Semiparametric Two-Step Estimation Using Doubly Robust Moment Conditions

Christoph Rothe and Sergio Firpo[*]

Abstract

An estimator of a finite-dimensional parameter is said to be doubly robust if it imposes parametric specifications on two unknown nuisance functions, but only requires that one of these two specifications is correct in order for the estimator to be consistent for the object of interest. In this paper, we study semiparametric versions of such estimators that use nonparametric methods for estimating the nuisance functions. We show that in many practically relevant models the particular structure of these estimators automatically removes the largest "second order" terms that otherwise adversely affect finite sample performance. The estimators are also remarkably robust with respect to the choice of smoothing parameters. By studying an abstract setting, we also clarify which features of these models are responsible for these favorable properties.

JEL Classification: C14, C21, C31, C51

Keywords: *Semiparametric estimation, missing data, treatment effects, average derivatives, partial linear model, policy effects, double robustness, higher order asymptotics*

[*] First version: December 20, 2012. This version: October 27, 2015. 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 the editor, associate editor, three anonymous referees, Matias Cattaneo, Michael Jansson, Marcelo Moreira, Ulrich Müller, Whitney Newey, Cristine Pinto, and audiences at numerous seminar and conference presentations for their helpful comments; and Bernardo Modenesi for excellent research assistance. Sergio Firpo gratefully acknowledges financial support from CNPq-Brazil.

1. Introduction

1.1. Motivation. An estimator of a finite-dimensional parameter in a semiparametric model is said to be doubly robust (DR) if it imposes parametric specifications on two unknown nuisance functions, and is consistent if at most one of these two specifications is incorrect. Estimators with a DR property feature prominently in the statistics and biostatistics literature on missing data and treatment effect models, but have also been developed for other settings. See for example Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995, 2001), Scharfstein, Rotnitzky, and Robins (1999), Robins, Rotnitzky, and van der Laan (2000), Van der Laan and Robins (2003), Bang and Robins (2005), van der Laan and Daniel (2006), Tan (2006, 2010), Kang and Schafer (2007) or Tsiatis (2007), among many others. DR methods are used much more rarely in econometrics; but see Wooldridge (2007), Graham, Pinto, and Egel (2012) or Sloczynski and Wooldridge (2014) for a discussion of some applications.

Doubly robust estimators achieve their eponymous property by combining the estimates of the two unknown nuisance functions in a particular way. To illustrate that, consider a simple missing data model where X is a vector of covariates that is always observed, and Y^* is a scalar outcome variable that is observed if $D = 1$, and unobserved if $D = 0$. The data consist of a random sample of size n from the distribution of $Z = (Y, X, D)$, where $Y = DY^*$, and the parameter of interest is $\theta^o = \mathbb{E}(Y^*)$. Also assume that the data are missing at random, that is $\mathbb{E}(D|Y^*, X) = \mathbb{E}(D|X)$, and define the regression function $\xi_1^o(x) = \mathbb{E}(Y|D = 1, X = x)$ and the propensity score $\xi_2^o(x) = \mathbb{E}(D|X = x)$. Then a class of DR estimators of θ^o is of the form

$$\widehat{\theta}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \widehat{\xi}_1(X_i))}{\widehat{\xi}_2(X_i)} + \widehat{\xi}_1(X_i)\right), \tag{1.1}$$

where $\widehat{\xi}_1$ and $\widehat{\xi}_2$ are estimates of ξ_1^o and ξ_2^o, respectively, based on "working" parametric models for these two functions.[1] Writing $\xi_g = \text{plim}_{n\to\infty} \widehat{\xi}_g$ for $g = 1, 2$, one sees that $\widehat{\theta}_{DR}$ is indeed doubly robust because

$$\operatorname*{plim}_{n\to\infty} \widehat{\theta}_{DR} = \mathbb{E}\left(\frac{D_i(Y_i - \xi_1(X_i))}{\xi_2(X_i)} + \xi_1(X_i)\right) = \theta^o \quad \text{if} \quad \xi_1 = \xi_1^o \quad \text{or} \quad \xi_2 = \xi_2^o.$$

Common alternative approaches to estimating θ^o, which only require an estimate of one of the two nuisance functions, include for example the regression-adjustment (REG) or the Inverse Probability Weighting (IPW) estimator, defined as

$$\widehat{\theta}_{REG} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\xi}_1(X_i) \quad \text{and} \quad \widehat{\theta}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{\widehat{\xi}_2(X_i)}, \tag{1.2}$$

respectively. While $\widehat{\theta}_{REG}$ is only consistent if the "working" parametric model for the regression function is correctly specified; and $\widehat{\theta}_{IPW}$ is only consistent if the "working" parametric model for the propensity score is correctly specified, for $\widehat{\theta}_{DR}$ to be consistent it suffices that either of these two "working" parametric specifications is correct. Given the high risk of model misspecification in practice, DR procedures thus have the desirable feature that they "give the analyst two chances, instead of one, to make a valid inference" (Bang and Robins, 2005, p. 962).

In the econometrics literature, semiparametric models like the above-mentioned missing data example are often estimated using semiparametric two-stage (STS) estimators, which are based on nonparametric estimates of unknown nuisance functions. For example, an STS version of the estimators in (1.2) can be obtained by using local polynomial smoothing or orthogonal series regression to estimate the regression function or the propensity score,

[1] For example, $\widehat{\xi}_1$ could be an Ordinary Least Squares estimate based on a linear specification of the regression function, and $\widehat{\xi}_2$ could be a Maximum Likelihood estimate based on a Probit or Logit specification of the propensity score. Here the notion of a "working" model means that the parametric specifications of ξ_1^o and ξ_2^o are not considered to be part of the semiparametric missing data model, but only the construction of the estimator. For example, they would not be used for the purpose of calculating the semiparametric efficiency bound for estimating θ^o. Also note that in the econometrics literature sometimes estimators are only referred to as being semiparametric if they depend on nonparametric estimates of nuisance functions. In this sense, the estimator $\widehat{\theta}_{DR}$ would be fully parametric.

respectively, instead of postulating parametric restrictions. Under suitable regularity conditions, such versions of $\widehat{\theta}_{REG}$ and $\widehat{\theta}_{IPW}$ are $\sqrt{n}$-consistent and asymptotically normal; and in both cases the limiting variance also coincides with the semiparametric efficiency bound for estimating θ^o. See Hirano, Imbens, and Ridder (2003) or Ichimura and Linton (2005) for theoretical results on STS-IPW estimation; and Imbens, Newey, and Ridder (2007) or Chen, Hong, and Tarozzi (2008) for results on STS-REG estimators.

Of course, one can also construct an STS analogue of $\widehat{\theta}_{DR}$. Standard calculations (e.g. Newey, 1994) show that the asymptotic variance of such an STS-DR estimator would be identical to those of STS versions of either $\widehat{\theta}_{REG}$ or $\widehat{\theta}_{IPW}$. In particular, all three estimators would be efficient. However, to construct an STS-DR estimator one has to compute two nonparametric estimators instead of one. This raises the question whether one should at all consider a DR approach to building an STS estimator in the above missing data example – or any other models in which DR estimators exist – given that it involves additional computational costs and leads to a procedure that may not dominate existing ones in terms of first-order asymptotic variance. Put differently: is the robustness against parametric misspecification of nuisance functions that constructions like (1.1) have of any use in a nonparametric setting where nuisance functions are always consistently estimated?

1.2. Contributions of this Paper. Motivated by these questions, this paper makes three main contributions. First, it shows that there are indeed strong reasons to prefer an STS version of $\widehat{\theta}_{DR}$ over $\widehat{\theta}_{REG}$ or $\widehat{\theta}_{IPW}$ in the context of a general missing data model where the data are missing at random (MAR). This model covers our simple example from above, but also many more realistic applications. Using kernel smoothing for estimating the nuisance functions, we show that the special structure of a construction like (1.1) automatically removes the largest "second order" terms in a linear expansion of the estimator. That is, while a standard linear expansion of $\widehat{\theta}_{DR}$, $\widehat{\theta}_{REG}$ and $\widehat{\theta}_{IPW}$ produces the same first-order terms, the corresponding remainder term is substantially smaller for $\widehat{\theta}_{DR}$ than it is for the competing procedures. As a consequence, the DR estimator can have smaller first order bias and second order variance, and its stochastic behavior is more accurately described by the usual Gaussian approximation based on first-order asymptotics. The latter feature implies for example that the coverage probability of the usual 95% confidence intervals of the form "point estimate±1.96×standard error" should be closer to its nominal value in finite samples when it is based on $\widehat{\theta}_{DR}$ instead of one of the other estimators (which all have the same standard error). Through simulations, we show that these theoretical predictions also translate well into actual finite sample gains of practically relevant magnitude.[2]

We also argue that having to choose a second smoothing parameter is not a substantial practical downside of the DR estimator. This is an important concern because it is well-known that many STS estimators, including $\widehat{\theta}_{REG}$ and $\widehat{\theta}_{IPW}$ in models like the one described in the introduction, can be highly sensitive with respect to the implementation details of the nonparametric stage in finite samples (e.g. Linton, 1995; Robins and Ritov, 1997). This sensitivity arises because the value of the smoothing parameters affects the magnitude of the remainder terms in a linear expansion of these estimators; and while these remainders are "second order" terms in an asymptotic sense, they can often be quite large for samples of the size typically encountered in empirical practice. Since the construction (1.1) automatically removes the largest of these second order terms, however, $\widehat{\theta}_{DR}$ should be rather insensitive to variation in smoothing parameters. This is confirmed by our simulations, which find that the properties of STS-DR estimators hardly vary over a wide range of smoothing parameters for estimators of the regression function and the propensity score, whereas the bias and variance of procedures like the IPW or REG estimator can be extremely sensitive in this regard.

[2] In a recent large-scale Monte Carlo study, Frölich, Huber, and Wiesenfarth (2015) report finite-sample properties of STS-DR estimators for class of complex data generating processes calibrated to mimic a realistic data set. Their results are qualitatively very similar to those that we find in our simulations; see Remark 3 below.

As a second contribution, we analyze which particular features of the semiparametric missing data model are responsible for the gain in performance of the DR estimator relative to its competitors; thus clarifying under which conditions we would expect to be able to construct estimators with analogous properties in other semiparametric models in which DR procedures are known to exist. To answer this question, we study a generic class of STS estimators based on a doubly robust moment condition. Here a moment condition is said to be 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.

Our main finding in this regard is that DR moments alone are not able to generate STS estimators with the same desirable properties that we obtained under the missing data model. Achieving analogous results requires some additional structure which, roughly speaking, ensures that the residuals from estimating the two nuisance functions are asymptotically uncorrelated. Such an orthogonality condition can follow for example from some feature of the semiparametric model, but it could also be ensured through an appropriate construction of the two nonparametric function estimates. In the case of the missing data model, for example, the orthogonality property follows from the data being missing at random.

As a third contribution, we use the results on STS estimators based on generic DR moment conditions to study estimation in semiparametric models for which the existence of doubly robust estimators is rather less well known relative to the missing data case. Specifically, we study the partially linear regression model, a model for nonparametric policy analysis, and weighted average derivatives. Our theory produces new results for some of these models, and reproduces familiar findings for others. For example, it turns out that Robinson's (1988) estimator of the parametric component of a partially linear model can in fact be interpreted as an example of an STS estimator based on a DR moment condition.

1.3. Related Literature. Semiparametric estimators that depend on nonparametrically estimated functions, such as densities or conditional expectations, are central to econometrics; and 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. Often these estimators are first-order asymptotically equivalent to sample average. In particular, an STS estimator $\widehat{\theta}$ of some parameter θ^o based on an i.i.d. sample $\{Z_i\}_{i=1}^n$ from the distribution of some random vector Z is said to be asymptotically linear with influence function $\phi(\cdot)$ if

$$R_n(\widehat{\theta}) \equiv \widehat{\theta} - \theta^o - \frac{1}{n}\sum_{i=1}^n \phi(Z_i) = o_P(n^{-1/2}), \quad \mathbb{E}(\phi(Z_i)) = 0, \quad \mathbb{E}(\phi(Z_i)\phi(Z_i)^\top) < \infty.$$

Our focus in this paper is not on the form of $\phi(\cdot)$, but on the accuracy of the first-order approximation that $R_n(\widehat{\theta}) \approx 0$, which together with an application of the Central Limit Theorem to $\frac{1}{n}\sum_{i=1}^n \phi(Z_i)$ justifies the Gaussian approximation that $\widehat{\theta} \overset{a}{\sim} N(\theta^o, \mathbb{E}(\phi(Z_i)\phi(Z_i)^\top)/n)$.

Several papers have obtained results about the magnitude of $R_n(\widehat{\theta})$ under various conditions; and in order to have a point of reference for the findings presented in this paper it is useful to review some of them. One class of results applies to settings where $\widehat{\theta}$ depends on an $(l+1)$-times differentiable regression or density function with d-dimensional argument that is estimated by kernel-type methods[3] using a bandwidth h. It is well-known that under standard regularity conditions the nonparametric estimator has bias of order h^{l+1} and (pointwise) variance of order $n^{-1}h^{-d}$ in this case. Newey and McFadden (1994) show that an STS estimator $\widehat{\theta}$ in this class generally satisfies

$$R_n(\widehat{\theta}) = O_P(h^{l+1}) + O_P(n^{-1}h^{-d}).$$

For $\widehat{\theta}$ to be asymptotically linear, one thus requires both a "small bias" and a "small variance"

[3] By kernel-type methods, we mean methods like the Rosenblatt-Parzen kernel density estimator, the Nadaraya-Watson estimator, Local Linear or Local Polynomial regression, and local parametric models.

condition on the first step nonparametric estimator. Hall and Marron (1987), Powell, Stock, and Stoker (1989) and Powell and Stoker (1996) show that if $\widehat{\theta}$ is a *linear* transformation of a "leave-one-out" kernel weighted average[4] the "small variance" condition can be relaxed because

$$R_n(\widehat{\theta}) = O_P(h^{l+1}) + O_P(n^{-1}h^{-d/2})$$

in this case. Techniques for explicitly removing the term of order $O_P(n^{-1}h^{-d})$ for estimators that are *nonlinear* transformations of a nonparametrically estimated function are discussed in the context of specific applications by Ichimura and Linton (2005) or Cattaneo, Crump, and Jansson (2013), for example. Newey, Hsieh, and Robins (2004) show that if a twicing kernel is used instead of a regular one, or if $\widehat{\theta}$ is based on an influence function in the corresponding semiparametric model, then one can weaken the "small bias" condition for asymptotic linearity since

$$R_n(\widehat{\theta}) = O_P(h^{2(l+1)}) + O_P(n^{-1}h^{-d})$$

in this case. Bickel and Ritov (2003) show that the same degree of accuracy can be achieved with a generic higher-order kernel if $\phi(\cdot)$ is sufficiently smooth and the nonparametrically estimated function is a density; see also Ichimura and Newey (2015).[5] Newey (1994) shows that when using an orthogonal series estimator in the first stage, an analogous less stringent "small bias" condition suffices for asymptotic linearity of a general class of STS estimators; a result that is extended by Shen et al. (1997), Chen and Shen (1998) and Ai and Chen (2003) to more general classes of sieve estimators. To preview one of our results from below,

[4] Functions that can be estimated by kernel-weighted averages are those of the form $\xi(a) = f(a)\mathbb{E}(B|A = a)$, where A, B are generic random variables and f is the density of of A. This class does thus not contain conditional expectation functions, for example.

[5] This result can to some extent be combined with the ones mentioned above. For example, if $\widehat{\theta}$ is a *linear* transformation of a "leave-one-out" kernel weighted average and a twicing kernel is being used, then $R_n(\widehat{\theta}) = O_P(h^{2(l+1)}) + O_P(n^{-1}h^{-d/2})$; see Newey et al. (2004). Note that estimators based on higher-order kernels typically have poor finite sample properties, and are therefore hardly used in practice. A twicing kernel is a particular type of higher-order kernel, and thus the same comment applies.

it will turn out that for kernel-based STS versions of DR estimators it typically holds that the magnitude of both second order terms is reduced relative to the baseline result of Newey and McFadden (1994); that is,

$$R_n(\widehat{\theta}) = O_P(h^{2(l+1)}) + O_P(n^{-1}h^{-d/2})$$

if the bandwidth is within an appropriate range.

STS versions of DR estimators have been used before in some papers. As mentioned above, Robinson's (1988) estimator of the parametric component of a partially linear model is in fact such a case. In a treatment effect context, Cattaneo (2010) proposed an STS estimator with the structure of a DR estimator, but did not prove that this approach has any formal advantages. The construction that leads to the DR property is explicitly exploited in Belloni, Chernozhukov, Fernández-Val, and Hansen (2014a), Belloni, Chernozhukov, and Hansen (2014b) and Farrell (2014) for treatment evaluation in a very high-dimensional setting, where a LASSO-type estimator is used in the first stage.

In parametric problems, concerns about the accuracy of first order distributional approximations are often 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 STS 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 (2014) for a cautionary tale regarding the robustness of such refinements.

1.4. Outline of the Paper. The remainder of the paper is structured as follows. In the next section, we study an STS analogue of DR estimator in a general missing data model, and show that it has favorable theoretical and practical properties relative to other commonly used STS estimators. In Section 3, we study semiparametric DR estimation in a stylized model, and show that the DR property alone is not enough to generate the type of results we obtained for the missing data model, and clarify which additional structure is needed. In Section 4, we consider three other classes of specific models which can be studied using our framework: the partially linear model, a model for policy analysis, and weighted average derivatives. Finally, Section 5 concludes. All proofs are collected in the Appendix.

2. Semiparametric Estimation in a General Missing Data Model

In this section, we study a general semiparametric missing data model that contains the simple example outlined in the introduction as a special case, but also covers for example regression models with missing covariates and/or outcome variables (e.g. Scharfstein et al., 1999; Robins and Rotnitzky, 1995; Chen et al., 2008), and a large class of treatment effect models (e.g. Heckman, Ichimura, and Todd, 1998; Hahn, 1998; Hirano et al., 2003; Imbens, 2004). We construct an STS version of a DR estimator and show that it has favorable theoretical and practical properties relative to other commonly used STS estimators.

2.1. Model. 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 $\{Z_i\}_{i=1}^n = \{(Y_i, X_i, D_i)\}_{i=1}^n$ 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 $\mathbb{R}^{d_\theta}$. Identification is achieved by assuming that Y^* is missing at random, that is $\mathbb{E}(D|Y^*, X) = \mathbb{E}(D|X)$ with probability 1. Now define the *regression function* $\xi_1^o(x, \theta) = \mathbb{E}(m(Y, X, \theta)|D = 1, X = x)$ and the *propensity score* $\xi_2^o(x) = \mathbb{E}(D|X = x)$. The propensity score is assumed to be bounded away from zero over the support of X. Also, define

$$\phi_{MD}(Z) = \mathbb{E}(\nabla_\theta m(Y^*, X, \theta^o))^{-1} \left(\frac{D(m(Y, X, \theta^o) - \xi_1^o(X, \theta^o))}{\xi_2^o(X)} + \xi_1^o(X, \theta^o) - \theta^o \right),$$

$$\Sigma_{MD} = \mathbb{E}(\phi_{MD}(Z)\phi_{MD}(Z)^\top),$$

which are, respectively, the efficient influence function and the asymptotic variance bound for estimating θ^o in this model (cf. Robins et al., 1994; Hahn, 1998; Chen et al., 2008).

2.2. Estimator. We estimate the two nuisance functions ξ_1^o and ξ_2^o by "leave-one-out" local polynomial regression. This class of kernel-based smoothers has been studied extensively by Fan (1993), Ruppert and Wand (1994), Fan and Gijbels (1996) and others. It is well-known to have attractive bias properties relative to other kernel-based methods, such as the Nadaraya-Watson estimator. For generic vectors $b = (b_1, \ldots, b_d)$ and $\alpha = (\alpha_{(0,\ldots,0)}, \alpha_{(1,0,\ldots,0)}, \ldots, \alpha_{(0,\ldots,0,l)})$, let $\mathcal{P}_{l,\alpha}(b) = \sum_{0\leq|s|\leq l} \alpha_s b^s$ be a polynomial of order l. Here $\sum_{0\leq|s|\leq l}$ denotes the summation over all d-vectors s of positive integers with $0 \leq |s| \leq l$. Also let $\mathcal{K}$ be a univariate density function, put $K_h(b) = \prod_{j=1}^d \mathcal{K}(b_j/h)/h$ for any bandwidth $h \in \mathbb{R}_+$, and define

$$\widehat{\alpha}^{-i}(x, \theta) = \underset{\alpha}{\operatorname{argmin}} \sum_{j\neq i} \left(m(Y_j, X_j, \theta) - \mathcal{P}_{l_1,\alpha}(X_j - x)\right)^2 K_{h_1}(X_j - x)\mathbb{I}\{D_j = 1\},$$

$$\widehat{\beta}^{-i}(x) = \underset{\beta}{\operatorname{argmin}} \sum_{j\neq i} \left(D_j - \mathcal{P}_{l_2,\beta}(X_j - x)\right)^2 K_{h_2}(X_j - x).$$

With this notation, the "leave-one-out" local polynomial estimates of $\xi_1^o(X_i, \theta)$ and $\xi_2^o(X_i)$ are given by

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

respectively, for $i = 1, \ldots, n$. Note that we are allowing for different orders of the local polynomial and different bandwidths when estimating ξ_1^o and ξ_2^o, but in practice they might well be the same. The estimator $\widehat{\theta}_{DR}$ is defined as the value of θ that solves the following equation:

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

We will refer to $\widehat{\theta}_{DR}$ as an STS-DR estimator in the following. For $g=1,2$, we also define the sequences $b_{gn}=h_g^{l_g+1}$ and $s_{gn}=(\log(n)/(nh_g^d))^{1/2}$, which under the conditions that we impose below correspond to the (uniform) order of the bias and the stochastic part, respectively, of our two nonparametric estimators (cf. Masry, 1996).

2.3. Main Result. We study the theoretical properties of the estimator $\widehat{\theta}_{DR}$ under the following assumptions.

Assumption K1. (i) The kernel $\mathcal{K}$ 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 MD1. (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) $\sup_{\theta\in\Theta}\mathbb{E}(\|\nabla_\theta m(Y^*,X,\theta)\|)<\infty$.

Assumption MD2. (i) X is continuously distributed both unconditionally and conditional on $D=1$, with compact and convex 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) $\xi_2^o(x)$ is (l_2+1)-times continuously differentiable; (iv) $\xi_1^o(x,\theta)$ is (l_1+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 K1 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 MD1 is a set of regularity conditions 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 MD2 collects a number smoothness and regularity conditions of the form commonly imposed in the context of nonparametric regression. We then obtain the following asymptotic normality result.

Theorem 1. *Suppose that Assumptions K1 and MD1–MD2 hold; and that h_1, h_2 are such that $b_{1n}b_{2n} = o(n^{-1/2})$, $b_{gn} = o(n^{-1/6})$ and $s_{gn} = o(n^{-1/6})$ for $g = 1, 2$. Then $\sqrt{n}(\widehat{\theta}_{DR} - \theta^o) \overset{d}{\to} N(0, \Sigma_{MD})$.*

2.4. Discussion. Theorem 1 differs from other asymptotic normality results for STS estimators (e.g. Newey, 1994; Newey and McFadden, 1994; Chen et al., 2003; Ichimura and Lee, 2010) in that it only imposes relatively weak conditions on the accuracy of the nonparametric first stage estimates. The bandwidth restrictions allow each of the smoothing biases from estimating ξ_1^o and ξ_2^o to be of the order $o(n^{-1/6})$ as long as their product is of the order $o(n^{-1/2})$, and only require the respective stochastic parts to be of the order $o_P(n^{-1/6})$. This is result is better understood by looking at the difference between $\widehat{\theta}_{DR}$ and its asymptotically linear representation. Consider the case that $h_1 = h_2 \equiv h$ and $l_1 = l_2 \equiv l$ to simplify the exposition. Then by following the proof of Theorem 1 one can see that this difference is minimized if h is chosen such from the permissible range of bandwidths that $n^{1/2d}h \to \infty$, in which case

$$\widehat{\theta}_{DR} - \theta^o - \frac{1}{n}\sum_{i=1}^{n} \phi_{MD}(Z_i) = O_P(h^{2(l+1)}) + O_P(n^{-1}h^{-d/2}). \tag{2.1}$$

The magnitudes of the two terms on the right-hand side of the previous equation correspond to those of the squared bias and $h^{d/2}$ times the (pointwise) variance of the $\widehat{\xi}_g$, respectively.

Their sum can be as small as $O_P(n^{-4(l+1)/(4(l+1)+d)})$. If $l = d = 1$, for example, the right-hand-side of (2.1) is minimized by choosing $h \propto n^{-2/9}$, and is of the order $O_P(n^{-8/9})$ in this case. If $d \leq 3$, the range of bandwidths that satisfy the conditions of Theorem 1 includes those of the form $h_g \propto n^{-1/(2(l+1)+d)}$, which minimize the Integrated Mean Squared Error (IMSE) for estimating ξ_1^o and ξ_2^o, respectively. This is convenient, as there are well-known methods such as cross-validation to construct such bandwidths. However, such bandwidths do not necessarily have any optimality properties for estimating θ^o.

It is instructive to compare these properties to that of the popular Inverse Probability Weighting (IPW) estimator $\widehat{\theta}_{IPW}$ (e.g. Hirano et al., 2003; Firpo, 2007), which is defined as the value of θ that solves the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{D_i m(Y_i, X_i, \theta)}{\widehat{\xi}_2(X_i)} = 0.$$

Following arguments like in Ichimura and Linton (2005), who study this estimator in a slightly simpler setting, we would need h_2 to be such that $b_{2n} = o(n^{-1/2})$ and $s_{2n} = o(n^{-1/4})$ to ensure that $\sqrt{n}(\widehat{\theta}_{IPW} - \theta^o) \xrightarrow{d} N(0, \Sigma_{MD})$. These conditions, which are the familiar from Newey and McFadden (1994), are called for because an expansion of $\widehat{\theta}_{IPW}$ only gives that

$$\widehat{\theta}_{IPW} - \theta^o - \frac{1}{n}\sum_{i=1}^{n}\phi_{MD}(Z_i) = O_P(h^{l+1}) + O_P(n^{-1}h^{-d}). \tag{2.2}$$

The difference between $\widehat{\theta}_{IPW}$ and its asymptotically linear representation is thus at best of the order $O_P(n^{-(l+1)/(l+1+d)})$. For example, if $l = d = 1$ the difference is at least of the order $O_P(n^{-2/3})$, which is bigger than what we obtained for the STS-DR estimator. As a consequence, we can expect standard Gaussian approximations based on first-order asymptotic theory to be more accurate in finite samples for $\widehat{\theta}_{DR}$ than for $\widehat{\theta}_{IPW}$.

Remark 1. Using a linear smoother to estimate the propensity score has the practical disadvantage that the estimates are not constrained to be between zero and one. A way to address this problem is to use a local polynomial Probit estimator instead. That is, one could redefine

$$\begin{aligned}\widehat{\beta}^{-i}(x) = \underset{\beta}{\operatorname{argmax}} \sum_{j \neq i} (D_j \log(\Phi(\mathcal{P}_{l,\beta}(X_j - x))) \\ + (1 - D_j) \log(1 - \Phi(\mathcal{P}_{l,\beta}(X_j - x))))K_h(X_j - x),\end{aligned}$$

where Φ is the CDF of the standard normal distribution, and let $\widehat{\xi}_2(X_i) = \Phi(\beta^{-i}_{(0,\ldots,0)}(X_i))$. A local Logit estimator could be defined similarly. In view of the results of Fan, Heckman, and Wand (1995) and Kong, Linton, and Xia (2010), we conjecture that this would not affect the result of our asymptotic analysis, as the local polynomial Probit estimator and the usual local polynomial estimator should have Bahadur expansions that share a common structure.

Remark 2. If one would use a "leave-in" version of the local polynomial regression estimator to construct $\widehat{\theta}_{DR}$, this would give rise to an additional bias term of order $O(n^{-1}h^{-d})$ in the right-hand-side of expansion (2.1). This bias would not vanish faster than $n^{-1/2}$ under the conditions of the Theorem. Using "leave-one-out" estimators to avoid this type of bias is a standard technique in the literature on STS estimation; see for example Hall and Marron (1987), Powell et al. (1989) or Powell and Stoker (1996).

2.5. Simulation Evidence. In this subsection, we study the finite sample properties of the STS-DR estimator through a Monte Carlo experiment, and compare them to those of other STS estimators of the same parameter. Our aim is to illustrate that the theoretical results obtained above provide a realistic picture of the behavior of the estimator in practice. The data generating process in our simulations is the special case of our general missing data model described in the introduction. The covariate X is scalar and uniformly distributed on the interval $[0, 1]$. The outcome variable Y^* is normally distributed given X with mean $\xi_1^o(X) = 1/(1 + 16 \cdot X^2)$ and variance .5. The indicator D for a complete observation is a Bernoulli random variable with mean $\xi_2^o(X) = 1 - .8 \cdot \xi_1^o(X)$. With these choices $\theta^o = \mathbb{E}(Y^*) \approx .331$ and $\Sigma_{MD} \approx .188$. We consider the sample size $n = 500$, and set the number of replications to 5,000.

To investigate the robustness of the properties of the STS-DR estimator with respect to the implementation of the nonparametric first stage, we do actually not only consider the kernel-based version described above, but also two alternative versions based on different types of nonparametric first stage estimators; and also use a variety of different smoothing parameters. Specifically, we consider estimators of the form

$$\widehat{\theta}_{DR-s} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \widehat{\xi}_1(X_i))}{\widehat{\xi}_2(X_i)} + \widehat{\xi}_1(X_i)\right)$$

with $s \in \{K, OS, SP\}$. Here $\widehat{\theta}_{DR-K}$ is the kernel based estimator described above, which uses a "leave-one-out" local linear estimator with bandwidth $h_1 \in \{.05, .08, \ldots, .5\}$ for the regression function ξ_1^o; and a "leave-one-out" local linear Logit estimator with bandwidth $h_2 \in \{.05, .08, \ldots, .5\}$ for the propensity score ξ_2^o. In both instances a Gaussian kernel is used. By $\widehat{\theta}_{DR-OS}$ we denote an orthogonal series based STS-DR estimator, which uses a linear series regression with a standard polynomial basis and $h_1 \in \{1, 2, \ldots, 11\}$ terms to estimate the regression function; and a series Logit estimator using the same set of basis functions and $h_2 \in \{1, 2, \ldots, 11\}$ terms to estimate the propensity score. Finally, we consider a spline based STS-DR estimator $\widehat{\theta}_{DR-SP}$, which uses cubic smoothing splines with smoothing parameters $h_1, h_2 \in \{.5, .55, \ldots, 1.25\}$ for the regression function and the propensity score, respectively.[6] For all combinations of nonparametric estimators and smoothing parameters,

[6] To give a point of reference, note that the smoothing parameters for estimating the regression function and the propensity score, respectively, that would be obtained by minimizing a least-squares cross-validation criterion are roughly equal for these two functions, and take a numerical value of about .1 for kernel estimation, 3 for series estimation, and 1 for splines.

we also compute nominal $(1-\alpha)$ confidence intervals of the usual form

$$CI_{DR-s}^{1-\alpha} = \left[\widehat{\theta}_{DR-s} \pm z_\alpha \cdot \left(\widehat{\Sigma}_s/n\right)^{1/2}\right],$$

where z_α is the $1-\alpha/2$ quantile of the standard normal distribution and

$$\widehat{\Sigma}_s = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \widehat{\xi}_1(X_i))}{\widehat{\xi}_2(X_i)} + \widehat{\xi}_1(X_i) - \widehat{\theta}_{DR-s}\right)^2$$

is an estimate of the asymptotic variance Σ_{MD}, for $s \in \{K, OS, SP\}$. We consider the usual confidence level $1-\alpha = .95$ for our simulations.

[TABLE 1 ABOUT HERE]

In Table 1 we report the Mean Squared Error (MSE), absolute bias (BIAS) and variance (VAR) for the various implementations of the STS-DR estimator for a subset of smoothing parameters that we considered. We scale these quantities by appropriate transformations of the sample size to make them more easily comparable to the predictions from asymptotic theory. Our results show that uniformly over all implementations the STS-DR estimators are essentially unbiased, and their variance is very close to the efficiency bound $\Sigma_{MD} \approx .188$. Correspondingly, the MSEs are also very similar to their theoretically predicted value Σ_{MD}. We also report the empirical coverage probability of the corresponding confidence intervals, which are all very close to the nominal level 95%. We interpret these results as evidence that first order asymptotic theory provides a reliable approximation to the finite sample distribution of the STS-DR estimators, and that this approximation is robust with respect to the construction of the nonparametric first stage. The last point covers both the general type of nonparametric procedure (kernel, series, splines) and the choice of smoothing parameters.

To put these results into perspective, we also study the performance of a number of alternative estimators that are not STS analogues of a DR procedure. To begin by considering inverse probability weighting (IPW) and regression (REG) type estimators using the same range of nonparametric first step procedures and smoothing parameters as for STS-DR estimators above. That is, we consider the estimators

$$\widehat{\theta}_{IPW-s} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i}{\widehat{\xi}_2(X_i)} \quad \text{and} \quad \widehat{\theta}_{REG-s} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}_1(X_i),$$

with $s \in \{K, OS, SP\}$, and the corresponding nominal $(1-\alpha)$ confidence intervals $CI_{IPW-s}^{1-\alpha}$ and $CI_{REG-s}^{1-\alpha}$. Note that these two confidence intervals by construction have the same length as $CI_{DR-s}^{1-\alpha}$ for each s and every value of the smoothing parameters, and only differ in the point at which they are centered.[7]

From the practitioner's perspective, if inference is based on calculation of analytical formulae of asymptotic variance, then it will be necessary to estimate both nuisance functions. No way out of it, unless inference is implemented using alternative methods. This point is to emphasize that the "computational" burden of our method does not really exist, as it requires the same amount of

In addition to these estimators, we also consider two modifications of the kernel-based procedures. First, we consider estimators $\widehat{\theta}_{IPW-TK}$ and $\widehat{\theta}_{REG-TK}$ that are obtained in the same way as the ordinary kernel-based IPW and REG estimator, respectively, except for using a Gaussian twicing kernel instead of a regular one (Newey et al., 2004). Second, we consider bootstrap bias corrected versions $\widehat{\theta}_{IPW-BS}$ and $\widehat{\theta}_{REG-BS}$ of the two kernel-based estimators, following recommendations in Cattaneo and Jansson (2014).[8] We also consider bootstrap-based confidence intervals for θ^o calculated using the usual percentile method. Note that the calculation of these intervals does not involve estimating the asymptotic vari-

[7] From the practitioner's perspective, STS-DR might actually not be seen as computationally more costly than, say, STS-REG or STS-IPW estimators since conducting inference based on an estimate of the asymptotic variance requires estimates of the regression function and the propensity score for all three procedures.

[8] Specifically, the estimators $\widehat{\theta}_{IPW-BS}$ and $\widehat{\theta}_{REG-BS}$ are obtained in three steps. First one computes and ordinary kernel-based IPW or REG estimator, except for *not* using a "leave-one-out" procedure in the nonparametric stage. Second, a bootstrap distribution is created by re-computing the estimator on i.i.d. draws of size n from the empirical distribution function of the data, and centering at the original estimate. Third, the mean of the bootstrap distribution is taken as an estimate of the bias, and subtracted from the original estimator.

ance, and thus do not depend on a second smoothing parameter.

[TABLE 2 ABOUT HERE]

[TABLE 3 ABOUT HERE]

The results of our simulations are given in Table 2 for IPW estimators, and Table 3 for REG estimators. They show that the properties of these estimators can differ substantially from the predictions of first-order asymptotic theory. The finite-sample distribution of the estimators also varies a lot over the various nonparametric first stage procedures we consider, and thus inference is generally less robust relative to the STS-DR estimators. For example, the kernel-based IPW estimator $\widehat{\theta}_{IPW-K}$ is strongly biased for most values of the bandwidth. Its finite sample variance is well above the theoretical asymptotic variance for all bandwidth values, exceeding it by almost 70% for $h_2 = .05$. In consequence, the coverage properties of the corresponding confidence intervals tend to be poor. We also illustrate this point in Figure 1 by plotting the MSE, bias and variance of $\widehat{\theta}_{IPW-K}$ against the results for the kernel-based DR estimator $\widehat{\theta}_{DR-K}$. One can see, for example, that the finite-sample MSE of $\widehat{\theta}_{DR-K}$ viewed as a function of the bandwidths is close to flat and near the theoretically predicted value of .188. On the other hand, the MSE of $\widehat{\theta}_{IPW-K}$ viewed as a function of the bandwidth has a pronounced "U-shape" and is always well above the theoretically predicted value.

Turning to the remaining estimators, we find that using a twicing kernel does not lead to a substantial improvement. Our results show some minor bias reduction but also a large increase in variance relative to the kernel-based IPW.[9] Bootstrap bias correction is effective at reducing the bias for small values of the smoothing parameter, but tends to increase it for larger bandwidth values. It also tends to increase the finite sample variance of the

[9] The the REG and IPW estimators based on twicing kernels also produce a number of substantial outliers, which we removed for the calculations of our summary statistics. Specifically, we discarded all realizations which differed from the median over all simulations runs by more than four times the interquartile range (we proceeded like this with all estimators to keep the results comparable).

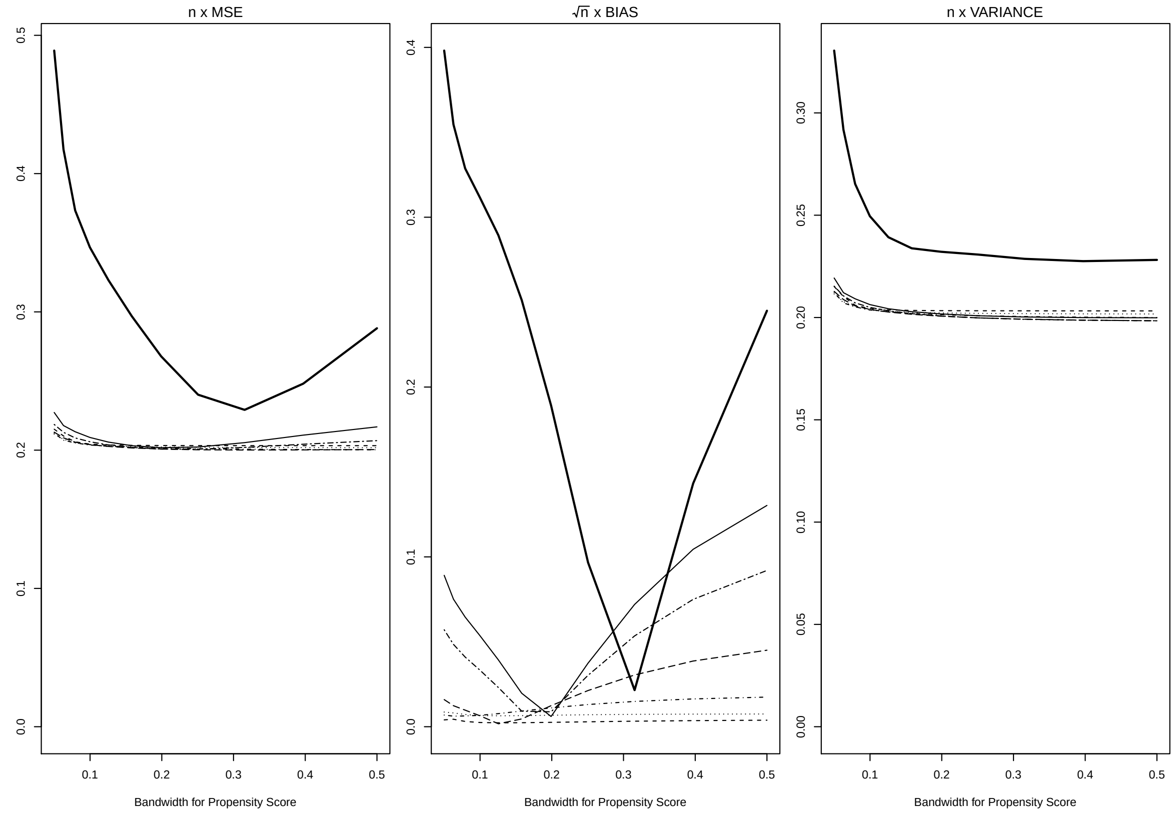

Figure 1: Simulation results: MSE, absolute bias and variance of $\widehat{\theta}_{IPW-K}$ for various values of h_2 (bold solid line), compared to results for $\widehat{\theta}_{DR-K}$ with bandwidth h_1 equal to .05 (short-dashed line), .08 (dotted line), .13 (dot-dashed line), .2 (long dashed line), .32 (long dashed dotted line), and .5 (thin solid line).

estimator, but bootstrap-based confidence intervals have good coverage properties for small and moderate values of the smoothing parameter. The orthogonal series estimator does very well in our study in terms of bias, which is small except for the implementation using a single series term. Still, its variance exceeds the predicted one by roughly 20%. Finally, the spline based estimator's bias also depends heavily on the smoothing parameter, whereas its variance properties are similar to those of the series-based one. Table 3 shows that REG-type estimators generally perform somewhat better than IPW estimators in this setting. Variances are relatively close to the efficiency bound over all nonparametric first stage procedures that we consider. Still, the bias of the kernel, twicing kernel, and bootstrap-corrected kernel estimators all vary strongly with the smoothing parameter.

Remark 3. While it would of course be possible to study the properties of STS-DR estimators under other DGPs, and to compare them to an even wider class of alternative procedures, we confine ourselves to the above results due to space constraints. Frölich et al. (2015) conduct a large-scale simulation study of STS program evaluation estimators, using a DGP designed to mimic a realistic data set. They report properties of STS-DR estimators that are qualitatively very similar to ours.

3. Semiparametric Estimation Using Doubly Robust Moment Conditions

The results we derived in the previous section prompt the question whether they are specific to the model that we considered there, or whether we should generally expect STS versions of DR estimators to have analogous favorable properties. To answer this question, we study a stylized class of STS-DR estimators that are constructed as solutions to a sample analogue of a DR moment condition; a concept that we formally define below. We show that this construction alone is not enough to obtain an asymptotic normality result like the one in the previous section, but that instead an additional orthogonality condition on the nonparametric first-stage estimation errors is needed. In our missing data model, this condition follows from the assumption that the data are missing at random.

3.1. Setup. We consider a stylized setup in which the problem is to estimate 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 some random vector $Z \in \mathbb{R}^{d_z}$. We suppose that one way (of potentially many different ones) to characterize θ^o is through a moment condition containing an infinite dimensional nuisance parameter. That is, we assume that the model which determines the distribution of Z is such that there exists a known moment function

ψ taking values in $\mathbb{R}^{d_\psi}$, with $d_\psi = d_\theta$, which satisfies the following relationship:

$$\Psi(\theta, \xi^o) := \mathbb{E}(\psi(Z, \theta, \xi^o)) = 0 \text{ if and only if } \theta = \theta^o. \tag{3.1}$$

Here ξ^o is an unknown (but identified) nuisance function that could in principle also depend on θ. We also assume that the functional Ψ in (3.1) is *doubly robust* for estimating θ^o. This means that ξ^o can be partitioned as $\xi^o = (\xi_1^o, \xi_2^o) \in \Xi_1 \times \Xi_2$ such that

$$\Psi(\theta, \xi_1^o, \xi_2) = 0 \text{ and } \Psi(\theta, \xi_1, \xi_2^o) = 0 \text{ if and only if } \theta = \theta^o \tag{3.2}$$

for all functions $\xi_1 \in \Xi_1$ and $\xi_2 \in \Xi_2$. The function ψ is called a *doubly robust moment function* in this case. To simplify the exposition, we focus on the case that the nuisance functions are two conditional expectations that do not depend on θ; that is

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

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. We also assume that the moment function is such that $\psi(Z, \theta, \xi_1, \xi_2)$ depends on ξ_g through $\xi_g(U_g)$ only, where U_g is a subvector of Z. With U denoting the union of distinct elements of U_1 and U_2, we write $\xi(U) = (\xi_1(U_1), \xi_2(U_2))$ and, with some abuse of notation, $\psi(Z, \theta, \xi_1, \xi_2) = \psi(Z, \theta, \xi_1(U_1), \xi_2(U_2))$ and $\Psi(\theta, \xi_1, \xi_2) = \mathbb{E}(\psi(Z, \theta, \xi_1(U_1), \xi_2(U_2)))$.

Remark 4. In this paper, we do not consider the question whether a DR moment condition exists in any given semiparametric model, or whether the asymptotic variance of an estimator of θ^o based on such a moment condition achieves the semiparametric efficiency bound. See Robins and Rotnitzky (2001) for some results in this regard. Our focus is on conditions for asymptotic linearity of STS estimators based on a DR moment condition.

Remark 5. Robins and Rotnitzky (2001) show that if a DR moment function exists, it has to be an element of the space of influence functions of the corresponding semiparametric model. Since every influence function for estimating a d_θ-dimensional parameter takes values in $\mathbb{R}^{d_\theta}$ by construction, without loss of generality we can focus on exactly identified settings where (3.1) holds with $d_\psi = d_\theta$.

Remark 6. Our analysis below can be carried out along similar lines if one of the nuisance functions is a density instead of a conditional expectation; or a derivative of a density or conditional expectation. In each case, the functions can be estimated by kernel-type estimators that share a common structure, and only this structure is used in our proofs.

3.2. Estimator.

Our interest is in the properties of STS estimators based on a sample analogue of a DR moment condition, to which we will refer as STS-DR estimators. Such an estimator $\widehat{\theta}_{DR}$ of θ^o can be constructed as the value of θ which solves the equation

$$\Psi_n(\theta, \widehat{\xi}) \equiv \frac{1}{n}\sum_{i=1}^{n} \psi(Z_i, \theta, \widehat{\xi}(U_i)) = 0, \tag{3.3}$$

where $\widehat{\xi} = (\widehat{\xi}_1, \widehat{\xi}_2)$ is a suitable nonparametric estimate of $\xi^o = (\xi^o_1, \xi^o_2)$. We focus on estimating ξ^o_g by "leave-one-out" local polynomial regression of order l_g using bandwidth h_g, where $g = 1, 2$. Using notation analogous to that introduced in Section 2.2, we put

$$\widehat{\alpha}_g^{-i}(x) = \underset{\alpha}{\operatorname{argmin}} \sum_{j \neq i} \left(Y_{gj} - \mathcal{P}_{l_g,\alpha}(X_{gj} - x)\right)^2 K_{h_g}(X_{gj} - x), \quad g = 1, 2.$$

The estimate $\widehat{\xi}_g(U_{gi})$ of $\xi^o_g(U_{gi})$ is then given by

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

For $g = 1, 2$, we also define the sequences $b_{gn} = h_g^{l_g+1}$ and $s_{gn} = (\log(n)/(nh_g^d))^{1/2}$, which under the conditions that we impose below correspond to the (uniform) order of the bias and the stochastic part, respectively, of our two nonparametric estimators.

3.3. Main Results.

We introduce the following further assumptions for our asymptotic analysis, which are similar to the regularity conditions imposed in Section 2.3.

Assumption G1. (i) The DR moment function $\psi(z, \theta, \xi(u))$ is three times continuously differentiable with respect to $\xi(u)$, with derivatives that are uniformly bounded; (ii) there exists $\alpha > 0$ and an open neighborhood $\mathcal{N}(\theta^o)$ of θ^o such that $\sup_{\theta \in \mathcal{N}(\theta^o)} \|\nabla_\theta \psi(Z, \theta, \xi(U)) - \nabla_\theta \psi(Z, \theta, \xi^o(U))\| \leq b(z) \|\xi - \xi^o\|^\alpha$; (iii)) the matrix $H = \mathbb{E}(\nabla_\theta \psi(Z, \theta^o, \xi^o(U))$ has full rank.

Assumption G2. 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 $c > 2$.

Following Remark 5, we expect $\widehat{\theta}_{DR}$ to be adaptive, in the sense that its own influence function and asymptotic variance are identical to that of an infeasible estimator which uses the true functions (ξ_1^o, ξ_2^o) instead of the corresponding nonparametric estimates. That is, we expect the influence function and asymptotic variance of $\widehat{\theta}_{DR}$ to be

$$\phi_G(Z) = H^{-1} \psi(Z, \theta^o, \xi_1^o, \xi_2^o) \quad \text{and} \quad \Sigma_G = \mathbb{E}(\phi_G(Z) \phi_G(Z)^\top),$$

respectively (cf. Newey, 1994). The following theorem gives conditions the for asymptotic normality of $\widehat{\theta}_{DR}$.

Theorem 2. *Suppose that Assumptions K1 and G1–G2 hold. Then* $\sqrt{n}(\widehat{\theta}_{DR} - \theta^o) \xrightarrow{d} N(0, \Sigma_G)$ *if in addition either of the following conditions is satisfied:*

(a) h_1, h_2 *are such that* $b_{gn} = o(n^{-1/4})$ *and* $s_{gn} = o(n^{-1/4})$ *for* $g = 1, 2$.

(b) the distribution of the random vector Z is such that

$$\mathbb{E}((Y_1 - \xi_1^o(X_1)) \cdot (Y_2 - \xi_2^o(X_2)) | X_1, X_2) = 0; \tag{3.4}$$

and h_1, h_2 *are such that* $b_{1n} b_{2n} = o(n^{-1/2})$, $b_{gn} = o(n^{-1/6})$ *and* $s_{gn} = o(n^{-1/6})$ *for* $g = 1, 2$.

3.4. Discussion. Theorem 2(a) shows that simply being based on a DR moment condition is not enough for an STS estimator to be asymptotically linear under the same type of weak restrictions that we imposed in the case of a semiparametric missing data model. Consider the case that $h_1 = h_2 \equiv h$ and $l_1 = l_2 \equiv l$ to simplify the exposition. From the proof of Theorem 2(a) we then see that under its conditions we only get that

$$\widehat{\theta}_{DR} - \theta^o - \frac{1}{n}\sum_{i=1}^{n}\phi_G(Z_i) = O_P(nh^{2(l+1)}) + O_P(n^{-1}h^{-d}). \tag{3.5}$$

Following Newey et al. (2004), we would expect such a result for *any* adaptive STS estimator, so there is nothing particularly special about the DR property here.[10] Under the conditions of Theorem 2(b), however, $\widehat{\theta}_{DR}$ has theoretical properties analogous to those established in Theorem 1. In particular, in this case it holds that

$$\widehat{\theta}_{DR} - \theta^o - \frac{1}{n}\sum_{i=1}^{n}\phi_G(Z_i) = O_P(nh^{2(l+1)}) + O_P(n^{-1}h^{-d/2}). \tag{3.6}$$

if h is chosen such from the permissible range of bandwidths that $n^{1/2d}h \to \infty$. The improvement comes from the fact that, roughly speaking, the term of order $O_P(n^{-1}h^{-d})$ in equation (3.5) is driven by the asymptotic covariance between the residuals of $\widehat{\xi}_1$ and $\widehat{\xi}_2$. The orthogonality condition (3.4) then essentially ensures that $\widehat{\xi}_1 - \xi_1^o$ and $\widehat{\xi}_2 - \xi_2^o$ are asymptotically uncorrelated, and the $O_P(n^{-1}h^{-d})$ term drops out.

Remark 7. In our missing data model, condition (3.4) is implied by the assumption that the data are missing at random. Using the Law of Iterated Expectations, we find that

$$\begin{aligned}&\mathbb{E}(D(m(Y,X,\theta^o) - \mathbb{E}(m(Y,X,\theta^o)|D=1,X)) \cdot (D - \mathbb{E}(D|X))|X)\\ &\quad = \mathbb{E}((m(Y,X,\theta^o) - \mathbb{E}(m(Y,X,\theta^o)|D=1,X))|D=1,X) \cdot (1-\mathbb{E}(D|X)) \cdot \mathbb{E}(D|X) = 0.\end{aligned}$$

Remark 8. The condition (3.4) is not the only way to achieve (3.6). In principle, any

[10] To be more precise, an inspection of the proof of Theorem 2 shows that the DR property alone removes *some* but not *all* terms of order $O_P(n^{-1}h^{-d})$ from an expansion of the estimator.

construction that ensures that $\widehat{\xi}_1 - \xi_1^o$ and $\widehat{\xi}_2 - \xi_2^o$ are asymptotically uncorrelated would deliver the same finding. For example, a trivial way to obtain fully independent estimation errors would be to split the data into two parts at random, and then calculate $\widehat{\xi}_1$ and $\widehat{\xi}_2$ from different subsamples. A result like in Theorem 2(b) can therefore in principle be obtained in any model in which a DR moment condition exists.[11] We discuss other alternatives to condition (3.4) in the context of specific examples where one of the nuisance functions is a density below.

Remark 9. We are not aware of an example of a semiparametric model in which there exists a DR moment condition of the form studied above but the orthogonality condition (3.4) fails. We therefore think of the result in Theorem 2(b) as the "standard" one.

4. Additional Applications

In this section, we study three additional examples in which semiparametric analogues of DR estimators have properties analogous to those obtained in Theorem 2(b): the partially linear regression model, a model for policy effects, and weighted average derivatives. In each case, the theoretical properties either follow from the same or largely similar arguments as those used in the proof of Theorem 2(b).

4.1. Partial Linear Model. Suppose that the data consist of a sample $\{Z_i\}_{i=1}^n = \{(Y_i, X_i, W_i)\}_{i=1}^n$ 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 $Y = \lambda^o(X) + W^\top \theta^o + \varepsilon$, where λ^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 has been extensively studied in the literature (e.g. Robinson, 1988; Donald and

[11] Of course, sample splitting is not very attractive from a practical point of view, and we do not mean to recommend this technique; but it would work for theoretical purposes here. 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).

Newey, 1994; Linton, 1995; Cattaneo, Jansson, and Newey, 2012) and 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). Let $\xi_1^o(x,\theta) = \mathbb{E}(Y - W^\top \theta | X = x)$ 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; and it follows from the model structure and the assumption that $\mathbb{E}(\varepsilon|X, W) = 0$ that the orthogonality condition (3.4) holds:

$$\begin{aligned}&\mathbb{E}((Y - W^\top \theta^o - \mathbb{E}(Y - W^\top \theta^o|X)) \cdot (W - \mathbb{E}(W|X))|X) \\ &\quad = \mathbb{E}(\mathbb{E}(\varepsilon|X, W) \cdot (W - \mathbb{E}(W|X))|X) = 0.\end{aligned}$$

To construct an STS-DR estimator of θ^o, we first estimate the two nuisance functions ξ_1^o and ξ_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

$$\begin{aligned}\widehat{\alpha}^{-i}(x) &= \operatorname*{argmin}_{\alpha} \sum_{j \neq i} (Y_j - \mathcal{P}_{l,\alpha}(X_j - x))^2 K_h(X_j - x), \\ \widehat{\beta}^{-i}(x) &= \operatorname*{argmin}_{\beta} \sum_{j \neq i} (W_j - \mathcal{P}_{l,\beta}(X_j - x))^2 K_h(X_j - x),\end{aligned}$$

and then putting

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

Note that to simplify the exposition the same order of the local polynomial l and bandwidth h are used in both local polynomial smoothing steps. We also define the sequences $b_n = h^{l+1}$ and $s_n = (\log(n)/(nh^d))^{1/2}$, which under the conditions that we impose below correspond to the (uniform) order of the bias and the stochastic part, respectively, of the two nonparametric estimators. The estimator $\widehat{\theta}_{DR}$ is then defined as the value of θ that solves the following equation:

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

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

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

which is identical (up to trimming terms) to the estimator proposed by Robinson (1988). We study its theoretical properties under the following assumptions.

Assumption PLM1. $H = \mathbb{E}((W - \xi_2^o(X))(W - \xi_2^o(X))^\top)$ is positive definite.

Assumption PLM2. (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 $\lambda^o(x)$ and $\xi_2^o(x)$ 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 $c > 4$.

Assumption PLM 1 is a full rank condition that ensures point identification of θ^o, and Assumption PLM2 is similar to Assumption G2 above. We expect $\widehat{\theta}_{DR}$ to have influence function and asymptotic variance

$$\phi_{PLM}(Z) = H^{-1}\psi_{PLM}(Z,\theta^o,\xi^o) \quad \text{and} \quad \Sigma_{PLM} = \mathbb{E}(\phi_{PLM}(Z)\phi_{PLM}(Z)^\top),$$

respectively. The following proposition gives a formal asymptotic normality result.

Proposition 1. *Suppose that Assumptions K1 and PLM1–PLM2 hold, and that h_1, h_2 are such that $b_n = o(n^{-1/4})$ and $s_n = o(n^{-1/6})$. Then $\sqrt{n}(\widehat{\theta}_{DR} - \theta^o) \xrightarrow{d} N(0,\Sigma_{PLM})$.*

The statement of the proposition is essentially analogous to that of Theorem 1 and Theorem 2(b) above. It is also similar to results obtained by Linton (1995) and Li (1996), who studied the higher-order properties of Robinson's estimator. Establishing a connection to DR estimation gives a new interpretation for the generally favorable properties Robinson's estimator.

4.2. Policy Effects. Suppose that the data consist of a sample $\{Z_i\}_{i=1}^n = \{(Y_i, X_i)\}_{i=1}^n$ from the distribution of $Z = (Y, X)$, where Y is a scalar dependent variable and X is a vector of continuous explanatory variables. 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 $\theta^o = \mathbb{E}(\mathbb{E}(Y|X = x)|_{x=\pi(X)})$. See e.g. Stock (1989), DiNardo, Fortin, and Lemieux (1996), Donald, Green, and Paarsch (2000), Gosling, Machin, and Meghir (2000), Rothe (2010, 2012, 2015), or Chernozhukov, Fernández-Val, and Melly (2013) for details and applications. Now let $\xi_1^o = (\xi_{11}^o, \xi_{12}^o)$, where $\xi_{11}^o(x) = \mathbb{E}(Y|X = x)$, $\xi_{12}^o(x) = \mathbb{E}(Y|X = \pi(x))$, and $\xi_2^o = (\xi_{21}^o, \xi_{22}^o)$, where $\xi_{21}^o(x)$ and $\xi_{22}^o(x)$ denote the densities of X and $\pi(X)$, respectively, at x. Then

$$\psi_{PE}(Z, \theta, \xi) = \xi_{12}(X) + (Y - \xi_{11}(X))\frac{\xi_{22}}{\xi_{21}(X)} - \theta$$

is a doubly robust moment function for estimating θ^o. This model differs slightly from the generic setup in Section 3 since ξ_1^o and ξ_2^o each have more than one component. It also differs in the sense that one of the nuisance functions is now a density instead of a conditional expectation.

To construct an STS-DR estimator, we can again estimate (ξ_{11}^o, ξ_{12}^o) by a "leave-one-out" local polynomial regression of order l_1 with bandwidth h_1. That is, we define

$$\widehat{\alpha}^{-i}(x) = \operatorname*{argmin}_{\alpha} \sum_{j \neq i} \left(Y_j - \mathcal{P}_{l_1,\alpha}(X_j - x)\right)^2 K_{h_1}(X_j - x),$$

and put

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

To estimate (ξ^o_{21}, ξ^o_{22}), we use standard "leave-one-out" kernel density estimators with bandwidth h_2, allowing a kernel function of order $l_2 + 1$ 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

$$\widehat{\xi}_{21}(X_i) = \frac{1}{n}\sum_{j\neq i} K^*_{h_2}(X_j - X_i) \quad \text{and} \quad \widehat{\xi}_{22}(X_i) = \frac{1}{n}\sum_{j\neq i} K^*_{h_2}(\pi(X_j) - X_i).$$

For $g = 1, 2$, we also define the sequences $b_{gn} = h_g^{l_g+1}$ and $s_{gn} = (\log(n)/(nh_g^d))^{1/2}$, which under the conditions that we impose below correspond to the (uniform) order of the bias and the stochastic part, respectively, of the nonparametric estimators $\widehat{\xi}_{gj}$, $j = 1, 2$. With this notation, our estimator of θ^o is then given by:

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

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

Assumption K 2. (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 PE 1. (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 $\xi^o_{21}(x)$ and $\xi^o_{22}(x)$ are bounded, $l_2 + 1$ times continuously differentiable, and bounded away from zero uniformly over $\mathcal{S}(X)$ and $\mathcal{S}(\pi(X))$, respectively; (iii) the function $\xi^o_{11}(x)$ is $l_1 + 1$ times continuously differentiable, and $\sup_{x\in\mathcal{S}(\pi(X))} \mathbb{E}(|Y|^c|X = x) < \infty$ for some constant $c > 2$.

Assumption K2 describes a kernel function of order $l_2 + 1$, which is used to control the asymptotic bias of the two density estimates. Assumption PE1 is again similar to Assumption G2 above. We expect $\widehat{\theta}_{DR}$ to have influence function and asymptotic variance

$$\phi_{PE}(Z) = \phi_{PE}(Z, \theta^o, \xi^o) \quad \text{and} \quad \Sigma_{PE} = \mathbb{E}(\phi_{PE}(Z)^2),$$

respectively. The following proposition gives a formal asymptotic normality result.

Proposition 2. *Suppose that Assumptions K1–K2 and PE1 hold, and that h_1, h_2 are such that $b_{1n}b_{2n} = o(n^{-1/2})$, $b_{gn} = o(n^{-1/6})$ and $s_{gn} = o(n^{-1/6})$ for $g = 1, 2$. Then $\sqrt{n}(\widehat{\theta} - \theta^o) \xrightarrow{d} N(0, \Sigma_{PE})$.*

The statement of the proposition is essentially analogous to that of Theorem 2(b), even though as explained above the setup is slightly different. Since one of the nuisance functions is a density here, equation (3.4) is clearly not applicable. Asymptotic non-correlation of $\widehat{\xi}_{1j} - \xi^o_{1j}$ and $\widehat{\xi}_{2j'} - \xi^o_{2j'}$ follows, roughly speaking, because (i) $\widehat{\xi}_{1j} - \xi^o_{1j}$ is essentially a kernel-weighted sum of the $\varepsilon_i = Y_i - \mathbb{E}(Y_i|X_i)$, (ii) $\widehat{\xi}_{2j'} - \xi^o_{2j'}$ is essentially a kernel weighted sum of functions of the X_i, and (iii) the ε_i are uncorrelated with any function of the X_i by the properties of conditional expectations. With this insight, one can prove the proposition similarly to Theorem 2(b).[12]

4.3. Weighted Average Derivatives. Suppose that the data consist of sample $\{Z_i\}_{i=1}^n = \{(Y_i, X_i)\}_{i=1}^n$ 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^o_2. Then the weighted average derivative (WAD) of the regression function $\mathbb{E}(Y|X = x)$ is defined as $\theta^o = \mathbb{E}(w(X)\nabla_x \mathbb{E}(Y|X = x)|_{x=X})$, where w is a known scalar weight function. WADs are important for estimating the coefficients in linear single-index models, and as a summary

[12] The fact that one of the nuisance functions is a density can easily be accommodated. The kernel density estimator proposed above has the same structure (or, actually, a much simpler one) as the Bahadur expansion of a local polynomial regression estimator, and only that structure is used in the proof of Theorem 2(b).

measure of nonparametrically estimated regression functions more generally (e.g. Stoker, 1986; Powell et al., 1989; Stoker, 1991; Newey and Stoker, 1993; Cattaneo et al., 2013).

Let $\xi^o_{11}(x) = \mathbb{E}(Y|X = x)$, denote the density of X by $\xi^o_{21}(x)$, and denote the vectors of partial derivatives of those two functions as $\xi^o_{12}(x) = \nabla_x \xi^o_{11}(x)$ and $\xi^o_{22}(x) = \nabla_x \xi^o_{21}(x)$, respectively. With this notation, we have that

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

is a DR moment function for estimating θ^o. Note that the need to estimate derivatives distinguishes this application from the other ones considered in this paper. 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 ξ^o_{11} and ξ^o_{12} 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

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

for $g = 1, 2$, and put

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

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

To estimate the density function ξ^o_{21}, 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 ξ^o_{22} 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

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

Note that we use same l_g and h_g to estimate ξ^o_{1g} and ξ^o_{2g} for $g \in \{1, 2\}$ for notational simplicity only. We also define the sequences $b_{gn} = h_g^{l_g+1}$ for $g = 1, 2$, $s_{1n} = (\log(n)/(nh_1^d))^{1/2}$, and $s_{2n} = (\log(n)/(nh_2^{d+2}))^{1/2}$. These sequences are chosen such that b_{gn} and s_{gn} will correspond to the (uniform) order of the bias and the stochastic part for estimating ξ_{jg}, $j = 1, 2$. With this notation, our estimator of θ_o is given by:

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

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

Assumption K3. (i) $\mathcal{K}^*$ and $\mathcal{K}^{**}$ are twice continuously differentiable; (ii) $\int \mathcal{K}^*(u)du = 1$ and (ii) $\int \mathcal{K}^{**}(u)du = 1$; (iii) $\int u^k \mathcal{K}^*(u)du = 0$ for $k = 1, \ldots, l_1 + 1$ and $\int u^k \mathcal{K}^{**}(u)du = 0$ for $k = 1, \ldots, l_2 + 1$; (iv) $\int |u^2 \mathcal{K}^*(u)|du < \infty$ and $\int |u^2 \mathcal{K}^{**}(u)|du < \infty$; and (v) $\mathcal{K}^*(u) = \mathcal{K}^{**}(u) = 0$ for u not contained in some compact set, say $[-1, 1]$.

Assumption WAD1. (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) $\xi^o_{11}(x)$ 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 K3 describes a kernel functions of order $l_1 + 1$ and $l_2 + 1$; and Assumption WAD1 is a standard smoothness condition similar to Assumption G2. We expect $\widehat{\theta}_{DR}$

to have influence function and asymptotic variance

$$\phi_{WAD}(Z) = \phi_{WAD}(Z, \theta^o, \xi^o) \quad \text{and} \quad \Sigma_{WAD} = \mathbb{E}(\phi_{WAD}(Z)\phi_{WAD}(Z)^\top),$$

respectively. Note that Σ_{WAD} is the semiparametric efficiency bound for estimating θ^o in this model (Newey and Stoker, 1993). The following proposition gives a formal asymptotic normality result.

Proposition 3. *Suppose that Assumptions K1, K3 and WAD1 hold, and that h_1, h_2 are such that $b_{1n}b_{2n} = o(n^{-1/2})$, $b_{gn} = o(n^{-1/6})$ and $s_{gn} = o(n^{-1/6})$ for $g = 1, 2$. Then $\sqrt{n}(\widehat{\theta} - \theta^o) \xrightarrow{d} N(0, \Sigma_{WAD})$.*

The statement of the proposition is essentially analogous to that of Theorem 1 and Theorem 2(b). Even though the estimator $\widehat{\theta}_{DR}$ 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, under the conditions of the proposition the difference between $\widehat{\theta}_{DR}$ and the efficient linear representation is of smaller order than that of various other efficient estimators discussed in Stoker (1991), or that of the jackknife bias corrected estimator in Cattaneo et al. (2013).

5. Concluding Remarks

In this paper, we have explored the possibility of constructing semiparametric two-step estimators as analogues of standard doubly robust estimators. We have shown that in the context of semiparametric missing data models such STS-DR estimators have favorable theoretical and practical properties relative to other commonly used STS estimators such as Inverse Probability Weighting. These results suggest that STS-DR estimators should become part of the standard toolkit of applied researchers working with these kinds of models. We have also shown that the DR property alone is unable to generate estimators with similarly favorable properties. Instead, it needs to be combined with an orthogonality condition on the estimation residuals from the nonparametric first stage. We also studied STS-DR estimation in three other models that satisfy these criteria. This analysis provided a new interpretation of Robinson's estimator for the partially linear model, and suggests that it should be advisable to combine estimates of conditional expectations and densities to estimate policy effects and weighted average derivatives.

A. Proofs of Main Results

In this appendix, we give the proofs of Theorem 1–2 and Proposition 1–3. Since Theorem 1 and Theorem 2(b) follow from essentially the same arguments, we only give a detailed account of the derivation of Theorem 2(b). Our proof of Theorem 1 then simply proceeds by arguing that the missing data model is a special case of the setup of Theorem 2(b). The derivations of Proposition 1–3 are sketched along similar lines. We begin by stating two auxiliary results about the rate of convergence of U-Statistics and a certain expansion of the local polynomial regression estimator that are used repeatedly in this Appendix.

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

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

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

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

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

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

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

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

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

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

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

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

A.2. Stochastic Expansion of the Local Polynomial Estimator. Our proofs use a particular stochastic expansion of the local polynomial regression estimators $\widehat{\xi}_g$. 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.

$\tau_s(1) = (0, \dots, 0, s)$, $\dots$, $\tau_s(n_s) = (s, 0, \dots, 0)$. For each $s \in \{0, 1, \dots, l_g\}$ 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

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

To better understand this notation, note that for the simple case that $l_g = 0$, i.e. when $\widehat{\xi}_g$ 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 $\widehat{\xi}_g(U_{gi})$ as

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

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 2. *Under Assumptions K1 and G1–G2 the following statements hold for $g \in \{1, 2\}$ if $h_g \to 0$ and $\log(n)/(nh_g^{d_g}) \to 0$ as $n \to \infty$:*

(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. Follows from well-known arguments as in e.g. Masry (1996) or Kong et al. (2010). □

A.3. Proof of Theorem 2(b). Having stated the auxiliary results, we now turn to the proof of Theorem 2(b). We give a proof of part (a) below. For notational simplicity, we drop the “DR” subscript on our estimator and write $\widehat{\theta}$ instead of $\widehat{\theta}_{DR}$. It is straightforward to show that $\widehat{\theta} \xrightarrow{p} \theta^o$ under either condition (a) or (b), and thus we omit the details of proving this step. Now it follows from the differentiability of ψ with respect to θ and the definition of $\widehat{\theta}$ that

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

for some θ^* between θ^o and $\widehat{\theta}$, and $H_n(\theta, \xi) = \sum_{i=1}^{n} \partial_\theta \psi(Z_i\theta, \xi_1(U_{1i}), \xi_2(U_{2i}))$. It then follows from standard arguments that $H_n(\theta^*, \widehat{\xi}) = H + o_P(1)$. Next, we consider an expansion of the term

$$\Psi_n(\theta^o, \widehat{\xi}) = n^{-1} \sum_{i=1}^{n} \psi(Z_i, \theta^o, \widehat{\xi}_1(U_{1i}), \widehat{\xi}_2(U_{2i})).$$

Using the notation that

$$\begin{aligned}
&\psi_i^1 = \partial\psi(Z_i, \theta^o, t, \xi_2^o(U_{2i}))/\partial t|_{t=\xi_1^o(U_i)}, \quad \psi_i^{11} = \partial^2\psi(Z_i, \theta^o, t, \xi_2^o(U_{2i}))/\partial t|_{t=\xi_1^o(U_i)}, \\
&\psi_i^2 = \partial\psi(Z_i, \theta^o, \xi_1^o(U_i), t)/\partial t|_{t=\xi_2^o(U_{2i})}, \quad \psi_i^{22} = \partial^2\psi(Z_i, \theta^o, \xi_1^o(U_i), t)/\partial t|_{t=\xi_2^o(U_{2i})}, \text{ and} \\
&\psi_i^{12} = \partial^2\psi(Z_i, \theta^o, t_1, t_2)/\partial t_1 \partial t_2|_{t_1=\xi_1^o(U_i), t_2=\xi_2^o(U_{2i})},
\end{aligned}$$

we find that because of differentiability conditions on the moment function ψ we have that

$$\begin{aligned}
\Psi_n(\theta^o, \widehat{\xi}) - \Psi_n(\theta^o, \xi^o) = &\frac{1}{n} \sum_{i=1}^{n} \psi_i^1(\widehat{\xi}_1(U_{1i}) - \xi_1^o(U_{1i})) + \frac{1}{n} \sum_{i=1}^{n} \psi_i^2(\widehat{\xi}_2(U_{2i}) - \xi_2^o(U_{2i})) \\
&+ \frac{1}{n} \sum_{i=1}^{n} \psi_i^{11}(\widehat{\xi}_1(U_{1i}) - \xi_1^o(U_{1i}))^2 + \frac{1}{n} \sum_{i=1}^{n} \psi_i^{22}(\widehat{\xi}_2(U_{2i}) - \xi_2^o(U_{2i}))^2 \\
&+ \frac{1}{n} \sum_{i=1}^{n} \psi_i^{12}(\widehat{\xi}_1(U_{1i}) - \xi_1^o(U_{1i}))(\widehat{\xi}_2(U_{2i}) - \xi_2^o(U_{2i})) \\
&+ O_P(\|\widehat{\xi}_1 - \xi_1^o\|_\infty^3) + O_P(\|\widehat{\xi}_2 - \xi_2^o\|_\infty^3).
\end{aligned}$$

By Lemma 2(i), the two "cubic" remainder terms are both of the order $o_P(n^{-1/2})$ under the conditions of Theorem 2(b), and thus also under those of Theorem 2(a). In Lemma 3–5 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 Theorem 2(b). The asymptotic normality result then follows from a simple application of the Central Limit Theorem.

The proofs of the following Lemmas repeatedly use the result that the smoothness conditions on the moment function ψ combined with the DR property imply 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 $\xi_1^o + t\lambda_1 \in \Xi_1$ and $\xi_2^o + t\lambda_2 \in \Xi_2$ 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

$$\mathbb{E}(\psi_i^1 \lambda_1(U_{1i})) = \lim_{t \to 0} \frac{\Psi(\theta^o, \xi_1^o + t\lambda_1, \xi_2^o) - \Psi(\theta^o, \xi_1^o, \xi_2^o)}{t} = 0$$

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

Lemma 3. *Under the conditions of Theorem 2(b), the following statements hold:*

$$(i) \quad \frac{1}{n}\sum_{i=1}^n \psi^1(Z_i)(\widehat{\xi}_1(U_{1i}) - \xi_1^o(U_{1i})) = o_P(n^{-1/2}),$$
$$(ii) \quad \frac{1}{n}\sum_{i=1}^n \psi^2(Z_i)(\widehat{\xi}_2(U_{2i}) - \xi_2^o(U_{2i})) = o_P(n^{-1/2}).$$

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

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

and since the second term on the right-hand side of the previous equation is of the order $o_P(n^{-1/2})$ due to the restrictions on the bandwidth, 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 2 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 1 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

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

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

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

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 due to the restrictions on the bandwidth. This completes our proof. □

Remark 10. Without any restrictions on the structure of the moment condition, 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. The fact that DR moment condition has a zero functional derivative with respect to the nuisance functions is what removes this term here. Note however that there are also non-DR moment conditions with this property, such as those where the corresponding moment function is an influence function in the underlying semiparametric model.

Lemma 4. *Under the conditions of Theorem 2(b), the following statements hold:*

$$(i) \quad \frac{1}{n}\sum_{i=1}^n \psi_i^{11}(\widehat{\xi}_1(U_{1i}) - \xi_1^o(U_{1i}))^2 = o_P(n^{-1/2}),$$
$$(ii) \quad \frac{1}{n}\sum_{i=1}^n \psi_i^{22}(\widehat{\xi}_2(U_{2i}) - \xi_2^o(U_{2i}))^2 = o_P(n^{-1/2}).$$

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

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

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 $o_P(n^{-1/2})$ due to the restrictions on the bandwidth, 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 $k \in \{1, \dots, 6\}$. 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 equation A.1 and Lemma 1, and crude bounds for $k \in \{3, 6\}$ simply using the uniform rates derived in Lemma 2. First, for $k = 1$ we find that

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

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

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

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 (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 1 to show that

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

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 3 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 2:

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

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

Remark 11. 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. For a non-DR moment condition based on an influence function this term would not vanish.

Lemma 5. *Under the conditions of Theorem 2(b), the following statement holds:*

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i^{12}(\widehat{\xi}_1(U_{1i}) - \xi_1^o(U_{1i}))(\widehat{\xi}_2(U_{2i}) - \xi_2^o(U_{2i})) = o_P(n^{-1/2}).$$

Proof. By Lemma 2, one can see that uniformly over $u = (u_1, u_2)$ we have that

$$\begin{aligned}
&(\widehat{\xi}_1(u_1) - \xi_1^o(u_1))(\widehat{\xi}_2(u_2) - \xi_2^o(u_2)) \\
&= \sum_{k=1}^{9} T_{n,k}(u) + O_P\left(\left(\frac{\log(n)}{nh_1^{d_1}}\right)^{3/2}\right)\left(O_P(h_2^{l_2+1}) + O_P\left(\frac{\log(n)}{nh_2^{d_2}}\right)\right) \\
&+ O_P\left(\left(\frac{\log(n)}{nh_2^{d_2}}\right)^{3/2}\right)\left(O_P(h_1^{l_1+1}) + O_P\left(\frac{\log(n)}{nh_1^{d_1}}\right)\right)
\end{aligned}$$

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 $o_P(n^{-1/2})$ due to the restrictions on the bandwidth, 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 4, 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 2. First, arguing as in the proof of Lemma 3 and 4 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. Lemma 1 and some variance calculations then imply 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 3 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 $k=5$. This term is the only one for which we exploit the condition (3.4). 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

$$T_{n,5,A} = \frac{1}{n^3} \sum_i \sum_{j \neq i} \psi_i^{12} (e_1^\top N_{1,n}(U_{1i})^{-1} w_{1j}(U_{1i}) K_{h_1}(X_{1j} - U_{1i}) \varepsilon_{1j})$$

$$\cdot (e_1^\top N_{2,n}(U_{2i})^{-1} w_{2j}(U_{2i}) K_{h_2}(X_{2j} - U_{2i}) \varepsilon_{2j}),$$

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

$$\cdot e_1^\top N_{2,n}(U_{2i})^{-1} w_{2l}(U_{2i}) K_{h_2}(X_{2l} - U_{2i}) \varepsilon_{2l}.$$

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 1 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. Since this step of the proof is important, we are providing some more details about this calculation. Let $\lambda_{ijl} = K_{h_1}(X_{1j} - U_{1i}) \varepsilon_{1j} K_{h_2}(X_{2l} - U_{2i}) \varepsilon_{2l}$. It is easy to see that the variance of $\tilde{T}_{n,5,B} = n^{-3} \sum_i \sum_{j \neq i} \sum_{l \neq i,j} \lambda_{ijl}$ is of the same order as $T_{n,5,B}$, and thus we focus on the former. Define $\mathcal{Z}_i = (U_{1i}, U_{2i})$, $\mathcal{Z}_j = (X_{1j}, \varepsilon_{1j})$ and $\mathcal{Z}_l = (X_{1l}, \varepsilon_{1l})$. It is easy to see that for distinct values of i, j and l we have that $\mathbb{E}(\lambda_{ijl}) = \mathbb{E}(\lambda_{ijl} | \mathcal{Z}_i) = \mathbb{E}(\lambda_{ijl} | \mathcal{Z}_j) = \mathbb{E}(\lambda_{ijl} | \mathcal{Z}_l) = 0$, and thus $\tilde{T}_{n,5,B}$ has mean zero and first-order degenerate kernel. It also holds that $\mathbb{E}(\lambda_{ijl} | \mathcal{Z}_i, \mathcal{Z}_j) = \mathbb{E}(\lambda_{ijl} | \mathcal{Z}_i, \mathcal{Z}_l) = 0$. Using the notation from Lemma 1, we thus have that

$$\begin{aligned} \rho_{n,2}^2 &= \mathrm{Var}(\mathbb{E}(\lambda_{ijl} | \mathcal{Z}_l, \mathcal{Z}_j)) \\ &= \mathrm{Var}\left(\int K_{h_1}(X_{1j} - u_1) K_{h_2}(X_{2l} - u_2) f_U(u_1, u_2) du_1 du_2 \varepsilon_{1j} \varepsilon_{2l} \right). \end{aligned}$$

The order of $\rho_{n,2}$ thus depends on the number of joint components of U_1 and U_2, or, more precisely, the effective dimension of the support of (U_1, U_2). The "best case" would be that (U_1, U_2) has effective support of dimension $d_1 + d_2$, in which case $\rho_{n,2} = O(1)$. The "worst case" would be that $U_1 = U_2$, in which case $\rho_{n,2} = O(\max\{h_1^{-d_1/2}, h_2^{-d_2/2})$. This "worst case" bound is sufficient for our purposes. Now consider

$$\begin{aligned}
\rho_{n,3}^2 &= \mathrm{Var}(\mathbb{E}(\lambda_{ijl} | \mathcal{Z}_i, \mathcal{Z}_j, \mathcal{Z}_l)) \\
&= \mathbb{E}\left(h_1^{-2d_1} K((X_{1j} - U_{1i})/h_1)^2 h_2^{-2d_2} K((X_{2l} - U_{2i})/h_2)^2 \varepsilon_{1j}^2 \varepsilon_{2l}^2\right) \\
&= \mathbb{E}\Bigg(\int h_1^{-2d_1} K((x_1 - U_{1i})/h_1)^2 \sigma_1^2(x_1) f_{X_1}(x_1) dx_1 \\
&\qquad\qquad \cdot \int h_2^{-2d_2} K((x_2 - U_{2i})/h_2)^2 \sigma_2^2(x_2) f_{X_2}(x_2) dx_2\Bigg) \\
&= \mathbb{E}\Bigg(\int h_1^{-d_1} K(t_1)^2 \sigma_1^2(U_{1i} + t_1 h_1) f_{X_1}(U_{1i} + t_1 h_1) dt_1 \\
&\qquad\qquad \cdot \int h_2^{-d_2} K(t_2)^2 \sigma_2^2(U_{2i} + t_2 h_2) f_{X_2}(U_{2i} + t_2 h_2) dt_2\Bigg) \\
&= O(h_1^{-d_1} h_2^{-d_2}).
\end{aligned}$$

From Lemma 1 we then obtain the desired result that

$$\begin{aligned}
\tilde{T}_{n,5,B} &= O_P(n^{-1}\rho_{n,2}) + O_P(n^{-3/2}\rho_{n,3}) \\
&= O_P(n^{-1} \max\{h_1^{-d_1/2}, h_2^{-d_2/2}\}) + O_P(n^{-3/2} h_1^{-d_1/2} h_2^{-d_2/2}),
\end{aligned}$$

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 (3.4) 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 similar to those detailed above, we obtain the bound that

$$\begin{aligned}T_{n,5,A} &= n^{-1}\cdot O_P(n^{-1/2}\max\{h_1^{-d_1}, h_2^{-d_2}\} + O_P(n^{-1}\max\{h_1^{-d_1/2}h_2^{-d_2}, h_1^{-d_1}h_2^{-d_2/2}\}))\\ &= n^{-1/2}O_P(\max\{nh_1^{-d_1}, nh_2^{-d_2}\} + \max\{n^{-1/2}h_1^{-d_1/2}nh_2^{-d_2}, nh_1^{-d_1}n^{-1/2}h_2^{-d_2/2}\})\\ &= o_P(n^{-1/2})\end{aligned}$$

under our restrictions on the bandwidths (in fact, our restrictions imply the much stronger result that $T_{n,5,A} = O_P(n^{-5/6})$). We are again providing some more details about this calculation. Let $\lambda_{ij} = K_{h_1}(X_{1j} - U_{1i})\varepsilon_{1j}K_{h_2}(X_{2j} - U_{2i})\varepsilon_{2j}$. It is easy to see that $\tilde{T}_{n,5,A} = n^{-3}\sum_i\sum_{j\neq i}\lambda_{ij}$ is of the same order as $T_{n,5,A}$, and thus we focus on the former. Define $\mathcal{Z}_i = (U_{1i}, U_{2i})$ and $\mathcal{Z}_j = (X_{1j}, X_{2j}, \varepsilon_{1j}, \varepsilon_{2j})$. We then have that for $i \neq j$

$$\begin{aligned}\mathbb{E}(\lambda_{ij}) &= \mathbb{E}[\mathbb{E}(\varepsilon_{1j}\varepsilon_{2j}|\mathcal{Z}_i, X_j)K_{h_1}(X_{1j} - U_{1i})K_{h_2}(X_{2j} - U_{2i})]\\ &= \mathbb{E}[\mathbb{E}(\varepsilon_{1j}\varepsilon_{2j}|X_j)K_{h_1}(X_{1j} - U_{1i})K_{h_2}(X_{2j} - U_{2i})] = 0,\end{aligned}$$

where the last equality follows from (3.4) and the fact that the data are i.i.d. Hence $\tilde{T}_{n,5,A}$ is mean zero. We also clearly have that $\mathbb{E}(\lambda_{ij}|\mathcal{Z}_i) = 0$. Using notation from Lemma 1, it then follows that

$$\begin{aligned}\rho_{n,1}^2 &= \mathrm{Var}(\mathbb{E}(\lambda_{ij}|\mathcal{Z}_j))\\ &= \mathrm{Var}(\int K_{h_1}(X_{1j} - u_1)K_{h_2}(X_{2j} - u_2)f_U(u_1, u_2)du_1du_2\varepsilon_{1j}\varepsilon_{2j})\end{aligned}$$

The order of $\rho_{n,1}$ thus depends on the number of joint components of U_1 and U_2, and of X_1 and X_2; or, more precisely, the effective dimension of the support of (U_1, U_2) and (X_1, X_2). The "best case" would be that (U_1, U_2) has effective support of dimension $d_1 + d_2$, in which case $\rho_{n,1} = O(1)$. The "worst case" would be that $U_1 = U_2$ and $X_1 = X_2$, in which case $\rho_{n,1} = O(\max\{h_1^{-d_1}, h_2^{-d_2})$. This "worst case" bound is sufficient for our purposes. Now consider

$$\begin{aligned}\rho_{n,2}^2 &= \mathrm{Var}(\mathbb{E}(\lambda_{ij}|\mathcal{Z}_l, \mathcal{Z}_j)) \\ &= \mathbb{E}\left(h_1^{-2d_1}K((X_{1j}-U_{1i})/h_1)^2 h_2^{-2d_2}K((X_{2j}-U_{2i})/h_2)^2\varepsilon_{1j}^2\varepsilon_{2j}^2\right).\end{aligned}$$

Under the "worst case" scenario that that $U_1 = U_2$ and $X_1 = X_2$ we then find that

$$\rho_{n,2}^2 = O(\max\{h_1^{-d_1}h_2^{-2d_2}, h_1^{-2d_1}h_2^{-d_2}\}).$$

From Lemma 1 we then obtain the desired result that

$$\begin{aligned}\tilde{T}_{n,5,A} &= n^{-1}\left(O_P(n^{-1/2}\rho_{n,1}) + O_P(n^{-1}\rho_{n,2})\right) \\ &= O_P(n^{-3/2}\max\{h_1^{-d_1}, h_2^{-d_2}\}) + O_P(n^{-2}\max\{h_1^{-d_1/2}h_2^{-d_2}, h_1^{-d_1}h_2^{-d_2/2}\}).\end{aligned}$$

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

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

The statement of the Lemma then follows from the restrictions on the bandwidth. This completes our proof. □

Remark 12. The derivation of the order of the term $T_{n,5,A}$ is the only step in our proof that requires the orthogonality condition (3.4). 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.4. Proof of Theorem 2(a). This result can be shown by following the steps of the proof of Theorem 2(b), and adapting the argument as indicated in the Remarks 10–12.

A.5. Proof of Theorem 1. The estimator has the same structure as the one studied in Theorem 2(b), and thus the statement follows from the same kind of arguments. Note that the condition (3.4) is satisfied here for reasons already explained in the main part of the paper.

A.6. Proof of Proposition 1. The estimator has the same structure as the one studied in Theorem 2(b), and thus the statement of the proposition follows from the same kind of arguments. See also Linton (1995) for a similar derivation.

A.7. Proof of Proposition 2. The main difference between this estimator and the one studied in Theorem 2 is that only one of the nuisance functions is a conditional expectation, whereas the other is a density function. This actually simplifies the problem, as a stochastic expansion of the kind given in Lemma 2 is easier to obtain of a substantially simpler form for kernel density estimators relative to the local polynomial estimator. In particular, it is easy to see that under the conditions of the proposition we have that

$$\widehat{\xi}_{2s}(X_i) = \xi^o_{2s}(X_i) + B_{sn}(X_i) + S_{sn}(X_i) \text{ for } s = 1, 2.$$

Here $B_{1n}(X_i) = \mathbb{E}(K_h(X - X_i)|X_i) = O(h^{l+1})$ is a deterministic bias function and $S_{1n}(X_i) = \sum_{j\neq i}(K_h(X_j - X_i) - \mathbb{E}(K_h(X - X_i)|X_i))/n = O_P((nh^d/\log(n))^{-1/2})$ is a mean zero stochastic term; and the terms B_{2n} and S_{2n} are defined analogously. The proof then follows from using the same arguments as in the one of Theorem 2(b), but using this simpler expansion of the kernel density estimator.

A.8. Proof of Proposition 3. This estimator differs from the one studied in Theorem 2 in that only one of the nuisance functions is a conditional expectation, whereas the other is a density function. Moreover, these two nuisance function enter the moment function through both their levels and their first order derivatives. The proof then follows from using the same arguments as in the one of Theorem 2(b), using an expansion for the estimate of the derivative of the conditional expectation function function similar to the one in Lemma 2. See Kong et al. (2010) for such a result. The estimates of the density and its derivative can be handled as described in the sketch of the proof of Proposition 2.

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 and Economic Statistics*, 13, 225–235.

Bang, H. and J. M. Robins (2005): "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–973.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2014a): "Program evaluation with high-dimensional data," *arXiv preprint arXiv:1311.2645*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650.

Bickel, P. J. and Y. Ritov (2003): "Nonparametric estimators which can be" plugged-in"," *Annals of Statistics*, 1033–1053.

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," *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.

Cattaneo, M. and M. Jansson (2014): “Bootstrapping Kernel-Based Semiparametric Estimators,” *Working Paper.*

Cattaneo, M., M. Jansson, and W. Newey (2012): “Alternative asymptotics and the partially linear model with many regressors,” *Working Paper.*

Cattaneo, M. D., R. K. Crump, and M. Jansson (2014): “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.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): “Inference on counterfactual distributions,” *Econometrica*, 81, 2205–2268.

DiNardo, J., N. Fortin, and T. Lemieux (1996): “Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach,” *Econometrica*, 64, 1001–1044.

Donald, S., D. Green, and H. Paarsch (2000): “Differences in wage distributions between Canada and the United States: An application of a flexible estimator of distribution functions in the presence of covariates,” *Review of Economic Studies*, 67, 609–633.

Donald, S. G. and W. Newey (1994): “Series estimation of semilinear models,” *Journal of Multivariate Analysis*, 50, 30–40.

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.

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.

Farrell, M. H. (2014): “Robust inference on average treatment effects with possibly more covariates than observations,” *arXiv preprint arXiv:1309.4686.*

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

Frölich, M., M. Huber, and M. Wiesenfarth (2015): “The finite sample performance of semi-and nonparametric estimators for treatment effects and policy evaluation,” *Working Paper.*

Gosling, A., S. Machin, and C. Meghir (2000): “The Changing Distribution of Male Wages in the UK,” *Review of Economic Studies*, 67, 635–666.

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.

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

Heckman, J., H. Ichimura, and P. Todd (1998): “Matching as an econometric evaluation estimator,” *Review of Economic Studies*, 65, 261–294.

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.

Ichimura, H. and W. Newey (2015): “The Influence Function of Semiparametric Estimators,” *Working Paper.*

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

Imbens, G., W. Newey, and G. Ridder (2007): "Mean-square-error calculations for average treatment effects," *Working Paper*.

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

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.

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. and T. 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.

Powell, J. L. and T. M. Stoker (1996): "Optimal bandwidth choice for density-weighted averages," *Journal of Econometrics*, 75, 291–316.

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.

——— (2015): "Decomposing the composition effect," *Journal of Business and Economic Statistics*, 33, 323–337.

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.

SHEN, X. ET AL. (1997): "On methods of sieves and penalization," *Annals of Statistics*, 25, 2555–2591.

SLOCZYNSKI, T. AND J. WOOLDRIDGE (2014): "A General Double Robustness Result for Estimating Average Treatment Effects," *Working Paper*.

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.

——— (2010): "Bounded, efficient and doubly robust estimation with inverse weighting," *Biometrika*, 97, 661–682.

Tsiatis, A. (2007): *Semiparametric theory and missing data*, Springer Science & Business Media.

van der Laan, M. and R. Daniel (2006): "Targeted maximum likelihood learning," *The International Journal of Biostatistics*, 2, 1–40.

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.

Table 1: Simulation results for DR: mean-squared-error, absolute bias, variance and empirical coverage rate of confidence intervals for various nonparametric first-stage estimators and smoothing parameters

		n×MSE						**$\sqrt{\text{n}}$×BIAS**						**n×VAR**						**Coverage Rate**					
$\widehat{\theta}_{DR-K}$	h_1/h_2	**.05**	**.08**	**.13**	**.20**	**.32**	**.50**	**.05**	**.08**	**.13**	**.20**	**.32**	**.50**	**.05**	**.08**	**.13**	**.20**	**.32**	**.50**	**.05**	**.08**	**.13**	**.20**	**.32**	**.50**
	.05	.215	.206	.204	.203	.203	.203	.004	.003	.002	.003	.003	.004	.215	.206	.204	.203	.203	.203	.95	.94	.94	.94	.94	.93
	.08	.212	.205	.203	.202	.202	.202	.009	.007	.006	.007	.007	.007	.212	.205	.203	.202	.202	.202	.95	.94	.94	.94	.94	.93
	.13	.212	.205	.203	.202	.201	.200	.007	.006	.008	.011	.015	.017	.212	.205	.203	.201	.200	.200	.95	.94	.94	.94	.94	.94
	.20	.213	.206	.203	.201	.200	.200	.016	.010	.002	.012	.030	.045	.213	.206	.203	.201	.199	.198	.95	.94	.94	.94	.94	.94
	.32	.219	.209	.204	.201	.202	.207	.057	.041	.023	.009	.053	.092	.215	.207	.203	.201	.199	.198	.95	.95	.95	.95	.94	.93
	.50	.227	.213	.206	.202	.205	.217	.089	.065	.039	.006	.072	.130	.219	.209	.204	.202	.200	.200	.96	.96	.96	.96	.94	.93
$\widehat{\theta}_{DR-OS}$	h_1/h_2	**1**	**2**	**3**	**4**	**7**	**10**	**1**	**2**	**3**	**4**	**7**	**10**	**1**	**2**	**3**	**4**	**7**	**10**	**1**	**2**	**3**	**4**	**7**	**10**
	1	.263	.205	.206	.206	.206	.209	.247	.027	.007	.002	.002	.006	.202	.204	.206	.206	.206	.209	.90	.95	.96	.96	.96	.96
	2	.200	.201	.203	.203	.204	.205	.046	.012	.004	.002	.001	.003	.198	.200	.203	.203	.204	.205	.93	.94	.94	.94	.94	.94
	3	.203	.202	.203	.203	.205	.205	.013	.000	.008	.004	.003	.007	.202	.202	.203	.203	.205	.205	.93	.94	.94	.94	.94	.94
	4	.203	.203	.203	.203	.204	.205	.008	.001	.005	.002	.003	.007	.203	.203	.203	.203	.204	.205	.93	.94	.94	.94	.94	.94
	7	.206	.206	.206	.206	.207	.207	.001	.001	.001	.001	.001	.003	.206	.206	.206	.206	.207	.207	.93	.93	.93	.93	.93	.93
	10	.218	.218	.218	.217	.217	.217	.004	.004	.004	.004	.004	.004	.217	.217	.217	.217	.217	.217	.92	.93	.93	.93	.93	.93
$\widehat{\theta}_{DR-SP}$	h_1/h_2	**.50**	**.65**	**.80**	**.95**	**1.10**	**1.25**	**.50**	**.65**	**.80**	**.95**	**1.10**	**1.25**	**.50**	**.65**	**.80**	**.95**	**1.10**	**1.25**	**.50**	**.65**	**.80**	**.95**	**1.10**	**1.25**
	.50	.212	.210	.209	.210	.209	.209	.005	.004	.004	.004	.004	.004	.212	.210	.209	.210	.209	.209	.92	.93	.93	.93	.92	.92
	.65	.209	.206	.205	.207	.205	.205	.005	.004	.003	.002	.002	.003	.209	.206	.205	.207	.205	.205	.93	.93	.93	.93	.93	.93
	.80	.207	.205	.204	.204	.203	.203	.005	.004	.001	.001	.000	.002	.207	.205	.204	.204	.203	.203	.93	.93	.93	.93	.93	.93
	.95	.207	.204	.204	.204	.202	.201	.005	.003	.000	.004	.004	.003	.206	.204	.204	.204	.202	.201	.94	.94	.94	.94	.93	.93
	1.10	.206	.204	.204	.204	.201	.200	.007	.003	.003	.005	.017	.050	.206	.204	.204	.204	.201	.198	.93	.94	.94	.94	.93	.93
	1.25	.209	.205	.204	.207	.203	.231	.043	.021	.010	.005	.050	.188	.207	.204	.204	.207	.201	.195	.94	.94	.95	.95	.94	.91

Results from 5,000 replications for first-stage kernel (K), orthogonal series (OS) and spline (SP) estimation. Outliers deviating from the simulation median by more than four times the interquartile range were removed for the computation of the summary statistics.

Table 2: Simulation results for IPW: mean-squared-error, absolute bias, variance and empirical coverage rate of confidence intervals for various nonparametric first-stage estimators and smoothing parameters

	$\mathbf{h_2}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_1}$)**					
					.05	**.08**	**.13**	**.20**	**.32**	**.50**
$\widehat{\theta}_{IPW-K}$	**.05**	0.489	0.398	0.330	0.83	0.83	0.82	0.83	0.84	0.87
	.08	0.373	0.329	0.265	0.87	0.87	0.87	0.87	0.89	0.90
	.13	0.323	0.289	0.239	0.89	0.88	0.88	0.88	0.90	0.91
	.20	0.268	0.189	0.232	0.91	0.91	0.91	0.91	0.91	0.93
	.32	0.229	0.022	0.229	0.93	0.92	0.92	0.92	0.93	0.94
	.50	0.288	0.245	0.228	0.88	0.88	0.88	0.88	0.88	0.89
	$\mathbf{h_2}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_1}$)**					
					.05	**.08**	**.13**	**.20**	**.32**	**.50**
$\widehat{\theta}_{IPW-TK}$	**.05**	0.639	0.282	0.560	0.83	0.80	0.84	0.92	0.99	0.79
	.08	0.934	0.020	0.934	0.76	0.73	0.77	0.89	0.99	0.70
	.13	0.788	0.205	0.747	0.76	0.73	0.77	0.90	0.99	0.67
	.20	0.646	0.176	0.615	0.81	0.79	0.83	0.94	0.99	0.74
	.32	0.397	0.354	0.272	0.90	0.88	0.92	0.98	1.00	0.86
	.50	0.252	0.153	0.229	0.94	0.93	0.94	0.98	1.00	0.91
	$\mathbf{h_2}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_1}$)**					
					1	**2**	**3**	**4**	**7**	**10**
$\widehat{\theta}_{IPW-OS}$	**1**	0.497	0.515	0.232	0.74	0.73	0.73	0.73	0.73	0.73
	2	0.244	0.063	0.240	0.93	0.91	0.91	0.91	0.91	0.91
	3	0.232	0.028	0.231	0.95	0.93	0.92	0.92	0.92	0.92
	4	0.221	0.009	0.221	0.95	0.93	0.93	0.93	0.92	0.92
	7	0.222	0.011	0.222	0.95	0.93	0.93	0.93	0.93	0.92
	10	0.236	0.028	0.236	0.94	0.92	0.92	0.92	0.92	0.91
	$\mathbf{h_2}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_1}$)**					
					.50	**.65**	**.80**	**.95**	**1.10**	**1.25**
$\widehat{\theta}_{IPW-SP}$	**.50**	0.359	0.339	0.244	0.82	0.82	0.82	0.83	0.83	0.84
	.65	0.252	0.170	0.223	0.89	0.90	0.90	0.90	0.90	0.91
	.80	0.226	0.077	0.220	0.91	0.92	0.92	0.92	0.92	0.93
	.95	0.233	0.025	0.232	0.91	0.92	0.92	0.92	0.92	0.93
	1.10	0.258	0.181	0.225	0.89	0.89	0.90	0.90	0.90	0.91
	1.25	0.613	0.633	0.212	0.64	0.65	0.65	0.65	0.65	0.65
	$\mathbf{h_2}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate**					
$\widehat{\theta}_{IPW-BS}$	**.05**	0.271	0.011	0.271	0.95					
	.08	0.301	0.062	0.298	0.96					
	.13	0.328	0.185	0.294	0.96					
	.20	0.275	0.156	0.251	0.95					
	.32	0.264	0.209	0.220	0.91					
	.50	0.512	0.552	0.207	0.74					

Results from 5,000 replications for first-stage kernel (K), twicing kernel (TK), orthogonal series (OS) and spline (SP), and bootstrap bias corrected kernel (BS) estimation. Outliers deviating from the simulation median by more than four times the interquartile range were removed for the computation of the summary statistics.

Table 3: Simulation results for REG: mean-squared-error, absolute bias, variance and empirical coverage rate of confidence intervals for various nonparametric first-stage estimators and smoothing parameters

	$\mathbf{h_1}$	**n×MSE**	$\sqrt{\mathbf{n}}$**×BIAS**	**n×VAR**	**Coverage Rate ($\mathbf{h_2}$)**					
					.05	**.08**	**.13**	**.20**	**.32**	**.50**
$\widehat{\theta}_{REG-K}$	**.05**	0.210	0.052	0.207	0.95	0.95	0.94	0.94	0.94	0.93
	.08	0.227	0.136	0.208	0.94	0.94	0.93	0.93	0.93	0.92
	.13	0.282	0.266	0.211	0.91	0.91	0.91	0.90	0.90	0.89
	.20	0.308	0.303	0.216	0.89	0.89	0.89	0.88	0.88	0.87
	.32	0.211	0.030	0.210	0.95	0.95	0.95	0.94	0.94	0.94
	.50	0.295	0.313	0.197	0.92	0.92	0.91	0.91	0.90	0.88
	$\mathbf{h_1}$	**n×MSE**	$\sqrt{\mathbf{n}}$**×BIAS**	**n×VAR**	**Coverage Rate ($\mathbf{h_2}$)**					
					.05	**.08**	**.13**	**.20**	**.32**	**.50**
$\widehat{\theta}_{REG-TK}$	**.05**	0.926	0.021	0.925	0.84	0.83	0.83	0.83	0.84	0.82
	.08	0.643	0.011	0.643	0.87	0.86	0.86	0.87	0.87	0.86
	.13	0.856	0.152	0.833	0.90	0.89	0.89	0.89	0.89	0.88
	.20	3.713	0.496	3.467	0.82	0.81	0.81	0.82	0.83	0.81
	.32	8.592	0.281	8.514	0.99	0.99	0.98	0.99	0.99	0.99
	.50	0.228	0.157	0.203	0.95	0.94	0.93	0.94	0.94	0.93
	$\mathbf{h_1}$	**n×MSE**	$\sqrt{\mathbf{n}}$**×BIAS**	**n×VAR**	**Coverage Rate ($\mathbf{h_2}$)**					
					1	**2**	**3**	**4**	**7**	**10**
$\widehat{\theta}_{REG-OS}$	**1**	0.612	0.652	0.187	0.67	0.73	0.75	0.75	0.75	0.75
	2	0.204	0.081	0.197	0.93	0.94	0.94	0.94	0.94	0.94
	3	0.203	0.016	0.203	0.93	0.94	0.94	0.94	0.94	0.94
	4	0.203	0.009	0.203	0.93	0.93	0.94	0.94	0.94	0.94
	7	0.206	0.001	0.206	0.93	0.93	0.93	0.93	0.93	0.93
	10	0.218	0.004	0.217	0.92	0.93	0.93	0.93	0.93	0.93
	$\mathbf{h_1}$	**n×MSE**	$\sqrt{\mathbf{n}}$**×BIAS**	**n×VAR**	**Coverage Rate ($\mathbf{h_2}$)**					
					.50	**.65**	**.80**	**.95**	**1.10**	**1.25**
$\widehat{\theta}_{REG-SP}$	**.50**	0.209	0.004	0.209	0.93	0.93	0.93	0.93	0.93	0.92
	.65	0.205	0.003	0.205	0.93	0.93	0.93	0.93	0.93	0.93
	.80	0.203	0.003	0.203	0.93	0.94	0.93	0.94	0.93	0.93
	.95	0.201	0.006	0.200	0.94	0.94	0.94	0.94	0.94	0.93
	1.10	0.203	0.091	0.195	0.94	0.94	0.94	0.94	0.94	0.93
	1.25	0.331	0.381	0.186	0.86	0.87	0.87	0.87	0.86	0.84
	$\mathbf{h_1}$	**n×MSE**	$\sqrt{\mathbf{n}}$**×BIAS**	**n×VAR**	**Coverage Rate**					
$\widehat{\theta}_{REG-BS}$	**.05**	0.210	0.050	0.207	0.94					
	.08	0.227	0.134	0.209	0.93					
	.13	0.282	0.267	0.211	0.90					
	.20	0.311	0.309	0.216	0.89					
	.32	0.212	0.038	0.210	0.94					
	.50	0.290	0.305	0.197	0.87					

Results from 5,000 replications for first-stage kernel (K), twicing kernel (TK), orthogonal series (OS) and spline (SP), and bootstrap bias corrected kernel (BS) estimation. Outliers deviating from the simulation median by more than four times the interquartile range were removed for the computation of the summary statistics.