Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically

Christoph Rothe and Sergio Firpo

Abstract

An estimator of a finite-dimensional parameter is said to be doubly robust (DR) 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 versions of such estimators that use local polynomial smoothing for estimating the nuisance functions. We show that such semiparametric two-step (STS) versions of DR estimators have favorable theoretical and practical properties relative to other commonly used STS estimators. We also show that these gains are not generated by the DR property alone. Instead, it needs to be combined with an orthogonality condition on the estimation residuals from the nonparametric first stage, which we show to be satisfied in a wide range of models.

JEL Classification: C14, C21, C31, C51

Keywords: *Semiparametric two-step estimation, missing data, treatment effects, average derivatives, partial linear model, policy effects, double robustness*

First version: December 20, 2012. This version: December 20, 2017. Christoph Rothe, University of Mannheim, Department of Economics, D-68161 Mannheim, Germany. Email: rothe@vwl.uni-mannheim.de. Sergio Firpo, Insper Institute of Education and Research, R. Quata 300, Sao Paulo-SP, 04546-042, Brasil. E-Mail: firpo@insper.edu.br. An earlier working version of this paper was circulated under the title "Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions". We would like to thank 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. Christoph Rothe gratefully acknowledges financial support from German Scholars Organization & Carl-Zeiss-Stiftung. Sergio Firpo gratefully acknowledges financial support from CNPq-Brazil.

1. INTRODUCTION

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 (e.g. Scharfstein et al., 1999; Robins and Rotnitzky, 2001). Such estimators, which feature prominently in the literature on missing data and causal inference models, achieve their eponymous property by combining 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 $\theta^o = \mathbb{E}(Y^*)$. Assume that the data are missing at random, so $\mathbb{E}(D|Y^*, X) = \mathbb{E}(D|X)$, 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 one well-known 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] Alternative estimators of θ^o, which only require an estimate of one of the two nuisance functions, include the regression-adjustment (REG) or the Inverse Probability Weighting (IPW) estimator, defined as

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

[1]The construction (1.1) is generally referred to as an Augmented Inverse Probility Weighting (AIPW) estimator in the DR literature (Robins et al., 1994). There are other ways to construct estimators with a DR property in such setups (e.g. Wooldridge, 2007; Graham et al., 2012; Vermeulen and Vansteelandt, 2015). We focus on "AIPW-type" estimators in the following.

respectively. In general, the latter estimators are only consistent if the "working" parametric model for the respective nuisance function is correctly specified, whereas for $\widehat{\theta}_{DR}$ to be consistent it suffices that either of the two specifications is correct.

This paper considers the properties of estimators following a construction like (1.1) when the nuisance functions are estimated nonparametrically via local polynomial smoothing. Such estimators fall within the class of semiparametric two-stage (STS) estimators (e.g. Newey, 1994; Ai and Chen, 2003), and standard calculations show that under appropriate regularity conditions the influence function of such a STS-DR estimator is identical to that of STS versions of either $\widehat{\theta}_{REG}$ or $\widehat{\theta}_{IPW}$. In particular, all three estimators are asymptotically efficient. 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 model in which DR estimators exist – given that it involves additional computational costs, requires a second smoothing parameter, 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 that constructions like (1.1) enjoy of any use in a nonparametric setting where nuisance functions are always consistently estimated?

In this paper, we show that there are indeed strong reasons to prefer an STS version of $\widehat{\theta}_{DR}$ over estimators like $\widehat{\theta}_{REG}$ or $\widehat{\theta}_{IPW}$. We first study a general missing data model, which covers the example from above as a special case. Using local polynomial regression to estimate the nuisance functions, a standard linear expansion of STS versions of $\widehat{\theta}_{DR}$, $\widehat{\theta}_{REG}$ and $\widehat{\theta}_{IPW}$ produces the same first-order terms, but the corresponding remainder is substantially smaller for $\widehat{\theta}_{DR}$ than it is for the competing procedures. As a consequence, the otherwise commonly required restriction that the estimation error of the nonparametric component is of the order $o_P(n^{-1/4})$ is not necessary. STS-DR estimators can also have smaller first-order bias and second-order variance than other first-order equivalent STS estimators, and their 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 the STS version of $\widehat{\theta}_{DR}$ instead of one of the other estimators. Through simulations, we show that these theoretical predictions 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 STS-DR estimator. This is an important concern because many STS estimators, including STS versions of $\widehat{\theta}_{REG}$ and $\widehat{\theta}_{IPW}$ in models like the one described above, can be highly sensitive with respect to the implementation details of the nonparametric stage in finite samples (e.g. 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 terms are of "second order" 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, STS-DR estimators should be rather insensitive to variation in smoothing parameters, which is confirmed by our simulations. We also show that in many cases the bandwidth can be selected via cross-validation.

As a second contribution, we analyze which particular features of the semiparametric missing data model are responsible for the gain in performance of the STS-DR estimator relative to its competitors in order to clarify under which conditions we would expect to be able to construct estimators with analogous properties in other semiparametric models in which doubly robust procedures are known to exist. To answer this question, we study a simple class of STS estimators in a general model about which we essentially only assume

[2]While our theoretical results only cover kernel-based estimators for the nuisance functions, in our simulations we also consider other nonparametric estimation procedures, such as orthogonal series regression and smoothing splines. Our simulation results confirm that STS versions of the DR estimator have favorable finite sample properties relative to many other STS estimation approaches, irrespective of the type of nonparametric estimation procedure being used.

that it gives rise to a doubly robust moment condition.[3] Our main, somewhat surprising finding is that the DR property alone is actually 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. In the case of the missing data model, for example, this property follows from the assumption that the data being missing at random; and it follows through analogous arguments for a wide range of causal inference models that are similar in structure. We also show that this property is satisfied in three other semiparametric models in which DR estimators are known to exist: the partially linear regression model, a model for nonparametric policy analysis, and weighted average derivatives.

1.1. **Related Literature.** Two-step estimators that depend on nonparametrically estimated functions, such as densities or conditional expectations, feature prominently in a wide range of applications. General results for such estimators have been obtained for example by Goldstein and Messer (1992), Newey (1994), Newey and McFadden (1994), Andrews (1994), Chen et al. (2003), Chen and Shen (1998), Ai and Chen (2003), Ichimura and Lee (2010), Escanciano et al. (2014, 2016), and Mammen et al. (2016), 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 CLT to $(1/\sqrt{n})\sum_{i=1}^n \phi(Z_i)$

[3]Here a moment condition is said to be doubly robust 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.

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 $R_n(\widehat{\theta})$ for various estimators under various conditions. In order to have a point of reference for the findings presented in this paper it is useful to review some of them.[4] 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 using a bandwidth h.[5] It is well-known that under standard regularity conditions the nonparametric first-stage 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

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

Asymptotic linearity of $\widehat{\theta}$ thus requires a "small bias" and a "small variance" condition on the first step nonparametric estimator. Hall and Marron (1987), Powell et al. (1989) and Powell and Stoker (1996) show that if $\widehat{\theta}$ is a *linear* transformation of a "leave-one-out" kernel weighted average[6] 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

[4]Note that if $R_n(\widehat{\theta})$ vanishes faster for one estimator relative to another as the sample size increases, this does not necessarily imply that the asymptotic linear approximation to is more accurate for a particular finite sample sizefor for the one estimator relative to the other, as the constants associated with the rates are generally different for different estimators.

[5]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. Moreover, we note that d refers to the number of continuously distributed covariates. Discrete covariates can be accommodated via the usual frequency method without affecting the following results, or as in Racine and Li (2004).

[6]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.

the context of specific applications by Ichimura and Linton (2005) or Cattaneo et al. (2013). Newey et al. (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).[7] 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 (1997), Chen and Shen (1998) and Ai and Chen (2003) to more general classes of sieve estimators. In contrast, our findings imply that the magnitude of both second order terms is reduced for kernel-based STS versions of DR estimators 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. We also show that this result does not follow from the DR property alone.

STS versions of DR estimators have been used before in some papers. One example is Robinson's (1988) estimator of the parametric component of a partially linear model. In a causal inference context, Cattaneo (2010) proposes an STS-DR estimator, but does not

[7]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 tend to have poor finite sample properties for reasons not reflected by second-order asymptotic expansions. Due to their instability, they are hardly used in practice. A twicing kernel is a particular type of higher-order kernel, and thus the same comment applies.

prove that this approach has any formal advantages. The construction that leads to the DR property is explicitly exploited in Belloni et al. (2014), Farrell (2015) and Belloni et al. (2016) for causal inference in a very high-dimensional setting, where a LASSO-type estimator is used in the first stage. Of course, our article also builds on the extensive literature on doubly robust estimation. See Robins et al. (1994), Scharfstein et al. (1999), Robins and Rotnitzky (1995), Robins and Rotnitzky (2001), Bang and Robins (2005), Kang and Schafer (2007), Tan (2010) or Vermeulen and Vansteelandt (2015), among many others.

1.2. **Outline of the Paper.** The remainder of the paper is structured as follows. In the next section, we study an STS analogue of a DR estimator in a 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 estimation in a general class of models which gives rise to a DR moment condition. We 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. We then argue that this structure is also present in three other models in which DR estimators are known to exist: the partially linear model, a model for policy analysis, and weighted average derivatives. Finally, Section 4 concludes. Regularity conditions and proofs are collected in the Appendix.

2. ESTIMATION IN A 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; Chen et al., 2008), and causal inference models (e.g. Hahn, 1998; Hirano et al., 2003). We consider this setting because it is the one in which double robust estimation features most prominently in the literature.

2.1. **Model.** Suppose that the underlying full data are a sample from the distribution of $(Y^*, X) \in \mathbb{R} \times \mathbb{R}^d$, 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 $Y = DY^*$. 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 $\xi_2^o(x) = \mathbb{E}(D|X = x)$, and note that the propensity score is assumed to be bounded away from zero over the support of X. Next, define

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

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

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).

2.2. **Estimator and Main Result.** We consider estimating the parameter θ^o based on the moment condition that the efficient influence function's expectation is equal to zero. Since the matrix $\mathbb{E}(\nabla_\theta m(Y^*, X, \theta^o))$ is assumed to be of full rank, this is equivalent to considering the moment condition

$$\mathbb{E}(\psi_{MD}(Z, \theta, \xi^o)) = 0 \text{ if and only if } \theta = \theta^o,$$

where $Z = (Y, X')'$, $\xi^o = (\xi_1^o, \xi_2^o)$, and $\psi_{MD}(Z, \theta, \xi) = (D(m(Y, X, \theta) - \xi_1^o(X, \theta^o)))/\xi_2(X) + \xi_1^o(X, \theta)$. Given an estimate $\widehat{\xi}$ of ξ^o, our estimator $\widehat{\theta}_{DR}$ of θ^o is then given by the value of θ

that solves the equation

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

We refer to ψ_{MD} as a DR moment function, and to $\widehat{\theta}_{DR}$ as an STS-DR estimator in the following, as analogous estimators of θ^o based on parametric specifications for the two nuisance functions are known to be doubly robust.

However, we deviate from the usual DR literature since we obtain our estimates $\widehat{\xi}_1$ and $\widehat{\xi}_2$ of ξ_1^o and ξ_2^o by "leave-one-out" local polynomial regression (Fan, 1993; Ruppert and Wand, 1994).[8] This class of kernel-based smoothers 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 we use the notation that $|a|=\sum_{i=1}^d a_i$ and $a^b=\prod_{i=1}^d a_i^{b_i}$ for generic d-dimensional vectors a,b, and that $\sum_{0\leq|s|\leq l}$ 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

$$\widehat{\xi}_1(X_i,\theta)=e_1^\top\operatorname*{argmin}_{\alpha}\sum_{j\neq i}\left(m(Y_j,X_j,\theta)-\mathcal{P}_{l_1,\alpha}(X_j-X_i)\right)^2K_{h_1}(X_j-X_i)\mathbb{I}\{D_j=1\},$$

$$\widehat{\xi}_2(X_i)=e_1^\top\operatorname*{argmin}_{\beta}\sum_{j\neq i}\left(D_j-\mathcal{P}_{l_2,\beta}(X_j-X_i)\right)^2K_{h_2}(X_j-X_i),$$

with e_1 the first unit vector (of appropriate dimension). 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. 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 rate of convergence of the bias and the stochastic part, respectively, of our two nonparametric estimators (cf. Masry, 1996). We then obtain the following result about the

[8] We assume for simplicity that all d components of X are continuously distributed. Discrete covariates can be accommodated via the usual frequency method, or as in Racine and Li (2004).

asymptotic linearity of $\widehat{\theta}_{DR}$.

Theorem 1. *Under suitable regularity conditions (see Appendix A.1),* $\sqrt{n}(\widehat{\theta}_{DR} - \theta^o) = n^{-1/2}\sum_{i=1}^n \phi_{MD}(Z_i) + o_P(1) \xrightarrow{d} N(0, \Sigma_{MD})$ *if* 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* $g = 1, 2$.

2.3. **Discussion.** Theorem 1 differs from other asymptotic linearity results for STS estimators 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 $o_P(n^{-1/6})$. Asymptotic linearity of a STS estimator otherwise typically requires an $o_P(n^{-1/4})$ rate of convergence for the nonparametric component (e.g. Newey, 1994).

Weaker conditions are possible here because the difference between $\widehat{\theta}_{DR}$ and its asymptotically linear representation is particularly small. To see this, consider the case that $h_1 = h_2 \equiv h$ and $l_1 = l_2 \equiv l$ to simplify the exposition. By following the proof of Theorem 1 we find 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 interesting 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 Ichimura and Linton (2005) and Bravo and Jacho-Chávez (2010), who study a slightly simpler version of this estimator, 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 $\sqrt{n}(\widehat{\theta}_{IPW} - \theta^o) = n^{-1/2}\sum_{i=1}^{n}\phi_{MD}(Z_i) + o_P(1)$. These conditions are called for because

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

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. In nonparametric regression problems with a binary dependent variable, one can use local Probit or Logit models, as in Fan et al. (1995), instead of standard local polynomial regression in order to ensure that the final estimator is constrained to take values in the unit interval. We conjecture that the use of such alternative estimation techniques for the propensity score (or the regression function, in case of a binary outcome) would not change the substantive conclusions of Theorem 1; subject of course to a suitable adaptation of regularity conditions. This conjecture is based on results by Kong et al. (2010), which show that a large class of local M-estimators possesses a stochastic expansion that is analogous in

structure to the one of the local polynomial regression estimator we use in our proofs.

Remark 2. If one were to 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 Theorem 1. 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).

2.4. **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 $[0, 1]$.[9] 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.

While our theoretical results above only cover kernel-based estimation of nuisance functions, for our simulations we also other types of nonparametric first stage estimators. We also use a

[9]While real empirical applications typically involve several covariates, we confine ourselves to the case of single one for our simulations. This is because the finite-sample performance of the estimators that we consider does not so much depend on the number of covariates, but on the overall complexity of the functions that are being estimated. We refer to Frölich et al. (2015) and Naimi and Kennedy (2017) for extensive simulation studies of STS-DR estimators in program evaluation settings with multiple covariates. These studies find properties of STS-DR estimators that are qualitatively very similar to the simulation results in this paper. In the supplemental material, we also report simulation results for several modified versions of our DGP in which the density of the covariate is not bounded away from zero. Again, the results are qualitatively very similar to the ones reported here.

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.[10] For all combinations of nonparametric estimators and smoothing parameters, 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]

[10] 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.

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 first-order 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 (both the type of estimator 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 $CI^{1-\alpha}_{IPW-s}$ and $CI^{1-\alpha}_{REG-s}$. Note that these two confidence intervals by construction have the same length as $CI^{1-\alpha}_{DR-s}$ for each s and every value of the smoothing parameters, and only differ in the point at which they are centered.[11]

[11]From the practitioner's perspective, using STS-DR is not more costly computationally than, say, using the STS-REG or STS-IPW estimator 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.

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).[12] We also consider bootstrap-based confidence intervals for θ^o calculated using the usual percentile method. Note that the calculation of these confidence intervals does not involve estimating the asymptotic variance, and thus do not depend on a second smoothing parameter.

[TABLES 2 AND 3 ABOUT HERE]

Our simulation results 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. In particular, their finite-sample distribution 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 5 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

[12]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.

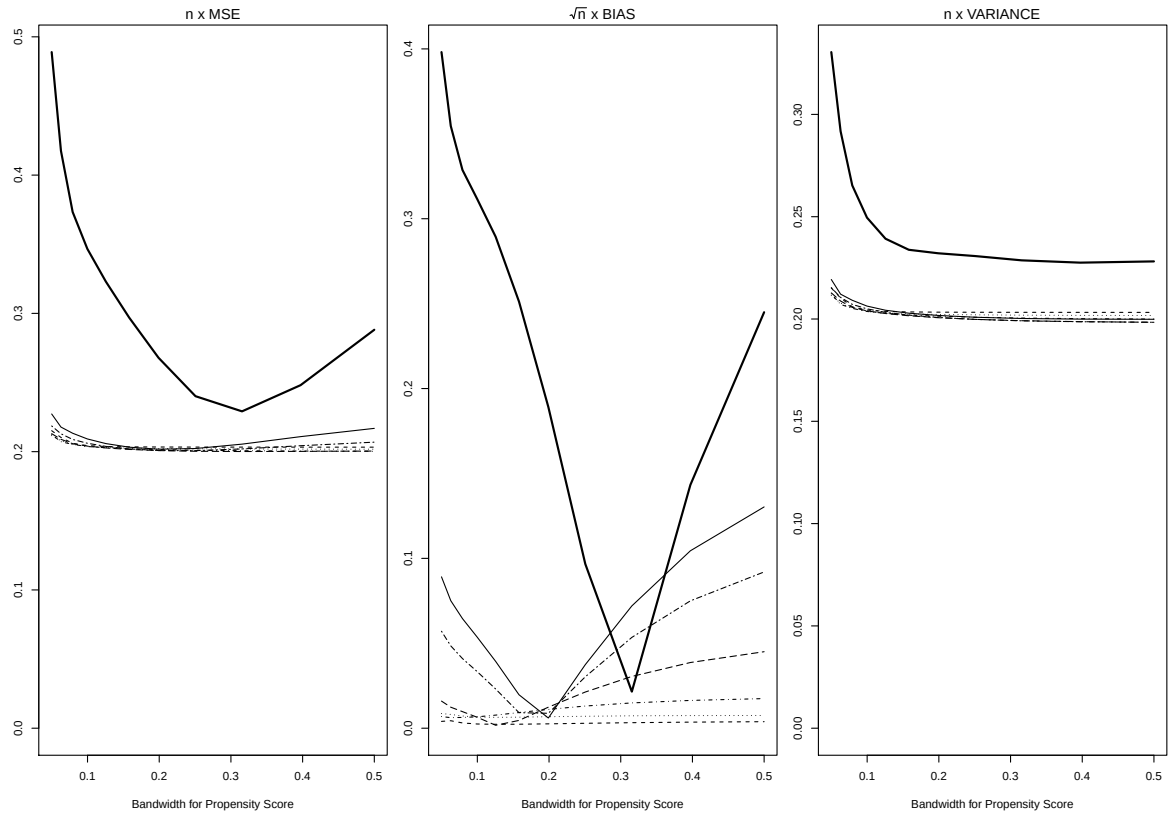

Figure 1: Simulation results: MSE, absolute bias and variance of $\widehat{\theta}_{IPW-K}$ for various values of h_2 (bold solid line), compared to results for $\widehat{\theta}_{DR-K}$ with bandwidth h_1 equal to .05 (short-dashed line), .08 (dotted line), .13 (dot-dashed line), .2 (long dashed line),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

.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.[13] 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

[13]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).

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.

3. ESTIMATION IN A GENERAL SEMIPARAMETRIC MODEL

The results 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 general 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. This setup covers the model from the previous section, but also several others; see Section 3.4 below. We show that this construction alone is not enough to obtain an asymptotic linearity result like the one in the previous section, but that instead some additional structure is needed. However, we argue that this additional structure is present in many models in which DR estimators are known to exist.

3.1. **Setup.** Consider the problem of estimating a parameter θ^o, contained in the interior of some compact parameter space $\Theta \subset \mathbb{R}^{d_\theta}$, using an i.i.d. sample $\{Z_i\}_{i=1}^n$ from the distribution of 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 function ψ that takes values in $\mathbb{R}^{d_\theta}$ and satisfies the following relationship:

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

Here ξ^o is an unknown (but identified) nuisance function that could in principle also depend on θ. The functional Ψ is also assumed to be *doubly robust* for estimating θ^o, in the sense 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. We do not consider the issue whether such a functions exists in any given semiparametric model.[14]

To simplify the exposition, we focus on two special cases; one where both ξ_1^o and ξ_2^o are conditional expectations, and one where ξ_1^o is the density function of a continuously distributed random vector and ξ_2^o is a conditional expectation:

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

Case 2: $\xi_1^o(x_1) = \partial_{x_1} P(X_1 \leq x_1)$ and $\xi_2^o(x_2) = \mathbb{E}(Y_2 | X_g = x_2)$.

Here $Y_g \in \mathbb{R}$, $X_g \in \mathbb{R}^{d_g}$ and (Y_1, Y_2, X_1, X_2) is a random subvector of Z that might have duplicate elements. We also assume for simplicity that ξ^o does not depend on θ, and 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 $\xi(U) =$

[14] See Robins and Rotnitzky (2001) for some results in this regard. One of their results is 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, we can focus on settings where ψ takes values in $\mathbb{R}^{d_\theta}$ without loss of generality.

$(\xi_1(U_1), \xi_2(U_2))$ 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 $\Psi(\theta, \xi_1, \xi_2) = \mathbb{E}(\psi(Z, \theta, \xi_1(U_1), \xi_2(U_2)))$.

3.2. **Estimator and Main Result.** Our interest is in the properties of STS estimators based on a sample analogue of a DR moment condition, to which we 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 $\xi^o = (\xi_1^o, \xi_2^o)$. Under Case 1, we estimate ξ_g^o by "leave-one-out" local polynomial regression of order l_g using bandwidth h_g, for $g = 1, 2$. Using notation analogous to that introduced in Section 2.2, we put

$$\widehat{\xi}_g(U_{gi}) = e_1^\top \operatorname*{argmin}_{\alpha} \sum_{j \neq i} \Big(Y_{gj} - \mathcal{P}_{l_g,\alpha}(X_{gj} - U_{gi})\Big)^2 K_{h_g}(X_{gj} - U_{gi}), \quad g = 1, 2.$$

For Case 2, we use a standard "leave-one-out" kernel density estimators to estimate the density ξ_1^o, using a kernel function of order $l_1 + 1$ for the purpose of bias control. That is, with $\mathcal{K}^*$ a symmetric function on $\mathbb{R}$ whose exact properties are stated in the Appendix, and $K_h^*(b) = \prod_{j=1}^d \mathcal{K}^*(b_j/h)/h$, we define

$$\widehat{\xi}_1(U_{1i}) = \frac{1}{n} \sum_{j \neq i} K_{h_1}^*(U_{1j} - X_{1i}).$$

Our estimate of ξ_2^o is the same as for Case 1. 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 in both Case 1 and 2.

Due to the DR property of ψ 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, where $H = \mathbb{E}(\partial_\theta \psi(Z, \theta^o, \xi^o))$. The following theorem gives conditions the for asymptotic linearity of $\widehat{\theta}_{DR}$.

Theorem 2. *Under suitable regularity conditions (see Appendix A.2),* $\sqrt{n}(\widehat{\theta}_{DR} - \theta^o) = n^{-1/2}\sum_{i=1}^n \phi_G(Z_i) + o_P(1) \xrightarrow{d} N(0, \Sigma_G)$ *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) we are in Case 1, the distribution of 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* $g = 1, 2$.

(c) we are in Case 2, the distribution of Z *is such that* $X_1 = r(X_2)$ *for some fixed function* r, *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* $g = 1, 2$.

3.3. **Discussion.** Theorem 2(a) is the main result of this section. It shows that simply being based on a DR moment condition is not enough for an STS estimator to be asymptotically linear under the same kind of weak restrictions that we imposed for the 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 see that under its conditions we only get that

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

Following Newey and McFadden (1994), we would expect such a result for any adaptive kernel-based STS estimator. We remark that STS-DR still performs somewhat better than adaptive

STS estimators, but in a way that is not apparent from (3.5). To explain this, consider the functionals $\xi_1 \mapsto \Psi(\theta^o, \xi_1, \xi_2^o)$ and $\xi_2 \mapsto \Psi(\theta^o, \xi_1^o, \xi_2)$. If ψ was only an influence function without the DR property, these functionals would each have zero first order derivatives. With the DR property, these functionals are actually constant and equal to zero. An inspection of the proof of Theorem 2 shows that this property removes some but not all terms of order $O_P(n^{-1}h^{-d})$ from an expansion of the estimator $\widehat{\theta}_{DR}$ (relative to that of an STS estimator based on moment condition with two nuisance function without the DR property).

So which features of a semiparametric model, in addition to the presence of a DR moment condition, deliver properties of STS-DR estimators of the kind we found in Section 2? This question in answered by Theorem 2(b)–(c). Under the conditions stated there, $\widehat{\theta}_{DR}$ has theoretical properties analogous to those established in Theorem 1. In particular, the nonparametric component is not required to converge with a rate of $o_P(n^{-1/4})$, and

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

if h is chosen such from the permissible range of bandwidths that $n^{1/2d}h \to \infty$. This improvement over part (a) comes from the fact that for an STS-DR estimator, roughly speaking, the remaining 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 $\widehat{\xi}_2$. For our Case 1, the orthogonality condition (3.4) 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. For our Case 2, the estimation error $\widehat{\xi}_1 - \xi_1^o$ is essentially a weighted sum of functions of the X_{1i}, whereas $\widehat{\xi}_2 - \xi_2^o$ is essentially a weighted sum of the "true" residuals $Y_{2i} - \xi_2^o(X_{2i})$. Since the latter are uncorrelated with any function of X_{2i}, the $O_P(n^{-1}h^{-d})$ term vanishes.

While the conditions of Theorem 2(b)–(c) might at first seem obscure and rather restrictive, they are indeed satisfied by a wide range of models. In fact, we are not aware of a commonly used model in which a DR procedure exist that is not covered by our Cases 1 or 2, or some

minor variation thereof. For the missing data model in Section 2, which falls under Case 1, the orthogonality property required by Theorem 2(b) follows from the fact that the data are missing at random. Theorem 2(b) also covers various models for causal inference that are similar in structure, in the sense that identification is achieved through some particular conditional independence restriction. Our results thus extend, for example, to the DR estimators of average treatment effects under unconfoundedness reviewed in Bang and Robins (2005), but also to the DR estimators of local average treatment effects using instrumental variables proposed by Tan (2006) and Ogburn et al. (2015). Moreover, as pointed out in the following subsection, Theorem 2(b)–(c) also cover the partially linear model, a policy effects model, and weighted average derivatives.

Remark 3. We are not aware of a model with a DR moment condition for which the conditions of Theorem 2(b)–(c) are not naturally satisfied. However, even if such a model existed, one can always construct a modified nonparametric first-stage estimators such that $\widehat{\xi}_1 - \xi_1^o$ and $\widehat{\xi}_2 - \xi_2^o$ are asymptotically uncorrelated. A simple way to achieve this, for example, would be to split the data into two parts at random, and then calculate $\widehat{\xi}_1$ and $\widehat{\xi}_2$ from different subsamples.

3.4. **Application to Further Specific Models.** We now briefly review three additional specific examples that are widely used in applied work, and that are covered by the theory presented in this section. In particular, in all cases a STS-DR estimator based on standard kernel-based estimators of the respective nuisance functions would be asymptotically linear under conditions on the order of the bias and the stochastic parts analogous to those in Theorem 1 and Theorem 2(b)–(c).

Partial Linear Model. Suppose that $Z = (Y, X, W)$, where Y is a scalar outcome variable and both X and W are vectors of explanatory variables. Then a partially linear regression model (e.g. Robinson, 1988) assumes that $Y = \lambda^o(X) + W^\top \theta^o + \varepsilon$, where λ^o is a smooth

function, θ^o is a vector of parameters, and ε is an unobserved random variable that satisfies $\mathbb{E}(\varepsilon|X,W)=0$. Now 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 DR moment function for estimating θ^o. This model is a minor variation of our Case 1, and it follows from the model structure and the assumption that $\mathbb{E}(\varepsilon|X,W)=0$ that the orthogonality condition (3.4) holds. The statement of Theorem 2(b) then applies analogously to the STS-DR estimator under appropriately adapted regularity conditions. See Linton (1995) for similar result using different arguments. Note that the STS-DR estimator is easily seen to be identical (up to trimming terms) to the estimator proposed by Robinson (1988).

Policy Effects. Suppose that $Z=(Y,X)$, where Y is a scalar outcome variable and X is a vector of explanatory variables. Stock (1989) studies the problem of predicting the effect of a change in the distribution of X to that of $\pi(X)$, where π is some known *policy function*, on the expectation of the outcome variable. Under certain assumptions, this parameter of interest is given by $\theta^o=\mathbb{E}(\mathbb{E}(Y|X=x)|_{x=\pi(X)})$. 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 DR moment function for estimating θ^o. This setup is a minor variation of our Case 2, the difference being that ξ_1^o and ξ_2^o each have more than one component. This however, has no effect on the structure of the proof of Theorem 2(c), and thus its statement applies analogously to the STS-DR estimator here under appropriately adapted regularity conditions.

Weighted Average Derivatives. Suppose that $Z=(Y,X)$, where Y is a scalar dependent variable and X is a vector of continuously distributed explanatory variables with density

function ξ_2^o. 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 measure of nonparametrically estimated regression functions more generally (e.g. Powell et al., 1989; Newey and Stoker, 1993; Cattaneo et al., 2013). Let $\xi_{11}^o(x) = \mathbb{E}(Y|X = x)$, denote the density of X by $\xi_{21}^o(x)$, and denote the vectors of partial derivatives of those two functions by $\xi_{12}^o(x) = \partial_x \xi_{11}^o(x)$ and $\xi_{22}^o(x) = \partial_x \xi_{21}^o(x)$, respectively. Then

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

is a DR moment function for estimating θ^o. This setup is a minor variation of our Case 2. Both ξ_1^o and ξ_2^o again have more than one component, and in addition one of the components is a derivative of a density or a conditional expectation, respectively. Again, it is easy to see that this has no effect on the structure of the proof of Theorem 2(c), and thus its statement applies analogously to the STS-DR estimator here under appropriately adapted regularity conditions.

4. CONCLUSIONS

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 such STS-DR estimators have favorable theoretical and practical properties relative to other commonly used STS estimators. 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, which we show to be satisfied in a wide range of models.

A. REGULARITY CONDITIONS FOR MAIN RESULTS

In this Appendix, we collect the regularity conditions necessary to establish the statements of Theorems 1 and 2.

A.1. **Regularity Conditions for Theorem 1.** The statement of the Theorem holds under the following regularity conditions.

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 $\theta \in \Theta \setminus \{\theta^o\}$, with $\Theta \subset \mathbb{R}^{d_\theta}$ 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 $c > 6$.

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.

A.2. **Regularity Conditions for Theorem 2.** The statement of the Theorem holds under Case 1 if Assumptions K1 and G1–G2 are satisfied, and under Case 2 if Assumptions K1–K2, G1 and G3 are satisfied. Here Assumptions K1 is as stated in the previous subsection, and all other regularity conditions are as follows.

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 G 1. (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 $c > 6$.

Assumption G3. The statements of Assumption G2 hold for $g = 2$. In addition: (i) X_1 and U_1 are continuously distributed with support $\mathcal{S}(X_1) = \mathbb{R}^{d_1}$ and $\mathcal{S}(U_1) \supseteq \mathcal{S}(X_1)$, respectively; (ii) the density function ξ_1^o of X_1 has continuous and uniformly bounded derivatives up to order $(l_1 + 1)$.

Assumption K2 describes a standard higher-order kernel function, and Assumption G1 is analogous to Assumption MD1 above. Assumptions G2–G3 collect various smoothness

conditions that are standard in the literature on kernel-based nonparametric regression and density estimation. Note that the restriction that X_1 has unbounded support in Assumptions G3 can easily be weakened if instead a boundary kernel is used in the construction of the estimator of ξ_1^o. This would ensure a bias of uniform order $O(h_1^{l_1+1})$ without affecting the basic structure of the stochastic part, which is all we need for the proof of the theorem.

B. **PROOFS OF MAIN RESULTS**

In this Appendix, we give the proofs of Theorems 1 and 2. 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.

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

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

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

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

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

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

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

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

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

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

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

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

B.2. **Stochastic Expansion of the Local Polynomial Estimator.** 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. $\tau_s(1) = (0, \ldots, 0, s)$, $\ldots$, $\tau_s(n_s) = (s, 0, \ldots, 0)$. For each $s \in \{0, 1, \ldots, l_g\}$ we also define a

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

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

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 $M_{gn,i}(u) = n^{-1}\sum_{j \neq i} K_{h_g}(X_{gj} - u)$ is the leave-one-out version of the usual Rosenblatt-Parzen density estimator, that $N_{gn}(u) = \mathbb{E}(K_{h_g}(X_{gi} - u))$ is its expectation, and that $\eta_{gn,j}(u) = K_{h_g}(X_{gj} - u) - \mathbb{E}(K_{h_g}(X_{gj} - u))$ 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 odd $l_g \geq 1$ the bias B_{gn} satisfies

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

and the first- and second-order stochastic terms satisfy

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

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

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

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

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

Proof. The statement about the bias in part (i) follows from standard Taylor expansion based arguments. All remaining statements of the Lemma follow from well-known arguments in e.g. Masry (1996) or Kong et al. (2010) for the usual "leave-in" version of the local polynomial regression estimator. It therefore only remains to be shown that the difference between such a "leave-in" estimator and the "leave-one-out" estimator that we consider in this paper is

of sufficiently small order. To see this for the case of the statement about the first-order stochastic term in part (i), let

$$\tilde{S}_{gn}(u) = \frac{1}{n}\sum_j e_1^\top N_{gn}(u)^{-1} w_{gj}(u) K_{h_g}(X_{gj} - u)\varepsilon_{gj}$$

be the "leave-in" analogue of our first-order stochastic term, and note that

$$\max_{i\in\{1,\ldots,n\}} |S_{gn}(U_{gi})| \leq \max_{i\in\{1,\ldots,n\}} |S_{gn}(U_{gi}) - \tilde{S}_{gn}(U_{gi})| + \sup_u |\tilde{S}_{gn}(u)|$$

It follows from e.g. Masry (1996) or Kong et al. (2010) that

$$\sup_{u\in\mathcal{S}(U_g)} |\tilde{S}_{gn}(u)| = O_P((nh_g^{d_g}/\log n)^{-1/2}).$$

Moreover, it holds that

$$\begin{aligned}
&\max_{i\in\{1,\ldots,n\}} |S_{gn}(U_{gi}) - \tilde{S}_{gn}(U_{gi})| = \max_{i\in\{1,\ldots,n\}} |\frac{1}{n} e_1^\top N_{gn}(U_{gi})^{-1} w_{gi}(U_{gi}) K_{h_g}(X_{gi} - U_{gi})\varepsilon_{gi}| \\
&\quad \leq \max_{i\in\{1,\ldots,n\}} |\frac{1}{n} e_1^\top N_{gn}(U_{gi})^{-1} w_{gi}(U_{gi}) K_{h_g}(X_{gi} - U_{gi})| \max_{i\in\{1,\ldots,n\}} |\varepsilon_{gi}| \\
&\quad = O_P((nh_g^{d_g})^{-1}) \times o_P(n^{1/c}),
\end{aligned}$$

with c as defined in Assumption 2. The desired result then follows from the restrictions on the bandwidth in the statement of the theorem, and the restrictions on c from Assumption 2. The remaining statements of the Lemma follow by similar arguments, and we hence omit the details.

□

B.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 parts (a) and (c) below. For notational simplicity, we drop the "DR" subscript of the estimator. It is straightforward to show that $\widehat{\theta} \xrightarrow{p} \theta^o$ under either condition (a), (b) or (c), and thus we omit the details of proving this step. From the

differentiability of ψ with respect to θ and the definition of $\widehat{\theta}$, it follows 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 $H_n(\theta, \xi) = (1/n) \sum_{i=1}^{n} \partial_\theta \psi(Z_i, \theta, \xi_1(U_{1i}), \xi_2(U_{2i}))$. It then follows from standard arguments that $H_n(\theta^*, \widehat{\xi}) = H + o_P(1)$. Next, we expand

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

Using the notation that

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

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

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

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:*

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

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

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

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

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

where the first equality follows from Chebyscheff's inequality, and the second equality follows from Lemma 2 and the fact that by equation (B.1) we have that $\mathbb{E}(\psi_i^g B_{gn}(U_{gi})) = 0$. Next,

consider the term

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

This is a second order U-Statistic (up to a bounded, multiplicative term), and since by equation (B.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. Lemma 1 and some simple variance calculations then imply that

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

Finally, we consider the term

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

where

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

Using equation (B.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 4. 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 $k \in \{1,\ldots,6\}$. Our proof proceeds by obtaining sharp bounds on $n^{-1}\sum_{i=1}^{n}\psi_i^{gg}T_{n,k}(U_{gi})$ for $k \in \{1,2,4,5\}$ using equation B.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 (B.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 (B.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 (B.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:

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

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

Remark 5. 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

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

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

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

where $T_{n,1}(u)=B_{1,n}(u_1)B_{2,n}(u_2)$, $T_{n,2}(u)=B_{1,n}(u_1)S_{2,n}(u_2)$, $T_{n,3}(u)=B_{1,n}(u_1)R_{2,n}(u_2)$, $T_{n,4}(u)=S_{1,n}(u_1)B_{2,n}(u_2)$, $T_{n,5}(u)=S_{1,n}(u_1)S_{2,n}(u_2)$, $T_{n,6}(u)=S_{1,n}(u_1)R_{2,n}(u_2)$, $T_{n,7}(u)=R_{1,n}(u_1)B_{2,n}(u_2)$, $T_{n,8}(u)=R_{1,n}(u_1)S_{2,n}(u_2)$, and $T_{n,9}(u)=R_{1,n}(u_1)R_{2,n}(u_2)$. Since the last two terms on the right-hand side of the previous equation are easily of the order $o_P(n^{-1/2})$ due to the restrictions on the bandwidth, it suffices to show that for any for $k\in\{1,\dots,9\}$ 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 $k\in\{1,\dots,5,7\}$ 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

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

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

$$\rho_{n,2}^2 = \mathrm{Var}(\mathbb{E}(\lambda_{ijl}|\mathcal{Z}_l, \mathcal{Z}_j)) = \mathrm{Var}\left(\int K_{h_1}(X_{1j} - u_1)K_{h_2}(X_{2l} - u_2)f_U(u_1, u_2)du_1du_2\varepsilon_{1j}\varepsilon_{2l}\right).$$

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

$$\begin{aligned}\rho_{n,3}^2 &= \mathrm{Var}(\mathbb{E}(\lambda_{ijl}|\mathcal{Z}_i, \mathcal{Z}_j, \mathcal{Z}_l)) \\ &= \mathbb{E}\left(h_1^{-2d_1}K((X_{1j} - U_{1i})/h_1)^2h_2^{-2d_2}K((X_{2l} - U_{2i})/h_2)^2\varepsilon_{1j}^2\varepsilon_{2l}^2\right) = O(h_1^{-d_1}h_2^{-d_2}).\end{aligned}$$

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

$$\rho_{n,1}^2 = \mathrm{Var}(\mathbb{E}(\lambda_{ij}|\mathcal{Z}_j)) = \mathrm{Var}(\int K_{h_1}(X_{1j} - u_1) K_{h_2}(X_{2j} - u_2) f_U(u_1, u_2) du_1 du_2 \varepsilon_{1j}\varepsilon_{2j})$$

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

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

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 6. 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 $T_{n,5,A} = O_P(n^{-1}\max\{h_1^{-d_1}, h_2^{-d_2}\})$.

B.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 4–6.

B.5. **Proof of Theorem 2(c).** The difference between Case 1 and Case 2 is that ξ_1^o is now a density function. This actually simplifies the problem, as a stochastic expansion of the kind given in Lemma 2 is easier to obtain and 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 Theorem we have that

$$\widehat{\xi}_1(X_i) = \xi_1^o(X_i) + B_1(X_i) + S_1(X_i)$$

Here $B_1(X_i) = \mathbb{E}(K_h(U_1 - X_i)|X_i) = O(h^{l+1})$ is a deterministic bias function and $S_1(X_i) = \sum_{j\neq i}(K_h(U_{1j} - X_i) - \mathbb{E}(K_h(U_1 - X_i)|X_i))/n = O_P((nh^d/\log(n))^{-1/2})$ is a mean zero stochastic term. 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.

B.6. **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 since

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

by the Law of Iterated Expectations.

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.

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 (2016): "Program evaluation with high-dimensional data," *Econometrica*, to appear.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *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.

Bravo, F. and D. T. Jacho-Chávez (2010): "Empirical likelihood for efficient semiparametric average treatment effects," *Econometric Reviews*, 30, 1–24.

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*.

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.

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

——— (2016): "Identification and estimation of semiparametric two-step models," *Quantitative Economics*, 7, 561–589.

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

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. (2015): "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189, 1–23.

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*.

Goldstein, L. and K. Messer (1992): "Optimal Plug-in Estimators for Nonparametric Functional Estimation," *Annals of Statistics*, 1306–1328.

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

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*.

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.

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

Mammen, E., C. Rothe, and M. Schienle (2016): "Semiparametric estimation with generated covariates," *Econometric Theory*, 32, 1140–1177.

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

Naimi, A. I. and E. H. Kennedy (2017): "Nonparametric Double Robustness," *Working Paper*.

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.

Ogburn, E. L., A. Rotnitzky, and J. M. Robins (2015): "Doubly robust estimation of the local average treatment effect curve," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 373–396.

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.

Racine, J. and Q. Li (2004): "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics*, 119, 99–130.

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.

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

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

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

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.

Vermeulen, K. and S. Vansteelandt (2015): "Bias-reduced doubly robust estimation," *Journal of the American Statistical Association*, 110, 1024–1036.

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						**$\sqrt{n}$×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**	.489	.398	.330	.832	.826	.823	.826	.843	.868
	.08	.373	.329	.265	.872	.870	.869	.871	.885	.903
	.13	.323	.289	.239	.885	.884	.883	.884	.895	.911
	.20	.268	.189	.232	.908	.906	.906	.907	.914	.926
	.32	.229	.022	.229	.926	.924	.924	.923	.928	.937
	.50	.288	.245	.228	.878	.879	.878	.876	.879	.887
	$\mathbf{h_2}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_1}$)**					
					.05	**.08**	**.13**	**.20**	**.32**	**.50**
$\widehat{\theta}_{IPW-TK}$	**.05**	.639	.282	.560	.825	.801	.840	.921	.993	.788
	.08	.934	.020	.934	.761	.734	.774	.885	.990	.698
	.13	.788	.205	.747	.756	.729	.773	.899	.988	.669
	.20	.646	.176	.615	.810	.786	.832	.935	.994	.744
	.32	.397	.354	.272	.903	.879	.920	.980	1.000	.860
	.50	.252	.153	.229	.938	.929	.944	.980	.998	.909
	$\mathbf{h_2}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_1}$)**					
					1	**2**	**3**	**4**	**7**	**10**
$\widehat{\theta}_{IPW-OS}$	**1**	.497	.515	.232	.744	.725	.731	.730	.726	.725
	2	.244	.063	.240	.934	.912	.914	.912	.909	.909
	3	.232	.028	.231	.950	.926	.924	.923	.920	.919
	4	.221	.009	.221	.954	.928	.928	.926	.924	.924
	7	.222	.011	.222	.951	.929	.929	.928	.925	.923
	10	.236	.028	.236	.942	.919	.921	.919	.916	.914
	$\mathbf{h_2}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_1}$)**					
					.50	**.65**	**.80**	**.95**	**1.10**	**1.25**
$\widehat{\theta}_{IPW-SP}$	**.50**	.359	.339	.244	.818	.821	.824	.826	.827	.837
	.65	.252	.170	.223	.894	.897	.900	.901	.901	.909
	.80	.226	.077	.220	.913	.916	.919	.920	.922	.930
	.95	.233	.025	.232	.913	.915	.916	.918	.919	.931
	1.10	.258	.181	.225	.889	.893	.897	.898	.899	.908
	1.25	.613	.633	.212	.644	.647	.651	.652	.648	.651
	$\mathbf{h_2}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate**					
$\widehat{\theta}_{IPW-BS}$	**.05**	.271	.011	.271	0.95					
	.08	.301	.062	.298	0.96					
	.13	.328	.185	.294	0.96					
	.20	.275	.156	.251	0.95					
	.32	.264	.209	.220	0.91					
	.50	.512	.552	.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

	$\mathbf{h_1}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_2}$)**					
					.05	**.08**	**.13**	**.20**	**.32**	**.50**
$\widehat{\theta}_{REG-K}$	**.05**	.210	.052	.207	.948	.945	.943	.940	.936	.932
	.08	.227	.136	.208	.938	.936	.934	.929	.926	.923
	.13	.282	.266	.211	.910	.907	.905	.901	.896	.889
	.20	.308	.303	.216	.892	.889	.888	.882	.877	.873
	.32	.211	.030	.210	.952	.950	.948	.944	.940	.935
	.50	.295	.313	.197	.923	.920	.914	.906	.895	.884
	$\mathbf{h_1}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_2}$)**					
					.05	**.08**	**.13**	**.20**	**.32**	**.50**
$\widehat{\theta}_{REG-TK}$	**.05**	.926	.021	.925	.837	.829	.826	.831	.836	.822
	.08	.643	.011	.643	.871	.863	.858	.866	.869	.857
	.13	.856	.152	.833	.897	.892	.888	.893	.894	.884
	.20	3.713	.496	3.467	.816	.810	.809	.819	.827	.806
	.32	8.592	.281	8.514	.993	.989	.981	.986	.991	.987
	.50	.228	.157	.203	.949	.938	.931	.938	.942	.927
	$\mathbf{h_1}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_2}$)**					
					1	**2**	**3**	**4**	**7**	**10**
$\widehat{\theta}_{REG-OS}$	**1**	.612	.652	.187	.671	.731	.749	.747	.751	.753
	2	.204	.081	.197	.931	.938	.940	.941	.940	.942
	3	.203	.016	.203	.931	.936	.939	.939	.940	.940
	4	.203	.009	.203	.930	.934	.936	.937	.938	.940
	7	.206	.001	.206	.926	.931	.932	.931	.931	.932
	10	.218	.004	.217	.919	.926	.926	.927	.927	.927
	$\mathbf{h_1}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate ($\mathbf{h_2}$)**					
					.50	**.65**	**.80**	**.95**	**1.10**	**1.25**
$\widehat{\theta}_{REG-SP}$	**.50**	.209	.004	.209	.925	.926	.927	.927	.925	.920
	.65	.205	.003	.205	.932	.932	.931	.931	.929	.925
	.80	.203	.003	.203	.934	.935	.934	.935	.932	.927
	.95	.201	.006	.200	.937	.937	.939	.939	.935	.929
	1.10	.203	.091	.195	.936	.937	.938	.938	.935	.928
	1.25	.331	.381	.186	.864	.868	.870	.872	.862	.837
	$\mathbf{h_1}$	$\mathbf{n\times MSE}$	$\mathbf{\sqrt{n}\times BIAS}$	$\mathbf{n\times VAR}$	**Coverage Rate**					
$\widehat{\theta}_{REG-BS}$	**.05**	.210	.050	.207	0.94					
	.08	.227	.134	.209	0.93					
	.13	.282	.267	.211	0.90					
	.20	.311	.309	.216	0.89					
	.32	.212	.038	.210	0.94					
	.50	.290	.305	.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.

SUPPLEMENTAL MATERIAL (NOT FOR PUBLICATION)

In this section, we consider again the simulation study in Section 2.4 in the main text. In particular, we study how the performance of the estimators considered in Section 2.4 changes in response to changes in the the distribution of the covariate X, which is uniform on $[0, 1]$ in the main text. Specifically, we consider a symmetric triangular distribution, a Beta(2,4) distribution, and a Beta(4,2) distribution. The three alternative distributions differ from a uniform one in that their density is not bounded away from zero over their support; and the latter two are also asymmetric. All other aspects of the data generating process remain the same (but of course the implied values of θ_o and Σ_{MD} change accordingly).

In Figures 2–5 we plot the simulated MSE, absolute bias and variance of the kernel-based IPW estimator against the results for the kernel-based STS-DR estimator. For convenience, Figure 2 replicates the results for the uniform covariate distribution from the main text, while Figures 3–5 show results for a triangular, a Beta(2,4), and a Beta(4,2) distribution, respectively. As one can see, the results are very similar qualitatively. In particular, the value of the MSE of the STS-DR estimator as a function of the two bandwidths is rather flat over the range of bandwidths considered here for all simulation designs (there is a spike in the MSE of the STS-DR estimator for very small values of the bandwidth used to estimate the propensity score in case of the trinagular and the Beta(4,2) covariate distribution, but it is much less pronouced than the one of the IPW estimator). Our exercise thus shows that the findings reported in Section 2.4 are not an artefact of the specific covariate distribution considered there.

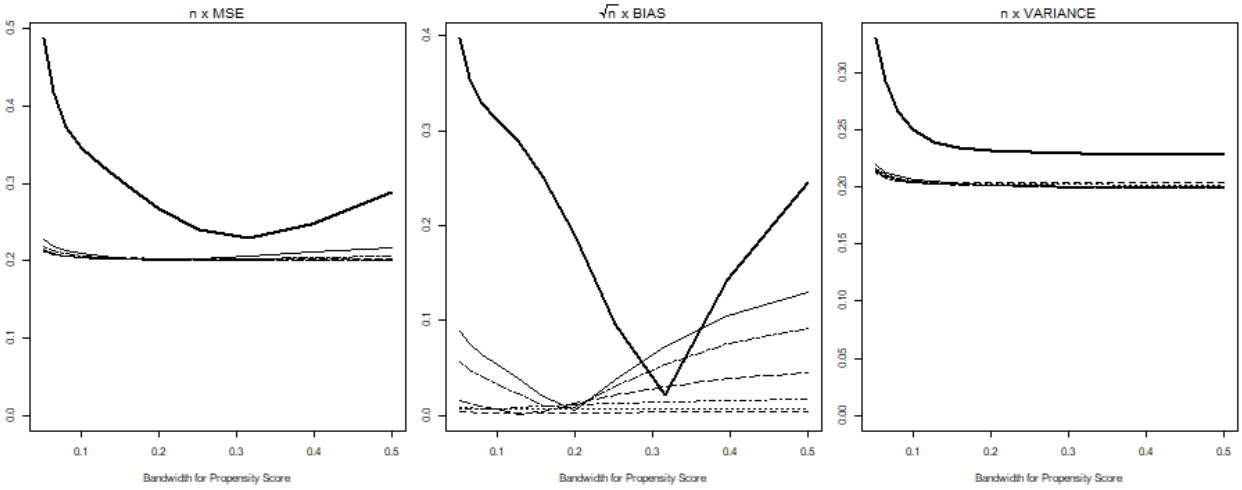

Figure 2: Simulation results for $U \sim U[0,1]$: MSE, absolute bias and variance of $\widehat{\theta}_{IPW-K}$ for various values of h_2 (bold solid line), compared to results for $\widehat{\theta}_{DR-K}$ with bandwidth h_1 equal to .05 (short-dashed line), .08 (dotted line), .13 (dot-dashed line), .2 (long dashed line), .32 (long dashed dotted line), and .5 (thin solid line).

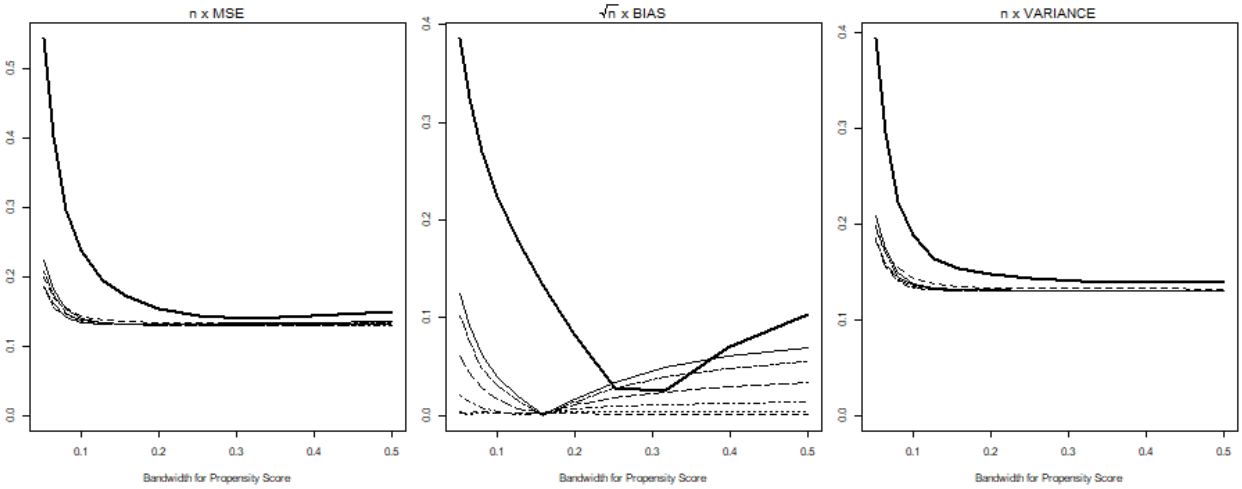

Figure 3: Simulation results for $U \sim$ Triangular$[0,1]$: MSE, absolute bias and variance of $\widehat{\theta}_{IPW-K}$ for various values of h_2 (bold solid line), compared to results for $\widehat{\theta}_{DR-K}$ with bandwidth h_1 equal to .05 (short-dashed line), .08 (dotted line), .13 (dot-dashed line), .2 (long dashed line), .32 (long dashed dotted line), and .5 (thin solid line).

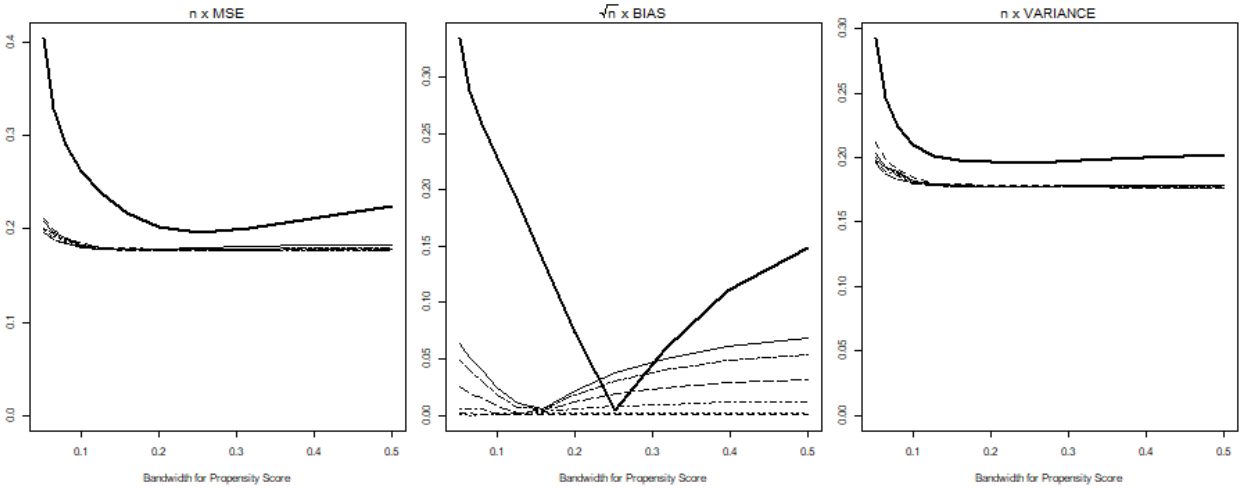

Figure 4: Simulation results for $U \sim \text{Beta}(2,4)$: MSE, absolute bias and variance of $\widehat{\theta}_{IPW-K}$ for various values of h_2 (bold solid line), compared to results for $\widehat{\theta}_{DR-K}$ with bandwidth h_1 equal to .05 (short-dashed line), .08 (dotted line), .13 (dot-dashed line), .2 (long dashed line), .32 (long dashed dotted line), and .5 (thin solid line).

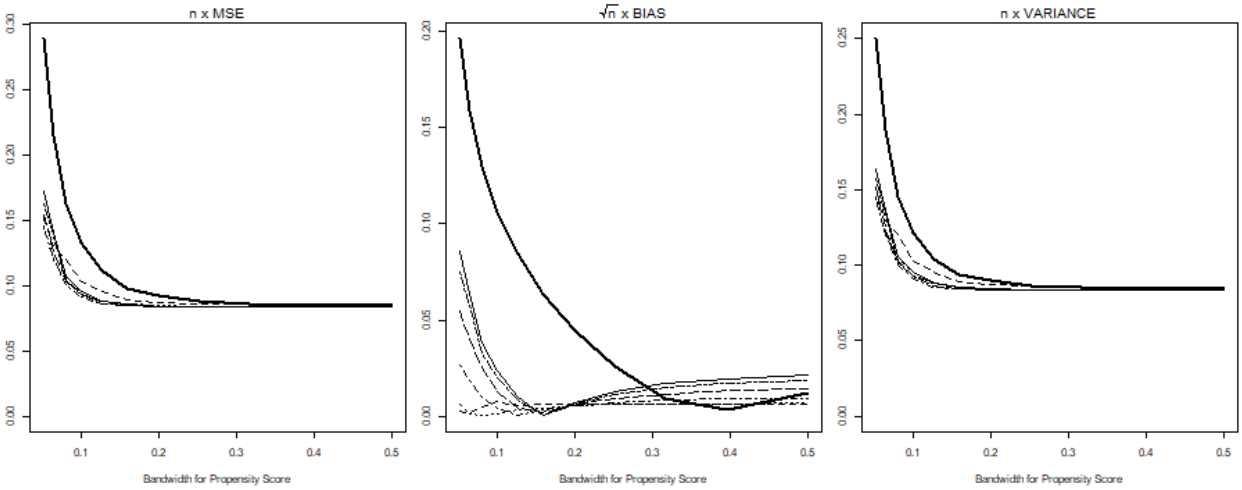

Figure 5: Simulation results for $U \sim \text{Beta}(4,2)$: MSE, absolute bias and variance of $\widehat{\theta}_{IPW-K}$ for various values of h_2 (bold solid line), compared to results for $\widehat{\theta}_{DR-K}$ with bandwidth h_1 equal to .05 (short-dashed line), .08 (dotted line), .13 (dot-dashed line), .2 (long dashed line), .32 (long dashed dotted line), and .5 (thin solid line).