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 model in the first stage. We show that these *semiparametric doubly robust estimators* (SDREs) are asymptotically linear under conditions that are much weaker than those necessary for generic semiparametric two-step estimators, and thus distributional approximations based on classical first-order asymptotic theory are substantially more reliable. In practice, this means that SDREs generally have smaller first-order bias, are less sensitive to the implementation of the nonparametric first stage, can allow for rate-optimal choices of smoothing parameters and data-driven estimates thereof, and do not require the use of bias reducing nonparametric estimators (such as those based on higher-order kernels) in settings with moderate dimensionality. SDREs exist for many interesting parameters in a wide range of missing data and treatment effect models. Given their theoretical and practical advantages, SDREs constitute an attractive alternative to other estimators commonly employed in those areas, such as e.g. Inverse Probability Weighting. We illustrate our method with a simulation exercise.

JEL Classification: C14, C21, C31, C51

Keywords: *Semiparametric estimation, missing data, treatment effects, double robustness*

[*] First version: December 20, 2012. This Version: January 23, 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

1. Introduction

Semiparametric models are of great importance for applied econometric research. These models often imply that the finite-dimensional parameter of interest can be characterized through a moment condition that contains an unknown nuisance function. This structure then leads to a two-step semiparametric estimation approach. In the first step, the nuisance function is estimated nonparametrically. In the second step, the parameter of interest is estimated from an empirical version of the moment condition, with the unknown nuisance function replaced by its first-step estimate. Such estimators are used in a wide range of applications, and their theoretical properties have been studied extensively (e.g. Newey, 1994; Newey and McFadden, 1994; Andrews, 1994; Chen, Linton, and Van Keilegom, 2003; Ichimura and Lee, 2010).

In this paper, we consider semiparametric two-step estimators that are based on a moment condition that exhibits 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 Robins, Rotnitzky, and van der Laan (2000) and Robins and Rotnitzky (2001), we refer to such moment conditions as *doubly robust*, and call the corresponding estimators *semiparametric doubly robust estimators* (SDREs). These estimators differ from the usual doubly robust procedures used widely in statistics (e.g. Van der Laan and Robins, 2003), which rely on additional parametric restrictions on the nuisance functions. We impose no such restrictions in our paper.

Our main contribution is to show that a SDRE possesses several attractive theoretical and practical properties relative to a generic semiparametric two-step estimator, even if the two have the same asymptotic variance. SDREs are generally root-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. Their stochastic behavior can thus be better approximated by classical first-order asymptotics. In practice, this means that SDREs generally have smaller first-order bias, are less sensitive to the implementation of the nonparametric first stage, can allow for rate-optimal choices of smoothing parameters and data-driven estimates thereof, and do not require the use of bias reducing nonparametric estimators (such as those based on higher-order kernels) in settings with moderate dimensionality. SDREs are also adaptive, in the sense that by construction their asymptotic variance does not contain adjustment terms for the nonparametric first step. This substantially simplifies the calculation of standard errors.

Doubly robust moment conditions are known to exist for many interesting parameters in a wide range of semiparametric models. Examples include regression coefficients in models with missing outcomes and/or covariates (e.g. Robins, Rotnitzky, and Zhao, 1994; Robins and Rotnitzky, 1995), average treatment effects in potential outcome models with unconfounded assignment (Scharfstein, Rotnitzky, and Robins, 1999), and local average treatment effects in instrumental variable models (Tan, 2006), amongst many others. It is thus straightforward to construct our SDREs in these settings. In all the aforementioned examples, and several others, doubly robust moment conditions take the form of an expectation of the respective efficient influence (or "score") function. The asymptotic variance of the SDRE is thus equal to the semiparametric efficiency bound in these settings (Newey, 1994). SDREs therefore have favorable properties even relative to other efficient estimators that are commonly used in such settings, such as e.g. Inverse Probability Weighting estimators in missing data and treatment effect models (e.g. Hirano, Imbens, and Ridder, 2003; Chen, Hong, and Tarozzi, 2008).

As mentioned above, doubly robust moment conditions are traditionally used in connection with fully parametric specifications for each of the two nuisance functions, as this ensures that the resulting estimator is consistent if at least one of the parametric specifications is correct. The use of parametric doubly robust estimators is typically motivated by valid concerns about the reliability of semiparametric two-step estimators, especially the accuracy of conventional approximations of their finite sample distribution based on first-order asymptotics (e.g. Robins and Ritov, 1997). Such approximations are typically derived under strong smoothness conditions on the nuisance function, which cannot be effectively exploited by nonparametric estimation procedures in moderate samples, even if the conditions are actually satisfied.

Our use of doubly robust moment conditions is different from the traditional one, since we always retain a fully nonparametric first stage. However, our method addresses the same concerns about the accuracy of distributional approximation based on conventional asymptotic theory. Our results show that the same structure that safeguards parametric doubly robust estimators against misspecification is also benefitial when using a nonparametric first stage. It considerably reduces the impact of both the smoothing bias and the stochastic variation from nonparametrically estimating the nuisance functions on the final SDRE. Our estimators can therefore be shown to be root-n-consistent and asymptotically normal under substantially weaker smoothness conditions than those used to derive similar results for generic semiparametric two-step procedures. As a consequence, we expect inferential procedures justified by this asymptotic theory, such as hypothesis tests or confidence intervals, to be more reliable in settings with moderate samples and not too high-dimensional nuisance functions.

Our paper is not the first to be concerned with improving the properties of semiparametric two-stage estimators. In other 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, Crump, and Jansson (2012a) 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, showing that the use of doubly robust moment conditions reduces both types of bias simultaneously. An alternative approach to improve inference, which we do not consider in this paper, would be to derive "non-root-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 (2012b), 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 derived in a general setting. Section 4 applies our findings to the important special case of estimating average treatment effects under unconfoundedness. Section 5 shows evidence that SDREs have superior properties compared to other methods in a simulation study. Finally, Section 6 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 vector-valued 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 that exhibits a particular structure: 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) nuisance 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. Moreover, we assume that

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

for all functions $q \in \mathcal{Q}$ and $p \in \mathcal{P}$. Following 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.* We give a number of examples of settings in which DR moment conditions exist in the following subsection. Note that restricting attention to "just-identified" cases with $\psi(\cdot)$ taking values in $\mathbb{R}^{d_\theta}$ is without loss of generality, as Robins and Rotnitzky (2001) show that for all DR moment conditions the number of moments is equal to the number of components of θ_o.

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 $\Psi(\theta, p_o, \widetilde{q})$ or $\Psi(\theta, \widetilde{p}, q_o)$ equal to zero, where $\widetilde{p} \in \mathcal{P}$ and $\widetilde{q} \in \mathcal{Q}$ are arbitrary known and fixed functions. The properties of such standard semiparametric two-step estimators could be analyzed using general results in e.g. Newey (1994), Andrews (1994), Ai and Chen (2003) or Chen et al. (2003). In this paper, we argue in favor of an estimator of θ_o that solves a direct sample analogue of (2.1), using estimates of both infinite-dimensional nuisance parameters. We refer to such estimators as *semiparametric doubly robust estimators* (SDREs), and show that they possess certain favorable theoretical properties that should offset the additional computational costs due to estimating two functions nonparametrically instead of just one.

2.2. Examples. Before discussing the specific form and implementation of the estimator, we give a number of examples of DR moment conditions for various parameters of interest in missing data and causal inference models. This should illustrate the broad applicability of the methodology. We remark that in all the examples that we give below the moment function ψ, on which the DR moment condition is based, is the semiparametrically efficient influence function for the respective parameter of interest. As we show in Section 3, this implies that the asymptotic variance of SDREs is equal to the respective semiparametric efficiency bound in these settings (under suitable regularity conditions).

Example 1 (Population Means 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 $\theta_o = \mathbb{E}(Y)$. Define the functions $\pi_o(x) = \mathbb{E}(D|X = x)$ and $\mu_o(x) = \mathbb{E}(Y|D = 1, X = x)$, and assume that $E(D|Y, X) = \pi_o(X) > 0$ with probability 1. Then $\Psi(\theta, \pi, \mu) = \mathbb{E}(\psi(Z, \theta, \pi(X), \mu(X)))$ with

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

is a DR moment condition for estimating θ_o. □

Example 2 (Linear Regression with Missing Covariates)**.** Let $X = (X_1', X_2')'$ 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 $Z = (Y, X_1D, X_2, D)$. 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 $\mu_o(x_2, \theta) = \mathbb{E}(\varphi(Y, X, \theta)|D = 1, X_2 = x_2)$ with $\varphi(Y, X, \theta) = (1, X')'(Y - (1, X')\theta)$, and assume that $\mathbb{E}(D|Y, X) = \pi_o(Y, X_2) > 0$ with probability 1. Then $\Psi(\theta, \pi, \mu) = \mathbb{E}(\psi(Z, \theta, \pi(X), \mu(X)))$ with

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

is a DR moment condition for estimating θ_o. □

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. □

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. □

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

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

is a DR moment condition for estimating θ_o. □

2.3. Semiparametric Estimation. In this paper, we consider an estimator $\widehat{\theta}$ of θ_o that solves a direct sample analogue of (2.1). That is, we take $\widehat{\theta}$ as the value that solves the equation

$$0 = \Psi_n(\theta, \widehat{p}, \widehat{q}) := \frac{1}{n} \sum_{i=1}^{n} \psi(Z_i, \theta, \widehat{p}(U_i), \widehat{q}(V_i)), \tag{2.3}$$

in θ, where $\widehat{p}$ and $\widehat{q}$ are suitable nonparametric estimates of p_o and q_o, respectively. Since Ψ_n takes values in $\mathbb{R}^{d_\theta}$, such a solution exists with probability one. We also define the following quantities, which will be important for estimating the asymptotic variance of the estimator $\widehat{\theta}$:

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

For simplicity, we focus on the important special case that both infinite-dimensional nuisance parameters are conditional expectation functions. That is, 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, and both X_p and X_q are assumed to be continuously distributed. 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. 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 $\lambda_p(u) = [u_1, u_1^2, \ldots, u_1^{l_p}, \ldots, u_{d_p}, u_{d_p}^2, \ldots, u_{d_p}^{l_p}]^T$ and $\lambda_q(v) = [v_1, v_1^2, \ldots, v_1^{l_q}, \ldots, v_{d_q}, v_{d_q}^2, \ldots, v_{d_q}^{l_q}]^T$, 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

$$(\widehat{a}_p(U_i), \widehat{b}_p(U_i)) = \operatorname*{argmin}_{a,b} \sum_{j\neq i} \left(Y_{p,j} - a - b'\lambda_p(X_{p,j} - U_i)\right)^2 K_{h_p}(X_{p,j} - U_i),$$
$$(\widehat{a}_q(V_i), \widehat{b}_q(V_i)) = \operatorname*{argmin}_{a,b} \sum_{j\neq i} \left(Y_{q,j} - a - b'\lambda_p(X_{q,j} - V_i)\right)^2 K_{h_q}(X_{q,j} - V_i).$$

Here $K_{h_p}(u) = \prod_{j=1}^{d_p} \mathcal{K}(u_j/h_p)/h_p$ is a d_p-dimensional product kernel built from the univariate 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 $K_{h_q}(v)$ and h_q are defined similarly. 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). Also note that 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

3.1. Informal Overview of Results. In this section, we derive a number of theoretical properties of SDREs. Our main result, stated in Theorem 1 below, gives conditions under which these estimators are consistent, regular and asymptotically linear (RAL), asymptotically unbiased, and asymptotically normal. We also derive the form of the asymptotic variance and propose a consistent estimate thereof. The main advantage of SDREs is that these types of results can be shown under assumptions about the accuracy of the first-stage nonparametric estimators that are weak relative to those commonly invoked to obtain similar findings for generic semiparametric two-step estimators. We therefore expect that for SDREs these asymptotic results provide more accurate guidance about the estimators' finite-sample properties than they do in general.

To understand the role of the DR property in obtaining these result, it is instructive to consider an informal sketch of the asymptotic normality argument. First, it is generally easy to show that the estimator $\widehat{\theta}$ has the following representation that depends on $\widehat{p}$ and $\widehat{q}$:

$$\sqrt{n}(\widehat{\theta} - \theta_o) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Gamma_o^{-1} \psi(Z_i, \theta_o, \widehat{p}(U_i), \widehat{q}(V_i)) + o_P(1).$$

where $\Gamma_o = \partial \mathbb{E}(\psi(Z, \theta, p_o(U), q_o(V)))/\partial\theta|_{\theta=\theta_o}$. Next, we show that $n^{-1}\sum_{i=1}^{n}(\psi(Z_i, \theta_o, \widehat{p}(U_i), \widehat{q}(V_i)) - \psi(Z_i, \theta_o, p_o(U_i), q_o(V_i))) = o_P(n^{-1/2})$ using the DR property. We start by considering a second-order Taylor expansion of this term in $(\widehat{p}, \widehat{q})$ around (p_o, q_o), which yields that

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

Here $\psi^p(Z_i)$ and $\psi^{pp}(Z_i)$ are the first and second derivative of $\psi(Z_i, p_o(U_i), q_o(V_i))$ 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 mixed partial derivative of $\psi(Z_i, p_o(U_i), q_o(V_i))$ with respect to $p_o(U_i)$ and $q_o(V_i)$. Clearly, the two "cubic" remainder terms are both of the order $o_P(n^{-1/2})$ if the estimation error of the two nonparametric estimates is uniformly of the order $o(n^{-1/6})$. However, if one would be using a generic moment condition, the five "leading" terms in the above equation would generally not be of the order $o_P(n^{-1/2})$ if the nonparametric component converges that slowly (this typically requires the first stage estimation error be of the order $o_P(n^{-1/4})$, and the smoothing bias to be of the order $o(n^{-1/2})$). At this point, we exploit that the DR property of Ψ implies that

$$\frac{\partial^k}{\partial t}\Psi(\theta_o, p_o + t\bar{p}, q_o)|_{t=0} = \frac{\partial^k}{\partial t}\Psi(\theta_o, p_o, q_o + t\bar{q})|_{t=0} = 0 \qquad (3.1)$$

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. This property can be used as follows. The estimators $\widehat{p}$ and $\widehat{q}$ generally satisfy a certain linear stochastic expansion (e.g. Kong, Linton, and Xia, 2010). When substituting these expansions into the above quadratic expansion, we obtain a number of U-Statistics of various orders, that can be shown to be degenerate because of (3.1). These terms thus vanish at a considerably faster rate than they would without the DR property. They can be shown to be of the order $o_P(n^{-1/2})$ as long as the *product* of the two smoothing bias terms from estimating p_o and q_o is of the order $o(n^{-1/2})$, and the respective overall estimation errors are uniformly of the order $o_P(n^{-1/6})$. Taken together, these arguments show that

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

The estimator is thus regular and asymptotically linear (RAL), and its asymptotic normality follows directly from the central limit theorem if the summands on the right-hand side of the last equation have finite second moments.

3.2. Asymptotic Properties. We now formalize the argument given above. For our theoretical analysis of the large-sample properties of SDREs, we impose the following conditions.

Assumption 1. *(i) the random vector U is continuously distributed with compact support I_U (ii) $\sup_{u \in I_U} \mathbb{E}(|Y_p|^c | X_p = u) < \infty$ for some constant $c > 2$, (iii) the random vector X_p is continuously distributed with support $I_p \supseteq I_U$, (iv) the corresponding density function f_p is bounded with bounded first order derivatives, and satisfies $\inf_{u \in \mathcal{I}_U} f_p(u) \geq \delta$ for some constant $\delta > 0$, (v) the function p_o is $(l_p + 1)$ times continuously differentiable.*

Assumption 2. *(i) the random vectors V is continuously distributed with compact support I_V, (ii) $\sup_{v \in I_V} \mathbb{E}(|Y_q|^c | X_q = v) < \infty$ for some constant $c > 2$, (iii) the random vector X_q is continuously distributed with support $I_q \supseteq I_V$, (iv) the corresponding density function f_q is bounded with bounded first order derivatives, and satisfies $\inf_{v \in I_V} f_q(v) \geq \delta$ for some constant $\delta > 0$ (v) the function q_o is $(l_q + 1)$ times continuously differentiable.*

Assumption 3. *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 4. *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, and (iii) such that the matrix $\Omega_o := \mathbb{E}(\psi_o(Z)\psi_o(Z)')$ is finite, where $\psi_o(Z) = \psi(Z, \theta_o, p_o(U), q_o(V))$*

Assumption 5. *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 1–2 are standard smoothness conditions in the context of nonparametric regression. The restrictions on the kernel function $\mathcal{K}$ in Assumption 3 could be weakened to allow for kernels with unbounded support. Parts (i)-(ii) of Assumption 4 impose some weak smoothness restrictions on the function ψ, which are needed to later justify a certain 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 (see Chen et al., 2003, for example). Parts (iii) of Assumption 4 ensures that the term $n^{1/2}\Psi_n(\theta_o, p_o, q_o)$ satisfies a central limit theorem. Finally, Assumption 5 imposes restrictions on the rate at which the bandwidths h_p and h_q tend to zero given the number of derivatives of the unknown regression functions and the dimension of the covariates. As argued above, these conditions are rather weak. Parts (i)–(iii) of Assumption 5 allow the smoothing bias from estimating either p_o or q_o to be as large as $o(n^{-1/6})$ as long as the *product* of the two bias terms is of the order $o(n^{-1/2})$, and parts (iv)–(v) only require the respective stochastic parts to be of the order $o_P(n^{-1/6})$. In contrast, to show that a generic semiparametric two-step estimator is RAL, it is generally required that the bias from estimating *each* nonparametric component alone is of the order $o(n^{-1/2})$, whereas the respective stochastic parts are generally required to be of the order $o_P(n^{-1/4})$ (see Newey and McFadden, 1994, for example).[1] Our assumptions yield the following theorem.

Theorem 1. *Under Assumption 1– 5, the following statements hold as $n \to \infty$.*

(i) $\widehat{\theta} \xrightarrow{p} \theta_o$;

(ii) $\sqrt{n}(\widehat{\theta} - \theta_o) = n^{-1/2}\sum_{i=1}^n \lambda_o(Z_i) + o_P(1)$, *where* $\lambda_o(z) = \Gamma_o^{-1}\psi(z, \theta_o, p_o(u), q_o(u))$ *and* $\Gamma_o = \partial\mathbb{E}(\psi(Z, \theta, p_o(U), q_o(V)))/\partial\theta|_{\theta=\theta_o}$;

(iii) $\sqrt{n}(\widehat{\theta} - \theta_o) \xrightarrow{d} N(0, \Gamma_o^{-1}\Omega_o\Gamma_o^{-1})$; *and*

(iv) $\widehat{\Gamma}^{-1}\widehat{\Omega}\widehat{\Gamma}^{-1} \xrightarrow{p} \Gamma_o^{-1}\Omega_o\Gamma_o^{-1}$.

Theorem 1 has a number of important implications. Taken together, parts (iii) and (iv) can e.g. be used to construct asymptotically valid confidence regions for θ_o or some of its components, or to conduct various large-sample testing procedures. Theorem 1 also shows that SDREs are adaptive semiparametric estimators, in the sense that they have the same first-order limiting

[1] Of course, such an estimator would often only use an estimate of either p_o or q_o, but not both, and thus require these rates to hold for estimates of one of the two functions only.

distribution as an infeasible estimator that uses the true values of the infinite-dimensional nuisance parameters instead of their nonparametric estimates. This is a property that SDREs share with all semiparametric estimators defined through a moment condition based on an influence function in the corresponding semiparametric problem (e.g. Newey, 1994). A further implication of this fact is that it is easy to verify whether the asymptotic variance of an SDRE achieves the corresponding semiparametric efficiency bound. This is the case if and only if the corresponding moment condition is based on the respective *efficient* influence function. This is the case for all settings that we listed in Section 2.2.

3.3. Advantages Relative to Generic Semiparametric Procedures. To illustrate in which sense Theorem 1 improves upon well-known results for semiparametric two-step estimators, it is useful to explicitly consider a "non-DR" estimator of θ_o that only uses a nonparametric estimate of one of the two nuisance functions. As pointed out above, such an estimator always exists in settings where there exists a DR moment condition. To be specific, we consider the estimator $\widehat{\theta}^*$ that solves

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

where $\psi^*(z, \theta, p(u)) = \psi(z, \theta, p(u), \bar{q}(v))$ for some arbitrary function $\bar{q} \in \mathcal{Q}$. Estimators like $\widehat{\theta}^*$ have been studied extensively in the literature. Without loss of generality, one could choose $\bar{q}$ such that $\widehat{\theta}^*$ and $\widehat{\theta}$ have the same influence function, but this is not important for the following discussion, which focuses on conditions under which the estimator is RAL. We introduce the following variations of Assumption 4–5.

Assumption 6. *The function $\psi^*(z, \theta, p(u))$ is (i) continuously differentiable with respect to θ, (ii) three times continuously differentiable with respect to $p(u)$, with derivatives that are uniformly bounded, and (iii) such that the matrix $\Omega_o^* := \mathbb{E}(\widetilde{\psi}_o^*(Z)\widetilde{\psi}_o^*(Z)')$ is finite, where $\widetilde{\psi}_o^*(Z) = \psi^*(Z, \theta_o, p_o(U)) + \alpha(Z)$ for $\alpha(z) = (y_p - p_o(x_p))\rho(x_p) f_U(x_p)/f_p(x_p)$ and $\rho(u) = \mathbb{E}(\partial\psi^*(Z, \theta, p)/\partial p|_{p=p_o(U)}|U = u)$.*

Assumption 7. *The bandwidth sequence h_p satisfies the following conditions as $n \to \infty$: (i) $nh_p^{2(l_p+1)} \to 0$ and (ii) $nh_p^{2d_q}/\log(n)^2 \to \infty$.*

Under these two conditions and Assumption 1 and 3, it follows from standard arguments from the literature on semiparametric estimation with first-stage kernel estimators (e.g. Newey and McFadden, 1994) that $\widehat{\theta}^*$ is RAL. For completeness, we formally state this result as a separate Theorem.[2]

Theorem 2. *Under Assumption 1, 3 and 6–7, we have that*

$$\sqrt{n}(\widehat{\theta}^* - \theta_o) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \lambda_o^*(Z_i) + o_P(1),$$

where $\lambda_o^*(z) = \Gamma_o^{*-1}(\psi^*(z, \theta_o, p_o(u)) + \alpha(z))$ *with* $\Gamma_o^* = \partial \mathbb{E}(\psi^*(Z, \theta, p_o(U)))/\partial\theta|_{\theta=\theta_o}$ *and* $\alpha(z)$ *as defined in Assumption 6, as* $n \to \infty$*.*

Comparing the conditions of Theorem 1 and Theorem 2 illustrates the benefits of using SDREs relative to generic semiparametric estimators. Strictly speaking, the conditions of Theorem 2 are neither weaker nor stronger than those used to establish Theorem 1. Showing that $\widehat{\theta}^*$ is RAL requires smoothness conditions on one functional nuisance parameter, in this case p_o, only, and also requires slightly weaker smoothness restrictions on the moment condition. Showing that $\widehat{\theta}$ is RAL requires smoothness conditions on both functional nuisance parameters, and stronger smoothness restrictions on the moment condition. On the other hand, if one is willing to impose such conditions, SDREs have a number of practically important advantages.

First, SDREs can generally make do with less stringent smoothness conditions on the functional nuisance parameters, and thus with lower order local polynomials, than generic semiparametric estimators. 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 5 is satisfied even if $l_p = l_q = 1$. For a generic estimator like $\widehat{\theta}^*$, using local linear smoothing in the first stage is typically only permissible if the unknown function is a one-dimensional nonparametric regression, since Assumption 7 cannot hold with $l_p = 1$ and $d_p > 1$. This is an important aspect, as higher-order local polynomial regression is well-known to have poor finite sample properties, especially when the covariates are multi-dimensional. This phenomenon is analogous to that of poor finite-sample performance of Nadaraya-Watson (or local constant) regression using higher-order kernels.

Second, in a semiparametric model where both SDREs and a generic semiparametric esti-

[2] We remark that the following Theorem can be shown under slightly weaker conditions if the term $\psi^*(z, \theta, p(u))$ is linear in $p(u)$.

mator are RAL given suitable nonparametric first-step estimates, the former generally allow for a much wider range of bandwidth values than the latter. To illustrate this point, consider the simple case that $d_p = d_q = l_p = l_q = 1$, where Assumption 5 certainly holds for $h_p \propto n^{-\delta_p}$ and $h_q \propto n^{-\delta_q}$ with $1/8 < \delta_p < 2/3$ and $1/8 < \delta_q < 2/3$. On the other hand, for $\widehat{\theta}^*$ to be RAL in this setting, one would require a bandwidth $h_p \propto n^{-\delta_p}$ with $1/4 < \delta_p < 1/2$, as otherwise Assumption 7 would be violated. Given the greater flexibility of the theory with respect to bandwidth choice, we would expect the finite-sample distribution of SDREs to be more robust to this issue than the finite-sample distribution of generic semiparametric estimators. This is useful for applications, as it simplifies the task of finding a bandwidth such that the standard large-sample inference procedures are approximately valid.

Third, in many instances the range of bandwidths that satisfy Assumption 5 includes the values that minimize the Integrated Mean Squared Error (IMSE) for estimating p_o and q_o, respectively. This is not the case for Assumption 7. For example, in the simple case that $d_p = d_q = l_p = l_q = 1$, choosing $h_p \propto h_q \propto n^{-1/5}$ is sufficient to satisfy Assumption 5. We could thus in principle choose an estimate of the bandwidths that minimize

$$\int (p_o(u) - \widehat{p}(u))^2 du \quad \text{and} \quad \int (q_o(v) - \widehat{q}(v))^2 dv, \tag{3.2}$$

respectively, which are well known to be proportional to $n^{-1/5}$ in this case. While strictly speaking these bandwidth do not have any optimality properties for estimating θ_o, they have the 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 is important, since the lack of an objective method for bandwidth selection is one of the major obstacles for applying semiparametric methods in practice.

Fourth, and finally, the difference between a SDRE and its RAL approximation is typically of smaller order than the difference between a generic semiparametric estimator and its RAL approximation. As a consequence, one can expect the Gaussian approximation resulting from an RAL result to be more accurate for SDREs than for generic estimators. To see this, consider first the generic estimator $\widehat{\theta}^*$ for the simple case that $d_p = l_p = 1$, and that the other conditions of Theorem 2 hold. Then one can show (e.g. Ichimura and Linton, 2005; Cattaneo et al., 2012a) that

$$\widehat{\theta}^* - \theta_o - \frac{1}{n}\sum_{i=1}^n \lambda_o^*(Z_i) = O_P(h_p^2) + O_P(n^{-1}h_p^{-1}). \tag{3.3}$$

The first term on the right-hand side of this equation is related to the smoothing bias from estimating the nonparametric component, whereas the second term is related to the variance from the nonparametric estimation step. Cattaneo et al. (2012a) refer to the latter term as a “nonlinearity bias”, as its occurrence is related to whether the final estimator depends nonlinearly on first-stage nonparametric estimator. Its magnitude is not affected by the use of techniques that reduce smoothing bias, and would increase with the dimension of the covariate vector in settings with $d_p > 1$. The bandwidth that minimizes the magnitude of the two second order terms on the right-hand side of the last equation satisfies $h_p \propto n^{-1/3}$. With this choice of bandwidth one finds that

$$\widehat{\theta}^* - \theta_o - \frac{1}{n}\sum_{i=1}^n \lambda_o^*(Z_i) = O_P(n^{-2/3}). \tag{3.4}$$

Thus, while the second order terms vanish faster than $n^{-1/2}$, they are still relatively large. On the other hand, suppose that $d_p = d_q = l_p = l_q = 1$, that $h_p \propto h_q$, and that the other conditions of Theorem 1 hold. Then, by following the steps of the proof of Theorem 1, one finds that

$$\widehat{\theta} - \theta_o - \frac{1}{n}\sum_{i=1}^n \lambda_o(Z_i) = O_P(h_p^4) + O_P(n^{-1}h_p^{-1/2}). \tag{3.5}$$

We can see that the DR structure of the moment condition has reduced the impact of both the smoothing bias and the nonlinearity bias, making the RAL approximation more accurate for the SDRE than for the generic estimator. The bandwidth that minimizes the magnitude of the two second order terms on the right-hand side of the last equation satisfies $h_p \propto n^{-2/9}$. With such a choice of bandwidth, we have that

$$\widehat{\theta} - \theta_o - \frac{1}{n}\sum_{i=1}^n \lambda_o(Z_i) = O_P(n^{-8/9}),$$

and the second order terms thus vanish faster than the in the best-possible case we could achieve for a generic estimator. As a consequence, we would thus e.g. expect confidence intervals based on a normality result to have better finite-sample coverage properties for SDREs than for generic estimators.

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. 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)

$$\gamma_0 = \mathbb{E}(Y(1) - Y(0)|D = 1)$$

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 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 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, and to ensure that there exists a RAL semiparametric estimator (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 in the absence of knowledge of the propensity score the asymptotic variance of any regular estimator of the ATE and ATT is bounded from below by

$$\begin{aligned} V_{ate} &= \mathbb{E}\left(\frac{\sigma^2(1,X)}{\pi_o(X)} + \frac{\sigma^2(0,X)}{1-\pi_o(X)} + (\mu_o^Y(1,X) - \mu_o^Y(0,X) - \tau)^2\right) \text{ and} \\ V_{att} &= \mathbb{E}\left(\frac{\pi_o(X)}{\Pi_o^2}\left(\sigma^2(1,X) + \frac{\pi_o(X)\sigma^2(0,X)}{1-\pi_o(X)} + (\mu_o^Y(1,X) - \mu_o^Y(0,X) - \gamma_o)^2\right)\right), \end{aligned}$$

respectively, where $\sigma^2(d,x) = \mathrm{Var}(Y|D = d, X = x)$. 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), amongst 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 many others, and are widely used in applied work. Cattaneo (2010) was the first to propose a SDRE for ATE in a model with multiple treatment levels. However, he did not formally prove that SDREs have favorable properties relative to generic 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 $\mu_o(x) = (\mu_o^Y(1,x), \mu_o^Y(0,x))$ 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. 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:

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

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, writing $\lambda_\pi(x) = [x_1, x_1^2, \ldots, x_1^{l_\pi}, \ldots, x_{d_X}, x_{d_X}^2, \ldots, x_{d_X}^{l_\pi}]^T$ and $\lambda_\mu(x) = [x_1, x_1^2, \ldots, x_1^{l_\mu}, \ldots, x_{d_X}, x_{d_X}^2, \ldots, x_{d_X}^{l_\mu}]^T$, 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

$$(\widehat{a}_\pi(X_i),\widehat{b}_\pi(X_i)) = \operatorname*{argmin}_{a,b}\sum_{j\neq i}\left(D_j - \Phi(a - b'\lambda_\pi(X_j-X_i))\right)^2 K_{h_\pi}(X_j-X_i),$$
$$(\widehat{a}_\mu(d,X_i),\widehat{b}_\mu(d,X_i)) = \operatorname*{argmin}_{a,b}\sum_{j\neq i}\mathbb{I}\{D_j=d\}\left(Y_j - a - b'\lambda_\mu(X_j-X_i)\right)^2 K_{h_\mu}(X_j-X_i),$$

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 behaviour.

To study the asymptotic properties of the SDRE $\widehat{\tau}_{DR}$, we impose the following assumptions, which essentially restate the content of Assumption 1–2 using the notation of the present treatment effects setting.

Assumption 8. *(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 9. *(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 3. *Suppose Assumption 8–9 hold, and that Assumption 3–5 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) $\widehat{\tau}_{DR} \xrightarrow{p} \tau_o$,

ii) $\sqrt{n}(\widehat{\tau}_{DR} - \tau_o) = n^{-1/2}\sum_{i=1}^n \psi_{ate}(Z_i, \tau_o, \pi_o(X_i), \mu_o(X_i)) + o_P(1)$,

iii) $\sqrt{n}(\widehat{\tau}_{DR} - \tau_o) \xrightarrow{d} N(0, V_{ate}^*)$.

iv) $\widehat{\tau}_{DR}$ achieves the semiparametric efficiency bound for estimating τ_o.

v) $\widehat{V}(\widehat{\tau}_{DR}) \xrightarrow{p} V_{ate}$.

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 $\tau_o = \mathbb{E}(\mu_o^Y(1,X) - \mu_o^Y(0,X))$. 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] We therefore recommend using the SDRE in practice.

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 $d_X < 4$, 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

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

[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.

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. 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:

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

The following Theorem establishes the estimator's asymptotic properties.

Theorem 4. *Suppose Assumption 8–9 hold, and that Assumption 3–5 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) $\widehat{\gamma}_{DR}\xrightarrow{p}\gamma_o$,

ii) $\sqrt{n}(\widehat{\gamma}_{DR}-\gamma_o)=n^{-1/2}\sum_{i=1}^n\psi_{att}(Z_i,\gamma_o,\pi_o(X_i),\mu_o^Y(0,X_i),\Pi_o)+o_P(1)$,

iii) $\sqrt{n}(\widehat{\gamma}_{DR}-\gamma_o)\xrightarrow{d}N(0,V_{att})$

iv) $\widehat{\gamma}_{DR}$ *achieves the semiparametric efficiency bound* γ_o *in the absence of knowledge of the propensity score.*

v) $\widehat{V}(\widehat{\gamma}_{DR})\xrightarrow{p}V_{att}$.

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 $\mu_o^Y(0,\cdot)$ (e.g. Imbens et al., 2005).

5. Monte Carlo

In this section, we illustrate the properties of SDREs relative to other semiparametric two-step estimators through a small scale Monte Carlo experiment. We consider the simple missing data model present 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 standard deviation .5, and the missingness indicator D is generated as a Bernoulli random variable with mean $\pi_o(X) = .2 + .8(1 - X^2)$. Our parameter of interest is $\theta_o = \mathbb{E}(Y) = 1$, and the semiparametric variance bound for estimating this parameter is $V^* \approx 1.644$. We study the sample size $n = 200$, and set the number of replications to 1,000. 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):

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

We defined $\widehat{\pi}$ as the "leave-i-out" local linear Probit estimator of $\pi_o(x)$ using the bandwidth $h \in \{.05, .15, \ldots, .45\}$, and $\widehat{\mu}(x)$ as the "leave-i-out" local linear estimator of $\mu_o(x)$ using a bandwidth $g \in \{.02, .04, \ldots, .1\}$.[4] 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

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

[4] We determined the range for the bandwidths as follows. In a preliminary simulation study, we generated 1,000 samples of size $n = 200$ from the distribution of (YD, D, X), and estimated the bandwidths that minimize a least squares cross-validation criterion for estimating π_o and μ_o, respectively. The range for h and g we then consider for our main simulation study corresponds roughly to that between the 5% and 95% empirical quantiles of the distribution of the respective cross-validation bandwidths. We expect that the range of bandwidths we consider covers most rules of thumb that practitioners might use in applications. We also experimented with a wider range of bandwidths to ensure that our choice contains the values that approximately minimize the respective MSE of the three estimators we consider in this simulation study.

Figure 1: Simulation results: MSE, bias and variance of the IPW estimator for various values of h (solid line), compared to results for the DR estimator with bandwidth g equal to .02 (dashed line), .04 (dotted line), .06 (dot-dashed line), .08 (long dashed line), and .10 (long dashed dotted line).

Figure 2: Simulation results: MSE, bias and variance of the REG estimator for various values of g (solid line), compared to results for the DR estimator with bandwidth h equal to .05 (dashed line), .15 (dotted line), .25 (dot-dashed line), .35 (long dashed line), and .45 (long dashed dotted line).

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 bias, and the variance of the IPW estimator as a function of the bandwidth h, and compare the results to those of the DR estimator for 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 for the DR estimator for various values of the bandwidth h. We also present the same results in table form in Table 1–3. One can clearly see that 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 $h = .15$ and $g = .02$, respectively, but these values would be very difficult to determine in an empirical application. For the DR estimators, we observe that those using one of the two smallest bandwidths, i.e. either $h = .05$ or $g = .02$, exhibit somewhat different behavior from the remaining ones. For DR estimators using $h > .05$ and $g > .02$, the MSE, bias and variance all exhibit only minimal variation with respect to the bandwidth. Their variance is substantially lower than that of IPW, and somewhat lower than that of REG for larger values of g. It is also very close to the semiparametric efficiency bound, which is equal to about 1.644 in our simulation design. The DR estimators are also essentially unbiased for all bandwidth choices. DR estimators using either $h = .05$ or $g = .02$ have somewhat higher variance than those using larger bandwidths, but are also essentially unbiased. As a consequence, they also compare favorably to both IPW and REG in terms of MSE. In applications, we would recommend to implement DR estimators using bandwidths that are relatively large.

We also computed 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 4. 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 for small and large values of h and good coverage properties for intermediate values, irrespective of

Table 1: Simulation Results: MSE of the IPW, REG and DR estimator for various bandwidth values (all results scaled by the sample size).

Bandwidth	*h*	*.05*	*.15*	*.25*	*.35*	*.45*	
g	Estimator	DR					REG
.02	DR	1.839	1.737	1.733	1.731	1.730	1.725
.04		1.721	1.680	1.680	1.679	1.679	1.747
.06		1.696	1.663	1.663	1.662	1.662	1.982
.08		1.683	1.656	1.657	1.656	1.656	2.497
.10		1.675	1.653	1.655	1.655	1.655	3.297
	IPW	2.199	1.730	1.826	1.958	2.049	

Table 2: Simulation Results: Bias of the IPW, REG and DR estimator for various bandwidth values (all results scaled by the square root of the sample size).

Bandwidth	*h*	*.05*	*.15*	*.25*	*.35*	*.45*	
g	Estimator	DR					REG
.02	DR	0.020	0.012	0.011	0.011	0.011	0.065
.04		0.018	0.015	0.014	0.014	0.014	0.264
.06		0.012	0.011	0.011	0.011	0.011	0.558
.08		0.005	0.009	0.010	0.010	0.009	0.906
.10		0.003	0.006	0.008	0.008	0.007	1.269
	IPW	0.509	0.013	0.170	0.177	0.122	

Table 3: Simulation Results: Variance of the IPW, REG and DR estimator for various bandwidth values (all results scaled by the sample size).

Bandwidth	*h*	*.05*	*.15*	*.25*	*.35*	*.45*	
g	Estimator	DR					REG
.02	DR	1.838	1.737	1.732	1.731	1.730	1.721
.04		1.720	1.680	1.679	1.679	1.678	1.677
.06		1.696	1.663	1.663	1.662	1.662	1.671
.08		1.683	1.656	1.656	1.656	1.656	1.677
.10		1.675	1.653	1.655	1.655	1.655	1.686
	IPW	1.939	1.730	1.797	1.926	2.034	

the choice of g. REG-based confidence intervals have good coverage properties for $g = .02$ and $g = .04$, and substantial under-coverage for large values of g, irrespective of the choice of h.

6. 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 very favorable properties relative to other semiparametric estimators that are currently widely used in such settings, such as e.g. Inverse Probability Weighting, and

Table 4: 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*	*.05*	*.15*	*.25*	*.35*	*.45*
	.02	0.949	0.946	0.945	0.945	0.945
	.04	0.948	0.948	0.948	0.948	0.948
	.06	0.948	0.949	0.949	0.948	0.948
	.08	0.949	0.949	0.949	0.949	0.949
	.10	0.951	0.949	0.950	0.950	0.950
IPW	*g / h*	*.05*	*.15*	*.25*	*.35*	*.45*
	.02	0.928	0.940	0.936	0.926	0.924
	.04	0.924	0.943	0.936	0.926	0.922
	.06	0.923	0.942	0.937	0.926	0.923
	.08	0.922	0.941	0.937	0.926	0.923
	.10	0.921	0.941	0.937	0.926	0.923
REG	*g / h*	*.05*	*.15*	*.25*	*.35*	*.45*
	.02	0.950	0.948	0.946	0.946	0.946
	.04	0.946	0.942	0.941	0.941	0.941
	.06	0.933	0.930	0.928	0.928	0.929
	.08	0.895	0.891	0.890	0.890	0.890
	.10	0.839	0.836	0.835	0.835	0.835

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 root-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., 2012a,b).

A. Proofs of Main Results

A.1. Proof of Theorem 1. Statement (i) is immediately implied by statement (ii), and could also be derived under weaker conditions. Statement (iii) follows from (ii) and a simple application of a Central Limit Theorem. Statement (iv) follows from standard arguments, and we thus omit an extensive proof for brevity. It thus remains to show statement (ii). To prove that result, note that it follows from the differentiability of ψ with respect to θ and the definition of $\widehat{\theta}$ that

$$\sqrt{n}(\widehat{\theta} - \theta_o) = \Gamma_n(\theta^*, \widehat{p}, \widehat{q})^{-1}\sqrt{n}\Psi_n(\theta_o, \widehat{p}, \widehat{q})$$

for some intermediate value θ^* between θ_o and $\widehat{\theta}$, and $\Gamma_n(\theta, p, q) = \partial\Psi_n(\theta, p_o, q_o)/\partial\theta$. 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 $\Psi_n(\theta_o, \widehat{p}, \widehat{q})$. 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 4 we have that

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

By Lemma 2(i) and Assumption 5, 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 our proof. □

A.2. Proof of Theorem 2. Follows standard arguments from the literature on semiparametric estimation with first-stage kernel estimators (e.g. Newey and McFadden, 1994).

A.3. Proof of Theorem 3 and 4. 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 seperately 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 et al. (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. To state the expansion, we define the following quantities:

$$\begin{aligned}
w(u) &= (1, u_1, ..., u_1^{l_p}, u_2, ..., u_2^{l_p}, \dots, u_{d_p}, ..., u_{d_p}^{l_p})^T \\
w_j(u) &= w((X_{p,j} - u)/h_p)\cdot \\
M_{p,n}(u) &= \frac{1}{n}\sum_{j\neq i}^{n} w_j(u)w_j(u)^\top K_{h_p}(X_{p,j} - u), \\
N_{p,n}(u) &= \mathbb{E}(w_j(u)w_j(u)^\top K_{h_p}(X_{p,j} - u)), \\
\eta_{p,n,j}(u) &= w_j(u)w_j(u)^\top K_{h_p}(X_{p,j} - u) - \mathbb{E}(w_j(u)w_j(u)^\top K_{h_p}(X_{p,j} - u)).
\end{aligned}$$

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 behaviour of the two stochastic terms, it is again instructive to consider simple case that $l_p = 0$, for which

$$S_{p,n}(U_i) = \frac{1}{n\bar{f}_{p,n}(U_i)} \sum_{j\neq i} K_{h_p}(X_{p,j} - U_i)\varepsilon_{p,j} \text{ and}$$

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

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 1–3, the following statements hold:*

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

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

and the first- and second-order stochastic terms satisfy

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

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

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

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

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

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 4 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 1–5, the following statements hold:*

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

Proof. We only show the first statement, as the proof for the second one is fully analogous. From Lemma 2 and Assumption 5, it follows that

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

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 5, 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 5. 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 $O_P(h_p^{l_p+1})$, 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 2 below).

Lemma 5. *Under Assumption 1–5, 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 5, 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

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

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:

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

The statement of the lemma thus follows if $h_p \to 0$ and $n^2h_p^{3d_p}/\log(n)^3 \to \infty$ as $n \to \infty$, which holds by Assumption 5. 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

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

The term of the order $O_P(n^{-1}h_p^{-d_p})$ is the "degrees of freedom bias" in Ichimura and Linton (2005), and analogous to the "nonlinearity bias" or "curse of dimensionality bias" in Cattaneo et al. (2012a). In our context, this term is removed by the DR property of the moment conditions, which illustrates how the structure of the latter acts like a bias correction method.

Lemma 6. *Under Assumption 1–5, the following statement holds:*

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

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 5, 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

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

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. On the other hand, the term $T_{n,5,A}$ is equal to n^{-1} times a non-degenerate second order U-statistic (up to a bounded, multiplicative term), and thus

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

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

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

The statement of the Lemma then follows from Assumption 5. This completes our proof. □

References

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

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 (2012a): "Generalized Jackknife Estimators of Weighted Average Derivatives," *Working Paper.*

——— (2012b): "Small bandwidth asymptotics for density-weighted average derivatives," *Econometric Theory.*

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics*, 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*, vol. 66, 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.

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

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," *The 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*, 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.

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.

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, pp. 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.

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.