cdot;\beta) - \Pi_{\infty n}\mathcal{E}(\cdot;\beta)||_{\infty,\omega} = o(1)$. Also $E_a[q^{k(n_a)}(X)q^{k(n_a)}(X)']$ is non-singular.*

THEOREM 4. *Let $\widehat{\beta}$ be the CEP-GMM estimator given in (3). Under Assumptions 1, 2 and 3, if $k(n_a) \to \infty$, $\frac{k(n_a)}{n_a} \to 0$, then $\hat{\beta} - \beta_0 = o_p(1)$.*

Additional regularity conditions are required for stating the asymptotic normality results. Let $E_p(\cdot) = E(\cdot|D=1)$ and $E_a(\cdot) = E(\cdot|D=0)$. Denote $||h||_{2,a}^2 = \int h(x)^2 f_{X_a}(x)dx = E_a\{h(X)^2\}$ and $\Pi_{2n}h$ be the projection of h onto the closed linear span of $q^{k(n_a)}(x) = (q_1(x),...,q_{k(n_a)}(x))'$ under the norm $||\cdot||_{2,a}$.

ASSUMPTION 4. *Let $\beta_0 \in int(B)$, $E_p[\mathcal{E}(X;\beta_0)\mathcal{E}(X;\beta_0)']$ be positive definite, and the following hold: (1) Assumption 3.1 is satisfied with $\gamma > d_x/2$ and Assumption 3.2 is satisfied with $\omega > \omega_1 + \gamma$; (2) For each fixed x, and for some $\delta > 0$ $\frac{\partial\mathcal{E}(x;\beta)}{\partial\beta'}$ is continuous in $\beta \in B$ with $|\beta - \beta_0| \leq \delta$, $E_p\left[\sup_{\beta:|\beta-\beta_0|\leq\delta}\left|\frac{\partial\mathcal{E}(X_p;\beta)}{\partial\beta'}\right|\right] < \infty$; (3) There exist a constant $\epsilon \in (0,1]$, a $\delta > 0$ and a measurable function $b(\cdot)$ with $E_p[b(X_p)] < \infty$ such that $|\frac{\partial\widetilde{\mathcal{E}}(x;\beta)}{\partial\beta'} - \frac{\partial\mathcal{E}(x;\beta)}{\partial\beta'}| \leq b(x)[||\widetilde{\mathcal{E}} - \mathcal{E}||_{\infty,\omega}]^\epsilon$ for all $\beta \in B$ with $|\beta - \beta_0| \leq \delta$ and all $\widetilde{\mathcal{E}} \in \Lambda_c^\gamma(\mathcal{X},\omega_1)$ with $||\widetilde{\mathcal{E}} - \mathcal{E}||_{\infty,\omega} \leq \delta$. (4) $E_a\left[\left(\frac{f_{X_p}(X)}{f_{X_a}(X)}\right)^2\right] < \infty$; (5) $k(n_a) = O\left(n_a^{\frac{d_x}{2\gamma+d_x}}\right)$, $n_a^{-\frac{\gamma}{2\gamma+d_x}} \times \left\|\frac{f_{X_p}(\cdot)}{f_{X_a}(\cdot)} - \Pi_{2n}\frac{f_{X_p}(\cdot)}{f_{X_a}(\cdot)}\right\|_{2,a} = o(n^{-1/2})$.*

THEOREM 5. *Let $\widehat{\beta}$ be the CEP-GMM estimator given in (3). Under Assumptions 1, 2, 3 and 4, we have $\sqrt{n}(\widehat{\beta}-\beta_0) \Rightarrow \mathcal{N}(0,V)$, with $V = (\mathcal{J}_\beta' W \mathcal{J}_\beta)^{-1}\mathcal{J}_\beta' W \Omega_\beta W \mathcal{J}_\beta (\mathcal{J}_\beta' W \mathcal{J}_\beta)^{-1}$, where Ω_β is given in Theorem 1. Furthermore, if $W = \Omega_\beta^{-1}$, then $\sqrt{n}(\widehat{\beta}-\beta_0) \Rightarrow \mathcal{N}(0,V_0)$, with $V_0 = (\mathcal{J}_\beta'\Omega_\beta^{-1}\mathcal{J}_\beta)^{-1}$, where $\mathcal{J}_\beta = \mathcal{J}_\beta^1$ and $\Omega_\beta = \Omega_\beta^1$ under moment condition (1), and $\mathcal{J}_\beta = \mathcal{J}_\beta^2$ and $\Omega_\beta = \Omega_\beta^2$ under moment condition (2).*

Remark 2: **(i)** Assumptions 3 and 4 allow for $m(Z;\beta)$ to be non-smooth such as in quantile based moment functions. **(ii)** The weights ω and ω_1 are needed since the support of the conditioning variable X is allowed to include the entire Euclidean space. When X has bounded support and f_X is bounded above and below over its support, we

can simply set $\omega = 0 = \omega_1$ in Assumptions 3 and 4, and replace Assumption 4.1 with the assumption that 3.1 holds with $\gamma > d_x/2$. **(iii)** Since $\frac{f_{X_p}(X)}{f_{X_a}(X)} = \frac{p(X)(1-p)}{(1-p(X))p}$, Assumption 5.4 is automatically satisfied under the condition $0 < \underline{p} \leq p(x) \leq \overline{p} < 1$, which is a condition typically imposed in the program evaluation literature such as in [18]. Assumption 4.5 will be satisfied under mild smoothness conditions imposed on $\frac{p(X)}{1-p(X)}$. In particular, if we let $k(n_a) = O\left(n_a^{\frac{d_x}{2\gamma+d_x}}\right)$, the growth order which leads to the optimal convergence rate of $||\widehat{\mathcal{E}}(\cdot;\beta_0) - \mathcal{E}(\cdot;\beta_0)||_{2,a} = O_p\left(n_a^{-\frac{\gamma}{2\gamma+d_x}}\right)$, then Assumption 4.5 is satisfied with $\left\|\frac{f_{X_p}(\cdot)}{f_{X_a}(\cdot)} - \Pi_{2n}\frac{f_{X_p}(\cdot)}{f_{X_a}(\cdot)}\right\|_{2,a} = o\left(n_a^{-\frac{d_x}{2(2\gamma+d_x)}}\right) = o\left(k(n_a)^{-\frac{1}{2}}\right)$. For example, both Assumptions 4.4 and 4.5 will be satisfied as long as $\frac{p(\cdot)}{1-p(\cdot)} \in \Lambda^{\gamma_1}(\mathcal{X}, \omega_1)$ with $\gamma_1 > d_x/2$.

The proofs of Theorems 4 and 5 follow directly from those in CHT, who also provide simple consistent estimators of V and V_0:

$$\widehat{V} = (\widehat{\mathcal{J}}'W\widehat{\mathcal{J}})^{-1}\widehat{\mathcal{J}}'W\widehat{\Omega}W\widehat{\mathcal{J}}(\widehat{\mathcal{J}}'W\widehat{\mathcal{J}})^{-1} \quad \text{and} \quad \widehat{V}_0 = (\widehat{\mathcal{J}}'\widehat{\Omega}^{-1}\widehat{\mathcal{J}})^{-1},$$

where for both moment conditions (1) and (2),

$$\widehat{\mathcal{J}} = \frac{1}{n_p}\sum_{i=1}^{n_p}\frac{\partial\widehat{\mathcal{E}}(X_{pi};\widehat{\beta})}{\partial\beta'},$$

$$\widehat{\Omega} = \frac{1}{n_a}\sum_{j=1}^{n_a}\left(\widehat{v^*_{aj}}\widehat{U}_{aj}\right)\left(\widehat{v^*_{aj}}\widehat{U}_{aj}\right)' + \frac{n_a}{n_p^2}\sum_{i=1}^{n_p}\left(\widehat{\mathcal{E}}(X_{pi};\widehat{\beta})\widehat{\mathcal{E}}(X_{pi};\widehat{\beta})'\right),$$

$$\widehat{U}_{aj} = m(Y_{aj}, X_{aj};\widehat{\beta}) - \widehat{\mathcal{E}}(X_{aj};\widehat{\beta}), \quad \widehat{v^*_{aj}} = \left[\frac{1}{n_p}\sum_{i=1}^{n_p}q^{k(n_a)}(X_{pi})\right]'\left(\frac{Q_a'Q_a}{n_a}\right)^{-1}q^{k(n_a)}(X_{aj}).$$

3.2. *CEP estimation with parametric or known propensity score.* Suppose now that the propensity score $p(X)$ is correctly parameterized as $p(X;\gamma)$ up to a finite-dimensional unknown parameter γ. Theorems 2 and 5 show that the optimally weighted CEP-GMM estimator defined in (3) still achieves the semiparametric efficiency bound for β defined by moment condition (2). However, according to Theorems 3 and 5, such an estimator is no longer efficient for β defined through moment condition (1).

Rewriting moment condition (1) as $E\left[\mathcal{E}(X_i;\beta_0)\frac{p(X_i)}{p}\right] = 0$, we can again construct an efficient estimator for β_0 based on the sieve estimate $\hat{\mathcal{E}}(X;\beta)$ and the correctly specified parametric form $p(X;\gamma)$. In particular, the optimally weighted GMM estimator using the following sample moment condition will achieve the efficiency bound in Theorem 3 for β defined through (1):

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{E}}(X_i;\beta)\frac{p(X_i;\hat{\gamma})}{\hat{p}}, \tag{6}$$

where $\hat{p} = \frac{n_p}{n}$ and $\hat{\gamma}$ is the parametric MLE estimator that solves the score equation for γ:

$$\frac{1}{n}\sum_{i=1}^{n} S_{\hat{\gamma}}(D_i, X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i - p(X_i;\hat{\gamma})}{p(X_i;\hat{\gamma})(1-p(X_i;\hat{\gamma}))} p_{\hat{\gamma}}(X_i) = 0.$$

THEOREM 6. *Let $p(X;\gamma)$ be the parametric propensity score function known up to the parameters γ and let $E[S_{\gamma_0}(D,X)S_{\gamma_0}(D,X)']$ be positive definite. Let β_0 satisfy the moment condition (1) and $\widehat{\beta}$ be its CEP-GMM estimator using the sample moment (6). Under Assumptions 1, 2, 3 and 4, we have $\sqrt{n}(\widehat{\beta} - \beta_0) \Rightarrow \mathcal{N}(0, V)$, with*

$$V = (\mathcal{J}_\beta^{1\prime} W \mathcal{J}_\beta^1)^{-1} \mathcal{J}_\beta^{1\prime} W \tilde{\Omega}_\beta W \mathcal{J}_\beta^1 (\mathcal{J}_\beta^{1\prime} W \mathcal{J}_\beta^1)^{-1},$$

where $\tilde{\Omega}_\beta$ is given in Theorem 3. Further, if $W = \tilde{\Omega}_\beta^{-1}$, then $\sqrt{n}(\widehat{\beta} - \beta_0) \Rightarrow \mathcal{N}(0, V_0)$, where $V_0 = (\mathcal{J}_\beta^{1\prime} \tilde{\Omega}_\beta^{-1} \mathcal{J}_\beta^1)^{-1}$ is the efficiency variance bound given in Theorem 3.

The proof of this theorem is very similar to the previous ones and hence is omitted. It suffices to point out that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\mathcal{E}}(X_i;\beta_0)\frac{p(X_i;\hat{\gamma})}{\hat{p}}$ can be shown to be asymptotically equivalent to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{p(X_i)}{p}\left\{\mathcal{E}(X_i;\beta_0) + \frac{1-D_i}{1-p(X_i)}\left[m(Z_i;\beta_0) - \mathcal{E}(X_i;\beta_0)\right]\right\} + E\left[\mathcal{E}(X_i;\beta_0)\frac{p_\gamma(X_i)}{p}\right]\sqrt{n}(\hat{\gamma} - \gamma_0),$$

where

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = \left[E S_{\gamma_0}(D,X) S_{\gamma_0}(D,X)'\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{\gamma_0}(D_i, X_i) + o_p(1).$$

We remark that even when a parametric assumption is being made about the propensity score $p(X;\gamma)$ (in fact even if in addition $f(Y)$ is assumed to be a parametric likelihood), the inference about β is still semiparametric. This is because the marginal density $f(X)$ is still nonparametric and contains semiparametric information about β. This explains why nonparametric estimation is still needed to achieve the efficiency bound for β.

The case where the propensity is fully known can be considered a special case of parametric propensity score where the parameters are known. In this case, the efficient moment condition is as in (6) after replacing $p(X_j;\hat{\gamma})$ with the known $p(X_j)$.

Remark: When the auxiliary data set is a validation data set, e.g. $p(X) = p$, the parameters β defined by both moment conditions (1) and (2) coincide. Therefore the CEP-GMM estimator defined in (3) when we take $n_p = n$ and the summation to be over the all observations will achieve semiparametric efficiency.

4. IPW-GMM Estimation. An alternative estimation method for β is the inverse probability weighting based GMM (IPW-GMM). Several authors have considered inverse probability weighting paired with a conditional independence assumption for estimation in presence of missing information. Recent examples include parametric IPW as in [24], [34], [35] and [32], for missing data models, and nonparametric inverse probability weighting as in [18] for the case of mean treatment effect analysis. In this section, we extend existing results and first show that the optimally weighted IPW-GMM estimator of β is semiparametrically efficient when the propensity score is unknown. The same estimator, however, will be generally inefficient when the propensity score is known or belongs to a correctly specified parametric family; combinations of nonparametric and known or parametric estimated propensity scores are needed to achieve the semiparametric efficiency bounds for these cases.

4.1. *Efficient estimation with unknown propensity score.* The IPW-GMM method uses the fact that under Assumption 2, moment condition (1) can be rewritten as:

$$E\left[m(Z;\beta)|\, D=1\right] = E\left[m(Z;\beta)\frac{p(X)(1-p)}{(1-p(X))p}\middle|\, D=0\right]; \tag{7}$$

while moment condition (2) is equivalent to:

$$E\left[m(Z;\beta)\right] = E\left[m(Z;\beta)\frac{1-p}{1-p(X)}\middle|\, D=0\right]. \tag{8}$$

Let $\hat{p}(X)$ be a consistent estimate of the true propensity score. Then we can estimate β_0 defined by case (1) using GMM with the following sample moment:

$$\sqrt{n}\frac{1}{n_a}\sum_{j=1}^{n_a} m(Z_j;\beta)\frac{\hat{p}(X_j)}{1-\hat{p}(X_j)}\frac{1-\hat{p}}{\hat{p}}, \tag{9}$$

and estimate β_0 defined by case (2) using GMM with the following sample moment:

$$\sqrt{n}\frac{1}{n_a}\sum_{j=1}^{n_a} m(Z_j;\beta)\frac{1-\hat{p}}{1-\hat{p}(X_j)}. \tag{10}$$

The inverse probability weighting approach is considered semiparametric when $\hat{p}(X)$ is estimated nonparametrically. In this case, it can be shown that the sample moment (9) evaluated at β_0 is asymptotically equivalent to

$$\frac{1}{p}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(1-D_i)m(Z_i;\beta_0)\frac{p(X_i)}{1-p(X_i)} + \mathcal{E}(X_i;\beta_0)\frac{D_i-p(X_i)}{1-p(X_i)}\right] + o_p(1).$$

The two components of this influence function are negatively correlated. Because of this, the asymptotic variance might be smaller than that of the estimator of β_0 based

on moment condition (9) with the known $p(X)$. The influence function can be rewritten in terms of two orthogonal terms as

$$\frac{1}{p}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{(1-D_i)\left[m(Z_i;\beta_0)-\mathcal{E}(X_i;\beta_0)\right]\frac{p(X_i)}{1-p(X_i)}+D_i\mathcal{E}(X_i;\beta_0)\right\}+o_p(1),$$

which is identical to the influence function in (4). Therefore,

$$Avar(\sqrt{n}\frac{1}{n_a}\sum_{j=1}^{n_a}m(Z_j;\beta_0)\frac{\hat{p}(X_j)}{1-\hat{p}(X_j)}\frac{1-\hat{p}}{\hat{p}})=\Omega^1_\beta,$$

where Ω^1_β is given in Theorem 1. An optimally weighted GMM estimator for β_0 defined by case (1) using sample moment (9) should then achieve the semiparametric efficiency bound stated in Theorem 1.

The influence function representation for sample moment (10) can be calculated as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(1-D_i)m(Z_i;\beta_0)\frac{1}{1-p(X_i)}+\mathcal{E}(X_i;\beta_0)\frac{D_i-p(X_i)}{1-p(X_i)}\right]+o_p(1),$$

whose two components are again negatively correlated. As in the previous case, the influence function can be written in terms of two orthogonal components as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{(1-D_i)\left[m(Z_i;\beta_0)-\mathcal{E}(X_i;\beta_0)\right]\frac{1}{1-p(X_i)}+\mathcal{E}(X_i;\beta_0)\right\}+o_p(1),$$

which is identical to the influence function in (5). Hence, an optimally weighted GMM estimator for β_0 defined by case (2) using sample moment (10) achieves the semiparametric efficiency bound for case (1) stated in Theorem 1.

In this subsection, to emphasize that the true propensity score function is unknown and has to be estimated nonparametrically, we use $p_o(x)\equiv E(D|X=x)$ to indicate the true propensity score and $p(x)$ to denote any candidate function. (Note that to save notations in the rest of the main text $p(x)$ denotes the true propensity score.) Let $\widehat{p}(\cdot)$ be a sieve estimator of $p_o(x)$ that uses the combined sample $\{(D_i,X_i):i=1,...,n\}$. Let $\{Z_{ai}=(Y_{ai},X_{ai}):i=1,...,n_a\}$ be the auxiliary (i.e. $D=0$) data set. We define the IPW-GMM estimator $\widehat{\beta}$ for moment condition (1) as

$$\widehat{\beta}=\arg\min_{\beta\in B}\left(\frac{1}{n_a}\sum_{i=1}^{n_a}m(Z_{ai};\beta)\frac{\widehat{p}(X_{ai})}{1-\widehat{p}(X_{ai})}\right)'\widehat{W}\left(\frac{1}{n_a}\sum_{i=1}^{n_a}m(Z_{ai};\beta)\frac{\widehat{p}(X_{ai})}{1-\widehat{p}(X_{ai})}\right) \tag{11}$$

and the IPW-GMM estimator $\widehat{\beta}$ for moment condition (2) as

$$\widehat{\beta}=\arg\min_{\beta\in B}\left(\frac{1}{n_a}\sum_{i=1}^{n_a}m(Z_{ai};\beta)\frac{1}{1-\widehat{p}(X_{ai})}\right)'\widehat{W}\left(\frac{1}{n_a}\sum_{i=1}^{n_a}m(Z_{ai};\beta)\frac{1}{1-\widehat{p}(X_{ai})}\right). \tag{12}$$

There are two popular sieve nonparametric estimators of $p_o()$:

(i) a sieve Least Squares (LS) estimator $\widehat{p}_{ls}(x)$ as in [16]:

$$\widehat{p}_{ls} = \arg \min_{p() \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} (D_i - p(X_i))^2/2.$$

In the appendix we establish the consistency and convergence rate of $\widehat{p}_{ls}(x)$ under the assumption that the variables in X have unbounded support.

(ii) a sieve Maximum Likelihood (ML) estimator $\widehat{p}_{mle}(x)$ as in [18]:

$$\begin{aligned} \widehat{p}_{mle} &= \arg \max_{p() \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \{D_i \log[p(X_i)] + (1 - D_i) \log[1 - p(X_i)]\}, \\ \mathcal{H}_n &= \left\{ h \in \Lambda_c^\gamma(\mathcal{X}) : h(x) = [A^{k_n}(x)'\pi]^2 \right\} \text{ or } \left\{ h \in \Lambda_c^\gamma(\mathcal{X}) : h(x) = \exp(A^{k_n}(x)'\pi) \right\}. \end{aligned}$$

Recall $E_a(\cdot) = \int (\cdot) f_{X|D=0}(x)dx$. Define a weighted sup-norm as, for some $\omega > 0$,

$$||h||_{\infty,\omega} \equiv \sup_{x \in \mathcal{X}} \left| h(x)[1 + |x|^2]^{-\omega/2} \right|.$$

Assumption 5. *Let $\widehat{W} - W = o_p(1)$ for a positive semidefinite matrix W, and the following hold: (1) $p_o()$ belongs to a Hölder ball $\mathcal{H} = \{p() \in \Lambda_c^\gamma(\mathcal{X}) : 0 < \underline{p} \leq p(x) \leq \overline{p} < 1\}$ for some $\gamma > 0$; (2) $\int (1 + |x|^2)^\omega f_X(x)dx < \infty$ for some $\omega > 0$; (3) there is a non-increasing function $b(\cdot)$ such that $b(\delta) \to 0$ as $\delta \to 0$ and*

$$E_a \left[\sup_{||\beta - \widetilde{\beta}|| < \delta} ||m(Z_i; \beta) - m(Z_i, \widetilde{\beta})||^2 \right] \leq b(\delta)$$

for all small positive value δ; (4) $E_a \left[\sup_{\beta \in B} ||m(Z_i; \beta)||^2 \right] < \infty$; (5) for any $h \in \mathcal{H}$, there is a sequence $\Pi_{\infty n} h \in \mathcal{H}_n$ such that $||h - \Pi_{\infty n} h||_{\infty,\omega} = o(1)$.

Theorem 7. *Let $\widehat{\beta}$ be the IPW-GMM estimator given in (11) or (12). Under Assumptions 2, 1 and 5, if $\frac{k_n}{n} \to 0$, $k_n \to \infty$, then: $\widehat{\beta} - \beta_0 = o_p(1)$.*

Let $E(\cdot) = \int (\cdot) f_X(x)dx$, $||h||_2 = \sqrt{\int h(x)^2 f_X(x)dx}$, and $\Pi_{2n} h$ be the projection of h onto the closed linear span of $q^{k_n}(x) = (q_1(x), ..., q_{k_n}(x))'$ under the norm $||\cdot||_2$. We need the following additional assumptions to obtain asymptotic normality.

Assumption 6. *: Let $\beta_0 \in int(B)$, $E[\frac{p_o(X)}{1-p_o(X)} \mathcal{E}(X; \beta_0) \mathcal{E}(X; \beta_0)']$ be positive definite, and the following hold: (1) Assumptions 5.1 and 5.2 are satisfied with $\gamma > d_x/2$ and $\omega > \gamma$; (2) There exist a constant $\epsilon \in (0, 1]$ and a small $\delta_0 > 0$ such that*

$$E_a \left[\sup_{||\beta - \widetilde{\beta}|| < \delta} ||m(Z_i; \beta) - m(Z_i, \widetilde{\beta})||^2 \right] \leq const.\delta^\epsilon$$

for any small positive value $\delta \leq \delta_0$*; (3)* $E_a\left[\sup_{\beta\in B:||\beta-\beta_0||\leq\delta_0}||m(Z_i;\beta)||^2(1+|X_i|^2)^\omega\right] < \infty$ *for some small* $\delta_0 > 0$*; (4)* $E\left[\left\|\frac{\partial\mathcal{E}(X;\beta_0)}{\partial\beta}\right\|(1+|X|^2)^{\frac{\omega}{2}}\right] < \infty$*, and for all* $x \in \mathcal{X}$*,* $\frac{\partial\mathcal{E}(x;\beta)}{\partial\beta}$ *is continuous around* β_0*; (5)* $k_n = O\left(n^{\frac{d_x}{2\gamma+d_x}}\right)$*,* $n^{-\frac{\gamma}{2\gamma+d_x}} \times \left\|\frac{\mathcal{E}(\cdot;\beta_o)}{1-p_o(\cdot)} - \Pi_{2n}\frac{\mathcal{E}(\cdot;\beta_o)}{1-p_o(\cdot)}\right\|_2 = o(n^{-1/2})$*. (6) either one of the following is satisfied: (6a)* $\sup_{\beta\in B:||\beta-\beta_0||\leq\delta_0}\sup_{x\in\mathcal{X}}||\mathcal{E}(x,\beta)|| \leq const. < \infty$ *for some small* $\delta_0 > 0$*; (6b)* $E_a\left[\sup_{\beta\in B:||\beta-\beta_0||\leq\delta_0}||\mathcal{E}(X,\beta)||^4\right] \leq const. < \infty$ *for some small* $\delta_0 > 0$*, and* $f_{X|D=0}(\cdot) \in \Lambda_c^\gamma(\mathcal{X})$ *with* $\gamma > 3d_x/4$*; (6c)* $E_a\left[\sup_{\beta\in B:||\beta-\beta_0||\leq\delta_0}||\mathcal{E}(X,\beta)||^2\right] \leq const. < \infty$ *for some small* $\delta_0 > 0$*, and* $f_{X|D=0}(\cdot) \in \Lambda_c^\gamma(\mathcal{X})$ *with* $\gamma > d_x$*.*

THEOREM 8. *Let* $\widehat{\beta}$ *be the IPW-GMM estimator given in (11) or (12). Under Assumptions 1, 2, 5 and 6, we have* $\sqrt{n}(\widehat{\beta}-\beta_0) \Rightarrow \mathcal{N}(0,V)$*, with* V *the same as in Theorem 5.*

Remark 3: **(i)** The weighting ω is needed since the support of the conditioning variable X is assumed to be the entire Euclidean space. When X has bounded support and $f_{X|D=0}$ is bounded above and below over its support, we can simply set $\omega = 0$ in Assumptions 5 and 6 and replace 6.1 with the assumption that 5.1 holds with $\gamma > d_x/2$. Note that Assumption 6.6a is easily satisfied when X has compact support. When $\mathcal{X} = \mathcal{R}^{d_x}$, Assumption 6.6a rules out $\mathcal{E}(x,\beta)$ being linear in x; Assumptions 6.6b or 6.6c allow for linear $\mathcal{E}(x,\beta)$ but need smoother propensity score $p(x)$ and density $f_{X|D=0}$. **(ii)** Assumptions 5 and 6 again allow for non-smooth moment conditions. **(iii)** Since $\frac{f_{X|D=0}(X)}{f_X(X)} = \frac{1-p_o(X)}{1-p}$, the assumption $0 < \underline{p} \leq p_o(x) \leq \overline{p} < 1$ implies that $\frac{1-\overline{p}}{1-p} \leq \frac{f_{X|D=0}(X)}{f_X(X)} \leq \frac{1-\underline{p}}{1-p}$, hence $E()$ and $E_a()$ in Assumptions 5 and 6 are effectively equivalent. **(iv)** Although Assumption 5.1 imposes the same strong condition $0 < \underline{p} \leq p_o(x) \leq \overline{p} < 1$ as that typically assumed in the program evaluation literature, unlike most existing papers on estimation of average treatment effects, our paper allows for unbounded support of X and assumes weaker smoothness on $p_o(x)$ and $\mathcal{E}(\cdot;\beta_o)$. In particular, if we let $k_n = O\left(n^{\frac{d_x}{2\gamma+d_x}}\right)$, the growth order which leads to the optimal convergence rate of $||\widehat{p}()-p_o()||_2 = O_p\left(n^{-\frac{\gamma}{2\gamma+d_x}}\right)$, then Assumption 6.5 is satisfied with $\left\|\frac{\mathcal{E}(\cdot;\beta_o)}{1-p_o(\cdot)} - \Pi_{2n}\frac{\mathcal{E}(\cdot;\beta_o)}{1-p_o(\cdot)}\right\|_2 = o\left(n^{-\frac{d_x}{2(2\gamma+d_x)}}\right) = o\left(k_n^{-\frac{1}{2}}\right)$.

4.2. *IPW Estimation with parametric or known propensity score.* The case of moment condition (2) is simpler and therefore we briefly discuss it first. Theorems 1 and 2 have shown that knowledge about the propensity score does not change the semiparametric efficiency bound. Furthermore, theorems 5 and 8 show that both a nonparametric CEP-GMM estimator and a nonparametric IPW-GMM estimator for β achieve this semiparametric efficiency bound regardless of whether the propensity score is unknown, known or parametrically specified. The following theorem also states, without proof, the interesting result that the parametric IPW estimator using $p(X;\hat{\gamma})$ is in fact less

efficient than the one using a nonparametric estimate $\hat{p}(X)$ in (10), but is more efficient than the one using the known $p(X)$.

THEOREM 9. *Suppose the parametric model $p(X_i;\gamma)$ is correctly specified. Suppose also that $E[S_\gamma(D,X)S_\gamma(D,X)']$ is positive definite. Under moment condition (2) and using the optimally weighted sample moment condition (10), an IPW-GMM estimator for β using a parametric estimate of $p(X_i;\hat{\gamma})$ in place of $\hat{p}(X_i)$ in (10) is more efficient than the one using the known $p(X_i)$, but is less efficient than the one using a nonparametric estimate $\hat{p}(X_i)$ of the propensity score.*

This result is based on the following relations, which hold asymptotically

$$Avar(\frac{\sqrt{n}}{n_a}\sum_{j=1}^{n_a} m(Z_j;\beta_0)\frac{1-\hat{p}}{1-p(X_i;\hat{\gamma})}) \leq Avar(\frac{\sqrt{n}}{n_a}\sum_{j=1}^{n_a} m(Z_j;\beta_0)\frac{1-\hat{p}}{1-p(X_i)})$$

and

$$Avar(\frac{\sqrt{n}}{n_a}\sum_{j=1}^{n_a} m(Z_j;\beta_0)\frac{1-\hat{p}}{1-p(X_i;\hat{\gamma})}) \geq Avar(\frac{\sqrt{n}}{n_a}\sum_{j=1}^{n_a} m(Z_j;\beta_0)\frac{1-\hat{p}}{1-\hat{p}(X_i)}).$$

Now consider the more interesting case where moment condition (1) holds and sample moment condition (9) is used. Consider the case when the parametric propensity score is correctly specified. First, it is clear that the optimally weighted IPW-GMM estimator of β based on (9) that uses a nonparametric estimate of $\hat{p}(X)$ does not achieve the efficiency bound in Theorem 3, because we see from Theorem 8 that this estimator achieves instead the variance bound in Theorem 1, which is larger than the variance bound in Theorem 3.

However, the parametric two step IPW estimator that uses a parametric first step for $p(X;\gamma)$ does not achieve the efficiency bound in Theorem 3 either. To see this, note that the parametric two step IPW estimator is based on the moment condition

$$\sqrt{n}\frac{1}{n_a}\sum_{j=1}^{n_a} m(Z_j;\beta)\frac{p(X_j;\hat{\gamma})}{1-p(X_j;\hat{\gamma})}\frac{1-\hat{p}}{\hat{p}},$$

which has a linear influence function representation of

$$\frac{1}{p}\left[m(Z_i;\beta_0)\frac{(1-D_i)p(X_i)}{1-p(X_i)} + \text{Proj}(\mathcal{E}(X_i;\beta_0)\frac{D_i-p(X_i)}{1-p(X_i)}\Big| S_\gamma(D_i,X_i))\right],$$

where

$$\begin{aligned}\text{Proj}(\mathcal{E}(X_i;\beta_0)\frac{D_i-p(X_i)}{1-p(X_i)}\Big| S_\gamma(D_i,X_i)) &= E\left[\mathcal{E}(X;\beta_0)\frac{p_\gamma(X)}{1-p(X)}\right] \\ &\quad \times E\left[S_\gamma(D_i,X_i)S_\gamma(D_i,X_i)'\right]^{-1} S_\gamma(D_i,X_i)\end{aligned}$$

is the influence function from the first step estimation of γ. The difference between this influence function and the influence function for Theorem 3 can be verified to be equal to

$$\text{Res}((D-p(X))\frac{p(X)}{1-p(X)}\mathcal{E}(X;\beta_0)\Big|S_\gamma(D_i,X_i)),$$

which is obviously orthogonal to the influence function of Theorem 3. Therefore, the two step parametric IPW estimator has a variance larger than the efficiency bound under the assumption of correct specification of the parametric model for $p(X;\gamma)$.

An IPW type estimator that achieves the efficiency bound under correct specification can be obtained by combining both nonparametric and parametric estimates of the propensity score. Such an efficient moment condition is given by

$$\sqrt{n}\frac{1}{n_a}\sum_{j=1}^{n_a} m(Z_j;\beta)\frac{p(X_j;\hat{\gamma})}{1-\hat{p}(X_j)}\frac{1-\hat{p}}{\hat{p}}, \tag{13}$$

where $\hat{\gamma}$ is the maximum likelihood estimator for γ_0 and $\hat{p}(X)$ is the sieve estimate of the propensity score. This moment condition has the following asymptotic linear representation:

$$\begin{aligned}\tfrac{1}{p}\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n\left[(1-D_i)(m(Z_i;\beta_0)-\mathcal{E}(X_i;\beta_0))\tfrac{p(X_i)}{1-p(X_i)}+p(X_i)\mathcal{E}(X_i;\beta_0)\right]\\ +E\left[\tfrac{\mathcal{E}(X;\beta_0)}{p}p_\gamma(X_i)\right]\sqrt{n}(\hat{\gamma}-\gamma),\end{aligned}$$

which is identical to the influence function under correct parametric specification of $p(X;\gamma)$ leading to the semiparametric efficiency bound in Theorem 3.

The case where the propensity score is fully known can be considered a special case of parametric propensity score where the parameters are known. In this case, the efficient moment condition is as in (13) after replacing $p(X_j;\hat{\gamma})$ with the known $p(X_j)$.

It is finally worth noting that Assumption 2 is an identification assumption that is not testable. Therefore both the CEP-GMM estimator and the IPW-GMM estimator will converge to the same population limit regardless of whether Assumption 2 holds, as long as the same weighting matrix is being used. The population difference between CEP and IPW can only arise from the parametric mis-specification of the approximating models for $\mathcal{E}(X;\beta)$ and $p(X)$.

5. Empirical Illustration. We illustrate our method empirically using data from the Indian National Sample Survey (NSS), a data set routinely used by the Indian Government to monitor changes in the distribution of private consumption. Several researchers have argued that due to changes in the expenditure questionnaire adopted for data collection in the 1999-2000 round of the NSS, poverty estimates from this round are likely to be non-comparable with those from previous years. A change in the questionnaire section where food expenditures are recorded likely led to the overestimation of food consumption, and hence to the underestimation of poverty ([14], [32]). In other

words, a missing data problem arises because the variable of interest (total expenditure as it would have been recorded using a standard questionnaire) is not observed. [13], [12] and [32] argue that total expenditure in a set of miscellaneous items for which the questionnaire was not modified ("comparable items" hereafter) can be used as a proxy variable to produce an estimate of poverty for 1999-2000 that is comparable with previous years.

In this section, we assume that the researcher is interested in estimating the cumulative distribution function (cdf) for rural India in 1999-2000 of a measure of total monthly expenditure that is comparable with previous NSS rounds. In the terminology used in this article, this situation corresponds to a verify-out-of-sample case (1), where the parameter of interest β is identified in terms of a variable Y that is not observed in the primary sample (the 1999-2000 survey). The moment function takes the form $m(Z;\beta_0) = 1(Y \leq y) - \beta_0$, where y is a given threshold. We use the previous round of the NSS (1993-94) as auxiliary survey, and expenditure in "comparable items" as proxy variable X. The two rounds are independent cross-sections from the Indian population. The crucial identifying assumption is that the distribution of Y conditional on X remained stable between 1993-94 and 1999-2000 (for a discussion and indirect evidence is support of this assumption see [32]).

Table 1 reports point estimates and standard errors for the cdf of total (log) household expenditure at selected thresholds, using different estimators. The first column reports the cdf estimated using the noncomparable data from the primary sample. Column 2 reports CEP-GMM estimates, calculated using 3^{rd} order polynomial splines in expenditure in comparable items as sieve basis, with 10 knots at the equal range quantiles of the empirical distribution of the proxy variables. Column 3 reports estimates obtained using moment condition (7), but with a nonparametric first step where we estimate $P(X)$ using sieve-logit, including the basis functions we used for CEP-GMM as regressors. In Column 4 we impose a parametric model, and we estimate the propensity score using logit, with X entered linearly in the single index. Column 5 reports the results for the estimator described in Section 3.2, which is efficient when a parametric model is correctly specified for $P(X)$.

For values of Y below 7 the adjusted estimates of the cdf in columns 2 to 4 are larger than the unadjusted figures presented in column 1, even if only by one-two percentage points. This is consistent with what would be expected based on the hypothesis that the change in the questionnaire introduced in 1999-2000 led to an overstatement of total reported expenditure with respect to previous NSS rounds. As expected, CEP and IPW non-parametric estimators produce virtually identical results. The estimates in columns 4 and 5 impose a simple logit for the propensity score, but they are still very similar. In the verify-out-of-sample case, knowledge of a parametric form for $P(X)$ lowers the semiparametric efficiency bound, and this may explain why in some cases the standard errors in column 4 are lower than for the estimators in columns 2 and 3, which are only efficient when $P(X)$ is unknown. Notice also that when the parametric assumption is correct the efficient estimator is the one in column 5. Indeed the standard errors for this estimator are always lower or virtually identical to those in column 4 every time this

latter estimator is more precise than the nonparametric estimators in columns 2 and 3.

6. Conclusions. We derive semiparametric efficiency bounds for the estimation of parameters defined through general nonlinear, possibly non-smooth and over-identified moment conditions, when variables in the primary sample of interest are missing. For identification we rely on the validity of a conditional independence assumption and on the availability of an auxiliary sample that contains information on the relation between missing variables and other proxy variables that are also observed in the primary sample. We study two alternative frameworks. In the first case ("verify-out-of-sample") validation is done with an auxiliary data set which is independent from the primary data set of interest. In the second case ("verify-in-sample") a subset of the observations in the primary sample is validated.

We show that the optimally weighted CEP-GMM estimators achieve the semiparametric efficiency bounds when the propensity score is unknown, or is known or belongs to a correctly specified parametric family. These estimators only use a nonparametric estimate of the conditional expectation of the moment functions, and their asymptotic efficiency is obtained under regularity conditions weaker than the existing ones in the literature. In particular, these CEP-GMM estimators still achieve efficiency bounds when proxy (conditioning) variables have unbounded supports and moment conditions are not smooth.

We also prove that an optimally weighted IPW-GMM estimator is semiparametrically efficient with fully unknown propensity score. However, this estimator is not efficient when the propensity score is either known, or is parametrically estimated using a *correctly* specified parametric model; in such instances, appropriate combinations of nonparametric and parametric estimates of the propensity score are needed to achieve the efficiency bounds.

We have also demonstrated that, from the theoretical point of view, the CEP-GMM estimators are more attractive than the IPW-GMM estimators. Recently and independently [19] advocated a similar sieve conditional expectation projection based estimator for the average treatment effect parameter in program evaluation applications. Also, for the estimation of the average treatment effects in missing data models, [33] suggested that a semiparametrically specified propensity score, such as a single index or a partially linear form, can be used to reduce the curse of dimensionality in the nonparametric estimation of the propensity score. An interesting topic for future research is to study the efficiency implications of these semiparametric restrictions on the propensity score.

REFERENCES

[1] Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71** 1795–1843.

[2] Bickel, P. J., Klaassen, C. A., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer.

[3] Breslow, N. E., McNeney, B. and Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Annals of Statistics*, **31** 1110–1139.

[4] Breslow, N. E., Robins, J. M. and Wellner, J. A. (2000). On the semiparametric efficiency of logistic regression under case-control sampling. *Bernoulli*, **6** 447–455.
[5] Carroll, R., Ruppert, D. and Stefanski, L. (1995). *Measurement error in nonlinear models*. No. 63 in Monographs on Statistics and Applied Probability, Chapman and Hall, New York.
[6] Carroll, R. and Wand, M. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society*, **53** 573–585.
[7] Chen, J. B. and Breslow, N. E. (2004). Semiparametric efficient estimation for the auxiliary outcome problem with the conditional mean model. *Canadian Journal of Statistics*, **32** 359–372.
[8] Chen, X., Hong, H. and Tamer, E. (2005). Measurement Error Models with Auxiliary Data. *Review of Economic Studies*, **72** 343–366.
[9] Chen, X., Linton, O. and van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71** 1583–1600.
[10] Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, **66** 289–314.
[11] Clogg, C., Rubin, D., Schenker, N., Schultz, B. and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using bayesian logistic regression. *Journal of The American Statistical Association*, **86** 68–78.
[12] Deaton, A. (2003). Adjusted Indian Poverty Estimates for 1999-2000. *Economic and Political Weekly* 322–326.
[13] Deaton, A. and Drèze, J. (2002). Poverty and Inequality in india, a Re-Examination. *Economic and Political Weekly* 3729–3748. September 7.
[14] Deaton, A. and Kozel, V. (2005). Data and Dogma: The Great indian Poverty Debate. New Delhi, India, MacMillian.
[15] Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, **55** 363–390.
[16] Hahn, J. (1998). On the role of propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, **66** 315–332.
[17] Heckman, J., LaLonde, R. and Smith, J. (1999). The economics and econometrics of active labor market programs. In *Handbook of Labor Economics, Vol. 3A* (O. Ashenfelter and D. Card, eds.). Elsevier Science.
[18] Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **71** 1161–1189.
[19] Imbens, G., Newey, W. and Ridder, G. (2005). Mean-squared-error calculations for average treatment effects. Working Paper.
[20] Lee, L. and Sepanski, J. (1995). Estimation of linear and nonlinear errors-invariables models using validation data. *Journal of the American Statistical Association*, **90** 130–140.
[21] Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. 2nd ed. John Wiley & Sons, New York.
[22] Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, **5** 99–135.
[23] Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, **62** 1349–82.
[24] Robins, J., Mark, S. and Newey, W. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, **48** 479–95.
[25] Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90** 122–129.
[26] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89** 846–866.
[27] Rotnitzky, A. and Robins, J. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, **82** 805–820.
[28] Schenker, N. (2003). Assessing variability due to race bridging: application to census counts and vital rates for the year 2000. *Journal of The American Statistical Association*, **98** 818–828.
[29] Sepanski, J. and Carroll, R. (1993). Semiparametric quasi-likelihood and variance estimation

in measurement error models. *Journal of Econometrics*, **58** 223–256.
[30] Shen, X. (1997). On Methods of Sieves and Penalization. *Annals of Statistics*, **25** 2555–2591.
[31] Shen, X. and Wong, W. (1994). Convergence Rates of Sieve Estimates. *Annals of Statistics*, **22** 580–615.
[32] Tarozzi, A. (2006). Calculating comparable statistics from incomparable surveys, with an application to poverty in india. *Journal of Business and Economic Statistics*. Forthcoming.
[33] Wang, Q., Linton, O. and Hardle, W. (2004). Semiparametric regression analysis for missing response data. *Journal of the American Statistical Association*, **99** 334–345.
[34] Wooldridge, J. (2002). Inverse probability weighted m-estimators for sample selection, attrition and stratification. *Portuguese Economic Journal*, **1** 117–139.
[35] Wooldridge, J. (2003). Inverse probability weighted estimation for general missing data problems. Manuscript, Michigan State University.

Appendix A - Calculation of Efficiency Bounds.

Proof. **Theorem 1**

We follow closely the structure of semiparametric efficiency bound derivation of [22] and [2].

Case (1). Consider a parametric path θ for the joint distribution of Y, D and X. Define $p_\theta = P_\theta(D=1)$. The joint density function for Y, D and X is given by

$$f_\theta(y,x,d) = p_\theta^d (1-p_\theta)^{1-d} f_\theta(x|D=1)^d f_\theta(x|D=0)^{1-d} f(y|x)^{1-d}. \tag{14}$$

The resulting score function is given by

$$S_\theta(d,y,x) = \frac{d-p_\theta}{p_\theta(1-p_\theta)}\dot{p}_\theta + (1-d)s_\theta(x|D=0) + ds_\theta(x|D=1) + (1-d)s_\theta(y \mid x),$$

where $s_\theta(y \mid x) = \frac{\partial}{\partial\theta}\log f_\theta(y \mid x)$, $\dot{p}_\theta = \frac{\partial}{\partial\theta}p_\theta$, $s_\theta(x|d) = \frac{\partial}{\partial\theta}\log f_\theta(x|d)$. The tangent space of this model is therefore given by:

$$\mathcal{T} = a(d-p_\theta) + (1-d)s_\theta(x|D=0) + (1-d)s_\theta(y|x) + ds_\theta(x|D=1), \tag{15}$$

where $\int s_\theta(y \mid x) f_\theta(y \mid x)\, dy = 0$, $\int s_\theta(x|d) f_\theta(x|d)\, dx = 0$, and a is a finite constant.

Consider first the case when the model is exactly identified. In this case β is uniquely identified by condition (1). Differentiating under the integral gives

$$\frac{\partial\beta(\theta)}{\partial\theta} = -\left(\mathcal{J}_\beta^1\right)^{-1} E\left[m(Z;\beta)\frac{\partial\log f_\theta(Y,X \mid D=1)}{\partial\theta'} \mid D=1\right]. \tag{16}$$

The second component of the right hand side of this expression can be calculated as

$$E\left[m(Z;\beta)s_\theta(Y \mid X)' \mid D=1\right] + E\left[m(Z;\beta)s_\theta(X \mid D=1)' \mid D=1\right] \tag{17}$$

Pathwise differentiability follows if we can find $\Psi^1(Y,X,D) \in \mathcal{T}$ such that

$$\partial\beta(\theta)/\partial\theta = E\left[\Psi^1(Y,X,D)\, S_\theta(Y,X,D)'\right] \tag{18}$$

Define $p_\theta = \int p_\theta(x) f_\theta(x)\,dx$, $\mathcal{E}_\theta(X) = E[m(Z;\beta) \mid X]$. It can be verified that pathwise differentiability is satisfied by choosing: $\Psi^1(Y,X,D) = -\left(\mathcal{J}_\beta^1\right)^{-1} F_\beta^1(Y,X,D)$ where

$$F_\beta^1(Y,X,D) = \frac{1-D}{p}\frac{p(X)}{1-p(X)}\left[m(Z;\beta) - \mathcal{E}(X)\right] + \frac{\mathcal{E}(X)}{p}D \tag{19}$$

Since $\mathcal{J}_\beta^1$ is a nonsingular transformation, this can be shown proving that

$$E\left[m(Z;\beta)\frac{\partial}{\partial\theta'}\log f_\theta(Y,X \mid D=1) \mid D=1\right] = E\left[F_\beta^1(Y,X,D)\,S_\theta(Y,X,D)'\right]. \tag{20}$$

This can in turn be verified by checking that

$$\begin{aligned} E\left[m(Z;\beta)s_\theta(Y \mid X)' \mid D=1\right] &= E\left[\frac{1-D}{p}\frac{p(X)}{1-p(X)}\left[m(Z;\beta) - \mathcal{E}(X)\right]s_\theta(Y|X)'\right], \\ E\left[m(Z;\beta)s_\theta(X \mid D=1)' \mid D=1\right] &= E\left[\frac{\mathcal{E}(X)}{p}Ds_\theta(X|D=1)'\right]. \end{aligned}$$

Now one can also verify that $F_\beta^1(Y,X,D)$ belongs to the tangent space $\mathcal{T}$ in equation (15), with the first and second terms of $F_\beta^1(Y,X,D)$ taking the role of $(1-d)s_\theta(y|x)$ and $ds_\theta(X|D=1)$ respectively, and the two other components in (15) being identically equal to 0.

Therefore all the conditions of Theorem 3.1 in [22] hold, and the efficiency bound for regular estimators of the parameter β is given by

$$V_1 = \left(\mathcal{J}_\beta^1\right)^{-1} E\left[F_\beta^1(Y,X,D)\,F_\beta^1(Y,X,D)'\right]\left(\mathcal{J}_\beta^1\right)'^{-1} = \left(\mathcal{J}_\beta^1\right)^{-1}\Omega_\beta^1\left(\mathcal{J}_\beta^1\right)'^{-1}. \tag{21}$$

Case (2). For this case we use an alternative factorization of the likelihood function. Define $p_\theta(x) = P_\theta(D=1 \mid x)$. The joint density function for Y, D and X is given by

$$f_\theta(y,x,d) = f_\theta(x)\,p_\theta(x)^d\left[1-p_\theta(x)\right]^{1-d} f_\theta(y \mid x)^{1-d}. \tag{22}$$

The resulting score function is then given by

$$S_\theta(d,y,x) = (1-d)\,s_\theta(y \mid x) + \frac{d-p_\theta(x)}{p_\theta(x)\left(1-p_\theta(x)\right)}\dot{p}_\theta(x) + t_\theta(x),$$

where $s_\theta(y \mid x) = \frac{\partial}{\partial\theta}\log f_\theta(y \mid x)$, $\dot{p}_\theta(x) = \frac{\partial}{\partial\theta}p_\theta(x)$, $t_\theta(x) = \frac{\partial}{\partial\theta}\log f_\theta(x)$.

The tangent space of this model is therefore given by:

$$\mathcal{T} = \left\{(1-d)\,s_\theta(y \mid x) + a(x)\left(d - p_\theta(x)\right) + t_\theta(x)\right\} \tag{23}$$

where $\int s_\theta(y \mid x) f_\theta(y \mid x)\,dy = 0$, $\int t_\theta(x) f_\theta(x)\,dx = 0$, and $a(x)$ is any square integrable function.

In case (2), equation (16) is replaced by:

$$\begin{aligned}\frac{\partial\beta(\theta)}{\partial\theta} &= -\left(\mathcal{J}_\beta^2\right)^{-1} E\left[m(Z;\beta)\frac{\partial\log f_\theta(Y,X)}{\partial\theta'}\right]. \\ &= -\left(\mathcal{J}_\beta^2\right)^{-1}\left\{E\left[m(Z;\beta)\, s_\theta(Y\mid X)'\right]+E\left[\mathcal{E}(X)\, t_\theta(X)'\right]\right\}.\end{aligned} \tag{24}$$

Now we replace $F_\beta^1(Y,X,D)$ in (19) with the following:

$$F_\beta^2(Y,X,D)=\frac{1-D}{1-p(X)}\left[m(Z;\beta)-\mathcal{E}(X)\right]+\mathcal{E}(X) \tag{24}$$

and then it can be verified that $E\left[F_\beta^2(Y,X,D)\, S_\theta(Y,X,D)'\right]=E\left[m(Z;\beta)\frac{\partial\log f_\theta(Y,X)}{\partial\theta'}\right]$. Then the efficient influence function for case (2) is equal to $-\left(\mathcal{J}_\beta^2\right)^{-1}F_\beta^2(Y,X,D)$ with the two terms being orthogonal to each other, and the second result in Theorem 1 follows.

Now consider overidentified moment conditions. We only consider case (1), as the derivation for case (2) is analogous. When $d_m > d_\beta$, the moment conditions in (1) is equivalent to the requirement that for any matrix $\mathcal{A}$ of dimension $d_\beta \times d_m$ the following exactly identified system of moment conditions holds $\mathcal{A}E\left[m(Z;\beta)\mid D=1\right]=0$. Differentiating again,

$$\frac{\partial\beta(\theta)}{\partial\theta}=-\left(\mathcal{A}E\left[\frac{\partial m(Z;\beta)}{\partial\beta}\mid D=1\right]\right)^{-1}E\left[\mathcal{A}m(Z;\beta)\frac{\partial\log f_\theta(Y,X\mid D=1)}{\partial\theta'}\mid D=1\right].$$

Therefore, any regular estimator for β is asymptotically linear with influence function of the form

$$-\left(\mathcal{A}E\left[\frac{\partial m(Z;\beta)}{\partial\beta}\mid D=1\right]\right)^{-1}\mathcal{A}m(z;\beta).$$

For a given matrix $\mathcal{A}$, the projection of the above influence function onto the tangent set follows from the previous calculations, and is given by $-\left[\mathcal{A}\mathcal{J}_\beta^1\right]^{-1}F_\beta^1(y,x,d)$. The asymptotic variance corresponding to this efficient influence function for fixed $\mathcal{A}$ is therefore

$$\left[\mathcal{A}\mathcal{J}_\beta^1\right]^{-1}\mathcal{A}\Omega\mathcal{A}'\left[\mathcal{J}_\beta^{1\prime}\mathcal{A}'\right]^{-1}, \tag{23}$$

where $\Omega=E\left[F_\beta^1(Y,X,D)\,F_\beta^1(Y,X,D)'\right]$ as calculated above. Therefore, the efficient influence function is obtained when $\mathcal{A}$ minimizes (23). It is easy to show that such matrix $\mathcal{A}$ is equal to $\mathcal{J}_\beta^{1\prime}\Omega^{-1}$, so that the asymptotic variance becomes $V=\left(\mathcal{J}_\beta^{1\prime}\Omega^{-1}\mathcal{J}_\beta^1\right)^{-1}$. In fact, a standard textbook calculation shows

$$\begin{aligned}&\mathcal{J}_\beta^{1\prime}\Omega^{-1}\mathcal{J}_\beta^1-\mathcal{J}_\beta'\mathcal{A}'\left(\mathcal{A}\Omega\mathcal{A}'\right)^{-1}\mathcal{A}\mathcal{J}_\beta \\ =\;&\left(\mathcal{J}_\beta^{1\prime}\Omega^{-1/2}-\mathcal{J}_\beta^{1\prime}\Omega^{-1/2}\Omega^{1/2\prime}\left(\Omega^{1/2}\Omega^{1/2\prime}\right)^{-1}\Omega^{1/2}\right) \\ &\qquad\left(\Omega^{-1/2}\mathcal{J}_\beta^1-\Omega^{1/2\prime}\left[\Omega^{1/2}\Omega^{1/2\prime}\right]^{-1}\Omega^{1/2}\Omega^{-1/2}\mathcal{J}_\beta^1\right)\geq 0.\end{aligned}$$

□

PROOF. **Theorem 2**

As for Theorem 1, it suffices to present the proof for the case of exact identification, since the overidentified case follows from choosing the optimal linear combination matrix. If the propensity score $p(x)$ is known, the score becomes (c.f. [16]) $S_\theta(d,y,x) = (1-d)\,s_\theta(y \mid x) + t_\theta(x)$, so that the tangent space becomes $\mathcal{T} = \{(1-d)\,s_\theta(y \mid x) + t_\theta(x)\}$ where $\int s_\theta(y \mid x)\, f_\theta(y \mid x)\, dy = 0$, and $\int t_\theta(x)\, f_\theta(x)\, dx = 0$. Consider case (1) first. The pathwise derivative becomes

$$E\left[\frac{p(X)}{p} m(Z;\beta) s(Y \mid X)'\right] + E\left[\frac{p(X)}{p}\mathcal{E}(X)\, t(X)'\right]$$

Pathwise differentiability is established by verifying that equation (18) holds, with

$$F^1_\beta(y,x,d) = \frac{1-d}{p}\frac{p(x)}{1-p(x)}\left(m(z;\beta) - \mathcal{E}(x)\right) + \frac{\mathcal{E}(x)}{p} p(x). \tag{20}$$

Then the efficient influence function is as before equal to $-\left(\mathcal{J}^1_\beta\right)^{-1} F^1_\beta(y,x,d)$, and the result of Theorem 2 follows using Theorem 3.1 of [22].

Since $p(x)$ does not enter the definition of β in case (2), there is no change to the efficient influence function and to the semiparametric efficiency bound for that case.

□

PROOF. **Theorem 3**

When $p(X)$ belongs to a correctly specified parametric family $p(X;\gamma)$, the score function for moment (1) becomes

$$S_\theta(d,y,x) = (1-d)\, s_\theta(y \mid x) + \frac{d - p_\theta(x)}{p_\theta(x)\left(1-p_\theta(x)\right)}\frac{\partial p(x;\gamma)}{\partial \gamma'}\frac{\partial \gamma}{\partial \theta} + t_\theta(x).$$

The tangent space is therefore $\mathcal{T} = \{(1-d)\, s_\theta(y \mid x) + c' S_\gamma(d;x) + t_\theta(x)\}$ where c is a finite vector of constants and $S_\gamma(d;x)$ is the parametric score function. Now define $F^1_\beta(Y,X,D)$ as

$$\frac{1-D}{p}\frac{p(X)}{1-p(X)}\left[m(Z;\beta) - \mathcal{E}(X)\right] + \text{Proj}(\mathcal{E}(X)\frac{D-p(X)}{p}\Big| S_\gamma(D,X))$$

It is clear that $F^1_\beta(Y,X,D)$ lies in the tangent space. Also note that $\frac{\partial \beta(\theta)}{\partial \theta}$ can be written as

$$-\left(\mathcal{J}^1_\beta\right)^{-1}\left\{E\left[m(Z;\beta) s_\theta(Y \mid X)' \mid D=1\right] + E\left[m(Z;\beta)(t_\theta(x)' + S_\gamma(d;x)'\frac{\partial \gamma}{\partial \theta})\Big| D=1\right]\right\}.$$

The second term in curly brackets can also be written as

$$\frac{E(D-p(X))\mathcal{E}(X) S_\gamma(D;X)'}{p}\frac{\partial \gamma}{\partial \theta} + \frac{p(X)\mathcal{E}(X) t_\theta(X)}{p}.$$

With these calculations it can be verified that

$$\frac{\partial\beta(\theta)}{\partial\theta} = -\left(\mathcal{J}_\beta^1\right)^{-1} E\left[F_\beta^1(Y,X,D)\, S_\theta(Y,X,D)\right].$$

In particular,

$$E\left[\frac{(D-p(X))\mathcal{E}(X)S_\gamma(D;X)'}{p}\right] = E\left[\text{Proj}(\mathcal{E}(X)\frac{D-p(X)}{p}\middle| S_\gamma(D,X))S_\theta(Y,X,D)'\right].$$

Therefore $-\left(\mathcal{J}_\beta^1\right)^{-1} F_\beta^1(Y,X,D)$ is the desired efficient influence function and its variance is given as the efficient variance of Theorem 3. □

Appendix B - Proofs of Asymptotic Properties. In this appendix we establish the large sample properties for the IPW-GMM estimator with nonparametrically estimated propensity score function. Again to stress the fact that the true propensity score is unknown, in this appendix we denote the true propensity score by $p_o(x) \equiv E[D|X=x]$ and any candidate function by $p(x)$.

Denote $\mathcal{L}_2(\mathcal{X}) = \{h : \mathcal{X} \to \mathcal{R} : ||h||_2 = \sqrt{\int h(x)^2 f_X(x)dx} < \infty\}$ and $\mathcal{L}_{2,a}(\mathcal{X}) = \{h : \mathcal{X} \to \mathcal{R} : ||h||_{2,a} = \sqrt{\int h(x)^2 f_{X_a}(x)dx} < \infty\}$ as the two Hilbert spaces. We use $||h||_2 \asymp ||h||_{2,a}$ to mean that there are two positive constants c_1, c_2 such that $c_1||h||_2 \leq ||h||_{2,a} \leq c_2||h||_2$, which is true under the assumption $0 < \underline{p} \leq p_o(x) \leq \overline{p} < 1$.

Proposition B.1 provides large sample properties for the sieve LS estimator $\widehat{p}(x)$ of $p_o(x)$.

Proposition B.1: *Under Assumptions 5.1, 5.2 and 5.5, and* $\frac{k_n}{n} \to 0$, $k_n \to \infty$, *we have (i)*

$$||\widehat{p}(\bullet) - p_o(\bullet)||_{\infty,\omega} = o_p(1); \quad ||\widehat{p}(\bullet) - p_o(\bullet)||_{2,a} \asymp ||\widehat{p}(\bullet) - p_o(\bullet)||_2 = o_p(1);$$

(ii) in addition, if Assumption 6.1 holds, then

$$||\widehat{p}(\bullet) - p_o(\bullet)||_{2,a} \asymp ||\widehat{p}(\bullet) - p_o(\bullet)||_2 = O_p\left(\sqrt{\frac{k_n}{n}} + (k_n)^{-\gamma/d_x}\right).$$

PROOF. **(Proposition B.1):(i)** Recall that $\widehat{p}(x)$ is the sieve LS estimator of $p_o(\cdot) \in \Lambda_c^\gamma(\mathcal{X})$ based on the entire sample. That is,

$$\widehat{p}() = \arg\min_{p()\in\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^n \{D_i - p(X_i)\}^2/2,$$

where $\mathcal{H}_n$ increases with sample size n, and is dense in $\Lambda_c^\gamma(\mathcal{X})$ as $k_n \to \infty$ (by assumption 5.5). Moreover, by Assumptions 5.1 and 5.2 we have the following results: (1) the parameter space is compact under the norm $||\cdot||_{\infty,\omega}$ for $\omega > 0$, see [1];

(2) $E\left[\{D_i - p(X_i)\}^2/2\right]$ is uniquely maximized at $p_o(x) = E[D|X = x] \in \mathcal{H}$; (3) $E\left[\{D_i - p(X_i)\}^2/2\right]$ is continuous in $p()$ under the metric $||\cdot||_{\infty,\omega}$; and (4)

$$\sup_{p()\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^{n} \{D_i - p(X_i)\}^2/2 - E\{D_i - p(X_i)\}^2/2 \right| = o_p(1);$$

where both results (3) and (4) are due to the fact that for any $p()$, $\widetilde{p}() \in \mathcal{H}$,

$$\begin{aligned} &\left|\{D_i - p(X_i)\}^2 - \{D_i - \widetilde{p}(X_i)\}^2\right| = |\{2D_i - [p(X_i) + \widetilde{p}(X_i)]\}\,[p(X_i) - \widetilde{p}(X_i)]| \\ \leq\ & const. \left|[p(X_i) - \widetilde{p}(X_i)](1 + X_i'X_i)^{-\frac{\omega}{2}}\right| \times (1 + X_i'X_i)^{\frac{\omega}{2}}. \end{aligned}$$

Now $E[(1 + X_i'X_i)^{\frac{\omega}{2}}] < \infty$ by assumption 5.2.

Hence by either Theorem 0 in [15] or Lemma 2.9 and Theorem 2.1 in [23], $||\widehat{p}(\bullet) - p_o(\bullet)||_{\infty,\omega} = o_p(1)$. Now

$$\begin{aligned} ||\widehat{p}(\bullet) - p_o(\bullet)||_2 &= \sqrt{\int [\widehat{p}(x) - p_o(x)]^2 f_X(x)dx} \\ &\leq \sqrt{(||\widehat{p}(\bullet) - p_o(\bullet)||_{\infty,\omega})^2 \int (1 + x'x)^{\omega} f_X(x)dx} = o_p(1) \quad \text{(by assumption 5.2).} \end{aligned}$$

(ii) We can obtain the convergence rate of $||\widehat{p}(\bullet) - p_o(\bullet)||_2$ by applying Theorem 1 in [10] or Theorem 2 in [31]. Let $L_n(p()) = \frac{1}{n}\sum_{i=1}^{n} \ell(D_i, X_i, p())$ with $\ell(D_i, X_i, p()) = -\{D_i - p(X_i)\}^2/2$. Since all the assumptions of [10] Theorem 1 are satisfied given our Assumptions 5.1 and 5.2. We obtain

$$||\widehat{p}(\bullet) - p_o(\bullet)||_2 = O_p\left(\max\left\{\sqrt{\frac{k_n}{n}}, ||p_o - \Pi_{2n}p_o||_2\right\}\right).$$

Under Assumption 6.1, for $p_o \in \Lambda_c^{\gamma}(\mathcal{X})$, there exists $\Pi_{\infty n}p_o \in \Lambda_c^{\gamma}(\mathcal{X})$ such that for any fixed $\omega > \gamma$,

$$||p_o - \Pi_{\infty n}p_o||_{\infty,\omega} = \sup_x \left|[p_o(x) - \Pi_{\infty n}p_o(x)](1 + |x|^2)^{-\omega/2}\right| \leq const.(k_n)^{-\gamma/d_x},$$

see [1]. Hence by Assumption 6.1 with $\omega = \gamma + \epsilon$ for a small $\epsilon > 0$,

$$\begin{aligned} ||p_o - \Pi_{2n}p_o||_2 &\leq ||p_o - \Pi_{\infty n}p_o||_2 = \sqrt{\int [p_o(x) - \Pi_{\infty n}p_o(x)]^2 f_X(x)dx} \\ &\leq \sqrt{(||p_o(\cdot) - \Pi_{\infty n}p_o(\cdot)||_{\infty,\omega})^2 \int (1 + x'x)^{\omega} f_X(x)dx} \leq c'(k_n)^{-\gamma/d_x} \end{aligned}$$

Then $||\widehat{p}(\bullet) - p_o(\bullet)||_2 = O_p\left(\sqrt{\frac{k_n}{n}} + (k_n)^{-\gamma/d_x}\right) = o_p(1)$. □

PROOF. (**Theorem 7**): We only provide the proof of the IPW-GMM estimator for moment condition (1), since the one for moment condition (2) is very similar. We establish this theorem by applying Theorem 1 in [9] (hereafter CLK) with their θ being our β and their h being our $p()$. Define

$$\begin{aligned} M_n(\beta, p()) &= \frac{1}{n_a}\sum_{i=1}^{n_a} m(Z_i,\beta)\frac{p(X_i)}{1-p(X_i)}; \\ M(\beta, p()) &= E_a\left[m(Z_i,\beta)\frac{p(X_i)}{1-p(X_i)}\right] = E\left[m(Z,\beta)\frac{p(X)}{1-p(X)}|D=0\right]. \end{aligned}$$

CLK's conditions (1.1) and (1.2) are directly implied by our Assumptions 1.1, 2 and moment condition (1). Note that for any $p() \in \mathcal{H}$, $0 < \frac{1}{1-\underline{p}} \leq \frac{1}{1-p(X)} \leq \frac{1}{1-\overline{p}} < \infty$, we have

$$\begin{aligned} & |M(\beta,p()) - M(\beta,p_o())| = \left|E[m(Z,\beta)\{\frac{p(X)}{1-p(X)} - \frac{p_o(X)}{1-p_o(X)}\}|D=0]\right| \\ \leq\ & \frac{1}{(1-\overline{p})^2} E_a\left[||m(Z,\beta)||(1+|X|^2)^{\frac{\omega}{2}}\right] \times \sup_{x\in\mathcal{X}}\left|[p(x)-p_o(x)](1+|x|^2)^{-\frac{\omega}{2}}\right| \\ \leq\ & \frac{1}{(1-\overline{p})^2}\left\{E_a[\sup_{\beta\in B}||m(Z,\beta)||^2] \times E_a[(1+|X|^2)^{\omega}]\right\}^{1/2} \times ||p()-p_o()||_{\infty,\omega}, \end{aligned}$$

where the last inequality is due to our Assumptions 5.1, 5.2 and 5.4, hence CLK's condition (1.3) is satisfied with respect to the norm $||\cdot||_{\mathcal{H}} = ||\cdot||_{\infty,\omega}$. CLK's condition (1.4) $||\widehat{p}() - p_o()||_{\infty,\omega} = o_p(1)$ is implied by Proposition B.1(i). Note that

$$\begin{aligned} & E_a\left[\sup_{||\beta-\widetilde{\beta}||<\delta, ||p()-\widetilde{p}()||_{\infty,\omega}<\delta}\left|m(Z_i,\beta)\frac{p(X_i)}{1-p(X_i)} - m(Z_i,\widetilde{\beta})\frac{\widetilde{p}(X_i)}{1-\widetilde{p}(X_i)}\right|\right] \\ \leq\ & E_a\left[\sup_{||\beta-\widetilde{\beta}||<\delta}||m(Z_i,\beta)-m(Z_i,\widetilde{\beta})|| \times \sup_{p()\in\mathcal{H}}\left|\frac{p(X_i)}{1-p(X_i)}\right|\right] \\ & + E_a\left[\sup_{\widetilde{\beta}\in B}||m(Z_i,\widetilde{\beta})|| \times \sup_{||p()-\widetilde{p}()||_{\infty,\omega}<\delta}\left|\frac{p(X_i)}{1-p(X_i)} - \frac{\widetilde{p}(X_i)}{1-\widetilde{p}(X_i)}\right|\right] \\ \leq\ & E_a\left[\sup_{||\beta-\widetilde{\beta}||<\delta}||m(Z_i,\beta)-m(Z_i,\widetilde{\beta})||\right] \times \frac{\overline{p}}{1-\overline{p}} \\ & + E_a\left[\sup_{\widetilde{\beta}\in B}||m(Z_i,\widetilde{\beta})||(1+|X_i|^2)^{\omega/2}\right] \frac{\sup_{||p()-\widetilde{p}()||_{\infty,\omega}<\delta}\sup_{x\in\mathcal{X}}\left|[p(x)-\widetilde{p}(x)](1+|x|^2)^{-\frac{\omega}{2}}\right|}{(1-\overline{p})^2} \\ \leq\ & const.b(\delta) + const.\delta, \end{aligned}$$

where the last inequality is due to our Assumptions 5.1 - 5.4 and Proposition B.1(i). Then CLK's condition (1.5) is satisfied, hence $\widehat{\beta} - \beta_0 = o_p(1)$. □

Lemma B.2: *Under assumptions 1, 2, 5 and 6, we have*

$$\sqrt{n}E\left\{\mathcal{E}(X,\beta_o)\frac{\widehat{p}(X)-p_o(X)}{1-p_o(X)}\right\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i-p_o(X_i)}{1-p_o(X_i)}\mathcal{E}(X,\beta_o)+o_p(1).$$

PROOF. (**Lemma B.2**): To establish this result, we follow the approach in [30] and [10]. Recall $p_o(x)=E[D|X=x]\in\Lambda_c^\gamma(\mathcal{X})$ and

$$\widehat{p}()=\arg\min_{p()\in\mathcal{H}_n}\frac{1}{n}\sum_{i=1}^{n}\{D_i-p(X_i)\}^2/2.$$

Define the inner product associated with the space $\mathcal{L}_2(\mathcal{X})$ as

$$\langle h,g\rangle=E\{h(X)g(X)\}\text{ hence }||h()||_2^2=\langle h,h\rangle=E[\{h(X)\}^2].$$

Then the Riesz representor υ^* for functional $E\left\{\mathcal{E}(X,\beta_o)\frac{p(X)-p_o(X)}{1-p_o(X)}\right\}$ is simply given by

$$\upsilon^*(X)=\frac{\mathcal{E}(X,\beta_o)}{1-p_o(X)},$$

this is because

$$||\upsilon^*||^2=\sup_{p()\in\mathcal{H}:p\neq p_o}\frac{\left[E\left\{\mathcal{E}(X,\beta_o)\frac{p(X)-p_o(X)}{1-p_o(X)}\right\}\right]^2}{E\left[(p(X)-p_o(X))^2\right]}=E\left[\left(\frac{\mathcal{E}(X,\beta_o)}{1-p_o(X)}\right)^2\right]$$

and

$$E\left\{\mathcal{E}(X,\beta_o)\frac{p(X)-p_o(X)}{1-p_o(X)}\right\}=\langle\upsilon^*,p()-p_o()\rangle=E\left\{\upsilon^*(X)[p(X)-p_o(X)]\right\}.$$

Let $L_n(p())=\frac{1}{n}\sum_{i=1}^{n}\ell(D_i,X_i,p())$ with $\ell(D_i,X_i,p())=-\{D_i-p(X_i)\}^2/2$. Let $U_i\equiv D_i-p_o(X_i)$. Then by definition $E[U_i|X_i]=0$, and $\ell(D_i,X_i,p())=-\{U_i-[p(X_i)-p_o(X_i)]\}^2/2$. We denote $\mu_n(g)=\frac{1}{n}\sum_{i=1}^{n}[g(D_i,X_i)-E(g(D_i,X_i))]$ as the empirical process indexed by g, and ε_n be any positive sequence with $\varepsilon_n=o(\frac{1}{\sqrt{n}})$. Then by definition,

$$\begin{aligned}0&\leq L_n(\widehat{p})-L_n(\widehat{p}\pm\varepsilon_n\Pi_{2n}\upsilon^*)\\&=\mu_n\left(\ell(D_i,X_i,\widehat{p})-\ell(D_i,X_i,\widehat{p}\pm\varepsilon_n\Pi_{2n}\upsilon^*)\right)+E\left(\ell(D_i,X_i,\widehat{p})-\ell(D_i,X_i,\widehat{p}\pm\varepsilon_n\Pi_{2n}\upsilon^*)\right).\end{aligned}$$

A simple calculation yields

$$\begin{aligned}&E\left(\ell(D_i,X_i,\widehat{p})-\ell(D_i,X_i,\widehat{p}\pm\varepsilon_n\Pi_{2n}\upsilon^*)\right)\\=\;&\pm\varepsilon_nE[\Pi_{2n}\upsilon^*(X_i)\{\widehat{p}(X_i)-p_o(X_i)\}]+\frac{1}{2}\varepsilon_n^2E[\{\Pi_{2n}\upsilon^*(X_i)\}^2]\end{aligned}$$

$$\begin{aligned} & \mu_n\left(\ell(D_i, X_i, \widehat{p}) - \ell(D_i, X_i, \widehat{p} \pm \varepsilon_n \Pi_{2n} \upsilon^*)\right) \\ = \; & \mp \varepsilon_n \times \mu_n\left(\Pi_{2n}\upsilon^* U_i\right) \pm \varepsilon_n \times \mu_n\left(\Pi_{2n}\upsilon^* \frac{2\{\widehat{p}() - p_o()\} \pm \varepsilon_n \Pi_{2n}\upsilon^*}{2}\right) \end{aligned}$$

hence

$$\begin{aligned} 0 \; \leq \; & \mp \mu_n\left(\Pi_{2n}\upsilon^*(X_i)U_i\right) \pm E[\Pi_{2n}\upsilon^*(X_i)\{\widehat{p}(X_i) - p_o(X_i)\}] \\ & \pm \mu_n\left(\Pi_{2n}\upsilon^*(X_i)\{\widehat{p}(X_i) - p_o(X_i)\}\right) + \frac{\varepsilon_n}{2n}\sum_{i=1}^n \{\Pi_{2n}\upsilon^*(X_i)\}^2 \\ = \; & \mp \mu_n\left([\Pi_{2n}\upsilon^* - \upsilon^*]U_i\right) \pm \mu_n\left(\upsilon^* U_i\right) \pm E[[\Pi_{2n}\upsilon^* - \upsilon^*]\{\widehat{p} - p_o\}] \mp E[\upsilon^*\{\widehat{p} - p_o\}] \\ & \pm \mu_n\left(\Pi_{2n}\upsilon^*(X_i)\{\widehat{p}(X_i) - p_o(X_i)\}\right) + \frac{\varepsilon_n}{2n}\sum_{i=1}^n \{\Pi_{2n}\upsilon^*(X_i)\}^2 \end{aligned}$$

In the following we shall establish **(B2.1)-(B2.4)**:

$$\textbf{(B2.1)} \quad \mu_n\left([\Pi_{2n}\upsilon^*(X_i) - \upsilon^*(X_i)]U_i\right) = o_p(\frac{1}{\sqrt{n}})$$

$$\textbf{(B2.2)} \quad E\left([\Pi_{2n}\upsilon^*(X_i) - \upsilon^*(X_i)]\{\widehat{p}(X_i) - p_o(X_i)\}\right) = o_p(\frac{1}{\sqrt{n}})$$

$$\textbf{(B2.3)} \quad \mu_n\left(\Pi_{2n}\upsilon^*(X_i)\{\widehat{p}(X_i) - p_o(X_i)\}\right) = o_p(\frac{1}{\sqrt{n}})$$

$$\textbf{(B2.4)} \quad \frac{1}{n}\sum_{i=1}^n \{\Pi_{2n}\upsilon^*(X_i)\}^2 = O_p(1)$$

Note that **(B2.1)** is implied by Chebychev inequality, i.i.d. data, and $||\Pi_n\upsilon^* - \upsilon^*||_2 = o(1)$ which is satisfied given the expression for υ^* and Assumptions 5.1 and 6.5. **(B2.2)** is implied by Assumption 6.5 and $||\widehat{p}() - p_o()||_2 = O_p\left(n^{-\frac{\gamma}{2\gamma+d_x}}\right)$ from Proposition B.1(ii). **(B2.4)** is implied by Markov inequality, i.i.d. data, and Assumptions 5.1 and 6.5. Finally for **(B2.3)**, let $\mathcal{F}_n = \{\Pi_{2n}\upsilon^*(\cdot)h(\cdot) : h(\cdot) \in \Lambda_c^\gamma(\mathcal{X})\}$, then by Assumption 6.1, $\log N_{[]}(\delta, \mathcal{F}_n, ||\cdot||_2) \leq const.(\frac{c}{\delta})^{d_x/\gamma}$ for any $\delta > 0$. Applying Theorem 3 in [10] with their $\delta_n = n^{-\gamma/(2\gamma+d_x)}$, we have

$$\sup_{h\in\mathcal{F}_n : ||h()-p_o()||_2 \leq \delta_n} \left|\sqrt{n}\mu_n\left(\Pi_{2n}\upsilon^*\{h() - p_o()\}\right)\right| = O_p\left(n^{-\frac{2\gamma-d_x}{2(2\gamma+d_x)}}\right) = o_p(1).$$

Hence we obtain **(B2.3)**. Now (B2.1)-(B2.4) imply $0 \leq \pm\mu_n\left(\upsilon^* U_i\right) \mp E[\upsilon^*\{\widehat{p} - p_o\}] + o_p(\frac{1}{\sqrt{n}})$, that is $\sqrt{n}E[\upsilon^*(X)\{\widehat{p}(X) - p_o(X)\}] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \upsilon^*(X_i)U_i + o_p(1)$, hence the result follows. □

PROOF. (**Theorem 8**): Again we only provide the proof of the IPW-GMM estimator for moment condition (1). We establish this theorem by applying Theorem 2 in CLK (2003). Given the definition of β_0 and Theorem 7, CLK's Condition (2.1) is directly

satisfied. Note that their $\Gamma_1(\beta, p_o) = \frac{p}{1-p} J^1_\beta$, hence their Condition (2.2) is satisfied with our Assumption 1.1.

Following the proof of CLK's Theorem 2, we note that the conclusion of CLK's Theorem 2 remains true when CLK's Conditions (2.3)(i) and (2.4) are replaced by the following one:

$$\textbf{(*)} \qquad \sup_{\beta \in B: ||\beta - \beta_0|| \leq \delta_0} \|M(\beta, \widehat{p}()) - M(\beta, p_o()) - \Gamma_2(\beta, p_o)[\widehat{p}() - p_o()]\| = o_p(n^{-1/2}),$$

where

$$\begin{aligned}
\Gamma_2(\beta, p_o)[p() - p_o()] &= E\left\{ m(Z, \beta) \frac{p(X) - p_o(X)}{(1 - p_o(X))^2} \middle| D = 0 \right\} \\
&= E_a\left\{ \mathcal{E}(X, \beta) \frac{p(X) - p_o(X)}{(1 - p_o(X))^2} \right\} \\
&= E\left\{ \mathcal{E}(X, \beta) \frac{p(X) - p_o(X)}{(1 - p_o(X))^2} \frac{f_{X|D=0}(X)}{f_X(X)} \right\} \\
&= \frac{1}{1-p} E\left\{ \mathcal{E}(X, \beta) \frac{p(X) - p_o(X)}{1 - p_o(X)} \right\},
\end{aligned}$$

and the last equality is due to $f_{X|D=0}(X)/f_X(X) = (1 - p_o(X))/(1 - p)$.

Before we apply Assumptions 6.6a or 6.6b or 6.6c to verify Condition **(*)**, let us check CLK's Conditions (2.3)(ii), (2.5) and (2.6). Since for all β with $||\beta - \beta_0|| \leq \delta_0$ and all $p()$ with $||p() - p_o()||_{\infty,\omega} \leq \delta_0$, we have

$$\begin{aligned}
& |\Gamma_2(\beta, p_o)[p() - p_o()] - \Gamma_2(\beta_o, p_o)[p() - p_o()]| \\
= & \left| \frac{1}{1-p} E\left\{ [\mathcal{E}(X, \beta) - \mathcal{E}(X, \beta_0)] \frac{p(X) - p_o(X)}{1 - p_o(X)} \right\} \right| = \left| \frac{\beta - \beta_0}{1-p} E\left\{ \frac{\partial \mathcal{E}(X, \overline{\beta})}{\partial \beta} \frac{p(X) - p_o(X)}{1 - p_o(X)} \right\} \right| \\
\leq & \frac{||\beta - \beta_0||}{(1-p)(1-\overline{p})} E\left[\left\| \frac{\partial \mathcal{E}(X, \overline{\beta})}{\partial \beta} \right\| (1 + |X|^2)^{\frac{\omega}{2}} \right] \times \sup_{x \in \mathcal{X}} \left| [p(x) - p_o(x)](1 + |x|^2)^{-\frac{\omega}{2}} \right|,
\end{aligned}$$

where $\overline{\beta}$ is in between β and β_0. Thus, under our Assumptions 5.2, 6.4, Proposition B.1(i) and Theorem 7, $|\Gamma_2(\beta, p_o)[p() - p_o()] - \Gamma_2(\beta_o, p_o)[p() - p_o()]| \leq const. ||\beta - \beta_0|| \times ||p() - p_o()||_{\infty,\omega}$ hence CLK's Condition (2.3)(ii) is satisfied.

Now we verify CKL's Condition (2.5) by applying their Theorem 3. In fact, given our Theorem 7 and Proposition B.1(i), it suffices to consider some neighborhood around (β_o, p_o). Let $\delta_0 > 0$ be a small value, then for all $(\widetilde{\beta}, \widetilde{p}) \in B \times \mathcal{H}$ with $||\widetilde{\beta} - \beta_o|| \leq \delta_0$ and

$||\widetilde{p} - p_o||_{\infty,\omega} \leq \delta_0$, we have for any $\delta \in (0, \delta_0]$,

$$
\begin{aligned}
& E_a\left[\sup_{||\beta-\widetilde{\beta}||<\delta, ||p()-\widetilde{p}()||_{\infty,\omega}<\delta}\left|m(Z_i,\beta)\frac{p(X_i)}{1-p(X_i)} - m(Z_i,\widetilde{\beta})\frac{\widetilde{p}(X_i)}{1-\widetilde{p}(X_i)}\right|^2\right] \\
\leq\ & E_a\left[\sup_{||\beta-\widetilde{\beta}||<\delta}||m(Z_i,\beta)-m(Z_i,\widetilde{\beta})||^2 \times \sup_h\left|\frac{p(X_i)}{1-p(X_i)}\right|^2\right] \\
& + E_a\left[\sup_{\widetilde{\beta}\in B: ||\widetilde{\beta}-\beta_o||\leq\delta_0}||m(Z_i,\widetilde{\beta})||^2 \times \sup_{||p()-\widetilde{p}()||_{\infty,\omega}<\delta}\left|\frac{p(X_i)}{1-p(X_i)} - \frac{\widetilde{p}(X_i)}{1-\widetilde{p}(X_i)}\right|^2\right] \\
\leq\ & E_a\left[\sup_{||\beta-\widetilde{\beta}||<\delta}||m(Z_i,\beta)-m(Y_i,X_i,\widetilde{\beta})||^2\right] \times \left(\frac{\overline{p}}{1-\overline{p}}\right)^2 \\
& + E_a\left[\sup_{\widetilde{\beta}\in B||\widetilde{\beta}-\beta_o||\leq\delta_0}||m(Z_i,\widetilde{\beta})||^2(1+|X_i|^2)^{\omega}\right] \\
\times\ & \sup_{||p()-\widetilde{p}()||_{\infty,\omega}<\delta}\sup_{x\in\mathcal{X}}\left|[p(x)-\widetilde{p}(x)](1+|x|^2)^{-\frac{\omega}{2}}\right|^2\frac{1}{(1-\overline{p})^2} \\
\leq\ & const.\delta^{2\epsilon} + const.\delta^2 \quad \text{for some } \epsilon \in (0,1],
\end{aligned}
$$

where the last inequality is due to our Assumptions 6.2, 6.3 and Proposition B.1(i). In the following we let $N(\varepsilon, \Lambda_c^\gamma(\mathcal{X}), ||\cdot||_{\infty,\omega})$ denote the $||\cdot||_{\infty,\omega}-$covering number of $\Lambda_c^\gamma(\mathcal{X})$ [i.e., the minimal number of N for which there exist $\varepsilon-$balls $\{h : ||h-u_j||_{\infty,\omega} \leq \varepsilon\}$, $j = 1, \ldots, N$ to cover $\Lambda_c^\gamma(\mathcal{X})$]. Then our assumption 6.1 implies

$$
\log N(\delta, \Lambda_c^\gamma(\mathcal{X}), ||\cdot||_{\infty,\omega}) \leq const.\left(\frac{c}{\delta}\right)^{d_x/\gamma}, \quad \int_0^1 \sqrt{\log N(\delta, \Lambda_c^\gamma(\mathcal{X}), ||\cdot||_{\infty,\omega})}d\delta < \infty.
$$

Thus by applying CLK's Theorem 3, CLK's Condition (2.5) is satisfied.

It remains to verify CLK's Condition (2.6). First we note

$$
\begin{aligned}
\sqrt{n_a}M_n(\beta_o, p_o) &= \frac{1}{\sqrt{n_a}}\sum_{i=1}^{n_a} m(Z_i,\beta_o)\frac{p_o(X_i)}{1-p_o(X_i)} = \frac{1}{\sqrt{n_a}}\sum_{i=1}^{n}(1-D_i)m(Z_i,\beta_o)\frac{p_o(X_i)}{1-p_o(X_i)} \\
&= \sqrt{\frac{n}{n_a}} \times \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-D_i)m(Z_i,\beta_o)\frac{p_o(X_i)}{1-p_o(X_i)},
\end{aligned}
$$

Next we notice

$$
\begin{aligned}
\sqrt{n_a}\Gamma_2(\beta_o, p_o)[p() - p_o()] &= \sqrt{n_a}E_a\left\{\mathcal{E}(X,\beta_o)\frac{p(X)-p_o(X)}{(1-p_o(X))^2}\right\} \\
&= \sqrt{\frac{n_a}{n}}\frac{1}{1-p} \times \sqrt{n}E\left\{\mathcal{E}(X,\beta_o)\frac{p(X)-p_o(X)}{1-p_o(X)}\right\}.
\end{aligned}
$$

By Lemma B.2 and $n_a/n = 1-p+o_p(1)$, we obtain

$$\begin{aligned}
&\sqrt{n_a}\{M_n(\beta_o,p_o)+\Gamma_2(\beta_o,p_o)[\widehat{p}()-p_o()]\} \\
= &\sqrt{\frac{1}{1-p}}\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{(1-D_i)m(Z_i,\beta_o)\frac{p_o(X_i)}{1-p_o(X_i)}+\frac{D_i-p_o(X_i)}{1-p_o(X_i)}\mathcal{E}(X,\beta_o)\right\}+o_p(1),
\end{aligned}$$

thus CLK's Condition (2.6) is satisfied. Moreover from the proof of CLK's Theorem 2 we obtain:

$$\begin{aligned}
\sqrt{n_a}(\widehat{\beta}-\beta_o) &= -(\Gamma_1'W\Gamma_1)^{-1}\Gamma_1'W\sqrt{n_a}\{M_n(\beta_o,p_o)+\Gamma_2(\beta_o,p_o)[\widehat{p}()-p_o()]\}+o_p(1) \\
&= -\frac{1-p}{p}(J_\beta^{1\prime}WJ_\beta^1)^{-1}J_\beta^{1\prime}W\sqrt{n_a}\{M_n(\beta_o,p_o)+\Gamma_2(\beta_o,p_o)[\widehat{p}()-p_o()]\}+o_p(1).
\end{aligned}$$

Since $\frac{n}{n_a}=\frac{1}{1-p}+o_p(1)$,

$$\begin{aligned}
\sqrt{n}(\widehat{\beta}-\beta_o) &= -\frac{1-p}{p}(J_\beta^{1\prime}WJ_\beta^1)^{-1}J_\beta^{1\prime}W\sqrt{n}\{M_n(\beta_o,p_o)+\Gamma_2(\beta_o,p_o)[\widehat{p}()-p_o()]\}+o_p(1) \\
&= -(J_\beta^{1\prime}WJ_\beta^1)^{-1}J_\beta^{1\prime}W\frac{1}{p}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{m(Z_i,\beta_o)\frac{[1-D_i]p_o(X_i)}{1-p_o(X_i)}+\frac{D_i-p_o(X_i)}{1-p_o(X_i)}\mathcal{E}(X,\beta_o)\right\}+o_p(1).
\end{aligned}$$

thus we obtain Theorem 8 after we establish condition **(*)**.

We now apply Assumption 6.6a or 6.6b or 6.6c to verify condition **(*)**. Since

$$\begin{aligned}
&M(\beta,p())-M(\beta,p_o())-\Gamma_2(\beta,p_o)[p()-p_o()] \\
= &E_a\left\{m(Z,\beta)\left[\frac{p(X)}{1-p(X)}-\frac{p_o(X)}{1-p_o(X)}-\frac{p(X)-p_o(X)}{(1-p_o(X))^2}\right]\right\} \\
= &E_a\left\{\frac{m(Z,\beta)[p(X)-p_o(X)]}{1-p_o(X)}\left[\frac{1}{1-p(X)}-\frac{1}{1-p_o(X)}\right]\right\} \\
= &E_a\left\{\frac{\mathcal{E}(X,\beta)[p(X)-p_o(X)]^2}{(1-p(X))(1-p_o(X))^2}\right\},
\end{aligned}$$

we have under Assumption 5.1,

$$\begin{aligned}
&\sup_{\beta\in B:||\beta-\beta_0||\le\delta_0}||M(\beta,p())-M(\beta,p_o())-\Gamma_2(\beta,p_o)[p()-p_o()]|| \\
= &\sup_{\beta\in B:||\beta-\beta_0||\le\delta_0}\left\|E_a\left\{\frac{\mathcal{E}(X,\beta)[p(X)-p_o(X)]^2}{(1-p(X))(1-p_o(X))^2}\right\}\right\| \\
\le &\frac{1}{(1-\overline{p})^3}E_a\left\{\sup_{\beta\in B:||\beta-\beta_0||\le\delta_0}||\mathcal{E}(X,\beta)||\times[p(X)-p_o(X)]^2\right\}.
\end{aligned}$$

If Assumption 6.6a holds, then

$$\begin{aligned}
&E_a\left\{\sup_{\beta\in B:||\beta-\beta_0||\le\delta_0}||\mathcal{E}(X,\beta)||\times[p(X)-p_o(X)]^2\right\} \\
\le &\sup_{\beta\in B:||\beta-\beta_0||\le\delta_0}\sup_x||\mathcal{E}(x,\beta)||\times E_a\left\{[p(X)-p_o(X)]^2\right\}\le const.[||p()-p_o()||_{2,a}]^2.
\end{aligned}$$

Now Proposition B.1(ii), $k_n = O\left(n^{\frac{d_x}{2\gamma+d_x}}\right)$ and $\gamma > d_x/2$ imply $[||\widehat{p}() - p_o()||_{2,a}]^2 = o_p(n^{-1/2})$, hence condition **(*)** is satisfied.

If Assumption 6.6b holds, then

$$\begin{aligned} & E_a\left\{\sup_{\beta\in B:||\beta-\beta_0||\leq\delta_0}||\mathcal{E}(X,\beta)||\times[p(X)-p_o(X)]^2\right\} \\ \leq & \left(E_a\left[\sup_{\beta\in B:||\beta-\beta_0||\leq\delta_0}||\mathcal{E}(X,\beta)||^4\right]\right)^{1/4}\left(E_a\left\{[p(X)-p_o(X)]^4\right\}\right)^{1/4}\sqrt{E_a\left\{[p(X)-p_o(X)]^2\right\}} \\ \leq & const.\times[||p()-p_o()||_{2,a}]^{2-\frac{d_x}{4\gamma}} \quad \text{for all } ||p()-p_o()||_{2,a}=o(1), \end{aligned}$$

where the last inequality is due to the following inequalities for any $s \in [\frac{d_x}{4}, \gamma)$:

$$\begin{aligned} \left(E_a\left\{[p(X)-p_o(X)]^4\right\}\right)^{1/4} &\leq const.\left(||p()-p_o()||_{2,a}+||\nabla^s\{p()-p_o()\}||_{2,a}\right), \\ ||\nabla^s\{p()-p_o()\}||_{2,a} &\leq const.[||p()-p_o()||_{2,a}]^{1-\frac{s}{\gamma}}. \end{aligned}$$

Now Proposition B.1(ii), $k_n = O\left(n^{\frac{d_x}{2\gamma+d_x}}\right)$ and $\gamma > 3d_x/4$ imply $[||\widehat{p}()-p_o()||_{2,a}]^{2-\frac{d_x}{4\gamma}} = o_p(n^{-1/2})$, hence condition **(*)** is satisfied.

If Assumption 6.6c holds, then

$$\begin{aligned} & E_a\left\{\sup_{\beta\in B:||\beta-\beta_0||\leq\delta_0}||\mathcal{E}(X,\beta)||\times[p(X)-p_o(X)]^2\right\} \\ \leq & \sqrt{E_a\left[\sup_{\beta\in B:||\beta-\beta_0||\leq\delta_0}||\mathcal{E}(X,\beta)||^2\right]}\times\sqrt{E_a\left\{[p(X)-p_o(X)]^4\right\}} \\ \leq & const.\times[||p()-p_o()||_{2,a}]^{2(1-\frac{d_x}{4\gamma})} \quad \text{for all } ||p()-p_o()||_{2,a}=o(1). \end{aligned}$$

Now Proposition B.1(ii), $k_n = O\left(n^{\frac{d_x}{2\gamma+d_x}}\right)$ and $\gamma > d_x$ imply $[||\widehat{p}()-p_o()||_{2,a}]^{2(1-\frac{d_x}{4\gamma})} = o_p(n^{-1/2})$, hence condition **(*)** is satisfied. □

Acknowledgements. We thank J. Hahn, J. Ham, G. Imbens, O. Linton, W. Newey, B. Salanié, the associate editor and three referees, as well as seminar participants at various universities for insightful comments and suggestions. We are solely responsible for any remaining errors and omissions.

X. Chen
Dept of Economics
New York University
269 Mercer Street
New York, NY 10003
e-mail: xiaohong.chen@nyu.edu

H. Hong and A. Tarozzi
Dept of Economics
Duke University
Social Sciences Building, PO Box 90097
Durham, NC 27708
E-mail: {hanhong,taroz}@duke.edu

TABLE 1
Cumulative Distribution Functions (×100) of total (log) household expenditure

y	(1) Unadjusted (primary)	(2) Adjusted NP CEP	(3) Adjusted NP IPW	(4) Adjusted Par. IPW	(5) Adjusted Eff. CEP
6	2.92 (.067)	3.388 (.0695)	3.387 (.0694)	3.15 (.0594)	3.23 (.0598)
6.25	5.67 (.092)	6.521 (.0948)	6.522 (.0948)	6.31 (.0846)	6.38 (.0845)
6.50	11.06 (.125)	12.272 (.1237)	12.273 (.1234)	12.21 (.1165)	12.21 (.1149)
6.75	20.28 (.161)	21.679 (.1588)	21.674 (.1587)	21.89 (.1645)	21.76 (.1575)
7	34.06 (.189)	35.052 (.1763)	35.041 (.1772)	35.53 (.1794)	35.28 (.1738)
7.25	50.75 (.200)	50.600 (.1967)	50.592 (.1975)	51.19 (.1948)	50.88 (.1920)
7.50	66.98 (.188)	65.682 (.1925)	65.687 (.1929)	66.15 (.1973)	65.91 (.1880)

Source: Authors' calculations from Indian National Sample Survey, rounds 50 (1993-94, $n = 58,846$) and 55 (1999-2000, $n = 62,679$), rural sector only from the major Indian states, which account for more than 95% of the total population. Column (1) - Calculated from the unadjusted primary sample. Column (2) - CEP-GMM cubic sieve Estimator, with 10 knots, using "comparable items" as predictor. Column (3) - IPW-GMM. Flexible logit with cubic sieve, with 10 knots, using "comparable items" as predictor. Column (4) - Parametric IPW Estimator. The propensity score is estimated using logit and including total expenditure in "comparable items" as sole predictor. Column (5) - Semiparametric estimator efficient for the case of correctly specified propensity score.