DEPARTMENT OF ECONOMICS
YALE UNIVERSITY

P.O. Box 208268
New Haven, CT 06520-8268

http://www.econ.yale.edu/

Economics Department Working Paper No. 42

Cowles Foundation Discussion Paper No. 1644

Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects

Xiaohong Chen, Han Hong, and Alessandro Tarozzi

March 2008

Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects*

Xiaohong Chen†, Han Hong‡, and Alessandro Tarozzi§

First version: September 2004, This version: November 2004

Abstract

We study semiparametric efficiency bounds and efficient estimation of parameters defined through general nonlinear, possibly non-smooth and over-identified moment restrictions, where the sampling information consists of a primary sample and an auxiliary sample. The variables of interest in the moment conditions are not directly observable in the primary data set, but the primary data set contains proxy variables which are correlated with the variables of interest. The auxiliary data set contains information about the conditional distribution of the variables of interest given the proxy variables. Identification is achieved by the assumption that this conditional distribution is the same in both the primary and auxiliary data sets. We provide semiparametric efficiency bounds for both the "verify-out-of-sample" case, where the two samples are independent, and the "verify-in-sample" case, where the auxiliary sample is a subset of the primary sample; and the bounds are derived when the propensity score is unknown, or known, or belongs to a correctly specified parametric family. These efficiency variance bounds indicate that the propensity score is ancillary for the "verify-in-sample" case, but is not ancillary for the "verify-out-of-sample" case. We show that sieve conditional expectation projection based GMM estimators achieve the semiparametric efficiency bounds for all the above mentioned cases, and establish their asymptotic efficiency under mild regularity conditions. Although inverse probability weighting based GMM estimators are also shown to be semiparametrically efficient, they need stronger regularity conditions and clever combinations of nonparametric and parametric estimates of the propensity score to achieve the efficiency bounds for various cases. Our results contribute to the literature on non-classical measurement error models, missing data and treatment effects.

JEL: C1, C3

Key words: Auxiliary data, Measurement Error, Missing Data, Treatment Effect, Semiparametric Efficiency Bound, GMM, Sieve Estimation.

*We thank John Ham, Guido Imbens, Oliver Linton, Whitney Newey, Bernard Salanié and seminar participants in Columbia University and New York University for insightful comments. We are especially grateful to Whitney Newey for pointing out relevant papers in the statistics literature. We are responsible for any remaining errors and omissions.

†Dept of Economics, New York University, 269 Mercer Street, New York, NY 10003. xiaohong.chen@nyu.edu.

‡Dept of Economics, Duke University. Social Sciences Building, PO Box 90097, Durham, NC 27708, hanhong@econ.duke.edu.

§Dept of Economics, Duke University. Social Sciences Building, PO Box 90097, Durham, NC 27708, taroz@econ.duke.edu.

1 Introduction

Many empirical studies in economics are complicated by the presence of relevant variables that are not observed, either because they are unobservable by their own nature, or because they are only available in an incomplete or corrupted way. A chief example of the former case arises in the program evaluation literature, where the estimation of treatment effects has to overcome the fact that one never observes individual outcomes with *and* without treatment. Important examples of the latter case include attrition in panel data analysis and the ubiquitous presence of measurement error which can potentially be correlated with the true unobserved variables. In such circumstances, identifying assumptions become necessary to overcome the lack of identification that results from the missing information in what we will refer to as the *primary* data set.

One solution to this identification problem is based on the assumption that the missing information can be recovered using *auxiliary* data sources under a conditional independence assumption. The key element of the identification strategy is that the auxiliary data set must provide information about the conditional distribution of the true variables of interest given a set of proxy variables, where the proxy variables are observed in both the primary sample and the auxiliary sample. In other words, conditional on the proxy variables, the distributions of the variables of interest are assumed to be independent of whether they belong to the primary sample or the auxiliary sample.

In this paper, we study semiparametric efficiency bounds and efficient estimation of parameters defined through general nonlinear, possibly non-smooth and over-identified moment conditions under this conditional independence assumption. We provide semiparametric efficiency bounds for the cases when the propensity score is unknown, or known, or belongs to a correctly specified parametric family. We define the propensity score as the probability that one observation belongs to the subsample where only the proxy variables are observed. Moreover, these efficiency bounds are applicable to both the "verify-out-of-sample" case, where the auxiliary sample and the primary sample are independent, and the "verify-in-sample" case, where the auxiliary sample is a subset of the primary sample. These efficiency variance bounds indicate that the propensity score is ancillary for the "verify-in-sample" case but is not ancillary for the "verify-out-of-sample" case. That is, more information on propensity score will not affect the asymptotic efficiency variance bounds for parameters defined in the "verify-in-sample" case, but will improve the asymptotic efficiency for parameters defined in the "verify-out-of-sample" case.

Semiparametric efficiency bounds are important for understanding the limits of semiparametric estimation methods, as they represent the equivalent of the Cramer-Rao lower bound in semiparametric models. Standard references for this literature are given by Newey (1990b) and Bickel, Klaassen, Ritov, and Wellner (1993) among others. Hahn (1998) pioneered the semiparametric efficiency bound calculation in mean treatment effect analysis, under the conditional independence assumption of the latent outcomes conditional on observables covariates. Within the program evaluation context, Hahn (1998), Heckman, Ichimura, and Todd (1998) and Hirano, Imbens, and Ridder (2003) study the semiparametric efficiency of different estimators of both the average treatment effect and the average treatment effect for the treated under the assumption of unknown or known propensity score (here interpreted as the probability of treatment conditional on the covariates). However, they do not consider the efficiency bounds when the propensity score takes a correctly specified parametric form. Several papers study the role of the propensity score in

semiparametric estimation of missing data and measurement error models. In the missing data literature and when the propensity score is assumed to be unknown, or known or parametric, Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995) and Rotnitzky and Robins (1995) have presented a unified framework for calculating the semiparametric efficiency bounds for nonlinear regression models in the "verify-in-sample" case. However they do not study the efficiency bounds for the "verify-out-of-sample" case.

Our efficiency bound calculations build on the insights of the existing results in program evaluation and missing data literature and generalize them in several ways. First, we study semiparametric efficiency variance bounds for parameters implicitly defined through general nonlinear, possibly non-smooth and over-identified moment restrictions for non-classical measurement error models, missing data and program evaluation. Second, we calculate efficiency bounds under the assumptions of unknown, or known or parametric propensity score, for both the "verify-in-sample" case and the "verify-out-of-sample" case.[1] Our new results for missing data and non-classical measurement error models under the"verify-out-of-sample" case are interesting since they imply that more information on the propensity score leads to smaller efficiency variance bounds.

We also develop sieve conditional expectation projection based GMM (hereafter CEP-GMM) estimators that achieve the semiparametric efficiency bounds for parameters defined through general moment restrictions when the propensity score is unknown, or known or belongs to a correctly specified parametric family. A sieve inverse probability weighting based GMM (hereafter IPW-GMM) estimator is also shown to achieve the semiparametric efficiency bounds. Each estimator relies only on one nonparametric estimate; the CEP-GMM estimator only requires the nonparametric estimation of a conditional expectation, while the IPW-GMM estimator only needs a nonparametric estimate of the propensity score. The asymptotic normality and efficiency properties of both estimators are established under weaker regularity conditions than the existing ones in the literature. In particular, we allow for non-smooth moment conditions, and for unbounded support of conditioning (or proxy) variables, which is very important for measurement error applications. Moreover, the root-n asymptotic normality and efficiency of the CEP-GMM estimators are obtained without the strong assumption that the unknown propensity score is uniformly bounded away from zero and one. Also, the CEP-GMM estimators are characterized by a simple common format that achieves the relevant efficiency bound for all the cases we consider, and the contributions to the variance deriving from each of the two stages of the estimation procedure are orthogonal to each other. Instead, the parametric inverse probability weighting based GMM estimator will be generally inefficient when the propensity score is known or belongs to a correctly specified parametric family; clever combinations of nonparametric and parametric estimates of propensity score are needed to achieve the semiparametric efficiency bounds for these cases.

The CEP-GMM estimator was proposed by Chen, Hong, and Tamer (2003) in the context of non-classical measurement error models, but it was previously unknown whether the optimal weighted version can reach the semiparametric efficiency bounds for measurement error models under either the "verify-in-sample" case or the "verify-out-of-sample" case. In the causal inference and program evaluation context, Robins, Mark, and Newey (1992) first developed the conditional expectation projection estimator for the average treatment effect parameter, and also discussed its semiparametric efficiency properties. The IPW-

[1] See Section 3 and Appendix C for further discussion.

GMM estimator extends the propensity score based estimators of Hahn (1998), Heckman, Ichimura, and Todd (1998) and Hirano, Imbens, and Ridder (2003) for the mean treatment effect parameter to any parameters defined by general nonlinear, possibly non-smooth and over-identified moment restrictions. When the propensity score is known, Hirano, Imbens, and Ridder (2003) used a nonparametric estimate of the propensity score and the known propensity score to obtain semiparametrically efficient estimation of the mean treatment effect for the treated parameter. For missing data models that correspond to the "verify-in-sample" case discussed below, Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995) and Robins, Rotnitzky, and Zhao (1995) have proposed various semiparametric estimators using parametric estimation of the propensity score. Wooldridge (2002) and Wooldridge (2003) showed that estimating propensity score parametrically is asymptotically more efficient than using the known propensity score in M-estimation of parameters for missing data models. Yet the issue of how to use a correctly specified parametric propensity score to obtain asymptotically efficient estimators for parameters defined by a general moment restriction under the "verify-out-of-sample" case has not been addressed before. We develop optimally weighted sieve CEP-GMM and IPW-GMM estimators inspired by the existing literature. They are shown to be semiparametrically efficient for parameters defined through general moment restrictions for both the "verify-in-sample" and the "verify-out-of-sample" cases under the assumption of unknown, or known or parametric propensity score.

In section 2 we describe the model, discuss the literature and give motivating examples. Section 3 calculates the semiparametric efficiency bounds. Section 4 shows that the optimally weighted CEP-GMM estimators achieve the semiparametric efficiency bounds when the propensity score is unknown, or known, or belongs to a parametric family. Section 5 shows that the optimally weighted IPW-GMM estimators achieve the semiparametric efficiency bounds when the propensity score is unknown, and cleverly modified versions can also achieve the efficiency bounds when the propensity score is known or correctly parameterized. In Section 6 we illustrate empirically the performance of the different estimators in the estimation of earnings quantiles and cumulative distribution functions, when non-classical measurement errors are present. Section 7 concludes. All proofs are given in the appendixes. Appendix C presents some new results for nonlinear regression models.

2 The Model and the Literature

We assume the researchers have access to two data sets: the **primary** data set is a random sample from the population of interest, while an **auxiliary** sample will serve the purpose of ensuring the identification of parameters that would not be identified by the primary data set alone. The researchers observe whether each observation belongs to the primary or the auxiliary dataset. Let D denote a binary variable equal to zero for observations belonging to the auxiliary sample, and equal to one otherwise.

We distinguish two cases. The first case is called the "verify-out-of-sample" case, in which case the primary data and the auxiliary data are two different and independent data sets. In this case $D = 1$ denotes the fact that an observation belongs to the primary data. The second case is called "verify-in-sample" case, where the auxiliary data set is a subset of the primary data set. In this case the primary data set includes both $D = 1$ and $D = 0$ observations. These terminologies are motivated from non-classical measurement error models described below. They are related to the parameters of "mean untreated outcome of the

treated" and the "average untreated outcome of the population" in the program evaluation literature.

We are interested in the estimation of parameters $\beta \in R^{d_\beta}$ defined implicitly in terms of general nonlinear moment conditions:

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

or

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

where $Z=(Y,X)$ and $m\left(\cdot;\beta\right)$ is a set of moment conditions with dimension $d_m \geq d_\beta$. (1) is the "verify-out-of-sample" case, while (2) is the "verify-in-sample" case. These conditions make clear that the moment conditions are assumed to hold in the *primary* sample. The main challenge is that *at least some* of the variables in Y are *not* observed when $D=1$, which is the primary sample in case (1) and a subset of the primary sample in case (2). Identification is ensured by the availability of an auxiliary data set ($D=0$) where both Y, the variables of interest, and X, a set of proxy variables that are also potentially of interest, are available. The following assumption of conditional independence is important:

Assumption 1 $Y \perp D \mid X$.

If assumption 1 holds, identification follows by noting that, under case (1)

$$\begin{aligned} E\left[m\left(Z;\beta\right) \mid D=1\right] &= E\left[E\left[m\left(Z;\beta\right) \mid X, D=1\right] \mid D=1\right] \\ &= \int E\left[m\left(Z;\beta\right) \mid x, D=0\right] f\left(x \mid D=1\right) dx, \end{aligned}$$

where the second equality follows immediately from assumption 1. Under case (2),

$$\begin{aligned} E\left[m\left(Z;\beta\right)\right] &= E\left[E\left[m\left(Z;\beta\right) \mid X\right]\right] \\ &= \int E\left[m\left(Z;\beta\right) \mid x, D=0\right] f\left(x\right) dx. \end{aligned}$$

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

These identifying assumptions have been extensively used in the literature. Examples include the following.

Example 1: Recent works on nonlinear but classical measurement error models include Hausman, Ichimura, Newey, and Powell (1991), Hausman, Newey, and Powell (1995), Newey (2001), Hsiao and Wang (1995), Li (2002) and Schennach (2004). On the other hand, the presence of non-classical measurement errors in economic data is well documented by Bound and Krueger (1991a), Bound, Brown, Duncan, and Rodgers (1994), Bound, Brown, and Mathiowetz (2001) and Bollinger (1998). These models are studied in Carroll and Wand (1991a), Sepanski and Carroll (1993), Lee and Sepanski (1995) and Chen, Hong, and Tamer (2003) under assumption 1. In the measurement error setup, the moment condition typically depends only on Y, which contains the unobserved true variables, e.g. income or union status. X contains *reported* variables of interest. In general, X do not necessarily have to be the reported values of Y, but

can also represent some proxy variables that contain information about Y. Assumption 1 in these models is stated as

$$f(Y|X, D = 1) = f(Y|X, D = 0).$$

For example, assumption 1 is satisfied in the **stratified sampling** design where a *nonrandom response based subsample* of the primary data is validated. In a typical example of this stratified sampling design, one first oversamples a certain subpopulation of the mismeasured variables X, and then validates the true variables Y corresponding to this nonrandom stratified subsample of X. It is common to oversample a subpopulation of the primary data set where more severe measurement error is suspected to be present. Assumption 1 is valid as long as the sampling procedure adopted to create the auxiliary data set is based only on information provided by the distribution of the primary data set. The stratified sampling procedure can be illustrated as follows. Let U be i.i.d $U(0,1)$ random variables independent of both X and Y, and let $p(X) \in (0,1)$ be a measurable function of the primary data. The stratified sample is obtained by validating every observation for which $U < p(X)$. In other words, $p(X)$ specifies the probability of validating an observation after X is observed. This sampling scheme corresponds to case (2). If $p(X)$ is a constant, the auxiliary data set Y, X is characterized by the same distribution of Y, X as the primary data set. In this case assumption 1 is easily seen satisfied, and this special case is typically referred to as the **validation** data set case. This framework is common in the statistics literature, where usually a random subset of the primary data is validated. The issue of semiparametric efficiency bound for the estimation of β has not been addressed in this literature. We derive these efficiency bounds and also present semiparametric efficient estimators under a variety of model assumptions.

Example 2: In the program evaluation literature, assumption 1 corresponds to unconfoundedness of treatment assignment, whereby treatment status (here represented by D) is assumed to be independent of potential outcomes, conditionally on observed covariates X (See e.g. the references surveyed in Heckman, LaLonde, and Smith (1999)). Let Y_0 denote the outcome if an individual is not treated. Y_0 is only observed when $D = 0$, while it is not observed when $D = 1$. Suppose one is interested in the average counter-factual untreated outcome Y_0 for people who are treated ($D = 1$), then the parameter of interest β is the solution to the moment condition (1) where $m(Z;\beta) = Y_0 - \beta$. On the other hand, if one is interested in the mean untreated outcome for the entire population, then the parameter of interest β is the solution to the moment condition (2) with the same $m(Z;\beta) = Y_0 - \beta$.

The analysis of efficient estimation of mean treatment effects has been pioneered by Hahn (1998), Hirano, Imbens, and Ridder (2003) and Heckman, Ichimura, and Todd (1998). Hahn (1998) laid down the cornerstone of semiparametric efficiency bounds and semiparametric efficient estimation for the estimation of average treatment effects and average treatment effects for the treated. Hahn (1998) and Hirano, Imbens, and Ridder (2003) developed semiparametric propensity score weighting estimators that achieve the efficiency bound with and without the knowledge of the true propensity score. Recently, Firpo (2004) considered efficient estimation of quantile treatment effects when the propensity score is unknown. Section (3.5) discusses the relation of our results to the treatment effect model, extending the mean analysis to general parameters defined through a possibly over-identified nonlinear moment condition when the propensity score is unknown, or known, or has a parametric specification.

Example 3: Several authors have shown that inverse probability weighting (also called propensity score weighting), under appropriate ignorability assumptions, allows identification and $\sqrt{n}$-consistent estimation of parameters in models with attrition or nonresponse. In this context, D is a binary variable indicating whether an observation is missing. Robins, Rotnitzky, and Zhao (1994) and Robins, Rotnitzky, and Zhao (1995) find the efficiency bound for nonlinear regression model with missing regressors. Rotnitzky and Robins (1995) obtains result with missing dependent variables. In panel data attrition analysis, Wooldridge (2002) and Wooldridge (2003) consider estimation of parameters identified by a condition such as (2) in presence of missing data. The main identifying assumption is that the probability of having a complete observation conditional on a set of auxiliary variables is independent on Y. GMM models with incomplete data have also been studied by Tripathi (2003) and Tripathi (2004), who analyze models where a true validation data set allows identification which is lost due to certain forms of censoring, truncation, or stratification. Wooldridge (2003) also considers the case where the auxiliary variables are not always observed (for example due to censored survival time), and the effect of misspecification of the parametric propensity score. He showed that when the propensity score assumes parametric form, estimation of the parameters leads to efficiency gains with respect to the case where a known propensity score is used. Interestingly, we find that estimation of the parametric propensity score does not achieve the efficient variance bound even if this parametric assumption is correctly specified. Instead, the semiparametric estimators of this paper achieve the variance bound under correct parametric propensity score specification.

Example 4: In Tarozzi (2004), the missing data problem stems from a change in survey methodology that leads to the non-comparability of reported statistics with those calculated from previous waves of the same survey. In his case, $D = 1$ denotes the observations that belong to the current (revised) survey, while $D = 0$ indicates observations from previous (auxiliary) surveys. Hence, the sampling process does not identify the parameters β defined by a moment condition such as (1). In this framework, Assumption 1 corresponds to the stability over time—that is, for different values of D—of the distribution of Y conditional on other observed auxiliary variables X, whose reports have not been affected by the change in survey design. Tarozzi (2004) describes a method of moments estimator based on parametric propensity score reweighting, where the propensity score is assumed to be appropriately described by a parametric model, and obtains large sample distributions that allow for the presence of complex survey design. In this paper we develop semiparametric estimators that are robust against the misspecification bias due to the incorrectly specified parametric form of propensity score.

Example 5: In poverty and inequality analysis, it is often desirable to have statistically precise estimates of income or consumption-based measures of welfare for small areas such as town or counties. However, this is rarely possible. While censuses possess the required "sample" size, they typically do not measure the necessary quantities. On the other hand, while most household surveys record consumption and/or income, they have a sample size which is too small to guarantee representativeness for small areas. Precise estimation can be achieved if the distribution of Y (income or consumption) is the same in the census and in the household survey, *conditional* on some covariates X that are recorded in both data sources.[2] In this framework, $D = 1$ for observations from the census (a larger sample where only X is observed), while

[2]Such covariates may include, for example, education achievements, housing characteristics, household size and demographic composition.

$D = 0$ for observations from the household survey (a smaller sample where both X and Y are observed). The identifying assumption is then again compatible with assumption 1 in our model. Elbers, Lanjouw, and Lanjouw (2003) develop a parametric simulation procedure for the micro-level estimation of poverty and inequality measures, and the calculation of correct standard errors. An alternative is the inverse probability weighting estimator described in Tarozzi (2004), which is easier to compute.

3 Semiparametric Efficiency Bounds

In this section we calculate the efficiency bound for the estimation of β defined by either moment conditions (1) or (2). The derivation is very closely related to Hahn (1998), except that we use a different factorization of the likelihood function for case (1). The semiparametric efficiency bounds are derived by calculating the efficient influence function associated with the pathwise derivatives of the parameters β. These efficient influence functions are the projection of the moment conditions onto the tangent space of all regular parametric submodels satisfying the moment restrictions.

To state the efficiency bounds we introduce some notations. Let n denote the size of a sample of observations on $Z_i = (Y_i, X_i), D_i$, where Y_i is only observed when $D_i = 0$. Let $p = Pr(D = 1)$ and $p(X) = Pr(D = 1|X)$. In this paper we use β to mean an arbitrary value in the parameter space, but to save notation β is sometimes used as the true parameter value β_0 in this section. Define

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

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, also define

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

Assumption 2 *(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 comes from an i.i.d. sample; (iii) $p = Pr(D = 1) \in (0, 1)$.*

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

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

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

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

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

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

3.1 Information content of the propensity score

It is interesting to analyze whether knowing $p(X)$ decreases the semiparametric efficiency bounds for the parameters β. Hahn (1998) showed that it does for estimation of the average effect of treatment on the treated, while the propensity score is ancillary for the average treatment effect. A similar result holds for the GMM model discussed here.

Theorem 2 *Under assumptions 1 and 2, if $p(X)$ is known, then the asymptotic variance bound for estimating β is*

$$\left(\mathcal{J}_\beta'\tilde{\Omega}_\beta^{-1}\mathcal{J}_\beta\right)^{-1}.$$

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

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

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 efficiency bound for β when it is defined by the moment condition (1), but plays no role in its variance bound when β is defined by the moment condition (2). The following argument provides an intuition for this result. The joint density function of the observed data $f(Y,X,D)$ can be factorized as

$$f(Y,X,D) = f(X)\,p(X)^D f(Y|X)^{1-D}(1-p(X))^{1-D}.$$

When (2) holds, β is defined through the relation

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

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

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

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. The following theorem provides an answer. First we assume that the propensity score takes a parametric form $p(X;\gamma)$, and denote $p_\gamma(X) = \frac{\partial p(X;\gamma)}{\partial \gamma'}$. Also define the score function for γ as

$$S_\gamma(D_i;X_i) = \frac{D_i - p(X_i;\gamma)}{p(X_i;\gamma)(1-p(X_i;\gamma))}p_\gamma(X_i).$$

Theorem 3 *Under assumptions 1 and 2, if $p(X) = p(X;\gamma)$ belongs to a correctly specified parametric family indexed by γ and $E[S_\gamma(D,X)S_\gamma(D,X)']$ is positive definite, then the efficient variance bound for estimating β defined by moment condition (1) is given by $\left(\mathcal{J}_\beta'\tilde{\Omega}_\beta^{-1}\mathcal{J}_\beta\right)^{-1}$ where $\mathcal{J}_\beta = \mathcal{J}_\beta^1$ and*

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

This variance bound is clearly larger than $\tilde{\Omega}_\beta^1$ stated in Theorem 2. The bound in Theorem 3 can also be described in terms of the variance of the following influence function:

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

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

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

so the variance bound stated in Theorem 3 is smaller than that in Theorem 1 for moment condition (1).

Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995) and Rotnitzky and Robins (1995) presented a unified framework for calculating the efficient projection into tangent spaces of Newey (1990b) and Bickel, Klaassen, Ritov, and Wellner (1993) when the tangent spaces are defined by general patterns of missing data in conditional mean models with missing regressors. The parts concerning moment condition (2) in theorems 1 and 2 can also be derived from proposition 8.2 of Robins, Rotnitzky, and Zhao (1994) when there is a single hierarchy in the case of monotone missing data patterns and when the instrument functions are given in the conditional mean model.

3.2 Information content of the stratifying scheme

A related question is which components of the joint density, other than $p(X)$, might be ancillary for estimating β when it is defined by moment conditions (1) or (2). For this purpose, consider an alternative factorization of the joint density function:

$$f\,(Y,X,D)=p^D(1-p)^{1-D}f\,(X|D=1)^D\,f\,(X|D=0)^{1-D}\,f\,(Y|X)^{1-D}\,.$$

Moment condition (1) can now be written as

$$\int\int m\,(y,x;\beta)\,f\,(y|x)\,dyf\,(x|D=1)\,dx=0.$$

This relation clearly depends only on $f\,(y|x)$ and $f\,(x|D=1)$, and not on $f\,(x|D=0)$ or p. Therefore, both the conditional distribution of X in the auxiliary sample and the size of the auxiliary data set relative to the primary data are ancillary to the estimation of β under (1). On the other hand, $f\,(x|D=0)$ and p are not ancillary for β under (2), since

$$\int\int m\,(y,x;\beta)\,f\,(y|x)\,dy\,[pf\,(x|d=1)+(1-p)\,f\,(x|D=0)]\,dx=0.$$

Knowledge of the stratifying sampling scheme does not reduce the variance bound if the auxiliary data set is disjoint from the primary data set, but does reduce the variance bound if the auxiliary data set is a subset of the primary data set. The following theorem formalizes the previous discussion.

Theorem 4 *Under assumptions 1 and 2, if p and $f\,(x|D=0)$ are known, the asymptotic variance bound for estimating β is*

$$\left(\mathcal{J}_\beta'\tilde{\Omega}_\beta^{-1}\mathcal{J}_\beta\right)^{-1}.$$

When moment condition (1) holds, $\mathcal{J}_\beta \equiv \mathcal{J}_\beta^1$ *and* $\tilde{\Omega}_\beta = \Omega_\beta^1$ *given in Theorem 1. When moment condition (2) holds,* $\mathcal{J}_\beta \equiv \mathcal{J}_\beta^2$ *and* $\tilde{\Omega}_\beta = \tilde{\Omega}_\beta^2$ *(which is smaller than* Ω_β^2 *in theorem 1) where*

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

The two components in Ω_β^1 and Ω_β^2 stated in Theorem 1 are due to the lack of knowledge of relevant elements of the joint density $f(Y,X,D)$ that contribute to the definition of β. More precisely, the two terms in Ω_β^1 can be attributed to the fact that $f(y|x)$ and $f(x|D=1)$ are unknown, while the two components in Ω_β^2 can be attributed to the fact that $f(y|x)$ and $f(x)$ have to be estimated, respectively. The following theorem formalizes the reduction in the variance bound due to knowledge of these components.

Theorem 5 *Under assumptions 1 and 2, if* $f(x|D=1)$ *is known but* $f(y|x)$ *unknown, the asymptotic variance for estimating* β *defined by the moment condition (1) is* $\left(\mathcal{J}_\beta'\tilde{\Omega}_\beta^{-1}\mathcal{J}_\beta\right)^{-1}$ *where* $\mathcal{J}_\beta = \mathcal{J}_\beta^1$ *and*

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

If $f(y|x)$ *is known but* $f(x|D=1)$ *unknown, the efficient variance for estimating* β *defined by the moment condition (1) is* $\left(\mathcal{J}_\beta'\tilde{\Omega}_\beta^{-1}\mathcal{J}_\beta\right)^{-1}$ *where* $\mathcal{J}_\beta = \mathcal{J}_\beta^1$ *and*

$$\tilde{\Omega}_\beta = E\left[\frac{p(X)}{p^2}\mathcal{E}(X;\beta)\,\mathcal{E}(X;\beta)'\right].$$

If $f(x)$ *is known but* $f(y|x)$ *unknown, the efficient variance for* β *defined by (2) is* $\left(\mathcal{J}_\beta'\tilde{\Omega}_\beta^{-1}\mathcal{J}_\beta\right)^{-1}$ *where* $\mathcal{J}_\beta = \mathcal{J}_\beta^2$ *and*

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

If $f(y|x)$ *is known but* $f(x)$ *unknown, the efficient variance for* β *defined by (2) is* $\left(\mathcal{J}_\beta'\tilde{\Omega}_\beta^{-1}\mathcal{J}_\beta\right)^{-1}$ *where* $\mathcal{J}_\beta = \mathcal{J}_\beta^2$ *and*

$$\tilde{\Omega}_\beta = E\left[\mathcal{E}(X;\beta)\,\mathcal{E}(X;\beta)'\right].$$

3.3 Validation samples

A special case of assumption 1 is when the auxiliary sample is randomly drawn from the same population of the primary sample. In this case the auxiliary sample is called a *validation* sample (e.g. Carroll and Wand (1991a), Sepanski and Carroll (1993), Lee and Sepanski (1995)), and the following assumption holds.

Assumption 3 $Y, X \perp D$.

This assumption is also called "random assignment" in the treatment effect literature (Hahn (1998)), or "missing completely at random" in panel data attrition analysis. It is easy to see that assumption 3 is equivalent to adding to assumption 1 the statement that $X \perp D$, or that $p(X) = p$ (an unknown constant). Within this framework, the moment conditions (1) and (2) coincide with each other, so that the semiparametric variance bound stated in the next theorem applies to both conditions.

Theorem 6 *Under assumption 2 and assumption 3, the semiparametric variance bound for the parameter β defined through the moment condition of (1) or (2) is given by $\left(\mathcal{J}_\beta'\Omega_v^{-1}\mathcal{J}_\beta\right)^{-1}$, where $\mathcal{J}_\beta = \mathcal{J}_\beta^1 = \mathcal{J}_\beta^2$ and*

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

The semiparametric variance bound Ω_v in theorem 6 is identical to Ω_β^2 given in theorem 1. This is because $p\left(X\right)$ is ancillary to β defined through the moment condition (2) as shown in Theorem 1. However, knowledge of assumption 3 does decrease the semiparametric variance bound for β defined through moment condition (1). Intuitively, assumption 3 implies that $f\left(X|D=0\right)$ provides useful information about β in addition to those provided by $f\left(Y|X,D=0\right)$ in the auxiliary sample. On the one hand, when moment condition (2) holds, the entire sample is used in the estimation, so that knowledge of assumption 3 in addition to assumption 1 does not help increase estimation efficiency. On the other hand, when moment condition (1) holds, if only assumption 1 is being used when in fact assumption 3 also holds, then only the information contained in $f\left(X|D=1\right)$ and $f\left(Y|X,D=0\right)$ is being used in the estimation of β, and the information contained in $f\left(X|D=0\right)$ is not being utilized. Therefore in this case assumption 3 provides additional information for estimation efficiency.

3.4 Parametric maximum likelihood models

Researchers are frequently interested in estimating a parametric likelihood model in the primary data set, see e.g. Wooldridge (2002). In this case β_0 is defined by either

$$\beta_0 = \max_\beta E\left[\log g\left(Y;\beta\right)|D=1\right] \tag{3}$$

or

$$\beta_0 = \max_\beta E\left[\log g\left(Y;\beta\right)\right], \tag{4}$$

where $g\left(Y;\beta\right)$ is a parametric likelihood for Y, or a conditional likelihood for a subvector of Y given the other components of Y. Since Y is not observed when $D=1$, assumption 1 provides an identification strategy for β by rewriting the expected log likelihood function as

$$E\left[E\left[\log g\left(Y;\beta\right)|D=0,X\right]|D=1\right] \quad \text{or} \quad E\left[E\left[\log g\left(Y;\beta\right)|D=0,X\right]\right].$$

Given each β, the conditional expectation $E\left[\log g\left(Y;\beta\right)|D=0,X\right]$ can be recovered from the auxiliary sample, and it can then be averaged over the primary data set to obtain the unconditional expectations in either (3) or (4). The following Theorem 7, whose proof is omitted, shows that the semiparametric efficiency bounds for models (3) and (4) are special cases of the framework analyzed in theorem 1 where the score functions for the likelihood model $g\left(\cdot\right)$ serve as the set of exactly identified moment conditions in (1) and (2). First of all define,

$$\mathcal{J}_\beta^3 = \frac{\partial^2}{\partial\beta\partial\beta'}E\left[\log g\left(Y;\beta\right)|D=1\right] \quad \text{and} \quad \mathcal{J}_\beta^4 = \frac{\partial^2}{\partial\beta\partial\beta'}E\left[\log g\left(Y;\beta\right)\right].$$

Also define $\mathcal{E}(X;\beta) = E\left[\frac{\partial}{\partial\beta}\log g(Y;\beta)\,|X\right]$ and

$$\Omega_\beta^3 = E\left[\frac{p(X)^2}{p^2(1-p(X))}V\left(\frac{\partial}{\partial\beta}\log g(Y;\beta)\,\middle|X\right) + \frac{p(X)}{p^2}\mathcal{E}(X;\beta)\,\mathcal{E}_\beta(X)'\right],$$

and

$$\Omega_\beta^4 = E\left[\frac{1}{1-p(X)}V\left(\frac{\partial}{\partial\beta}\log g(Y;\beta)\,\middle|X\right) + \mathcal{E}(X;\beta)\,\mathcal{E}(X;\beta)'\right].$$

Theorem 7 *Under assumptions 1 and 2, the asymptotic semiparametric variance bounds for the parametric likelihood models (3) and (4) are given by* $\mathcal{J}_\beta^{-1}\Omega_\beta\mathcal{J}_\beta^{-1}$, *where* $\mathcal{J}_\beta = \mathcal{J}_\beta^3$ *and* $\Omega_\beta = \Omega_\beta^3$ *under model (3), and* $\mathcal{J}_\beta = \mathcal{J}_\beta^4$ *and* $\Omega_\beta = \Omega_\beta^4$ *under model (4).*

3.5 Relation to treatment effect models

In the language of the program evaluation literature, the results described above are sufficient to describe the semiparametric bounds for the estimation of either mean *treated* or mean *untreated* outcomes. However, the above framework has to be modified to accommodate the calculation of bounds for *mean treatment effects* (for the treated, or for the whole population), which are typically the real parameters of interest. In the models we have considered so far, the auxiliary data set identifies the joint distribution of all the variables included in Z. However, in program evaluation the researcher only observes *either* the untreated outcome Y_0 (when $D=0$) *or* the treated outcome Y_1 (when $D=1$), but never both. The GMM framework can be adapted to generalize the calculation of nonparametric efficiency bounds for the estimation of average treatment effects if we let $Y = DY_1 + (1-D)Y_0$, and if we assume that the parameters of interest are identified by the following separable moment condition:

$$m(Z;\beta) = m_1(Y_1, X;\beta) - m_0(Y_0, X;\beta). \tag{5}$$

In the following, define $\mathcal{E}(X;\beta) = \mathcal{E}_1(X;\beta) - \mathcal{E}_0(X;\beta)$, where $\mathcal{E}_j(X;\beta) = E[m_j(Y_j, X;\beta) \mid X]$, $j = 0, 1$. Also, define

$$V_j(X) = E\left(m_j(Y_j, X;\beta)\,m_j(Y_j, X;\beta)' \mid X\right) - \mathcal{E}_j(X;\beta)\,\mathcal{E}_j(X;\beta)', \quad j = 0, 1$$

to be the conditional variances of the moment functions given X. The following results can be proved similarly to theorem 1, using the property that the linear projection of a sum is the sum of the linear projections.

Theorem 8 *Under assumptions 1 and 2, if the moment condition takes the form of (5), the semiparametric variance bound for* β *is* $\left(\mathcal{J}_\beta'\Omega_\beta^{-1}\mathcal{J}_\beta\right)^{-1}$. *When the moment condition (1) holds,* $\mathcal{J}_\beta = \mathcal{J}_\beta^1$ *and* $\Omega_\beta = \Omega_\beta^1$ *where*

$$\Omega_\beta^1 = E\left[\frac{p(X)}{p}V_1(X) + \frac{p(X)^2}{p^2(1-p(X))}V_0(X) + \frac{p(X)}{p^2}\mathcal{E}(X;\beta)\,\mathcal{E}(X;\beta)'\right].$$

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

$$\Omega_\beta^2 = E\left[\frac{1}{p(X)}V_1(X) + \frac{1}{1-p(X)}V_0(X) + \mathcal{E}(X;\beta)\,\mathcal{E}(X;\beta)'\right].$$

These results extend the mean analysis in Hahn (1998) and the weighted mean treatment in Hirano, Imbens, and Ridder (2003) to parameters defined by general additive separable moment conditions in the form of (5), and also include as a special case the quantile treatment effect analysis of Firpo (2004).

4 CEP-GMM Estimation

There are two alternative approaches for semiparametric estimation of the parameter β when assumption 1 holds and if Y is only observed when $D = 0$. The first one is based on a conditional expectation projection method. The second one is based on the inverse probability weighting (IPW) or propensity score weighting method. In this section we show that the optimally weighted GMM estimator of β using a sieve conditional expectation projection approach is semiparametrically efficient. We discuss IPW in the next section.

4.1 Efficient estimation with unknown propensity score

4.1.1 The estimator and heuristics for asymptotic variances

Under assumption 1, $\mathcal{E}(X;\beta) = E[m(Z;\beta)|X, D = 0]$ for all β, and the moment condition (1) is equivalent to

$$E[\mathcal{E}(X_i;\beta_0)|D_i = 1] = 0,$$

while the moment condition (2) is simply

$$E[\mathcal{E}(X_i;\beta_0)] = 0.$$

The projection method first estimates $\mathcal{E}(X;\beta)$ nonparametrically from the auxiliary sample, and then averages this nonparametric estimator over the primary sample. 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. Chen, Hong, and Tamer (2003) (hereafter CHT) used a sieve based method for this nonparametric estimation. Let $\{q_l(X), l = 1, 2, ...\}$ denote a sequence of known basis functions that can approximate any square-measurable function of X arbitrarily well. Also let

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

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,

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

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

$$\hat{\beta} = \arg\min_{\beta \in B} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \hat{\mathcal{E}}\left(X_{pi}; \beta\right) \right)' \hat{W} \left(\frac{1}{n_p} \sum_{i=1}^{n_p} \hat{\mathcal{E}}\left(X_{pi}; \beta\right) \right). \tag{6}$$

This projection based GMM method (CEP-GMM) is closely related to the imputation method of Hahn (1998). The difference is that Hahn (1998)'s imputation method plugs in Y whenever it is observed while the projection method only uses $\hat{\mathcal{E}}\left(X; \beta\right)$.

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

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

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

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

$$\frac{f_{X_p}(X_{aj})}{f(X_{aj}|D=0)} = \frac{\left(1-p\right) p\left(X\right)}{p\left(1-p\left(X\right)\right)}.$$

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

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

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

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

where Ω_β^1 is given in Theorem 1.

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

$$\frac{f_{X_p}(X_{aj})}{f(X_{aj}|D=0)} = \frac{\left(1-p\right)}{\left(1-p\left(X\right)\right)}.$$

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

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

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

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

where Ω_β^2 is given in Theorem 1.

The optimally weighted GMM estimator $\hat{\beta}$ for β_0 that uses a weighting matrix $\hat{W} = \Omega_\beta^{-1} + o_p(1)$ will achieve the semiparametric efficiency bound of $\left(\mathcal{J}_\beta' \Omega_\beta^{-1} \mathcal{J}_\beta\right)^{-1}$ given in Theorem 1.

4.1.2 Asymptotic properties

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 finite 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 modification of the presentation of the assumptions, the large sample results stated below remain valid. To avoid tedious notation yet to allow for some unbounded support elements of X, we assume $\mathcal{X} = \mathcal{X}_c = \mathcal{R}^{d_x}$ in this paper. 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 $\underline{\gamma}$ -th 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 $q^{k_{na}}(x) = (q_1(x), ..., q_{k_{na}}(x))'$ 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 4 *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 9 *Let $\widehat{\beta}$ be the CEP-GMM estimator given in (6). Under assumptions 1, 2 and 4, 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 5 *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 4.1 is satisfied with $\gamma > d_x/2$ and assumption 4.2 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. *$k_{na} = O\left((n_a)^{\frac{d_x}{2\gamma+d_x}}\right)$, $(n_a)^{-\frac{\gamma}{2\gamma+d_x}} \times \left\|\frac{f_{X_p}(\cdot)}{f_{X_a}(\cdot)} - \Pi_{2n}\frac{f_{X_p}(\cdot)}{f_{X_a}(\cdot)}\right\|_{2,a} = o(n^{-1/2})$.*

Theorem 10 *Let $\widehat{\beta}$ be the CEP-GMM estimator given in (6). Under Assumptions 1, 2, 4 and 5, 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

$$V_0 = \left(\mathcal{J}_\beta'\Omega_\beta^{-1}\mathcal{J}_\beta\right)^{-1},$$

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 1: **(i)** Assumptions 4 and 5 allow for $m(Z;\beta)$ to be non-smooth such as in quantile based moment functions. **(ii)** The weightings ω and ω_1 are needed since the support of the conditioning variable X is assumed to be the entire Euclidean space in this paper. 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 4 and 5, and replace assumption 5.1 by that assumption 4.1 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 $0 < \underline{p} \leq p(x) \leq \overline{p} < 1$ imposed in Hirano, Imbens, and Ridder (2003) and Firpo (2004). Assumption 5.5 will be satisfied under mild smoothness conditions imposed on $\frac{p(X)}{1-p(X)}$ that are weaker than those imposed in Hirano, Imbens, and Ridder (2003) and Firpo (2004). In particular, if we let $k_{na} = O\left((n_a)^{\frac{d_x}{2\gamma+d_x}}\right)$, 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 5.5 is satisfied with $\left\|\frac{f_{X_p}(\cdot)}{f_{X_a}(\cdot)} - \Pi_{2n}\frac{f_{X_p}(\cdot)}{f_{X_a}(\cdot)}\right\|_{2,a} = o\left((n_a)^{-\frac{d_x}{2(2\gamma+d_x)}}\right) = o\left((k_{na})^{-\frac{1}{2}}\right)$. For example, both assumptions 5.4 and 5.5 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 9 and 10 follow directly from those in CHT, who also provide simple consistent estimators of V and V_0:

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

where for moment condition (1),

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

and

$$\begin{aligned}
\widehat{\Omega} &= \frac{1}{n_a}\sum_{j=1}^{n_a}\left(\widehat{v^*_{aj}}\widehat{U}_{aj}\right)\left(\widehat{v^*_{aj}}\widehat{U}_{aj}\right)' + \frac{n_a}{n_p^2}\sum_{i=1}^{n_p}\left(\widehat{\mathcal{E}}(X_{pi};\widehat{\beta})\widehat{\mathcal{E}}(X_{pi};\widehat{\beta})'\right) \\
\widehat{U}_{aj} &= m(Y_{aj}, X_{aj};\widehat{\beta}) - \widehat{\mathcal{E}}(X_{aj};\widehat{\beta}), \quad \widehat{v^*_{aj}} = [\frac{1}{n_p}\sum_{i=1}^{n_p}q^{k_{na}}(X_{pi})]'\left(\frac{Q_a'Q_a}{n_a}\right)^{-1}q^{k_{na}}(X_{aj}).
\end{aligned}$$

4.2 Efficient estimation with known propensity score

Suppose now that the propensity score $p(X)$ is known. Theorems 2 and 10 show that the optimally weighted CEP-GMM estimator given in (6) still achieves the semiparametric efficiency bound for β defined by moment condition (2). Even if the estimator is no longer efficient for β defined through moment condition (1), it is possible to construct an efficient estimator for case (1) using the sieve estimate $\hat{\mathcal{E}}(X;\beta)$ and the known $p(X)$. Notice that under assumption 1, the moment condition (1) is equivalent to

$$E\left[\mathcal{E}(X_i;\beta_0)\frac{p(X_i)}{p}\right] = 0.$$

Hence the optimally weighted GMM using the following sample moment condition will give an efficient estimator for β_0 defined through (1):

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

where $\hat{p} = \frac{n_p}{n}$. From the proof in CHT, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\mathcal{E}}(X_i;\beta_0)\frac{p(X_i)}{\hat{p}}$ is asymptotically equivalent to

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

which has asymptotic variance $\tilde{\Omega}^1_\beta$, the semiparametric efficiency bound stated in Theorem 2.

4.3 Efficient estimation with a validation sample

Suppose now that assumption 3 holds, that is, the auxiliary data set is actually a validation data set, so that $p(X) = p$. Theorems 6 and 10 show that the optimally weighted CEP-GMM estimator defined in (6) still achieves the semiparametric efficiency bound for β defined by moment condition (2), but not for β defined through moment condition (1), $\Omega^1_\beta > \Omega^2_\beta$ in Theorem 1. Intuitively, this is because the estimator using (6) for case (1) does not exploit the fact that

$$\frac{1}{n_a}\sum_{j=1}^{n_a} m\left(Z_j;\beta_0\right) \tag{7}$$

should be close to zero. This set of moment conditions is not orthogonal to $\frac{1}{n_p}\sum_{i=1}^{n_p}\hat{\mathcal{E}}\left(X_{pi};\beta\right)$, which are used in (6). But it does contain additional information about $f\left(X|D=0\right)$ that is useful for estimating β_0 defined by (1).

Under assumption 3, one might be tempted to estimate β_0 using the validation data set moment condition alone:

$$\check{\beta} = \arg\min_{\beta}\left(\frac{1}{n_a}\sum_{j=1}^{n_a} m\left(Z_j;\beta\right)\right)' \hat{W}\left(\frac{1}{n_a}\sum_{j=1}^{n_a} m\left(Z_j;\beta\right)\right).$$

When $\hat{W}$ is optimally chosen, $\sqrt{n}\left(\check{\beta}-\beta_0\right)$ has asymptotic variance $\left(\mathcal{J}'_\beta\check{\Omega}^{-1}_\beta\mathcal{J}_\beta\right)^{-1}$ where

$$\check{\Omega}_\beta = \frac{1}{1-p}Var\left(m\left(Z_j;\beta_0\right)\right) > \Omega_v,$$

where

$$\Omega_v = E\left[\frac{1}{1-p}V\left[m\left(Z;\beta_0\right)\mid X\right] + \mathcal{E}\left(X;\beta_0\right)\mathcal{E}\left(X;\beta_0\right)'\right] = \frac{1}{1-p}\left[V\left(m\left(Z;\beta_0\right)\right) - pV\left(\mathcal{E}\left(X;\beta_0\right)\right)\right]$$

is the efficient bound stated in theorem 6 under assumption 3. Therefore the estimator $\check{\beta}$ that uses the validation sample alone is not efficient.

Efficiency can be achieved by an estimator that optimally combines the two moment conditions (6) and (7). However, a simpler alternative efficient estimator is the optimally weighted GMM using the following sample moment

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{E}}\left(X_i;\beta\right),$$

which is the same moment condition as (6) but taking n_p to be the entire sample instead of just the subsample where $D=1$. This is because under assumption 3, moment condition (1) becomes the same as case (2): $E\left[\mathcal{E}\left(X_i;\beta_0\right)\right]=0$, and therefore one should use all the observations in the sample. Again from the proofs in CHT and under assumption 3, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\mathcal{E}}\left(X_i;\beta_0\right)$ has the asymptotic representation

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

which has asymptotic variance Ω_v, the semiparametric efficiency bound stated in Theorem 6. During a recent private conversation, Newey has suggested another simple efficient GMM estimator of β_0 using an increasing set of moments $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_i m(Z_i;\beta)$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i-\hat{p})A(X_i)$, where $A(X_i)$ is a growing set of measurable function of X_i.

4.4 Efficient estimation with parametric 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 10 show that the optimally weighted CEP-GMM estimator defined in (6) still achieves the semiparametric efficiency bound for β defined by moment condition (2). However, according to Theorems 3 and 10, such an estimator is no longer efficient for β defined through moment condition (1).

Using the equivalent expression of moment condition (1): $E\left[\mathcal{E}(X_i;\beta_0)\frac{p(X_i)}{p}\right]=0$, 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 efficient bound in Theorem 3 for β defined through (1):

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{E}}(X_i;\beta)\frac{p(X_i;\hat{\gamma})}{\hat{p}}, \tag{8}$$

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 11 *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 (8). Under assumptions 1, 2, 4 and 5, 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 $V_0=\left(\mathcal{J}_\beta^{1\prime}\tilde{\Omega}_\beta^{-1}\mathcal{J}_\beta^1\right)^{-1}$ is the efficiency variance bound given in Theorem 3.

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

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

where

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

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 efficient variance bound for β.

We also remark that if we are willing to assume that a parametric assumption $f_\gamma(Y|X)$ is correctly specified, then we can replace $\hat{\mathcal{E}}(X;\beta)$ by a parametric estimate:

$$\mathcal{E}(X;\beta,\hat{\gamma}) = \int m(y,x;\beta)\, f_{\hat{\gamma}}(y|X)\, dy,$$

where $\hat{\gamma}$ is a maximum likelihood estimate for γ using the subsample where $D=0$:

$$\sum_{i=1}^{n}(1-D_i)\, S_{\hat{\gamma}}(Y_i|X_i) = \sum_{i=1}^{n}(1-D_i)\frac{\partial \log f(Y_i|X_i;\hat{\gamma})}{\partial\gamma} = 0.$$

It is easy to show that this achieves the semiparametric efficiency bound when $f_\gamma(Y|X)$ is correctly parameterized. For illustration, consider only the case when $p(X_i)$ is unknown. In this case, the efficient influence functions for the moment conditions given a correctly specified parametric model of $f_\gamma(Y|X)$ are

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

under moment condition (1) and

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

under moment condition (2). Next, we note that the parametric estimated moment condition has the following influence function:

$$\frac{\sqrt{n}}{n_p}\sum_{i=1}^{n_p}\mathcal{E}(X_{pi};\beta_0,\hat{\gamma}) = \frac{\sqrt{n}}{n_p}\sum_{i=1}^{n_p}\mathcal{E}(X_{pi};\beta_0,\gamma_0)+\frac{\partial}{\partial\gamma}E_{n_p}\mathcal{E}(X_{pi};\beta_0,\gamma_0)\sqrt{n}(\hat{\gamma}-\gamma)+o_p(1),$$

where E_{n_p} denotes expectation taken with respect to the primary sample. Under moment condition (1), it can be calculated that

$$\begin{aligned}&\tfrac{\partial}{\partial\gamma}E_{n_p}\mathcal{E}(X_{pi};\beta_0,\gamma_0) = \tfrac{\partial}{\partial\gamma}E\left[\mathcal{E}(X_{pi};\beta_0,\gamma_0)\,|D=1\right]\\ =\;& E\left(\tfrac{p(X)}{p}\tfrac{1-D}{1-p(X)}\left[m(Z;\beta_0)-\mathcal{E}(X;\beta_0)\right]\times(1-D)\,S_\gamma(Y|X)\right),\end{aligned}$$

and hence up to $o_p(1)$, $\frac{\sqrt{n}}{n_p}\sum_{i=1}^{n_p}\mathcal{E}(X_{pi};\beta_0;\hat{\gamma})$ is

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{D_i}{p}\mathcal{E}(X_i;\beta_0)+\text{Proj}\left(\frac{1-D}{1-p(X)}\left[m(Z;\beta_0)-\mathcal{E}(X;\beta_0)\right]\middle|\,(1-D_i)\,S_\gamma(Y_i|X_i)\right)\right],$$

which is identical to the desired efficient influence function.

Similarly, under moment condition (2), it can be calculated that

$$\begin{aligned}&\tfrac{\partial}{\partial\gamma}E_{n_p}\mathcal{E}(X_{pi};\beta_0,\gamma_0) = \tfrac{\partial}{\partial\gamma}E\left[\mathcal{E}(X_{pi};\beta_0,\gamma_0)\right]\\ =\;& E\left(\tfrac{1-D}{1-p(X)}\left[m(Z;\beta_0)-\mathcal{E}(X;\beta_0)\right]\times(1-D)\,S_\gamma(Y|X)\right),\end{aligned}$$

and that up to $o_p(1)$, $\frac{\sqrt{n}}{n_p}\sum_{i=1}^{n_p}\mathcal{E}(X_{pi};\beta_0,\hat{\gamma})$ is

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\mathcal{E}(X_i;\beta_0)+\text{Proj}\left(\frac{1-D}{1-p(X)}\left[m(Z;\beta_0)-\mathcal{E}(X;\beta_0)\right]\middle|\,(1-D_i)\,S_\gamma(Y_i|X_i)\right)\right].$$

This also achieves the efficient influence function. Similarly, using the parametric $\mathcal{E}(X;\beta,\hat{\gamma})$ also achieves the semiparametric efficiency bound in cases when the propensity score is known, or is correctly parametrically specified.

5 IPW-GMM Estimation

A popular alternative estimation method for β is the inverse probability weighting based GMM (IPW-GMM). Examples include parametric inverse probability weighting as in Wooldridge (2002), Wooldridge (2003) and Tarozzi (2004) for missing data models, and nonparametric inverse probability weighting as in Hirano, Imbens, and Ridder (2003) for the case of mean treatment effect analysis. In this section, we extend their 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; clever combinations of nonparametric and known or parametric estimated propensity scores are needed to achieve the semiparametric efficiency bounds for these cases.

5.1 Efficient estimation with unknown propensity score

5.1.1 the estimator and heuristics for asymptotic variances

The IPW-GMM method is based on the fact that under assumption 1, Moment condition (1) is equivalent to:

$$E\left[m(Z;\beta_0)\frac{p(X)(1-p)}{(1-p(X))p}|D=0\right]=0;$$

while moment condition (2) is equivalent to:

$$E\left[m(Z;\beta_0)\frac{1-p}{1-p(X)}|D=0\right]=0.$$

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

$p(X)$. This is pointed out by Hirano, Imbens, and Ridder (2003) in the treatment effect literature. The influence function can be orthogonalized as

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

This is identical to the efficient influence function calculated in Theorem 1 for moment condition (1). Therefore,

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

where Ω_β^1 is given in Theorem 1. An optimally weighted GMM estimator for β_0 defined by case (1) using this sample moment (9) should then achieve the semiparametric efficiency bound of $\left(\mathcal{J}_\beta^{1\prime}(\Omega_\beta^1)^{-1}\mathcal{J}_\beta^1\right)^{-1}$.

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

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

The two components of this influence function are again negatively correlated. The influence function can be orthogonalized by writing it as

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

which is also identical to the efficient influence function calculated in the proof of Theorem 1 for moment condition (2). Therefore,

$$Avar\left(\sqrt{n}\frac{1}{n_a}\sum_{j=1}^{n_a}m\left(Z_j;\beta_0\right)\frac{1-p}{1-\hat{p}\left(X_j\right)}\right)=\Omega_\beta^2,$$

where Ω_β^2 is given in Theorem 1. An optimally weighted GMM estimator for β_0 defined by case (2) using this sample moment (10) should then achieve the semiparametric efficiency bound of $\left(\mathcal{J}_\beta^{2\prime}(\Omega_\beta^2)^{-1}\mathcal{J}_\beta^2\right)^{-1}$.

5.1.2 Asymptotic properties

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.[3] Let $\widehat{p}(\cdot)$ be a sieve estimator of $p_o(x)$ using the combined sample $\{(D_i,X_i):i=1,...,n=n_a+n_p\}$. 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

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

[3]Note that to save notations in the rest of the main text $p(x)$ denotes true propensity score function.

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 LS estimator $\widehat{p}_{ls}(x)$ as in Hahn (1998), Das, Newey, and Vella (2003):

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

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 ML estimator $\widehat{p}_{mle}(x)$ as in Hirano, Imbens, and Ridder (2003):

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

Recall that $E_a(\cdot) = \int (\cdot) f_{X|D=0}(x) dx$. We define a weighted sup-norm as

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

for some $\omega > 0$.

Assumption 6 *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 12 *Let $\widehat{\beta}$ be the IPW-GMM estimator given in (11) or (12). Under assumptions 1, 2 and 6, 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 7 *: 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 6.1 and 6.2 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 13 *Let $\widehat{\beta}$ be the IPW-GMM estimator given in (11) or (12). Under Assumptions 1, 2, 6 and 7, we have $\sqrt{n}(\widehat{\beta}-\beta_0) \Rightarrow \mathcal{N}(0,V)$, with V the same as that in Theorem 10.*

Remark 2: **(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 6 and 7 and replace assumption 7.1 by that assumption 6.1 holds with $\gamma > d_x/2$. Note that assumption 7.6a is easily satisfied when X has compact support. When $\mathcal{X} = \mathcal{R}^{d_x}$, assumption 7.6a rules out $\mathcal{E}(x,\beta)$ being linear in x; assumption 7.6b or 7.6c allows for linear $\mathcal{E}(x,\beta)$ but needs smoother propensity score $p(x)$ and density $f_{X|D=0}$. **(ii)** Assumptions 6 and 7 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 6 and 7 are effectively equivalent. **(iv)** Although assumption 6.1 imposes the same strong condition $0 < \underline{p} \leq p_o(x) \leq \overline{p} < 1$ as that in Hirano, Imbens, and Ridder (2003) and Firpo (2004), we relax their other conditions by allowing for unbounded support of X and assume 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 7.5 is satisfied with $\left\|\frac{\mathcal{E}(\cdot;\beta_o)}{1-p_o(\cdot)} - \Pi_{2n}\frac{\mathcal{E}(\cdot;\beta_o)}{1-p_o(\cdot)}\right\|_2 = o\left((n)^{-\frac{d_x}{2(2\gamma+d_x)}}\right) = o\left((k_n)^{-\frac{1}{2}}\right)$.

5.2 Efficient estimation with known propensity score

Suppose now that the propensity score $p(X)$ is known. According to Theorems 1, 2 and 13, the optimally weighted IPW-GMM estimator using the sample moment (10) with a nonparametric estimate $\hat{p}(X)$ will still be efficient for β_0 defined by case (2). However, the optimally weighted IPW-GMM estimator using

the sample moment (9) with nonparametric estimate $\hat{p}(X)$ will not be efficient for β defined by case (1). In section 4, we already pointed out that the CEP-GMM estimator using the sample moment condition $\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{E}}\left(X_i;\beta\right)\frac{p(X_i)}{\hat{p}}$ will give an efficient estimator for β defined by case (1). Another efficient moment condition that will achieve the same purpose is the one using both the sieve estimate $\hat{p}\left(X\right)$ and the known $p\left(X\right)$:

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

This is inspired by the work of Hirano, Imbens, and Ridder (2003), and it has the efficient asymptotic linear representation of

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

5.3 Estimation with parametric propensity score

In section 4.4 we have shown that one can construct sieve based CEP-GMM estimators that achieve the efficiency bound even if the propensity score has a known parametric form. This raises the interesting question of whether the parametric propensity score versions of (9) and (10) take advantage of the efficiency improvement of the parametric assumptions. The answer turns out to be no.

5.3.1 Moment condition (2) case

For the sake of simplicity, we consider the case of moment condition (2) first. As theorem 1 and theorem 2 show that the semiparametric efficiency bound is not affected by knowledge of the propensity score $p\left(X\right)$, it is no surprise that making a parametric assumption about $p\left(X;\gamma\right)$ does not change this semiparametric efficiency bound either. In addition, we already know from theorems 10 and 13 that both an optimally weighted CEP-GMM estimator and a nonparametric IPW-GMM estimator for β achieve this semiparametric efficiency bound when the true propensity score is not used.

In what follows, we show the interesting result that the parametric inverse probability weighting estimator using $p\left(X;\hat{\gamma}\right)$ is in fact less efficient than the one using a nonparametric estimate $\hat{p}\left(X\right)$ (but more efficient than the one using the known $p(X)$) in (10). For this purpose, first note that

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

has the following influence function representation

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

where

$$\begin{aligned}&\text{Proj}\left(\mathcal{E}\left(X_i;\beta\right)\tfrac{D_i-p(X_i)}{1-p(X_i)}\middle|S_\gamma\left(D_i,X_i\right)\right)\\ =\;&E\left[\mathcal{E}\left(X_i;\beta\right)\tfrac{D_i-p(X_i)}{1-p(X_i)}S_\gamma\left(D_i,X_i\right)\right]E\left[S_\gamma\left(D_i,X_i\right)S_\gamma\left(D_i,X_i\right)'\right]^{-1}S_\gamma\left(D_i,X_i\right),\end{aligned}$$

which is the influence function representation for the estimation error in $\hat{\gamma}$

$$\frac{\partial}{\partial \gamma} E\left[(1-D_i)\, m\left(Z_i;\beta\right) \frac{1}{1-p\left(X_i;\gamma\right)}\right] \sqrt{n}\left(\hat{\gamma}-\gamma\right).$$

Now we have the following relative efficiency comparison.

Theorem 14 *Suppose the parametric model $p\left(X_i;\gamma\right)$ is correctly specified and $E[S_\gamma(D,X)S_\gamma(D,X)']$ 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 $\hat{p}\left(X_i;\hat{\gamma}\right)$ is more efficient than the one using the known $p\left(X_i\right)$, but is less efficient than the one using a nonparametric estimate $\hat{p}\left(X_i\right)$ of the propensity score.*

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

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

and

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

To understand the first relation, note that the difference between the influence functions of the two sides under comparison is given by

$$\text{Proj}\left(\mathcal{E}\left(X_i;\beta_0\right)\frac{D_i-p\left(X_i\right)}{1-p\left(X_i\right)}\middle| S_\gamma\left(D_i,X_i\right)\right). \tag{13}$$

Now the influence function for $\frac{\sqrt{n}}{n_a}\sum_{j=1}^{n_a} m\left(Z_j;\beta_0\right)\frac{1-\hat{p}}{1-p(X_i)}$ can be rewritten as

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

where we have used $\text{Res}\left(Z_1|Z_2\right)$ to denote the residual of a projection of Z_1 into the linear space spanned by Z_2. It is clear that each of the above three terms are orthogonal to the projection in (13). Hence the first relation between the two variances holds.

A similar logic can be used to demonstrate the second relation. Note that the difference between the influence functions for the two terms under comparison is simply

$$\text{Res}\left(\mathcal{E}\left(X_i;\beta_0\right)\frac{D_i-p\left(X_i\right)}{1-p\left(X_i\right)}\middle| S_\gamma\left(D_i,X_i\right)\right).$$

This is orthogonal to the influence function for the second term with nonparametric $\hat{p}\left(X_i\right)$, which is

$$\frac{(1-D_i)}{1-p\left(X_i\right)}\left(m\left(Z_i;\beta_0\right)-\mathcal{E}\left(X_i;\beta_0\right)\right)+\mathcal{E}\left(X_i;\beta_0\right).$$

Hence the second relation also holds.

5.3.2 Moment condition (1) case

Now consider the more interesting case where moment condition (1) holds and sample moment condition (9) is used. 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 13 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

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

where

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

is the influence function from the first step estimation of γ. Using this influence function and the influence function for Theorem 3, it can be verified that the difference between them is given by

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

which is obviously orthogonal to the influence function of Theorem 3. Therefore, the two step parametric IPW estimator has a variance larger than the efficient variance 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

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

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

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

which is the efficient influence function under correct parametric specification of $p(X;\gamma)$ in Theorem 3.

It is also worth noting that assumption 1 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 1 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 $p(X)$.

6 Empirical Application

We illustrate our method empirically using a data set that matches self-reported earning from the Current Population Survey (CPS) to employer-reported social security earnings (SSR) from 1978 (the CPS/SSR Exact Match File). For individuals who accepted to report their Social Security Number, reported earnings were matched against social security records. Bound and Krueger (1991b) and Bollinger (1998) use this dataset to argue that the classical measurement error (CME) model is not appropriate for reporting errors in earnings. As in Chen, Hong, and Tamer (2003), our maintained assumption is that Social Security (SS) earnings are accurate, so that we treat observations for which SS reports are available as the auxiliary dataset. We analyze the effect of correcting for measurement error in the estimation of earnings quantiles, and in the estimation of the cumulative distribution function of earnings over a grid of points.

We use individual-specific observations for persons that worked either full time or part time for all or part of 1977. We include in the sample both men and women, regardless of their age. Because we want to use SSR as a validation dataset, we exclude observations for which the gap between SSR and reported earnings are likely to be due to reasons different from reporting error. For this reason, out of the 78,227 observations that compose the initial sample, we exclude 8,817 individuals whose earnings are reported to derive also from sources different from wages and salaries. We also drop 7,100 observations with zero SS Earnings in 1977. The remaining sample is composed of 62,296 observations, of which 38,759 are complete with SSR. Clearly, individuals who provided their SS number may be a selected group. However, the conditional independence assumption does not require equality of the *marginal* distribution of the proxy variables X in the primary and auxiliary sample. Sample selection would invalidate identification only if the *conditional* distribution of Y given X were different in the two datasets. Table 1 reports selected summary statistics for the sample. Demographic characteristics of the two groups appear to be overall similar, as well as their respective educational achievements. Individuals matched to their SS data are more likely to be married, and slightly older. However, these individuals appear to report higher earnings, a finding also reported by Bound and Krueger (1991b). Mean reported earnings are \$9,435 for matched workers, and \$7,927 for unmatched ones.

The results in Table 1 also show that mean and standard deviation of reported earnings are not very dissimilar from SS earnings if the statistics are calculated for earnings below the top-coding value for SS reports, that is, \$16,500. However, measurement error in earnings are not well represented by a classical measurement error model. Figure 1 shows histograms of $(100\times)$ the ratio between reported and SS earnings by quartile of Social Security Earnings.[4] This figure suggests that errors are less common for individuals with higher earnings. In the fourth quartile, approximately 40 percent of observations are clustered around 100, while the corresponding fraction decreases for lower quartiles, reaching below 30 percent for the first quartile.

Figure 2 shows that measurement error is negatively correlated with true earnings. The downward sloping line represents a nonparametric locally weighted regression (see Fan (1992)) on SS earnings of the difference between reported and SS earnings. The evidence presented in Figures 1 and 2 is consistent with the findings in Bound and Krueger (1991b), who conclude that measurement error in earnings is not independent of true earnings, so that methods that allow for non-classical errors are needed to recover consistent estimates of parameters of interest using the primary sample.

[4]In this figure, as well as in Figure 2, we only use observations that are not top coded.

We first consider the estimation of lower quantiles. The focus on low and middle income individuals allows us to avoid the complications that would arise from the fact that SS earnings are top-coded at \$16,500.[5] In fact, we calculated that top coding of SS earnings affected less than 1 percent of all reported earnings below \$8,000, a value that was *above* the median of reported earnings in 1977. Limiting the analysis to low and middle earners will not limit the interest of the results, as researchers interested in poverty analysis are naturally interested in the lower tail of the distribution of earnings (or income, or expenditure), and it is well known that poverty estimates can be severely distorted by the presence of measurement error (see, for example, Ravallion (1988)). Letting Y denote true earnings, and letting β_α indicate the α-quantile of the distribution of Y, it is easy to see that the parameter of interest is identified by the following moment condition:

$$E\left[1\left(Y < \beta_\alpha\right) - \alpha\right] = E\left[E\left[1\left(Y < \beta_\alpha\right) - \alpha \mid X, D = 0\right]\right] = 0,$$

where the first equality follows from the conditional independence assumption (1). The moment condition can also be written as

$$E\left[\left[1\left(Y < \beta_\alpha\right) - \alpha\right] \frac{(1-p)}{\left[1 - P\left(X\right)\right]} \mid D = 0\right] = 0. \tag{14}$$

Table 2 reports point estimates and standard errors for selected quantiles, using different estimators. The first two columns report the quantiles estimated using the unadjusted primary sample (column 1), and the auxiliary sample alone (column 2). Note that both these columns will contain biased estimates: the former because of measurement error, and the latter because the validation sample is not a random subset from the primary data set. Column 3 reports estimates obtained using two-step parametric IPW, where in the first step we estimate the propensity score using logit, and including only reported earnings as predictor, and then we estimate β_α with a grid search, using the sample analogue of (14). The results do not change when we estimate β_α that minimizes the appropriate objective function. Column (4) reports CEP-GMM estimates, calculated using 3^{rd} order polynomial splines in reported earnings as sieve basis, with 10 knots at the equal range quantiles of the empirical distribution of reported earnings. Finally, in column (5) we report estimates obtained using moment condition (14), but with a nonparametric first step where we estimate the propensity score using again logit, but including the basis functions we used for CEP-GMM as regressors. The nonparametric results appear to change only marginally if we change the degree of the polynomial, or the number of knots. Also, the results do not change substantially if we include other possible predictors such as education, age, marital status, or race in the estimation of the propensity score. The asymptotic variances are estimated as described in section 4. Notice that, in principle, top coding would require modelling right-censoring explicitly. However, right censoring at a point c is of no consequence, as long as we focus, as we do here, on lower quantiles (so that $\min\left(\beta_\alpha, c\right) = \beta_\alpha$).

The adjusted estimates suggest that measurement error leads to an underestimation of earnings in the lower tail of the distribution (up to the 15^{th} percentile of the distribution), while unadjusted quantiles that are closer to the middle of the distribution appear to be biased upwards. Nonparametric inverse probability weighting and CEP-GMM lead to almost identical point estimates. Consistently with the theoretical results in the paper, the nonparametric estimators appear to be more efficient than the parametric one. Up to

[5] There is also a small number of observations for which *reported* earnings are top-coded, but these account for less than 1% of the sample, and we ignore them.

the 20^{th} percentile, the standard errors for the parametric IPW estimator are 50% or more larger than the corresponding figures for the nonparametric estimators. The results in column (3) also signal that the parametric IPW is likely to be misspecified, as the point estimates remain very close to those from the auxiliary sample alone. This is to be expected when the propensity score is estimated using regressors that do not predict adequately whether the observation has been validated, so that the reweighting function in (14) remains always very close to one.

We turn next to the estimation of the cumulative distribution function (cdf) of earnings, evaluated over a grid of points. The results are displayed in Table 3. Again, we focus on low earnings, and we use estimators analogous to those used to construct Table 2, ordered in the same way. Not surprisingly, the results confirm the finding described in the previous table. The two nonparametric estimators lead to almost indistinguishable results, while parametric IPW leads to standard errors that are generally larger, when contrasted with the corresponding point estimates, which are always smaller. As in Table 2, a probably misspecified parametric IPW estimator does not succeed in moving the distribution of the validation sample towards the primary sample. The point estimates in columns (4) and (5) suggest that the use of reported earnings does not lead to a large bias in the estimation of the cdf for earnings below \$2,000, while it leads to an *under*estimation of 2 to 3 percentage points of the CDF for earnings between two and five thousand dollars. For very low earnings (below \$1,000), both IPW and CEP suggest a small but *positive* bias of the cdf estimated using the primary sample.

[Figure 1 about here.]

[Table 1 about here.]

7 Conclusions

We derive semiparametric efficiency variance bounds for the estimation of parameters defined through general nonlinear, possibly non-smooth and over-identified moment conditions, where the sampling information consists of a primary sample and an auxiliary sample. We distinguish the "verify-out-of-sample" case and the "verify-in-sample" case, which are closely related to the estimation of the average treatment effect on the treated and the average treatment effect in the program evaluation literature. These two cases also have natural interpretations in non-classical measurement error models and missing data models. Our efficiency bounds are derived for both cases.

The semiparametric CEP-GMM estimators that we develop make use of a sieve based nonparametric estimate of the conditional expectation of the general moment condition given variables that are observable in both the primary sample and the auxiliary sample. We show that the optimally weighted CEP-GMM estimators achieve the semiparametric efficiency bounds when the propensity score is unknown, or known or belongs to a correctly specified parametric family. These estimators only use one nonparametric estimate and do not require nonparametric estimations of both the conditional expectation of the moment functions and the propensity score. They also require weaker regularity conditions than the existing ones in the literature. They allow for unbounded support of conditional variables and non-smooth moment conditions. Our results are applicable to a wide variety of models, including non-classical measurement error models, missing data models and treatment effect models. Our results may also suggest useful guidance to the design of survey data sets, which generates the crucial data input for the analysis of econometric models.

While we have constructed efficient semiparametric estimators in this paper, their efficiency are proved by comparing their asymptotic variance with the calculated semiparametric efficiency bound. While beyond the scope of this paper, an interesting extension will be to consider estimators based on nonparametric maximum likelihood principles that can be proved to achieve the semiparametric efficiency variance bound without knowledge of its particular form. Recently, for estimation of the average treatment effect parameter or/and of the mean parameter in missing data models, Wang, Linton, and Hardle (2004) suggested that 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. It will also be interesting to study the efficiency implications of these semiparametric restrictions on the propensity score.

Appendix A: Calculation of Efficiency Bounds

Proof. Theorem 1

We follow closely the structure of efficiency derivation of Hahn (1998) and Newey (1990b). The proof proceeds as follows: first, we characterize the tangent set $\mathcal{T}$ for all (regular) parametric submodels satisfying the semiparametric assumptions. Then, we conjecture a form for the efficient influence function, proving pathwise differentiability of the parameter β and verifying that the efficient influence function lies in the tangent set. Then an application of Theorem 3.1 in Newey (1990b) concludes that the semiparametric efficiency bound is the expectation of the outer product of the efficient influence function.

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

$$f_\theta(y,x,d) = p_\theta^d (1-p_\theta)^{1-d} f_\theta(x|D=1)^d f_\theta(x|D=0)^{1-d} f(y|x)^{1-d}. \tag{15}$$

The resulting score function is given by

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

where

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

The tangent space of this model is therefore given by:

$$\mathcal{T} = a(d-p_\theta) + (1-d)\, s_\theta(x|D=0) + (1-d)\, s_\theta(y|x) + d s_\theta(x|D=1), \tag{16}$$

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

$$\frac{\partial\beta(\theta)}{\partial\theta} = -\left(\mathcal{J}_\beta^1\right)^{-1} E\left[m(Z;\beta)\frac{\partial \log f_\theta(Y,X \mid D=1)}{\partial\theta'} \mid D=1\right]. \tag{17}$$

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

$$E\left[m(Z;\beta)\, s_\theta(Y \mid X)' \mid D=1\right] + E\left[m(Z;\beta)\, s_\theta(X \mid D=1)' \mid D=1\right] \tag{18}$$

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

$$\partial\beta(\theta)/\partial\theta = E\left[\Psi^1(Y,X,D)\, S_\theta(Y,X,D)'\right] \tag{19}$$

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 such property is satisfied by choosing:

$$\Psi^1(Y,X,D) = -\left(\mathcal{J}_\beta^1\right)^{-1} F_\beta^1(Y,X,D)$$

where

$$F_\beta^1(Y,X,D) = \frac{1-D}{p}\frac{p(X)}{1-p(X)}\left[m(Z;\beta) - \mathcal{E}(X)\right] + \frac{\mathcal{E}(X)}{p}D \tag{20}$$

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

$$E\left[m(Z;\beta)\frac{\partial}{\partial\theta'}\log f_\theta(Y,X \mid D=1) \mid D=1\right] = E\left[F_\beta^1(Y,X,D)\,S_\theta(Y,X,D)'\right]. \tag{21}$$

This can in turn be verified by checking that both

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

and

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

Now one can also verify that $F_\beta^1(Y,X,D)$ belongs to the tangent space $\mathcal{T}$ stated in equation (16), with the first term of $F_\beta^1(Y,X,D)$ taking the role of $(1-d)\,s_\theta(y|x)$ and the second term of $F_\beta^1(Y,X,D)$ taking the role of $ds_\theta(X|D=1)$, and the two other components in (16) being identically equal to 0.

Therefore all the conditions of Theorem 3.1 in Newey (1990b) hold, and the efficiency bound for regular estimators of the parameter β is given by

$$\begin{aligned} V_1 &= \left(\mathcal{J}_\beta^1\right)^{-1} E\left[F_\beta^1(Y,X,D)\,F_\beta^1(Y,X,D)'\right]\left(\mathcal{J}_\beta^1\right)'^{-1} \\ &= \left(\mathcal{J}_\beta^1\right)^{-1} E\left[\frac{p(X)^2}{p^2(1-p(X))}V(m(Z;\beta) \mid X) + \frac{p(X)}{p^2}\mathcal{E}(X)\,\mathcal{E}(X)'\right]\left(\mathcal{J}_\beta^1\right)'^{-1}. \end{aligned} \tag{22}$$

Case (2). For this case we use an alternative factorization of the likelihood function. Consider the parametric path θ of the joint distribution of Y, D and X defined in Hahn (1998). Define $p_\theta(x) = P_\theta(D=1 \mid x)$. The joint density function for Y, D and X is given by

$$f_\theta(y,x,d) = f_\theta(x)\,p_\theta(x)^d\left[1-p_\theta(x)\right]^{1-d} f_\theta(y \mid x)^{1-d}. \tag{23}$$

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

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

The tangent space of this model is therefore given by:

$$\mathcal{T} = \left\{(1-d)\,s_\theta(y \mid x) + a(x)\left(d-p_\theta(x)\right) + t_\theta(x)\right\} \tag{24}$$

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.

The results for the efficiency bound for the estimation of parameters β defined by the *un*conditional moment condition (2) follow using arguments parallel to those described above. In case (2), equation (17) is replaced by:

$$\begin{aligned} \frac{\partial\beta(\theta)}{\partial\theta} &= -\left(\mathcal{J}_\beta^2\right)^{-1} E\left[m(Z;\beta)\frac{\partial\log f_\theta(Y,X)}{\partial\theta'}\right]. \\ &= -\left(\mathcal{J}_\beta^2\right)^{-1}\left\{E\left[m(Z;\beta)\,s_\theta(Y \mid X)'\right] + E\left[\mathcal{E}(X)\,t_\theta(X)'\right]\right\}. \end{aligned} \tag{25}$$

Now we replace $F_\beta^1(Y,X,D)$ in (20) with the following:

$$F_\beta^2(Y,X,D) = \frac{1-D}{1-p(X)}\left[m(Z;\beta) - \mathcal{E}(X)\right] + \mathcal{E}(X) \tag{26}$$

and then it can be shown that

$$E\left[F_\beta^2(Y,X,D)\,S_\theta(Y,X,D)'\right] = E\left[m(Z;\beta)\frac{\partial\log f_\theta(Y,X)}{\partial\theta'}\right],$$

which follows from

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

Then the efficient influence function for case (2) is equal to $-\left(\mathcal{J}_{\beta}^{2}\right)^{-1}F_{\beta}^{2}\left(Y,X,D\right)$ with the two terms being orthogonal to each other. So that the asymptotic variance bound is equal to

$$\begin{aligned}V_{2} &= \left(\mathcal{J}_{\beta}^{2}\right)^{-1}E\left[F_{\beta}^{2}\left(Y,X,D\right)F_{\beta}^{2}\left(Y,X,D\right)'\right]\left(\mathcal{J}_{\beta}^{2}\right)'^{-1}\\ &= \left(\mathcal{J}_{\beta}^{2}\right)^{-1}E\left[\frac{1}{1-p(X)}Var\left(m\left(Z;\beta\right)\mid X\right)+\mathcal{E}\left(X\right)\mathcal{E}\left(X\right)'\right]\left(\mathcal{J}_{\beta}^{2}\right)'^{-1}.\end{aligned}$$

For both moment conditions (1) and (2), the case of overidentification is a straightforward extension of the results for exact identification described above. Here we only consider the case of moment condition (1). The case of moment condition (2) can be derived following an analogous argument. 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

$$\mathcal{A}E\left[m\left(Z;\beta\right)\mid D=1\right]=0.$$

Differentiating under the integral again, we have

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

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

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

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

$$-\left[\mathcal{A}\mathcal{J}_{\beta}^{1}\right]^{-1}F_{\beta}^{1}\left(y,x,d\right).$$

The asymptotic variance corresponding to this efficient influence function for fixed $\mathcal{A}$ is therefore

$$\left[\mathcal{A}\mathcal{J}_{\beta}^{1}\right]^{-1}\mathcal{A}\Omega\mathcal{A}'\left[\mathcal{J}_{\beta}^{1\prime}\mathcal{A}'\right]^{-1} \tag{27}$$

where

$$\Omega=E\left[F_{\beta}^{1}\left(Y,X,D\right)F_{\beta}^{1}\left(Y,X,D\right)'\right]$$

as calculated above. Therefore, the efficient influence function is obtained when $\mathcal{A}$ minimizes (27). 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

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

■

Proof. Theorem 2

As for Theorem 1, it suffices to present the proof for the case of exact identification. The overidentified case follows from choosing the optimal linear combination matrix. If the propensity score $p\left(x\right)$ is known, the score becomes (c.f. Hahn (1998))

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

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 can then be established as in Theorem 1, verifying that equation (19) holds, with

$$F_\beta^1(y,x,d) = \frac{1-d}{p}\frac{p(x)}{1-p(x)}\left(m(z;\beta) - \mathcal{E}(x)\right) + \frac{\mathcal{E}(x)}{p} p(x). \tag{28}$$

Then the efficient influence function is as before equal to $-\left(\mathcal{J}_\beta^1\right)^{-1} F_\beta^1(y,x,d)$, and using Theorem 3.1 of Newey (1990b), the semiparametric efficiency bound can be calculated as:

$$V_1 = \left(\mathcal{J}_\beta^1\right)^{-1}\left\{E\left[\frac{p(X)^2}{p^2(1-p(X))} V\left(m(Z;\beta) \mid X\right)\right] + E\left[\frac{\mathcal{E}(X)\,\mathcal{E}(X)'}{p^2} p(X)^2\right]\right\}\left(\mathcal{J}_\beta^1\right)^{\prime-1}. \tag{29}$$

A comparison of (22) and (29) shows that the bound is reduced when the propensity score is known.

The bound for case (2) is obtained using analogous arguments. Since $p(x)$ does not enter the definition of β in case (2), the efficient influence function is still $-\left(\mathcal{J}_\beta^2\right)^{-1} F_\beta^2(Y,X,D)$, as in Theorem 1, where

$$F_\beta^2(Y,X,D) = \frac{1-D}{1-p(X)}\left(m(Z;\beta) - \mathcal{E}(X)\right) + \mathcal{E}(X). \tag{30}$$

and the efficiency bound is the same as in Theorem 1:

$$\left(\mathcal{J}_\beta^2\right)^{-1} E\left[F_\beta^2(Y,X,D) F_\beta^2(Y,X,D)'\right]\left(\mathcal{J}_\beta^2\right)^{\prime-1}.$$

The proof for the case of overidentification proceeds as in Theorem 1.

■

Proof. Theorem 3

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

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

The tangent space is therefore

$$\mathcal{T} = \{(1-d)\, s_\theta(y \mid x) + cS_\gamma(d;x) + t_\theta(x)\}$$

where c is a finite vector of constants and

$$S_\gamma(d;x) = \frac{d - p(x)}{p(x)(1-p(x))}\frac{\partial p(x;\gamma)}{\partial \gamma}$$

is the parametric score function. Now define $F_\beta^1(Y,X,D)$ as

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

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

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

The second term in the curly bracket 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

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

In particular,

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

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. ■

Proof. Theorem 4

All the propositions in Theorems 4 and 5 can be proved following the same strategy as in Theorems 1 to 3, so here we only sketch the argument, specifying the score function, the tangent set, and the resulting efficient influence function for the case of exact identification. The score function will depend on which parts of the likelihood are known to the econometrician.

If both p and $f(x \mid D=0)$ are known, the score is

$$S_\theta(d,y,x) = (1-d)s_\theta(y \mid x) + ds_\theta(x \mid D=1)$$

so that the tangent space can be characterized as

$$\mathcal{T} = \{(1-d)\,s_\theta(y \mid x) + ds_\theta(x \mid D=1)\} \tag{31}$$

where $\int s_\theta(y \mid x)\,f_\theta(y \mid x)\,dy = 0$, and $\int s_\theta(x \mid D=1)\,f_\theta(x \mid D=1)\,dx = 0$.

Case (1). Since neither p nor $f(x|D=0)$ enters the definition of β in case (1), the efficient influence function is still $-\mathcal{J}_\beta^{-1}F_\beta^1(Y,X,D)$ as in Theorem 1 and the efficient variance bound is also the same as in Theorem 1.

Case (2). When both p and $f(x \mid D=0)$ are known, the pathwise derivative of β can be written as follows:

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

The following choice of $F_\beta^2(Y,X,D)$:

$$F_\beta^2(Y,X,D) = \frac{1-D}{1-p(X)}\left[m(Z;\beta) - \mathcal{E}(X)\right] + D\left[\mathcal{E}(X) - E\left(\mathcal{E}(X) \mid D=1\right)\right] \tag{32}$$

can be easily verified to belong to the tangent set (31) and satisfy

$$\frac{\partial\beta(\theta)}{\partial\theta} = -\left(\mathcal{J}_\beta^2\right)^{-1}E\left[F_\beta^2(Y,X,D)\,\frac{\partial\log f_\theta(Y,X)}{\partial\theta'}\right].$$

Then, the bound can be calculated as usual:

$$\begin{aligned}
\tilde{\Omega}_\beta^2 &= E\left[F_\beta^2(Y,X,D)\,F_\beta^2(Y,X,D)'\right] \\
&= E\left[\frac{1}{1-p(X)}Var\left(m(Z;\beta) \mid X\right)\right] + pVar\left(\mathcal{E}(X) \mid D=1\right) \\
&= E\left[\frac{1}{1-p(X)}Var\left(m(Z;\beta) \mid X\right) + p(X)\mathcal{E}(X)\,\mathcal{E}(X)'\right]
\end{aligned}$$

Which is different from Ω_β^2 in Theorem 1. ■

Proof. Theorem 5

Case (1), with $f(x \mid D=1)$ known.

By Theorem 3, $f(x \mid D=0)$ and p are ancillary for the estimation of β. Then the score function can be written as

$$S_\theta(d,y,x) = (1-d)\,s_\theta(y \mid x)$$

The efficient influence function is

$$\Psi_\beta^1(Y,X,D) = -\left(\mathcal{J}_\beta^1\right)^{-1}\left[\frac{1-D}{p}\,\frac{p(X)}{1-p(X)}\left(m(Z;\beta) - \mathcal{E}(X)\right)\right].$$

Case (1), with $f(y \mid x)$ known.

Using again the ancillarity of $f(x \mid D=0)$ and p, the relevant score can be written as

$$S_\theta(d,y,x) = ds_\theta(x|D=1),$$

and the efficient influence function becomes

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

Case (2), with $f(x)$ known.

By Theorem 2, the propensity score is ancillary for the estimation of the parameter β. The likelihood score can then be written as

$$S_{\theta}(d,y,x) = (1-d)\, s_{\theta}(y \mid x),$$

and the efficient influence function is

$$\Psi_{\beta}^{2}(Y,X,D) = -\left(\mathcal{J}_{\beta}^{2}\right)^{-1}\left[\frac{1-D}{1-p(X)}\left(m(Z;\beta) - \mathcal{E}(X)\right)\right].$$

Case (2), with $f(y \mid x)$ known.

Now the relevant score is $S_{\theta}(d,y,x) = t(x)$ and the efficient influence function becomes

$$\Psi_{\beta}^{2}(Y,X,D) = -\left(\mathcal{J}_{\beta}^{2}\right)^{-1}\mathcal{E}(X).$$

■

Proof. Theorem 6

If $p_{\theta}(x) = p$, the score becomes

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

so that the tangent set is:

$$\mathcal{T} = \left\{(1-d)\, s_{\theta}(y \mid x) + a(d-p_{\theta}) + t_{\theta}(x)\right\}, \tag{33}$$

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 is any real number. For both case (1) and (2), assumption 3 ensures that pathwise differentiation of β leads to

$$\begin{aligned}\frac{\partial\beta(\theta)}{\partial\theta} &= -(\mathcal{J}_{\beta})^{-1} E\left[m(Z;\beta)\frac{\partial \log f_{\theta}(Y,X)}{\partial\theta'}\right] \\ &= -(\mathcal{J}_{\beta})^{-1}\left\{E\left[m(Z;\beta)\, s(Y \mid X)'\right] + E\left[\mathcal{E}(X)\, t(X)'\right]\right\},\end{aligned}$$

where

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

It is easy to verify that the following function is the efficient influence function function as it satisfies equation (19),

$$\Psi_{\beta}(Y,X,D) = -(\mathcal{J}_{\beta})^{-1}\left\{\frac{1-D}{1-p}\left[m(Z;\beta) - \mathcal{E}(X)\right] + \mathcal{E}(X)\right\}.$$

Then the efficiency bound is calculated as usual as $E\left[\Psi_{\beta}(Y,X,D)\,\Psi_{\beta}(Y,X,D)'\right]$. ■

Proof. Theorem 8

Case (1). The likelihood of a parametric submodel takes the form

$$p_{\theta}^{d} f_{\theta}(x|D=1)^{d} f_{\theta}(y_1|x)^{d}(1-p_{\theta})^{d} f_{\theta}(x|D=0)^{1-d} f_{\theta}(y_0|x)^{1-d}.$$

The corresponding score is given by

$$S_{\theta}(d,y,x) = \frac{d-p_{\theta}}{p_{\theta}(1-p_{\theta})}\cdot p_{\theta} + d s_{\theta}(y_1|x) + d s_{\theta}(x|D=1) + (1-d)\, s_{\theta}(y_0|x) + (1-d)\, s_{\theta}(x|D=0).$$

The tangent space of this model is therefore given by:

$$\mathcal{T} = a(d-p_{\theta}) + (1-d)\, s_{\theta}(x|D=0) + (1-d)\, s_{\theta}(y_0|x) + d s_{\theta}(x|D=1) + d s_{\theta}(y_1|x).$$

where $\int s_{\theta}(y_i \mid x) f_{\theta}(y_i \mid x)\, dy = 0$ for $i = 1, 2$, $\int s_{\theta}(x|d) f_{\theta}(x|d)\, dx = 0$, and a is a finite constant.

Pathwise differentiation shows that

$$\begin{aligned}\frac{\partial\beta(\theta)}{\partial\theta} &= -\left(\mathcal{J}_\beta^1\right)^{-1} E\left[\left(m(Y_1;\beta)-m(Y_0;\beta)\right)\frac{\partial \log f_\theta(Y_1,Y_0,X|D=1)}{\partial\theta} \mid D=1\right] \\ &= -\left(\mathcal{J}_\beta^1\right)^{-1}\left[E\left[m(Y_1;\beta)\, s_\theta(Y_1|D=1)\,|D=1\right]-E\left[m(Y_0;\beta)\, s_\theta(Y_0,X|D=1)\,|D=1\right]\right].\end{aligned}$$

The second component of this expression can now be calculated as equal to

$$E\left[F_\beta^1(Y_1,Y_0,X,D)\, s_\theta(Y_1,Y_0,X,D)\right]$$

where the efficient influence function is

$$F_\beta^1(Y_1,Y_0,X,D)=\frac{D}{p}\left(m_1(Y_1;\beta)\right)-\frac{1-D}{p}\frac{p(X)}{1-p(X)}\left(m_0(Y_0;\beta)-\mathcal{E}_0(X)\right)-D\frac{\mathcal{E}_0(x)}{p},$$

which can also be rewritten as a sum of orthogonal terms:

$$\begin{aligned}F_\beta^1(Y,X,D) &= \frac{D}{p}\left(m_1(Y_1;\beta)-\mathcal{E}_1(X)\right)-\frac{1-D}{p}\frac{p(X)}{1-p(X)}\left(m_0(Y_0;\beta)-\mathcal{E}_0(X)\right) \\ &\quad +\frac{\left(\mathcal{E}_1(x)-\mathcal{E}_0(x)\right)}{p}D\end{aligned}$$

Then the semiparametric efficiency bound can be calculated as :

$$E\left[\Psi_\beta^1(Y,X,D)\,\Psi_\beta^1(Y,X,D)'\right]=\Omega_\beta^1=E\left[\frac{p(X)}{p}V_1(X)+\frac{p(X)^2}{p^2(1-p(X))}V_0(X)+\frac{p(X)}{p^2}\mathcal{E}(X;\beta)\,\mathcal{E}(X;\beta)'\right].$$

Case (2). The likelihood of a parametric submodel takes the form

$$f_\theta(x)\left[f_\theta(y_1\mid x)\,p_\theta(x)\right]^d\left[f_\theta(y_0\mid x)\left(1-p_\theta(x)\right)\right]^{1-d}. \tag{34}$$

The pathwise differentiation leads to

$$\frac{\partial\beta(\theta)}{\partial\theta}=-\left(\mathcal{J}_\beta^2\right)^{-1}E\left[m_1(Y_1;\beta)\, s_\theta(Y_1,X)-m_0(Y_0;\beta)\, s_\theta(Y_0,X)\right],$$

and the second expectation can be written as

$$E\left[F_\beta^2(Y_1,Y_0,X,D)\, S_\theta(Y_1,Y_0,X,D)'\right],$$

where the efficient influence function is

$$\begin{aligned}F_\beta^2(Y_1,Y_0,X,D) &= \frac{D}{p(X)}\left(m_1(Y_1;\beta)-\mathcal{E}_1(X)\right)-\frac{1-D}{1-p(X)}\left(m_0(Y_0;\beta)-\mathcal{E}_0(X)\right) \\ &\quad +\left(\mathcal{E}_1(x)-\mathcal{E}_0(x)\right).\end{aligned}$$

The variance of the efficient influence function $F_\beta^2(Y,X,D)$ is Ω_β^2 given in Theorem 8.

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 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 $c_1||h||_2\le||h||_{2,a}\le c_2||h||_2$, which is true under the assumption $0<\underline{p}\le p_o(x)\le\overline{p}<1$.

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

Proposition B.1: *Under Assumptions 6.1, 6.2 and 6.5, 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 7.1 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 6.5). Moreover, by Assumptions 6.1 and 6.2 we have the following results: (1) the parameter space is compact under the norm $||\cdot||_{\infty,\omega}$ for $\omega > 0$, see Chen, Hansen, and Scheinkman (1997) or Ai and Chen (2003); (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 for any $p()$, $\widetilde{p}() \in \mathcal{H}$,

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

Now $E[(1 + X_i'X_i)^{\frac{\omega}{2}}] < \infty$ by assumption 6.2.

Hence by either Theorem 0 in Gallant and Nychka (1987) or Lemma 2.9 and theorem 2.1 in Newey (1994), $||\widehat{p}(\bullet) - p_o(\bullet)||_{\infty,\omega} = o_p(1)$. Now

$$\begin{aligned} ||\widehat{p}(\bullet) - p_o(\bullet)||_2 &= \sqrt{\int [\widehat{p}(x) - p_o(x)]^2 f_X(x)dx} \\ &\leq \sqrt{(||\widehat{p}(\bullet) - p_o(\bullet)||_{\infty,\omega})^2 \int (1 + x'x)^\omega f_X(x)dx} \\ &= o_p(1) \quad \text{(by assumption 6.2).} \end{aligned}$$

(ii) We can obtain the convergence rate of $||\widehat{p}(\bullet) - p_o(\bullet)||_2$ by applying theorem 1 in Chen and Shen (1998). 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 Chen and Shen (1998) theorem 1 are satisfied given our Assumptions 6.1 and 6.2. 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 7.1, 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 Chen, Hansen, and Scheinkman (1997) or Ai and Chen (2003). Hence by Assumption 7.1 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 12**): 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 Chen, Linton, and van Keilegom (2003)(hereafter CLK) with their θ being our β and their h being our $p()$. (Note that if $m(Z_i, \beta)$ is pointwise smooth in β, then Theorem 12 also follows from Newey (1994) (Lemma 5.2)). 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, 2.1 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 6.1, 6.2 and 6.4, 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 6.1 - 6.4 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, 6 and 7, 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**): Although one can apply the approach in Newey (1994) to establish this result, here we follow the approach in Chen and Shen (1998) and Ai and Chen (2003). 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 v^* for functional $E\left\{ \mathcal{E}(X,\beta_o)\frac{p(X)-p_o(X)}{1-p_o(X)} \right\}$ is simply given by

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

this is because

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

and

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

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

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

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

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 6.1 and 7.5. **(B2.2)** is implied by Assumption 7.5 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, Assumptions 6.1 and 7.5. 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 7.1, $\log N_{[]}(\delta, \mathcal{F}_n, ||\cdot||_2) \leq const.(\frac{c}{\delta})^{d_x/\gamma}$ for any $\delta > 0$. Applying theorem 3 in Chen and Shen (1998) 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 13**): 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), again one could also apply Lemma 5.3 in Newey (1994) when

$m(Z_i,\beta)$ is pointwise smooth in β. Given the definition of β_0 and Theorem 12, 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 2.1.

Chen (2003) points out 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

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

and the last equality is due to

$$\frac{f_{X|D=0}(X)}{f_X(X)}=\frac{1-p_o(X)}{1-p}.$$

Before we apply assumptions 7.6a or 7.6b or 7.6c 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 6.2, 7.4, Proposition B.1(i) and Theorem 12,

$$|\Gamma_2(\beta,p_o)[p()-p_o()]-\Gamma_2(\beta_o,p_o)[p()-p_o()]|\leq const.||\beta-\beta_0||\times||p()-p_o()||_{\infty,\omega}$$

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 12 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 $\delta\in(0,\delta_0]$,

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

where the last inequality is due to our assumptions 7.2, 7.3 and Proposition B.1(i). In the following we let $N(\varepsilon,\Lambda^\gamma_c(\mathcal{X}),||\cdot||_{\infty,\omega})$ denote the $||\cdot||_{\infty,\omega}-$covering number of $\Lambda^\gamma_c(\mathcal{X})$ [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 7.1 implies

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

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

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

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

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

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

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

thus we obtain Theorem 13 after we establish condition **(*)**.

We now apply assumption 7.6a or 7.6b or 7.6c to verify condition **(*)**. Since

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

we have under assumption 6.1,

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

If assumption 7.6a holds, then

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

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 7.6b 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 7.6c 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. ■

Appendix C: Nonlinear Regressions

Our semiparametric efficiency bound calculation and semiparametric CEP and IPW estimation methods can be extended to nonlinear regression models. In regression models, either the dependent variable or the independent variables can be missing or measured with error. Previously Robins, Rotnitzky, and Zhao (1994) developed a unified framework for efficient bound calculation with missing regressors which corresponds to the case of moment condition (2). Their efficient instrument function does not have explicit close form solutions in general, except in the special cases they considered, which include normal errors and discrete regressors. In the following we give several new results. The first set of results are concerned with the case of measurement errors in dependent variables only. The efficiency bound for moment condition (C.2) was derived in Robins and Rotnitzky (1995) and Rotnitzky and Robins (1995). Our result for moment condition (C.1) is new to our knowledge. For completeness and comparison we present them both below. The second set of results, which are new to our knowledge, are concerned with measurement errors in regressors only, where the propensity score depends only on the observed reported regressors.

We will only present the explicit forms of semiparametric efficiency bounds. Semiparametric estimators that achieve these efficiency bounds using the CEP approach can be derived and proved along the lines of section 4, by either nonparametrically estimating the efficient instruments in a one-step updating procedure (Newey (1990a)) or by invoking the semiparametric minimum distance principle of Ai and Chen (2003). The CEP semiparametric estimation principle achieves the orthogonalization of the influence functions.

We only state the efficiency bound results under the conditions of theorem 1. We omit presenting analogs for theorems 2 and 3 due to space limitation.

Measurement errors in dependent variables

Consider the nonlinear regression model under assumption 1,

$$Y = g(X_1, \beta) + \epsilon,$$

such that the mean of ϵ given X_1 is zero in the relevant sample. In other words, the dependent variables Y can be either missing or measured with errors. In this case $Z = (Y, X_1)$ and $X = (X_1, X_2)$ where X_2 contains either the reported value or a proxy for Y. Moment conditions (1) and (2) are now restated as, respectively,

$$E[(Y - g(X_1, \beta))|X_1, D = 1] = 0, \tag{C.1}$$

and

$$E[(Y - g(X_1, \beta))|X_1] = 0. \tag{C.2}$$

These conditional moments can be translated into unconditional ones by specifying that for any measurable function $A(X_1)$,

$$E[A(X_1)(Y - g(X_1; \beta))|D = 1] = 0, \quad \text{and} \quad E[A(X_1)(Y - g(X_1; \beta))] = 0.$$

Consider the conditional analog of theorem 1. For given instrumental functions $A(X_1)$, the efficient influence functions (20) and (26) can be factorized into

$$A(X_1)\left[\frac{1-D}{p}\frac{p(X)}{1-p(X)}[Y - E(Y|X)] + \frac{E(Y|X) - g(X_1, \beta)}{p}D\right]$$

for the analog of (20). The analog for (26) is

$$A(X_1)\left[\frac{1-D}{1-p(X)}[Y - E(Y|X_1)] + E(Y|X) - g(X_1; \beta)\right].$$

Then following Newey and McFadden (1994), the efficient instrument functions can be found by simplifying the sandwich form of the limiting variance matrix, which results in the following optimal $A^1(X_1)$ for moment condition (C.1) and $A^2(X_1)$ for moment condition (C.2):

$$\begin{aligned} A^1(X_1) = &\quad \tfrac{\partial}{\partial\beta}g(X_1; \beta)\, Var\left(\tfrac{1-D}{p}\tfrac{p(X)}{1-p(X)}[Y - E(Y|X)] + \tfrac{E(Y|X)-g(X_1,\beta)}{p}D|X_1\right)^{-1}, \\ A^2(X_1) = &\quad \tfrac{\partial}{\partial\beta}g(X_1; \beta)\, Var\left(\tfrac{1-D}{1-p(X)}[Y - E(Y|X_1)] + \mathcal{E}(Y|X) - g(X_1; \beta)|X_1\right)^{-1}. \end{aligned}$$

These efficient instruments can be nonparametrically estimated. Efficient estimators for β can be derived in analogy with section 4 by replacing $m(Z; \beta)$ by $\hat{A}(X_1)(y - g(X_1; \beta))$. Under suitable regularity conditions, the variation from nonparametric estimation of the efficient instrument function will not contribute to the variation in $\hat{\beta}$.

Measurement errors in regressors

Now consider the case of measurement errors in regressors where the measurement errors are not informative of the dependent variables given knowledge of the true regressors. Semiparametric estimators have been proposed by Carroll and Wand (1991b), Lee and Sepanski (1995) among others, but the issue of efficiency was not previously addressed. Let

$$W = g(Y; \beta) + \epsilon,$$

where either

$$E(\epsilon|Y, X, D = 1) = E(\epsilon|Y, D = 1) = 0 \iff E(W|Y, D = 1) = E(W|Y, X, D = 1) = g(Y; \beta), \tag{D.1}$$

or

$$E(\epsilon|Y, X) = E(\epsilon|Y) = 0 \iff E(W|Y) = E(W|Y, X) = g(Y; \beta). \tag{D.2}$$

In other words, X is a reported value or a proxy variable for Y and does not contain more information about the mean of W beyond that contained in Y. Assumption 1 is assumed to continue to hold. These can be restated as for any measurable function $A(X)$, either

$$E[A(X)(W - g(Y; \beta))|D = 1] = 0, \quad \text{(D.1)} \quad \text{or} \quad E[A(X)(W - g(Y; \beta))] = 0. \quad \text{(D.2)}$$

Under condition (D.1), the likelihood of observed data can be factorized into

$$p^d (1-p)^d f(x|d=0)^{1-d} f(y|x)^{1-d} f(x|d=1)^d f(w|x,d=1)^d,$$

where $Y \perp D|X$, and the tangent space takes the form of

$$\frac{d-p}{p(1-p)} p_\theta + (1-d) s(x|d=0) + (1-d) s(y|x) + ds(x|d=1) + ds(w|x,d=1)$$

Next define $h(x;\beta) = E(g(Y;\beta)|X=x)$. It can be verified that for given $A(X)$, the efficient influence function resulting from the projection into this tangent space is given by

$$\frac{D}{p} A(X)(W - h(X;\beta)) + \frac{1-D}{p} \frac{p(X)}{1-p(X)} A(X)[h(X;\beta) - g(Y;\beta)].$$

The two terms above are the projections into the $ds(w|x,d=1)$ component and the $(1-d)s(y|x)$ component of the tangent space. Then following Newey and McFadden (1994), the efficient instrument function $A^1(X)$ is found to be

$$A^1(X) = \frac{\partial}{\partial\beta} h(X;\beta) Var\left(\frac{D}{p}(W - h(X;\beta)) + \frac{1-D}{p}\frac{p(X)}{1-p(X)}[h(X;\beta) - g(Y;\beta)] \middle| X\right)^{-1}.$$

The variance of an estimator $\hat{\beta}$ using $\hat{A}^1(X)$ to form the moment conditions will achieve the semiparametric efficiency bound.

In the case of condition (D.2), for a given choice of the instrument functions, we look for an efficient influence function $F(w,y,x)$ that lies in the tangent space and satisfies the relation

$$EA(X)(W - g(Y;\beta)) s(W,Y,X) = EF(W,Y,X) s(W,Y,X).$$

First of all, we can write the left hand side as

$$EA(X)(W - h(X,\beta)) s(W,X) + EA(X)(h(X,\beta) - g(Y,\beta)) s(Y,X).$$

Next note that the likelihood function can be factorized as either

$$f(X) p(X)^d (1-p(X))^{1-d} f(Y|X)^{1-d} f(W|Y,X,D=0)^{1-d} f(W|X,D=1)^d,$$

or as

$$f(X) f(W|X) p(W,X)^d (1-p(W,X))^{1-d} f(Y|W,X,D=0)^{1-d}.$$

Therefore the tangent space can be represented either as

$$\tau_1 = \left\{ s(X) + \frac{d-p(X)}{p(X)(1-p(X))}\dot{p}(X) + ds(W\ |X,D=1) + (1-d)s(Y|X) + (1-d)s(W|Y,X,D=0) \right\},$$

or as

$$\tau_2 = \left\{ s(X) + s(W|X) + \frac{d-p(W,X)}{p(W,X)(1-p(W,X))}\dot{p}(W,X) + (1-d)s(Y|W,X,D=0) \right\}.$$

Then it can be verified that $A(X)(W - h(X,\beta))$ is the projection of itself into the $s(W|X)$ component of τ_2, and that

$$\frac{1-D}{1-p(X)} A(X)(h(X,\beta) - g(Y,\beta))$$

is the projection of $A(X)(h(X,\beta) - g(Y,\beta))$ into the $(1-d)s(Y|X)$ component of τ_1, because

$$E\frac{1-D}{1-p(X)} A(X)(h(X,\beta) - g(Y,\beta)) s(Y,X) = E(h(X,\beta) - g(Y,\beta)) s(Y,X),$$

because of assumption 1. Therefore for a given $A(X)$ the efficient projection of the unconditional moment is

$$A(X)\left[W - h(X;\beta) + \frac{1-D}{1-p(X)}(h(X;\beta) - g(Y;\beta))\right].$$

Note that the two components are possibly correlated with each other. The efficient instrument function $A^2(X)$ can then be derived by following Newey and McFadden (1994) and is given by

$$A^2(X) = \frac{\partial}{\partial\beta} h(X;\beta) Var\left[W - h(X;\beta) + \frac{1-D}{1-p(X)}(h(X;\beta) - g(Y;\beta)) \middle| X\right].$$

Similarly, an efficient estimator $\hat{\beta}$ can be devised by replacing the moment condition $m(Z;\beta)$ by $\hat{A}^2(X)(W - g(Y;\beta))$. Under suitable regularity conditions, the variation from nonparametric estimation of the efficient instrument does not contribute to the variance of β.

References

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

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

Bollinger, C. (1998): "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, 16(3), 576–594.

Bound, J., C. Brown, G. Duncan, and W. Rodgers (1994): "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, 12, 345–368.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics, Vol. 5*, ed. by J.J.Heckman, and E. Leamer. North Holland.

Bound, J., and A. Krueger (1991a): "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right," *Journal of Labor Economics*, 12, 1–24.

——— (1991b): "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?," *Journal of Labor Economics*, 9, 1–24.

Carroll, R., and M. Wand (1991a): "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society*, 53, 573–585.

——— (1991b): "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society*, 53, 573–585.

Chen, X. (2003): "Large Sample Sieve Estimation of Semi-nonparametric Models," in *Handbook of Econometrics, Vol 6*. forthcoming, Elsevier Science.

Chen, X., L. Hansen, and J. Scheinkman (1997): "Shape-Preserving Estimation of Diffusions," Working paper, University of Chicago, Department of Economics.

Chen, X., H. Hong, and E. Tamer (2003): "Measurement Error Models with Auxiliary Data," Working Paper, forthcoming Review of Economic Studies.

Chen, X., O. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function is not Smooth," *Econometrica*, 71, 1583–1600.

Chen, X., and X. Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314.

Das, M., W. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

Elbers, C., J. Lanjouw, and P. Lanjouw (2003): "Micro-level estimation of poverty and inequality," *Econometrica*, 71(1), 355–364.

Fan, J. (1992): "Design-adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Firpo, S. (2004): "Efficient Semiparametric Estimation of Quantile Treatment Effects," manuscript, Department of Economics, University of British Columbia.

Gallant, A. R., and D. W. Nychka (1987): "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.

Hahn, J. (1998): "On the Role of Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–332.

Hausman, J., H. Ichimura, W. Newey, and J. Powell (1991): "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics*, 50, 271–295.

Hausman, J., W. Newey, and J. Powell (1995): "Nonlinear Errors in Variables: Estimation of Some Engle Curves," *Journal of Econometrics*, 65, 205–233.

Heckman, J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–94.

Heckman, J., R. LaLonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics, Vol. 3A*, ed. by O. Ashenfelter, and D. Card. Elsevier Science.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Hsiao, C., and L. Wang (1995): "A Simulation Based Semi-parametric Estimation of Nonlinear Errors-in-Variables Models," Working Paper, University of Southern California.

Lee, L., and J. Sepanski (1995): "Estimation of Linear and Nonlinear Errors-inVariables Models Using Validation Data," *Journal of the American Statistical Association*, 90(429), 130–140.

Li, T. (2002): "Robust and consistent estimation of nonlinear errors-in-variables models," *Journal of Econometrics*, 110, 1–26.

Newey, W. (1990a): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58(4), 809–837.

——— (1990b): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–82.

——— (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors in Variables Models," *Review of Economics and Statistics*.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2113–2241. North Holland.

Ravallion (1988): "Expected poverty under risk-induced welfare variability," *The Economic Journal*, 98, 1171–1182.

Robins, J., S. Mark, and W. Newey (1992): "Estimating exposure effects by modelling the expectation of exposure conditional on confounders," *Biometrics*, 48, 479–95.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90(429), 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of regression coefficients when some regressions are not always observed," *Journal of the American Statistical Association*, 89(427), 846–866.

——— (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90(429), 106–121.

Rotnitzky, A., and J. Robins (1995): "Semiparametric Regression Estimation in the Presence of Dependent Censoring," *Biometrika*, 82(4), 805–820.

Schennach, S. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72(1), 33–75.

Sepanski, J., and R. Carroll (1993): "Semiparametric Quasi-likelihood and Variance Estimation in Measurement Error Models," *Journal of Econometrics*, 58, 223–256.

Tarozzi, A. (2004): "Calculating Comparable Statistics from Incomparable Surveys, with an Application to Poverty in India," working paper, Duke University.

Tripathi, G. (2003): "GMM and Empirical Likelihood with Incomplete Data," Department of Economics, University of Wisconsin-Madison.

——— (2004): "Moment based inference with incomplete data," Department of Economics, University of Wisconsin-Madison.

Wang, Q., O. Linton, and W. Hardle (2004): "Semiparametric Regression Analysis for Missing Response Data," *Journal of the American Statistical Association*, 99, 334–345.

Wooldridge, J. (2002): "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification," *Portuguese Economic Journal*, 1, 117–139.

——— (2003): "Inverse Probability Weighted Estimation For General Missing Data Problems," manuscript, Michigan State University.

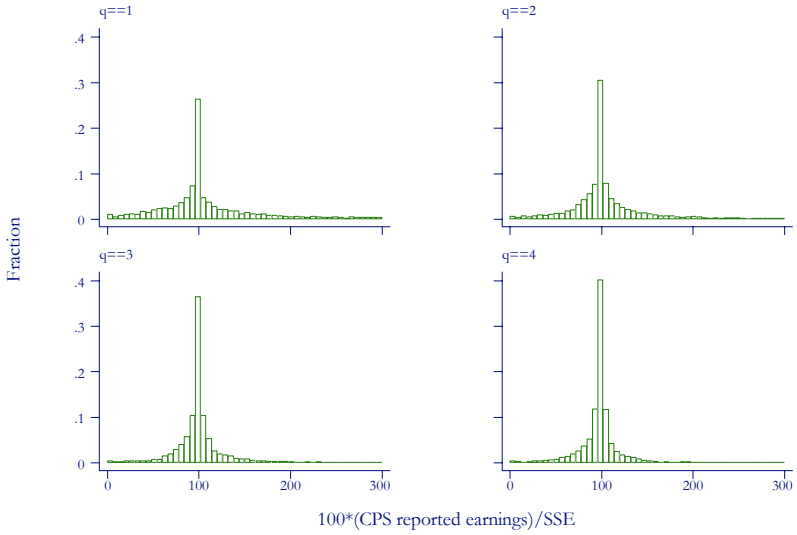

Figure 1: Histograms of measurement error in earnings, by quartile of true (Social Security) earnings

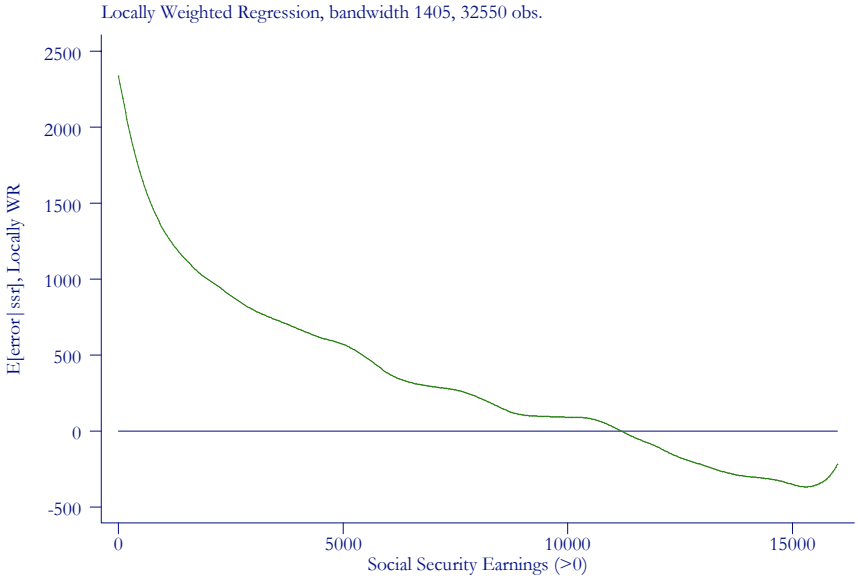

Figure 2: Measurement Error in Earnings as a function of true (Social Security) Earnings

Table 1: Summary Statistics – CPS March 1978

	Full sample		With SSR match		No SSR Match	
	mean	s.dev.	mean	s.dev.	mean	s.dev.
Reported Earnings	8865	7974	9435	8010	7927	7826
Reported Earnings (< $16500)	6363	4673	6759	4608	5741	4707
Social Security Earnings (< $16500)			6576	4639		
Years of education	13.3	2.84	13.4	2.78	13.1	2.93
Age	35.5	14.5	36.7	14.4	33.5	14.6
Female	.46	.50	.46	.50	.45	.50
Non-white	.11	.31	.10	.31	.11	.31
Married	.59	.50	.63	.48	.53	.50

All earnings refer to 1977, and are in 1977 US$. Full sample refers to the sample of 62296 individuals selected as described in the text.

Table 2: Earnings quantiles ('000$)

α	(1) unadjusted (primary)	(2) auxiliary data	(3) adjusted Par. IPW	(4) adjusted NP, CEP	(4) adjusted NP, IPW
0.05	0.30 (0.00533)	0.485 (0.01114)	0.45 (0.02209)	0.38 (0.01323)	0.38 (0.01323)
0.10	0.73 (0.01202)	0.996 (0.01668)	0.93 (0.02693)	0.78 (0.01580)	0.78 (0.01580)
0.15	1.20 (0.01724)	1.583 (0.02398)	1.47 (0.02915)	1.24 (0.01888)	1.25 (0.01891)
0.20	2.00 (0.01896)	2.251 (0.03188)	2.09 (0.03070)	1.80 (0.02382)	1.81 (0.02388)
0.25	2.70 (0.04291)	2.981 (0.02756)	2.78 (0.03250)	2.46 (0.03028)	2.47 (0.03035)

Column (1): Calculated from the unadjusted primary sample. Column (2): Calculated from the unadjusted auxiliary sample. Column (3): IPW Estimator, with a logit first step that includes only a constant and reported earnings. Column (4): CEP-GMM cubic sieve Estimator, with 10 knots, using reported earnings as predictor. Column (5): IPW-GMM. Flexible logit with cubic sieve, with 10 knots, using reported earnings as predictor.

Table 3: Cumulative Distribution function

$	(1) unadjusted (primary)	(2) auxiliary data	(3) adjusted Par. IPW	(4) adjusted NP, CEP	(4) adjusted NP, IPW
500	0.071 (0.00103)	0.051 (0.00112)	0.055 (0.00118)	0.065 (0.00129)	0.065 (0.00129)
1,000	0.124 (0.00132)	0.100 (0.00153)	0.107 (0.00159)	0.125 (0.00161)	0.124 (0.00161)
1,500	0.166 (0.00149)	0.143 (0.00178)	0.152 (0.00183)	0.174 (0.00180)	0.173 (0.00180)
2,000	0.199 (0.00160)	0.182 (0.00196)	0.193 (0.00199)	0.217(0.00192)	0.216 (0.00192)
2,500	0.236 (0.00170)	0.217 (0.00209)	0.230 (0.00210)	0.253 (0.00200)	0.252 (0.00200)
3,000	0.264 (0.00177)	0.252 (0.00220)	0.267 (0.00219)	0.289 (0.00205)	0.288 (0.00205)
3,500	0.300 (0.00184)	0.282 (0.00228)	0.298 (0.00225)	0.318 (0.00208)	0.317 (0.00208)
4,000	0.322 (0.00187)	0.310 (0.00235)	0.327 (0.00229)	0.346 (0.00210)	0.345 (0.00210)
4,500	0.352 (0.00191)	0.336 (0.00240)	0.355 (0.00232)	0.371 (0.00212)	0.370 (0.00212)
5,000	0.373 (0.00194)	0.363 (0.00244)	0.383 (0.00234)	0.397 (0.00214)	0.396 (0.00214)

Earnings refer to 1977, and are in 1977 US$. Full sample refers to the sample of 62296 individuals selected as described in the text. Column (1) - Calculated from the unadjusted primary sample. Column (2) - Calculated from the unadjusted auxiliary sample. Column (3) - IPW Estimator, with a logit first step that includes only a constant and reported earnings. Column (4) - CEP-GMM cubic sieve Estimator, with 10 knots, using reported earnings as predictor. Column (5) - IPW-GMM. Flexible logit with cubic sieve, with 10 knots, using reported earnings as predictor.