# Combining Strategies for the Estimation of Treatment Effects[*]

Sergio Firpo[**]
Rafael de Carvalho Cayres Pinto[***]

### Abstract

The estimation of the average effect of a program or treatment on a variable of interest is an important tool for the assessment of economic policies. In general, assignment of potential participants to treatment does not occur at random and could thus generate a selection bias in absence of some correction. A way to get around this problem is by assuming that the econometrician observes a set of determinant characteristics of participation up to a strictly random component. Under such an assumption, the literature contains semiparametric estimators of the average treatment effect that are consistent and can asymptotically reach the semiparametric efficiency bound. However, in frequently available samples, the performance of these methods is not always satisfactory. The aim of this paper is to investigate how the combination of two strategies may generate estimators with better properties in small samples. Therefore, we consider two ways of combining these approaches, based on the double robustness literature developed by James Robins et al. We analyze the properties of these combined estimators and discuss why they can outperform the separate use of each method. Finally, using a Monte Carlo simulation, we compare the performance of these estimators with that of the imputation and reweighting techniques. Our results show that the combination of strategies can reduce bias and variance, but this improvement depends on adequate implementation. We conclude that the choice of smoothing parameters is decisive for the performance of estimators in medium-sized samples.

*Keywords:* Average Treatment Effect, Ignorability, Imputation, Reweighting, Double Robust Estimation, Monte Carlo Simulation.

*JEL Codes:* C14, C21, C31, C51.

[*] Submitted in January 2013. Revised in March 2013.

[**] Sao Paulo School of Economics at FGV and IZA. E-mail: `sergio.firpo@fgv.br`

[***] PUC-Rio and BNDES. E-mail: `rafael.pinto@puc-rio.br`

## 1. Introduction

The identification of the causal effect of a treatment or program on a variable of interest is one of the major topics discussed in the econometric literature and it is an integral part of the development of public policies, such as active interventions in the labor market (Heckman et al., 1997, 1998). The key point in the analysis of this problem concerns the relationship between unobserved components that eventually determine the outcome and those which affect participation in the program. Given the inability to carry out randomized experiments, there must be self-selection for the treatment, which often follows patterns that are unknown to the econometrician. This leads to the so-called selection bias, since the effect of individual treatment is different for affected and unaffected units.

Under such circumstances, an additional assumption is necessary to identify a parameter of interest. The ignorability assumption (Rosenbaum and Rubin, 1983) postulates that all relevant information about heterogeneity can be obtained from auxiliary variables observed for all units. In other words, there is no systematic selection bias when comparing individuals which are similar in regard to given characteristics. Taken in isolation, this assumption defines a semiparametric model for the population, i.e., it imposes that the underlying probability distribution ought to belong to a given class which, even though it is a proper subset of the universe of all probability measures, is too big to be indexed by a finite dimensional parameter. When one is interested in inference about the treatment effect on the whole population, using non-experimental data, this is the main model available in the literature that allows for the exact identification of a parameter.

The theory addressing the semiparametric estimation of the treatment effect under ignorability in large samples is in an advanced developmental stage. A wide array of distinct methods has been thoroughly investigated, and each approach presents conditions under which asymptotically efficient estimation is assured. Two of the most important techniques are imputation (or regression) and reweighting. When utilizing imputation, data from each group are used to estimate the relationship between potential outcome and auxiliary variables, the so-called regression function; then, the estimated regression function is used as substitute for the unobserved outcome in the other group. The second method consists of estimating the conditional probability of selection given auxiliary variables' values, known as propensity score. This information describes the relative representation of the groups for any combination of the auxiliary characteristics, thus enabling reweighting, which allows the sample to be representative of a population in which participation was randomly assigned.

A peculiar aspect of the asymptotic theory for some semiparametric problems is that the procedures are equivalent, provided that relatively weak regularity conditions are satisfied (Newey, 1994). However, practical issues spark interest in inference performance in small samples. Given that, specially in this context, asymptotic results have little relation with the properties of the estimators in finite samples, a number of simulation studies have emerged. These analyses, besides comparing the different methods, underscore the importance of implementation of each estimator. In particular, a recent issue that has developed involves the potential benefit of combining different methods into procedures which share the strength of each individual technique under certain circumstances.

This paper seeks to provide better insight into the possibilities available from the combination of imputation and reweighting methods for treatment effects models.[1] To attain this goal, two approaches were employed. First, we discuss the theoretical rationale behind the superior performance of the mixed method, linked to the double robust estimation literature (Robins and Rotnitzky, 1995, Robins et al., 1995). However, it is important to highlight that the major reason for a combined approach in our case differs from that of most of these studies. As a matter of fact, double robust inference has been applied by most studies to combine parametric imputation and reweighting in order to obtain consistency, when either the regression function used in imputation or the propensity score in reweighting is well-specified. In this paper, we try to improve the performance in finite samples by combining semiparametric imputation and reweighting, a possibility that, albeit supported by Robins and Ritov (1997), is underexplored in the literature, exceptions being the work of Cattaneo (2010) and Rothe and Firpo (2013). Based on this discussion, we propose two double robust procedures. The first of them is a straightforward generalization of the estimator proposed by Scharfstein et al. (1999), in which we use preliminary sieve estimates in lieu of parametric ones. This method, which until recently had not been explicitly dealt with in the literature, coincides with that of Cattaneo (2010) and more generally with the one discussed by Rothe and Firpo (2013). The second method consists of a semiparametric imputation procedure in which the estimates of regression functions are obtained by weighted least squares, where the weights depend on estimated propensity score of each unit. This estimator was implemented by Hirano and Imbens (2001), but its properties lack further investigation.

As a complement to the theoretical analysis, we run Monte Carlo simulations, which compare the performance of different implementations of imputation, reweighting, and the proposed double robust estimators. As a way to assess the relevance of the latter estimators and to test the theoretical predictions and suggestions, this exercise is reproduced for several population models. The specifications differ in terms of the functional form of the true propensity score and regression functions, heteroskedasticity of the potential values, and size of the set of auxiliary variables. With respect to the functional forms, we seek to vary the smoothness of the model systematically, based on the concept of 'smooth function' relevant to the theory. By doing this, we thereby avoid arbitrary choices of functional forms, which are very frequent in simulation studies.

Results indicated that the combination of imputation and reweighting in dou-

[1] Rothe and Firpo (2013) provide a more general and theoretical treatment of this issue.

ble robust procedures allows reducing the mean squared error in all scenarios. In addition, in line with the theoretical analysis, potential efficiency gains were particularly high in less smooth models and when a multidimensional auxiliary variables' set was considered. Heteroskedasticity, in turn, did not have a clear effect on the advantage of combining methods. Finally, in most models simulated, combining the estimation of regression functions with weighting by the inverse of the true propensity score proved to be an effective way to use knowledge of the latter.

## 2. Identification of Treatment Effect Under Ignorability

In this section, we define the problem to which the methods discussed in the remainder of this paper can be applied. We consider a heterogeneous population of units (e.g., individuals, households, or firms) that may be subject to different alternative regimens or treatments (e.g., social benefits, geographical situation, or tax systems). We intend to estimate the 'treatment effect', which can be broadly defined as the causal effect of a treatment on a given attribute of units, i.e., on the outcome or variable of interest.

Following the work of Rubin (1973, 1977, 1978), causal effect is obtained by comparing the possible values of the variable of interest under different treatment conditions. The main difficulty associated with this approach is that, since each unit is observed under only one of the treatments, the necessary comparison involves unobserved values.

Several substitutes for the (impossible) comparison between counterfactual values have been suggested and investigated by the literature on program evaluation. The model studied in this paper assumes availability of additional information about the units, called auxiliary variables, covariables, or pre-treatment variables. Under a given type of hypothesis, generally referred to as the ignorability assumption, it is legitimate to compare the outcomes of units with the same auxiliary characteristics, but subjected to different treatments.

These conditions suggest performing experiments where groups of individuals with heterogeneous auxiliary variables are randomly submitted to different treatments. In this case, experimental data are made available. Notwithstanding, the analyst often cannot carry out such experiments, and relies only upon data obtained from the sampling of a population where treatment assignment is not within reach. The model introduced in this section illustrates the latter case, i.e. the problem of inference from non-experimental or observational data under the ignorability assumption.

### 2.1 Basic Elements

The available data originate from the observation of a random sample of $N$ units of the population of interest $(Y, X, T)$, indexed by $i = 1, 2, ..., N$. For every unit $i$, the observed value of the variable of interest is $Y_i$, which we assume to be real-valued. The treatment given to $i$ is denoted by $T_i$. As in most of the literature, we take into account a binary set of alternative treatments, which represent the participation or non-participation in the program we want to assess.[2] In this context, we have the treatment group, denoted by $T_i = 1$, which is composed of the units affected by the intervention, and the control group, denoted by $T_i = 0$, which includes the unaffected units. Finally, $X_i$ represents the set of auxiliary variables of unit $i$. The fundamental property of $X_i$ is that neither its observation nor its value depends, in terms of causality, upon the treatment $T_i$. This is logically verified when $X_i$ is observed before the determination of the received treatment. Because of this particular case, the components of $X$ are sometimes called "pre-treatment" variables.

Rubin's potential values approach leads us to define the variables $Y_i(t)$, for each possible value $t$ of $T_i$, representing the value of $Y_i$ if individual $i$ is subject to treatment $T_i = t$. This notation device is important for elaborating the model[3] and implicitly carries the assumption that each unit is not influenced by the treatment received by other units. This hypothesis, which is quite plausible in the original context of clinical trials (as in Rubin, 1978), demands careful interpretation of the model if applied to social phenomena.

Another concept derived from the work of Rubin is the propensity score (Rosenbaum and Rubin, 1983), defined as the probability of selection for treatment, conditional on the value of auxiliary variables.

$$p(x) \equiv Pr[T_i = 1|X_i = x] = E[T_i|X_i = x]$$

It is also convenient to introduce the regression functions

$$m_t(x) \equiv E[Y_i(t)|X_i = x]$$

and conditional variance functions

$$\sigma_t^2(x) \equiv V[Y_i(t)|X_i = x]$$

## 2.2 Parameter of interest

The parameter we intend to estimate is

$$\beta = E\left[Y(1) - Y(0)\right] = E\left[Y(1)\right] - E\left[Y(0)\right]$$

[2] The techniques used for analyzing a treatment effect which could take a finite number of values are similar to those of the binary case. The analysis of identification hypotheses akin to those we use in this paper can be seen in Imbens (2000). Cattaneo (2010) develops versions of two of the estimators we consider. The recent review by Imbens and Wooldridge (2009) provides a discussion on treatment methods with continuous values.

[3] Imbens and Wooldridge (2009) discuss the difficulties associated with the formulation of the problem using only observed values.

which is the expected difference between the potential values of the attribute of interest, called average treatment effect. The focus on the inference about the average effect should not be viewed as an important limitation, as there is an immediate generalization towards any given transformation $g(.)$ of $Y$. In particular, by taking $g(Y) = 1(Y \leq y)$, and because we effectively estimate $E[Y(1)]$ and $E[Y(0)]$ separately, the techniques we discuss allow to estimate the marginal distributions of each potential outcome. For instance, $Y(1)$, $F_1(y) = Pr(Y(1) \leq y) = E[g(Y(1))]$, at each point $y$. Another possibility is discussed in subsection 2.6.

A related family of parameters of interest that has commanded a lot of attention in the literature is that of treatment effects on the treated. These parameters consider outcomes' averages conditional on treatment assignment, thus measuring the effect of the program on the population of units that was actually submitted to the intervention. The main example of this type of parameter is the average treatment effect on the treated, $\beta_{ATT} = E\left[Y(1) - Y(0)|T = 1\right]$. Due to space restrictions, estimators for $\beta_{ATT}$ will not be addressed in our paper.

### 2.3 Identification

Because of the heterogeneity between groups, comparing population means of $Y(1)$ and $Y(0)$ by considering the subpopulations in which they are observed would be inappropriate due to the selection bias. Therefore, the identification of the parameter of interest depends on assumptions about the relationship between unobserved and observed variables.

In this paper, we analyze estimators based on the ignorability assumption, which corresponds to stating that relevant and/or systematic heterogeneity is captured by auxiliary variables. More specifically, we consider the strong ignorability assumption, which we will simply call ignorability, following Rosenbaum and Rubin (1983). Strong ignorability is composed by

(i) the conditional independence assumption:[4]

$$T_i \perp (Y_i(1), Y_i(0)) \mid X_i \tag{1}$$

which determines the independence of treatment assignment relative to potential values, given the auxiliary and

(ii) overlap assumption

$$0 < \varepsilon < p(x) < 1 - \varepsilon < 1, : \forall : x \in supp(X) \tag{2}$$

where $supp(X)$ represents the support of the distribution of $X$. Intuitively, overlap sets a uniform bound, greater than zero, for the probability of selection and non-selection.

[4] Also known as *Unconfounded Treatment Assignment* (Rosenbaum and Rubin, 1983).

A widely investigated analogous case is that of missing data, in which one seeks to infer the marginal distribution of a variable $Y$, observed only for some of the sample units. The missing at random assumption posits that, conditional on the auxiliary variables, the probability of observing $Y$ is independent from its value. The model thus defined is equivalent to that of treatment effect under ignorability, when $Y(0)$ is regarded as identically null. Conversely, estimating the average treatment effect is equivalent to solving two missing data problems in order to obtain estimates of the mean of each potential value, $E[Y(1)]$ and $E[Y(0)]$, and then taking their difference. In this case, the original ignorability assumption implies the missing at random assumption in both missing data problems it is decomposed into.

A way to demonstrate the identification of $\beta$ under ignorability is motivated by the following application of the law of iterated expectations:

$$\begin{aligned}\beta = E\left[Y_i(1) - Y_i(0)\right] &= E\left[E\left[Y_i(1) - Y_i(0)|X_i\right]\right] \\ &= E\left[E\left[Y_i(1)|X_i\right] - E\left[Y_i(0)|X_i\right]\right] \\ &= E\left[m_1(X_i) - m_0(X_i)\right]\end{aligned}$$

i.e., the parameter of interest can be represented as the population mean of a difference between the functions of observed variables $X$. Although the regression functions are presumably unknown and involve potential outcomes, as an immediate consequence of conditional independence, we have:

$$E\left[Y_i(t)|X_i\right] = E\left[Y_i(t)|X_i, T_i = t\right] = E\left[Y_i|X_i, T_i = t\right], : t = 0, 1$$

Hence, it is possible to rewrite $m_1(.)$ and $m_0(.)$ using, respectively, the data on treated and control observations:

$$\beta = E\left[m_1(X_i) - m_0(X_i)\right] = E\left[E\left[Y_i|X_i, T_i = 1\right] - E\left[Y_i|X_i, T_i = 0\right]\right] \tag{3}$$

The overlap assumption allows estimating $E\left[Y_i(1)|X_i\right] = E\left[Y_i|X_i, T_i = 1\right]$ and evaluating it over the empirical distribution of $X$. In fact, (2) implies that any region to which the marginal distribution of $X$ assigns a strictly positive probability, say $q$, has also a strictly positive probability conditional on $T = 1$, at least $q\frac{\varepsilon}{1-\varepsilon}$. In this way, the observation of the behavior of $Y(1)$ (and analogously , of $Y(0)$) for all values of $X$ is (probabilistically) assured.

Another way to show the identification is by observing that, once the ignorability assumption holds, we get

$$\begin{aligned}E\left[\frac{TY}{p(X)}\right] &= E\left[E\left[\frac{TY}{p(X)}|X\right]\right] = E\left[\frac{E\left[TY|X\right]}{p(X)}\right] \\ &= E\left[\frac{E\left[Y(1)|X\right]Pr(T=1|X)}{p(X)}\right] \\ &= E\left[E\left[Y(1)|X\right]\right] = E\left[Y(1)\right] \end{aligned} \tag{4}$$

where existence of the first expression follows from equation (2) and, for the equality in the second line, we use

$$\begin{aligned}E\left[TY|X\right] &= E\left[Y|X,T=1\right]Pr(T=1|X) + 0Pr(T=0|X) \\ &= E\left[Y(1)|X,T=1\right]Pr(T=1|X)\end{aligned}$$

and the conditional independence assumption, which implies $E\left[Y(1)|X,T=1\right] = E\left[Y(1)|X\right]$. Analogously, we have $E\left[\frac{(1-T)Y}{1-p(X)}\right] = E\left[Y(0)\right]$ and, therefore,

$$\beta = E\left[Y_i(1) - Y_i(0)\right] = E\left[\frac{TY}{p(X)}\right] - E\left[\frac{(1-T)Y}{1-p(X)}\right] \tag{5}$$

which is a representation of the parameter of interest as the difference of the weighted averages of the observed values. Equation 4 shows that the weights proportional to the inverse propensity score render the mean of the variable of interest among treated individuals representative of the mean potential value $Y(1)$ for the population.

## 2.4 Criticism and alternatives to the ignorability assumption

Among the components of ignorability, the overlap assumption is the testable part, and verifying it is highly recommended. Violation of this assumption means that certain units of a group have few or no correspondents in the other, and thus at least one of the counterfactual outcomes cannot have its mean properly estimated for some part of the population. Busso et al. (2013) demonstrate that in this case the accuracy of ignorability-based estimators is poor and the asymptotic theory provides little information about performance. This problem can be circumvented by excluding observations with extreme propensity score (estimated) values, as is usually done in applied studies. Crump et al. (2009) systematically discuss this possibility and show how to minimize the asymptotic variance of average treatment effect estimators by discarding observations outside a given subset $A^*$of auxiliary variable profiles. The same study demonstrates conditions under which $A^*$ is determined solely by the propensity score of the observations, i.e., it takes the form $A^* = \{x \in supp(X)|p(x) \in [\alpha, 1-\alpha]\}$, for some $\alpha \in (0,1)$. However, we observe that this method requires redefining the population or changing the parameter of interest, since it estimates $E[Y(1) - Y(0)|X \in A^*]$ instead of the average treatment effect $\beta = E[Y(1) - Y(0)]$ in relation to the original population. This reflects the fact that ignorability, by relying solely on comparison of similar units, would clearly not be expected to be successful if some unit types were not observed in both conditions (treated and control).

The greater concern in using the ignorability assumption is validity of conditional independence. The reason for this issue is that in econometric applications the value of variable $Y$ matters to the units, who are typically economic agents. It is plausible that these agents have more information on the potential values than the knowledge of auxiliary variables, and that they self-select for the treatment according to their expectations. Thus, conditional independence is threatened by the possibility that the additional information is not independent of treatment assignment, given $X$.

This argument suggests that conditional independence may be an excessively strong assumption. Nonetheless, Scharfstein et al. (1999) show that conditional independence is a minimum requirement for identification in the absence of assumptions about the propensity score and the joint distribution of $(Y(t), X)$. More precisely, they consider a missing data model, where $(YT, X, T)$ is observed, i.e., the value of $Y$ is known only when $T = 1$. The joint distribution of complete data $(Y, X)$, $F_{X,Y}(.,.)$ is allowed to be any valid bivariate probability function, and the observation probability, $P(T = 1|Y, X)$, is represented by the product between an arbitrary component, dependent solely upon $X$, $\lambda(X)$, and a function that is potentially dependent on $Y$, $r(Y, X; \alpha_0)$, known to belong to a parametric family $\{r(.,.;\alpha)|\alpha \in A\}$.[5] In this model, it is demonstrated that neither the distribution $F_{X,Y}(.,.)$ nor the components of observation probability $\lambda(.)$ and $\alpha_0$ are identified. Additionally, when an arbitrary value $\tilde{\alpha} \in A$ is fixed for the parametric component, any joint distribution of $(YT, X, T)$ (satisfying certain regularity conditions) can be exactly reproduced by the suitable choice of nonparametric components. Consequently,

(i) in the absence of assumptions about complete data $(F_{X,Y}(.,.))$ or about the relationship between the observed variables and selection $(\lambda(X))$, identification requires an assumption about the relationship between the potentially unobserved variable and the observation probability, i.e., that an $\alpha$ be fixed, and

(ii) it is impossible to infer $\alpha_0$, based on observable data, i.e., it is not possible to reject any value for this parameter by statistical testing.

The difficulty in justifying the ignorability assumption must therefore be put into perspective by considering its role in identification. In this sense, it is useful

[5] The model considered by Scharfstein et al. (1999) is a bit more complex, allowing variables $X$ and $T$ to be time-dependent, but contains the version discussed here as a particular case.

to discuss some identification alternatives. Manski (1990, 2003) shows that, under weaker assumptions, it is possible to determine bounds to the treatment effect, based on the distribution of observable data. Another possibility – the use of instrumental variables (Imbens and Angrist, 1994) – requires the observation of instruments that exogenously lead to the participation of certain units. This strategy allows estimating the average treatment effect for this subpopulation without restricting the relationship between potential values and participation. However, the exogeneity condition imposed on the instruments is often strong. Some alternatives are concerned with the identification of the treatment effect in even more specific subpopulations. The 'Regression Discontinuity' approach (Hahn et al., 2000) also allows identification under arbitrary heterogeneity, but only for observations near a point along the distribution of a given variable, in which there exists discontinuity in the probability of selection for a treatment. Methods that include data from more than one period control for unobserved factors affecting potential values, thus identifying the treatment effect for the group of units that changed their treatment status.

### 2.5 Semiparametric efficiency bound

Strong ignorability by itself defines a semiparametric model for the data, as the set of possible distributions, albeit limited by the assumption, cannot be parameterized by a finite dimensional set. Therefore, even if one is interested in estimating the finite-dimensional parameter $\beta$, the traditional definition of Cramer-Rao's lower bound does not apply. However, an analogous concept, initially proposed by Stein (1956) and developed by Bickel et al. (1993), among others, allows establishing the maximum accuracy a given class of semiparametric estimators may achieve.

The rationale behind the semiparametric efficiency bound is to consider estimation in certain parametric submodels, i.e., families $(\Theta, \{f(z;\theta); \theta \in \Theta\})$ of distributions, parameterized by finite-dimensional set $\Theta$, such that $f(z;\theta)$ satisfies the semiparametric constraint for all $\theta \in \Theta$, and that for some $\theta_0 \in \Theta$, $f(z;\theta_0)$ is the true distribution. For each $\theta_0$, Cramer-Rao's bound, should it exist and be well-defined, must be smaller than the variance of a valid estimator in the semiparametric model. Formally, the semiparametric efficiency bound is defined by the supremum of the efficiency bounds of regular parametric submodels (see definition in Newey, 1990).

The semiparametric efficiency bound applies to regular estimators. An estimator is said to be regular if, for every regular parametric submodel $(\Theta, \{f(z;\theta); \theta \in \Theta\})$, and every sequence $(\theta_n) \subseteq \Theta$ such that $\sqrt{n}(\theta_n - \theta_0)$ is limited, the sequence of distributions of $\sqrt{n}(\hat{\beta} - \beta(\theta_n))$ under $\theta_n$ converges to the same limit. This class excludes the so-called 'superefficient estimators' and those that use more information than is contained in the semiparametric model. However, regularity is less of a stringent requirement than uniform convergence (even if locally) in distribution.[6] Hence, in general, approximations implied by the asymptotic properties of regular estimators depend on unknown sample sizes.

Following Newey (1990), a parameter is said to be differentiable if

(i) it is differentiable with respect to the parameters of any smooth parametric submodel, and

(ii) there exists a function $d$, of finite variance, which satisfies the following expression for any regular parametric submodel $(\Theta, \{f(z;\theta); \theta \in \Theta\})$:

$$\frac{\partial \beta(\theta_0)}{\partial \theta} = E\left[dS'_\theta\right] \tag{6}$$

the so-called pathwise differentiability equation, where $S_\theta$ is the score (the derivative of the log-likelihood for a single observation) of the submodel. Newey shows that the efficiency bound for a differentiable parameter is given by $E[\delta\delta']$, where $\delta$ is the projection of $d$ (component for component) onto the closed linear space spanned by the scores, also known as tangent space.

As $\delta$ has a zero mean, its variance is given by $E[\delta\delta']$. Therefore, an asymptotically linear estimator with influence function $\delta$ is asymptotically efficient. For this reason, $\delta$ is called the efficient influence function of the model.

Hahn (1998) demonstrates that, under strong ignorability, the average treatment effect is a differentiable parameter, with a derivative

$$\begin{aligned}\frac{\partial \beta(\theta_0)}{\partial \theta} = E\Bigg[\bigg(&m_1(X) - m_0(X) - \beta + \\ &\frac{T}{p(X)}(Y - m_1(X)) + \frac{1-T}{1-p(X)}(Y - m_0(X))\bigg) S_\theta\Bigg]\end{aligned} \tag{7}$$

whereas the tangent space contains all the functions of the form

$$a(X)(T - p(X)) + b(X) + Ts_1(Y,X) + (1-T)s_0(Y,X) \tag{8}$$

where $a(.)$ is an arbitrary square integrable function, $b(X)$ has a zero mean under the true distribution of X, and $s_t(Y,X)$, $t = 0,1$, has a zero mean under the distribution of $Y$ conditional on any value of $X$. It can be verified that the expression for $d$ implied by comparing equations (6) and (7) belongs to the tangent space described by expression (8). Therefore, this expression coincides with its own projection and determines the efficient influence function

[6] Bickel et al. (1993) show that the existence of uniformly convergent estimators imposes extremely strong constraints on the models, thus ruling out many semiparametric models.

$$\psi^* = m_1(X) - m_0(X) - \beta + \frac{T}{p(X)}(Y - m_1(X)) - \frac{1-T}{1-p(X)}(Y - m_0(X)) \quad (9)$$

and the efficiency bound

$$E[\psi^{*2}] = E\left[(m_1(X) - m_0(X) - \beta)^2 + \frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1-p(X)}\right] \quad (10)$$

## 2.6 Generalization for other parameters of interest

As was noted in subsection 2.2, an inference method for the average treatment effect can be easily extended to the estimation of the distribution of each potential value at any point. This allows approximating arbitrarily well the marginal distributions of the both conditional outcomes and making almost any comparison between them. Such approximations, however, will depend on the estimates obtained for a large number of points, which evidently make the assessment of the properties of such procedure less practical.

Nevertheless, if the intention is to assess how the distributions differ in terms of a single parameter defined by an unconditional moment condition, another generalization can be used. Let $\mu^* = \mu_1 - \mu_0$ be the quantity to be estimated, with $\mu_t$, $t = 0, 1$, solutions of the moment conditions

$$G(\mu_t) = E[g(X, Y(t); \mu_t)] = 0$$

where $g(.)$ is known. Since for data $X$ and $\mu$, we have $g(X, Y(1); \mu)$ and $g(X, Y(0); \mu)$ are known functions of $Y(1)$ and $Y(0)$, the conditional independence assumption

$$T_i \perp (Y_i(1), Y_i(0)) \mid X_i$$

implies

$$T_i \perp (g(X, Y(1); \mu), g(X, Y(0); \mu)) \mid X_i$$

Thus, if the ignorability assumption is satisfied by $(Y(1), Y(0), X, T)$, then it also holds for $(g(X, Y(1); \mu), g(X, Y(0); \mu), X, T)$.[7] Then, it is possible to estimate, for example, the left-hand side of the moment condition $E[g(X, Y(1); \mu] = 0$ for all $\mu$, and to obtain an estimate $\hat{\mu}_1$ for $\mu_1$ from the value that solves the estimated moment condition. Moreover, the consistency and asymptotic efficiency of this type of estimator require that the estimation of the moment condition have the same properties.

For instance, Firpo (2007) develops an estimator for the quantile treatment effect – difference of the quantiles of potential distributions – which shares the

[7] Note that, trivially, the overlap assumption remains unchanged.

desirable properties of average treatment effects estimators addressed in the next section. Chen et al. (2008) discuss the estimation of solutions of very general – not necessarily smooth – moment conditions.

## 3. Estimation by Imputation and Reweighting

This section discusses two methods that have received a great deal of attention in the literature on the estimation of treatment effect under ignorability.[8]

The first method consists of the imputation of an estimate of the regression function as a substitute for potential values. This allows for the estimation of average treatment effects by the mean of differences between imputed treatment and control outcomes for each sample unit. The second method involves reweighting both control and treated groups, to make each of them representative of the whole population. Thus, by subtracting the weighted mean of outcomes across the control units from that of the treated, one would obtain an estimate for the average treatment effect.

Despite their distinct origins, both estimators can be analyzed under the same theoretical framework developed by Newey (1994). This theory guarantees that both approaches have good asymptotic properties under the additional assumptions of (*i*) smoothness of the estimated functions and (*ii*) regularity of the underlying probabilistic model.

### 3.1 Regression/Imputation

An estimation method is motivated by noting that the last term in equation (3) represents the average treatment effect in terms of functions that can be estimated from observable data. This suggests estimating $\beta$ by:

$$\hat{\beta}_{imp} = \frac{1}{N}\sum_{i=1}^{N} \hat{m}_1(X_i) - \hat{m}_0(X_i) \tag{11}$$

where $\hat{m}_t(.)$ is an estimate for $m_t(.)$. Thus, the procedure involves two steps: the first estimates the regression functions, while the second one integrates the difference between these estimates over the empirical distribution of $X_i$. Intuitively, this estimator replaces the difference between the potential outcomes – which would require unavailable counterfactual data – with the difference between the values of the regression functions, which can be estimated and imputed.

The imputation method was implemented by Rubin (1977), in a setup where $m_1(.)$ and $m_0(.)$ are known to belong to a given parametric family. More recently,

[8] In the review of Imbens and Wooldridge (2009), the major methods are classified into three groups: regression, propensity score-based methods, and matching. In this paper, we deal with regression and reweighting, which belongs to the second group. Matching methods, in general inefficient, will not be addressed here as they are not closely related to the method proposed herein.

imputation has been considered in the setting discussed in this paper, where the functional forms of the regression functions are unknown. Heckman et al. (1997) and Heckman et al. (1998) worked out this case, utilizing kernel methods for the first stage.

Another alternative approach consists in obtaining $\hat{m_t}$ by a sieve method (Grenander, 1981, Gemam and Huang, 1982), which estimates increasingly flexible parametric models as sample size increases. An example of this is the estimation by sieve/least squares, in which an increasing sequence of bases of functions $B_K = \{q_k(X), k = 1, 2, ..., K\}$ is chosen, and, for each sample size $n$, we obtain the estimate $\hat{m_t}(X_i) = q^{k(n)}(X_i)'\hat{\gamma}_{k(n),t}$by orthogonal projection of $Y$ onto the span of $q_1(X), ..., q_{k(n)}(X)$, considering the subsample $T_i = t$. In other words, introducing the notation $(X_i^1, Y_i^1)_{i=1}^{N_1}$ for the treated subsample, and $(X_i^0, Y_i^0)_{i=1}^{N_0}$ for the control subsample,

$$
\begin{aligned}
q^k(X) &= (q_1(X), ..., q_k(X)) \\
\hat{\gamma}_{k,t} &= \left(Q^{k,t\prime k,t}\right)^{-1} Q^{k,t\prime} Y^t \\
Q^{k,t} &= (q^k(X_1^t)'^k(X_{n_t}^t)')
\end{aligned}
$$

A necessary condition for the convergence in probability of $\hat{m_t}$ is that $m_t(.)$ may be approximated arbitrarily well by some element of the span of $B_K$, for sufficiently large $K$, and dimension $k(n)$ of the projection space increases arbitrarily with, but less than proportionately to, the sample size.

This technique was adopted by Chen et al. (2005) using a setting similar to ours.[9] Imbens et al. (2005) investigated a version of the same estimator, with $B_K$ chosen as being polynomial bases, in the context of average treatment effect under ignorability.

### 3.2 Reweighting

As an alternative to regression/imputation, a group of methods was developed based on the work of Rosenbaum and Rubin (1983). Those authors showed that the propensity score contains all the necessary information about the difference between the treatment and control groups. Formally, Rosenbaum and Rubin demonstrated that, under ignorability:

$$T \perp (Y(0), Y(1)) \mid p(X) \tag{12}$$

[9] Chen et al. (2005) analyze the problem with the estimation of the solution to a moment condition using latent variables, $X^*$, observed only in an auxiliary sample. The identification hypothesis that the distribution of $X^*$ conditional on observed variables $X$ does not depend on whether the observation is about the auxiliary sample, and is analogous to the ignorability assumption, where $X^* = Y(t)$ and the units that receive $T = t$ correspond to the auxiliary sample in the estimation of $E[Y(t)]$.

and that, therefore, there is no bias in comparing units with the same propensity score. This observation prompted the search for estimators based on this function and avoided using imputation methods that, due to the high-dimensionality of their first step estimations, often had computational difficulties and poor performance in small samples.

Of particular interest for our study is the method that uses the propensity score to reweight the observations, as presented in the second identification equation (5). Reweighting can be interpreted as a way to even out the distribution of $X$ between treated and control groups.

Reweighting procedures were initially utilized in contexts with known propensity score. Horvitz and Thompson (1952), often regarded as the forerunners of this technique, used weighting with inverse probability of selection in stratified samples.

In the case of average treatment effects with observational data, the propensity score is generally unknown, and should therefore be estimated. By making an analogy with equation (5), a reweighting estimator takes the form:

$$\hat{\beta}_{rew} = \frac{1}{N}\sum_{i=1}^{N} \frac{T_i Y_i}{\hat{p}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(X_i)} \tag{13}$$

or

$$\tilde{\beta}_{rew} = \frac{\sum_{i=1}^{N} T_i Y_i/\hat{p}(X_i)}{\sum_{i=1}^{N} T_i/\hat{p}(X_i)} - \frac{\sum_{i=1}^{N}(1-T_i)Y_i/(1-\hat{p}(X_i))}{\sum_{i=1}^{N}(1-T_i)/(1-\hat{p}(X_i))} \tag{14}$$

where $\hat{p}(.)$ is an estimate of the propensity score. The difference between the two estimators is that $\tilde{\beta}_{rew}$ requires that the sum of the weights be equal to one. If normalization of the weights is included in the first step, based on how $\hat{p}(.)$ is estimated, we have $\hat{\beta}_{rew} = \tilde{\beta}_{rew}$.

Since the propensity score is actually a conditional mean (of the treatment indicator $T$), it could be estimated by sieve/least squares methods. This is similar to the regression functions in the imputation estimator. The properties of the weighting estimator with this specification were assessed by Hirano et al. (2000), using polynomial bases and, later on by Chen et al. (2008), in a more general setting.

However, the least squares method allows the estimated values $\hat{p}(X_i)$ not to belong to the unit interval, thereby allowing for negative weights.[10] Because of this disadvantage, Hirano et al. (2003) propose estimating $p(.)$ by an alternative sieve method. They replace least squares projection by a maximum likelihood estimation of the logistic regression of $T$ on the functions of basis $B_{k(n)}$, i.e., $\hat{p}(X_i) = L(q^{k(n)}(X_i)'\gamma_{k(n)})$, where $L(.)$ is the logistic function and

[10] A similar criticism is leveled at the estimator in the first step of imputation when $Y$ is a limited variable.

$$\hat{\gamma_k} = \arg\max_{\gamma \in \mathbb{R}^k} \sum_{i=1}^{N} T_i \log(L(q^k(X_i)'\gamma)) + (1 - T_i)(1 - \log(L(q^k(X_i)'\gamma))$$

An interesting fact observed in several studies on the topic (see, for instance, Hahn, 1998, Hirano et al., 2003), is that substituting the true propensity score $p(.)$ for the estimated one $\hat{p}(.)$ does not yield an efficient estimator of average treatment effects. On the other hand, the necessity to obtain these estimates makes reweighting as complex as imputation, unlike what had been initially suggested for the study of propensity score-based methods.

## 3.3 Asymptotic properties

Both approaches presented in the previous sections make use of two-step procedures: a nonparametric estimation and the solution to a moment condition based on this first estimate. Estimators like this are of particular interest amongst the available semiparametric methods, and several studies propose general frameworks for their analysis, such as those conducted by Bickel et al. (1993), Newey (1994), Newey and McFadden (1994), and Chen et al. (2003).

In the terminology of this literature, the function estimated in the first step is called the nuisance function. A fundamental issue concerns how the estimation of the parameter of interest is affected by the lack of knowledge about the nuisance function, using as reference the method of moments estimator obtained if that function was known.

Following established results about the method of moments, consistency of the estimators is implied by the convergence of the sample moment condition, uniformly with respect to the parameter of interest, to the population moment condition. In other words, in the case of imputation, $\hat{\beta} \xrightarrow{p} \beta$ if

$$\begin{aligned} & \sup_{\beta} \left[ \frac{1}{N} \sum_{i=1}^{N} (\hat{m}_1(X_i) - \hat{m}_0(X_i)) - \beta - (E\left[m_1(X) - m_0(X)\right] - \beta) \right] \\ = \; & \frac{1}{N} \sum_{i=1}^{N} (\hat{m}_1(X_i) - \hat{m}_0(X_i)) - E\left[m_1(X) - m_0(X)\right] \xrightarrow{p} 0 \qquad (15) \end{aligned}$$

and, in the case of reweighting, if

$$\frac{1}{N} \sum_{i=1}^{N} \frac{T_i Y_i}{\hat{p}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(X_i)} - E\left[\frac{TY}{p(X)} - \frac{(1-T)Y}{1-p(X)}\right] \xrightarrow{p} 0 \qquad (16)$$

A way to prove equations (15) and (16) is by observing that the contribution of each unit to the sample moments $\hat{m}_1(X_i) - \hat{m}_0(X_i)$ and $\frac{T_i Y_i}{\hat{p}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(X_i)}$

are continuous with respect to the nuisance functions $(\hat{m_0}(.), \hat{m_1}(.))$ and $\hat{p}(.)$, if we endow these parameters' spaces with the supremum norm. Therefore, under conditions that guarantee convergence (in supremum norm) of estimated nuisance functions to their true population values, application of the uniform law of large numbers verifies these limits. Newey (1994, Lemma 5.2) suggests this strategy, which is used by Hirano et al. (2000) and Chen et al. (2005), for reweighting and imputation, respectively.

The study of the properties related to $\sqrt{N}$-consistency is facilitated by Newey's (1994) observation that, when the semiparametric model is sufficiently large (in a sense that we will explain shortly), any regular asymptotically linear (RAL) estimator turns out to be efficient. This result stems from another one (Newey, 1990): the influence function of a RAL estimator, $\psi$, solves the pathwise differentiability equation (6) for $d$. Thus, using the notation in subsection 2.5 for any regular parametric submodel,

$$E[(\psi - \delta)S'_\theta] = \frac{\partial \beta(\theta_0)}{\partial \theta} - \frac{\partial \beta(\theta_0)}{\partial \theta} = 0 \tag{17}$$

i.e., the difference $\psi - \delta$ is orthogonal to any score. The semiparametric model is sufficiently large if the tangent space contains all the zero mean functions of the data. When this condition holds, $\psi - \delta$ simultaneously belongs and is orthogonal to the tangent space; so, it must be identical to zero.

In fact, the result by Scharfstein et al. (1999) discussed in subsection 2.4 implies – as is pointed out by the authors themselves[11] – that the model which only assumes ignorability is large in the sense just described. Hence, regardless of the approach (reweighting or imputation), under regularity conditions, the influence function of an asymptotically linear estimator of $\beta$ will be the efficient one. In view of that, it is interesting to write this influence function in the following two ways:

$$\begin{aligned}\psi^* &= m_1(X) - m_0(X) - \beta + \frac{T}{p(X)}(Y - m_1(X)) - \frac{1-T}{1-p(X)}(Y - m_0(X)) \\ &= \frac{T}{p(X)}Y - \frac{1-T}{1-p(X)}Y - \beta - (T - p(X))(\frac{m_1(X)}{p(X)} + \frac{m_0(X)}{1-p(X)})\end{aligned}$$

Each of the representations contains two parts. The first component represents the influence function of imputation and reweighting estimators that would be obtained were the true values of the regression function and the propensity score used, instead of their estimates. The remaining terms can be viewed as a correction for the estimation of these functions. In the case of estimation by reweighting, the correction term is negatively correlated with the first part; hence, there are

[11] Scharfstein et al. (1999, p. 1118).

efficiency gains from using an estimate of the propensity score, even when this function is known.

Thus, the properties of the asymptotic distribution depend on the convergence of the 'residuals'

$$\frac{1}{N}\sum_{i=1}^{N}\hat{m_1}(X_i) - m_1(X_i) - (\hat{m_0}(X_i) - m_0(X_i))$$

from the estimation of $m_t(.)$ in regression, and

$$\frac{1}{N}\sum_{i=1}^{N}\frac{T_iY_i}{\hat{p}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(X_i)} - \left(\frac{T_iY_i}{p(X_i)} - \frac{(1-T_i)Y_i}{1-p(X_i)}\right)$$

from the estimation of $p(.)$ in reweighting, for the respective correction terms at a rate faster than $1/\sqrt{N}$. This condition imposes additional constraints on the rate at which the nuisance function estimators should converge. In the case of estimation by reweighting, a consequence of the inefficiency of using the true propensity score is that the convergence of its estimate should, on the other hand, be slow enough in order to yield the remainder of order $1/\sqrt{N}$ shown above.[12] However, by simply incorporating the information as a constraint in the first step, the desired convergence rate for the remainder term would not be attained. For example, if we know the propensity score and use this value instead of performing the first step, the remainder would be zero.

The convergence rate that a nonparametric method for the estimation of a function can achieve depends on the smoothness of this function. For sieve/least squares, for example, Newey (1997) shows that the maximum rates of uniform and squared mean convergence are decreasing in the dimension of regressors' set and increasing with the number of continuous derivatives of the estimated function. Therefore, the larger the set of auxiliary variables, the greater is the smoothness demanded from the nuisance functions.

For their reweighting estimator, Hirano et al. (2003) show asymptotic normality and efficiency, assuming the propensity score to be $7r$ times differentiable, where $r$ is the number of the auxiliary variables. Chen et al. (2005) demonstrate the same asymptotic properties for the imputation estimator, provided that $m_t(.)$ belong to certain Hölder spaces[13] and $p(.)/(1-p(.))$ is smooth enough to admit certain approximations in the sieve spaces.

[12] Chen et al. (2008) discuss how to use that information without losing asymptotic efficiency.

[13] A function $f(.)$ belongs to the Hölder space $(\gamma, p)$ if it contains derivatives up to order $\lfloor\gamma\rfloor$ (largest integer less than or equal to $\gamma$, its whole part), in which highest-order derivatives are Hölder-contínuous functions with exponent $\alpha = (\gamma - \lfloor\gamma\rfloor)$, i.e., with limited $\frac{\|f(x)-f(y)\|}{\|x-y\|^{\alpha}}$.

## 4. Combining Imputation and Reweighting

The preceding section showed that imputation and reweighting can yield consistent and asymptotically efficient estimators of the average treatment effect. Although these properties provide an important argument in favor of these methods, some remarks are necessary.

Because of their asymptotic character, consistency and efficiency do not allow for immediate conclusions about the performance for a given amount of observations. Moreover, as was pointed out in the discussion about efficiency bound, regular estimators do not require uniform convergence in the distribution. Thus, in principle, the concept of 'sufficiently large sample' depends on unknown values and, therefore, it is not possible to verify whether a given sample satisfies such a requirement. Confirming this theoretical reasoning against the exclusive recommendation of asymptotic optimality, simulation studies (discussed in the next section) present distinct results for the performance of several 'efficient' and asymptotically equivalent methods.

In practice, a combination of imputation and reweighting methods may be preferable and it is typically used to reduce bias (Imbens and Wooldridge, 2009). A way to implement that is suggested by double robust estimation methods, developed by James Robins and co-authors, discussed in what follows.

### 4.1 Double Robust Estimation

Methods that combine the estimation of both the probability of selection and regression functions were initially proposed for the case of parametric models by Robins and Rotnitzky (1995) and Robins et al. (1995) in their studies on inference based on missing data. In this context, the purpose is to protect against the misspecification of these models. An example of this is the 'augmented' reweighting estimator, discussed by Scharfstein et al. (1999):

$$\begin{aligned}\hat{\beta}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_iY_i}{p(X_i;\hat{\gamma})} - \frac{(1-T_i)Y_i}{1-p(X_i;\hat{\gamma})} \\ - (T_i - p(X_i;\hat{\gamma}))\left(\frac{m_1(X_i;\hat{\delta_1})}{p(X_i;\hat{\gamma})} + \frac{m_0(X_i;\hat{\delta_0})}{1-p(X_i;\hat{\gamma})}\right) \qquad (18)\end{aligned}$$

where $p(.;\hat{\gamma})$ and $m_t(.;\hat{\delta_t})$ are, respectively, estimates for $p(.)$ and $m_t(.)$ based on parametric models $\{p(.;\gamma);\gamma\in\Gamma\}$ and $\{m_t(.;\delta_t);\delta\in\Delta\}$, which must satisfy regularity conditions.

The similarity between equation (18) and the efficient influence function presented in Section 2 is quite evident. In fact, equation (18) is a sample analog of the equality $E[\psi^*]=0$. The additional term is similar to the correction for the estimation of $p(.)$, and serves to make $\hat{\beta}_{DR}$ consistent provided that at least one of the parametric models is well specified, i.e., that it contains a true value. We can check that by first observing that the population equation

$$\beta = E\left[\frac{TY}{p(X)} - \frac{(1-T)Y}{1-p(X)}\right] - E\left[(T-p(X))(\frac{\bar{m}_1(X)}{p(X)} + \frac{\bar{m}_0(X)}{1-p(X)})\right] \tag{19}$$

analogous to equation (18), is satisfied for the true values of $\beta$ and $p(.)$, regardless of functions $\bar{m}_t(.)$. In this case, the second expectation on the right-hand side is zero, since $T - p(X) = T - E[T|X]$ is the residual of the orthogonal projection of $T$ onto the space of the functions of $X$. Therefore, equation (19) is identical to equation (5) when $p(.)$ assumes its population value. The consistency of $\hat{\beta}_{DR}$ follows, under regularity conditions, from the usual arguments of the theory of estimation by the method of moments.

To check consistency when $m_0(.)$ and $m_1(.)$ are correctly estimated, it is useful to rewrite equation (19) as follows:

$$\beta = E\left[m_1(X) - m_0(X)\right] - E\left[\frac{T}{\bar{p}(X)}(Y - m_1(X)) + \frac{1-T}{1-\bar{p}(X)}(Y - m_0(X))\right] \tag{20}$$

If $m_0$ and $m_1$ are the true regression functions, the second term on the right-hand side of equation (20) is zero for any function $\bar{p}(.)$, as

$$\begin{aligned} E\left[\frac{T}{\bar{p}(X)}(Y - m_1(X))|X\right] &= \frac{1}{\bar{p}(X)}E\left[E\left[T(Y - m_1(X))|X,T\right]|X\right] \\ &= \frac{1}{\bar{p}(X)}E[E[Y_1 - m_1(X)|X, T=1]Pr(T=1|X)|X] \\ &= 0 \end{aligned}$$

and, analogously,

$$E\left[\frac{1-T}{1-\hat{p}(X)}(Y - m_0(X))|X\right] = 0$$

Thus equation (20) corresponds to equation (3) and, as in the previous case, well known results establish the consistency of $\hat{\beta}_{DR}$. Estimators like this, which satisfy consistency even when one out of two parametric models fails, are called double robust models. Besides $\hat{\beta}_{DR}$, other estimators with this property are known, including those introduced by Robins et al. (2007) and Egel et al. (2012).

### 4.2 Relevance for semiparametric estimation

Clearly, a double robust procedure is preferable to alternatives that rely on a single, unreliable parametric model. This method does not show, however, how regression and imputation can be combined to produce estimators with desirable properties when there are no reliable parametric models. Furthermore, the semiparametric imputation and reweighting methods described in the previous section allow for consistent and efficient estimation by imposing no other restrictions than the identification hypothesis and regularity conditions.

A reason for combining approaches is that, depending on the sample size available, a nonparametric estimator is often biased. In particular, in sieve estimation, for any given sample size one actually estimates a misspecified parametric model. Therefore, the discussion about double robustness suggests that including a sample analog of the correction term might reduce the bias. This argument is outlined in the simulation study by Bang and Robins (2005), who suggest the validity of an approximate version of double robustness: if at least one of the models is approximately correct, the estimation bias is small. It is also interesting to note that recently studied methods for optimal implementation of imputation (Imbens et al., 2005) and reweighting estimators (Ichimura and Linton, 2005) use both the propensity score and the regression functions to estimate the bias and variance.

Formally, for the case of missing data with binary $Y$, Robins and Ritov (1997) provide a reason to use an estimator analogous to $\hat{\beta}_{DR}$ under a semiparametric specification. According to the authors, a problem with the asymptotic results of semiparametric estimators is the dependence on smoothness assumptions on the observation probability /propensity score and regression function. Robins and Ritov show that, by assuming only that these functions are measurable, there are no estimators that converge to algebraic rates, i.e, of the form $n^{-\alpha}$, $\alpha > 0$. In demonstrating this fact, alternative models are used, which differ from the true population model due to a large number of irregularities in $p(X)$ and $m(X)$ introduced within small intervals of the distribution of $X$. The construction of these regions restricts the probability of each one of them having more than one point in the sample.

This idea is compared to the performance problems in a finite sample. As the dimension of covariate set $X$ increases, the probability of observations having similar $X$ values decreases quickly. Therefore, important local behaviors of selection and/or regression models are difficult to detect, eventually resulting in substantial bias. Based on this discussion of the weaknesses of asymptotic properties, Robins and Ritov argue that caution is needed when the nonparametric step is susceptible to this 'curse of dimensionality'. They then develop the 'curse of dimensionality appropriate asymptotic theory', which requires the researcher to subjectively identify those nuisance functions that are potentially misspecified, due to either low smoothness or high dimensionality. This theory suggests estimating the parameter of interest from moment conditions, which are unbiased under misspecification of the susceptible nuisance functions.

For the missing data model, the equation that sets the expected value of the influence function to zero is appropriate for the case of 'well-specified selection'/'misspecified regression' as well as for the opposite situation, the case of 'well-specified regression'/'misspecified selection'. Robins and Ritov show that it is possible to obtain a RAL estimator in both cases. Additionally, in the first case, the estimator is also uniformly asymptotically normal and unbiased (UANU). The latter property is defined by the existence of a sequence $s_n(F)$, such that

$$\sup_F |Pr_F\left[N^{1/2}(\hat{\beta} - \beta(F))/s_N(F) < t\right] - \phi(t)| \to 0$$

where $F$ indexes the distributions that are admissible by the model. This condition is weaker than requiring the estimator to be uniformly regular Gaussian, but it suffices to build uniformly asymptotically valid confidence intervals, where $s_N$ is estimated by bootstrapping. The estimator corresponds to equation (18), for $Y(0) \equiv 0$, but to (nonparametric) histogram-type estimators for $p(.)$ and $m(.) = m_1(.)$, where each of first-step estimates and the final step are obtained from independent subsamples.

### 4.3 Semiparametric double robust estimators

Based on the discussion above, we propose combinations of imputation and reweighting techniques in estimators that:

(i) are based on moment condition $E[\psi^*] = 0$ which, as was shown in subsection 4.1, is verified by either $m_1(.)$ and $m_0(.)$, or $p(.)$ taking their true values; and, at the same time,

(ii) employ the nonparametric estimation of these functions. By using these combinations, we obtain better performance in finite samples, given the smaller vulnerability to nuisance functions.

The first suggestion we analyze is the use of nonparametric estimators in an augmented reweighting procedure such as in $\hat{\beta}_{DR}$:

$$\begin{aligned}\hat{\beta}_{IF} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_iY_i}{\hat{p}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(X_i)} \\ - (T_i - \hat{p}(X_i))\left(\frac{\hat{m}_1(X_i)}{\hat{p}(X_i)} + \frac{\hat{m}_0(X_i)}{1-\hat{p}(X_i)}\right) \qquad (21)\end{aligned}$$

where $\hat{p}(.)$ and $\hat{m}_t(.)$ are estimates obtained from sieve/logit and sieve/least squares, respectively. Although this procedure was suggested in some studies, it had not been explicitly investigated until recently. Cattaneo (2010) assesses the properties of an estimator analogous to $\hat{\beta}_{IF}$ in the context of a multilevel average treatment effect. Rothe and Firpo (2013) present a more general treatment of this issue.

An interesting fact about $\hat{\beta}_{IF}$ is that it can be represented as the linear combination of three well-known estimators. Rewriting equation (21) in a more convenient way, we get:

$$\begin{aligned}\hat{\beta}_{IF} =: & \frac{1}{N}\sum_{i=1}^{N}\frac{T_iY_i}{\hat{p}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{p}(X_i)} \\ + & \frac{1}{N}\sum_{i=1}^{N}\hat{m}_1(X_i) - \hat{m}_0(X_i) \\ - & \frac{1}{N}\sum_{i=1}^{N}\frac{T_i\hat{m}_1(X_i)}{\hat{p}(X_i)} - \frac{(1-T_i)\hat{m}_0(X_i)}{1-\hat{p}(X_i)}\end{aligned}$$

which corresponds exactly to the sum of imputation and reweighting estimators subtracted from the 'modified' estimator proposed by Imbens et al. (2005). This representation allows the immediate establishment of conditions for consistency and asymptotic efficiency. Under differentiability assumptions, Imbens et al. (2005) demonstrate that the three estimators are asymptotically linear with an efficient influence function. With these conditions, therefore:

$$\begin{aligned}\sqrt{N}(\hat{\beta}_{IF} - \beta) &= \sqrt{N}(\hat{\beta}_{rew} + \hat{\beta}_{imp} - \hat{\beta}_m - \beta) \\ &= \sqrt{N}(\hat{\beta}_{rew} - \beta) + \sqrt{N}(\hat{\beta}_{imp} - \beta) - \sqrt{N}(\hat{\beta}_m - \beta) \\ &= N^{-1/2}\sum_{i=1}^{N}\psi^*(Z_i) + N^{-1/2}\sum_{i=1}^{N}\psi^*(Z_i) - N^{-1/2}\sum_{i=1}^{N}\psi^*(Z_i) \\ &+ o_p(1) \\ &= N^{-1/2}\sum_{i=1}^{N}\psi^*(Z_i) + o_p(1)\end{aligned}$$

and, thus, $\hat{\beta}_{IF}$ is asymptotically efficient as well. Hence, we conclude that, in the semiparametric context accounted for in this paper, double robust estimators are not necessarily inefficient. This contrasts with parametric settings, where robustness against the violation of specification assumptions implies inefficiency if there is no misspecification.[14] Cattaneo (2010) presents another proof of efficiency for this

[14] It should be noted that here we refer to distinct efficiency bounds. A double robust estimator based on certain parametric specifications for $m_t(.)$ and $p(.)$ does not reach the efficiency bound of the model for which they are valid, which is different from the semiparametric efficiency bound. However, there are estimators, like the one developed by Egel et al. (2012), that reach the semiparametric efficiency bound when both specifications are correct. These estimators are known as locally efficient.

estimator and observes that the necessary conditions are weaker, since the similarity between the estimated moment condition and the efficient influence function implies that the estimation of the nuisance functions does not have to generate a remainder of order $1/\sqrt{N}$ seen in subsection 3.3. Therefore, it is interesting to observe that double robust procedures can provide a simple way to take advantage of additional information on the propensity score.

Another way to obtain a double robust procedure is by the adequate selection of the preliminary estimator in an imputation procedure, as was pointed out by Robins et al. (2007). In this sense, we note that an estimator derived from condition $E[\psi^*] = 0$ takes the form

$$\begin{aligned}\hat{\beta} =: &\frac{1}{N}\sum_{i=1}^{N} \hat{m}_1(X_i) - \hat{m}_0(X_i) \\ &- \left(\frac{T_i(Y_i - \hat{m}_1(X_i))}{\hat{p}(X_i)} - \frac{(1-T_i)(Y_i - \hat{m}_0(X_i))}{1-\hat{p}(X_i)}\right)\end{aligned}$$

Note that $\hat{\beta}$ differs from a regression estimator because of the term in the second line, corresponding to the correction arising from the use of the estimated nuisance functions $\hat{m}_t$. Note also that this term consists of products of the weights used in the reweighting procedure with the residuals from the regression function estimation. Thus, if estimation method for the regression functions is designed to yield residuals that are orthogonal to the weights, then the correction term vanishes and $\hat{\beta}$ will be identical to the regression estimator.

This can be obtained by the inclusion of weights as regressors. In this case, normal equations directly lead to the required orthogonality. However, this procedure is not recommended, because the distribution of weights, by construction, is concentrated at different points between the groups. For example, $\hat{p}(X_i)$ will assume higher values in the treatment group and possibly very low values in the control group. As a result, the inclusion of regressor $1/\hat{p}(X_i)$ may produce a severe case of extrapolation.

The correction term can also be eliminated if we replace $\hat{m}_1(.)$ and $\hat{m}_0(.)$ with the sieve/weighted least squares estimators $\tilde{m}_1(.;\hat{p}(.))$ and $\tilde{m}_0(.;\hat{p}(.))$, which use, respectively, $1/\sqrt{\hat{p}(X_i)}$ and $1/\sqrt{1-\hat{p}(X_i)}$ as weights. If the chosen basis of the functions contains a constant, the normal equation for the corresponding coefficient is, for the calculation of $\tilde{m}_1(.;\hat{p}(.))$:

$$-\frac{2}{N_1}\sum_{i=1}^{N_1}\frac{1}{\hat{p}(X_i)}(Y_i - \tilde{m}_1(X_i;\hat{p}(.))) = 0$$

or

$$\sum_{i=1}^{N} \frac{T_i}{\hat{p}(X_i)}(Y_i - \tilde{m_1}(X_i;\hat{p}(.))) = 0$$

with an analogous equation holding for $\tilde{m_0}(.;\hat{p}(.))$. In this case, we define the weighted regression estimator:

$$\hat{\beta}_{WR} = \frac{1}{N}\sum_{i=1}^{N} \tilde{m_1}(X_i;\hat{p}(.)) - \tilde{m_0}(X_i;\hat{p}(.)) \tag{22}$$

where $\tilde{m_t}(.;\hat{p}(.))$ are obtained by the coefficients of the weighted least squares projection of $Y$ onto functions of $X$ in the subsamples, using the above-mentioned weights.

This estimator was implemented by Hirano and Imbens (2001),[15] who also proposed the following algorithm for the selection of preliminary estimators. First, the researcher establishes a large set of variables (possibly including transformations and interactions) and a pair of positive real values $(t_{prop}, t_{reg})$. Then, simple logistic regressions of the treatment indicator on each variable are performed. Following this, the null hypothesis that the explanatory variable has no effect is tested for each regression. For the estimation of $p(.)$, one should use the variables whose absolute value of the $t$-statistic is greater than $t_{prop}$. Similarly, each variable of the large set is used in a separate regression of $Y_i$ on $T_i$. Those whose $t$-statistics have an absolute value greater than $t_{reg}$ are included in the estimations of $m_1(.)$ and $m_0(.)$. Hirano and Imbens remark that this procedure includes imputation and reweighting estimators as particular cases when only constants are utilized for the estimation of $p(.)$ and $m_t(.)$, respectively.

The estimator $\hat{\beta}_{WR}$ also allows a new interpretation of the double robustness property, based on the idea of omitted variable bias in the regression analysis (as noted by Imbens and Wooldridge, 2009). Imputation is equivalent to regressing $Y$ on $T$, with $X$ (and interactions between $X$ and $T$) as controls. A reweighting estimator, on the other hand, is equivalent to regressing $Y$ on $T$ alone, after reweighting the sample in a way that eliminates correlation between $T$ and $X$, and thus the omission of the latter no longer causes bias. If the correct specification of the effect of controls $X$ is not perfectly known, then the double robust estimator, which uses both forms of correction, reduces bias.

## 5. Monte Carlo Simulation

The variety of alternative methods for the estimation of treatment effects under ignorability and of possible implementations of each one has encouraged a broad range of simulation studies in recent years, comparing the performance of these

[15] Previously, Robins et al. (1995) had considered a similar estimator, using nonparametric estimation of the propensity score, but a with parametric specification for the regression functions.

methods in finite samples. In this section, we review the simulation literature on combined methods similar to the ones proposed in subsection 4.3. Next, we carry out a new simulation study.

### 5.1 Simulation literature results for the combination of methods

Simulation studies on combined methods have been motivated by theoretical studies about these estimators and, as result, they are quite recent. A considerable strand of this literature was developed by the same authors who propose the use of double robust methods (Bang and Robins, 2005, Robins et al., 2007, Neugebauer and der Laan, 2005), as a way disseminate these estimators. However, the results published to date are clearly distinct from each other.

Lunceford and Davidian (2004) assess the finite sample behavior of several propensity score-based estimators. Their study focuses on the comparison between reweighting and sample stratification methods, finding that the former method performs better. An interesting point for our discussion is that both reweighting and stratification are shown to be improved when adjusted by a regression. The final recommendation is the use of a double robust estimator in the form of augmented reweighting as proposed by Scharfstein et al. (1999).

Bang and Robins (2005) introduce extensions to the double robust estimation for different contexts involving missing data. In the simulations, special attention is given to the possibility of misspecification of the models used to estimate the observation probability and the potential values. The major result is that double robust estimators may have good performance, even when both specifications are incorrect.

In a similar work, Kang and Schaeffer (2007) discuss double robust methods and compare them with the separate use of regression and reweighting. The authors also highlight the behavior of different procedures when selection and regression models are mildly misspecified. The results obtained suggest the sensitivity of the propensity score-based methods (i.e. reweighting and double robust) to the misspecification of this function. Unlike in the results of Bang and Robins (2005), the simple regression models perform better, whereas double robust estimators outperform the reweighting estimators in some cases. In response to the previously mentioned study, Robins et al. (2007) challenge its results, arguing that they are rather dependent on the choice of population model, which is particularly favorable to regression estimators. Robins and his coauthors point out that a symmetric setup, which was obtained from Kang and Schaeffer's model by switching the $T$ indicator, provides very different results.

Busso et al. (2013) compare reweighting, matching, and double robust estimators, the latter of which is implemented by weighted regression. In their simulations, reweighting estimators yield the best results, except when the overlap assumption is violated or when the propensity score reaches values close to zero or one, i.e. when overlap almost fails. When this occurs, all methods produce considerable bias and predictions made by the asymptotic theory fail. Busso et al. (2013) then relate the performance of propensity score-based estimators to validity of overlap, thus explaining some of the apparently conflicting results in simulation literature and highlighting the importance of this assumption.

## 5.2 A new simulation study

The following simulation study aims to assess the interaction between the elements of the population model and the relative performance of the combination of reweighting and imputation, when compared to the separate use of each of these methods. Additionally, we examine how characteristics of the true model relate to the optimal implementation of double robust methods. In this regard, four main elements are analyzed.

First, we try to underline the effect of the functional form of regression functions and the propensity score, especially with respect to smoothness. As was discussed in the previous section, the reason for using the inverse propensity score as a weight in a regression is that it eliminates the correlation between the regressors and the omitted variables. Misspecification can be included in this context if it is interpreted as the omission of some part of $m_t(X)$. Therefore, projecting rough regression functions onto a small basis of auxiliary variables should result in considerable error, and thus may be associated to important gains from previously reweighting the sample.

The second element we investigate is the effect of heteroskedasticity on the convenience of running weighted regressions. Since reweighting of the regression in $\hat{\beta}_{WR}$ follows a criterion that is totally independent from the precision of this first step (which in turn should take into account only the heteroskedasticity pattern), the effect on the efficiency of the treatment effect estimator might be adverse.

Thirdly, we take into account the relevance of the number of dimensions of auxiliary variables' space. The discussion by Robins and Ritov (1997) suggests that 'the curse of dimensionality' is more likely to impair the performance of non-robust estimators. On the other hand, the ignorability assumption requires that the econometrician use all the available information to control for heterogeneity, which is likely to result in a large number of auxiliary variables. In spite of that, most of the simulations available in the literature consider univariate $X$. Of the studies mentioned earlier, only the one conducted by Lunceford and Davidian (2004) addresses this concern.

Finally, we assess the use of the true propensity score as an element of the calculation of weights, in the weighted regression, and the correction term, in augmented reweighting. As we underscored in the previous section, a double robust estimator that uses the populational value of $p(.)$, instead of an estimate, can achieve the semiparametric efficiency bound. This suggests that the available information about propensity score can be readily incorporated in the first step, easing the computational burden of the procedure without influencing the asymptotic variance.

### 5.3 Estimators and models

We examine the performance of reweighting ($\hat{\beta}_{rew}$), imputation ($\hat{\beta}_{imp}$) and double robust methods of types $\hat{\beta}_{IF}$ (augmented reweighting or influence function) and $\hat{\beta}_{WR}$ (weighted regression). The preliminary estimates of $p(.)$ are obtained by sieve/logit for all estimators, while regression functions $m_t(.)$, $t = 0, 1$ are estimated by sieve/ordinary least squares for imputation and $\hat{\beta}_{IF}$. Estimates of $m_t(.)$ by sieve/weighted least squares (with the weights described in subsection 4.3) are utilized for $\hat{\beta}_{WR}$. In addition, we also calculate estimators $\tilde{\beta}_{IF}$ and $\tilde{\beta}_{WR}$, analogous to $\hat{\beta}_{IF}$ and $\hat{\beta}_{WR}$, but combining the true propensity score and the estimated regression functions.

As a way to analyze the selection of the number of bases' terms for the sieve estimators, we take into account the whole interval of values $L = 1, 2, 3, ...\bar{L}$ for the estimation of $m_t(.)$ and $K = 1, 2, 3, ...\bar{K}$ for the estimation of $p(.)$. The bases consist of polynomials of the auxiliary variables in the following order:

(i) increasing degree,

(ii) decreasing largest exponent, and

(iii) decreasing exponent of the first variable, then decreasing exponent of the second variable and so on.

As noted in the previous section, regression and reweighting estimators are particular cases of double robust estimators, respectively, when the propensity score and regression functions are estimated by a constant. Therefore, the calculation of $\hat{\beta}_{IF}$ and $\hat{\beta}_{WR}$ includes that of $\hat{\beta}_{rew}$, when $L = 1$ and $\hat{\beta}_{imp}$, when $K = 1$. Moreover, for $L = 1$, $\tilde{\beta}_{FI}$ and $\tilde{\beta}_{RP}$ are identical to the (inefficient) reweighting estimator by the populational propensity score.

To highlight the effect of the smoothness of $p(.)$ and $m_t(.)$, we initially consider a model with only one auxiliary variable $X_i$, uniformly distributed on the interval $[-1, 1]$, and

$$\begin{aligned}
Y_i(1) &= m_1(X_i) + u_{1,i}\\
Y_i(0) &= m_0(X_i) + u_{0,i}\\
Y_i &= T_i(m_1(X_i) + u_{1,i}) + (1 - T_i)(m_0(X_i) + u_{0,i})
\end{aligned}$$

with homoskedastic $Y_i$, i.e., $u_i = (u_{0,i}, u_{1,i})$ is independent and identically distributed according to a bivariate normal distribution with zero mean and variance $\sigma^2$. The variation in the specifications of $p(.)$ and $m_t(.)$ is obtained by manipulating the coefficients of the following representations:

$$
\begin{aligned}
log\left(\frac{p(x)}{1-p(x)}\right) &= \sum_{k=0}^{\infty} a_k\lambda_k(x) \\
m_1(x)-m_0(x) &= \sum_{k=0}^{\infty} b_k\lambda_k(x) \\
m_0(x) &= \sum_{k=0}^{\infty} c_k\lambda_k(x)
\end{aligned}
$$

where $\lambda_k$ is the Legendre polynomial of order $k$. Legendre polynomials are constructed to be orthogonal with respect to the inner product $< f,g >= \int fgdx$; thus, they define uncorrelated functions of a uniformly distributed variable. The first polynomial is constant ($\lambda_0(x) \equiv 1$); so, by orthogonality, the other polynomials have zero mean, and the average treatment effect is then given by

$$\beta = E[Y_i(1) - Y_i(0)] = \sum_{k=0}^{\infty} b_k E[\lambda_k(x)] = b_0$$

The decay rate of the coefficients of the series simulates the smoothness of the corresponding functions. In fact, the role of the smoothness assumptions in the asymptotic theory of the analyzed estimators is to guarantee some convergence rates of the first-step nonparametric estimator. In the case of estimation by the sieve method, the relevant consequence of the smoothness is to guarantee the existence of a sequence of (non-random) approximations $\sum_{k=0}^{K} \tilde{a}_k^K \lambda_k(x)$ converging at a given rate (in terms of $K$) to the estimated function (Newey, 1997). This is the exact effect simulated by the manipulation of the specifications carried out in models 1 to 3.

In model 1, we used $\sigma^2 = 9.5114I_2$ ($I_2$ is the 2x2 identity matrix); $a_0 = 0.5$; $a_k = \frac{k^{-1}}{2}$, if $1 \leq k \leq 10$; $a_k = 0$, if $k > 10$; $b_0 = 1$; $b_k = 20a_k$, if $k \geq 1$; $c_k \equiv 1$. With these parameters, the propensity score is a monotonic transformation of the regression function, which intuitively seems to make the reweighting and imputation procedures redundant when combined. The model is homoskedastic, which would make the correctly specified regression estimator attractive.

In models 2 and 3, the specifications are altered by changing the coefficients. In the first case, the decay is faster: $a_k = 1.0922\frac{k^{-2}}{2}$, if $1 \leq k \leq 10$; in the second, there is no decay: $a_k = \frac{0.5926}{2}$, if $1 \leq k \leq 10$. The corresponding adjustment is made to maintain $b_k = 20a_k$, if $k \geq 1$, in each case. The constants multiplied in the redefinition of $a_k$ keep the semiparametric efficiency bound constant (equal to that of model 1).

Models 4 and 5 introduce heteroskedasticity. The specifications of model 1 are considered for $a_k$, $b_k$ and $c_k$, but $u_i$ is independently distributed with zero mean and variance $\sigma^2(X_i)$. In model 4, we define $\sigma^2(X_i) = 20.7767diag(1-p(X_i), p(X_i))$ ($diag(d_1, d_2)$ is the matrix with $d_1$ and $d_2$ on the main diagonal and zero in the other entries). This specification is favorable to the weighted regression as, for instance, treated observations with a high propensity score, and hence low weights, have a larger variance. If the weights are correctly specified, they correspond to those that optimize the regressions of $Y(t)$ in $X$ by weighted least squares. In model 5, the situation is the opposite, with $\sigma^2(X_i) = 17.5421diag(p(X_i), 1 - p(X_i))$. The constants in the expressions for $\sigma^2(X_i)$ are introduced to preserve the same semiparametric efficiency bound, as we did in models 2 and 3.

For the analysis of the effect of the dimension of $X_i$, model 6 considers three components for this variable. All of the components are identically distributed, uniformly on interval $[-1, 1]$, and independent from each other. The selection models for the treatment assignment and the conditional mean are described below:

$$\begin{aligned}
&p(x) = 0.8x_3L(x_1) + (1 - 0.8x_3)L(x_2)\\
&L(.) = \frac{\exp(.)}{1 - \exp(.)}\\
&m_1(x) - m_0(x) = 5\exp(2x_3)(1.5 + \sin(\frac{\pi}{4}(x_1 + x_2)) - (x_1 - x_2)^2)\\
&m_0(x) \equiv 0
\end{aligned}$$

The pair of the potential outcomes has a random component $u_i$ independently and identically distributed according to a bivariate normal distribution with zero mean and variance $10I_2$.

### 5.4 Results

For the simulations with one-dimensional $X$, we considered samples with $N = 100$ and $N = 200$ observations. Above these values, the estimators had indistinguishable performances. Tables 1 through 4 report the simulated mean squared errors of the estimators for the two forms used for combining the methods within model 1 specification.

It is remarkable that the best estimator, in all cases, uses more than one term to estimate the propensity score ($K > 1$) and the regression functions ($L > 1$). In contrast with the observation made in the previous section, that model 1 is particularly unfavorable, this result provides strong evidence in favor of the mix of methods. The gain from using the best combination, in relation to reweighting or imputation, is approximately 5% for $N = 100$ and 1% for $N = 200$, in terms of mean squared error.

Another interesting aspect concerns the proximity between $\hat{\beta}_{IF}$ and $\hat{\beta}_{WR}$, when the same number of terms $K$ and $L$ are used. Curiously enough, while in most cases the optimal estimator is $\hat{\beta}_{WR}$ for very high values of $K$ and $L$, $\hat{\beta}_{IF}$ is more stable.

Table 1
Model 1, Weighted Regression, N=100

| Simulated Mean Squared Errors | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 3.578 | 1.004 | 0.906 | 0.869 | 0.914 | 1.015 | 1.969 | 3.767 | > 10 |
| 2 | 0.957 | 1.047 | 0.900 | 0.864 | 0.899 | 1.003 | 1.952 | 4.420 | > 10 |
| 3 | 0.890 | 0.880 | 0.906 | 0.867 | 0.902 | 0.998 | 1.961 | 4.541 | > 10 |
| 4 | 0.889 | 0.862 | 0.868 | 0.870 | 0.902 | 1.006 | 1.914 | 4.520 | > 10 |
| 5 | 0.900 | 0.855 | 0.866 | 0.857 | 0.905 | 1.012 | 1.922 | 4.219 | > 10 |
| 6 | 0.911 | 0.855 | 0.865 | **0.855** | 0.901 | 1.012 | 1.952 | 4.745 | > 10 |
| 7 | 0.930 | 0.863 | 0.873 | 0.860 | 0.906 | 1.015 | 1.928 | 4.670 | > 10 |
| 8 | 0.949 | 0.868 | 0.882 | 0.865 | 0.908 | 1.015 | 1.984 | 4.639 | > 10 |
| 9 | 0.991 | 0.882 | 0.899 | 0.876 | 0.923 | 1.023 | 2.046 | 4.624 | > 10 |
| 10 | 1.018 | 0.930 | 9.286 | > 10 | > 10 | > 10 | > 10 | > 10 | > 10 |
| 11 | 1.070 | 0.917 | 0.941 | 3.951 | > 10 | > 10 | > 10 | > 10 | > 10 |
| 12 | 1.120 | 1.114 | > 10 | > 10 | > 10 | > 10 | > 10 | > 10 | > 10 |

N=100, Semiparametric Efficiency Bound = 0.831, 5,000 replications

Table 2
Model 1, Augmented Reweighting, N=100

| Simulated Mean Squared Errors | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 3.578 | 1.004 | 0.906 | 0.869 | 0.914 | 1.015 | 1.969 | 3.767 | > 10 |
| 2 | 0.957 | 1.062 | 0.902 | 0.869 | 0.914 | 1.015 | 1.969 | 3.767 | > 10 |
| 3 | 0.890 | 0.881 | 0.912 | 0.871 | 0.913 | 1.015 | 1.969 | 3.767 | > 10 |
| 4 | 0.889 | 0.867 | 0.872 | 0.875 | 0.915 | 1.016 | 1.969 | 3.767 | > 10 |
| 5 | 0.900 | 0.860 | 0.874 | 0.859 | 0.918 | 1.018 | 1.969 | 3.767 | > 10 |
| 6 | 0.911 | 0.859 | 0.873 | **0.858** | 0.911 | 1.017 | 1.968 | 3.767 | > 10 |
| 7 | 0.930 | 0.867 | 0.880 | 0.864 | 0.917 | 1.017 | 1.965 | 3.766 | > 10 |
| 8 | 0.949 | 0.873 | 0.888 | 0.870 | 0.923 | 1.020 | 1.968 | 3.767 | > 10 |
| 9 | 0.991 | 0.887 | 0.905 | 0.883 | 0.941 | 1.032 | 1.985 | 3.773 | > 10 |
| 10 | 1.018 | 0.903 | 0.914 | 0.891 | 0.947 | 1.039 | 1.980 | 3.765 | > 10 |
| 11 | 1.070 | 0.928 | 0.932 | 0.911 | 0.965 | 1.062 | 1.995 | 3.777 | > 10 |
| 12 | 1.120 | 0.936 | 0.947 | 0.918 | 0.977 | 1.066 | 2.009 | 3.782 | > 10 |

N=100, Semiparametric Efficiency Bound = 0.831, 5,000 replications

Table 3
Model 1, Weighted Regression, N=200

| | Simulated Mean Squared Errors | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 2.557 | 0.512 | 0.458 | 0.439 | 0.439 | 0.436 | 0.445 | 0.450 | 0.481 |
| 2 | 0.491 | 0.542 | 0.457 | 0.438 | 0.438 | 0.437 | 0.444 | 0.450 | 0.476 |
| 3 | 0.449 | 0.449 | 0.459 | 0.439 | 0.438 | 0.436 | 0.444 | 0.451 | 0.476 |
| 4 | 0.440 | 0.437 | 0.438 | 0.440 | 0.438 | 0.437 | 0.444 | 0.450 | 0.475 |
| 5 | 0.438 | **0.433** | 0.435 | 0.434 | 0.438 | 0.437 | 0.444 | 0.450 | 0.475 |
| 6 | 0.439 | 0.435 | 0.436 | 0.434 | 0.436 | 0.437 | 0.443 | 0.450 | 0.476 |
| 7 | 0.443 | 0.436 | 0.437 | 0.435 | 0.437 | 0.436 | 0.444 | 0.450 | 0.478 |
| 8 | 0.445 | 0.435 | 0.437 | 0.434 | 0.436 | 0.436 | 0.443 | 0.451 | 0.478 |
| 9 | 0.449 | 0.437 | 0.439 | 0.436 | 0.438 | 0.437 | 0.445 | 0.452 | 0.478 |
| 10 | 0.450 | 0.437 | 0.438 | 0.435 | 0.437 | 0.436 | 0.445 | 0.451 | 0.477 |
| 11 | 0.458 | 0.439 | 0.442 | 0.437 | 0.440 | 0.439 | 0.447 | 0.453 | 0.480 |
| 12 | 0.467 | 0.442 | 0.444 | 0.438 | 0.442 | 0.440 | 0.448 | 0.455 | 0.483 |

N=200, Semiparametric Efficiency Bound = 0.4155, 5,000 replications

Table 4
Model 1, Augmented Reweighting, N=200

| | Simulated Mean Squared Errors | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 2.557 | 0.512 | 0.458 | 0.439 | 0.439 | 0.436 | 0.445 | 0.450 | 0.481 |
| 2 | 0.491 | 0.548 | 0.457 | 0.439 | 0.439 | 0.436 | 0.445 | 0.450 | 0.481 |
| 3 | 0.449 | 0.449 | 0.461 | 0.439 | 0.439 | 0.436 | 0.445 | 0.450 | 0.481 |
| 4 | 0.440 | 0.438 | 0.438 | 0.441 | 0.439 | 0.437 | 0.445 | 0.450 | 0.481 |
| 5 | 0.438 | **0.433** | 0.436 | 0.433 | 0.439 | 0.437 | 0.445 | 0.451 | 0.481 |
| 6 | 0.439 | 0.435 | 0.437 | 0.434 | 0.437 | 0.437 | 0.445 | 0.450 | 0.481 |
| 7 | 0.443 | 0.436 | 0.438 | 0.435 | 0.438 | 0.436 | 0.445 | 0.451 | 0.481 |
| 8 | 0.445 | 0.436 | 0.438 | 0.434 | 0.437 | 0.436 | 0.444 | 0.451 | 0.482 |
| 9 | 0.449 | 0.438 | 0.440 | 0.436 | 0.439 | 0.437 | 0.446 | 0.452 | 0.481 |
| 10 | 0.450 | 0.437 | 0.440 | 0.434 | 0.438 | 0.436 | 0.445 | 0.451 | 0.480 |
| 11 | 0.458 | 0.440 | 0.443 | 0.437 | 0.441 | 0.438 | 0.448 | 0.453 | 0.482 |
| 12 | 0.467 | 0.443 | 0.447 | 0.439 | 0.445 | 0.440 | 0.450 | 0.455 | 0.484 |

N=200, Semiparametric Efficiency Bound = 0.4155, 5,000 replications

Table 5 shows the simulation results for model 2, which is smoother, using $N = 200$. Due to the similarity exhibited by the combined methods, for this and the subsequent models, we report only the values of $\hat{\beta}_{WR}$. Note that, when the model is too smooth, the difference between all estimators almost disappears. In this specification, although the best estimator strictly utilizes a combination of methods $(K, L > 1)$, loss of accuracy from using either reweighting $(L = 1)$ or imputation $(K = 1)$ alone is smaller than 3% of the efficiency bound.

Table 5
Model 2, Weighted Regression

| | Simulated Mean Squared Errors | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 2.417 | 0.442 | 0.428 | 0.428 | 0.430 | 0.432 | 0.437 | 0.443 | 0.466 |
| 2 | 0.430 | 0.446 | 0.428 | 0.427 | 0.430 | 0.432 | 0.437 | 0.443 | 0.467 |
| 3 | 0.429 | 0.427 | 0.428 | 0.428 | 0.430 | 0.432 | 0.437 | 0.443 | 0.467 |
| 4 | 0.430 | 0.427 | **0.427** | 0.428 | 0.429 | 0.432 | 0.437 | 0.443 | 0.468 |
| 5 | 0.433 | 0.429 | 0.429 | 0.429 | 0.429 | 0.433 | 0.437 | 0.443 | 0.468 |
| 6 | 0.438 | 0.430 | 0.431 | 0.430 | 0.431 | 0.433 | 0.437 | 0.444 | 0.467 |
| 7 | 0.448 | 0.431 | 0.432 | 0.432 | 0.433 | 0.434 | 0.438 | 0.444 | 0.467 |
| 8 | 0.446 | 0.433 | 0.433 | 0.433 | 0.433 | 0.435 | 0.438 | 0.444 | 0.467 |
| 9 | 0.452 | 0.434 | 0.434 | 0.433 | 0.434 | 0.436 | 0.439 | 0.445 | 0.468 |
| 10 | 0.456 | 0.436 | 0.437 | 0.436 | 0.437 | 0.438 | 0.442 | 0.447 | 0.469 |
| 11 | 0.475 | 0.440 | 0.440 | 0.439 | 0.440 | 0.441 | 0.445 | 0.449 | 0.472 |
| 12 | 0.474 | 0.442 | 0.443 | 0.442 | 0.443 | 0.443 | 0.447 | 0.450 | 0.474 |

N=200, Semiparametric Efficiency Bound = 0.4155, 5,000 replications

The least smooth specification deteriorates the performance of all estimators, as is shown in Table 6. Of significance is the fact that the best estimators use a large number of terms for the estimation of $p(.)$, and that the introduction of at least one term, in addition to the constant, in the estimation of $m_t(.)$, often reduces the mean squared error. There is also a change in the relative performance between pure imputation and pure reweighting, the latter yielding a more favorable result.

A possible explanation for this fact is that $p(.)$ is a transformation of $m_1(.) - m_0(.)$ by the logistic function, $L(.)$. The loss of smoothness may have unevenly affected imputation and reweighting methods as an application of the $L(.)$ function usually attenuates the lack of smoothness.

The models used to assess the effect of heteroskedasticity did not provide evidence in favor of the suggestion that this feature could affect the performance of $\hat{\beta}_{WR}$. Tables 7 and 8 report the results, respectively, for the cases presumably favorable and unfavorable to propensity score-based weighting of the regression. Optimal estimators, in both cases, use the same number of terms, contrasting with

Table 6
Model 3, Weighted Regression

| Simulated Mean Squared Errors | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 2.683 | 1.346 | 0.892 | 0.674 | 0.598 | 0.533 | 0.532 | 0.503 | 0.548 |
| 2 | 1.357 | 1.435 | 0.908 | 0.678 | 0.595 | 0.530 | 0.524 | 0.495 | 0.524 |
| 3 | 0.879 | 0.900 | 0.950 | 0.703 | 0.600 | 0.531 | 0.524 | 0.495 | 0.524 |
| 4 | 0.679 | 0.685 | 0.702 | 0.734 | 0.610 | 0.536 | 0.520 | 0.492 | 0.520 |
| 5 | 0.578 | 0.576 | 0.586 | 0.595 | 0.627 | 0.545 | 0.525 | 0.496 | 0.521 |
| 6 | 0.523 | 0.520 | 0.523 | 0.526 | 0.544 | 0.556 | 0.532 | 0.497 | 0.525 |
| 7 | 0.491 | 0.486 | 0.491 | 0.489 | 0.506 | 0.505 | 0.538 | 0.501 | 0.527 |
| 8 | 0.479 | 0.470 | 0.473 | 0.471 | 0.482 | 0.479 | 0.504 | 0.504 | 0.533 |
| 9 | 0.458 | 0.453 | 0.462 | 0.457 | 0.469 | 0.461 | 0.486 | 0.479 | 0.537 |
| 10 | 0.449 | 0.448 | 0.453 | 0.449 | 0.457 | 0.452 | 0.469 | 0.468 | 0.512 |
| 11 | 0.446 | 0.442 | 0.446 | **0.440** | 0.448 | 0.441 | 0.459 | 0.456 | 0.499 |
| 12 | 0.451 | 0.445 | 0.450 | 0.442 | 0.452 | 0.444 | 0.464 | 0.459 | 0.504 |

N=200, Semiparametric Efficiency Bound = 0.4155, 5,000 replications

the expectation that, in the second case, a more accurate reweighting procedure (higher $K$) would be recommendable.

The introduction of multiple auxiliary variables reinforces the gain from combining methods, when we use the weighted regression estimator ($\hat{\beta}_{WR}$), as Table 9 shows. In this case, the best double robust estimator has a squared error 10% and 8% lower than the best imputation and reweighting estimators, respectively. Unlike the previous specifications, however, the way the methods are combined influences the outcome. In Table 10, which uses the augmented reweighting estimator, there are considerable losses in some combinations of $K$ and $L$, and the best estimator uses only the reweighting procedure. Nonetheless, the estimator $\hat{\beta}_{IF}$ which utilizes the optimal combination for $\hat{\beta}_{WR}$ ($K = 10$, $L = 8$) still performs relatively well.

Finally, as was expected, reweighting by inverse true propensity score produced much larger deviations than the semiparametric efficiency bound, as is demonstrated in Table 11. On the other hand, the use of $p(.)$ combined with imputation procedures allowed for a performance similar to the optimal implementations of $\hat{\beta}_{IF}$ and $\hat{\beta}_{WR}$, except in models 3 and 6. In summary, this indicates that the use of information about the propensity score, although it deteriorates asymptotic performance in simple reweighting estimators, allows for the achievement of asymptotic efficiency when combined with the estimation of regression functions. In models that are less affected by 'the curse of dimensionality,' the use of this information was particularly effective in obtaining an optimal estimator. This is because it precluded the selection of the number of terms to be used in $\hat{p}(.)$ and,

Table 7
Model 4, Weighted Regression

| Simulated Mean Squared Errors | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 2.620 | 0.531 | 0.457 | 0.445 | 0.440 | 0.438 | 0.450 | 0.448 | 0.477 |
| 2 | 0.504 | 0.562 | 0.458 | 0.443 | 0.439 | 0.438 | 0.448 | 0.448 | 0.471 |
| 3 | 0.450 | 0.451 | 0.461 | 0.445 | 0.439 | 0.438 | 0.448 | 0.448 | 0.470 |
| 4 | 0.444 | 0.441 | 0.442 | 0.445 | 0.439 | 0.438 | 0.447 | 0.448 | 0.470 |
| 5 | 0.440 | **0.436** | 0.437 | 0.437 | 0.440 | 0.438 | 0.447 | 0.448 | 0.471 |
| 6 | 0.443 | 0.436 | 0.437 | 0.436 | 0.438 | 0.438 | 0.447 | 0.448 | 0.473 |
| 7 | 0.446 | 0.438 | 0.440 | 0.438 | 0.440 | 0.440 | 0.447 | 0.448 | 0.473 |
| 8 | 0.448 | 0.438 | 0.440 | 0.437 | 0.440 | 0.439 | 0.446 | 0.449 | 0.476 |
| 9 | 0.451 | 0.440 | 0.442 | 0.439 | 0.441 | 0.440 | 0.447 | 0.450 | 0.477 |
| 10 | 0.457 | 0.443 | 0.443 | 0.440 | 0.443 | 0.442 | 0.449 | 0.451 | 0.480 |
| 11 | 0.462 | 0.445 | 0.445 | 0.442 | 0.445 | 0.443 | 0.451 | 0.453 | 0.482 |
| 12 | 0.466 | 0.446 | 0.446 | 0.443 | 0.445 | 0.444 | 0.451 | 0.453 | 0.483 |

N=200, Semiparametric Efficiency Bound = 0.4155, 5,000 replications

Table 8
Model 5, Weighted Regression

| Simulated Mean Squared Errors | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 2.564 | 0.521 | 0.453 | 0.432 | 0.431 | 0.431 | 0.447 | 0.453 | 0.483 |
| 2 | 0.494 | 0.550 | 0.454 | 0.432 | 0.431 | 0.432 | 0.446 | 0.452 | 0.472 |
| 3 | 0.443 | 0.446 | 0.457 | 0.433 | 0.431 | 0.432 | 0.446 | 0.455 | 0.471 |
| 4 | 0.432 | 0.432 | 0.432 | 0.435 | 0.431 | 0.432 | 0.446 | 0.455 | 0.473 |
| 5 | 0.432 | **0.430** | 0.430 | 0.430 | 0.432 | 0.433 | 0.447 | 0.455 | 0.473 |
| 6 | 0.434 | 0.430 | 0.431 | 0.430 | 0.431 | 0.434 | 0.447 | 0.455 | 0.474 |
| 7 | 0.441 | 0.433 | 0.434 | 0.432 | 0.434 | 0.435 | 0.447 | 0.456 | 0.474 |
| 8 | 0.443 | 0.435 | 0.436 | 0.434 | 0.436 | 0.437 | 0.449 | 0.458 | 0.474 |
| 9 | 0.450 | 0.437 | 0.439 | 0.436 | 0.438 | 0.439 | 0.452 | 0.460 | 0.475 |
| 10 | 0.457 | 0.442 | 0.443 | 0.440 | 0.442 | 0.443 | 0.455 | 0.463 | 0.477 |
| 11 | 0.459 | 0.442 | 0.445 | 0.441 | 0.444 | 0.444 | 0.457 | 0.464 | 0.477 |
| 12 | 0.465 | 0.444 | 0.447 | 0.443 | 0.445 | 0.445 | 0.459 | 0.465 | 0.479 |

N=200, Semiparametric Efficiency Bound = 0.4155, 5,000 replications

Table 9
Model 6, Weighted Regression

Simulated Mean Squared Errors

| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6.313 | 5.336 | 4.547 | 4.250 | 4.201 | 4.463 | 4.519 | 3.826 | 3.906 |
| 2 | 5.319 | 5.280 | 4.520 | 4.184 | 4.093 | 4.349 | 4.393 | 3.791 | 3.892 |
| 3 | 4.443 | 4.505 | 4.478 | 4.160 | 4.089 | 4.183 | 4.234 | 3,738 | 3.803 |
| 4 | 3.995 | 4.068 | 4.026 | 4.178 | 4.094 | 4.164 | 4.189 | 3.677 | 3.760 |
| 5 | 3.878 | 3.967 | 3.925 | 4.078 | 4.100 | 4.172 | 4.195 | 3.671 | 3.763 |
| 6 | 3.805 | 3.892 | 3.928 | 4.093 | 4.090 | 4.189 | 4.208 | 3.695 | 3.783 |
| 7 | 3.825 | 3.899 | 3.940 | 4.089 | 4.075 | 4.154 | 4.197 | 3.668 | 3.768 |
| 8 | 3.746 | 3.725 | 3.630 | 3.785 | 3.762 | 3.836 | 3.887 | 3.673 | 3.779 |
| 9 | 3.755 | 3.727 | 3.622 | 3.750 | 3.728 | 3.796 | 3.868 | 3.636 | 3.758 |
| 10 | 4.124 | 3.910 | 3.696 | 3.669 | 3.640 | 3.720 | 3.779 | **3.475** | 3.600 |
| 11 | 4.006 | 3.844 | 3.657 | 3.663 | 3.636 | 3.726 | 3.791 | 3.513 | 3.625 |
| 12 | 3.851 | 3.847 | 3.738 | 3.744 | 3.712 | 3.796 | 3.862 | 3.547 | 3.669 |

N=100, 5,000 replications

Table 10
Model 6, Augmented Reweighting

Simulated Mean Squared Errors

| K \ L | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6.313 | 5.336 | 4.547 | 4.250 | 4.201 | 4.463 | 4.519 | 3.826 | 3.906 |
| 2 | 5.319 | 5.286 | 4.574 | 4.295 | 4.200 | 4.462 | 4.519 | 3.827 | 3.907 |
| 3 | 4.443 | 4.547 | 4.536 | 4.281 | 4.212 | 4.312 | 4.363 | 3.832 | 3.907 |
| 4 | 3.995 | 4.113 | 4.096 | 4.280 | 4.203 | 4.275 | 4.304 | 3.760 | 3.870 |
| 5 | 3.878 | 4.039 | 4.026 | 4.218 | 4.238 | 4.315 | 4.339 | 3.766 | 3.894 |
| 6 | 3.805 | 3.979 | 4.044 | 4.268 | 4.260 | 4.362 | 4.377 | 3.806 | 3.933 |
| 7 | 3.825 | 3.993 | 4.052 | 4.267 | 4.262 | 4.330 | 4.364 | 3.778 | 3.899 |
| 8 | **3.746** | 3.824 | 3.765 | 4.037 | 4.024 | 4.067 | 4.127 | 3.847 | 3.986 |
| 9 | 3.755 | 3.848 | 3.799 | 4.045 | 4.032 | 4.066 | 4.149 | 3.851 | 3.999 |
| 10 | 4.124 | 4.319 | 4.151 | 4.460 | 4.430 | 4.361 | 4.473 | 3.875 | 4.049 |
| 11 | 4.006 | 4.183 | 4.047 | 4.407 | 4.399 | 4.386 | 4.474 | 3.925 | 4.072 |
| 12 | 3.851 | 3.999 | 3.958 | 4.203 | 4.214 | 4.321 | 4.313 | 3.823 | 3.961 |

N=100, 5,000 replications

at the same time, allowed for obtaining a mean squared error practically identical to that of $\hat{\beta}_{IF}$ and $\hat{\beta}_{WR}$.

Table 11
Estimates using the true propensity score

| Simulated Mean Squared Errors | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | N \ L | 1 | 2 | 3 | 4 | 6 | 8 | 9 |
| 1, $\tilde{\beta}_{RP}$ | 100 | 1.348 | 0.906 | 0.870 | **0.856** | 0.992 | 3.899 | > 10 |
| 1, $\tilde{\beta}_{RP}$ | 200 | 0.661 | 0.459 | 0.442 | **0.430** | 0.434 | 0.448 | 0.502 |
| 2, $\tilde{\beta}_{RP}$ | 200 | 0.741 | 0.436 | **0.428** | 0.428 | 0.436 | 0.463 | 0.506 |
| 3, $\tilde{\beta}_{RP}$ | 200 | 0.510 | 0.494 | 0.501 | 0.491 | **0.484** | 0.484 | 0.789 |
| 4, $\tilde{\beta}_{RP}$ | 200 | 0.709 | 0.471 | 0.447 | **0.437** | 0.438 | 0.447 | 0.488 |
| 5, $\tilde{\beta}_{RP}$ | 200 | 0.656 | 0.452 | 0.436 | **0.429** | 0.432 | 0.452 | 0.484 |
| 6, $\tilde{\beta}_{RP}$ | 100 | 4.580 | 4.633 | 4.573 | 4.299 | 4.427 | **3.853** | 3.873 |
| 6, $\tilde{\beta}_{FI}$ | 100 | 4.580 | 4.687 | 4.649 | 4.408 | 4.587 | 3.946 | **3.924** |

## 6. Conclusion

In this paper we discussed the combination of regression/imputation and reweighting strategies for the estimation of treatment effects under the ignorability assumption. We argued that the combination of methods allows for the improvement of inference in finite samples, using the double robustness literature as theoretical background.

In the context of parametric inference, the theory of double robust estimation proposes a regression-based adjustment of the reweighting estimator. This allows for consistent estimation when there is risk of misspecification of one of the models. The semiparametric version of this literature shows that a similar procedure can be used to tackle the potential failure of the smoothness assumptions that underlie the asymptotic properties. The theoretical properties of the semiparametric double robust estimation suggest that this technique could help in mitigating the 'curse of dimensionality', thus improving efficiency at small samples.

With this motivation, two double robust estimators were introduced and tested by using Monte Carlo simulations. The simulations confirmed the possibility of a more accurate estimation of the parameter of interest through the combination of traditional methods. We also assessed the relationship between the 'curse of dimensionality' and the advantage of combining methods, as was proposed by Robins and Ritov (1997). A larger number of space dimensions of the auxiliary variables and/or lesser smoothness of the regression functions and propensity score were associated to a more significant performance improvement. On the other hand, unexpectedly, heteroskedasticity did not produce an effect on the gain from the weighting of regressions. Another advantage of double robust estimators observed in the Monte Carlo simulation was the capacity to include information about the propensity score in a trivial manner without losing efficiency.

An important issue that was not addressed is the selection of the number of terms in the sieve estimator base, or – more generally speaking – of the smoothness parameter of first-step semiparametric estimations. The simulation study simply revealed the possibility of making a sufficiently good choice, thus outperforming the imputation and reweighting methods. However, while these methods need only one smoothness parameter, the combination of methods requires two decisions. The ability to propose a data-dependent rule which ensures a proper selection of these parameters is therefore critical to the practical success of this estimation strategy. Rothe and Firpo (2013) show that data-dependent methods for bandwidth selection can in fact be valid under semipametric double robust estimation models.

A possible extension of this study is the analysis of the inference about parameters of treatment effect for the treated. Even though this topic has, in general, received a consideration akin to that of the average treatment effect in the literature, the theory is not completely equivalent. In particular, the literature on double robustness is, in this case, incipient. Of the works investigated in this paper, only that of Egel et al. (2012) and Rothe and Firpo (2013) present an estimator for this case.

## References

Bang, H. & Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972.

Bickel, P., Klassen, C., Ritov, Y., & Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer-Verlag, New York.

Busso, M., Dinardo, J., & McCrary, J. (2013). New evidence on the finite sample properties of propensity score matching and reweighting estimators. *Review of Economics and Statistics*, forthcoming.

Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155:138–154.

Chen, X., Hong, H., & Tamer, E. (2005). Measurement error models with auxiliary data. *Review of Economic Studies*, 72:343–366.

Chen, X., Hong, H., & Tarozzi, A. (2008). Semiparametric efficiency in gmm models with auxiliary data. *Annals of Statistics*, 36:808–843.

Chen, X., Linton, O., & Keilegom, I. V. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71:1591–1608.

Crump, R., Hotz, V., Imbens, G., & Mitnik, O. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96:1–13.

Egel, D., Graham, B., & Pinto, C. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79:1053–1079.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75:259–276.

Gemam, S. & Huang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10:401–414.

Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

Hahn, J. (1998). In the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331.

Hahn, J., Todd, P., & van der Klaauw, W. (2000). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69:201–209.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66:1017–1098.

Heckman, J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies*, 64:605–654.

Hirano, K. & Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application of data on right *Heart* catherization. *Health Services and Outcomes Research Methodology*, 2:259–278.

Hirano, K., Imbens, G., & Ridder, G. (2000). Efficient estimation of average treatment effects using the estimated propensity score. Working Paper Series T0251, NBER. Available at: `http://ssrn.com/abstract=228061\`.

Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71:1161–1189.

Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Ichimura, H. & Linton, O. (2005). Asymptotic expansions for some semiparametric program evaluation estimators. In Andrews, D. & Stock, J., editors, *Identification and Inference for Econometric Models*, chapter 8, pages 149–170. Cambridge University Press.

Imbens, G. (2000). The role of the propensity score in estimating dose response functions. *Biometrika*, 87:706–710.

Imbens, G. & Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 61:467–476.

Imbens, G., Newey, W., & Ridder, G. (2005). Mean-squared-error calculations for average treatment effects. Working Paper 05.34, IEPR. Available at: `http://ssrn.com/abstract=820205`.

Imbens, G. & Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47:5–86.

Kang, J. & Schaeffer, J. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22:523–539.

Lunceford, J. & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23:2937–2960.

Manski, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings*, 80:319–323.

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

Neugebauer, R. & der Laan, M. V. (2005). Why prefer double robust estimates? *Journal of Statistical Planning and Inference*, 129:405–426.

Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5:99–135.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62:1349–1382.

Newey, W. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168.

Newey, W. & McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. & McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 38, pages 2111–2242. North-Holland, Amsterdam.

Robins, J. & Ritov, Y. (1997). Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Semiparametric Efficiency in Multivariate Regression Models with Missing Data*, 16:285–319.

Robins, J. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129.

Robins, J., Rotnitzky, A., & Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121.

Robins, J., Sued, M., Lei-Gomez, O., & Ritnitzky, A. (2007). Comment: Performance of double-robust estimators when 'inverse probability' weights are highly variable. *Statistical Science*, 22:544–559.

Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Rothe, C. & Firpo, S. (2013). Semiparametric estimation and inference using doubly robust moment conditions. IZA Discussion Papers 7564, Institute for the Study of Labor (IZA). Available at: `http://ideas.repec.org/p/iza/izadps/dp7564.html`.

Rubin, D. (1973). Matching to remove bias in observational studies. *Biometrics*, 29:159–183.

Rubin, D. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2:1–26.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.

Scharfstein, D., Robins, J., & Rotnitzky, A. (1999). Adjusting for non-ignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1146.

Stein, C. (1956). Efficient nonparametric testing and estimation. In *Berkeley Symposium on Mathematical Statistics and Probability, 3. Proceedings*, pages 187–196. University of California Press, Berkeley.