Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions

Christoph Rothe and Sergio Firpo*

Abstract

We study semiparametric two-step estimators which have the same structure as parametric doubly robust estimators in their second step, but retain a fully nonparametric specification in the first step. Such estimators exist in many economic applications, including a wide range of missing data and treatment effect models. We show that these estimators are $\sqrt{n}$-consistent and asymptotically normal under weaker than usual conditions on the accuracy of the first stage estimates, have smaller first order bias and second order variance, and that their finite-sample distribution can be approximated more accurately by classical first order asymptotics. We argue that because of these refinements our estimators are useful in many settings where semiparametric estimation and inference are traditionally believed to be unreliable. We also illustrate the practical relevance of our approach through simulations and an empirical application.

JEL Classification: C14, C21, C31, C51

Keywords: *Semiparametric model, missing data, treatment effects, doubly robust estimation, higher order asymptotics*

*First version: December 20, 2012. This version: November 10, 2013. Christoph Rothe, Columbia University, Department of Economics, 420 W 118th St, New York, NY 10027, USA. Email: cr2690@columbia.edu. Sergio Firpo, Escola de Economia de Sao Paulo FGV-SP, R. Itapeva, 474/1215, Sao Paulo-SP, 01332-000, Brasil. E-Mail: sergio.firpo@fgv.br. We would like to thank Matias Cattaneo, Michael Jansson, Marcelo Moreira, Ulrich Müller, Cristine Pinto, and seminar audiences at Brown, Columbia, EPGE-FGV, University of Pennsylvania, Princeton, PUC-Rio, the 2012 Greater NY Metropolitan Colloquium and the 2013 North American Summer Meetings for their helpful comments; and Matias Cattaneo for making available the data used in the empirical application. Sergio Firpo gratefully acknowledges financial support from CNPq-Brazil.

1. Introduction

Semiparametric two-step estimators are by now available for a wide range of econometric applications. These estimators typically arise from a flexible model in which a finite-dimensional parameter of interest can be characterized through a moment condition that contains an unknown nuisance function. In a first step, the nuisance function is estimated nonparametrically. In a second step, the parameter of interest is then estimated from an empirical version of the moment condition, with the unknown nuisance function replaced by its first step estimate. Semiparametric two-step estimators are important for empirical research because they allow practitioners to remove many parametric restrictions, which could potentially mask important features of the data, from their specifications.

The first order asymptotic properties of semiparametric two-step estimators have been studied extensively (e.g. Newey, 1994; Newey and McFadden, 1994; Andrews, 1994; Chen, Linton, and Van Keilegom, 2003; Ichimura and Lee, 2010), and are widely used to justify large sample inference procedures. However, there is considerable evidence that first order asymptotic distributions provide poor approximations to the sampling behavior of semiparametric two-step estimators, at least for sample sizes typically encountered in empirical practice (e.g. Linton, 1995; Robins and Ritov, 1997; Cattaneo, Crump, and Jansson, 2013a). For instance, the standard first order approximation is invariant to the nonparametric estimation technique used in the first step, yet point estimates can be very sensitive to implementation details, such as the choice of smoothing parameters.

This discrepancy can be attributed to the fact that first order approximations are usually derived under strong smoothness conditions on the unknown nuisance function. Such an approach allows treating certain terms in an expansion of the estimator as negligible in an asymptotic sense (e.g. Robins and Ritov, 1997). However, in finite samples these higher order terms could still be of substantial magnitude, and thus considerably affect the properties of the final estimator. One way to address this issue would be to subtract estimates of these terms from the final estimator, but this generally adds an undesirable layer of complexity as higher order terms often depend on nonlinear transformations of nonparametric objects (e.g. Linton, 1995).

In this paper, we consider a different approach, which involves constructing simple alternative estimators for which higher order terms are small to begin with. We propose a new class of semiparametric two-step estimators that are based on a moment condition with a

particular structure: it depends on two unknown nuisance functions, but still identifies the parameter of interest if either one of the two functions is replaced by some arbitrary value. Following the terminology in Robins, Rotnitzky, and van der Laan (2000), we refer to such moment conditions as *doubly robust* (DR), and thus call the corresponding estimators *semiparametric doubly robust estimators* (SDREs). DR moment conditions exist for many interesting parameters, including regression coefficients in models with missing outcomes and/or covariates, average treatment effects in potential outcome models with unconfounded assignment, and local average treatment effects in instrumental variable models, amongst many others. Our estimators can thus be applied in a wide range of empirical applications.

Our main contribution is to show that SDREs have attractive theoretical and practical properties relative to generic semiparametric two-step estimators based on a moment condition without the DR property. We show that the special structure of DR moment conditions, together with a certain orthogonality condition that is not restrictive in all examples that we consider, removes the two largest second order terms in a traditional expansion of the estimator. This effect occurs automatically, and does not require choosing additional tuning parameters or involved numerical computations. As a consequence, SDREs have smaller first order bias and second order variance, and are $\sqrt{n}$-consistent and asymptotically normal under weaker conditions on the accuracy of the first step nonparametric estimates. Moreover, their finite sample distribution can be better approximated by classical first order asymptotics. Therefore any method for inference that is justified by large sample theory, such as the usual confidence intervals or hypothesis tests, should be more accurate in our case. In all examples that we consider in this paper, SDREs are also semiparametrically efficient. They have thus clear advantages even relative to other efficient estimators that are commonly used in such settings, such as Inverse Probability Weighting (IPW) estimators in missing data and treatment effect models (e.g. Hirano, Imbens, and Ridder, 2003; Firpo, 2007; Chen, Hong, and Tarozzi, 2008).

From a practitioner's perspective, our results imply that SDREs are generally more precise in finite samples than generic semiparametric estimators with the same asymptotic variance, and that their properties are less sensitive to the implementation of the nonparametric first stage. Moreover, in settings with moderate dimensionality, they can allow for rate-optimal choices of smoothing parameters (which are relatively easy to estimate from the data), and do not require the use of bias reducing nonparametric estimators (such as those

based on higher order kernels, for instance). These are important advantages that make SDREs attractive in applications. SDREs are also adaptive, in the sense that by construction their asymptotic variance does not contain adjustment terms for the nonparametric first step. This is a useful property, as it simplifies the calculation of standard errors.

Our SDREs differ from the usual doubly robust procedures used widely in statistics. See for example Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995), Scharfstein, Rotnitzky, and Robins (1999), Robins and Rotnitzky (2001), Van der Laan and Robins (2003) or Tan (2006), and Wooldridge (2007) or Graham, Pinto, and Egel (2012) for applications in econometrics. These estimators employ fully parametric specifications of the two nuisance functions, and the role of the DR property is to ensure consistency of the final estimator if at most one of these specifications is incorrect. In this paper we impose no such parametric restrictions on nuisance functions when computing our SDREs. Instead, we retain a fully nonparametric first stage.

Our paper is not the first to be concerned with improving the properties of semiparametric two-stage estimators. In very different contexts, Newey, Hsieh, and Robins (2004) and Klein and Shen (2010) propose methods that do not exploit higher order differentiability conditions to reduce the impact of the first stage smoothing bias on the properties of certain two-step estimators. Cattaneo et al. (2013a) study a jackknife approach to remove bias terms related to the variance of the first stage nonparametric problem in the specific context of weighted average derivative estimation. Our paper complements these findings in a general sense by showing that the use of doubly robust moment conditions achieves both goals simultaneously. An alternative approach to improve inference, which we do not consider in this paper, would be to derive "non-$\sqrt{n}$" asymptotic approximations. Examples of such a strategy include Robins, Li, Tchetgen, and Van Der Vaart (2008), who consider semiparametric inference in models with very high-dimensional functional nuisance parameters, and Cattaneo, Crump, and Jansson (2013b), who study so-called small bandwidth asymptotics for semiparametric estimators of density-weighted average derivatives.

The remainder of this paper is structured as follows. In the next section, we present the modeling framework and our estimation procedure, and give some concrete examples of doubly robust moment conditions. In Section 3, the estimators' asymptotics properties are studied. Section 4 applies our method to the estimation of treatment effects under unconfoundedness. Section 5 shows evidence that SDREs have superior properties compared

to other methods in a simulation study. In Section 6, we apply our method to study the effect of smoking on birth weight. Finally, Section 7 concludes. All proofs are collected in the Appendix.

2. Modeling Framework and Estimation Procedure

2.1. Doubly Robust Moment Conditions. We consider the problem of estimating a parameter θ_o, contained in the interior of some compact parameter space $\Theta \subset \mathbb{R}^{d_\theta}$, in a semiparametric model. The data consists of an i.i.d. sample $\{Z_i\}_{i=1}^n$ from the distribution of the random vector $Z \in \mathbb{R}^{d_z}$. We assume that one way to identify θ_o within the semiparametric model is through a moment condition with two nuisance functions. That is, there exists a known moment function $\psi(\cdot)$ taking values in $\mathbb{R}^{d_\theta}$ such that

$$\Psi(\theta, p_o, q_o) := \mathbb{E}(\psi(Z, \theta, p_o(U), q_o(V))) = 0 \text{ if and only if } \theta = \theta_o, \tag{2.1}$$

where $p_o \in \mathcal{P}$ and $q_o \in \mathcal{Q}$ are unknown (but identified) functions, and $U \in \mathbb{R}^{d_p}$ and $V \in \mathbb{R}^{d_q}$ are random subvectors of Z that might have common elements. We consider settings where the moment condition (2.1) exhibits a particular structure. First, we assume that

$$\Psi(\theta, p_o, q) = 0 \text{ and } \Psi(\theta, p, q_o) = 0 \text{ if and only if } \theta = \theta_o \tag{2.2}$$

for all functions $q \in \mathcal{Q}$ and $p \in \mathcal{P}$. Following the terminology in Robins et al. (2000), we refer to any moment condition that is of the form in (2.1) and satisfies the restriction (2.2) as a *doubly robust (DR) moment condition*. Second, we assume that $p_o(x) = \mathbb{E}(Y_p|X_p = x)$ and $q_o(x) = \mathbb{E}(Y_q|X_q = x)$, where $(Y_p, Y_q, X_p, X_q) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d_p} \times \mathbb{R}^{d_q}$ is a random subvector of Z that might have common elements, and that

$$\mathbb{E}((Y_p - p_o(X_p)) \times (Y_q - q_o(X_q))|X_p, X_q) = 0. \tag{2.3}$$

Equation (2.3) is an *orthogonality condition*, which ensures that one can construct nonparametric estimates of p_o and q_o that are asymptotically uncorrelated. In all applications that we consider in this paper, this condition is implied by the assumptions made to identify the parameter of interest, and is thus not restrictive. We explain this point, and how we exploit the property, in more detail below.

2.2. Examples. The conditions (2.2) and (2.3) are of course restrictive, but are jointly satisfied in a wide range of models that are widely used in empirical practice. Before discussing the specific form and implementation of the estimator, we now give a number of examples of models where there exists a DR moment condition with nuisance parameters satisfying the orthogonality condition, all of which are known well in the literature. These examples cover various parameters of interest in missing data and causal inference models, which should illustrate the broad applicability of the methodology. Note that the moment function ψ, on which the DR moment condition is based, is the semiparametrically efficient influence function for the respective parameter of interest in all these examples. This implies that the asymptotic variance of SDREs is equal to the respective semiparametric efficiency bound in these settings (under suitable regularity conditions; see Section 3).

Example 1 (Population Mean with Missing Data)**.** Let X be a vector of covariates that is always observed, and Y a scalar outcome variable that is observed if $D = 1$, and unobserved if $D = 0$. The data consists of a sample from the distribution of $Z = (DY, X, D)$, and the parameter of interest is $\theta_o = \mathbb{E}(Y)$. Assume that the data are missing at random (MAR), i.e. $\mathbb{E}(D|Y, X) = \mathbb{E}(D|X) > 0$ with probability 1, and define the functions $\pi_o(x) = \mathbb{E}(D|X = x)$ and $\mu_o(x) = \mathbb{E}(Y|D = 1, X = x)$. Then $\Psi(\theta, \pi, \mu) = \mathbb{E}(\psi(Z, \theta, \pi(X), \mu(X)))$ with

$$\psi(z, \theta, \pi(x), \mu(x)) = \frac{d(y - \mu(x))}{\pi(x)} + \mu(x) - \theta$$

is a DR moment condition for estimating θ_o. Moreover, because the MAR assumption implies that $\mu_o(x) = \mathbb{E}(Y|X = x)$, it follows from the law of iterated expectations that

$$\mathbb{E}((D - \pi_o(X)) \times (Y - \mu_o(X))|X) = 0,$$

and thus the orthogonality condition holds. □

Example 2 (Linear Regression with Missing Covariates)**.** Let $X = (X_1^\top, X_2^\top)^\top$ be a vector of covariates and Y a scalar outcome variable. Suppose that the covariates in X_1 are only observed if $D = 1$ and unobserved if $D = 0$, whereas (Y, X_2) are always observed. The data thus consists of a sample from the distribution of $Z = (Y, X_1 D, X_2, D)$. Here we consider the vector of coefficients θ_o from a linear regression of Y on X as the parameter of interest. Define the functions $\pi_o(y, x_2) = \mathbb{E}(D|Y = y, X_2 = x_2)$ and $\mu_o(y, x_2, \theta) = \mathbb{E}(\varphi(Y, X, \theta)|D = 1, Y = y, X_2 = x_2)$ with $\varphi(Y, X, \theta) = (1, X^\top)^\top (Y - (1, X^\top)\theta)$, and assume that $\pi_o(Y, X_2) > 0$

with probability 1. Then $\Psi(\theta, \pi, \mu) = \mathbb{E}(\psi(Z, \theta, \pi(Y, X_2), \mu(Y, X_2, \theta)))$ with

$$\psi(z, \theta, \pi(y, x_2), \mu(y, x_2, \theta)) = \frac{d(\varphi(y, x, \theta) - \mu(y, x_2, \theta))}{\pi(y, x_2)} + \mu(y, x_2, \theta)$$

is a DR moment condition for estimating θ_o, and it is easy to verify that the orthogonality condition holds. □

Example 3 (Average Treatment Effects)**.** Let $Y(1)$ and $Y(0)$ denote the potential outcomes with and without taking some treatment, respectively, with $D = 1$ indicating participation in the treatment, and $D = 0$ indicating non-participation in the treatment. Then the realized outcome is $Y = Y(D)$. The data consist of a sample from the distribution of $Z = (Y, D, X)$, where X is some vector of covariates that are unaffected by the treatment, and the parameter of interest is the Average Treatment Effect (ATE) $\theta_o = \mathbb{E}(Y(1)) - \mathbb{E}(Y(0))$. Define the functions $\pi_o(x) = \mathbb{E}(D|X = x)$ and $\mu_o^Y(d, x) = \mathbb{E}(Y|D = d, X = x)$, put $\mu_o(x) = (\mu_o^Y(1, x), \mu_o^Y(0, x))$, and assume that $1 > \mathbb{E}(D|Y(1), Y(0), X) = \pi_o(X) > 0$ with probability 1. Then $\Psi(\theta, \pi, \mu) = \mathbb{E}(\psi(Z, \theta, \pi(X), \mu(X)))$ with

$$\psi(z, \theta, \pi(x), \mu(x)) = \frac{d(y - \mu^Y(1, x))}{\pi(x)} - \frac{(1 - d)(y - \mu^Y(0, x))}{1 - \pi(x)} + (\mu^Y(1, x) - \mu^Y(0, x)) - \theta$$

is a DR moment condition for estimating θ_o, and it is easy to verify that the orthogonality condition holds. □

Example 4 (Average Treatment Effect on the Treated)**.** Consider the potential outcomes setting introduced in the previous example, but now suppose that the parameter of interest is $\theta_o = \mathbb{E}(Y(1)|D = 1) - \mathbb{E}(Y(0)|D = 1)$, the Average Treatment Effect on the Treated (ATT). Define the functions $\pi_o(x) = \mathbb{E}(D|X = x)$ and $\mu_o(x) = \mathbb{E}(Y|D = 0, X = x)$, put $\Pi_o = \mathbb{E}(D)$, $\Pi_o > 0$, and assume that $\mathbb{E}(D|Y(1), Y(0), X) = \pi_o(X) < 1$ with probability 1. Then $\Psi(\theta, \pi, \mu) = \mathbb{E}(\psi(Z, \theta, \pi(X), \mu(X)))$ with

$$\psi(z, \theta, \pi(x), \mu(x)) = \frac{d(y - \mu(x))}{\Pi_o} - \frac{\pi(x)}{\Pi_o} \cdot \frac{(1 - d)(y - \mu(x))}{1 - \pi(x)} - \theta$$

is a DR moment condition for estimating θ_o, and it is easy to verify that the orthogonality condition holds. □

Example 5 (Local Average Treatment Effects)**.** Let $Y(1)$ and $Y(0)$ denote the potential outcomes with and without taking some treatment, respectively, with $D = 1$ indicating participation in the treatment, and $D = 0$ indicating non-participation in the treatment. Furthermore, let $D(1)$ and $D(0)$ denote the potential participation decision given some realization

of a binary instrumental variable $W \in \{0,1\}$. That is, the realized participation decision is $D = D(W)$ and the realized outcome is $Y = Y(D) = Y(D(W))$. The data consist of a sample from the distribution of $Z = (Y, D, W, X)$, where X is some vector of covariates that are unaffected by the treatment and the instrument. Define the function $\pi_o(x) = \mathbb{E}(W|X = x)$, and suppose that $1 > \mathbb{E}(W|Y(1), Y(0), D(1), D(0), X) = \mathbb{E}(W|X) > 0$ and $P(D(1) \geq D(0)|X) = 1$ with probability 1. Under these conditions, it is possible to identify the Local Average Treatment Effect (LATE) $\theta_o = \mathbb{E}(Y(1) - Y(0)|D(1) > D(0))$, which serves as the parameter of interest in this example. Also define the functions $\mu_o^D(w,x) = \mathbb{E}(D|W = w, X = x)$ and $\mu_o^Y(w,x) = \mathbb{E}(Y|W = w, X = x)$, and put $\mu_o(x) = (\mu_o^D(1,x), \mu_o^D(0,x), \mu_o^Y(1,x), \mu_o^Y(0,x))$. Then $\Psi(\theta, \pi, \mu) = \mathbb{E}(\psi(Z, \theta, \pi(X), \mu(X)))$ with

$$\psi(z, \theta, \pi(x), \mu(x)) = \psi^A(z, \pi(x), \mu(x)) - \theta \cdot \psi^B(z, \pi(x), \mu(x)),$$

where

$$\psi^A(z, \pi(x), \mu(x)) = \frac{w(y - \mu^Y(1,x))}{\pi(x)} - \frac{(1-w)(y - \mu^Y(0,x))}{1 - \pi(x)} + \mu^Y(1,x) - \mu^Y(0,x),$$
$$\psi^B(z, \pi(x), \mu(x)) = \frac{w(d - \mu^D(1,x))}{\pi(x)} - \frac{(1-w)(d - \mu^D(0,x))}{1 - \pi(x)} + \mu^D(1,x) - \mu^D(0,x),$$

is a DR moment condition for estimating θ_o, and it is easy to verify that the orthogonality condition holds. □

2.3. Semiparametric Estimation. Equation (2.2) implies that knowledge of either p_o or q_o suffices for identifying θ_o. In principle, one could therefore construct semiparametric estimators of θ_o that only require an estimate of either p_o or q_o, but not both. For example, θ_o could be estimated by the value that sets a sample analogue of either $\Psi(\theta, p_o, \bar{q})$ or $\Psi(\theta, \bar{p}, q_o)$ equal to zero, where $\bar{p} \in \mathcal{P}$ and $\bar{q} \in \mathcal{Q}$ are arbitrary known and fixed functions. In this paper, we argue in favor of an estimator of θ_o that solves a direct sample analogue of the doubly robust moment condition (2.1). That is, we consider the estimator $\widehat{\theta}$ which solves the equation

$$0 = \frac{1}{n} \sum_{i=1}^{n} \psi(Z_i, \theta, \widehat{p}(U_i), \widehat{q}(V_i)), \tag{2.4}$$

where $\widehat{p}$ and $\widehat{q}$ are suitable nonparametric estimates of p_o and q_o, respectively. We refer to such an estimator as a *semiparametric doubly robust estimator* (SDRE). We also define the

following quantities, which will be important for estimating the asymptotic variance of the estimator $\widehat{\theta}$:

$$\widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} \partial\psi(Z_i, \widehat{\theta}, \widehat{p}(U_i), \widehat{q}(V_i))/\partial\theta$$
$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \psi(Z_i, \widehat{\theta}, \widehat{p}(U_i), \widehat{q}(V_i))\psi(Z_i, \widehat{\theta}, \widehat{p}(U_i), \widehat{q}(V_i))^\top.$$

It remains to define suitable nonparametric estimates of p_o and q_o. Recall that we consider the case that $p_o(x) = \mathbb{E}(Y_p|X_p = x)$ and $q_o(x) = \mathbb{E}(Y_q|X_q = x)$, where $(Y_p, Y_q, X_p, X_q) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d_p} \times \mathbb{R}^{d_q}$ is a random subvector of Z that might have common elements. For simplicity, both X_p and X_q are assumed to be continuously distributed in the following.[1] We propose to estimate both functions by local polynomial regression of order l_p and l_q, respectively. This class of kernel-based smoothers has been studied extensively by e.g. Fan (1993), Ruppert and Wand (1994) or Fan and Gijbels (1996). It is well-known to have attractive bias properties relative to the standard Nadaraya-Watson estimator with higher-order kernels. In applications where the dimension of X_p and X_q is not too large (in a sense made precise below), we will work with $l_p = l_q = 1$. using the notation that for d-dimensional vectors a, b we have $|a| = \sum_{i=1}^{d} a_i$ and $a^b = \prod_{i=1}^{d} a_i^{b_i}$, the "leave-$i$-out" local polynomial estimators of $p_o(U_i)$ and $q_o(V_i)$ are given by

$$\widehat{p}(U_i) = \widehat{a}_p(U_i) \quad \text{and} \quad \widehat{q}(V_i) = \widehat{a}_q(V_i),$$

respectively, where

$$(\widehat{a}_p(U_i), \widehat{b}_p(U_i)) = \underset{a,b}{\operatorname{argmin}} \sum_{j\neq i} \left(Y_{p,j} - a - \sum_{1\leq|s|\leq l_p} b_s (X_{p,j} - U_i)^s \right)^2 K_{h_p}(X_{p,j} - U_i),$$
$$(\widehat{a}_q(V_i), \widehat{b}_q(V_i)) = \underset{a,b}{\operatorname{argmin}} \sum_{j\neq i} \left(Y_{q,j} - a - \sum_{1\leq|s|\leq l_q} b_s (X_{q,j} - V_i)^s \right)^2 K_{h_q}(X_{q,j} - V_i),$$

Here $\sum_{1\leq|s|\leq l_p}$ denotes the summation over all d_p vectors of positive integers with $1 \leq |s| \leq l_p$, $K_{h_p}(u) = \prod_{j=1}^{d_p} \mathcal{K}(u_j/h_p)/h_p$ is a d_p-dimensional product kernel built from the univariate

[1] It would be straightforward to extend our results to other types of functions, including derivatives of conditional expectation functions, density functions, and conditional expectation functions with multivariate outcome variables and/or discrete covariates.

kernel function $\mathcal{K}$, and h_p is a one-dimensional bandwidth that tends to zero as the sample size n tends to infinity; and $\sum_{1\leq|s|\leq l_q}$, $K_{h_q}(v)$ and h_q are defined similarly. Note that under suitable regularity conditions (see e.g. Masry, 1996, or Appendix B) these estimators are uniformly consistent, and satisfy

$$\max_{i=1,\ldots,n} |\widehat{p}(U_i) - p_o(U_i)| = O(h_p^{l_p+1}) + O_P((nh_p^{d_p}/\log n)^{-1/2}), \tag{2.5}$$

$$\max_{i=1,\ldots,n} |\widehat{q}(V_i) - q_o(V_i)| = O(h_q^{l_q+1}) + O_P((nh_q^{d_q}/\log n)^{-1/2}), \tag{2.6}$$

where the terms on the right-hand side of each of the two previous equations correspond to the order of the respective bias and stochastic part. Also note that it would be straightforward to employ more general estimators using a matrix of smoothing parameters that is of dimension $d_p \times d_p$ or $d_q \times d_q$, respectively, at the cost of a much more involved notation (Ruppert and Wand, 1994). Moreover, using "leave-i-out" versions of the nonparametric estimators is only necessary for the results we derive below in applications where either U and X_p or V and X_q share some common elements.

3. Asymptotic Theory

In this section, we study the theoretical properties of SDREs, and compare them to those of generic semiparametric two-step estimators. To illustrate the nature of our results, we begin by writing our estimator as

$$\widehat{\theta} - \theta_o = \frac{1}{n}\sum_{i=1}^{n} \Gamma_o^{-1}\psi(Z_i, \theta_o, p_o(U_i), q_o(V_i)) + R_n, \tag{3.1}$$

where $\Gamma_o = \partial\mathbb{E}(\psi(Z, \theta, p_o(U), q_o(V)))/\partial\theta|_{\theta=\theta_o}$ is assumed to have full rank. Without saying anything about R_n, this representation is certainly without loss of generality. Note that the first term on the right-hand side of (3.1) is a sample average of n i.i.d. mean zero random vectors, and is thus asymptotically normal under standard conditions. Now our contribution is two-fold. First, we show that (3.1) holds with $R_n = o_P(n^{-1/2})$, which implies that $\widehat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal, under conditions that are substantially weaker than those commonly employed in the literature on semiparametric two-step estimation. In particular, the familiar requirement that the first stage nonparametric estimation error and bias are $o_p(n^{-1/4})$ and $o(n^{-1/2})$, respectively, in some suitable norm is relaxed. Second, we derive an explicit expression for the rate at which R_n tends to zero, and show that this rate

is substantially faster than the one that can be achieved by generic semiparametric two-step estimators. As a consequence, we can expect standard Gaussian approximations based on (3.1) to be more accurate in finite samples for our SDREs.

3.1. The Structure of the Argument. Before formally stating our results, we give a simplified explanation for how the particular structure of our model and corresponding estimation procedure help us achieving them. We first provide some intuition for why (3.1) holds with $R_n = o_P(n^{-1/2})$ under weak conditions on the accuracy of the first stage. Define

$$\begin{aligned}
T_{n,1} &= \frac{1}{n}\sum_{i=1}^{n} \psi^{p}(Z_i)(\widehat{p}(U_i) - p_o(U_i)) + \frac{1}{n}\sum_{i=1}^{n} \psi^{q}(Z_i)(\widehat{q}(V_i) - q_o(V_i)), \\
T_{n,2,A} &= \frac{1}{n}\sum_{i=1}^{n} \psi^{pp}(Z_i)(\widehat{p}(U_i) - p_o(U_i))^2 + \frac{1}{n}\sum_{i=1}^{n} \psi^{qq}(Z_i)(\widehat{q}(V_i) - q_o(V_i))^2 \text{ and} \\
T_{n,2,B} &= \frac{1}{n}\sum_{i=1}^{n} \psi^{pq}(Z_i)(\widehat{p}(U_i) - p_o(U_i))(\widehat{q}(V_i) - q_o(V_i)),
\end{aligned}$$

where $\psi^{p}(Z_i)$ and $\psi^{pp}(Z_i)$ are the first and second derivative of $\psi(Z_i, \theta_o, p_o(U_i), q_o(V_i))$ with respect to $p_o(U_i)$, respectively, $\psi^{q}(Z_i)$ and $\psi^{qq}(Z_i)$ are defined analogously, and $\psi^{pq}(Z_i)$ is the partial cross derivative of $\psi(Z_i, \theta_o, p_o(U_i), q_o(V_i))$ with respect to $p_o(U_i)$ and $q_o(V_i)$. The terms $T_{n,1}$ and $T_{n,2} = T_{n,2,A} + T_{n,2,B}$ are the linear and quadratic part, respectively, of a standard expansion of the functional $(\widehat{p}, \widehat{q}) \mapsto n^{-1}\sum_{i=1}^{n} \psi(Z_i, \widehat{p}(U_i), \widehat{q}(V_i))$ around (p_o, q_o). Under a weak smoothness condition on the function ψ, it then holds that

$$R_n = O(T_{n,1}) + O(T_{n,2,A}) + O(T_{n,2,B}) + O_P(\|\widehat{p} - p_o\|_\infty^3) + O_P(\|\widehat{q} - q_o\|_\infty^3).$$

Clearly, the two "cubic" remainder terms in this equation are both $o_P(n^{-1/2})$ even if the two first-stage nonparametric estimation errors are uniformly only $o_P(n^{-1/6})$. For a generic semiparametric two-step estimator, however, the remaining "linear" and "quadratic" terms would not be $o_P(n^{-1/2})$ if the nonparametric component converges that slowly.

The advantage of working with a DR moment condition is that its particular structure substantially improves the rate at with the linear and quadratic terms converge to zero. To see this, first note that the DR property (2.2) implies that

$$\frac{\partial^k}{\partial t}\Psi(\theta_o, p_o + t\bar{p}, q_o)|_{t=0} = \frac{\partial^k}{\partial t}\Psi(\theta_o, p_o, q_o + t\bar{q})|_{t=0} = 0 \tag{3.2}$$

for $k = 1, 2$ and all functions $\bar{p}$ and $\bar{q}$ such that $p_o + t\bar{p} \in \mathcal{P}$ and $q_o + t\bar{q} \in \mathcal{Q}$ for all $t \in \mathbb{R}$ with $|t|$ sufficiently small. Now write the the first summand in the definition of $T_{n,1}$ as $\Psi_n^p[\widehat{p} - p_o]$, where the operator Ψ_n^p is defined as $\Psi_n^p[\bar{p}] = n^{-1} \sum_{i=1}^n \psi^p(Z_i)\bar{p}(U_i)$ for any fixed function $\bar{p}$. Clearly, we have that $\Psi_n^p[\bar{p}] \xrightarrow{p} \partial\Psi(\theta_o, p_o + t\bar{p}, q_o)/\partial t|_{t=0} = 0$ for all $\bar{p}$. This explains why the term $\Psi_n^p[\widehat{p} - p_o]$ converges to zero at rate faster-than-usual rate: not only does the argument of the operator tend to zero, but by (3.2) also the operator itself. An analogous reasoning applies to all components of $T_{n,1}$ and $T_{n,2,A}$. This property is specific to SDREs, and does not hold for generic semiparametric estimators.

The term $T_{n,2,B}$ involves a different argument. After some calculations, one finds that the leading terms in a stochastic expansion of this quantity are equal to a constant times the *product* of the smoothing bias terms of the estimators of p_o and q_o, and to a term that is proportional to the asymptotic covariance between the estimation errors of $\widehat{p}$ and $\widehat{q}$. Due to the orthogonality condition (2.3), however, this asymptotic covariance is exactly equal to zero. In order to obtain the desired rate for $T_{n,2,B}$, it thus suffices that the *product* of the bias terms is uniformly $o(n^{-1/2})$.

3.2. Assumptions. We now state the assumptions that allow us to formalize the above arguments.

Assumption 1. *(i) the random vectors U and V are continuously distributed with compact support I_U and I_V, respectively (ii)* $\sup_u \mathbb{E}(|Y_p|^c | X_p = u) < \infty$ *and* $\sup_v \mathbb{E}(|Y_q|^c | X_q = v) < \infty$ *for some constant $c > 2$, (iii) the random vectors X_p and X_q are continuously distributed with support $I_p \supseteq I_U$ and $I_q \supseteq I_V$, respectively (iv) the corresponding density functions f_p and f_q are bounded with bounded first order derivatives, and satisfy* $\inf_{u \in \mathcal{I}_U} f_p(u) \geq \delta$ *and* $\inf_{v \in I_V} f_q(v) \geq \delta$ *for some constant $\delta > 0$, (v) the functions p_o and q_o are $(l_p + 1)$ and $(l_q + 1)$ times continuously differentiable, respectively.*

Assumption 2. *The kernel function $\mathcal{K}$ is twice continuously differentiable, and satisfies the following conditions: $\int \mathcal{K}(u)du = 1$, $\int u\mathcal{K}(u)du = 0$ and $\int |u^2\mathcal{K}(u)|du < \infty$, and $\mathcal{K}(u) = 0$ for u not contained in some compact set, say $[-1, 1]$.*

Assumption 3. *The function $\psi(z, \theta, p(u), q(v))$ is (i) continuously differentiable with respect to θ, (ii) three times continuously differentiable with respect to $(p(u), q(v))$, with derivatives that are uniformly bounded, (iii) such that the matrix $\Omega_o := \mathbb{E}(\psi_o(Z)\psi_o(Z)^\top)$ is finite, where $\psi_o(Z) = \psi(Z, \theta_o, p_o(U), q_o(V))$, (iv) such that* $\sup \|\partial_\theta \psi(Z, \theta, p(U), q(V)) - \partial_\theta \psi(Z, \theta, p_o(U), q_o(V))\| =$

$o_P(1)$, where the supremum is taken over the (θ, p, q) in some open neighborhood of (θ_o, p_o, q_o), and (v) such that $\Gamma = \mathbb{E}(\partial_\theta \psi(Z, \theta_o, p_o(U), q_o(V)))$ has full rank.

Assumption 4. *The bandwidth sequences h_p and h_q satisfy the following conditions as $n \to \infty$: (i) $nh_p^{2(l_p+1)} h_q^{2(l_q+1)} \to 0$, (ii) $nh_p^{6(l_p+1)} \to 0$, (iii) $nh_q^{6(l_q+1)} \to 0$, (iv) $n^2 h_p^{3d_p} / \log(n)^3 \to \infty$, and (v) $n^2 h_q^{3d_q} / \log(n)^3 \to \infty$.*

Assumption 1 collects smoothness conditions that are standard in the context of nonparametric regression. The restrictions on the kernel function $\mathcal{K}$ in Assumption 2 could be weakened to allow for kernels with unbounded support. Parts (i)-(ii) of Assumption 3 impose some weak smoothness restrictions on the function ψ, which are needed to justify a quadratic expansion. At the cost of a more involved theoretical argument, these assumptions could be relaxed by imposing smoothness conditions on the population functional Ψ instead (cf. Chen et al., 2003). Assumption 3(iii) ensures that the leading term in (3.1) satisfies a central limit theorem, and Assumption 3(iv)–(v) are standard smoothness and invertability conditions. Finally, Assumption 4 imposes restrictions on the rate at which the bandwidths h_p and h_q tend to zero that depend on the number of derivatives of the unknown regression functions and the dimension of the covariates.

3.3. Asymptotic Normality. Our first main result is concerned with the asymptotic normality of SDREs under the conditions that we just imposed.

Theorem 1. *Under Assumption 1–4, equation* (3.1) *holds with $R_n = o_P(n^{-1/2})$. Moreover, we have that $\sqrt{n}(\widehat{\theta} - \theta_o) \xrightarrow{d} N(0, \Gamma_o^{-1} \Omega_o \Gamma_o^{-1})$, and that $\widehat{\Gamma}^{-1} \widehat{\Omega} \widehat{\Gamma}^{-1} \xrightarrow{p} \Gamma_o^{-1} \Omega_o \Gamma_o^{-1}$.*

Theorem 1 shows that our SDRE is asymptotically linear, which immediately implies its $\sqrt{n}$-consistency and asymptotic normality. The asymptotic variance is of the usual sandwich form, and the theorem establishes consistency of a simple sample analogue variance estimator. Taken together, these results can be used to justify various large sample inference procedures, such as e.g. the construction of confidence regions for θ_o. The theorem also shows that SDREs are adaptive, in the sense that their asymptotic variance does not contain an adjustment term for the use of first-step nonparametric estimates. This is a property SDREs share with all semiparametric estimators that take the form of a sample analogue of an influence function in the corresponding model (e.g. Newey, 1994). It also implies that SDREs are

semiparametrically efficient if the DR moment condition is based on the respective *efficient* influence function. This is the case for all examples that we listed in Section 2.2.

Theorem 1 differs from other asymptotic normality results for semiparametric two-step estimators (e.g. Newey, 1994; Newey and McFadden, 1994; Chen et al., 2003; Ichimura and Lee, 2010), because it only imposes relatively weak conditions on the accuracy of the nonparametric first stage estimates. In particular, the bandwidth restrictions in Assumption 4 allow the smoothing bias from estimating either p_o or q_o to be go to zero as slow as $o(n^{-1/6})$ as long as the *product* of the two bias terms is $o(n^{-1/2})$, and only require the respective stochastic parts to be $o_P(n^{-1/6})$; see also (2.5)–(2.6). In contrast, for a generic estimator to be asymptotically normal, the first stage nonparametric estimation error and bias typically have to be $o_p(n^{-1/4})$ and $o(n^{-1/2})$, respectively, in some suitable norm. Another way to interpret this difference is that SDREs require less stringent smoothness conditions on the nuisance functions, which is very important in higher dimensional settings. For example, it is easily verified that if $d_p \leq 5$ and $d_q \leq 5$, there exist bandwidths h_p and h_q such that Assumption 4 is satisfied even if $l_p = l_q = 1$. For a generic estimator that uses an estimate of, say, p_o to be asymptotically normal one typically cannot allow for $l_p = 1$ if $d_p > 1$. SDREs can thus achieve the same first order asymptotic properties as generic semiparametric estimators with lower order local polynomials in the first stage. This is very important in empirical practice: while higher order local polynomial regression leads to estimates with small asymptotic bias, it is also well-known to have poor finite sample properties.

We also remark that in lower dimensional settings the range of bandwidths that is permitted by Assumption 4 includes the values that minimize the Integrated Mean Squared Error (IMSE) for estimating p_o and q_o, respectively. In contrast, a generic semiparametric estimator would not be asymptotically normal with such a choice. While these bandwidths do not have any optimality properties for estimating θ_o, they have the practical advantage that they can be estimated from the data via least-squares cross validation. For many SDREs, there thus exist an objective and feasible data-driven bandwidth selection method that does not rely on preliminary estimates of the nonparametric component. This might be important, since the lack of such a method is one of the major obstacles for applying semiparametric estimators in practice.

3.4. Higher Order Properties. We can strengthen the first part of Theorem 1 by deriving an explicit expression for the rate at which the remainder R_n in the linear representation (3.1) tends to zero. To simplify the exposition, we only state such a result for the case that the arguments of p_o and q_o have the same dimension, that is $d_p = d_q \equiv d$, and that the same bandwidth and order of the local polynomial are used to estimate these two functions, that is $l_p = l_q \equiv l$ and $h_p = h_q \equiv h$. Similar results could be established in more general settings.

Corollary 1. *Under Assumption 1– 4 and the above restrictions, we have that* $T_{n,1} + T_{n,2} = O_P(h^{2(l+1)}) + O_P(n^{-1}h^{-d/2})$*; and the bandwidth that minimizes the order of the sum of these two terms satisfies* $h \propto n^{-2/(4(l+1)+d)}$*. Moreover, with this choice of bandwidth equation* (3.1) *holds with* $R_n = O_P(n^{-4(l+1)/(4(l+1)+d)})$*.*

Again, Corollary 1 documents a substantial advantage of SDREs relative to generic semiparametric two-step estimators. For the latter, arguments analogous to those in Linton (1995), Ichimura and Linton (2005) or Cattaneo et al. (2013a) show that the sum of the "linear" and the "quadratic" term in an analogous expansion would generally only be $O_P(h^{l+1}) + O_P(n^{-1}h^{-d})$, where the two summands corresponds to the orders of the first-stage smoothing bias and variance, respectively.[2] A linear representation analogous to (3.1) could thus at best be obtained with $R_n = O_P(n^{-(l+1)/(l+1+d)})$, which is slower than the rate we get for SDREs. For the simple case with $d = l = 1$, for example, a generic semiparametric estimator differs from its asymptotically linear representation by a term that is at least $O_P(n^{-2/3})$, whereas for our SDREs the difference can be as small as $O_P(n^{-8/9})$. As a consequence, we can expect standard Gaussian approximations based on linear representations like (3.1) to be more accurate in finite samples for our SDREs.

It is common practice to approximate the first-order bias and second-order variance of semiparametric two-step estimators by the mean and variance of the leading terms in a quadratic expansion (Linton, 1995). Corollary 1 therefore implies a reduction of both quantities for SDREs relative to generic semiparametric estimators. Moreover, a careful inspection of the proof of Corollary 1 shows that the term of order $n^{-1}h^{-d/2}$ is actually mean zero, whereas for a generic estimator the term of order $n^{-1}h^{-d}$ is not (Linton, 1995; Cattaneo et al., 2013a). This means that the amount of bias reduction that is achieved by using an

[2]For a generic semiparametric estimator that is linear in the nonparametric component better rates could be obtained, because in this case the "quadratic" term is exactly equal to zero.

SDRE is even bigger than what is immediately apparent from the corollary.

4. Application to Estimation of Treatment Effects

In this section, we apply our theory to the problem of estimating the causal effect of a binary treatment on some outcome variable of interest. See Imbens (2004) and Imbens and Wooldridge (2009) for excellent surveys of the extensive literature on this topic.

4.1. Model and Parameters of Interest. We now provide a more detailed description of the model used in Examples 3 and 4. Following Rubin (1974), we define treatment effects in terms of potential outcomes. Let $Y(1)$ and $Y(0)$ denote the potential outcomes with and without taking some treatment, respectively, with $D = 1$ indicating participation in the treatment, and $D = 0$ indicating non-participation in the treatment. We observe the realized outcome $Y = Y(D)$, but never the pair $(Y(1), Y(0))$. The data consist of a sample from the distribution of $Z = (Y, D, X)$, where X is some vector of covariates that are unaffected by the treatment. We write $\Pi_o = \mathbb{E}(D)$, denote the propensity score by $\pi_o(x) = \mathbb{E}(D|X = x)$, and define the conditional expectation function $\mu_o^Y(d, x) = \mathbb{E}(Y|D = d, X = x)$. We focus on the Population Average Treatment Effect (ATE)

$$\tau_o = \mathbb{E}(Y(1) - Y(0))$$

and the Average Treatment Effect on the Treated (ATT)

$$\gamma_o = \mathbb{E}(Y(1) - Y(0)|D = 1)$$

as our parameters of interest. Since we observe either $Y(1)$ or $Y(0)$, but never both, we have to impose further restrictions on the mechanism that selects individuals into treatment to achieve identification. Here we maintain the assumptions that the selection mechanism is "unconfounded" and satisfies a "strict overlap" condition. Unconfoundedness means that conditional on the observed covariates, the treatment indicator is independent of the potential outcomes, i.e. $(Y(1), Y(0))\perp D|X$ (Rosenbaum and Rubin, 1983). This condition is sometimes also referred to as selection on observables (Heckman and Robb, 1985). Strict overlap means that the propensity score is bounded away from zero and one, i.e. $P(\underline{\pi} < \pi_o(X) < \overline{\pi}) = 1$ for $\underline{\pi} > 0$ and $\overline{\pi} < 1$. This condition is important to ensure that the semiparametric efficiency bounds for estimating our parameters of interest are finite (Khan

and Tamer, 2010). Hahn (1998) derived the semiparametric efficiency bounds for estimating the ATE and the ATT in this setting (under some additional smoothness conditions on the model). That is, he showed that the asymptotic variance of any regular estimator of the ATE and ATT is bounded from below by

$$V^*_{ate} = \mathbb{E}\left(\frac{\sigma^2(1,X)}{\pi_o(X)} + \frac{\sigma^2(0,X)}{1-\pi_o(X)} + (\mu^Y_o(1,X) - \mu^Y_o(0,X) - \tau_o)^2\right) \text{ and}$$
$$V^*_{att} = \mathbb{E}\left(\frac{\pi_o(X)}{\Pi^2_o}\left(\sigma^2(1,X) + \frac{\pi_o(X)\sigma^2(0,X)}{1-\pi_o(X)} + (\mu^Y_o(1,X) - \mu^Y_o(0,X) - \gamma_o)^2\right)\right),$$

respectively, where $\sigma^2(d,x) = \text{Var}(Y|D=d, X=x)$. Semiparametric two-step estimators that achieve these bounds have been studied by Heckman, Ichimura, and Todd (1997), Heckman, Ichimura, Smith, and Todd (1998), Hahn (1998), Hirano et al. (2003) or Imbens, Newey, and Ridder (2005), among others.

Doubly robust estimators of treatment effect parameters that impose additional parametric restrictions on nuisance functions have been studied by Robins et al. (1994), Robins and Rotnitzky (1995), Rotnitzky, Robins, and Scharfstein (1998) and Scharfstein et al. (1999), among others, and are widely used in applied work. Cattaneo (2010) proposed an estimator of the ATE that has the same structure as our SDRE, but did not formally show the favorable properties of this approach relative to other estimators.

4.2. Estimating the Average Treatment Effect for the Population. We now use the methodology developed in Section 2–3 to study a SDRE of the ATE $\tau_o = \mathbb{E}(Y(1) - Y(0))$. Straightforward calculations show that under unconfoundedness we can characterize τ_o through the moment condition

$$\mathbb{E}(\psi_{ate}(Z, \tau_o, \pi_o(X), \mu_o(X))) = 0,$$

where $\mu_o(x) = (\mu^Y_o(1,x), \mu^Y_o(0,x))$ and

$$\psi_{ate}(z, \tau, \pi(x), \mu(x)) = \frac{d(y - \mu^Y(1,x))}{\pi(x)} - \frac{(1-d)(y-\mu^Y(0,x))}{1-\pi(x)} + (\mu^Y(1,x) - \mu^Y(0,x)) - \tau$$

is the efficient influence function for estimating τ_o (Hahn, 1998). It is also easily verified that the above moment condition is doubly robust, and that the orthogonality condition holds because of unconfoundedness. Given nonparametric estimates of the propensity score π_o and

the regression function μ_o^Y, we estimate the ATE by the value that sets a sample version of this moment condition equal to zero. This leads to the estimator

$$\widehat{\tau}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \widehat{\mu}^Y(1,X_i))}{\widehat{\pi}(X_i)} - \frac{(1-D_i)(Y_i - \widehat{\mu}^Y(0,X_i))}{1-\widehat{\pi}(X_i)} + (\widehat{\mu}^Y(1,X_i) - \widehat{\mu}^Y(0,X_i))\right).$$

Since we can anticipate the asymptotic variance of $\widehat{\tau}_{DR}$ to be $\mathbb{E}(\psi_{ate}(Z,\tau_o,\pi_o(X),\mu_o(X))^2)$ from Theorem 1, we can also already define the corresponding estimator as follows:

$$\widehat{V}_{ate}^* = \frac{1}{n}\sum_{i=1}^{n}\psi_{ate}(Z_i,\widehat{\tau}_{DR},\widehat{\pi}_o(X_i),\widehat{\mu}_o(X_i))^2.$$

We define $\widehat{\pi}$ as the l_π-th order "leave-i-out" local polynomial Probit estimator of $\pi_o(x)$ using the bandwidth h_π, and $\widehat{\mu}^Y(d,x)$ as the usual l_μth order "leave-i-out" local polynomial estimator of $\mu_o^Y(d,x)$ using a bandwidth h_μ. That is, using notation analogous to that introduced in Section 2, we define

$$\widehat{\pi}(X_i) = \Phi(\widehat{a}_\pi(X_i)) \quad \text{and} \quad \widehat{\mu}(d,X_i) = \widehat{a}_\mu(d,X_i),$$

respectively, where

$$(\widehat{a}_\pi(X_i),\widehat{b}_\pi(X_i)) = \underset{a,b}{\operatorname{argmin}}\sum_{j\neq i}\left(D_j - \Phi\left(a - \sum_{1\leq|s|\leq l_\pi} b_s(X_j - X_i)^s\right)\right)^2 K_{h_\pi}(X_j - X_i),$$

$$(\widehat{a}_\mu(d,X_i),\widehat{b}_\mu(d,X_i)) = \underset{a,b}{\operatorname{argmin}}\sum_{j\neq i}\mathbb{I}\{D_j = d\}\left(Y_j - a - \sum_{1\leq|s|\leq l_\pi} b_s(X_j - X_i)^s\right)^2 K_{h_\mu}(X_j - X_i),$$

and $\Phi(\cdot)$ is the CDF of the standard normal distribution. Note that we slightly deviate from the general theory presented in Section 2 by using a local polynomial Probit estimator for the propensity score instead of a standard local polynomial smoother. This ensures that the estimator of π_o is bounded between 0 and 1, and should improve the finite-sample properties of the procedure. This choice has no impact on our asymptotic analysis, as it is well known from the work of e.g. Fan, Heckman, and Wand (1995), Hall, Wolff, and Yao (1999) or Gozalo and Linton (2000) that the asymptotic bias of the local polynomial Probit estimator is of the same order of magnitude as that of the usual local polynomial estimator uniformly over the covariates' support, and that the two estimators have the same stochastic behavior.

To study the asymptotic properties of the SDRE $\widehat{\tau}_{DR}$, we impose the following assumptions, which essentially restate the content of Assumption 1 using the notation of the present treatment effects setting.

Assumption 5. *(i) The random vector X is continuously distributed with compact support I_X, (ii) the corresponding density function f_X is bounded with bounded first-order derivatives, and satisfies $\inf_{x\in I_X} f_X(x) \geq \delta$ for some constant $\delta > 0$, and (iii) the function $\pi_o(x)$ is $(l_\pi+1)$ times continuously differentiable.*

Assumption 6. *(i) For any $d \in \{0,1\}$, the random vector X is continuously distributed conditional on $D = d$ with compact support I_X, (ii) the corresponding density functions $f_{X|d}$ are bounded with bounded first-order derivatives, and satisfy $\inf_{x\in I_X} f_{X|d}(x) \geq \delta$ for some constant $\delta > 0$ and any $d \in \{0,1\}$, (iii) $\sup_{x\in I_x, d\in\{0,1\}} \mathbb{E}(|Y|^c|X = x, D = d) < \infty$ for some constant $c > 2$ and any $d \in \{0,1\}$ (iv) the function $\mu_o(d,x)$ is $(l_\mu + 1)$ times continuously differentiable with respect to its second argument for any $d \in \{0,1\}$.*

The following Theorem establishes the asymptotic properties of the SDRE $\widehat{\tau}_{DR}$.

Theorem 2. *Suppose Assumption 5–6 hold, and that Assumption 2–4 hold with $(l_p, d_p, h_p) = (l_\pi, d_X, h_\pi)$ and $(l_q, d_q, h_q) = (l_\mu, d_X, h_\mu)$. Then the following holds:*

i) $\widehat{\tau}_{DR} \xrightarrow{p} \tau_o$, and $\sqrt{n}(\widehat{\tau}_{DR} - \tau_o) \xrightarrow{d} N(0, V^*_{ate})$, and thus $\widehat{\tau}_{DR}$ achieves the semiparametric efficiency bound for estimating τ_o.

ii) If the conditions of the theorem are satisfied with $l_\pi = l_\mu \equiv l$ and $h_\pi \propto h_\mu \propto n^{-2/(8l+d_X)}$, then $\widehat{\tau}_{DR} - \tau_o = n^{-1}\sum_{i=1}^n \psi_{ate}(Z_i, \tau_o, \pi_o(X_i), \mu_o(X_i)) + O_P(n^{-8l/(8l+d_X)})$.

iii) $\widehat{V}^*_{ate} \xrightarrow{p} V^*_{ate}$.

Theorem 3 shows that the semiparametric DR estimator $\widehat{\tau}_{DR}$ enjoys the same efficiency property as e.g. the Inverse Probability Weighting estimator of Hirano et al. (2003), which is based on the moment condition $\tau_o = \mathbb{E}(DY/\pi_o(X) + (1-D)Y/(1-\pi_o(X)))$, or the Regression estimator of Imbens et al. (2005), which is based on the moment condition $\tau_o = \mathbb{E}(\mu^Y_o(1,X) - \mu^Y_o(0,X))$. However, following the discussion after Theorem 1, the SDRE has a number of theoretical and practical advantages relative to kernel-based versions of these estimators,[3] that make it preferable to be used in practice.

[3]Both Hirano et al. (2003) and Imbens et al. (2005) consider series estimation in the first stage, and thus their results are not directly comparable to ours. See Ichimura and Linton (2005) for an analysis of the Inverse Probability Weighting estimator when the propensity score is estimated via local linear regression.

Remark 1 (Selection of Tuning Parameters)**.** Implementing the estimator $\widehat{\tau}_{DR}$ requires choosing two types of tuning parameters for the nonparametric estimation step: the bandwidths and the order of the local polynomials. We recommend using $l_\pi = l_\mu = 1$ as long as $d_X \leq 5$, as such a choice is compatible with the asymptotic theory and local linear regression estimators are well-known to have superior small-sample properties relative to higher order local polynomial smoothers. If $d_X \leq 3$, our theory also allows choosing the bandwidths that minimize a least-squares cross validation criterion, i.e. using

$$h_\pi = \underset{h}{\operatorname{argmin}} \sum_{i=1}^{n} (D_i - \widehat{\pi}(X_i))^2 \text{ and } h_\mu = \underset{h}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \widehat{\mu}^Y(D_i, X_i))^2.$$

As pointed out above, such a choice has no particular optimality properties for estimating τ_o, but it has the advantage of being objective, data-driven, and easily implementable.

4.3. Estimating the Average Treatment Effect for the Treated. In this section, we consider semiparametric DR estimation of the Average Treatment Effect for the Treated $\gamma_0 = \mathbb{E}(Y(1) - Y(0)|D = 1)$. Again, straightforward calculations show that under unconfoundedness we can characterize γ_o through the moment condition

$$\mathbb{E}(\psi_{att}(Z, \tau_{ate}, \pi_o(x), \mu_o^Y(0, x), \Pi_o) = 0,$$

where

$$\psi_{att}(z, \gamma, \pi(x), \mu^Y(0, x), \Pi) = \frac{d(y - \mu^Y(0, x))}{\Pi} - \frac{\pi(x)}{\Pi} \cdot \frac{(1 - d)(y - \mu^Y(0, x))}{1 - \pi(x)} - \gamma.$$

It is also easily verified that this moment condition is doubly robust with respect to the two nuisance functions, and that the orthogonality condition holds because of unconfoundedness. Given the same nonparametric estimators of the propensity score π_o and the regression function μ_o^Y we defined above, and setting $\widehat{\Pi} = \sum_{i=1}^n D_i/N$, the SDRE of the ATT is given by the value that sets a sample version of this moment condition equal to zero, namely

$$\widehat{\gamma}_{DR} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{D_i(Y_i - \widehat{\mu}^Y(0, X_i))}{\widehat{\Pi}} - \frac{\widehat{\pi}(X_i)}{\widehat{\Pi}} \cdot \frac{(1 - D_i)(Y_i - \widehat{\mu}^Y(0, X_i))}{1 - \widehat{\pi}(X_i)} \right).$$

Since from Theorem 1 we can anticipate the form of the asymptotic variance of $\widehat{\gamma}_{DR}$, we can also already define its estimator as follows:

$$\widehat{V}_{att}^* = \frac{1}{n} \sum_{i=1}^{n} \psi_{att}(Z_i, \widehat{\gamma}_{DR}, \widehat{\pi}_o(X_i), \widehat{\mu}_o^Y(0, X_i), \widehat{\Pi})^2.$$

The following Theorem establishes the estimator's asymptotic properties.

Theorem 3. *Suppose Assumption 5–6 hold, and that Assumption 2–4 hold with* $(l_p, d_p, h_p) = (l_\pi, d_X, h_\pi)$ *and* $(l_q, d_q, h_q) = (l_\mu, d_X, h_\mu)$. *Then*

i) $\widehat{\gamma}_{DR} \xrightarrow{p} \gamma_o$, *and* $\sqrt{n}(\widehat{\gamma}_{DR} - \gamma_o) \xrightarrow{d} N(0, V^*_{att})$, *and thus* $\widehat{\gamma}_{DR}$ *achieves the semiparametric efficiency bound* γ_o *in the absence of knowledge of the propensity score.*

ii) If the conditions of the theorem are satisfied with $l_\pi = l_\mu \equiv l$ *and* $h_\pi \propto h_\mu \propto n^{-2/(8l+d_X)}$, *then* $\widehat{\gamma}_{DR} - \gamma_o = n^{-1} \sum_{i=1}^n \psi_{att}(Z_i, \gamma_o, \pi_o(X_i), \mu^Y_o(0, X_i), \Pi_o) + o_P(n^{-8l/(8l+d_X)})$.

iii) $\widehat{V}^*_{att} \xrightarrow{p} V^*_{att}$.

The discussion after Theorem 3 applies analogously to the result in Theorem 4. The SDRE of the ATT is not only semiparametrically efficient, but its properties also compare favorably to those of other efficient estimators that use only a nonparametric estimate of either the propensity score $\pi_o(\cdot)$ (e.g. Hirano et al., 2003) or the regression function $\mu^Y_o(0, \cdot)$ (e.g. Imbens et al., 2005).

5. Monte Carlo

In this section, we illustrate the finite sample properties of SDREs through a small scale Monte Carlo experiment, and compare them to those of other semiparametric two-step estimators. We consider the simple missing data model presented in Example 1 above: the covariate X is uniformly distributed on the interval $[0, 1]$, the outcome variable Y is normally distributed with mean $\mu_o(X) = (3X - 1)^2$ and variance 1, and the missingness indicator D is generated as a Bernoulli random variable with mean $\pi_o(X) = 1 - .2 \times \mu_o(X)$. Our parameter of interest is $\theta_o = \mathbb{E}(Y) = 1$, and the semiparametric variance bound for estimating this parameter is $V^* \approx 2.632$. We study the sample size $n = 200$, and set the number of replications to 5,000. We consider three estimators of $\theta_o = \mathbb{E}(Y)$, namely the semiparametric doubly robust one based on a sample analogue of the efficient influence function (DR),

inverse probability weighting (IPW), and a regression-based estimator (REG):

$$\widehat{\theta}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \widehat{\mu}(X_i))}{\widehat{\pi}(X_i)} + \widehat{\mu}(X_i)\right)$$
$$\widehat{\theta}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i}{\widehat{\pi}(X_i)}$$
$$\widehat{\theta}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}(X_i).$$

We define $\widehat{\pi}$ as the "leave-i-out" local linear Probit estimator of $\pi_o(x)$ using the bandwidth $h \in \{.1, .15, \ldots, .6\}$, and $\widehat{\mu}(x)$ as the "leave-i-out" local linear estimator of $\mu_o(x)$ using a bandwidth $g \in \{.035, .05, \ldots, .185\}$. The construction of these nonparametric estimators is analogous to that described in Section 4. We also consider nominal $(1-\alpha)$ confidence intervals of the usual form

$$CI_j^{1-\alpha} = \left[\widehat{\theta}_j \pm \Phi^{-1}(1-\alpha/2)(\widehat{V}_j/n)^{1/2}\right]$$

with $\Phi^{-1}(\alpha)$ the α quantile of the standard normal distribution and

$$\widehat{V}_j = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \widehat{\mu}(X_i))}{\widehat{\pi}(X_i)} + \widehat{\mu}(X_i) - \widehat{\theta}_j\right)^2$$

an estimate of the asymptotic variance, for $j \in \{DR, IPW, REG\}$.

Our simulation results generally confirm the predictions of our asymptotic theory. In Figure 1, we plot the Mean Squared Error (MSE), the (absolute) bias, and the variance of the IPW estimator as a function of the bandwidth h, and compare the results to those of DR estimators using various values of the bandwidth g. In Figure 2, we plot the same three quantities for the REG estimator as a function of the bandwidth g, and compare the results to those of DR estimators using various values of the bandwidth h. Clearly, the bias of both IPW and REG varies substantially with the respective bandwidth. To a lesser extend, this applies also to the variances of the two estimators, especially in the case of IPW. As a consequence, the MSE shows strong dependence on the bandwidth in both cases. It is minimized for $h = .4$ and $g = .05$, respectively, but these values would be very difficult to determine by some rule of thumb in an empirical application.[4] Moreover, in both cases the

[4]To give some point of reference, in this setting the average bandwidth values selected by least squares cross validation are equal to about .1 for both the propensity score and the regression function. Our graphs show that both IPW and REG do not perform well with such a choice of bandwidth.

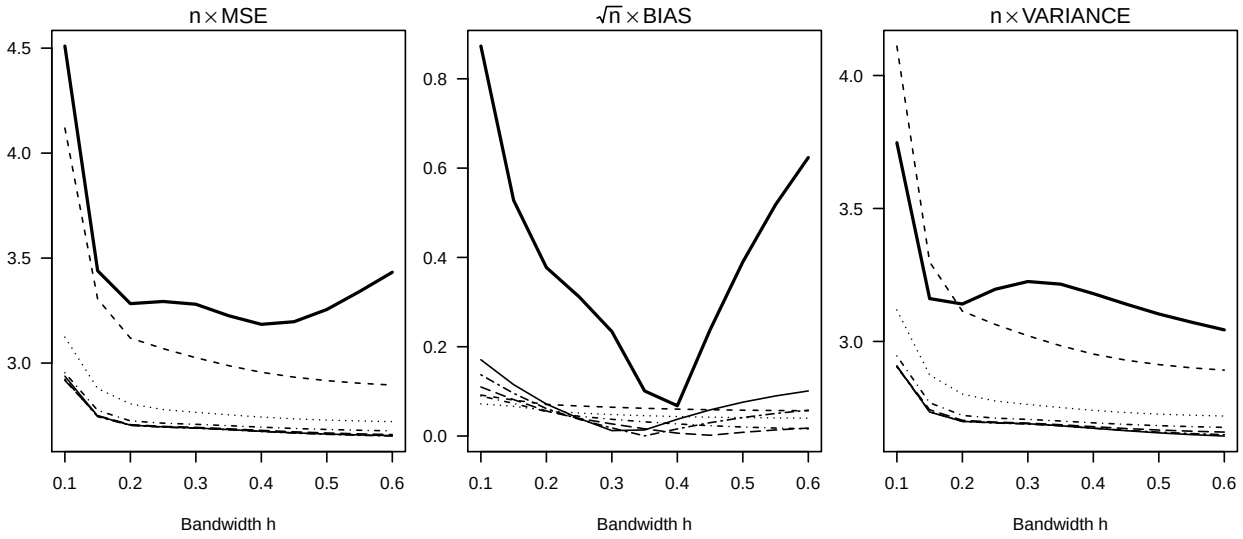

Figure 1: Simulation results: MSE, absolute bias and variance of the IPW estimator for various values of h (bold solid line), compared to results for the DR estimator with bandwidth g equal to .035 (short-dashed line), .065 (dotted line), .095 (dot-dashed line), .125 (long dashed line), .155 (long dashed dotted line), and .185 (thin solid line).

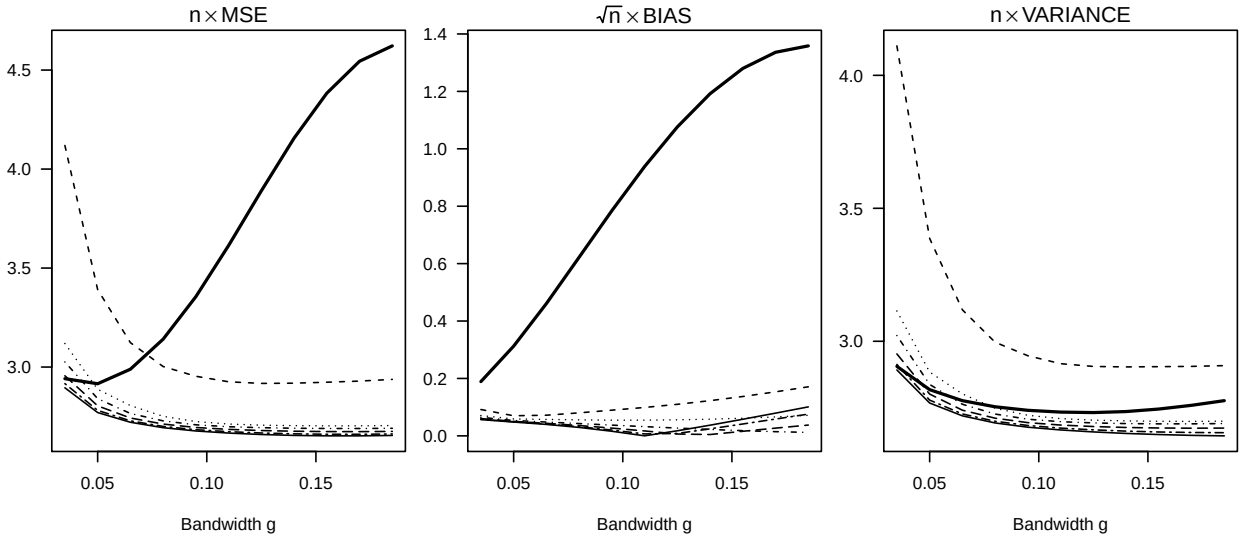

Figure 2: Simulation results: MSE, absolute bias and variance of the REG estimator for various values of g (bold solid line), compared to results for the DR estimator with bandwidth h equal to .1 (short-dashed line), .2 (dotted line), .3 (dot-dashed line), .4 (long dashed line), .5 (long dashed dotted line), and .6 (thin solid line).

minimum is much larger than the asymptotic variance $V^* \approx 2.632$ that it is supposed to achieve.

For the DR estimators, we observe that those using one of the two smallest bandwidths, i.e. either $h = .1$ or $g = .035$, exhibit somewhat different behavior from the remaining ones. For DR estimators using $h > .1$ and $g > .035$, the MSE, bias and variance are all very similar, and exhibit only little variation with respect to the bandwidth. The variance of these DR estimators is substantially lower than that of IPW, and broadly similar to that of REG (with some minor gains for larger values of g). The variance is also very close to the semiparametric efficiency bound $V^* \approx 2.632$. These DR estimators also have very little bias for all bandwidth choices. DR estimators that use either $h = .1$ or $g = .035$, the two smallest bandwidth values, have somewhat higher variance than those using larger bandwidths, but are also essentially unbiased. As a consequence, they still compare favorably to both IPW and REG in terms of MSE.

We also compute the empirical coverage probabilities of the confidence intervals $CI_j^{0.95}$ for $j \in \{DR, IPW, REG\}$, using again various bandwidths for estimating the nonparametric components. Results are reported in Table 1. Note that computing a confidence interval for θ_o based on the IPW estimator requires an estimate of μ_o, and similarly a confidence interval based on the REG estimator requires and estimate of π_o. Therefore all confidence intervals vary with respect to both bandwidth parameters. Our results show that the coverage probability of DR-based confidence intervals is extremely close to its nominal value for all combinations of bandwidths we consider. IPW-based confidence intervals exhibit slight under-coverage all values of the two bandwidth. REG-based confidence intervals have good coverage properties for $g = .035$ and increasing under-coverage for larger values of g, irrespective of the choice of h.

6. Empirical Application

In order to investigate the relative performance of SDREs using actual data, we conduct a small-scale study on the effect of maternal smoking during pregnancy on birth weight. We use the same dataset as in Almond, Chay, and Lee (2005) who study the economic costs of low birth weight using several non-experimental techniques. That same dataset was also used in a recent paper by Cattaneo (2010), who exploits the fact that mothers report their daily smoking intensity to apply his semiparametric method of estimation of multi-valued

Table 1: Simulation Results: Empirical coverage probability of nominal 95% confidence intervals based on either the DR, IPW or REG estimator, for various bandwidth values.

DR	*g / h*	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*
	.035	0.948	0.947	0.946	0.945	0.942	0.941
	.065	0.951	0.949	0.949	0.949	0.948	0.948
	.095	0.954	0.953	0.952	0.951	0.950	0.949
	.125	0.952	0.953	0.950	0.949	0.948	0.947
	.155	0.952	0.950	0.950	0.948	0.947	0.946
	.185	0.952	0.950	0.949	0.950	0.946	0.943
IPW	*g / h*	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*
	.035	0.920	0.936	0.933	0.930	0.922	0.913
	.065	0.912	0.933	0.926	0.930	0.920	0.907
	.095	0.909	0.930	0.926	0.928	0.917	0.905
	.125	0.908	0.928	0.926	0.927	0.915	0.902
	.155	0.909	0.927	0.927	0.926	0.917	0.902
	.185	0.910	0.928	0.928	0.927	0.916	0.901
REG	*g / h*	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*
	.035	0.949	0.946	0.945	0.943	0.941	0.941
	.065	0.942	0.939	0.941	0.940	0.937	0.937
	.095	0.935	0.932	0.933	0.930	0.928	0.926
	.125	0.909	0.909	0.907	0.904	0.903	0.900
	.155	0.883	0.881	0.878	0.877	0.870	0.868
	.185	0.875	0.869	0.869	0.865	0.861	0.855

treatment effects.

The original data is a very rich database of 497,139 singleton births that took place in Pennsylvania between 1989 and 1991. As we are using the data for illustration purposes only, we randomly selected 5,000 observations from the original data and kept a few covariates in order to decrease the complexity. Also, to have a relatively homogeneous sample, we only kept non-hispanic mothers who consumed no alcohol during pregnancy, and who were in the 14–38 age range. In our sample, both parents also have at least 8 years of schooling. After applying these filters, we ended up with 4,317 observations.

Table 2: Summary Statistics: Means and Standard Deviations (in paranthesis)

	Overall	*Smoker*	*Non-Smoker*	*Diff.*	*T-Stat*
Smoker	0.17	-	-	-	-
	(0.38)	-	-	-	-
Birth weight	3377.82	3164.65	3421.50	-256.85	-11.16
	(576.29)	(562.46)	(569.39)		
Mother Married	0.25	0.45	0,21	0.25	14.30
	(0.43)	(0.50)	(0.41)		
Mother's Age	26.76	25.34	27.05	-1.71	-8.05
	(5.27)	(5.1)	(5.26)		
Num. Obs.	4317	734	3583	-	-

Our treatment variable is a dummy variable that equals one if the mother smoked during pregnancy and zero otherwise, whereas our response variable is birth weight measured in grams. The covariates are mother's age and a dummy for being married, as those are in our sample the most important 'determinants' of both smoking during pregnancy and birth weight.[5] In Table 2, we report some summary statistics for the variables we use in this section.

Our parameter of interest is the Average Treatment Effect on the Treated (ATT). Almond et al. (2005) found a strong negative effect of about 200 to 250 grams of maternal smoking on birth weight using both subclassification on the propensity score and regression adjusted

[5]The reasoning for using that particular sample and maintaining these few covariates is the following. A quick inspection of data revealed bunching on father's and mother's education at 0, and that was mostly likely caused by misreporting, especially for father's information. We then dropped all births in which parents have less than 8 years of schooling. A multiple regression of treatment dummy on all other covariates revealed that maternal alcohol consumption, mother being hispanic and being married were important determinants of smoking habits, along with a few others variables. However, given that there were only few expecting mothers that were hispanic or had alcohol consumption habits, we dropped them from our data and re-ran the same regression but separately for the subsamples of married and unmarried mothers. We dropped all regressors with t-statistics below 2 in at least one subsample regression. We ended up with four covariates. We then ran a regression of birth weight on the treatment dummy and the remaining five regressors. Beyond the treatment dummy, only the dummy for being married and mother's age were significant at the 10% level, and those were the variables we kept in our analysis.

methods to estimate the ATT. For both (i) the conditional expectation of birth weight given covariates for the non-smoking mothers, $\mu_o^Y(0,x)$, and (ii) the probability of smoking given covariates (the propensity-score), $\pi_o(x)$, they used parametric specifications. We estimate both functions nonparametrically, using local linear regression for (i) and local linear probit regression for (ii). In both cases we apply the leave-i-out version of the estimator, as discussed previously in subsections 4.2 and 4.3, and used Gaussian kernels. Given that we have a continuous variable, mother's age, and a binary one, the dummy for being married, we therefore divided the sample between married and single mothers and estimated (i) and (ii) for each subsample. We then used those estimates to calculate the semiparametric doubly robust (DR), regression-based (REG) and inverse probability weighting (IPW) estimators. Following previous notation, these estimators are defined as

$$\widehat{\gamma}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \widehat{\mu}^Y(0,X_i))}{\widehat{\Pi}} - \frac{\widehat{\pi}(X_i)}{\widehat{\Pi}}\cdot\frac{(1-D_i)(Y_i - \widehat{\mu}^Y(0,X_i))}{1-\widehat{\pi}(X_i)}\right),$$

$$\widehat{\gamma}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\widehat{\Pi}} - \frac{\widehat{\pi}(X_i)}{\widehat{\Pi}}\cdot\frac{(1-D_i)Y_i}{1-\widehat{\pi}(X_i)}\right),$$

and

$$\widehat{\gamma}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \widehat{\mu}^Y(0,X_i))}{\widehat{\Pi}}\right).$$

We used the cross-validation criterion presented in subsection 4.2 to select a baseline bandwidth for our kernel estimators. For object (i), the conditional expectation, the baseline bandwidth, g^*, was 2.1 standard deviations of mother's age, whereas for object (ii), the propensity-score, the baseline bandwidth, h^*, was 0.81 standard deviation of mother's age. Results are presented for seven choices of bandwidths: 1/10, 1/2, 3/4, 1, 4/3, 2, and 10 times the respective baseline bandwidth.

In Table 3 we report ATT point estimates. For different choices of bandwidths, the results range from 198.8 to 207.4 for the REG estimator, from 153.8 to 457.5 for the IPW estimator and from 172.4 to 208.2 for the DR estimator. However, if we fix g, the bandwidth for (i), at g^* and let h assume different values, DR estimates range from 173.1 to 206, whereas IPW estimates vary from 153.8 to 457.5. And if we fix h, the bandwidth for (ii), at h^* and let g assume different values, DR estimates range from 204.0 to 207.8, whereas REG estimates vary from 198.8 to 207.4. In both cases, we can see that the DR estimator is much less sensitive to the choice of the bandwidth relative to the other two estimators. Finally, note

Table 3: Point estimates of ATT of maternal smoking on birth weight (in grams) using DR, IPW or REG estimators, for various bandwidth values (multiples of baseline bandwidth)

DR	*g / h*	*0.1*	*0.5*	*0.75*	*1*	*1.33*	*2*	*10*
	0.1	-178.7	-208.2	-207.8	-207.8	-207.7	-207.6	-207.5
	0.5	-173.8	-206.3	-205.1	-204.7	-204.4	-204.1	-203.9
	0.75	-173.5	-206.1	-204.9	-204.4	-203.9	-203.3	-202.7
	1	-173.1	-206.0	-204.8	-204.2	-203.7	-202.9	-202.0
	1.33	-172.8	-206.0	-204.7	-204.1	-203.5	-202.7	-201.5
	2	-172.6	-206.0	-204.6	-204.1	-203.4	-202.5	-201.1
	10	-172.4	-205.9	-204.6	-204.0	-203.4	-202.3	-200.8
IPW	*h*	*0.1*	*0.5*	*0.75*	*1*	*1.33*	*2*	*10*
		-457.5	-201.8	-182.0	-163.4	-153.8	-162.6	-206.0
REG	*g*	*0.1*	*0.5*	*0.75*	*1*	*1.33*	*2*	*10*
		-207.4	-198.8	-199.1	-199.8	-200.4	-200.4	-201.3

that when fixing one bandwidth at any value, not only at the baseline, DR estimates always have the smallest range of variation as a function of the other bandwidth.

Even though REG and IPW point estimates do not vary with h and g, respectively, their standard errors do. To see this, note that for each of the three estimators one has to compute the respective estimate of asymptotic variance

$$\widehat{V}^*_{att,j} = \frac{1}{n}\sum_{i=1}^{n} \psi_{att}(Z_i, \widehat{\gamma}_j, \widehat{\pi}(X_i), \widehat{\mu}^Y(0, X_i), \widehat{\Pi})^2,$$

where $j =$ DR, IPW, REG. One can see that the estimator of the asymptotic variance does depend on estimates of both $\mu^Y_o(0,x)$ and $\pi_o(x)$ and because standard errors are simply $se(\widehat{\gamma}_j) = (\widehat{V}^*_{att,j}/n)^{1/2}$ they will thus depend on the choice of first-stage smoothing parameters. In Table 4 we report 90% confidence intervals for γ_o based on the usual first order asymptotic approximation for each one of these three estimators. Specifically, they are calculated using the formula

$$CI^{0.9}_j = \left[\widehat{\gamma}_j \pm \Phi^{-1}(0.95) se\left(\widehat{\gamma}_j\right)\right].$$

Results from Table 4 point out that there is very little variation in terms of CI's width. In fact, if we fix g, the bandwidth for (i), at g^* and let h assume different values, width of DR confidence intervals ranges from 82.1 to 153.9, width of IPW confidence intervals varies

from 82.1 to 154.5 and width of REG confidence intervals varies from 82.1 to 153.9. In all three cases, the longest confidence intervals result from using the extremely small bandwidth $h = h^*/10$. On the other hand, if we fix h, the bandwidth for (ii), at h^* and let g assume different values, the width of confidence intervals based on DR, IPW and REG presents the same range values from 82.2 to 82.5. A similar pattern occurs for fixing g and h separately at any fixed value.

These results show that even though point estimates do depend on the choice of estimator and smoothing parameters, standard errors seem to be much more stable as they do not vary much with the type of estimator. The only occasion we obtain very high values for standard errors is when substantially undersmoothing the propensity-score. In that case, if we use $h = h^*/10$, then standard errors will be almost as twice the value obtained using all other bandwidth values considered. It seems that when using such a small bandwidth the estimated propensity score becomes close to one in some regions of the covariate space, which in turn gives very large weight to some observations when calculating IPW and DR.

7. Conclusions

Semiparametric two-step estimation based on a doubly robust moment condition is a highly promising methodological approach in a wide range of empirically relevant models, including many applications that involve missing data or the evaluation of treatment effects. Our results suggest that SDREs have favorable properties relative to other semiparametric estimators that are currently widely used in such settings, such as e.g. Inverse Probability Weighting, and should thus be of particular interest to practitioners in these areas. From a more theoretical point of view, we have shown that SDREs are generally $\sqrt{n}$-consistent and asymptotically normal under weaker conditions on the smoothness of the nuisance functions, or, equivalently, on the accuracy of the first step nonparametric estimates, than those commonly used in the literature on semiparametric estimation. As a consequence, the stochastic behavior of SDREs can be better approximated by classical first-order asymptotics. We view these results as an important contribution to a recent literature that aims at improving the accuracy of inference in semiparametric models (e.g. Robins et al., 2008; Cattaneo et al., 2013a,b).

Table 4: 90% confidence intervals for ATT using DR, IPW or REG estimators, for various bandwidth values (multiples of baseline bandwidth)

DR		LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB
	g / h	*0.1*		*0.5*		*0.75*		*1*		*1.33*		*2*		*10*	
	0.1	-246.9	-110.5	-249.7	-166.6	-249.2	-166.5	-249.0	-166.5	-248.9	-166.6	-248.9	-166.4	-249.1	-165.9
	0.5	-250.3	-97.4	-247.7	-164.8	-246.4	-163.9	-245.8	-163.6	-245.4	-163.3	-245.2	-163.0	-245.4	-162.4
	0.75	-250.1	-96.8	-247.6	-164.7	-246.1	-163.7	-245.5	-163.3	-244.9	-162.9	-244.4	-162.2	-244.2	-161.3
	1	-250.0	-96.2	-247.5	-164.6	-246.0	-163.5	-245.3	-163.1	-244.7	-162.6	-244.0	-161.8	-243.4	-160.5
	1.33	-250.0	-95.7	-247.4	-164.6	-245.9	-163.5	-245.2	-163.0	-244.6	-162.5	-243.8	-161.6	-242.9	-160.0
	2	-249.9	-95.2	-247.4	-164.5	-245.9	-163.4	-245.2	-163.0	-244.5	-162.4	-243.6	-161.4	-242.5	-159.6
	10	-249.8	-94.9	-247.3	-164.5	-245.8	-163.4	-245.1	-162.9	-244.4	-162.3	-243.4	-161.2	-242.2	-159.3
IPW		LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB
	g / h	*0.1*		*0.5*		*0.75*		*1*		*1.33*		*2*		*10*	
	0.1	-526.1	-389.0	-243.3	-160.2	-223.4	-140.6	-204.6	-122.2	-195.0	-112.6	-203.8	-121.3	-247.6	-164.4
	0.5	-534.3	-380.7	-243.2	-160.3	-223.2	-140.7	-204.5	-122.3	-194.9	-112.8	-203.7	-121.5	-247.5	-164.6
	0.75	-534.5	-380.5	-243.2	-160.3	-223.2	-140.8	-204.5	-122.3	-194.9	-112.8	-203.7	-121.5	-247.5	-164.6
	1	-534.8	-380.3	-243.2	-160.3	-223.2	-140.8	-204.5	-122.3	-194.9	-112.8	-203.7	-121.5	-247.5	-164.6
	1.33	-535.0	-380.1	-243.2	-160.3	-223.2	-140.8	-204.5	-122.3	-194.9	-112.8	-203.7	-121.5	-247.5	-164.6
	2	-535.2	-379.9	-243.2	-160.3	-223.2	-140.8	-204.5	-122.3	-194.9	-112.8	-203.7	-121.5	-247.5	-164.6
	10	-535.3	-379.7	-243.2	-160.3	-223.2	-140.8	-204.5	-122.3	-194.9	-112.8	-203.7	-121.5	-247.5	-164.6
REG		LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB
	g / h	*0.1*		*0.5*		*0.75*		*1*		*1.33*		*2*		*10*	
	0.1	-275.6	-139.1	-248.9	-165.8	-248.7	-166.0	-248.6	-166.1	-248.5	-166.2	-248.6	-166.1	-249.0	-165.7
	0.5	-275.2	-122.3	-240.2	-157.3	-240.0	-157.5	-239.9	-157.6	-239.8	-157.7	-239.9	-157.6	-240.2	-157.3
	0.75	-275.8	-122.4	-240.5	-157.7	-240.3	-157.9	-240.2	-158.0	-240.1	-158.1	-240.2	-158.0	-240.6	-157.6
	1	-276.7	-122.9	-241.2	-158.4	-241.0	-158.6	-240.9	-158.7	-240.8	-158.8	-240.9	-158.7	-241.3	-158.3
	1.33	-277.5	-123.2	-241.8	-158.9	-241.6	-159.1	-241.5	-159.3	-241.4	-159.3	-241.5	-159.3	-241.8	-158.9
	2	-278.2	-123.6	-242.3	-159.4	-242.1	-159.6	-242.0	-159.8	-241.9	-159.8	-242.0	-159.8	-242.3	-159.4
	10	-278.8	-123.8	-242.7	-159.9	-242.5	-160.1	-242.4	-160.2	-242.3	-160.3	-242.4	-160.2	-242.8	-159.8

A. Proofs of Main Results

A.1. Proof of Theorem 1. To prove the first statement, note that it follows from the differentiability of ψ with respect to θ and the definition of $\widehat{\theta}$ that

$$\widehat{\theta} - \theta_o = \Gamma_n(\theta^*, \widehat{p}, \widehat{q})^{-1} \frac{1}{n} \sum_{i=1}^{n} \psi(Z_i \theta_o, \widehat{p}(U_i), \widehat{q}(V_i))$$

for some intermediate value θ^* between θ_o and $\widehat{\theta}$, and $\Gamma_n(\theta, p, q) = \sum_{i=1}^{n} \partial\psi(Z_i\theta, \widehat{p}(U_i), \widehat{q}(V_i))/\partial\theta$. It also follows from standard arguments that $\Gamma_n(\theta^*, \widehat{p}, \widehat{q}) = \Gamma_o + o_P(1)$. Next, we consider an expansion of the term $n^{-1}\sum_{i=1}^{n} \psi(Z_i\theta_o, \widehat{p}(U_i), \widehat{q}(V_i))$. Using the notation that

$$\begin{aligned}
\psi^p(Z_i) &= \partial\psi(Z_i, t, q_o(V_i))/\partial t|_{t=p_o(U_i)}, \\
\psi^{pp}(Z_i) &= \partial^2\psi(Z_i, t, q_o(V_i))/\partial t|_{t=p_o(U_i)}, \\
\psi^q(Z_i) &= \partial\psi(Z_i, p_o(U_i), t)/\partial t|_{t=q_o(V_i)}, \\
\psi^{qq}(Z_i) &= \partial^2\psi(Z_i, p_o(U_i), t)/\partial t|_{t=q_o(V_i)}, \\
\psi^{pq}(Z_i) &= \partial^2\psi(Z_i, t_1, t_2)/\partial t_1 \partial t_2|_{t_1=p_o(U_i), t_2=q_o(V_i)},
\end{aligned}$$

we find that by Assumption 3 we have that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \psi(Z_i\theta_o, \widehat{p}(U_i), \widehat{q}(V_i)) - \frac{1}{n}\sum_{i=1}^{n} \psi(Z_i\theta_o, p_o(U_i), q_o(V_i)) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \psi^p(Z_i)(\widehat{p}(U_i) - p_o(U_i)) + \frac{1}{n}\sum_{i=1}^{n} \psi^q(Z_i)(\widehat{q}(V_i) - q_o(V_i)) \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n} \psi_i^{pp}(\widehat{p}(U_i) - p_o(U_i))^2 + \frac{1}{n}\sum_{i=1}^{n} \psi_i^{qq}(\widehat{q}(V_i) - q_o(V_i))^2 \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n} \psi^{pq}(Z_i)(\widehat{p}(U_i) - p_o(U_i))(\widehat{q}(V_i) - q_o(V_i)) \\
&\qquad + O_P(\|\widehat{p} - p_o\|_\infty^3) + O_P(\|\widehat{q} - q_o\|_\infty^3).
\end{aligned}$$

By Lemma 2(i) and Assumption 4, the two "cubic" remainder terms are both of the order $o_P(n^{-1/2})$. In Lemma 4–6 below, we show that the remaining five terms on the right hand side of the previous equation are also all of the order $o_P(n^{-1/2})$ under the conditions of the theorem. This completes the proof of the first statement of the theorem. The asymptotic normality result then follows from a simple application of the Central Limit Theorem. The proof of consistency of the variance estimator is standard, and thus omitted. □

A.2. Proof of Corollary 1. Following the proofs of Lemma 4–6 below, we find that $T_{n,1}+T_{n,2} = O(h^{2(l+1)}) + O_P(n^{-1}h^{-d/2})$ under the conditions of the corollary. The remainder of the result then follows from the convergence rate of the "cubic" remainder terms; see e.g. Lemma 2(i).

A.3. Proof of Theorem 2 and 3. These two results can be shown using the same arguments as for the proof of Theorem 1. □.

B. Auxiliary Results

In this section, we collect a number of auxiliary results that are used to prove our main theorems. The results in Sections B.1 and B.2 are minor variations of existing ones and are mainly stated for completeness. The result in Section B.3 is simple to obtain and stated separately again mainly for convenience. Section B.4 contains a number of important and original lemma that form the basis for our proof of Theorem 1.

B.1. Rates of Convergence of U-Statistics. For a real-valued function $\varphi_n(x_1, \ldots, x_k)$ and an i.i.d. sample $\{X_i\}_{i=1}^n$ of size $n > k$, the term

$$U_n = \frac{(n-k)!}{n!} \sum_{s \in \mathcal{S}(n,k)} \varphi_n(X_{s_1}, \ldots, X_{s_k})$$

is called a kth order U-statistic with kernel function φ_n, where the summation is over the set $\mathcal{S}(n,k)$ of all $n!/(n-k)!$ permutations $(s_1, \ldots, s_k)$ of size k of the elements of the set $\{1, 2, \ldots, n\}$. Without loss of generality, the kernel function φ_n can be assumed to be symmetric in its k arguments. In this case, the U-statistic has the equivalent representation

$$U_n = \binom{n}{k}^{-1} \sum_{s \in \mathcal{C}(n,k)} \varphi_n(X_{s_1}, \ldots, X_{s_k}),$$

where the summation is over the set $\mathcal{C}(n,k)$ of all $\binom{n}{k}$ combinations $(s_1, \ldots, s_k)$ of k of the elements of the set $\{1, 2, \ldots, n\}$ such that $s_1 < \ldots < s_k$. For a symmetric kernel function φ_n and $1 \leq c \leq k$, we also define the quantities

$$\varphi_{n,c}(x_1, \ldots, x_c) = \mathbb{E}(\varphi_n(x_1, \ldots, x_c, X_{c+1}, \ldots, X_k) \quad \text{and}$$
$$\rho_{n,c} = \mathrm{Var}(\varphi_{n,c}(X_1, \ldots, X_c))^{1/2}.$$

If $\rho_{n,c} = 0$ for all $c \leq c^*$, we say that the kernel function φ_n is c^*th order degenerate. With this notation, we give the following result about the rate of convergence of a kth order U-statistic with a kernel function that potentially depends on the sample size n.

Lemma 1. *Suppose that U_n is a kth order U-statistic with symmetric, possibly sample size dependent kernel function φ_n, and that $\rho_{n,k} < \infty$. Then*

$$U_n - \mathbb{E}(U_n) = O_P\left(\sum_{c=1}^{k} \frac{\rho_{n,c}}{n^{c/2}}\right).$$

In particular, if the kernel φ_n is c^*th order degenerate, then

$$U_n = O_P\left(\sum_{c=c^*+1}^{k} \frac{\rho_{n,c}}{n^{c/2}}\right).$$

Proof. The result follows from explicitly calculating the variance of U_n (see e.g. Van der Vaart, 1998), and an application of Chebyscheff's inequality. □

B.2. Stochastic Expansion of the Local Polynomial Estimator. In this section, we state a particular stochastic expansion of the local polynomial regression estimators $\widehat{p}$ and $\widehat{q}$. This is a minor variation of results given in e.g. Masry (1996) or Kong, Linton, and Xia (2010). For simplicity, we state the result only for the former of the two estimators, but it applies analogously to the latter by replacing p with q in the following at every occurrence. We require the following notation. For any $s \in \{0, 1, \ldots, l_p\}$ let $n_s = \binom{s+d_p-1}{d_p-1}$ be the number of distinct d_p-tuples u with $|u| = s$. Arrange these d_p-tuples as a sequence in a lexicographical order with the highest priority given to the last position, so that $(0, \ldots, 0, s)$ is the first element in the sequence and $(s, 0, \ldots, 0)$ the last element. Let τ_s denote this 1-to-1 mapping, i.e. $\tau_s(1) = (0, \ldots, 0, s)$, $\ldots$, $\tau_s(n_s) = (s, 0, \ldots, 0)$. For each $s \in \{0, 1, \ldots, l_p\}$ we also define a $n_s \times 1$ vector $w_{j,s}(u)$ with its kth element given by $(((X_{p,j} - u)/h_p)^{\tau_s(k)}$. Finally, we put

$$\begin{aligned}
w_j(u) &= (1, w_{j,1}(u)^\top, \ldots, w_{j,l_p}(u)^\top)^\top \\
M_{p,n}(u) &= \frac{1}{n}\sum_{j \neq i}^{n} w_j(u) w_j(u)^\top K_{h_p}(X_{p,j} - u), \\
N_{p,n}(u) &= \mathbb{E}(w_j(u) w_j(u)^\top K_{h_p}(X_{p,j} - u)), \\
\eta_{p,n,j}(u) &= w_j(u) w_j(u)^\top K_{h_p}(X_{p,j} - u) - \mathbb{E}(w_j(u) w_j(u)^\top K_{h_p}(X_{p,j} - u)).
\end{aligned}$$

To better understand this notation, note that for the simple case that $l_p = 0$, i.e. when $\widehat{p}$ is the Nadaraya-Watson estimator, we have that $w_j(u) = 1$, that the term $M_{p,n}(u) = n^{-1}\sum_{i=1}^{n} K_{h_p}(X_{p,i} - u)$ is the usual Rosenblatt-Parzen density estimator, that $N_{p,n}(u) = \mathbb{E}(K_{h_p}(X_{p,i} - u))$ is its expectation, and that $\eta_{p,n,i}(u) = K_{h_p}(X_{p,i} - u) - \mathbb{E}(K_{h_p}(X_{p,i} - u))$ is a mean zero stochastic term with variance of the order $O(h_p^{-d_p})$. Also note that with this notation we can write the estimator $\widehat{p}(U_i)$

as

$$\widehat{p}(U_i) = \frac{1}{n-1}\sum_{j\neq i} e_1^\top M_{p,n}(U_i)^{-1} w_j(U_i) K_{h_p}(X_{p,j} - U_i) Y_{p,j},$$

where e_1 denotes the $(1 + l_p d_p)$-vector whose first component is equal to one and whose remaining components are equal to zero. We also introduce the following quantities:

$$\begin{aligned}
B_{p,n}(U_i) &= e_1^\top N_{p,n}(U_i)^{-1} \mathbb{E}(w_j(U_i) K_{h_p}(X_{p,j} - U_i)(p_o(X_{p,j}) - p_o(U_i))|U_i) \\
S_{p,n}(U_i) &= \frac{1}{n}\sum_{j\neq i} e_1^\top N_{p,n}(U_i)^{-1} w_j(U_i) K_{h_p}(X_{p,j} - U_i)\varepsilon_{p,j} \\
R_{p,n}(U_i) &= \frac{1}{n}\sum_{j\neq i} e_1^\top \left(\frac{1}{n}\sum_{l\neq i}\eta_{p,n,l}(U_i)\right) N_{p,n}(U_i)^{-2} w_j(U_i) K_{h_p}(X_{p,j} - U_i)\varepsilon_{p,j}
\end{aligned}$$

We refer to these three terms as the bias, and the first- and second-order stochastic terms, respectively. Here $\varepsilon_{p,j} = Y_{p,j} - p_o(X_{p,j})$ is the nonparametric regression residual, which satisfies $\mathbb{E}(\varepsilon_{p,j}|X_{p,j}) = 0$ by construction. To get an intuition for the behavior of the two stochastic terms, it is again instructive to consider simple case that $l_p = 0$, for which

$$\begin{aligned}
S_{p,n}(U_i) &= \frac{1}{n\bar{f}_{p,n}(U_i)}\sum_{j\neq i} K_{h_p}(X_{p,j} - U_i)\varepsilon_{p,j} \text{ and} \\
R_{p,n}(U_i) &= \frac{1}{n\bar{f}_{p,n}(U_i)^2}\left(\frac{1}{n}\sum_{l\neq i}(K_{h_p}(X_{p,l} - U_i) - \bar{f}_{p,n}(U_i))\right)\sum_{j\neq i} K_{h_p}(X_{p,j} - U_i)\varepsilon_{p,j}
\end{aligned}$$

with $\mathbb{E}(K_{h_p}(X_{p,j} - u)) = \bar{f}_{p,n}(u)$. With this notation, we obtain the following result.

Lemma 2. *Under Assumptions 1–2, the following statements hold:*

(i) For uneven $l_p \geq 1$ the bias $B_{p,n}$ satisfies

$$\max_{i\in\{1,\dots,n\}} |B_{p,n}(U_i)| = O_P(h_p^{l_p+1}),$$

and the first- and second-order stochastic terms satisfy

$$\max_{i\in\{1,\dots,n\}} |S_{p,n}(U_i)| = O_P((nh_p^{d_p}/\log n)^{-1/2}) \text{ and } \max_{i\in\{1,\dots,n\}} |R_{p,n}(U_i)| = O_P((nh_p^{d_p}/\log n)^{-1}).$$

(ii) For any $l_p \geq 0$, we have that

$$\max_{i\in\{1,\dots,n\}} |\widehat{p}(U_i) - p_o(U_i) - B_{p,n}(U_i) - S_{p,n}(U_i) - R_{p,n}(U_i)| = O_P((nh_p^{d_p}/\log n)^{-3/2}).$$

(iii) For $\|\cdot\|$ a matrix norm, we have that

$$\max_{i\in\{1,\dots,n\}} \|n^{-1}\sum_{j\neq i}\eta_{p,n,j}(U_i)\| = O_P((nh_p^{d_p}/\log n)^{-1/2}).$$

Proof. The proof follows from well-known arguments in e.g. Masry (1996) or Kong et al. (2010). □

B.3. Functional Derivatives of DR moment conditions. In this section, we formally prove a result about the functional derivatives of DR moment conditions. Using the notation introduced in the proof of Theorem 1, we obtain the following result.

Lemma 3. *If the function ψ satisfies the Double Robustness Property in* (2.2)*, and Assumption 3 holds, then $\mathbb{E}(\psi^p(Z)\bar{p}(U)) = \mathbb{E}(\psi^{pp}(Z)\bar{p}(U)) = \mathbb{E}(\psi^q(Z)\bar{q}(U)) = \mathbb{E}(\psi^{qq}(Z)\bar{q}(U)) = 0$ for all functions $\bar{p}$ and $\bar{q}$ such that $p_o + t\bar{p} \in \mathcal{P}$ and $q_o + t\bar{q} \in \mathcal{Q}$ for any $t \in \mathbb{R}$ with $|t|$ sufficiently small.*

Proof. The proof is similar for all four cases, and thus we only consider the first one. By dominated convergence, we have that

$$\mathbb{E}(\psi^p(Z)\bar{p}(U)) = \lim_{t\to 0} \frac{\Psi(\theta_o, p_o + t\bar{p}, q_o) - \Psi(\theta_o, p_o, q_o)}{t} = 0$$

where the last equality follows since the numerator is equal to zero by the DR property. □

B.4. Further Helpful Results. In this subsection, we derive a number of intermediate results used in proof of Theorem 1.

Lemma 4. *Under Assumption 1–4, the following statements hold:*

$$(i) \quad \frac{1}{n}\sum_{i=1}^{n} \psi^p(Z_i)(\widehat{p}(U_i) - p_o(U_i)) = o_P(n^{-1/2}),$$
$$(ii) \quad \frac{1}{n}\sum_{i=1}^{n} \psi^q(Z_i)(\widehat{q}(V_i) - q_o(V_i)) = o_P(n^{-1/2}).$$

Proof. We only show the first statement, as the proof for the second one is fully analogous. From Lemma 2 and Assumption 4, it follows that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \psi^p(Z_i)(\widehat{p}(U_i) - p_o(U_i)) = \frac{1}{n}\sum_{i=1}^{n} \psi^p(Z_i)(B_{p,n}(U_i) + S_{p,n}(U_i) + R_{p,n}(U_i)) \\ + O_P(\log(n)^{3/2} n^{-3/2} h_p^{-3d_p/2}),\end{aligned}$$

and since the second term on the right-hand side of the previous equation is of the order $o_P(n^{-1/2})$ by Assumption 4, it suffices to study the first term. As a first step, we find that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \psi^p(Z_i) B_{p,n}(U_i) &= \mathbb{E}(\psi^p(Z_i) B_{p,n}(U_i)) + O_P(h_p^{l_p+1} n^{-1/2}) \\ &= O_P(h_p^{l_p+1} n^{-1/2}),\end{aligned}$$

where the first equality follows from Chebyscheff's inequality, and the second equality follows from Lemma 2 and the fact that by Lemma 3 we have that $\mathbb{E}(\psi^p(Z_i)B_{p,n}(U_i)) = 0$. Next, consider the term

$$\frac{1}{n}\sum_{i=1}^{n}\psi^p(Z_i)S_{p,n}(U_i) = \frac{1}{n^2}\sum_{i}\sum_{j\neq i}\psi^p(Z_i)e_1^\top N_{p,n}(U_i)^{-1}w_j(U_i)K_{h_p}(X_{p,j}-U_i)\varepsilon_{p,i}.$$

This is a second order U-Statistic (up to a bounded, multiplicative term), and since by Lemma 3 we have that $\mathbb{E}(\psi^p(Z_i)e_1^\top N_{p,n}(U_i)^{-1}w_j(U_i)K_{h_p}(X_{p,j}-U_i)|X_{p,j}) = 0$, its kernel is first-order degenerate. It then follows from Lemma 1 and some simple variance calculations that

$$\frac{1}{n}\sum_{i=1}^{n}\psi^p(Z_i)S_{p,n}(U_i) = O_P(n^{-1}h_p^{-d_p/2}).$$

Finally, we consider the term

$$\frac{1}{n}\sum_{i=1}^{n}\psi^p(Z_i)R_{p,n}(U_i) = T_{n,1} + T_{n,2},$$

where

$$T_{n,1} = \frac{1}{n^3}\sum_{i}\sum_{j\neq i}\psi^p(Z_i)e_1^\top\eta_{p,n,j}(U_i)N_n(u)^{-2}w_j(U_i)K_{h_p}(X_{p,j}-U_i)\varepsilon_{p,j} \text{ and}$$
$$T_{n,2} = \frac{1}{n^3}\sum_{i}\sum_{j\neq i}\sum_{l\neq i,j}\psi^p(Z_i)e_1^\top\eta_{p,n,j}(U_i)N_n(U_i)^{-2}w_l(U_i)K_{h_p}(X_{p,l}-U_i)\varepsilon_{p,l}.$$

Using Lemma 3, one can see that $T_{n,2}$ is equal to a third-order U-Statistic (up to a bounded, multiplicative term) with second-order degenerate kernel, and thus

$$T_{n,2} = O_P(n^{-3/2}h_p^{-d_p})$$

by Lemma 1 and some simple variance calculations. On the other hand, the term $T_{n,1}$ is equal to n^{-1} times a second order U-statistic (up to a bounded, multiplicative term), with first-order degenerate kernel, and thus

$$T_{n,1} = n^{-1}\cdot O_P(n^{-1}h_p^{-3d_p/2})) = n^{-1/2}h_p^{-d_p/2}O_P(T_{n,2}).$$

The statement of the lemma thus follows if $h_p \to 0$ and $n^2h_p^{3d_p} \to \infty$ as $n \to \infty$, which holds by Assumption 4. This completes our proof. □

Remark 2. Without the DR property, the term $n^{-1}\sum_{i=1}^{n}\psi^p(Z_i)B_{p,n}(U_i)$ in the above proof would be of the larger order $O(h_p^{l_p+1})$, which is the usual order of the bias due to smoothing the nonparametric component. This illustrates how the DR property of the moment conditions acts like a bias correction device (see also Remark 3 below).

Lemma 5. *Under Assumption 1–4, the following statements hold:*

$$(i) \quad \frac{1}{n}\sum_{i=1}^{n} \psi^{pp}(Z_i)(\widehat{p}(U_i) - p_o(U_i))^2 = o_P(n^{-1/2}),$$
$$(ii) \quad \frac{1}{n}\sum_{i=1}^{n} \psi^{qq}(Z_i)(\widehat{q}(V_i) - q_o(V_i))^2 = o_P(n^{-1/2}).$$

Proof. We only show the first statement, as the second statement is conceptually similar to establish. Note that by Lemma 2 we have that

$$(\widehat{p}(u) - p_o(u))^2 = \sum_{k=1}^{6} T_{n,k}(u) + O_P\left(\left(\frac{\log(n)}{nh_p^{d_p}}\right)^{3/2}\right)\left(O_P(h_p^{l_p+1}) + O_P\left(\frac{\log(n)}{nh_p}\right)\right),$$

where $T_{n,1}(u) = B_{p,n}(u)^2$, $T_{n,2}(u) = S_{p,n}(u)^2$, $T_{n,3}(u) = R_{p,n}(u)^2$, $T_{n,4}(u) = 2B_{p,n}(u)S_{p,n}(u)$, $T_{n,5}(u) = 2B_{p,n}(u)R_{p,n}(u)$, and $T_{n,6}(u) = 2S_{p,n}(u)R_{p,n}(u)$. Since the second term on the right-hand side of the previous equation is of the order $o_P(n^{-1/2})$ by Assumption 4, it suffices to show that we have that $n^{-1}\sum_{i=1}^{n} \psi^{pp}(Z_i)T_{n,k}(U_i) = o_P(n^{-1/2})$ for $k \in \{1, \ldots, 6\}$. Our proof proceeds by obtaining sharp bounds on $n^{-1}\sum_{i=1}^{n} \psi^{pp}(Z_i)T_{n,k}(U_i)$ for $k \in \{1, 2, 4, 5\}$ using Lemmas 3 and 1, and crude bounds for $k \in \{3, 6\}$ simply using the uniform rates derived in Lemma 2. First, for $k = 1$ we find that

$$\frac{1}{n}\sum_{i=1}^{n} \psi^{pp}(Z_i)T_{n,1}(U_i) = \mathbb{E}(\psi^{pp}(Z_i)B_{p,n}(U_i)^2) + O_P(n^{-1/2}h_p^{2l_p+2}) = O_P(n^{-1/2}h_p^{2l_p+2})$$

because $\mathbb{E}(\psi^{pp}(Z_i)B_{p,n}(U_i)^2) = 0$ by Lemma 3. Second, for $k = 2$ we can write

$$\frac{1}{n}\sum_{i=1}^{n} \psi^{pp}(Z_i)T_{n,2}(U_i) = T_{n,2,A} + T_{n,2,B}$$

where

$$T_{n,2,A} = \frac{1}{n^3}\sum_{i}\sum_{j\neq i} \psi^{pp}(Z_i)(e_1^\top N_{p,n}(U_i)^{-1}w_j(U_i))^2 K_{h_p}(X_{p,j} - U_i)^2\varepsilon_{p,j}^2$$
$$T_{n,2,B} = \frac{1}{n^3}\sum_{i}\sum_{j\neq i}\sum_{l\neq i,j} \psi^{pp}(Z_i)e_1^\top N_{p,n}(U_i)^{-1}w_j(U_i)K_{h_p}(X_{p,j} - U_i)\varepsilon_{p,j}$$
$$\cdot e_1^\top N_{p,n}(U_i)^{-1}w_l(U_i)K_{h_p}(X_{p,l} - U_i)\varepsilon_{p,l}$$

Using Lemma 3, one can see that $T_{n,2,B}$ is equal to a third-order U-Statistic with a second-order degenerate kernel function (up to a bounded, multiplicative term), and thus

$$T_{n,2,B} = O_P(n^{-3/2}h_p^{-d_p}).$$

On the other hand, the term $T_{n,2,A}$ is (again, up to a bounded, multiplicative term) equal to n^{-1} times a second order U-statistic with first-order degenerate kernel function, and thus

$$T_{n,2,A} = n^{-1} O_P(n^{-1} h_p^{-3d_p/2}) = O_P(n^{-2} h_p^{-3d_p/2}).$$

Third, for $k = 4$ we use again Lemma 3 and Lemma 1 to show that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \psi^{p}(Z_i) T_{n,4}(U_i) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \psi^{pp}(Z_i) B_{p,n}(U_i) e_1^\top N_{p,n}(U_i)^{-1} w_j(U_i) K_{h_p}(X_{p,j} - U_i)\varepsilon_{p,j} \\ &= O_P(n^{-1} h_p^{-d_p/2}) \cdot O(h_p^{l_p+1}),\end{aligned}$$

where the last equality follows from the fact that $n^{-1}\sum_{i=1}^{n} \psi^{p}(Z_i) T_{n,4}(U_i)$ is (again, up to a bounded, multiplicative term) equal to a second order U-statistic with first-order degenerate kernel function. Fourth, for $k = 5$, we can argue as in the final step of the proof of Lemma 4 to show that

$$\frac{1}{n}\sum_{i=1}^{n} \psi^{pp}(Z_i) T_{n,5}(U_i) = O_P(n^{-3/2} h_p^{-d_p} h_p^{l_p+1})$$

Finally, we obtain a number of crude bounds based on uniform rates in Lemma 2:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \psi^{pp}(Z_i) T_{n,3}(U_i) &= O_P(\|R_{p,n}\|_\infty^2) = O_P(\log(n)^2 n^{-2} h_p^{-2d_p}) \\ \frac{1}{n}\sum_{i=1}^{n} \psi^{pp}(Z_i) T_{n,6}(U_i) &= O_P(\|R_{p,n}\|_\infty) \cdot O_P(\|S_{p,n}\|_\infty) = O_P(\log(n)^{3/2} n^{-3/2} h_p^{-3d_p/2})\end{aligned}$$

The statement of the lemma thus follows if $h_p \to 0$ and $n^2 h_p^{3d_p} / \log(n)^3 \to \infty$ as $n \to \infty$, which holds by Assumption 4. This completes our proof. □

Remark 3. Without the DR property, the term $T_{n,2,B}$ in the above proof would be (up to a bounded, multiplicative term) equal to a third-order U-Statistic with a first-order degenerate kernel function (instead of a second order one). In this case, we would find that

$$T_{n,2,B} = O_P(n^{-1} h_p^{-d_p/2}) + O_P(n^{-3/2} h_p^{-d_p}) = O_P(n^{-1} h_p^{-d_p/2}).$$

On the other hand, in the absence of the DR property, the term $T_{n,2,A}$ would be (up to a bounded, multiplicative term) equal to a n^{-1} times a second-order U-Statistic with a non-degenerate kernel function, and thus we would have

$$T_{n,2,A} = O(n^{-1} h_p^{-d_p}) + O_P(n^{-3/2} h^{-d_p}) + O_P(n^{-2} h^{-2d_p}) = O(n^{-1} h^{-d_p}) + o_P(n^{-1} h^{-d_p}).$$

The leading term of an expansion of the sum $T_{n,2,A} + T_{n,2,B}$ is thus a pure bias term of order $n^{-1} h_p^{-d_p}$. This term is analogous to the "degrees of freedom bias" in Ichimura and Linton (2005),

and the "nonlinearity bias" or "curse of dimensionality bias" in Cattaneo et al. (2013a). In our context, the DR property of the moment conditions removes this term, which illustrates how our structure acts like a bias correction method.

Lemma 6. *Under Assumption 1–4, the following statement holds:*

$$\frac{1}{n}\sum_{i=1}^{n}\psi^{pq}(Z_i)(\widehat{p}(U_i)-p_o(U_i))(\widehat{q}(V_i)-q_o(V_i))=o_P(n^{-1/2}).$$

Proof. By Lemma 2, one can see that uniformly over (u,v) we have that

$$\begin{aligned}(\widehat{p}(u)-p_o(u))(\widehat{q}(v)-q_o(v))=\sum_{k=1}^{9}T_{n,k}(u,v)+O_P\left(\left(\frac{\log(n)}{nh_p^{d_p}}\right)^{3/2}\right)\left(O_P(h_q^{l_q+1})+O_P\left(\frac{\log(n)}{nh_q^{d_q}}\right)\right)\\+O_P\left(\left(\frac{\log(n)}{nh_q^{d_q}}\right)^{3/2}\right)\left(O_P(h_p^{l_p+1})+O_P\left(\frac{\log(n)}{nh_p^{d_p}}\right)\right)\end{aligned}$$

where $T_{n,1}(u,v)=B_{p,n}(u)B_{q,n}(v)$, $T_{n,2}(u,v)=B_{p,n}(u)S_{q,n}(v)$, $T_{n,3}(u,v)=B_{p,n}(u)R_{q,n}(v)$, $T_{n,4}(u,v)=S_{p,n}(u)B_{q,n}(v)$, $T_{n,5}(u,v)=S_{p,n}(u)S_{q,n}(v)$, $T_{n,6}(u,v)=S_{p,n}(u)R_{q,n}(v)$, $T_{n,7}(u,v)=R_{p,n}(u)B_{q,n}(v)$, $T_{n,8}(u,v)=R_{p,n}(u)S_{q,n}(v)$, and $T_{n,9}(u,v)=R_{p,n}(u)R_{q,n}(v)$. Since the last two terms on the right-hand side of the previous equation are easily of the order $o_P(n^{-1/2})$ by Assumption 4, it suffices to show that for any for $k\in\{1,\dots,9\}$ we have that $n^{-1}\sum_{i=1}^{n}\psi^{pp}(Z_i)T_{n,k}(U_i,V_i)=o_P(n^{-1/2})$. As in the proof of Lemma 5, we proceed by obtaining sharp bounds on $n^{-1}\sum_{i=1}^{n}\psi^{pp}(Z_i)T_{n,k}(U_i)$ for $k\in\{1,\dots,5,7\}$ using Lemma 1– 3, and crude bounds for $k\in\{6,8,9\}$ simply using the uniform rates derived in Lemma 2. First, arguing as in the proof of Lemma 4 and 5 above, we find that

$$\frac{1}{n}\sum_{i=1}^{n}\psi^{pp}(Z_i)T_{n,1}(U_i,V_i)=\mathbb{E}(\psi^{pq}(Z_i)B_{p,n}(U_i)B_{q,n}(V_i))+O_P(n^{-1/2}h_p^{l_p+1}h_q^{l_q+1})=O_P(h_p^{l_p+1}h_q^{l_q+1}),$$

where the last equation follows from the fact that $\mathbb{E}(\psi^{pq}(Z_i)B_{p,n}(U_i)B_{q,n}(V_i))=O(h_p^{l_p+1}h_q^{l_q+1})$. Second, for $k=2$ we consider the term

$$\frac{1}{n}\sum_{i}\psi^{pq}(Z_i)T_{n,2}(U_i,V_i)=\frac{1}{n^2}\sum_{i}\sum_{j\neq i}\psi^{pq}(Z_i)B_{p,n}(U_i)e_1^{\top}N_{p,n}(V_i)^{-1}w_j(V_i)K_{h_q}(X_{q,j}-V_i)\varepsilon_{q,j}.$$

This term is (up to a bounded, multiplicative term) equal to a second-order U-Statistic with non-degenerate kernel function. It thus follows from Lemma 1 and some variance calculations that

$$\frac{1}{n}\sum_{i}\psi^{pq}(Z_i)T_{n,2}(U_i,V_i)=O_P(n^{-1/2}h_p^{l_p+1})+O_P(n^{-1}h_q^{-d_q/2}h_p^{l_p+1})$$

Using the same argument, we also find that

$$\frac{1}{n}\sum_{i}\psi^{pq}(Z_i)T_{n,4}(U_i,V_i)=O_P(n^{-1/2}h_q^{l_q+1})+O_P(n^{-1}h_p^{-d_p/2}h_q^{l_q+1}).$$

For $k = 3$, we can argue as in the final step of the proof of Lemma 4 to show that

$$\frac{1}{n}\sum_{i=1}^{n}\psi^{pp}(Z_i)T_{n,3}(U_i,V_i) = O_P(n^{-1}h_q^{-d_q/2}h_p^{l_p+1}) + O_P(n^{-3/2}h_q^{-d_q}h_p^{l_p+1}),$$

and for the same reason we find that

$$\frac{1}{n}\sum_{i=1}^{n}\psi^{pp}(Z_i)T_{n,7}(U_i,V_i) = O_P(n^{-1}h_p^{-d_p/2}h_q^{l_q+1}) + O_P(n^{-3/2}h_p^{-d_p}h_q^{l_q+1}).$$

Next, we consider the case $k = 5$. Here we can write

$$\frac{1}{n}\sum_{i}\psi^{pq}(Z_i)T_{n,5}(U_i,V_i) = T_{n,5,A} + T_{n,5,B},$$

where

$$\begin{aligned}
T_{n,5,A} = \frac{1}{n^3}\sum_{i}\sum_{j\neq i}\psi^{pq}(Z_i)(e_1^\top N_{p,n}(U_i)^{-1}w_{p,j}(U_i)K_{h_p}(X_{p,j}-U_i)\varepsilon_{p,j}) \\
\cdot (e_1^\top N_{q,h_q}(V_i)^{-1}w_{q,j}(V_i)K_{h_q}(X_{q,j}-V_i)\varepsilon_{q,j}), \\
T_{n,5,B} = \frac{1}{n^3}\sum_{i}\sum_{j\neq i}\sum_{l\neq i,j}\psi^{pq}(Z_i)e_1^\top N_{p,n}(U_i)^{-1}w_j(U_i)K_{h_p}(X_{p,j}-U_i)\varepsilon_{p,j} \\
\cdot e_1^\top N_{q,h_q}(V_i)^{-1}w_l(V_i)K_{h_q}(X_{q,l}-V_i)\varepsilon_{q,l}.
\end{aligned}$$

One can easily see that $T_{n,5,B}$ is equal to a third-order U-Statistic (up to a bounded, multiplicative term) with first-order degenerate kernel, and thus

$$T_{n,5,B} = O_P(n^{-1}) + O_P(n^{-3/2}h_p^{-d_p/2}h_q^{-d_q/2})$$

by Lemma 1 and some straightforward variance calculations. To derive the order of the term $T_{n,5,A}$, we exploit the orthogonality condition (2.3), which implies that $\mathbb{E}(\varepsilon_p\varepsilon_q|X_p,X_q) = 0$. Clearly, $T_{n,5,A}$ is equal to n^{-1} times a second order U-statistic (up to a bounded, multiplicative term), and because of (2.3) the kernel of this U-Statistic is first-order degenerate. We thus find that

$$T_{n,5,A} = n^{-1}\cdot O_P(n^{-1}h_p^{-d_p/2}h_q^{-d_q/2}) = n^{-1/2}O_P(T_{n,5,B}).$$

by Lemma 1 and a simple variance calculation. Finally, we obtain a number of crude bounds based on uniform rates in Lemma 2 for the following terms:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\psi^{pp}(Z_i)T_{n,6}(U_i) &= O_P(\|S_{p,n}\|_\infty)\cdot O_P(\|R_{q,n}\|_\infty) = O_P(\log(n)^{5/2}n^{-5/2}h_p^{-d_p}h_q^{-3d_q/2}) \\
\frac{1}{n}\sum_{i=1}^{n}\psi^{pp}(Z_i)T_{n,8}(U_i) &= O_P(\|R_{p,n}\|_\infty)\cdot O_P(\|S_{q,n}\|_\infty) = O_P(\log(n)^{5/2}n^{-5/2}h_q^{-d_q}h_p^{-3d_p/2}) \\
\frac{1}{n}\sum_{i=1}^{n}\psi^{pp}(Z_i)T_{n,9}(U_i) &= O_P(\|R_{p,n}\|_\infty)\cdot O_P(\|R_{q,n}\|_\infty) = O_P(\log(n)^{3}n^{-3}h_p^{-3d_p/2}h_q^{-3d_q/2})
\end{aligned}$$

The statement of the Lemma then follows from Assumption 4. This completes our proof. □

Remark 4. The derivation of the order of the term $T_{n,5,A}$ is the only step in our proof that requires the orthogonality condition (2.3). Without this condition, the kernel of the respective U-Statistic would be non-degenerate, and in general we would only find that $T_{n,5,A} = O_P(n^{-1} \max\{h_p^{-d_p}, h_q^{-d_q}\})$.

References

Almond, D., K. Y. Chay, and D. S. Lee (2005): "The costs of low birth weight," *The Quarterly Journal of Economics*, 120, 1031–1083.

Andrews, D. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica*, 62, 43–72.

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

Cattaneo, M., R. Crump, and M. Jansson (2013a): "Generalized Jackknife Estimators of Weighted Average Derivatives," *Journal of the American Statistical Association*, to appear.

——— (2013b): "Small bandwidth asymptotics for density-weighted average derivatives," *Econometric Theory*, to appear.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics*, 36, 808–843.

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

Fan, J. (1993): "Local linear regression smoothers and their minimax efficiencies," *The Annals of Statistics*, 21, 196–216.

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications*, Chapman & Hall/CRC.

Fan, J., N. Heckman, and M. Wand (1995): "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions," *Journal of the American Statistical Association*, 90, 141–150.

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

Gozalo, P. and O. Linton (2000): "Local Nonlinear Least Squares: Using parametric information in nonparametric regression," *Journal of Econometrics*, 99, 63–106.

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

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

Hall, P., R. Wolff, and Q. Yao (1999): "Methods for estimating a conditional distribution function," *Journal of the American Statistical Association*, 94, 154–163.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

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

Heckman, J. and R. Robb (1985): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of Econometrics*, 30, 239–267.

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

Ichimura, H. and S. Lee (2010): "Characterization of the asymptotic distribution of semiparametric M-estimators," *Journal of Econometrics*, 159, 252–266.

Ichimura, H. and O. Linton (2005): "Asymptotic expansions for some semiparametric program evaluation estimators," in *Identifcation and Inference for Econometric Models: A Festschrift in Honor of Thomas J. Rothenberg*, ed. by D. Andrews and J. Stock, Cambridge, UK: Cambridge University Press, 149–170.

Imbens, G. (2004): "Nonparametric estimation of average treatment effects under exogeneity: A review," *Review of Economics and Statistics*, 86, 4–29.

Imbens, G., W. Newey, and G. Ridder (2005): "Mean-square-error calculations for average treatment effects," *Working Paper*.

Imbens, G. and J. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Khan, S. and E. Tamer (2010): "Irregular identification, support conditions, and inverse weight estimation," *Econometrica*, 78, 2021–2042.

Klein, R. and C. Shen (2010): "Bias Corrections in Testing and Estimating Semiparametric, Single Index Models," *Econometric Theory*, 26, 1683–1718.

Kong, E., O. Linton, and Y. Xia (2010): "Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model," *Econometric Theory*, 26, 1529–1564.

Linton, O. (1995): "Second order approximation in the partially linear regression model," *Econometrica*, 63, 1079–1112.

Masry, E. (1996): "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17, 571–599.

Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Newey, W., F. Hsieh, and J. Robins (2004): "Twicing kernels and a small bias property of semiparametric estimators," *Econometrica*, 72, 947–962.

Newey, W. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Robins, J., L. Li, E. Tchetgen, and A. Van Der Vaart (2008): "Higher order influence functions and minimax estimation of nonlinear functionals," *Probability and Statistics: Essays in Honor of David A. Freedman, ed. by D. Nolan, and T. Speed. Beachwood, OH: Institute of Mathematical Statistics*, 335–421.

Robins, J. and Y. Ritov (1997): "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theroy for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319.

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

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

Robins, J. M. and A. Rotnitzky (2001): "Comment on "Inference for semiparametric models: some questions and an answer" by P. Bickel and J. Kwon," *Statistica Sinica*, 11, 920–936.

Robins, J. M., A. Rotnitzky, and M. van der Laan (2000): "On Profile Likelihood: Comment," *Journal of the American Statistical Association*, 95, 477–482.

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

Rotnitzky, A., J. Robins, and D. Scharfstein (1998): "Semiparametric regression for repeated outcomes with nonignorable nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339.

Rubin, D. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688–701.

Ruppert, D. and M. Wand (1994): "Multivariate locally weighted least squares regression," *Annals of Atatistics*, 1346–1370.

Scharfstein, D., A. Rotnitzky, and J. Robins (1999): "Adjusting for nonignorable drop-out using semiparametric nonresponse models," *Journal of the American Statistical Association*, 94, 1096–1120.

Tan, Z. (2006): "Regression and weighting methods for causal inference using instrumental variables," *Journal of the American Statistical Association*, 101, 1607–1618.

Van der Laan, M. and J. Robins (2003): *Unified methods for censored longitudinal data and causality*, Springer.

Wooldridge, J. (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141, 1281–1301.