Semiparametric Two-Step Estimation Using Doubly Robust Moment Conditions

Christoph Rothe and Sergio Firpo*

Abstract

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

JEL Classification: C14, C21, C31, C51

*First version: December 20, 2012. This version: May 23, 2014. Christoph Rothe, Columbia University, Department of Economics, 420 W 118th St, New York, NY 10027, USA. Email: cr2690@columbia.edu. Sergio Firpo, Escola de Economia de Sao Paulo FGV-SP, R. Itapeva, 474/1215, Sao Paulo-SP, 01332-000, Brasil. E-Mail: sergio.firpo@fgv.br. We would like to thank Matias Cattaneo, Michael Jansson, Marcelo Moreira, Ulrich Müller, Whitney Newey, Cristine Pinto, and seminar audiences at Brown, Columbia, EPGE-FGV, University of Pennsylvania, Princeton, PUC-Rio, the 2012 Greater NY Metropolitan Colloquium and the 2013 North American Summer Meetings for their helpful comments. Sergio Firpo gratefully acknowledges financial support from CNPq-Brazil.

1. Introduction

Semiparametric two-step estimators are important tools in several areas of empirical economic research. They typically arise in models where the parameter of interest is identified by a moment condition that contains unknown nuisance functions. The estimator is then obtained by setting a sample analogue of the moment condition as close to zero as possible, with the nuisance functions being replaced by preliminary nonparametric estimates. Under certain restrictions, such semiparametric estimators are $\sqrt{n}$-consistent and asymptotically normal (CAN), and their asymptotic variance is invariant with respect to the choice of nonparametric estimator (e.g. Newey, 1994). This type of first order asymptotic theory is commonly used to justify a Gaussian approximation to the sampling distribution of the estimator, which in turn is the basis for most classical large sample inference procedures. While convenient, such approximations may be overly optimistic about the minuteness of asymptotically negligible "second order" terms, at least for the sample sizes typically encountered in empirical practice. As a consequence, classical inference can potentially be misleading in semiparametric settings (e.g. Linton, 1995; Robins and Ritov, 1997).

The absolute magnitude of second order terms, and thus the accuracy of first order distributional approximations, is known to vary not only with the type of nonparametric estimator being used in the first stage, but also with the structure of the moment condition (e.g. Newey, Hsieh, and Robins, 2004). Indeed, in most semiparametric models there exist many different moment conditions containing different nuisance functions that can in principle all be used to construct an estimator of the parameter of interest. Even if attention is restricted to moment conditions that lead to an estimator whose first order variance achieves the semiparametric efficiency bound, practitioners can still often choose from several approaches with potentially very different second order properties. This is for instance the case for average treatment effects under unconfoundedness (e.g. Imbens, 2004) or weighted average derivatives (e.g. Stoker, 1991).

In this paper, we argue that the class of *doubly robust* (DR) moment conditions is particularly attractive for the construction of semiparametric two-step estimators. Here we follow the terminology of Robins, Rotnitzky, and van der Laan (2000) and say that a moment condition is DR if it depends on two unknown nuisance functions, but still identifies the parameter of interest if either one of these functions is replaced by some arbitrary value. Moment conditions with this particular structure exist in many interesting (but by no means

all) semiparametric models, and are often found as efficient influence functions for the respective parameter of interest. Some examples that are particularly relevant in econometrics include missing data and treatment effect models, partially linear regression models, models for nonparametric policy analysis, and weighted average derivatives. We refer to a semiparametric two-stage estimator based on a DR moment condition as a *semiparametric doubly robust estimator* (SDRE).

Our first main result is that SDREs are CAN under weaker-than-usual conditions on the accuracy of the first step nonparametric estimates. This is because DR moment conditions are insensitive to variation in the value of the nuisance functions in a particular sense. More specifically, at the true parameter value their kth order functional derivatives with respect to each of the two nuisance functions are zero for any k. This means for example that the familiar requirement that the nonparametric component converges with a rate that is faster than $n^{-1/4}$ can be relaxed, as the particular structure of DR moments automatically removes certain "linear" and "quadratic" terms in a stochastic expansion of the final estimator that otherwise need to be controlled by this assumption. This result is not tied to the use of a particular nonparametric first stage estimator. We show explicitly in this paper how it can be verified for kernel-based smoothers, but a similar argument could also be made with series estimators for example. Our only substantial requirement is that the asymptotic correlation between the two nonparametric estimates and the product of their biases vanishes sufficiently quickly, in a sense which we make precise below.

Our second main result is an explicit characterization of the second order properties of SDREs when the first stage is a kernel estimator, such as a local polynomial regression. We derive explicit rates at which the largest second order components of SDREs tend to zero in this case, and approximate their mean and variance. Standard expansions of semiparametric two-step estimators contain "linear" and "quadratic" components that usually give rise to non-trivial and mostly bias related terms whose magnitudes depend on both the smoothness and the dimensionality of the nonparametric component (often called "smoothing" and "nonlinearity" bias, respectively; see Cattaneo, Crump, and Jansson (2013)). The structure of DR moments removes those terms, and SDREs thus generally have smaller first order bias and second order variance than many standard estimators, and differ less from their asymptotically linear representation. We present explicit rate results for kernel-based smoothers in the first stage as those are typically easier to be obtained than the ones for series smoothers.

From a practical perspective, our results imply that inference based on SDREs can generally be expected to more accurate in finite samples than that based on a generic semiparametric estimator. Moreover, the finite sample distribution of SDREs should be less sensitive to the implementation of the nonparametric first stage. In settings with moderate dimensionality, SDREs also allow for first stage estimates that converge at the optimal nonparametric rate. This can be useful in practice because smoothing parameters that lead to such a rate are relatively easy to estimate from the data via cross-validation, for example. For similar reasons, SDREs can also make do without the use of bias reducing nonparametric estimators, such as those based on higher order kernels. This is practically relevant, as such a reduction in asymptotic smoothing bias often comes at the cost of a substantial increase in finite sample variance (e.g. Marron and Wand, 1992). Finally, SDREs are also adaptive by construction, in the sense that their first order asymptotic variance does not contain adjustment terms for the nonparametric first step. This means that valid standard errors can be calculated without estimating any further nuisance parameters.

To illustrate the usefulness of our general results, we also apply them to the four substantial examples mentioned above: a general missing data model that also covers many popular program evaluation setups, the partially linear regression model, a model for nonparametric policy analysis, and weighted average derivatives. Our theory produces a number of new results for these settings. For example, it shows that in the missing data context the SDRE dominates many popular competitors such as the Inverse Probability Weighting estimator (e.g. Hirano, Imbens, and Ridder, 2003; Ichimura and Linton, 2005). The theory also reproduces some familiar results. For example, it turns out that Robinson's (1988) famous estimator of the parametric component of a partially linear model is an example of an SDRE, and that our general higher-order results yield conclusions similar to those of Linton's (1995) study of that estimator.

The remainder of this paper is structured as follows. In the next subsection, we provide a review of the related literature. In Section 2 we introduce the notion of DR moment conditions, give some concrete examples, and describe our estimation procedure. Section 3 gives a general asymptotic normality result. In Section 4, we show how the conditions of this result can be verified in a setting with first stage kernel regressions. Section 5 applies our results to the examples discussed in Section 2. Finally, Section 6 concludes. The appendix contains several technical arguments.

1.1. Related Literature. The first order asymptotic properties of semiparametric two-step estimators have been studied extensively by Newey (1994), Newey and McFadden (1994), Andrews (1994), Chen, Linton, and Van Keilegom (2003), Ai and Chen (2003), Chen and Shen (1998) and Ichimura and Lee (2010), among many others. Newey and McFadden (1994) show that with first step kernel estimation, sufficient conditions for the final estimator to be CAN include that the rate of convergence and bias of the nonparametric component must be $o_P(n^{-1/4})$ and $o(n^{-1/2})$, respectively. These conditions are also known to be necessary for many, but not all, semiparametric estimators. Since we also use first step kernel estimation when deriving the exact higher-order properties of SDREs in this paper, we use Newey and McFadden's result as our baseline when commenting on improvements that can be achieved from using SDREs instead of generic semiparametric estimators.

It is known that the just-mentioned restrictions on the rate of convergence and the bias can be weakened when using other types of nonparametric first stage estimators and/or a moment condition with a particular structure. Consequently, the relative gains from using SDREs will be smaller in such settings, even though they can still be substantial. Newey (1994) shows that when using an orthogonal series estimator in the first stage a less stringent "small bias" condition suffices for a general class of moment conditions; and Shen et al. (1997) and Chen and Shen (1998) establish a similar result for sieve maximum likelihood. Newey et al. (2004) argue that such an effect can also be achieved for kernel-based estimators by using a twicing kernel, or by using a moment condition based on an influence function in the corresponding semiparametric problem. Since DR moments (if it exists) are always based on an influence function (Robins and Rotnitzky, 2001), our SDREs benefit from this effect as well; but their advantages go substantially beyond that. Hall and Marron (1987) show that the rate requirement can be weakened by using "leave-one-out" kernel estimators in the special case that the final estimator is a linear kernel average. See also Powell, Stock, and Stoker (1989). Cattaneo et al. (2013) show that a jackknifed version of the weighted average derivative estimator of Stoker (1986) can also make do with a slower rate of convergence. For SDREs, such a result follows from the structure of the moment condition, together with any construction that ensures that the correlation between the two nonparametric components vanishes sufficiently quickly.

In parametric problems, concerns about the accuracy of first order distributional approximations are often be addressed by using the bootstrap, which is able to achieve asymptotic

refinements in many settings (e.g. Hall, 1992). Unfortunately, there exist hardly any comparable results for semiparametric estimators in the literature. Although suitably implemented bootstrap procedures are known to be first order valid in semiparametric settings (e.g. Chen et al., 2003), to the best of our knowledge the only paper that establishes an asymptotic refinement is Nishiyama and Robinson (2005), which studies the density-weighted average derivative estimator of Powell et al. (1989); but see Cattaneo, Crump, and Jansson (2014b) for a cautionary tale regarding the robustness of such refinements.

An alternative approach to improving inference, which we do not consider in this paper, would be to derive "non-$\sqrt{n}$" asymptotic approximations to the distribution of semiparametric two-step estimators. 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 (2014a), who study so-called small bandwidth asymptotics for the density-weighted average derivative estimator.

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

2. Semiparametric Doubly Robust Estimation

2.1. Model. We consider the problem of estimating a parameter θ^o, contained in the interior of some compact parameter space $\Theta \subset \mathbb{R}^{d_\theta}$, using an i.i.d. sample $\{Z_i\}_{i=1}^n$ from the distribution of the random vector $Z \in \mathbb{R}^{d_z}$. The distribution of Z belongs to some semiparametric model. We assume that one way to identify θ^o is through a moment condition containing an infinite dimensional nuisance parameter. That is, we assume that the model is such that there exists a known moment function ψ taking values in $\mathbb{R}^{d_\theta}$ which satisfies the

following relationship:

$$\Psi(\theta, \xi^o) := \mathbb{E}(\psi(Z, \theta, \xi^o)) = 0 \text{ if and only if } \theta = \theta^o. \tag{2.1}$$

Here $\xi^o \in \Xi$ is an unknown (but identified) nuisance function that might also depend on θ^o. Semiparametric two-step estimation of θ^o in this type of moment condition models has been studied extensively by e.g. Newey (1994), Andrews (1994), Chen et al. (2003), and Ichimura and Lee (2010).

In most semiparametric models, there exist several moment conditions of the form in (2.1), and in principle any of them could be used to construct a semiparametric estimator of θ^o. In this paper, we study a class of settings where there exists a moment condition with a particular structure. Specifically, we assume that for some moment of the form in (2.1) the function ξ^o can be partitioned as $\xi^o = (\xi_1^o, \xi_2^o) \in \Xi_1 \times \Xi_2$ such that

$$\Psi(\theta, \xi_1^o, \xi_2) = 0 \text{ and } \Psi(\theta, \xi_1, \xi_2^o) = 0 \text{ if and only if } \theta = \theta^o \tag{2.2}$$

for all functions $\xi_1 \in \Xi_1$ and $\xi_2 \in \Xi_2$. Following the terminology of Robins et al. (2000), we refer to any functional Ψ that is of the form in (2.1) and satisfies the restriction (2.2) as a *doubly robust moment condition* for estimating θ^o, and to the corresponding function ψ as a *doubly robust moment function*. Such moment functions often coincide with the efficient influence function for estimating θ^o. Indeed, Robins and Rotnitzky (2001) show that a DR moment function (if it exists) has to be an element of the space of influence functions of the corresponding semiparametric model. Since any influence function for estimating a d_θ-dimensional parameter takes values in $\mathbb{R}^{d_\theta}$ by construction, our focus on exactly identified settings in (2.1) is thus without loss of generality.

2.2. Applications. Doubly robust moment functions do not exist in all semiparametric models. To get a sense of the range of applications covered by this structure, it is helpful to consider a few specific examples.

Example 1 (Missing Data)**.** Suppose that the underlying full data are a sample from the distribution of (Y^*, X), and let D be an indicator variable with $D = 1$ if Y^* is observed and $D = 0$ otherwise. The observed data thus consist of a sample from the distribution of $Z = (Y, X, D)$, where $Y = DY^*$. Also suppose that the parameter θ_o is the *unique* solution

of the nonlinear moment condition $\mathbb{E}(m(Y^*, X, \theta)) = 0$, where $m(\cdot, \theta)$ is a known function taking values in $\mathbb{R}^{d_\theta}$. In order to achieve identification, assume that Y^* is missing at random, i.e. $Y^* \perp D|X$. This type of setup is the one in which parametric DR estimators are most commonly studied. It covers a number of important special cases, including linear regression models with missing covariates and/or outcome variables (Robins and Rotnitzky, 1995), average treatment effects under unconfoundedness (Imbens, 2004), and local average treatment effects with an instrument that is only valid conditional on some observed covariates (Tan, 2006). See Appendix C for more details. Now define $\xi_1^o(x, \theta) = \mathbb{E}(m(Y, X, \theta)|D = 1, X = x)$; and let $\xi_2^o(x) = \mathbb{E}(D|X = x)$ be the propensity score, which is assumed to be bounded away from zero. Then

$$\psi_{MD}(Z, \theta, \xi) = \frac{D(m(Y, X, \theta) - \xi_1(X, \theta))}{\xi_2(X)} + \xi_1(X, \theta)$$

is a doubly robust moment function for estimating θ^o. □

Example 2 (Partial Linear Model)**.** Suppose that the data consist of a sample from the distribution of $Z = (Y, X, W)$, where Y is a scalar outcome variable and both X and W are vectors of explanatory variables. Then a partially linear regression model assumes that $Y = \phi^o(X) + W^\top \theta^o + \varepsilon$, where ϕ^o is some smooth unknown function, θ^o is a vector of parameters, and ε is an unobserved random variable that satisfies $\mathbb{E}(\varepsilon|X, W) = 0$. This model is commonly used for example to estimate demand curves (e.g. Engle, Granger, Rice, and Weiss, 1986; Hausman and Newey, 1995; Blundell, Duncan, and Pendakur, 1998). Now define $\xi_1^o(x, \theta) = \mathbb{E}(Y - W^\top \theta|X = x)$ and $\xi_2^o(x) = \mathbb{E}(W|X = x)$. Then

$$\psi_{PLM}(Z, \theta, \xi) = (Y - W^\top \theta - \xi_1(X, \theta))(W - \xi_2(X))$$

is a doubly robust moment function for estimating θ^o. □

Example 3 (Policy Effects)**.** Suppose that the data consist of sample from the distribution of $Z = (Y, X)$, where Y is a scalar dependent variable and X is a vector of continuous explanatory variables with density ξ_2^o. The problem is to predict the effect of a change in the distribution of X to that of $\pi(X)$, where π is some known *policy function*, on the mean of the dependent variable. Under an exogeneity condition, this mean effect is given by $\theta^o = \mathbb{E}(\xi_1^o(\pi(X)))$, where $\xi_1^o(x) = \mathbb{E}(Y|X = x)$. See Stock (1989) and Rothe (2010, 2012) for

details and applications. In this context

$$\psi_{PE}(Z,\theta,\xi) = \xi_1(\pi(X)) + (Y - \xi_1(X))\frac{\xi_2^\pi(X)}{\xi_2(X)} - \theta$$

is a doubly robust moment function for estimating θ^o, where we use the notation that f^π is the density of $\pi(X)$ when X has density f. □

Example 4 (Weighted Average Derivatives)**.** Suppose that the data consist of sample from the distribution of $Z = (Y, X)$, where Y is a scalar dependent variable and X is a vector of continuously distributed random variables with density function ξ_2^o. Then the weighted average derivative (WAD) of the regression function $\xi_1^o(x) = \mathbb{E}(Y|X = x)$ is defined as $\theta^o = \mathbb{E}(w(X)\nabla_x\xi_1^o(X))$, where w is a known scalar weight function. WADs are important for estimating the coefficients in linear single-index models (e.g. Stoker, 1986; Powell et al., 1989; Stoker, 1991; Newey and Stoker, 1993), and as a summary measure of nonparametrically estimated regression functions more generally. In this context

$$\psi_{WAD}(Z,\theta,\xi) = w(X)\nabla_x\xi_1(X) - (Y - \xi_1(X))\left(\nabla_x w(X) + w(X)\frac{\nabla_x\xi_2(X)}{\xi_2(X)}\right) - \theta$$

is a doubly robust moment function for estimating θ^o. □

2.3. Estimator. A semiparametric doubly robust estimator (SDRE) is a two-step estimator of θ^o that solves a direct sample analogue of the doubly robust moment condition (2.1). Before formally defining the estimator, we introduce a final bit of structure. We assume that for all $\xi = (\xi_1, \xi_2) \in \Xi_1 \times \Xi_2$ and each $g \in \{1, 2\}$ there exists a t_g-vector $\zeta_g = (\zeta_{g1}, \ldots, \zeta_{gt_g}) \in \aleph_g$ of functions such that $\psi(Z, \theta, \xi_1, \xi_2)$ depends on ξ_g through $\zeta_g(U_g)$ only, where U_g is a subvector of Z. All the examples we mentioned above are of this form. We write ζ_g^o for the element of $\aleph_g$ corresponding to ξ_g^o. With U denoting the union of distinct elements of U_1 and U_2, we write $\zeta(U) = (\zeta_1(U_1), \zeta_2(U_2))$ and, with some abuse of notation, we also put $\psi(Z, \theta, \xi_1, \xi_2) = \psi(Z, \theta, \zeta_1(U_1), \zeta_2(U_2))$ and $\Psi(\theta, \zeta_1, \zeta_2) = \mathbb{E}(\psi(Z, \theta, \zeta_1(U_1), \zeta_2(U_2)))$. Note that the DR property of Ψ is clearly preserved with this notation since $\Psi(\theta, \zeta_1^o, \zeta_2) = 0$ and $\Psi(\theta, \zeta_1, \zeta_2^o) = 0$ if and only if $\theta = \theta^o$ for all functions $\zeta_1 \in \aleph_1$ and $\zeta_2 \in \aleph_2$. The main advantage of this notation is that in cases where the moment condition depends on a nuisance function in two conceptually different ways, say through both its level and its derivative at a certain point, we can accommodate this by using two different nonparametric estimators.

This flexibility is particularly useful for our Example 4 given above.

With this notation, we define the SDRE $\widehat{\theta}$ of θ^o as the value of θ which solves the equation

$$\Psi_n(\theta, \widehat{\zeta}) \equiv \frac{1}{n}\sum_{i=1}^{n} \psi(Z_i, \theta, \widehat{\zeta}(U_i)) = 0, \tag{2.3}$$

where $\widehat{\zeta} = (\widehat{\zeta}_1, \widehat{\zeta}_2)$ is a suitable nonparametric estimate of $\zeta^o = (\zeta_1^o, \zeta_2^o)$. We also define the following quantities, which will be important for estimating the asymptotic variance of $\widehat{\theta}$:

$$\widehat{H} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \psi(Z_i, \widehat{\theta}, \widehat{\zeta}(U_i))$$
$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \psi(Z_i, \widehat{\theta}, \widehat{\zeta}(U_i))\psi(Z_i, \widehat{\theta}, \widehat{\zeta}(U_i))^\top.$$

Here ∇_θ is the usual partial derivatives operator, and its use implicitly indicates that the corresponding partial derivatives are assumed to exist.

3. Asymptotic Properties under General Conditions

In this paper, we argue that SDREs have highly desirable properties relative to many standard semiparametric procedures even though they require the estimation of several nonparametric components.[1] To motivate why this should be the case, suppose that $\widehat{\theta}$ is consistent, and note that expanding equation (2.3) and solving for $\sqrt{n}(\widehat{\theta} - \theta^o)$ gives

$$\sqrt{n}(\widehat{\theta} - \theta^o) = -\left(\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \psi(Z_i, \theta^*, \widehat{\zeta}(U_i))\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi(Z_i, \theta^o, \widehat{\zeta}(U_i)), \tag{3.1}$$

for θ^* between $\widehat{\theta}$ and θ^o. Standard uniform convergence arguments and the consistency of $\widehat{\theta}$ and $\widehat{\zeta}$ imply that $\sum_{i=1}^n \nabla_\theta \psi(Z_i, \theta^*, \widehat{\zeta}(U_i))/n \xrightarrow{P} \mathbb{E}(\nabla_\theta \psi(Z, \theta^o, \zeta^o(U))) \equiv H$ for the Jacobian term in (3.1). Showing asymptotic normality of the score term $\sum_{i=1}^n \psi(Z_i, \theta^o, \widehat{\zeta}(U_i))/\sqrt{n}$ is more problematic. The arguments commonly used for this step in the literature require, among other things, certain restrictions on the accuracy of $\widehat{\zeta}$. In particular, one generally

[1] The DR property (2.2) implies that knowledge of either nuisance function suffices for identification of θ^o. In principle, one could thus always construct semiparametric estimators of θ^o that only requires estimating one of them, and holds the other fixed at some arbitrary known value.

needs that $\|\widehat{\zeta} - \zeta^o\|_\infty = o_P(n^{-1/4})$. If $\widehat{\zeta}$ is made up of kernel-based estimators it is also typically necessary that its bias is of the order $o(n^{-1/2})$. For other types of nonparametric estimators, such as orthogonal series, a less stringent "small bias" condition generally suffices (cf. Newey and McFadden, 1994; Newey, 1994).

Using a DR moment condition allows us to obtain an asymptotic normality result for the score term under relatively weak accuracy conditions. The result does not rely on the use of a particular nonparametric estimator in the first stage, but is a consequence of the geometry of the moment condition. To show this, we first to introduce the notion of a functional derivative of Ψ_n and Ψ in the nonparametric component. To that end, let $\lambda = (\lambda_1, \lambda_2)$ be a function such that $\zeta_o + t\lambda \in \aleph$ for all $t \in \mathbb{R}$ with $|t|$ sufficiently small, and define the derivative operators

$$\begin{aligned}
\Gamma_{1,n}^{(k)}(\lambda) &= \partial_t^k \Psi_n(\theta^o, \zeta_1^o + t\lambda_1, \zeta_2^o)|_{t=0}, \\
\Gamma_{2,n}^{(k)}(\lambda) &= \partial_t^k \Psi_n(\theta^o, \zeta_1^o, \zeta_2^o + t\lambda_2)|_{t=0}, \text{ and} \\
\Gamma_{12,n}(\lambda) &= \partial_{t_1,t_2} \Psi_n(\theta^o, \zeta_1^o + t_1\lambda_1, \zeta_2^o + t_2\lambda_2)|_{t_1=0,t_2=0}
\end{aligned}$$

for $k = 1, 2$. Note that in all three expressions Ψ_n is evaluated at the true parameter θ^o. With this notation, we can also define a second-order "Taylor approximation" of $\Psi_n(\theta^o, \widehat{\zeta}) - \Psi_n(\theta^o, \zeta^o)$ as follows:

$$\Psi_n^2(\widehat{\zeta} - \zeta^o) \equiv \sum_{k=1,2} \Gamma_{1,n}^{(k)}(\widehat{\zeta} - \zeta^o) + \sum_{k=1,2} \Gamma_{2,n}^{(k)}(\widehat{\zeta} - \zeta^o) + \Gamma_{12,n}(\widehat{\zeta} - \zeta^o).$$

Next, we introduce the population counterparts of the derivative operators given above:

$$\begin{aligned}
\Gamma_1^{(k)}(\lambda) &= \partial_t^k \Psi(\theta^o, \zeta_1^o + t\lambda_1, \zeta_2^o)|_{t=0}, \\
\Gamma_2^{(k)}(\lambda) &= \partial_t^k \Psi(\theta^o, \zeta_1^o, \zeta_2^o + t\lambda_2)|_{t=0}, \text{ and} \\
\Gamma_{12}(\lambda) &= \partial_{t_1,t_2} \Psi(\theta^o, \zeta_1^o + t_1\lambda_1, \zeta_2^o + t_2\lambda_2)|_{t_1=0,t_2=0};
\end{aligned}$$

and define a second-order "Taylor approximation" of $\Psi(\theta^o, \widehat{\zeta}) - \Psi(\theta^o, \zeta^o)$ by

$$\Psi^2(\widehat{\zeta} - \zeta^o) \equiv \sum_{k=1,2} \Gamma_1^{(k)}(\widehat{\zeta} - \zeta^o) + \sum_{k=1,2} \Gamma_2^{(k)}(\widehat{\zeta} - \zeta^o) + \Gamma_{12}(\widehat{\zeta} - \zeta^o).$$

Note that the "mixed partial derivative" Γ_{12} in this expansion can also be written as a

weighted integral of the pairwise products of estimation errors from estimating the various components of ζ_1^o and ζ_2^o, i.e.

$$\Gamma_{12}(\widehat{\zeta} - \zeta^o) = \sum_{s=1}^{t_1} \sum_{\bar{s}=1}^{t_2} \int \omega_{s\bar{s}}(u)(\widehat{\zeta}_{1s}(u_1) - \zeta_{1s}^o(u_1))(\widehat{\zeta}_{2\bar{s}}(u_2) - \zeta_{2\bar{s}}^o(u_2))dF_U(u_1, u_2), \tag{3.2}$$

where $\omega_{s\bar{s}}(u) = \mathbb{E}(\partial_{\zeta_{1s}^o(U_1)}\partial_{\zeta_{2\bar{s}}^o(U_2)}\psi(Z_i, \theta^o, \zeta^o(U))|U = u)$. Now the important point is that due to the DR property the functionals $\zeta_1 \mapsto \Psi(\theta^o, \zeta_1, \zeta_2^o)$ and $\zeta_2 \mapsto \Psi(\theta^o, \zeta_1^o, \zeta_2)$ are both constant, and thus the corresponding functional derivatives of Ψ are equal to zero:

$$\Gamma_g^{(k)}(\lambda) = 0 \text{ for } g = 1, 2 \text{ and } k = 1, 2, \ldots \tag{3.3}$$

This property implies that DR moment conditions are relatively insensitive with respect to variation in the nuisance functions, which is central for the following result.

Theorem 1. *Suppose that (i) the doubly robust moment function $\psi(z, \theta, \zeta(u))$ is three times continuously differentiable with respect to $\zeta(u)$, with derivatives that are uniformly bounded; (ii) $\|\widehat{\zeta} - \zeta^o\|_\infty^3 = o_P(n^{-1/2})$; (iii) $\Psi_n^2(\widehat{\zeta} - \zeta^o) - \Psi^2(\widehat{\zeta} - \zeta^o) = o_P(n^{-1/2})$; (iv) $\Gamma_{12}(\widehat{\zeta} - \zeta^o) = o_P(n^{-1/2})$. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Z_i, \theta^o, \widehat{\zeta}(U_i)) \xrightarrow{d} N(0, \Omega),$$

where $\Omega = Var(\psi(Z, \theta^o, \zeta^o(U)))$.

Proof. Assumption (i)–(ii) imply that $\Psi_n(\theta^o, \widehat{\zeta}) = \Psi_n(\theta^o, \zeta^o) + \Psi_n^2(\widehat{\zeta} - \zeta^o) + o_P(n^{-1/2})$, whereas (iii)–(iv) together with (3.3) imply that $\Psi_n^2(\widehat{\zeta} - \zeta^o) = o_P(n^{-1/2})$. The result then follows from the Central Limit Theorem. □

Theorem 1 is intentionally stated in a way similar to the famous Lemma 5.3 in Newey (1994). Its conclusion follows almost immediately from its "high-level" assumptions, and thus the important issue is how those assumptions can be verified. Assumption (i) is a convenient smoothness condition that is satisfied in all examples that we considered above. It ensures that the sample moment condition satisfies a quadratic expansion with well-behaved cubic remainder in its nonparametric component. This assumption could be weakened to allow for a non-smooth sample moment condition if an appropriate amount of smoothness is imposed on its population version instead (cf. Chen et al., 2003; Ichimura and Lee, 2010). Such an

approach would require additional technical arguments that are unrelated to the main point of this paper, and thus is not pursued here. Assumption (ii) weakens the usually required uniform rate of consistency for the nonparametric first stage from $o_P(n^{-1/4})$ to $o_P(n^{-1/6})$. Assumption (iii) is what is called a stochastic equicontinuity condition in the semiparametrics literature (e.g. Newey, 1994). This condition can typically either be verified by using general results from empirical process theory, or through direct calculations. For example, when $\widehat{\zeta}$ is a kernel-based estimator such a result follows from a project result for U-Statistics as in Newey and McFadden (1994); and with an orthogonal series estimator one can proceed as in Newey (1994).

Assumption (iv) is the most critical component of the result, and is not trivial when the rate of the nonparametric component is only $o_P(n^{-1/6})$. By equation (3.2), the term $\Gamma_{12}(\widehat{\zeta} - \zeta^o)$ is a weighted integral of the *pairwise products* of the nonparametric estimation errors of the respective components. The assumption is thus implied by a "small bias" and a "small covariance" condition on the bias terms and the purely stochastic parts of the estimators, respectively. Specifically, suppose that the first stage nonparametric estimators satisfy a uniform expansion of the form $\widehat{\zeta}_{gs}(u_g) = \zeta^o_{gs}(u_g) + b_{gs,n}(u_g) + v_{gs,n}(u_g) + o_P(n^{-1/2})$, where $b_{gs,n}(u_g)$ is a deterministic bias and $v_{gs,n}(u_g)$ is a mean zero stochastic term with finite second moments. Such an expansion can be established for many nonparametric estimators (e.g. Kong, Linton, and Xia, 2010). Then condition (iv) follows by Markov's inequality if $\sup_{(u_1,u_2)\in\mathcal{S}(U)} \|b_{1s,n}(u_1)b_{2\bar{s},n}(u_2)\| = o(n^{-1/2})$ and $\sup_{(u_1,u_2)\in\mathcal{S}(U)} \|\mathbb{E}(v_{1s,n}(u_1)v_{2\bar{s},n}(u_2))\| = o(n^{-1/2})$ for all $s = 1,\ldots,t_1$ and $\bar{s} = 1,\ldots,t_2$. In all the examples we considered above, the latter "small covariance" condition can be established by exploiting the particular structure of the nuisance functions. This is formally shown in Section 5. More generally, however, such a condition could always be achieved by randomly splitting the sample into separate parts, and then calculating each nonparametric estimator on only one of them, since such a construction obviously makes the estimators stochastically independent.

We complete this section by giving a formal result that establishes asymptotic normality of $\widehat{\theta}$ under certain standard conditions on the Jacobian term in (3.1).

Theorem 2. *Suppose $\widehat{\theta} \xrightarrow{p} \theta^o$, the conditions of Theorem 1 are satisfied, and (i) there exists $\alpha > 0$ and an open neighborhood $\mathcal{N}(\theta^o)$ of θ^o such that $\sup_{\theta\in\mathcal{N}(\theta^o)} \|\nabla_\theta\psi(Z,\theta,\zeta(U)) - \nabla_\theta\psi(Z,\theta,\zeta^o(U))\| \leq b(z)\|\zeta - \zeta^o\|^\alpha$; (iii) $H = \mathbb{E}(\nabla_\theta\psi(Z,\theta^o,\zeta^o(U)))$ has full rank. Then $\sqrt{n}(\widehat{\theta} - \theta^o) \xrightarrow{d} N(0, H^{-1}\Omega H^{-1})$, and $\widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1} \xrightarrow{P} H^{-1}\Omega H^{-1}$.*

Proof. Follows from standard arguments (e.g. Newey, 1994) □

In addition to asymptotic normality of $\widehat{\theta}$, Theorem 2 shows that the asymptotic variance of an SDRE is of the usual sandwich form, and establishes consistency of a simple sample analogue variance estimator. Taken together, these results can be used to justify various large sample inference procedures, such as e.g. the construction of confidence regions for θ^o. The theorem also shows that SDREs are adaptive, in the sense that their asymptotic variance does not contain an adjustment term for the use of first-step nonparametric estimates. This is a property SDREs share with all semiparametric estimators that take the form of a sample analogue of an influence function in the corresponding model (e.g. Newey, 1994). It also implies that SDREs are semiparametrically efficient if the DR moment condition is based on the respective *efficient* influence function. This is the case for several of the examples that we listed above.

4. Asymptotic Properties with First-Step Kernel Regressions

In this section, we show how the conditions of Theorem 1 can be verified under weak assumptions on the primitives of the model when using a specific nonparametric estimation procedure. Here we consider kernel-based smoothers, which are popular first-stage estimators for semiparametric two-step procedures. We also focus on a setting where the nuisance functions are two conditional expectations; that is

$$\zeta_g^o(x_g) = \mathbb{E}(Y_g|X_g = x_g) \text{ for } g \in \{1, 2\},$$

where $Y_g \in \mathbb{R}$, $X_g \in \mathbb{R}^{d_g}$ and (Y_1, Y_2, X_1, X_2) is a random subvector of Z that might have duplicate elements. Note that our Examples 1 and 2 are essentially of this structure. In our Examples 3 and 4, where one of the nuisance functions is a density, one can proceed similarly.

We consider estimating ζ_g^o by "leave-one-out" local polynomial regression of order l_g using bandwidth h_g. 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 Nadaraya-Watson estimator, for example. The local polynomial estimator is formally defined as follows. For a vector $b = (b_1, \ldots, b_{d_g})$, let $\mathcal{P}_{l_g,\alpha}(b) = \sum_{0 \leq |s| \leq l_g} \alpha_s b^s$ be a polynomial of order l_g. Here $\sum_{0 \leq |s| \leq l}$ denotes the summation over all

d_g-vectors s of positive integers with $0 \leq |s| \leq l_g$, and $\alpha = (\alpha_{(0,\ldots,0)}, \alpha_{(1,0,\ldots,0)}, \ldots, \alpha_{(0,\ldots,0,l_g)})$. Also let $\mathcal{K}$ be a density function on $\mathbb{R}$, put $K_{h_g}(b) = \prod_{j=1}^{d} \mathcal{K}(b_j/h_g)/h_g$, and define

$$\widehat{\alpha}_g^{-i}(b) = \underset{\alpha}{\operatorname{argmin}} \sum_{j \neq i} \left(Y_{gj} - \mathcal{P}_{l_g,\alpha}(X_{gj} - b)\right)^2 K_{h_g}(X_{gj} - b).$$

With this notation, the estimate $\widehat{\zeta}_g(U_{gi})$ of $\zeta_g^o(U_{gi})$ is given by

$$\widehat{\zeta}_g(U_{gi}) = \widehat{\alpha}_{g,(0,\ldots,0)}^{-i}(U_{gi}).$$

Under suitable regularity conditions (see e.g. Masry, 1996, or Appendix B) these estimates are uniformly consistent, and satisfy

$$\max_{i=1,\ldots,n} |\widehat{\zeta}_g(U_{gi}) - \zeta_g^o(U_{gi})| = O(h_g^{l_g+1}) + O_P((nh_g^{d_g}/\log n)^{-1/2}) \tag{4.1}$$

for $g \in \{1, 2\}$, where the two terms on the right-hand side of the previous equation correspond to the bias and stochastic part of the respective estimator. We also assume that underlying semiparametric model is such that

$$\mathbb{E}((Y_1 - \zeta_1^o(X_1)) \cdot (Y_2 - \zeta_2^o(X_2))|X) = 0, \tag{4.2}$$

where X denotes the union of distinct elements of X_1 and X_2. This is a non-trivial restriction which, together with the other assumptions on the primitives of the data generating process made below, will ensure that

$$\int \omega(u)(\widehat{\zeta}_1(u_1) - \zeta_1^o(u_1))(\widehat{\zeta}_2(u_2) - \zeta_2^o(u_2))dF_U(u_1, u_2) = o_P(n^{-1/2}) \tag{4.3}$$

for any smooth weighting function ω. We require this condition to verify assumption (iv) of Theorem 1. Note that equation (4.2) is satisfied in our Examples 1 and 2. In our Examples 3 and 4, where only one of the nuisance functions is a conditional expectation and the other is a density, it also follows naturally from the structure of the model that the corresponding analogue of (4.3) holds.[2]

[2]Strictly speaking, equation (4.2) is not necessary for deriving the results in this section. If this condition would not be satisfied, we could always construct asymptotically uncorrelated estimates $\widehat{\zeta}_1$ and $\widehat{\zeta}_2$ by splitting the data into two parts at random and calculating each estimator on a different one of them. Of course, sample splitting might not be attractive from a practical point of view in settings with samples of moderate

We introduce the following assumptions for our asymptotic analysis.

Assumption 1. *(i) The doubly robust moment function $\psi(z, \theta, \zeta(u))$ is three times continuously differentiable with respect to $\zeta(u)$, with derivatives that are uniformly bounded; (ii) $\widehat{\theta} \xrightarrow{p} \theta^o$; (iii) there exists $\alpha > 0$ and an open neighborhood $\mathcal{N}(\theta^o)$ of θ^o such that $\sup_{\theta \in \mathcal{N}(\theta^o)} \|\nabla_\theta \psi(Z, \theta, \zeta(U)) - \nabla_\theta \psi(Z, \theta, \zeta^o(U))\| \leq b(z)\|\zeta - \zeta^o\|^\alpha$; (iv) $H = \mathbb{E}(\nabla_\theta \psi(Z, \theta^o, \zeta^o(U)))$ has full rank.*

Assumption 1 restates the conditions of Theorem 1 and 2 that do not involve the nonparametric estimates of the nuisance functions.

Assumption 2. *(i) $\mathcal{K}$ is twice continuously differentiable; (ii) $\int \mathcal{K}(u)du = 1$; (iii) $\int u\mathcal{K}(u)du = 0$; (iv) $\int |u^2\mathcal{K}(u)|du < \infty$; and (v) $\mathcal{K}(u) = 0$ for u not contained in some compact set, say $[-1, 1]$.*

Assumption 2 describes a standard kernel function. The support restrictions on $\mathcal{K}$ could be weakened to allow for kernels with unbounded support at the expense of a more involved notation.

Assumption 3. *The following holds for $g \in \{1, 2\}$: (i) U_g is continuously distributed with compact support $\mathcal{S}(U_g)$; (ii) X_g is continuously distributed with support $\mathcal{S}(X_g) \supseteq \mathcal{S}(U_g)$; (iii) the corresponding density functions are bounded, have bounded first order derivatives, and are bounded away from zero uniformly over $\mathcal{S}(U_g)$; (iv) the function ξ_g^o is $(l_g + 1)$ times continuously differentiable; (v) $\sup_{u \in \mathcal{S}(U_g)} \mathbb{E}(|Y_g|^c | X_g = u) < \infty$ for some constant $c > 2$*

This assumption collects a number smoothness and regularity conditions that are standard in the context of nonparametric regression.

Assumption 4. *(i) $nh_1^{2(l_1+1)}h_2^{2(l_2+1)} \to 0$; (ii) $nh_1^{6(l_1+1)} \to 0$; (iii) $nh_2^{6(l_2+1)} \to 0$; (iv) $n^2 h_1^{3d_1} / \log(n)^3 \to \infty$; and (v) $n^2 h_2^{3d_2} / \log(n)^3 \to \infty$.*

Assumption 4 imposes restrictions on the rate at which the bandwidths h_1 and h_2 tend to zero that depend on the number of derivatives of the unknown regression functions and the dimension of the corresponding covariates. These restrictions are substantially weaker

size. That said, the idea has been found useful in other contexts for applied economic research (e.g. Angrist and Krueger, 1995; Card, Mas, and Rothstein, 2008).

than those commonly found in the literature on semiparametric two-step estimation; see the discussion below.

Using these assumptions, we can now prove the asymptotic normality of $\widehat{\theta}$ by verifying the conditions of Theorem 1 and Theorem 2.

Theorem 3. *Suppose that Assumption 1–4 hold. Then*

$$\sqrt{n}(\widehat{\theta} - \theta^o) \xrightarrow{d} N(0, H^{-1}\Omega H^{-1}) \text{ and } \widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1} \xrightarrow{P} H^{-1}\Omega H^{-1}.$$

Proof. See the Appendix. □

Theorem 3 differs from other asymptotic normality results for semiparametric two-step estimators (e.g. Newey, 1994; Newey and McFadden, 1994; Chen et al., 2003; Ichimura and Lee, 2010), because it only imposes relatively weak conditions on the accuracy of the nonparametric first stage estimates. The bandwidth restrictions in Assumption 4 allow the smoothing bias from estimating either ζ_1^o or ζ_2^o to be of the order $o(n^{-1/6})$ as long as the *product* of the two bias terms is of the order $o(n^{-1/2})$. They also only require the respective stochastic parts to be of the order $o_P(n^{-1/6})$; see (4.1). These conditions should be contrasted with those needed for generic kernel-based semiparametric two-step estimator. By this we mean estimators for which the conditions from Newey and McFadden (1994) that the first stage nonparametric estimation error and bias are of the order $o_P(n^{-1/4})$ and $o(n^{-1/2})$, respectively, are necessary.

The weaker accuracy conditions on first-stage estimates mean that SDREs can be asymptotically normal under less stringent smoothness conditions on the nuisance functions. For example, it is easily verified that if $d_1 \leq 5$ and $d_2 \leq 5$, there exist bandwidths h_1 and h_2 such that Assumption 4 is satisfied even if $l_1 = l_2 = 1$. For a generic kernel-based semiparametric two-step estimator that uses an estimate of, say, ζ_1^o to be asymptotically normal one typically cannot allow for $l_1 = 1$ if $d_1 > 1$. Less stringent smoothness conditions allow for the use of lower order local polynomials in the first stage. This is very important in empirical practice: while higher order local polynomial regression leads to estimates with small asymptotic bias, it is also well-known to have poor finite sample properties (Seifert and Gasser, 1996).

In lower dimensional settings the range of bandwidths that is permitted by Assumption 4 includes the values that minimize the Integrated Mean Squared Error (IMSE) for estimating ζ_1^o and ζ_2^o, respectively. This is generally not the case for a generic kernel-based

semiparametric two-step estimator. While these bandwidths do not have any optimality properties for estimating θ^o, they have the practical advantage that they can be estimated from the data via least-squares cross validation. For many SDREs, there thus exist a simple data-driven bandwidth selection method that does not rely on preliminary estimates of the nonparametric component.[3]

Given our specific first-stage estimator, we can actually improve the result in Theorem 3 and derive a more explicit characterization of the higher-order properties of $\widehat{\theta}$. To simplify the exposition, we only state such a result for the case that the arguments of ζ_1^o and ζ_2^o have the same dimension, that is $d_1 = d_2 \equiv d$, and that the same bandwidth and order of the local polynomial are used to estimate these two functions, that is $l_1 = l_2 \equiv l$ and $h_1 = h_2 \equiv h$. Similar results could be established in more general settings.

Theorem 4. *Suppose Assumption 1–4 hold, and assume for notational simplicity that $d_1 = d_2 \equiv d$, $h_1 = h_2 \equiv h$ and $l_1 = l_2 \equiv l$. Then*

$$\widehat{\theta} - \theta^o = \frac{1}{n}\sum_{i=1}^{n} H^{-1}\psi(Z_i, \theta^o, \zeta^o(U_i)) + R_n + o_P(R_n), \tag{4.4}$$

for some random sequence R_n such that (i) $R_n = O_P(h^{2(l+1)}) + O_P(n^{-1}h^{-d/2})$, and (ii) $\mathbb{E}(R_n) = O(h^{2(l+1)})$ and $Var(R_n) = O(n^{-2}h^{-d})$.

Proof. See the Appendix. □

Theorem 4(i) shows that the difference between our SDRE and its linear representation is of the order $O_P(h^{2(l+1)}) + O_P(n^{-1}h^{-d/2})$. The magnitudes of these two terms correspond to those of the squared bias and $h^{-d/2}$ times the (pointwise) variance of $\widehat{\zeta}$, respectively. The bandwidth h^* that minimizes the sum of these two terms satisfies $h^* \propto n^{-2/(4(l+1)+d)}$, and with this choice of bandwidth equation (4.4) holds with $R_n = O_P(n^{-4(l+1)/(4(l+1)+d)})$. This is a substantial improvement over generic kernel-based semiparametric two-step estimators, whose linear representation is generally only accurate up to terms of order $O_P(h^{l+1}) + O_P(n^{-1}h^{-d})$. See Ichimura and Linton (2005) or Cattaneo et al. (2013) for such calculations. A result analogous to (4.4) could thus at best be obtained with $R_n = O_P(n^{-(l+1)/(l+1+d)})$ for those

[3]Note that while our result formally do not allow for data-dependent bandwidths, such an extension could be obtained using arguments similar to those in e.g. Einmahl and Mason (2005). See also Escanciano, Jacho-Chávez, and Lewbel (2014).

estimators. For the simple case with $d = l = 1$, for example, a generic semiparametric two-step estimator would thus differs from its linear representation by a term that is at least of the order $O_P(n^{-2/3})$, whereas for our SDREs the difference can be of an order as small as $O_P(n^{-8/9})$. As a consequence, we can expect standard Gaussian approximations based on linear representations like (4.4) to be more accurate in finite samples for our SDREs.

Theorem 4(ii) gives the first two moments of the largest second-order terms in the expansion (4.4). It is common practice to use such quantities to approximate the first-order bias and second-order variance of a semiparametric two-step estimator (Linton, 1995). Theorem 4(ii) thus implies that the asymptotic mean squared error (MSE) of $\widehat{\theta}$ satisfies

$$MSE(\widehat{\theta}) \approx n^{-1}H^{-1}\Omega H^{-1} + O(n^{-2}h^{-d}) + O(h^{4(l+1)}),$$

which is minimized by choosing a bandwidth h^* with $h^* \propto n^{-2/(4(l+1)+d)}$. In this case, the second-order component of the asymptotic MSE is of the order $O(n^{-4(l+1)/(4(l+1)+d)})$. For the simple case that $d = l = 1$, this term is of the order $O(n^{-16/9})$ for example. On the other hand, in general both leading terms in an analogous expansion of a generic semiparametric two-step estimator have non-zero means (Ichimura and Linton, 2005; Cattaneo et al., 2013), and thus there is no tradeoff between first-order bias and second-order variance in this case. For such an estimator the second-order component of the asymptotic mean squared error is generally only of the order $O(h^{2(l+1)} + n^{-1}h^{-d})$. For the simple case that $d = l = 1$, this term is of the order $O(n^{-4/3})$ at best. This shows that using an SDRE should lead to a sizable reduction in asymptotic mean squared error relative to a generic estimator with the same asymptotic variance.

5. Applications

In this section, we revisit the examples introduced in Section 2 above; and study the properties of the SDRE of the respective parameter of interest. The assumptions and conclusions are very similar in structure to our generic results in Section 3, and thus we keep the discussion relatively brief.

5.1. Missing Data. In this subsection, we study an SDRE of the parameter θ^o in the Missing Data model described in Example 1. Let $\zeta_1^o(x,\theta) = \mathbb{E}(m(Y,X,\theta)|D = 1, X = x)$

and $\zeta_2^o(x) = \mathbb{E}(D|X = x)$, and recall that in this setting

$$\psi_{MD}(Z, \theta, \zeta(X)) = \frac{D(m(Y, X, \theta) - \zeta_1(X, \theta))}{\zeta_2(X)} + \zeta_1(X, \theta) - \theta$$

is a DR moment function for estimating θ^o. Also write $\psi_{MD}^o(Z) = \psi_{MD}(Z, \theta^o, \zeta^o(X))$. The data consist of an i.i.d. sample $\{Z_i\}_{i=1}^n = \{(Y_i, X_i, D_i)\}_{i=1}^n$ from the distribution of $Z = (Y, X, D)$. We estimate the two nuisance functions ζ_1^o and ζ_2^o by "leave-one-out" local polynomial regression, using the same order of the local polynomial l and bandwidth h in both cases to simplify the exposition. That is, using notation analogous to that introduced in Section 3, we define

$$\begin{aligned}
\widehat{\alpha}^{-i}(x, \theta) &= \underset{\alpha}{\operatorname{argmin}} \sum_{j \neq i} \left(m(Y_j, X_j, \theta) - \mathcal{P}_{l,\alpha}(X_j - x)\right)^2 K_h(X_j - x) \mathbb{I}\{D_j = 1\}, \\
\widehat{\beta}^{-i}(x) &= \underset{\beta}{\operatorname{argmin}} \sum_{j \neq i} \left(D_j - \mathcal{P}_{l,\beta}(X_j - x)\right)^2 K_h(X_j - x),
\end{aligned}$$

and put

$$\widehat{\zeta}_1(X_i, \theta) = \widehat{\alpha}^{-i}_{(0,\ldots,0)}(X_i, \theta) \quad \text{and} \quad \widehat{\zeta}_2(X_i) = \beta^{-i}_{(0,\ldots,0)}(X_i).$$

The estimator $\widehat{\theta}$ is then defined as the value of θ that solves the following equation:

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{D_i(m(Y_i, X_i, \theta) - \widehat{\zeta}_1(X_i, \theta))}{\widehat{\zeta}_2(X_i)} + \widehat{\zeta}_1(X_i, \theta) \right) = 0.$$

We derive the properties of this estimator under the following assumptions.

Assumption MD 1. (i) $\mathbb{E}(m(Y^*, X, \theta_o)) = 0$ and $\mathbb{E}(m(Y^*, X, \theta)) \neq 0$ for all $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$, with Θ a compact set and $\theta_o \in \text{int}(\Theta)$, (ii) there exists a non-negative function b such that $|m(Y^*, X, \theta)| < b(Y^*, X)$ with probability 1 for all $\theta \in \Theta$, and $\mathbb{E}(b(Y^*, X)) < \infty$, (iii) $m(Y^*, X, \theta)$ is continuous on Θ and continuously differentiable in an open neighborhood of θ_o, (iv) $\mathbb{E}(\|m(Y^*, X, \theta_o)\|^2) < \infty$ and, (v) $\sup_{\theta \in \Theta} \mathbb{E}(\|\partial_\theta m(Y^*, X, \theta)\|) < \infty$.

Assumption MD 2. (i) X is continuously distributed both unconditionally and conditional on $D = 1$, with compact support $\mathcal{S}(X)$ and $\mathcal{S}(X|D = 1)$, respectively; (ii) the corresponding density functions are bounded, have bounded first order derivatives, and are bounded away from zero, uniformly over $\mathcal{S}(X)$ and $\mathcal{S}(X|D = 1)$, respectively; (iii) $\zeta_2^o(x)$ is $(l + 1)$-times continuously differentiable; (iv) $\zeta_1^o(x, \theta)$ is $(l + 1)$-times continuously differentiable in x for

all $\theta \in \Theta$, and $\sup_{x \in \mathcal{S}(X|D=1)} \mathbb{E}(\|m(Y,X,\theta_o)\|^c | D=1, X=x) < \infty$ for some constant $c > 2$.

Assumption MD 3. $nh^{4(l+1)} \to 0$ and $n^2 h^{3d}/\log(n)^3 \to \infty$ as $n \to \infty$.

Assumption MD 1 is a set of regularity condition that ensures that a standard Method-of-Moments estimator of θ^o would be $\sqrt{n}$-consistent and asymptotically normal in the absence of missing data. Assumption MD 2–MD 3 are similar to Assumption 3–4 above. Under these conditions, we obtain the following proposition.

Proposition 1. *Suppose that Assumption 2 and Assumption MD 1–MD 3 hold. Then*

$$\widehat{\theta} \xrightarrow{P} \theta^o \text{ and } \sqrt{n}(\widehat{\theta} - \theta^o) \xrightarrow{d} N(0, H^{-1}\mathbb{E}(\psi^o_{MD}(Z)\psi^o_{MD}(Z)^\top)H^{-1}).$$

Moreover, if $h \propto n^{-2/(4(l+1)+d)}$, *then*

$$\widehat{\theta} - \theta^o = \frac{1}{n}\sum_{i=1}^{n} H^{-1}\psi^o_{MD}(Z_i) + R_n + o_P(R_n)$$

for a random sequence R_n *such that* $R_n = O_P(n^{-4(l+1)/(4(l+1)+d)})$ *and* $\mathbb{E}(R_n^2) = O(n^{-8(l+1)/(4(l+1)+d)})$, *where* $H = \mathbb{E}(\nabla_\theta m(Y^*, X, \theta_o))$.

Proof. See the Appendix. □

The statement of the proposition is essentially analogous to that of Theorems 3–4 above. Note that the asymptotic variance of $\widehat{\theta}$ coincides with the semiparametric efficiency bound for estimating θ^o in this model, which was derived by Robins et al. (1994); see also Hahn (1998) and Chen, Hong, and Tarozzi (2008) for related results. Being an efficient estimator with desirable second order properties, the SDRE has clear advantages relative to other efficient estimators that are commonly used in such settings, such as Inverse Probability Weighting (IPW) estimators (e.g. Hirano et al., 2003; Firpo, 2007), whose kernel-based version is an example of an estimator for which Newey and McFadden's sufficient conditions are also necessary (Ichimura and Linton, 2005). Note that in a closely related model Cattaneo (2010) proposed an estimator that has the same structure as our SDRE, but did not formally show favorable properties of this approach relative to other estimators.

Remark 1. Using a linear smoother to estimate the propensity score can be undesirable in practice, as the estimates are not constrained to be between zero and one. One way to

address this problem is to use a local polynomial Probit estimator instead. That is, one could redefine

$$\widehat{\beta}^{-i}(x) = \underset{\beta}{\operatorname{argmin}} \sum_{j \neq i} \left(D_j - \Phi(\mathcal{P}_{l,\beta}(X_j - x))\right)^2 K_h(X_j - x),$$

where Φ is the CDF of the standard normal distribution, and then put $\widehat{\zeta}_2(X_i) = \Phi(\beta^{-i}_{(0,\ldots,0)}(X_i))$. This change would not affect the result of our asymptotic analysis, as it is well known from the work of e.g. Fan, Heckman, and Wand (1995), Hall, Wolff, and Yao (1999) or Gozalo and Linton (2000) that the asymptotic bias of the local polynomial Probit estimator is of the same order of magnitude as that of the usual local polynomial estimator uniformly over the covariates' support, and that the two estimators have the same stochastic behavior.

5.2. Partial Linear Model. In this subsection, we study an SDRE of the parameter θ^o in the Partial Linear Model described in Example 2. Let $\zeta_1^o(x,\theta) = \mathbb{E}(Y - W^\top\theta | X = x)$ and $\zeta_2^o(x) = \mathbb{E}(W|X = x)$; and recall that in this setting

$$\psi_{PLM}(Z, \theta, \zeta(X)) = (Y - W^\top\theta - \zeta_1(X,\theta))(W - \zeta_2(X)) - \theta$$

is a DR moment function for estimating θ^o. We also write $\psi^o_{PLM}(Z) = \psi_{PLM}(Z, \theta^o, \zeta^o(X))$. The data consist of an i.i.d. sample $\{Z_i\}_{i=1}^n = \{(Y_i, X_i, W_i)\}_{i=1}^n$ from the distribution of $Z = (Y, X, W)$. We estimate the two nuisance functions ζ_1^o and ζ_2^o by "leave-one-out" local polynomial regression, using the same order of the local polynomial l and bandwidth h in both cases for notational simplicity. Since local polynomial regression is a linear smoothing procedure, these estimates can be obtained by first defining

$$\widehat{\alpha}^{-i}(x) = \underset{\alpha}{\operatorname{argmin}} \sum_{j \neq i} \left(Y_j - \mathcal{P}_{l,\alpha}(X_j - x)\right)^2 K_h(X_j - x),$$
$$\widehat{\beta}^{-i}(x) = \underset{\beta}{\operatorname{argmin}} \sum_{j \neq i} \left(W_j - \mathcal{P}_{l,\beta}(X_j - x)\right)^2 K_h(X_j - x),$$

and then putting

$$\widehat{\zeta}_1(X_i, \theta) = \widehat{\alpha}^{-i}_{(0,\ldots,0)}(X_i) - \widehat{\beta}^{-i}_{(0,\ldots,0)}(X_i)^\top\theta \quad \text{and} \quad \widehat{\zeta}_2 = \widehat{\beta}^{-i}_{(0,\ldots,0)}(X_i).$$

The estimator $\widehat{\theta}$ is then defined as the value of θ that solves the following equation:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - W_i^\top\theta - \widehat{\zeta}_1(X_i,\theta))(W_i - \widehat{\zeta}_2(X_i)) = 0.$$

It is easily seen that an explicit expression for $\widehat{\theta}$ is given by

$$\widehat{\theta} = \left(\sum_{i=1}^{n}(W_i - \widehat{\zeta}_2(X_i))(W_i - \widehat{\zeta}_2(X_i))^\top\right)^{-1}\sum_{i=1}^{n}(W_i - \widehat{\zeta}_2(X_i))(Y_i - \widehat{\alpha}^{-i}_{(0,\ldots,0)}(X_i)).$$

Note that $\widehat{\theta}$ is identical (up to trimming terms) to the estimator proposed by Robinson (1988). We study its theoretical properties under the following assumptions.

Assumption PLM 1. $H = \mathbb{E}((W - \zeta_2^o(X))(W - \zeta_2^o(X))^\top)$ is positive definite.

Assumption PLM 2. (i) X is continuously distributed with compact support $\mathcal{S}(X)$, and the corresponding density function is bounded, has bounded first order derivatives, and is bounded away from zero uniformly over $\mathcal{S}(Z)$; (ii) the functions $\phi^o(x)$ and $\zeta_2^o(x)$ are both $(l+1)$-times continuously differentiable; (iii) $\sup_{x\in\mathcal{S}(X)}\mathbb{E}(|W|^c|X = x) < \infty$ and $\sup_{x\in\mathcal{S}(X)}\mathbb{E}(|\varepsilon|^c|X = x) < \infty$ for some constant $c > 2$.

Assumption PLM 3. $nh^{4(l+1)} \to 0$ and $n^2h^{3d}/\log(n)^3 \to \infty$ as $n \to \infty$.

Assumption PLM 1 is a full rank condition that ensures point identification of θ^o. Assumption PLM 2–PLM 3 are similar to Assumption 3–4 above. Under these conditions, we obtain the following proposition.

Proposition 2. *Suppose that Assumption 2 and Assumption PLM 1–PLM 3 hold. Then*

$$\widehat{\theta} \xrightarrow{P} \theta^o \text{ and } \sqrt{n}(\widehat{\theta} - \theta^o) \xrightarrow{d} N(0, H^{-1}\mathbb{E}(\psi^o_{PLM}(Z)\psi^o_{PLM}(Z)^\top)H^{-1}).$$

Moreover, if $h \propto n^{-2/(4(l+1)+d)}$, *then*

$$\widehat{\theta} - \theta^o = \frac{1}{n}\sum_{i=1}^{n} H^{-1}\psi^o_{PLM}(Z_i) + R_n + o_P(R_n)$$

for a random sequence R_n *such that* $R_n = O_P(n^{-4(l+1)/(4(l+1)+d)})$ *and* $\mathbb{E}(R_n^2) = O(n^{-8(l+1)/(4(l+1)+d)})$.

Proof. See the Appendix. □

The statement of the proposition is essentially analogous to that of Theorems 3–4 above. Note that this result is similar to those obtained by Linton (1995) and Li (1996), who also studied the higher-order properties of the Robinson estimator. By establishing a connection to DR moment conditions, we show that the desirable properties of this estimator are not a coincidence, but follow from the fact that it is an example of an SDRE.

5.3. Policy Effects. In this subsection, we study an SDRE of the Policy Effect parameter θ^o described in Example 3. Let $\zeta_{11}^o(x) = \mathbb{E}(Y|X = x)$ and $\zeta_{12}^o(x) = \mathbb{E}(Y|X = \pi(x))$, and denote the densities of X and $\pi(X)$ by $\zeta_{21}^o(x)$ and $\zeta_{22}^o(x)$, respectively. With this notation, we have that

$$\psi_{PE}(Z, \theta, \zeta(X)) = \zeta_{12}(X) + (Y_i - \zeta_{11}(X))\frac{\zeta_{22}(X)}{\zeta_{21}(X)} - \theta$$

is a DR moment function for estimating θ^o. We also write $\psi_{PE}^o(Z) = \psi_{PE}(Z, \theta^o, \zeta^o(X))$. The data consist of an i.i.d. sample $\{Z_i\}_{i=1}^n = \{(Y_i, X_i)\}_{i=1}^n$ from the distribution of $Z = (Y, X)$. We again estimate the two components of ζ_1^o by a "leave-one-out" local polynomial regression of order l with bandwidth h. That is, we define

$$\widehat{\alpha}^{-i}(x) = \underset{\alpha}{\operatorname{argmin}} \sum_{j \neq i} \left(Y_j - \mathcal{P}_{l,\alpha}(X_j - x)\right)^2 K_h(X_j - x),$$

and put

$$\widehat{\zeta}_{11}(X_i) = \widehat{\alpha}^{-i}_{(0,\ldots,0)}(X_i) \quad \text{and} \quad \widehat{\zeta}_{12}(X_i) = \widehat{\alpha}^{-i}_{(0,\ldots,0)}(\pi(X_i)).$$

To estimate the two components of ζ_2^o, we use standard "leave-one-out" kernel density estimators with bandwidth h, allowing a kernel function of order $l + 1$ for the purpose of bias control. That is, with $\mathcal{K}^*$ a symmetric function on $\mathbb{R}$ whose exact properties are stated below, and $K_h^*(b) = \prod_{j=1}^d \mathcal{K}^*(b_j/h)/h$, we define

$$\widehat{\zeta}_{21}(X_i) = \frac{1}{n}\sum_{j \neq i} K_h^*(X_j - X_i) \quad \text{and} \quad \widehat{\zeta}_{22}(X_i) = \frac{1}{n}\sum_{j \neq i} K_h^*(\pi(X_j) - X_i).$$

With this notation, our estimator of θ^o is then given by:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\zeta}_{12}(X_i) + (Y_i - \widehat{\zeta}_{11}(X_i))\frac{\widehat{\zeta}_{22}(X_i)}{\widehat{\zeta}_{21}(X_i)}\right).$$

We study the theoretical properties of this estimator under the following assumptions.

Assumption PE 1. (i) $\mathcal{K}^*$ is twice continuously differentiable; (ii) $\int \mathcal{K}^*(u)du = 1$; (iii) $\int u^k \mathcal{K}^*(u)du = 0$ for $k = 1, \ldots, l+1$; (iv) $\int |u^2\mathcal{K}^*(u)|du < \infty$; and (v) $\mathcal{K}^*(u) = 0$ for u not contained in some compact set, say $[-1, 1]$.

Assumption PE 2. (i) X and $\pi(X)$ are continuously distributed with compact support $\mathcal{S}(X)$ and $\mathcal{S}(\pi(X)) \subset \mathcal{S}(X)$, respectively; (ii) the corresponding density functions $\zeta^o_{21}(x)$ and $\zeta^o_{22}(x)$ are bounded, $l+1$ times continuously differentiable, and bounded away from zero uniformly over $\mathcal{S}(X)$ and $\mathcal{S}(\pi(X))$, respectively; (iii) the function $\zeta^o_{11}(x)$ is $l+1$ times continuously differentiable, and $\sup_{x\in\mathcal{S}(\pi(X))}\mathbb{E}(|Y|^c|X = x) < \infty$ for some constant $c > 2$.

Assumption PE 3. $nh^{4(l+1)} \to 0$ and $n^2h^{3d}/\log(n)^3 \to \infty$ as $n \to \infty$.

Assumption PE 1 describes a kernel function of order $l+1$, which is used to control the asymptotic bias of the two density estimates. Note that such a kernel function must take on negative values. Assumption PE 2–PE 3 are again similar to Assumption 3–4 above. Under these conditions, we obtain the following proposition.

Proposition 3. *Suppose that Assumption 2 and Assumption PE 1–PE 3 hold. Then*

$$\widehat{\theta} \xrightarrow{P} \theta^o \text{ and } \sqrt{n}(\widehat{\theta} - \theta^o) \xrightarrow{d} N(0, \mathbb{E}(\psi^o_{PE}(Z)^2)).$$

Moreover, if $h \propto n^{-2/(4(l+1)+d)}$, *then*

$$\widehat{\theta} - \theta^o = \frac{1}{n}\sum_{i=1}^{n}\psi^o_{PE}(Z_i) + R_n + o_P(R_n)$$

for a random sequence R_n *such that* $R_n = O_P(n^{-4(l+1)/(4(l+1)+d)})$ *and* $\mathbb{E}(R_n^2) = O(n^{-8(l+1)/(4(l+1)+d)})$.

Proof. See the Appendix. □

The statement of the proposition is again essentially analogous to that of Theorems 3–4 above even though only one of the nuisance functions is a conditional expectation. This is

because the ratio $\widehat{\zeta}_{22}(X_i)/\widehat{\zeta}_{21}(X_i)$ satisfies a stochastic expansion that has the same structure as that of $\widehat{\zeta}_{11}(X_i)$; and thus the proof follows from the same type of arguments.

5.4. Weighted Average Derivatives. In this subsection, we study an SDRE of the Weighted Average Derivative θ^o described in Example 4. Let $\zeta_{11}^o(x) = \mathbb{E}(Y|X=x)$, denote the density of X by $\zeta_{21}^o(x)$, and define the vectors of partial derivatives of those two functions as $\zeta_{12}^o(x) \equiv \nabla_x \zeta_{11}^o(x)$ and $\zeta_{22}^o(x) \equiv \nabla_x \zeta_{21}^o(x)$, respectively. With this notation, we have that

$$\psi_{WAD}(Z,\theta,\zeta(X)) = w(X)\zeta_{12}(X) - (Y-\zeta_{11}(X))\left(\nabla_x w(X) + w(X)\frac{\zeta_{22}(X)}{\zeta_{21}(X)}\right) - \theta$$

is a DR moment function for estimating θ^o. We also write $\psi_{WAD}^o(Z) = \psi_{WAD}(Z,\theta^o,\zeta^o(X))$. The data consist of an i.i.d. sample $\{Z_i\}_{i=1}^n = \{(Y_i,X_i)\}_{i=1}^n$ from the distribution of $Z = (Y,X)$. Note that the need to estimate derivatives distinguishes this application from the other ones considered in this section. We address this issue by using standard derivative estimators of conditional expectations and densities. Moreover, we allow using different smoothing parameters for estimating levels and derivatives of a function. Specifically, we estimate ζ_{11}^o and ζ_{12}^o by "leave-one-out" local polynomial regression of order l_1 and l_2 with bandwidth h_1 and h_2, respectively. That is, we define

$$\widehat{\alpha}_g^{-i}(x) = \underset{\alpha}{\operatorname{argmin}} \sum_{j\neq i} \left(Y_j - \mathcal{P}_{l_g,\alpha}(X_j - x)\right)^2 K_{h_g}(X_j - x).$$

for $g = 1, 2$, and put

$$\widehat{\zeta}_{11}(X_i) = \widehat{\alpha}_{1,(0,\ldots,0)}^{-i}(X_i) \quad \text{and} \quad \widehat{\zeta}_{12}(X_i) = \left(\widehat{\alpha}_{2,(1,0,\ldots,0)}^{-i}(X_i), \ldots, \widehat{\alpha}_{2,(0,\ldots,0,1)}^{-i}(X_i)\right)^\top.$$

Note that ζ_{12}^o is estimated by the slope coefficients of a local polynomial approximation, whereas ζ_{11}^o is estimated by the intercept as usual.

To estimate the density function ζ_{21}^o, we proceed as we did in the Policy Effects example, and use a "leave-one-out" kernel density estimator with bandwidth h_1 and a kernel of order $l_1 + 1$. A natural approach to estimate the density derivative ζ_{22}^o is to take the derivative of a "leave-one-out" kernel density estimator with bandwidth h_2 and a kernel of order $l_2 + 1$.

That is, we put

$$\widehat{\zeta}_{21}(X_i) = \frac{1}{n}\sum_{j\neq i} K^*_{h_1}(X_j - X_i) \quad \text{and} \quad \widehat{\zeta}_{22}(X_i) = \frac{1}{n}\sum_{j\neq i} \nabla_x \, K^*_{h_2}(\pi(X_j) - x)\big|_{x=X_i}.$$

Note that we use same l_s and h_s to estimate ζ^o_{1s} and ζ^o_{2s} for $s \in \{1,2\}$ for notational simplicity only. With this notation, our estimator of θ_o is given by:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left(w(X_i)\widehat{\zeta}_{12}(X_i) - (Y_i - \widehat{\zeta}_{11}(X_i))\left(\nabla_x w(X_i) + w(X_i)\frac{\widehat{\zeta}_{22}(X_i)}{\widehat{\zeta}_{21}(X_i)}\right)\right).$$

We study the theoretical properties of this estimator under the following assumptions.

Assumption WAD 1. (i) $\mathcal{K}^*$ is twice continuously differentiable; (ii) $\int \mathcal{K}^*(u)du = 1$; (iii) $\int u^k \mathcal{K}^*(u)du = 0$ for $k = 1, \ldots, l_2 + 1$; (iv) $\int |u^2 \mathcal{K}^*(u)|du < \infty$; and (v) $\mathcal{K}^*(u) = 0$ for u not contained in some compact set, say $[-1, 1]$.

Assumption WAD 2. (i) w is bounded and has bounded and continuous first order derivatives, and $\mathcal{S}(w) \equiv \{x \in \mathbb{R}^d : w(x) > 0\}$ is a compact set; (ii) X is continuously distributed, and the corresponding density function is bounded, has bounded and continuous derivatives up to order $\max\{l_1, l_2\} + 1$, and is bounded away from zero uniformly over $\mathcal{S}(w)$; (iii) $\zeta^o_{11}(x)$ has bounded and continuous derivatives up to order $\max\{l_1, l_2\} + 1$, and $\sup_{x\in\mathcal{S}(w)} \mathbb{E}(|Y|^c|X = x) < \infty$ for some constant $c > 2$.

Assumption WAD 3. (i) $nh_1^{2(l_1+1)}h_2^{2(l_2+1)} \to 0$; (ii) $nh_1^{6(l_1+1)} \to 0$; (iii) $nh_2^{6(l_2+1)} \to 0$; (iv) $n^2 h_1^{3d}/\log(n)^3 \to \infty$; and (v) $n^2 h_2^{6d+12}/\log(n)^3 \to \infty$.

Assumption WAD 1 describes a kernel function of order $l_2 + 1$, Assumption WAD 2 is a standard smoothness condition similar to Assumption 3. To understand Assumption WAD 3, note that these restrictions on the bandwidths ensure that $\max_{i=1,\ldots,n} |\widehat{\zeta}_{gj}(X_i) - \zeta^o_{gj}(X_i)| = o_p(n^{-1/6})$ for $g \in \{1,2\}$ and $j \in \{1,2\}$; and that the product of the smoothing biases from estimating ζ^o_{gj} and $\zeta^o_{g'j'}$ is $o(n^{-1/2})$ for $g \neq g'$.

Proposition 4. *Suppose that Assumption 2 and Assumption WAD 1–WAD 3 hold. Then*

$$\widehat{\theta} \xrightarrow{P} \theta^o \text{ and } \sqrt{n}(\widehat{\theta} - \theta^o) \xrightarrow{d} N(0, \mathbb{E}(\psi^o_{WAD}(Z)\psi^o_{WAD}(Z)^\top)).$$

Moreover, if $l_1 = l_2 \equiv l$, $h_1 \propto h_2 \equiv h$ *and* $h \propto n^{-2/(4l+d+8)}$, *then*

$$\widehat{\theta} - \theta^o = \frac{1}{n}\sum_{i=1}^{n} \psi^o_{WAD}(Z_i) + R_n + o_P(R_n)$$

for a random sequence R_n *such that* $R_n = O_P(n^{-4(l+1)/(4l+d+8)})$ *and* $\mathbb{E}(R_n^2) = O(n^{-8(l+1)/(4l+d+8)})$.

Proof. See the Appendix. □

Note that the asymptotic variance of $\widehat{\theta}$ coincides with the semiparametric efficiency bound for estimating θ^o in this model, which was derived by Newey and Stoker (1993). Even though the estimator $\widehat{\theta}$ involves the estimation of four unknown functions, the proposition shows that it has attractive properties relative to other efficient estimators proposed in the literature. For example, the bandwidth restrictions in Assumption WAD 3 are weaker than those needed for the various efficient estimators discussed in Stoker (1991), and also weaker than those use by Cattaneo et al. (2013) for their jackknife bias corrected estimator.

6. Concluding Remarks

In this paper, we have explored the possibility of constructing semiparametric two-step estimators from a doubly robust moment condition. We have shown that this generally leads to robust procedures with desirable second order properties. While such estimators do not exist in all semiparametric models, they are available in a wide range of settings that are relevant in econometrics; and provide substantial improvements over existing approaches in some of them.

A. Proofs of Main Results

In this appendix, we give the proofs of Theorem 3–4 and Proposition 1–4. We give a detailed account of our argument in the case of Theorem 3. The proofs of the remaining results are very similar and structure, and are thus only sketched.

A.1. Proof of Theorem 3. To prove the first statement, note that it follows from the differentiability of ψ with respect to θ and the definition of $\widehat{\theta}$ that

$$\widehat{\theta} - \theta^o = H_n(\theta^*, \widehat{\zeta})^{-1} \frac{1}{n} \sum_{i=1}^{n} \psi(Z_i, \theta^o, \widehat{\zeta}_1(U_{1i}), \widehat{\zeta}_2(U_{2i}))$$

for some intermediate value θ^* between θ^o and $\widehat{\theta}$, and $H_n(\theta, \zeta) = \sum_{i=1}^{n} \partial_\theta \psi(Z_i\theta, \zeta_1(U_{1i}), \zeta_2(U_{2i}))$. It then follows from standard arguments that $H_n(\theta^*, \widehat{\zeta}) = H + o_P(1)$. Next, we consider an expansion of the term $\Psi_n(\theta^o, \widehat{\zeta}) = n^{-1} \sum_{i=1}^{n} \psi(Z_i, \theta^o, \widehat{\zeta}_1(U_{1i}), \widehat{\zeta}_2(U_{2i}))$. Using the notation that

$$\begin{aligned}
&\psi_i^1 = \partial\psi(Z_i, \theta^o, t, \zeta_2^o(U_{2i}))/\partial t|_{t=\zeta_1^o(U_i)}, \quad \psi_i^{11} = \partial^2\psi(Z_i, \theta^o, t, \zeta_2^o(U_{2i}))/\partial t|_{t=\zeta_1^o(U_i)}, \\
&\psi_i^2 = \partial\psi(Z_i, \theta^o, \zeta_1^o(U_i), t)/\partial t|_{t=\zeta_2^o(U_{2i})}, \quad \psi_i^{22} = \partial^2\psi(Z_i, \theta^o, \zeta_1^o(U_i), t)/\partial t|_{t=\zeta_2^o(U_{2i})}, \text{ and} \\
&\psi_i^{12} = \partial^2\psi(Z_i, \theta^o, t_1, t_2)/\partial t_1 \partial t_2|_{t_1=\zeta_1^o(U_i), t_2=\zeta_2^o(U_{2i})},
\end{aligned}$$

we find that because of assumption (i) we have that

$$\begin{aligned}
\Psi_n(\theta^o, \widehat{\zeta}) - \Psi_n(\theta^o, \zeta^o) = &\frac{1}{n} \sum_{i=1}^{n} \psi_i^1 (\widehat{\zeta}_1(U_{1i}) - \zeta_1^o(U_{1i})) + \frac{1}{n} \sum_{i=1}^{n} \psi_i^2 (\widehat{\zeta}_2(U_{2i}) - \zeta_2^o(U_{2i})) \\
&+ \frac{1}{n} \sum_{i=1}^{n} \psi_i^{11} (\widehat{\zeta}_1(U_{1i}) - \zeta_1^o(U_{1i}))^2 + \frac{1}{n} \sum_{i=1}^{n} \psi_i^{22} (\widehat{\zeta}_2(U_{2i}) - \zeta_2^o(U_{2i}))^2 \\
&+ \frac{1}{n} \sum_{i=1}^{n} \psi_i^{12} (\widehat{\zeta}_1(U_{1i}) - \zeta_1^o(U_{1i}))(\widehat{\zeta}_2(U_{2i}) - \zeta_2^o(U_{2i})) \\
&+ O_P(\|\widehat{\zeta}_1 - \zeta_1^o\|_\infty^3) + O_P(\|\widehat{\zeta}_2 - \zeta_2^o\|_\infty^3).
\end{aligned}$$

By Lemma 5(i) and Assumption 4, the two "cubic" remainder terms are both of the order $o_P(n^{-1/2})$. In Lemma 1–3 below, we show that the remaining five terms on the right hand side of the previous equation are also all of the order $o_P(n^{-1/2})$ under the conditions of the theorem. This completes the proof of the first statement of the theorem. The asymptotic normality result then follows from a simple application of the Central Limit Theorem. The proof of consistency of the variance estimator is standard, and thus omitted.

The proofs of following Lemmas repeatedly use the result that assumption (i) of Theo-

rem 3 combined with the double robustness property implies that

$$0 = \mathbb{E}(\psi_i^1 \lambda_1(U_{1i})) = \mathbb{E}(\psi_i^{11} \lambda_1(U_{1i})^2) = \mathbb{E}(\psi_i^2 \lambda_2(U_{2i})) = \mathbb{E}(\psi_i^{22} \lambda_2(U_{2i})^2) \tag{A.1}$$

for all functions λ_1 and λ_2 such that $\zeta_1^o + t\lambda_1 \in \aleph_1$ and $\zeta_2^o + t\lambda_2 \in \aleph_2$ for any $t \in \mathbb{R}$ with $|t|$ sufficiently small. To see why that is the case, consider the first equality (the argument is similar for the remaining ones). By dominated convergence, we have that

$$\mathbb{E}(\psi_i^1 \lambda_1(U_{1i})) = \lim_{t \to 0} \frac{\Psi(\theta^o, \zeta_1^o + t\lambda_1, \zeta_2^o) - \Psi(\theta^o, \zeta_1^o, \zeta_2^o)}{t} = 0$$

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

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

$$(i) \quad \frac{1}{n} \sum_{i=1}^n \psi^1(Z_i)(\widehat{\zeta}_1(U_{1i}) - \zeta_1^o(U_{1i})) = o_P(n^{-1/2}),$$
$$(ii) \quad \frac{1}{n} \sum_{i=1}^n \psi^2(Z_i)(\widehat{\zeta}_2(U_{2i}) - \zeta_2^o(U_{2i})) = o_P(n^{-1/2}).$$

Proof. We show the statement for a generic $g \in \{1, 2\}$. From Lemma 5 and Assumption 4, it follows that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \psi_i^g(\widehat{\zeta}_g(U_{gi}) - \zeta_g^o(U_{gi})) = \frac{1}{n} \sum_{i=1}^n \psi_i^g(B_{gn}(U_{gi}) + S_{gn}(U_{gi}) + R_{gn}(U_{gi})) \\ + O_P(\log(n)^{3/2} n^{-3/2} h_g^{-3d_g/2}), \end{aligned}$$

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

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \psi_i^g B_{gn}(U_{gi}) &= \mathbb{E}(\psi_i^g B_{gn}(U_{gi})) + O_P(h_g^{l_g+1} n^{-1/2}) \\ &= O_P(h_g^{l_g+1} n^{-1/2}), \end{aligned}$$

where the first equality follows from Chebyscheff's inequality, and the second equality follows from Lemma 5 and the fact that by equation (A.1) we have that $\mathbb{E}(\psi_i^g B_{gn}(U_{gi})) = 0$. Next,

consider the term

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i^g S_{gn}(U_{gi}) = \frac{1}{n^2}\sum_i\sum_{j\neq i}\psi_i^g e_1^\top N_{gn}(U_{gi})^{-1}w_{gj}(U_{gi})K_{h_g}(X_{gj}-U_{gi})\varepsilon_{gi}.$$

This is a second order U-Statistic (up to a bounded, multiplicative term), and since by equation (A.1) we have that $\mathbb{E}(\psi_i^g e_1^\top N_{gn}(U_{gi})^{-1}w_{gj}(U_{gi})K_{h_g}(X_{gj}-U_{gi})|X_{gj}) = 0$, its kernel is first-order degenerate. It then follows from Lemma 4 and some simple variance calculations that

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i^g S_{gn}(U_{gi}) = O_P(n^{-1}h_g^{-d_g/2}).$$

Finally, we consider the term

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i^g R_{gn}(U_{gi}) = T_{n,1} + T_{n,2},$$

where

$$T_{n,1} = \frac{1}{n^3}\sum_i\sum_{j\neq i}\psi_i^g e_1^\top \eta_{gn,j}(U_{gi})N_n(u)^{-2}w_{gj}(U_{gi})K_{h_g}(X_{gj}-U_{gi})\varepsilon_{gj} \text{ and}$$
$$T_{n,2} = \frac{1}{n^3}\sum_i\sum_{j\neq i}\sum_{l\neq i,j}\psi_i^g e_1^\top \eta_{gn,j}(U_{gi})N_n(U_{gi})^{-2}w_{gl}(U_{gi})K_{h_g}(X_{gl}-U_{gi})\varepsilon_{gl}.$$

Using equation (A.1), 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

$$T_{n,2} = O_P(n^{-3/2}h_g^{-d_g})$$

by Lemma 4 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

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

The statement of the lemma thus follows if $h_g \to 0$ and $n^2h_g^{3d_g} \to \infty$ as $n \to \infty$, which holds

by Assumption 4. This completes our proof. □

Remark 2. Without the DR property, the term $n^{-1}\sum_{i=1}^{n}\psi_i^g B_{gn}(U_{gi})$ in the above proof would be of the larger order $O(h_g^{l_g+1})$, 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 ?? below).

Lemma 2. *Under Assumption 1–4, the following statements hold:*

$$(i) \quad \frac{1}{n}\sum_{i=1}^{n}\psi_i^{11}(\widehat{\zeta}_1(U_{1i})-\zeta_1^o(U_{1i}))^2 = o_P(n^{-1/2}),$$
$$(ii) \quad \frac{1}{n}\sum_{i=1}^{n}\psi_i^{22}(\widehat{\zeta}_2(U_{2i})-\zeta_2^o(U_{2i}))^2 = o_P(n^{-1/2}).$$

Proof. We show the statement for a generic $g \in \{1,2\}$. Note that by Lemma 5 we have that

$$(\widehat{\zeta}_g(u)-\zeta_g^o(u))^2 = \sum_{k=1}^{6} T_{n,k}(u) + O_P\left(\left(\frac{\log(n)}{nh_g^{d_g}}\right)^{3/2}\right)\left(O_P(h_g^{l_g+1}) + O_P\left(\frac{\log(n)}{nh_g}\right)\right),$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i^{gg}T_{n,1}(U_{gi}) = \mathbb{E}(\psi_i^{gg}B_{gn}(U_{gi})^2) + O_P(n^{-1/2}h_g^{2l_g+2}) = O_P(n^{-1/2}h_g^{2l_g+2})$$

because $\mathbb{E}(\psi_i^{gg}B_{gn}(U_{gi})^2) = 0$ by equation (A.1). Second, for $k=2$ we can write

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i^{gg}T_{n,2}(U_{gi}) = T_{n,2,A} + T_{n,2,B}$$

where

$$T_{n,2,A} = \frac{1}{n^3}\sum_i \sum_{j\neq i} \psi_i^{gg}(e_1^\top N_{gn}(U_{gi})^{-1} w_{gj}(U_{gi}))^2 K_{h_g}(X_{gj}-U_{gi})^2 \varepsilon_{gj}^2$$

$$T_{n,2,B} = \frac{1}{n^3}\sum_i \sum_{j\neq i} \sum_{l\neq i,j} \psi_i^{gg} e_1^\top N_{gn}(U_{gi})^{-1} w_{gj}(U_{gi}) K_{h_g}(X_{gj}-U_{gi})\varepsilon_{gj}$$
$$\cdot e_1^\top N_{gn}(U_{gi})^{-1} w_{gl}(U_{gi}) K_{h_g}(X_{gl}-U_{gi})\varepsilon_{gl}$$

Using equation (A.1), 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

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

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

$$T_{n,2,A} = n^{-1} O_P(n^{-1/2} h^{-d_g} + n^{-1} h_g^{-3d_g/2}) = O_P(n^{-3/2} h^{-d_g}) = O_P(T_{n,2,B}).$$

Third, for $k = 4$ we use again equation (A.1) and Lemma 4 to show that

$$\frac{1}{n}\sum_{i=1}^n \psi_i^{gg} T_{n,4}(U_{gi}) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j\neq i} \psi_i^{gg} B_{gn}(U_{gi}) e_1^\top N_{gn}(U_{gi})^{-1} w_{gj}(U_{gi}) K_{h_g}(X_{gj}-U_{gi})\varepsilon_{gj}$$
$$= O_P(n^{-1} h_g^{-d_g/2}) \cdot O(h_g^{l_g+1}),$$

where the last equality follows from the fact that $n^{-1}\sum_{i=1}^n \psi_i^{gg} T_{n,4}(U_{gi})$ 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 1 to show that

$$\frac{1}{n}\sum_{i=1}^n \psi^{11}(Z_i) T_{n,5}(U_{gi}) = O_P(n^{-3/2} h_g^{-d_g} h_g^{l_g+1})$$

Finally, we obtain a number of crude bounds based on uniform rates in Lemma 5:

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i^{gg}T_{n,3}(U_{gi}) = O_P(\|R_{gn}\|_\infty^2) = O_P(\log(n)^2 n^{-2}h_g^{-2d_g})$$
$$\frac{1}{n}\sum_{i=1}^{n}\psi_i^{gg}T_{n,6}(U_{gi}) = O_P(\|R_{gn}\|_\infty)\cdot O_P(\|S_{gn}\|_\infty) = O_P(\log(n)^{3/2} n^{-3/2}h_g^{-3d_g/2})$$

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

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

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

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

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

The leading term of an expansion of the sum $T_{n,2,A} + T_{n,2,B}$ would thus be a pure bias term of order $n^{-1}h_g^{-d_g}$. This term is analogous to the "degrees of freedom bias" in Ichimura and Linton (2005), and the "nonlinearity bias" or "curse of dimensionality bias" in Cattaneo et al. (2013). In our context, the DR property of the moment conditions removes this term, which illustrates how our structure acts like a bias correction method.

Lemma 3. *Under Assumption 1–4, the following statement holds:*

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i^{12}(\widehat{\zeta}_1(U_{1i}) - \zeta_1^o(U_{1i}))(\widehat{\zeta}_2(U_{2i}) - \zeta_2^o(U_{2i})) = o_P(n^{-1/2}).$$

Proof. By Lemma 5, one can see that uniformly over $u = (u_1, u_2)$ we have that

$$\begin{aligned}
&(\widehat{\zeta}_1(u_1) - \zeta_1^o(u_1))(\widehat{\zeta}_2(u_2) - \zeta_2^o(u_2)) \\
&\quad = \sum_{k=1}^{9} T_{n,k}(u) + O_P\left(\left(\frac{\log(n)}{nh_1^{d_1}}\right)^{3/2}\right)\left(O_P(h_2^{l_2+1}) + O_P\left(\frac{\log(n)}{nh_2^{d_2}}\right)\right) \\
&\quad + O_P\left(\left(\frac{\log(n)}{nh_2^{d_2}}\right)^{3/2}\right)\left(O_P(h_1^{l_1+1}) + O_P\left(\frac{\log(n)}{nh_1^{d_1}}\right)\right)
\end{aligned}$$

where $T_{n,1}(u) = B_{1,n}(u_1)B_{2,n}(u_2)$, $T_{n,2}(u) = B_{1,n}(u_1)S_{2,n}(u_2)$, $T_{n,3}(u) = B_{1,n}(u_1)R_{2,n}(u_2)$, $T_{n,4}(u) = S_{1,n}(u_1)B_{2,n}(u_2)$, $T_{n,5}(u) = S_{1,n}(u_1)S_{2,n}(u_2)$, $T_{n,6}(u) = S_{1,n}(u_1)R_{2,n}(u_2)$, $T_{n,7}(u) = R_{1,n}(u_1)B_{2,n}(u_2)$, $T_{n,8}(u) = R_{1,n}(u_1)S_{2,n}(u_2)$, and $T_{n,9}(u) = R_{1,n}(u_1)R_{2,n}(u_2)$. Since the last two terms on the right-hand side of the previous equation are easily of the order $o_P(n^{-1/2})$ by Assumption 4, it suffices to show that for any for $k \in \{1, \ldots, 9\}$ we have that $n^{-1}\sum_{i=1}^n \psi_i^{12} T_{n,k}(U_i) = o_P(n^{-1/2})$. As in the proof of Lemma 2, we proceed by obtaining sharp bounds on $n^{-1}\sum_{i=1}^n \psi_i^{12} T_{n,k}(U_i)$ for $k \in \{1, \ldots, 5, 7\}$ using a similar strategy as in the proofs above, and crude bounds for $k \in \{6, 8, 9\}$ simply using the uniform rates derived in Lemma 5. First, arguing as in the proof of Lemma 1 and 2 above, we find that

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i^{12} T_{n,1}(U_i) = \mathbb{E}(\psi_i^{12} B_{1,n}(U_{1i})B_{2,n}(U_{2i})) + O_P(n^{-1/2}h_1^{l_1+1}h_2^{l_2+1}) = O_P(h_1^{l_1+1}h_2^{l_2+1}),$$

where the last equation follows from the fact that $\mathbb{E}(\psi_i^{12} B_{1,n}(U_{1i})B_{2,n}(U_{2i})) = O(h_1^{l_1+1}h_2^{l_2+1})$. Second, for $k = 2$ we consider the term

$$\frac{1}{n}\sum_{i} \psi_i^{12} T_{n,2}(U_i) = \frac{1}{n^2}\sum_{i}\sum_{j\neq i} \psi_i^{12} B_{1,n}(U_{1i}) e_1^\top N_{2,n}(U_{2i})^{-1} w_{2j}(U_{2i}) K_{h_2}(X_{2,j} - U_{2i})\varepsilon_{2,j}.$$

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 4 and some variance calculations that

$$\frac{1}{n}\sum_{i} \psi_i^{12} T_{n,2}(U_i) = O_P(n^{-1/2}h_1^{l_1+1}) + O_P(n^{-1}h_2^{-d_2/2}h_1^{l_1+1})$$

Using the same argument, we also find that

$$\frac{1}{n}\sum_i \psi_i^{12} T_{n,4}(U_i) = O_P(n^{-1/2} h_2^{l_2+1}) + O_P(n^{-1} h_1^{-d_1/2} h_2^{l_2+1}).$$

For $k = 3$, we can argue as in the final step of the proof of Lemma 1 to show that

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i^{12} T_{n,3}(U_i) = O_P(n^{-1} h_2^{-d_2/2} h_1^{l_1+1}) + O_P(n^{-3/2} h_2^{-d_2} h_1^{l_1+1}),$$

and for the same reason we find that

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i^{12} T_{n,7}(U_i) = O_P(n^{-1} h_1^{-d_1/2} h_2^{l_2+1}) + O_P(n^{-3/2} h_1^{-d_1} h_2^{l_2+1}).$$

Next, we consider the case $k = 5$. In some sense deriving the order of this term is the most critical step for proving Theorem 3, and it is the only one in which we exploit the condition (4.2). We start by considering the decomposition

$$\frac{1}{n}\sum_i \psi_i^{12} T_{n,5}(U_i) = T_{n,5,A} + T_{n,5,B},$$

where

$$\begin{aligned} T_{n,5,A} = \frac{1}{n^3}\sum_i \sum_{j\neq i} \psi_i^{12} (e_1^\top N_{1,n}(U_{1i})^{-1} w_{1j}(U_{1i}) K_{h_1}(X_{1j} - U_{1i}) \varepsilon_{1,j}) \\ \cdot (e_1^\top N_{2,n}(U_{2i})^{-1} w_{2j}(U_{2i}) K_{h_2}(X_{2,j} - U_{2i}) \varepsilon_{2j}), \\ T_{n,5,B} = \frac{1}{n^3}\sum_i \sum_{j\neq i} \sum_{l\neq i,j} \psi_i^{12} e_1^\top N_{1,n}(U_{1i})^{-1} w_{1j}(U_{1i}) K_{h_1}(X_{1,j} - U_{1i}) \varepsilon_{1j} \\ \cdot e_1^\top N_{2,n}(U_{2i})^{-1} w_{2l}(U_{2i}) K_{h_2}(X_{2l} - U_{2i}) \varepsilon_{2l}. \end{aligned}$$

Here term $T_{n,5,B}$ is equal to a third-order U-Statistic (up to a bounded, multiplicative term) with first-order degenerate kernel. Finding the variance of this U-Statistic is slightly more involved, as it depends on the number of joint components of U_1 and U_2. Using Lemma 4 and some tedious calculations, we obtain the following bound:

$$T_{n,5,B} = O_P(n^{-1} \max\{h_1^{-d_1/2}, h_2^{-d_2/2}\}) + O_P(n^{-3/2} h_1^{-d_1/2} h_2^{-d_2/2}).$$

This bound is sufficient four our purposes. Now consider the term $T_{n,5,A}$, which is equal to n^{-1} times a second order U-statistic (up to a bounded, multiplicative term). Since condition (4.2) implies that $\mathbb{E}(\varepsilon_1\varepsilon_2|X_1, X_2) = 0$, we find that this U-Statistic has mean zero. The calculation of its variance is again slightly more involved, and the exact result depends on the number of joint components of U_1 and U_2, and on the number of joint components of X_1 and X_2. After some calculations, we obtain the bound that We thus find that

$$T_{n,5,A} = n^{-1} \cdot O_P(n^{-1/2} \max\{h_1^{-d_1/2}, h_2^{-d_2/2}\} + n^{-1} h_1^{-d_1/2} h_2^{-d_2/2}) = n^{-1/2} O_P(T_{n,5,B}),$$

which again suffices for our purposes. Finally, we obtain a number of crude bounds based on uniform rates in Lemma 5 for the following terms:

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i^{12} T_{n,6}(U_i) = O_P(\|S_{1,n}\|_\infty) \cdot O_P(\|R_{2,n}\|_\infty) = O_P(\log(n)^{5/2} n^{-5/2} h_1^{-d_1} h_2^{-3d_2/2})$$
$$\frac{1}{n}\sum_{i=1}^{n} \psi_i^{12} T_{n,8}(U_i) = O_P(\|R_{1,n}\|_\infty) \cdot O_P(\|S_{2,n}\|_\infty) = O_P(\log(n)^{5/2} n^{-5/2} h_2^{-d_2} h_1^{-3d_1/2})$$
$$\frac{1}{n}\sum_{i=1}^{n} \psi_i^{12} T_{n,9}(U_i) = O_P(\|R_{1,n}\|_\infty) \cdot O_P(\|R_{2,n}\|_\infty) = O_P(\log(n)^{3} n^{-3} h_1^{-3d_1/2} h_2^{-3d_2/2})$$

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

Remark 4. The derivation of the order of the term $T_{n,5,A}$ is the only step in our proof that requires the orthogonality condition (4.2). Without this condition, the kernel of the respective U-Statistic would not be mean zero, and in general we would only find that $T_{n,5,A} = O_P(n^{-1} \max\{h_1^{-d_1}, h_2^{-d_2}\})$.

A.2. Proof of Theorem 4. Following the proofs of Lemma 1–3 below, we find that the two largest terms that appear in these calculations are $O(h^{2(l+1)})$ and $O_P(n^{-1}h^{-d/2})$ under the conditions of the theorem. The result then follows by verifying that these two terms are of larger order than the "cubic" remainder terms, which follows from Lemma 5(i).

A.3. Proof of Proposition 1. The estimator has the same structure as the one studied in Theorem 3–4, and thus the statement of the proposition follows from the same kind of arguments. Note that the condition (4.2) is satisfied here, since the assumption that the data are missing at random implies that $\mathbb{E}((Y - E(Y|D = 1, X)) \cdot (D - E(D|X))|D = 1, X) = 0$.

A.4. Proof of Proposition 2. The estimator has the same structure as the one studied in Theorem 3–4, and thus the statement of the proposition follows from the same kind of arguments. See also Linton (1995) for a similar derivation.

A.5. Proof of Proposition 3. To be completed (but follows using similar arguments as the proof of Theorem 3–4).

A.6. Proof of Proposition 4. To be completed (but follows using similar arguments as the proof of Theorem 3–4).

B. Auxiliary Results

In this section, we state two auxiliary results that are used repeatedly in the proof of Theorem 3.

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 4. *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 $\widehat{\zeta}_g$. This is a minor variation of results given in e.g. Masry (1996) or Kong et al. (2010). We require the following notation. For any $s \in \{0, 1, \ldots, l_g\}$ let $n_s = \binom{s+d_g-1}{d_g-1}$ be the number of distinct d_g-tuples u with $|u| = s$. Arrange these d_g-tuples as a sequence in a lexicographical order with the highest priority given to the last position, so that $(0, \ldots, 0, s)$ is the first element in the sequence and $(s, 0, \ldots, 0)$ the last element. Let τ_s denote this 1-to-1 mapping, i.e. $\tau_s(1) = (0, \ldots, 0, s)$, $\ldots$, $\tau_s(n_s) = (s, 0, \ldots, 0)$. For each $s \in \{0, 1, \ldots, l_g\}$ we also define a

$n_s \times 1$ vector $w_{gj,s}(u)$ with its kth element given by $((X_{gj}-u)/h_g)^{\tau_s(k)}$. Finally, we put

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

To better understand this notation, note that for the simple case that $l_g = 0$, i.e. when $\widehat{\zeta}_g$ is the Nadaraya-Watson estimator, we have that $w_{gj}(u) = 1$, that the term $M_{gn}(u) = n^{-1}\sum_{i=1}^{n} K_{h_g}(X_{gi}-u)$ is the usual Rosenblatt-Parzen density estimator, that $N_{gn}(u) = \mathbb{E}(K_{h_g}(X_{gi}-u))$ is its expectation, and that $\eta_{gn,i}(u) = K_{h_g}(X_{gi}-u) - \mathbb{E}(K_{h_g}(X_{gi}-u))$ is a mean zero stochastic term with variance of the order $O(h_g^{-d_g})$. Also note that with this notation we can write the estimator $\widehat{\zeta}_g(U_{gi})$ as

$$\widehat{\zeta}_g(U_{gi}) = \frac{1}{n-1}\sum_{j\neq i} e_1^\top M_{gn}(U_{gi})^{-1} w_{gj}(U_{gi}) K_{h_g}(X_{gj}-U_{gi}) Y_{gj},$$

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

$$\begin{aligned}
B_{gn}(U_{gi}) &= e_1^\top N_{gn}(U_{gi})^{-1}\mathbb{E}(w_{gj}(U_{gi})K_{h_g}(X_{gj}-U_{gi})(\xi_1^o(X_{gj}) - \xi_1^o(U_{gi}))|U_{gi}) \\
S_{gn}(U_{gi}) &= \frac{1}{n}\sum_{j\neq i} e_1^\top N_{gn}(U_{gi})^{-1} w_{gj}(U_{gi}) K_{h_g}(X_{gj}-U_{gi})\varepsilon_{gj} \\
R_{gn}(U_{gi}) &= \frac{1}{n}\sum_{j\neq i} e_1^\top \left(\frac{1}{n}\sum_{l\neq i}\eta_{gn,l}(U_{gi})\right) N_{gn}(U_{gi})^{-2} w_{gj}(U_{gi}) K_{h_g}(X_{gj}-U_{gi})\varepsilon_{gj}
\end{aligned}$$

We refer to these three terms as the bias, and the first- and second-order stochastic terms, respectively. Here $\varepsilon_{gj} = Y_{gj} - \xi_1^o(X_{gj})$ is the nonparametric regression residual, which satisfies $\mathbb{E}(\varepsilon_{gj}|X_{gj}) = 0$ by construction. To get an intuition for the behavior of the two stochastic

terms, it is again instructive to consider simple case that $l_g = 0$, for which

$$S_{gn}(U_{gi}) = \frac{1}{n\bar{f}_{gn}(U_{gi})} \sum_{j \neq i} K_{h_g}(X_{gj} - U_{gi})\varepsilon_{gj} \text{ and}$$
$$R_{gn}(U_{gi}) = \frac{1}{n\bar{f}_{gn}(U_{gi})^2} \left(\frac{1}{n} \sum_{l \neq i} (K_{h_g}(X_{gl} - U_{gi}) - \bar{f}_{gn}(U_{gi})) \right) \sum_{j \neq i} K_{h_g}(X_{gj} - U_{gi})\varepsilon_{gj}$$

with $\mathbb{E}(K_{h_g}(X_{gj} - u)) = \bar{f}_{gn}(u)$. With this notation, we obtain the following result.

Lemma 5. *Under Assumptions 2–3, the following statements hold for $g \in \{1, 2\}$:*

(i) For uneven $l_g \geq 1$ the bias B_{gn} satisfies

$$\max_{i \in \{1,\ldots,n\}} |B_{gn}(U_{gi})| = O_P(h_g^{l_g+1}),$$

and the first- and second-order stochastic terms satisfy

$$\max_{i \in \{1,\ldots,n\}} |S_{gn}(U_{gi})| = O_P((nh_g^{d_g}/\log n)^{-1/2}) \text{ and } \max_{i \in \{1,\ldots,n\}} |R_{gn}(U_{gi})| = O_P((nh_g^{d_g}/\log n)^{-1}).$$

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

$$\max_{i \in \{1,\ldots,n\}} |\widehat{\xi}_g(U_{gi}) - \xi_g^o(U_{gi}) - B_{gn}(U_{gi}) - S_{gn}(U_{gi}) - R_{gn}(U_{gi})| = O_P((nh_g^{d_g}/\log n)^{-3/2}).$$

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

$$\max_{i \in \{1,\ldots,n\}} \|n^{-1} \sum_{j \neq i} \eta_{gn,j}(U_{gi})\| = O_P((nh_g^{d_g}/\log n)^{-1/2}).$$

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

C. Some more Examples

In this section, we give a number of additional examples that essentially have the same structure as our Example 1. This is done to illustrate the broad applicability of our results in that setting.

Example 5 (Population Mean with Missing Data)**.** Let X be a vector of covariates that is always observed, and Y a scalar outcome variable that is observed if $D = 1$, and unobserved if $D = 0$. The data consists of a sample from the distribution of $Z = (DY, X, D)$, and the parameter of interest is $\theta^o = \mathbb{E}(Y)$. Assume that the data are missing at random (MAR), i.e. $\mathbb{E}(D|Y, X) = \mathbb{E}(D|X) > 0$ with probability 1, and define the functions and $\xi_1^o(x) = \mathbb{E}(Y|D = 1, X = x)$ and $\xi_2^o(x) = \mathbb{E}(D|X = x)$. Then

$$\psi(z, \theta, \xi) = \frac{d(y - \xi_1(x))}{\xi_2(x)} + \xi_2(x) - \theta$$

is a DR moment function for estimating θ^o. □

Example 6 (Linear Regression with Missing Covariates)**.** Let $X = (X_1^\top, X_2^\top)^\top$ be a vector of covariates and Y a scalar outcome variable. Suppose that the covariates in X_1 are only observed if $D = 1$ and unobserved if $D = 0$, whereas (Y, X_2) are always observed. The data thus consists of a sample from the distribution of $Z = (Y, X_1 D, X_2, D)$. Here we consider the vector of coefficients θ^o from a linear regression of Y on X as the parameter of interest. Define the functions $\xi_1^o(y, x_2, \theta) = \mathbb{E}(\varphi(Y, X, \theta)|D = 1, Y = y, X_2 = x_2)$ with $\varphi(Y, X, \theta) = (1, X^\top)^\top (Y - (1, X^\top)\theta)$ and $\xi_2^o(y, x_2) = \mathbb{E}(D|Y = y, X_2 = x_2)$ and , and assume that $\xi_2^o(Y, X_2) > 0$ with probability 1. Then

$$\psi(z, \theta, \xi) = \frac{d(\varphi(y, x, \theta) - \xi_1(y, x_2, \theta))}{\xi_2(y, x_2)} + \xi_1(y, x_2, \theta)$$

is a DR moment function for estimating θ^o. □

Example 7 (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 $\xi_1^o(d, x) = \mathbb{E}(Y|D = d, X = x)$ and $\xi_2^o(x) = \mathbb{E}(D|X = x)$, and assume that $1 > \mathbb{E}(D|Y(1), Y(0), X) = \xi_2^o(X) > 0$ with probability 1. Then

$$\psi(z, \theta, \xi) = \frac{d(y - \xi_1(1, x))}{\xi_2(x)} - \frac{(1 - d)(y - \xi_1(0, x))}{1 - \xi_2(x)} + (\xi_1(1, x) - \xi_1(0, x)) - \theta$$

is a DR moment function for estimating θ^o. □

Example 8 (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 $\xi_1^o(x) = \mathbb{E}(Y|D = 0, X = x)$ and $\xi_2^o(x) = \mathbb{E}(D|X = x)$, put $\Pi_o = \mathbb{E}(D)$, and assume that $\Pi_o > 0$ and $\mathbb{E}(D|Y(1), Y(0), X) = \xi_2^o(X) < 1$ with probability 1. Then

$$\psi(z, \theta, \xi) = \frac{d(y - \xi_1(x))}{\Pi_o} - \frac{\xi_2(x)}{\Pi_o} \cdot \frac{(1 - d)(y - \xi_1(x))}{1 - \xi_2(x)} - \theta$$

is a DR moment function for estimating θ^o. □

Example 9 (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 $\xi_2^o(x) = \mathbb{E}(W|X = x)$, and suppose that $1 > \mathbb{E}(W|Y(1), Y(0), D(1), D(0), X) = \xi_2^o(X) > 0$ 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 function $\xi_{1,1}^o(w, x) = \mathbb{E}(D|W = w, X = x)$ and $\xi_{1,2}^o(w, x) = \mathbb{E}(Y|W = w, X = x)$. Then

$$\psi(z, \theta, xi) = \psi^A(z, \xi) - \theta \cdot \psi^B(z, \xi),$$

where

$$\psi^A(z, \xi) = \frac{w(y - \xi_{1,2}(1, x))}{\xi_2(x)} - \frac{(1 - w)(y - \xi_{1,2}(0, x))}{1 - \xi_2(x)} + \xi_{1,2}(1, x) - \xi_{1,2}(0, x),$$
$$\psi^B(z, \xi) = \frac{w(d - \xi_{1,1}(1, x))}{\xi_2(x)} - \frac{(1 - w)(d - \xi_{1,1}(0, x))}{1 - \xi_2(x)} + \xi_{1,1}(1, x) - \xi_{1,1}(0, x),$$

is a DR moment condition for estimating θ^o. □

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.

Angrist, J. D. and A. B. Krueger (1995): "Split-sample instrumental variables estimates of the return to schooling," *Journal of Business & Economic Statistics*, 13, 225–235.

Blundell, R., A. Duncan, and K. Pendakur (1998): "Semiparametric estimation and consumer demand," *Journal of Applied Econometrics*, 13, 435–461.

Card, D., A. Mas, and J. Rothstein (2008): "Tipping and the Dynamics of Segregation," *The Quarterly Journal of Economics*, 123, 177–218.

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 (2013): "Generalized Jackknife Estimators of Weighted Average Derivatives," *Journal of the American Statistical Association*, 108, 1243–1268.

——— (2014a): "Small bandwidth asymptotics for density-weighted average derivatives," *Econometric Theory*, 30, 176–200.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2014b): "Bootstrapping density-weighted average derivatives," *Econometric Theory*, to appear.

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

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

Chen, X. and X. Shen (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, 289–314.

Einmahl, U. and D. M. Mason (2005): "Uniform in bandwidth consistency of kernel-type function estimators," *Annals of Statistics*, 33, 1380–1403.

Engle, R. F., C. W. Granger, J. Rice, and A. Weiss (1986): "Semiparametric estimates of the relation between weather and electricity sales," *Journal of the American Statistical Association*, 81, 310–320.

Escanciano, J. C., D. T. Jacho-Chávez, and A. Lewbel (2014): "Uniform convergence of weighted sums of non and semiparametric residuals for estimation and testing," *Journal of Econometrics*, 178, 426–443.

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

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

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

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

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

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

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

Hall, P. (1992): *The bootstrap and Edgeworth expansion*, Springer.

Hall, P. and J. S. Marron (1987): "Estimation of integrated squared density derivatives," *Statistics & Probability Letters*, 6, 109–115.

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

Hausman, J. A. and W. K. Newey (1995): "Nonparametric estimation of exact consumers surplus and deadweight loss," *Econometrica*, 1445–1476.

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.

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.

Li, Q. (1996): "On the root-N-consistent semiparametric estimation of partially linear models," *Economics Letters*, 51, 277–285.

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

Marron, J. S. and M. P. Wand (1992): "Exact mean integrated squared error," *Annals of Statistics*, 20, 712–736.

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.

Newey, W. K. and T. M. Stoker (1993): "Efficiency of weighted average derivative estimators and index models," *Econometrica*, 61, 1199–223.

Nishiyama, Y. and P. M. Robinson (2005): "The Bootstrap and the Edgeworth Correction for Semiparametric Averaged Derivatives," *Econometrica*, 73, 903–948.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica*, 1403–1430.

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

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

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

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

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

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

Robinson, P. (1988): "Root-N-consistent semiparametric regression," *Econometrica*, 931–954.

Rothe, C. (2010): "Nonparametric estimation of distributional policy effects," *Journal of Econometrics*, 155, 56–70.

——— (2012): "Partial distributional policy effects," *Econometrica*, 80, 2269–2301.

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.

Seifert, B. and T. Gasser (1996): "Finite-sample variance of local polynomials: analysis and solutions," *Journal of the American Statistical Association*, 91, 267–275.

Shen, X. et al. (1997): "On methods of sieves and penalization," *Annals of Statistics*, 25, 2555–2591.

Stock, J. H. (1989): "Nonparametric policy analysis," *Journal of the American Statistical Association*, 84, 567–575.

Stoker, T. M. (1986): "Consistent estimation of scaled coefficients," *Econometrica*, 1461–1481.

——— (1991): "Equivalence of direct, indirect, and slope estimators of average derivatives," in *Nonparametric and semiparametric methods in econometrics and statistics*, ed. by W. A. Barnett, J. L. Powell, and G. Tauchen, Cambridge, UK: Cambridge University Press, 99–118.

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

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

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