Semiparametric Estimation with Generated Covariates

Enno Mammen, Christoph Rothe, and Melanie Schienle[*]

Heidelberg University & National Research University Higher School of Economics,

Columbia University, and Karlsruhe Institute of Technology

Abstract

We study a general class of semiparametric estimators when the infinite-dimensional nuisance parameters include a conditional expectation function that has been estimated nonparametrically using generated covariates. Such estimators are used frequently to e.g. estimate nonlinear models with endogenous covariates when identification is achieved using control variable techniques. We study the asymptotic properties of estimators in this class, which is a non-standard problem due to the presence of generated covariates. We give conditions under which estimators are root-n consistent and asymptotically normal, derive a general formula for the asymptotic variance, and show how to establish validity of the bootstrap.

JEL Classification: C14, C31

Keywords: *Semiparametric estimation, generated covariates, profiling, propensity score*

[*] First Version: January 10, 2010. This Version: March 4, 2015. Enno Mammen: Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany. E-mail: mammen@math.uni-heidelberg.de. Christoph Rothe: Department of Economics, Columbia University, 420 W 118th Street, New York, NY 10027, USA. E-mail: cr2690@columbia.edu. Melanie Schienle: Institute of Economics, Karlsruhe Institute of Technology, Schlossbezirk 12, D-76131 Karlsruhe, Germany. E-mail: melanie.schienle@kit.edu. Enno Mammen acknowledges financial support from the Government of the Russian Federation within the framework of the implementation of the 5-100 Programme Roadmap of the National Research University Higher School of Economics. Melanie Schienle thanks Deutsche Forschungsgemeinschaft for financial support. The authors would like to thank Ingrid Van Keilegom, Oliver Linton, the referees, and attendants of numerous conference and seminar presentations for their helpful comments. All remaining errors are ours.

1. Introduction

In this paper, we study the theoretical properties of semiparametric estimators when estimation of the nonparametric component requires the use of generated covariates. Such "three-step" estimators are for example frequently used to estimate nonlinear models with endogenous covariates when identification is achieved using control variable techniques. Here we consider a general class of semiparametric optimization estimators with a criterion function that depends on a conditional expectation function that has been estimated nonparametrically using generated covariates. The nonparametric component may be profiled and thus depend on unknown finite-dimensional parameters. Generated covariates may originate from an either parametric, semiparametric or nonparametric first step, and we allow the function that generates them to also serve some other purpose within the model.

Deriving asymptotic properties of estimators in this class is a non-standard problem due to the presence of generated covariates. The contribution of this paper is to give conditions on the primitives of the model under which these semiparametric "three-step" estimators are root-n consistent and asymptotically normal, to derive a general formula for the asymptotic variance, and to show how to establish validity of the bootstrap. As an illustration, we apply our methods to estimation of average treatment effects under unconfoundedness via regression on the propensity score (Rosenbaum and Rubin, 1983); and our paper is the first to give explicit conditions for root-n consistency and asymptotic normality of the respective estimator.

Semiparametric estimation problems involving both finite- and infinite-dimensional parameters are central to econometrics, and are studied extensively under general conditions by e.g. Newey (1994), Andrews (1994), Chen and Shen (1998), Ai and Chen (2003, 2007), Chen, Linton, and Van Keilegom (2003), Chen and Pouzo (2009), or Ichimura and Lee (2010). None of these papers explicitly considers the case of generated covariates in the nonparametric component, but one can easily see that the "high-level" conditions they employ are typically sufficiently abstract to encompass the generation step. What needs to be adapted substantially, however, are the methods used to verify some of these abstract conditions in the context of a specific model. Compared to a standard analysis of a setting without generated covariates, the main difficulties occur when establishing a uniform rate of consistency for the nonparametric component (e.g. Newey, 1994, Assumption 5.1(ii); or Chen, Linton, and Van Keilegom, 2003, Condition (2.4)), and an asymptotic normality result for a linearized version of the objective function (e.g. Newey, 1994, Assumption 5.3 and Lemma 1; or Chen, Linton, and Van Keilegom, 2003, Condition (2.6)).

The main technical contribution of our paper is to provide two new stochastic expansions to verify such conditions. It thus establishes a connection between the extensive literature on estimation and inference in semiparametric models and the one on applications with generated covariates. Our expansions characterize the influence of generated covariates in the model's nonparametric component. We then show how to use this result to verify the above-mentioned uniform consistency and asymptotic normality conditions. Alternatively, our expansion could also be directly applied to a linearized version of the estimator. The expansions, which are proven using techniques from empirical process theory (e.g. Van der Vaart and Wellner, 1996; van de Geer, 2009), are related to results in Mammen, Rothe, and Schienle (2012) for purely nonparametric regression problems with generated covariates. The main difference is that in the present paper we derive sufficiently sharp bounds on weighted integrals of the remainder term instead of controlling its supremum norm. This requires substantially different mathematical methods. The new error bounds shrink at a considerably faster rate than those obtained in Mammen, Rothe, and Schienle (2012), which is critical for our development of a general theory of semiparametric estimation with generated covariates.

As a byproduct of the verification of the asymptotic normality condition mentioned above, we also obtain an explicit formula for the asymptotic variance of semiparametric estimators contained in the general class we consider. Compared to an infeasible procedure that uses the true values of the covariates, the influence function of such an estimator generally contains two additional terms: one that accounts for using generated covariates to estimate the nonparametric component, and one that accounts for the direct influence of generated covariates in other parts of the model, e.g. through determining the point of evaluation of the infinite-dimensional parameter. Additionally, we obtain a characterization of cases under which these two adjustment terms exactly offset each other, and thus do not affect first-order asymptotic theory. Our methods can also be used to verify conditions under which a bootstrap procedure leads to asymptotically valid inference. The latter aspect can be important in many applications where the asymptotic variance is difficult to estimate.

Our paper is related to an extensive literature on models with generated covariates. To the best of our knowledge, Newey (1984) and Murphy and Topel (1985) were among the first to study the theoretical properties of such two-step estimators in a fully parametric setting. Pagan (1984) and Oxley and McAleer (1993) provide extensive surveys. Nonparametric regression with (possibly nonparametrically) generated covariates is studied by Mammen, Rothe, and Schienle (2012) under general conditions. See their references for a list of examples, and Andrews (1995), Song (2008) and Sperlich (2009) for related results. Examples of semiparametric applications with generated covariates include Olley and Pakes (1996), Heckman, Ichimura, and Todd (1998), Li and Wooldridge (2002), Levinsohn and Petrin (2003), Blundell and Powell (2004), Linton, Sperlich, and Van Keilegom (2008), Rothe (2009), Escanciano, Jacho-Chávez, and Lewbel (2012) and Caetano, Rothe, and Yildiz (2014), among many others. Song (2013) studies a class of semiparametric models with generated covariates that have a single-index structure, and shows that adaptive estimation is possible in such models under weak conditions (see also Song (2012)). Hahn and Ridder (2013) study the form of the influence function of semiparametric GMM-type estimators with generated covariates. They start with the assumption that the estimator is $\sqrt{n}$-consistent and asymptotically linear (without explicitly specifying it), and then use arguments adapted from Newey (1994) to argue what the asymptotic variance of such a hypothetical estimator should be. In contrast, the focus of this paper is on giving explicit conditions that ensure that a concrete estimator is root-n consistent and asymptotic normal in the first place, and on showing how to establish validity of the bootstrap. Both aspects are important for implementing an estimator in practice. Escanciano, Jacho-Chávez, and Lewbel (2014) provide stochastic expansions for sample means of weighted residuals of semiparametric regressions with generated covariates, and certain uniform convergence results. Their results are useful for deriving asymptotic properties of certain semiparametric regression-type estimators, where the nonparametric component affects the final estimator solely through its value at the generated covariates. They also require a particular "index" condition, which can imply strong restrictions on the underlying economic model and affect the form of the asymptotic variance. No such restriction is necessary for our results

The remainder of the paper is structured as follows: In Section 2, we describe the class of models and estimators we consider, and outline how to establish their asymptotic properties. In Section 3, we present our main technical results: stochastic expansions that characterize the influence of generated covariates in the model's nonparametric component. Section 4 shows how these expansions can be used to verify classic conditions for $\sqrt{n}$-consistency and asymptotic normality of semiparametric estimators. In Section 5, we further study the asymptotic variance of our estimators, and show how to establish validity of the bootstrap. In Section 6, we discuss an application that makes use of our results. Finally, Section 7 concludes. All proofs and further details on the applications are collected in the Appendix.

Throughout the paper, we use the notation that for any vector $a \in \mathbb{R}^d$ the values $a_{min} = \min_{1\leq j\leq d} a_j$ and $a_{max} = \max_{1\leq j\leq d} a_j$ denote the smallest and largest of its elements, respectively, $a_+ = \sum_{j=1}^d a_j$ denotes the sum of its elements, $a_{-k} = (a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_d)$ denotes the $(d-1)$-dimensional subvector of a with the kth element removed, $a! = \prod_{j=1}^d a_j!$ denotes the product of the factorial of the elements of the vector (in case the latter are all non-negative integers), and $a^b = \prod_{j=1}^d a_j^{b_j}$ for any vector $b \in \mathbb{R}^d$. Except where specifically indicated otherwise, we denote the cumulative distribution function (CDF) and the density function of a generic random vector X by F_X and f_X, respectively.

2. Model, Estimation Procedure

We consider a general class of semiparametric optimization estimators where the criterion function depends on two types of infinite dimensional nuisance parameters: a conditional expectation function that has been estimated nonparametrically using generated covariates, and another estimated function that is used to compute the generated covariates in a first step. We allow the criterion function to depend on the latter estimated function directly to accommodate settings where it also serves another purpose other than determining the shape of the conditional expectation function. Our setup covers both parametrically and nonparametrically generated covariates, as well as intermediate cases. It also allows for non-smooth criterion functions and profiled estimation of the nonparametric components.

2.1. Model. Let $Z = (Y, X, W) \in \mathbb{R}^{d_Z}$ be a random vector defined on a complete probability space. Let $\Theta \subset \mathbb{R}^{d_\theta}$ denote a finite dimensional parameter space with generic element θ, and

$\Xi = \mathcal{M} \times \mathcal{R}$ be an infinite dimensional parameter space with generic element $\xi = (m, r)$. Denote by $\theta_0 \in \Theta$ and $\xi_0(\cdot, \theta) = (m_0(\cdot, \theta), r_0(\cdot)) \in \Xi$ the true values of the finite and infinite dimensional parameter, respectively.[1] We assume that there exists a nonrandom function $q : \text{supp}(Z) \times \Theta \times \Xi \to \mathbb{R}^{d_q}$ such that

$$Q(\theta, \xi_0(\cdot, \theta)) = \mathbb{E}(q(Z, \theta, \xi_0(\cdot, \theta))) = 0 \text{ if and only if } \theta = \theta_0.$$

The parametric component of our semiparametric model is thus identified via a moment condition. For simplicity, we also assume that for every $\xi \in \Xi$ the objective function $Q(\theta, \xi(\cdot, \theta))$ depends on the nuisance parameter ξ through its levels at values over some compact set $I_T^* \times I_R^*$ only, which is useful to later accommodate "fixed trimming" schemes into the estimation procedure. We also impose certain restrictions on the nature of the infinite dimensional parameter $\xi_0(\cdot, \theta) = (m_0(\cdot, \theta), r_0(\cdot))$. First, we assume that r_0 can be identified from the distribution of a random subvector W of Z alone. This allows for a consistent estimate of r_0 to be computed without knowledge of either θ_0 and m_0. Note that while in many applications r_0 will be a conditional expectation function, our setup does not require such a structure. Second, we assume that $m_0(\cdot, \theta)$ is a conditional expectation function that depends on $\theta \in \Theta$ and the true value r_0 through the relationship

$$m_0(\cdot, \theta) = \mathbb{E}(Y | T(X, \theta, r_0) = \cdot) \tag{2.1}$$

where $T(X, \theta, r) = t(X, r(X_r), \theta)$ is a random vector of dimension d_T, X_r is a random subvector of X that contains the covariates entering the function r, and $t : \mathbb{R}^{d_X} \times \mathbb{R}^{d_r} \times \Theta \to \mathbb{R}^{d_T}$ is a known function. The primary role of r_0 is thus to generate (some of) the covariates used to compute the function m_0. By allowing the shape of m_0 to depend on X and $r_0(X_r)$ through a known transformation indexed by θ, our setup includes a broad class of index models that require profiling of the nonparametric component.

Note that we allow r_0 to enter the objective function $Q(\theta, \xi_0(\cdot, \theta)) = Q(\theta, (m_0(\cdot, \theta), r_0(\cdot)))$ directly to accommodate settings where r_0 also serves a purpose other than determining the shape of m_0. The fact that m_0 is a functional of r_0 is *not imposed* at the level of the criterion function

[1] Note that the notation $\xi_0(\cdot, \theta) = (m_0(\cdot, \theta), r_0(\cdot))$ is understood to mean that $\xi_0(a, b, \theta) = (m_0(a, \theta), r_0(b))$ for every conformable (a, b).

Q. That is, for a generic function $r(\cdot)$ the expression $Q(\theta,(m_0(\cdot,\theta),r(\cdot)))$ is understood to mean $Q(\theta,(\bar{m}(\cdot,r_0,\theta),r(\cdot)))$ and *not* $Q(\theta,(\bar{m}(\cdot,r,\theta),r(\cdot)))$, where $\bar{m}(\cdot,r,\theta)=E(Y|T(X,\theta,r)=\cdot)$.

To make the notation more compact, we usually suppress the arguments of the infinite dimensional nuisance parameters, writing $(\theta,\xi)=(\theta,m,r)\equiv(\theta,m(\cdot,\theta),r(\cdot))$, $(\theta,\xi_0)=(\theta,m_0,r_0)\equiv(\theta,m_0(\cdot,\theta),r_0(\cdot))$, and $(\theta_0,\xi_0)=(\theta_0,m_0,r_0)\equiv(\theta_0,m_0(\cdot,\theta_0),r_0(\cdot))$. We also write $T(\theta,r)\equiv T(X,\theta,r)$, $T(\theta)\equiv T(\theta,r_0)$, $T(r)\equiv T(\theta_0,r)$ and $T\equiv T(\theta_0,r_0)$. We also write $\|B\|=(\text{tr}(B'AB))^{1/2}$ for any matrix B, where we suppress the dependence of the norm on the fixed symmetric positive definite matrix A for notational convenience.

2.2. Estimation Procedure. Given an i.i.d. sample $\{Z_1,\ldots,Z_n\}$ from the distribution of Z, a three-step semiparametric extremum estimator $\widehat{\theta}$ of θ_0 can be constructed as follows. In the first step, we compute a (possibly nonparametric) consistent estimate $\widehat{r}$ of r_0. We do not require a specific procedure for this step, but will only impose certain "high-level" restrictions below that cover a wide range of methods. These include nonparametric estimators based on either kernels or sieves, and fully parametric procedures, as well as intermediate cases. Given an estimate of r_0, we then compute an estimate of $m_0(\cdot,\theta)$ for every $\theta\in\Theta$ through a nonparametric regression of Y on the generated covariates $\widehat{T}(\theta)=t(X,\widehat{r}(X_r),\theta)$ using p-th order local polynomial smoothing. Our estimator is thus given by $\widehat{m}(t,\theta)=\widehat{\alpha}$, where

$$(\widehat{\alpha},\widehat{\beta})=\underset{\alpha,\beta}{\operatorname{argmin}}\sum_{i=1}^{n}\left(Y_i-\alpha-\sum_{1\leq u_+\leq p}\beta_u(\widehat{T}_i(\theta)-t)^u\right)^2 K_h(\widehat{T}_i(\theta)-t)\ . \tag{2.2}$$

Here $K_h(v)=\prod_{j=1}^{d_T}\mathcal{K}(v_j/h_j)/h_j$ is a d_T-dimensional product kernel built from the univariate kernel function $\mathcal{K}$, $h=(h_1,...,h_{d_T})$ is a vector of bandwidths that tend to zero as the sample size n tends to infinity, and $\sum_{1\leq u_+\leq p}$ denotes the summation over all vectors $u=(u_1,\ldots,u_p)$ of positive integers with $1\leq u_+\leq p$. For $p=1$, we get the usual local linear estimator, but our setup also allows for uneven orders $p>1$ for the purpose of bias control.[2] We focus on local polynomial

[2] Note that the definition of the estimator $\widehat{m}(\cdot,\theta)$ in (2.2) implicitly requires the generated covariates to be continuously distributed (see also Assumption 1(ii) below). This is not restrictive, however, as it would be straightforward to modify the estimator $\widehat{m}(\cdot,\theta)$ by the usual frequency method if some covariates are in fact discrete. Also note that for the special case that the objective function Q_n depends on $\widehat{m}(\cdot,\theta)$ through its values at the $\widehat{T}_i(\theta)$ only, one could

estimation for $m_0(\cdot,\theta)$ in this paper because the particular structure of the estimator facilitates controlling the presence of generated covariates (see Mammen, Rothe, and Schienle, 2012), and does not require a separate treatment of boundary regions. While it might be possible to conduct a similar analysis for other nonparametric procedures, such as orthogonal series estimators, we conjecture that this would require substantially more involved technical arguments. Finally, writing $(\theta,\widehat{\xi}) = (\theta,\widehat{m}(\cdot,\theta),\widehat{r}(\cdot))$, we define the estimator $\widehat{\theta}$ of θ_0 as any approximate solution to the problem of minimizing a semiparametric GMM-type objective function:

$$\|Q_n(\widehat{\theta},\widehat{\xi})\| = \inf_{\theta\in\Theta} \|Q_n(\theta,\widehat{\xi})\| + o_P(1/\sqrt{n}), \tag{2.3}$$

where $Q_n(\theta,\widehat{\xi}) = (1/n)\sum_{i=1}^n q(Z_i,\theta,\widehat{\xi})$. Our estimator is a semiparametric procedure involving generated covariates, in the sense that a preliminary estimate $\widehat{r}$ of the nuisance parameter r_0 is used to compute the covariates entering the nonparametric regression procedure to estimate $m_0(\cdot,\theta)$.

For the later asymptotic analysis, it will be useful to also consider an infeasible estimation procedure that uses the true value r_0 instead of an estimate $\widehat{r}$. Such an estimator $\widetilde{\theta}$ of θ_0 can be obtained by first computing an estimate $\widetilde{m}(\cdot,\theta)$ of $m_0(\cdot,\theta)$ via nonparametric regression of Y on $T(\theta)$ for every $\theta\in\Theta$. That is, it is given by $\widetilde{m}(t,\theta) = \widetilde{\alpha}$, where

$$(\widetilde{\alpha},\widetilde{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^n \left(Y_i - \alpha - \sum_{1\le u_+\le p} \beta_u (T_i(\theta)-t)^u \right)^2 K_h(T_i(\theta)-t).$$

One then defines $\widetilde{\theta}$ as an approximate minimizer of an infeasible version of the objective function:

$$\|Q_n(\widetilde{\theta},\widetilde{\xi})\| = \inf_{\theta\in\Theta} \|Q_n(\theta,\widetilde{\xi})\| + o_P(1/\sqrt{n}) \tag{2.4}$$

where $(\theta,\widetilde{\xi}) = (\theta,\widetilde{m}(\cdot,\theta),r_0(\cdot))$. In order to distinguish the two procedures, we refer to $\widehat{\theta}$ and $\widehat{m}$ in the following as the *real* estimators of θ_0 and m_0, respectively, and to $\widetilde{\theta}$ and $\widetilde{m}$ as the corresponding *oracle* estimators.

2.3. General Approach for Asymptotic Analysis. We now describe the general strategy of our asymptotic analysis (a formal result is given in Theorem 5 at the end of Section 4). Our approach

slightly simplify some technical arguments later by directly considering a "leave-one-out" version of $\widehat{m}(\cdot,\theta)$. Since our setup does not require such a structure, we proceed with the definition in (2.2).

builds on an extensive literature that has given "high-level" sufficient conditions for $\sqrt{n}$-consistency and asymptotic normality of semiparametric estimators. Examples include Newey (1994), Andrews (1994), Chen and Shen (1998), Ai and Chen (2003), Chen, Linton, and Van Keilegom (2003), or Ichimura and Lee (2010). A closer inspection of these conditions reveals that many of them can be verified irrespective of the presence of generated covariates, and thus through well-established (although still potentially highly involved) arguments. On the other hand, some of these sufficient conditions are strongly affected by the presence of generated covariates, and cannot be verified by standard techniques. Our main technical contribution in this paper is to develop stochastic expansions of nonparametric regression estimators for exactly this purpose.

We develop two types of results, which can be used to verify sufficient conditions for consistency and asymptotic normality of semiparametric estimators, respectively. For establishing consistency, the main problem that arises from the presence of generated covariates is to show that the nonparametric first-stage estimates are uniformly consistent. Under certain sufficient conditions that can be verified irrespective of the question whether generated covariates are present or not (see, e.g., Newey (1994) or Chen, Linton, and Van Keilegom (2003); see also Appendix C), one can show that $\widehat{\theta} \xrightarrow{p} \theta_0$ if

$$\|\widehat{\xi} - \xi_0\|_\Xi = o_P(1), \tag{2.5}$$

where $\|\cdot\|_\Xi$ denotes the pseudo-norm induced by the sup-norm on a class of continuous and bounded functions, i.e. we have $\|\xi\|_\Xi = \sup_\theta \sup_x |m(x,\theta)| + \sup_{x_r} |r(x_r)|$. In the following section, we provide new tools to verify a condition like (2.5), and then illustrate their application in Section 4.

In comparison, establishing $\sqrt{n}$-consistency and asymptotic normality of $\widehat{\theta}$ turns out to be more involved in the presence of generated covariates, as it requires stronger conditions on the nonparametric first stage. Under certain sufficient conditions that can be verified irrespective of the question whether generated covariates are present or not (see, e.g., Newey (1994) or Chen, Linton, and Van Keilegom (2003); see also Appendix C), the estimator $\widehat{\theta}$ satisfies the following representation:

$$\sqrt{n}(\widehat{\theta}-\theta_0) = -(Q_0^{\theta\top} A Q_0^{\theta})^{-1} Q_0^{\theta\top} A \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} q(Z_i,\theta_0,\xi_0) + \sqrt{n} Q_0^{\xi}[\widehat{\xi}-\xi_0] \right)$$
$$+ O_P(\sqrt{n}\|\widehat{\xi}-\xi_0\|_{\Xi}^2) + o_P(1).$$

Here $Q_0^{\theta} = Q^{\theta}(\theta_0,\xi_0)$ denotes the ordinary derivative of $Q(\theta,\xi)$ with respect to θ evaluated at (θ_0,ξ_0), and for any $\bar{\xi}$ such that $\xi_0+\tau\bar{\xi}\in\Xi$ for $|\tau|$ sufficiently small we put $Q_0^{\xi}[\bar{\xi}] = Q^{\xi}(\theta_0,\xi_0)[\bar{\xi}] = \lim_{\tau\to 0}(Q(\theta_0,\xi_0+\tau\bar{\xi})-Q(\theta_0,\xi_0))/\tau$ as the pathwise derivative of $Q(\theta_0,\xi)$ at ξ_0 in the direction $\bar{\xi}$.[3] This representation implies that $\widehat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal with asymptotic variance $\Omega = (Q_0^{\theta\top} A Q_0^{\theta})^{-1} Q_0^{\theta\top} A V A Q_0^{\theta} (Q_0^{\theta\top} A Q_0^{\theta})^{-1}$ if

$$\|\widehat{\xi}-\xi_0\|_{\Xi} = o_P(n^{-1/4}) \tag{2.6}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} q(Z_i,\theta_0,\xi_0) + \sqrt{n} Q_0^{\xi}[\widehat{\xi}-\xi_0] \xrightarrow{d} N(0,V) \tag{2.7}$$

for some positive definite variance matrix V. Verifying these two conditions is technically much more challenging than verifying (2.5). The estimate of the nuisance functions $\widehat{\xi}(\cdot,\theta) = (\widehat{m}(\cdot,\theta),\widehat{r}(\cdot))$ is required to be uniformly consistent with a particular rate, and a particular linear functional of $\widehat{\xi}-\xi_0$ needs to satisfy an asymptotic normality condition. In the following section, we provide new tools for verifying both (2.6) and (2.7), and we illustrate their application in Section 4.

3. Stochastic Expansions of the Nonparametric Component

This section contains our main technical results. We consider a stochastic expansion of nonparametrically estimated regression functions under very general conditions, deriving a bound on both weighted averages and the supremum norm of the remainder term that is sufficiently sharp for our purposes. These are the key tools for verifying conditions (2.7) and (2.5)–(2.6), respectively, in applications.

[3] To make the notation clear, define $\bar{m}(\cdot,r,\theta) = \mathbb{E}(Y|T(\theta,r)=\cdot)$, and note that $m_0(\cdot,\theta) = \bar{m}(\cdot,r_0,\theta)$. Then for any $\xi=(m,r)$ the pathwise derivative $Q_0^{\xi}[\xi]$ is defined as $Q_0^{\xi}[\xi] = \lim_{\tau\to 0}(Q(\theta_0,\bar{m}(\cdot,r_0,\theta_0)+\tau m(\cdot), r_0+\tau r) - Q(\theta_0,\bar{m}(\cdot,r_0,\theta_0),r_0))/\tau$, and *not* as $Q_0^{\xi}[\xi] = \lim_{\tau\to 0}(Q(\theta_0,\bar{m}(\cdot,r_0+\tau r,\theta_0)+\tau m(\cdot), r_0+\tau r) - Q(\theta_0,\bar{m}(\cdot,r_0,\theta_0),r_0))/\tau$. See also Linton, Sperlich, and Van Keilegom (2008).

3.1. Assumptions.

We now state our assumptions on the data generating process and the preliminary estimator $\widehat{r}$ of r_0. We begin by defining the generalized regression residual $\varepsilon(\theta) = Y - \mathbb{E}(Y|T(\theta))$. This definition allows us to write the dependent variable Y as $Y = m_0(T(\theta), \theta) + \varepsilon(\theta)$ with $\mathbb{E}(\varepsilon(\theta)|T(\theta)) = 0$.

Assumption 1 (Regularity). *We assume the following properties for the data distribution, the bandwidth, and kernel function $\mathcal{K}$.*

(i) The sample observations $\{Z_1, \ldots, Z_n\}$ are an independent and identically distributed sample from the distribution of Z.

(ii) The parameter space Θ is compact. For every $\theta \in \Theta$, the random vector $T(\theta) = t(X, r_0(X_r), \theta)$ is continuously distributed with support $I_{T,\theta}$ satisfying $I_T^* \subset int(I_{T,\theta})$ with I_T^* compact. The corresponding density function $f_T(x, \theta)$ has a continuous partial derivative with respect to x, and $\inf_{\theta \in \Theta, x \in I_T^*} f_T(x, \theta) > 0$.

(iii) The function $m_0(u, \theta)$ has continuous partial derivatives of order $p+1$ with respect to u for all $\theta \in \Theta$.

(iv) There exist a constant $C > 0$ and some constant $l > 0$ small enough such that for every $\theta \in \Theta$ the residuals $\varepsilon(\theta)$ satisfy the inequality $\mathbb{E}(\exp(l|\varepsilon(\theta)|)|T(\theta)) \leq C$.

(v) The function $\mathcal{K}$ is twice continuously differentiable and satisfies the following conditions: $\int \mathcal{K}(u)du = 1$, $\int u\mathcal{K}(u)du = 0$, and $\mathcal{K}(u) = 0$ for values of u not contained in some compact interval, say $[-1, 1]$.

(vi) The bandwidth $h = (h_1, \ldots, h_{d_T})$ satisfies $h_j \sim n^{-\eta_j}$ for all $j = 1, \ldots, d_T$, and $(1 - \eta_+)/2 > \eta_{\max}$.

Most restrictions imposed in Assumption 1 are standard for nonparametric kernel-type estimators of nuisance functions in semiparametric models. Part (i) is not necessary and could be relaxed to allow for certain forms of temporal dependence, albeit at the cost of substantially more involved theoretical arguments. Part (ii) states that the covariates $T(\theta)$ are continuously distributed, and that the density is bounded away from zero over some compact set I_T^*. The latter condition ensures that $\widehat{m}(\cdot,\theta)$ is a stable estimate over I_T^*. If it is known that the estimator $\widehat{\theta}$ is consistent, the parameter set Θ can be replaced in the assumptions by a local neighborhood of the true parameter. Then assumption (ii) is reasonable if the support $I_{T,\theta}$ smoothly changes with θ. The differentiability conditions in (iii) are used to control the magnitude of bias terms. Assuming subexponential tails of $\varepsilon(\theta)$ conditional on $T(\theta)$ in part (iv) is a regularity condition that is necessary to apply certain results from empirical process theory in our proofs. Note that conditions (ii)–(iv) involve the true function r_0 only. Unlike Escanciano, Jacho-Chávez, and Lewbel (2014, Assumption 3), we do not assume that e.g. the vector $T(\theta, r)$ or the conditional expectation $\mathbb{E}(Y|T(\theta, r))$ have particular distributional or smoothness properties for values of $r \in \mathcal{R}$ other than r_0. Part (v) describes a standard kernel function with compact support. Finally, the restrictions on the bandwidth in (vi) imply that the smoothing bias of the nonparametric regression estimator will be dominated by certain stochastic terms. As we will see from the next assumption, allowing the components of h to tend to zero at different rates can be useful in applications with multiple generated covariates that have different rates of convergence.

We remark that our setting can easily be extended to allow for random, data-dependent bandwidths. Allowing for a random bandwidth would only require to control the behavior of the mapping $(t,\theta) \mapsto \widehat{m}(t,\theta)$ as a function of h uniformly over some grid of bandwidth values that expands at a polynomial rate (Einmahl and Mason, 2005). To account for the presence of generated covariates, we are going to control the mapping $(t,\theta) \mapsto \widehat{m}(t,\theta)$ as a function of r uniformly over a much bigger space (see Assumption 3 below). Hence the extension to data-dependent bandwidths would cause no particular technical difficulties.

Assumption 2 (Accuracy). *We assume the following properties of the estimator $\widehat{r}$:*

(i) $\sup_s |\widehat{r}_j(s) - r_{0,j}(s)| = O_P(n^{-\delta_j^*})$ *for some* $\delta_j^* > 0$ *and all* $j = 1, \ldots, d_r$, *and*

(ii) $\sup_{\theta,x} |T_j(x,\theta,\widehat{r}) - T_j(x,\theta,r_0)| = o_P(n^{-\delta_j})$ *for some* $\delta_j > \eta_j$ *and all* $j = 1, \ldots, d_T$,

where in both cases the subscript j denotes the j-th component of the respective object.

Assumption 2 imposes restrictions on the accuracy of the first-step estimator $\widehat{r}$. Clearly, part (i) is a necessary condition for equation (2.5) to hold, and part (i) with $\delta_j^* > 1/4$ is necessary for condition (2.6). Part (ii) ensures that the difference between the respective components of $\widehat{T}(\theta)$ and $T(\theta)$ tend to zero in probability at a rate at least as fast as the corresponding bandwidth in the second stage of the estimation procedure, uniformly in θ. Such conditions can be verified for a wide range of nonparametric estimators (e.g. Masry (1996), Newey (1997)), and they trivially hold for regular parametric estimators.

Assumption 3 (Complexity)**.** *For every $j = 1, \dots, d_T$, there exist a sequence of sets of functions $\mathcal{T}_{n,j}$ that satisfies the following properties:*

- *(i)* $\Pr((x,\theta) \mapsto T_j(x,\theta,\widehat{r}) \in \mathcal{T}_{n,j}) \to 1$ *as $n \to \infty$.*
- *(ii) For some function r_n satisfying $\sup_{\theta,x} |T_j(x,\theta,r_n) - T_j(x,\theta,r_0)| = o(n^{-\delta_j})$ there exists a constant $C_T > 0$ such that the set $\mathcal{T}^*_{n,j} = \mathcal{T}_{n,j} \cap \{T_j(\cdot,r) : \sup_{\theta,x} |T_j(x,\theta,r) - T_j(x,\theta,r_n)| \leq n^{-\delta_j}$ and $r \in \mathcal{R}\}$ can be covered by at most $C_T \exp(\lambda^{-\alpha_j} n^{\chi_j})$ balls with $\|\cdot\|_\infty$-radius λ for all $\lambda \leq n^{-\delta_j}$, where $0 < \alpha_j \leq 2$, $\chi_j \in \mathbb{R}$ and $\|\cdot\|_\infty$ denotes the supremum norm.*

Assumption 3 restricts the complexity of the function space in which the mapping $(x,\theta) \mapsto T(x,\theta,\widehat{r})$ takes its values by imposing constraints on the cardinality of the covering sets. Since we have that $T(x,\theta,r) = t(x, r(x_r), \theta)$ for some known function t which, by Assumption 1(iii), is continuously differentiable with respect to its second component, the condition imposes implicit restrictions on the complexity of the first-stage estimator $\widehat{r}$. Indeed, if the function t has a partial derivative with respect to its second argument that is bounded and bounded away from zero, we could equivalently state a restriction similar to Assumption 3 on the set $\mathcal{R}^*_n = \{r \in \mathcal{R} : T_j(\cdot, r) \in \mathcal{T}^*_{n,j}$ for all $j = 1, \dots, d_T\}$.

Restrictions on covering numbers are a common requirement in the literature on empirical processes, that is typically fulfilled under suitable smoothness assumptions. Suppose for example that $\mathcal{R}^*_n$ is the set of smooth functions defined on the convex set $I_R \subset \mathbb{R}^{d_{X_r}}$, whose partial derivatives up to order k exist and are uniformly bounded by some multiple of $n^{\chi^*_j}$ for some $\chi^*_j \geq 0$, that $\|\partial^l T_j(x, r(x_r), \theta)/\partial x_r - \partial^l T_j(x, r(x_r), \theta^*)/\partial x_r\| \leq C_l \|\theta - \theta^*\|$ for every θ, θ^*, every value of x and r, and every $l \in \{0, \dots, k\}$, and that t has a absolutely bounded partial derivative with respect to its second argument. Then the set $\mathcal{T}_{n,j} = \{(x,\theta) \mapsto T_j(x,\theta,r) : r \in \mathcal{R}^*_n\}$ satisfies Assumption 3(ii) with $\alpha_j = d_{X_r}/k$ and $\chi_j = \chi_j^* \alpha_j$ (Van der Vaart and Wellner, 1996, Theorem 2.7.1). The same entropy bound applies if $\mathcal{R}_n^*$ consists of the sum of one fixed function and a smooth function from a respective smoothness class. This extension is useful if one chooses the fixed function as equal to the sum of r_0 and the bias of $\widehat{r}$, and thus does not require the bias term to be a smooth function. For further discussion of entropy bounds and additional references we refer to van de Geer (2009).

For kernel-based estimators of r_0, one can then verify Assumption 3(i) by explicitly calculating the derivatives. Consider for example the one-dimensional Nadaraya-Watson estimator $\widehat{r}_{n,j}$ with bandwidth of order $n^{-1/5}$ over some compact subset of the interior of the covariate's support where its density is bounded away from zero. Choose $r_{n,j}$ equal to $r_{0,j}$ plus its asymptotic bias term. Then one can check that the second derivative of $\widehat{r}_{n,j} - r_{n,j}$ is absolutely bounded by $O_P(\sqrt{\log n}) = o_P(n^{\chi_j^*})$ for all $\chi_j^* > 0$ over some compact set in the interior of the support of the respective conditioning variables. For sieve and orthogonal series estimators, Assumption 3(i) immediately holds when the set $\mathcal{T}_{n,j}$ is chosen as the image of the sieve set or a subset of the linear span of an increasing number of basis functions, respectively, under the functional $T(x, \theta, \cdot)$. That is, if $\widehat{r}(\cdot) = P_{k(n)}(\cdot)^\top \widehat{\gamma}$ for some $\widehat{\gamma} \in \mathbb{R}^{k(n)}$ and $P_{k(n)}(\cdot) = (p_1(\cdot), \ldots, p_{k(n)}(\cdot))^\top$ with $\{p_j\}_{j=1}^\infty$ a complete basis for the space $\mathcal{R}$, then Assumption 3(i) holds if we define $\mathcal{T}_{n,j} = \{(x, \theta) \mapsto T(x, \theta, P_k(\cdot)^\top \gamma) : \gamma \in \mathbb{R}^{k(n)}\}$. Note that in settings where r_0 is estimated by parametric or semiparametric methods verifying Assumption 3 is generally much more simple, and substantially smaller values can be established for the constants α_j and χ_j.

To state our final assumption, we define the "index bias" $\rho(X, \theta) = \mathbb{E}(Y|X) - \mathbb{E}(Y|T(\theta))$, which is the difference between the conditional expectations of Y given the underlying d_X-dimensional covariate vector X and the d_T-dimensional "index" $T(\theta)$, respectively.

Assumption 4 (Continuity)**.** *The elements of $\mathcal{R}_n^* = \{r \in \mathcal{R} : T_j(\cdot, r) \in \mathcal{T}_{n,j}^*$ for all $j = 1, \ldots, d_T\}$ satisfy the following properties for n large enough:*

(i) For all $r \in \mathcal{R}_n^*$ and $\theta \in \Theta$, the function $\tau^B(t, \theta, r) = \mathbb{E}(\rho(X, \theta)|T(r) = t)$ is $p + 1$ times differentiable with respect to its first argument, and the derivatives are uniformly bounded in absolute value over r, θ and t.

(ii) For a constant $C_B^* > 0$, all $\theta \in \Theta$ and all $r_1, r_2 \in \mathcal{R}_n^*$ it holds that

$$|\tau^B(T(r_1), \theta, r_1) - \tau^B(T(r_2), \theta, r_2)| \leq C_B^* \|T(r_1) - T(r_2)\| \ \textit{a.s.}$$

(iii) For a constant $C_B > 0$, all $\theta \in \Theta$, all $r_1, r_2 \in \mathcal{R}_n^*$ and all $t \in I_T^*$ it holds that

$$\begin{aligned} \big| \mathbb{E}\left[(T(\theta, r_1) - t)^u h^{-u} K_h(T(\theta, r_1) - t)\right] \\ - \mathbb{E}\left[(T(\theta, r_2) - t)^u h^{-u} K_h(T(\theta, r_2) - t)\right]\big| \leq C_B n^{-\delta_{min}} \end{aligned}$$

for $0 \leq u_+ \leq p$.

Assumption 4(i)–(ii) are technical conditions which ensure that a conditional expectation of the "index bias" $\rho(X, \theta)$ satisfies certain smoothness restrictions. Under the additional assumption that the mapping $r \mapsto T(r)$ is smooth, one important implication of these conditions and our previous assumptions is that the functional $r \mapsto \mathbb{E}(Y|T(X, \theta, r) = \cdot)$ is uniformly continuous in a neighborhood of r_0. To see this, drop the dependence on θ for a moment, put $\varepsilon^* = \varepsilon - \rho(X)$, and write

$$\begin{aligned} \mathbb{E}(Y|T(X, r)) &= \mathbb{E}(m_0(T(r_0)) - m_0(T(r))|T(r)) + \mathbb{E}(m_0(T(r))|T(r)) + \mathbb{E}(\varepsilon^*|T(r)) + \mathbb{E}(\rho(X)|T(r)) \\ &= \mathbb{E}(m_0(T(r_0)) - m_0(T(r))|T(r)) + m_0(T(r)) + \mathbb{E}(\rho(X)|T(r)). \end{aligned}$$

Continuity of this expression with respect to r then follows from our assumptions on smoothness of the mappings $t \mapsto m_0(t)$ and $r \mapsto E(\rho(X)|T(r))$. Note that our assumptions do not require the functional $r \mapsto \mathbb{E}(Y|T(X, \theta, r) = \cdot)$ to be pathwise differentiable.

It is difficult to give more "low-level" conditions for Assumption 4(i)–(ii) in general, but there are certain settings where $\rho(X, \theta) = 0$ and thus these conditions trivially hold . Examples of such settings include many instrumental variable models. In general, however, it is undesirable to impose that $\rho(X, \theta) = 0$, and we do not require such a condition for our analysis. See our Section 6 below for an application where this flexibility is important.

Assumption 4(iii) is a further smoothness condition. If the random vector $T(\theta, r)$ is continuously distributed for every r, conditions (ii) and (iii) hold if one has appropriate bounds for $\|f_1 - f_2\|_\infty$ for $r_1, r_2 \in \mathcal{R}_n^*$ where for $j = 1, 2$ the term f_j denotes the density function of either $T(\theta, r_j)$ or of $(\rho(X, \theta), T(\theta, r_j))$, and the density function $T(\theta, r_j)$ is bounded away from zero uniformly over θ on its support.

3.2. Main Results. Under the assumptions described in the previous subsection, we can now derive a stochastic approximation of the nonparametric estimator $\widehat{m}$. To state the result, we require some further notation. For any $s \in \{0, 1, \ldots, p\}$ let $n_s = \binom{s+d_T-1}{d_T-1}$ be the number of distinct d_T-tuples u with $u_+ = s$. Arrange these d_T-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 $s \in \{0, 1, \ldots, p\}$ we also define a $n_s \times 1$ vector $w_{i,s}(t, \theta, r)$ with its kth element given by $((T_i(\theta, r) - t)/h)^{\tau_s(k)}$, and write $w_i(t, \theta, r) = (1, w_{i,1}(t, \theta, r)^\top, \ldots, w_{i,p}(t, \theta, r)^\top)^\top$. Next, define $N_h(t, \theta, r) = \mathbb{E}(w_i(t, \theta, r) w_i(t, \theta, r)^\top K_h(T_i(\theta, r) - t))$ and let $m_{pol}(a, t, \theta)$ be the following polynomial approximation of $m_0(a, \theta)$ in a neighborhood of t:

$$m_{pol}(a, t, \theta) = \sum_{0 \leq u_+ \leq p} \frac{1}{u!} \frac{\partial^u m_0(t, \theta)}{\partial t_1^{u_1} \ldots \partial t_{d_T}^{u_{d_T}}} (a - t)^u.$$

Finally, let $m'_{pol}(a, t, \theta)$ denote the vector of partial derivatives of $m_{pol}(a, t, \theta)$ with respect to the components of its first argument, write $e_1 = (1, 0, \ldots, 0)^\top$ for the first unit vector in $\mathbb{R}^N$, where $N = \sum_{s=0}^p n_s$, put $K'_h(v) = (\mathcal{K}'_{h,1}(v), \ldots, \mathcal{K}'_{h,d_T}(v))^\top$ with elements

$$\mathcal{K}'_{h,j}(v) = (\mathcal{K}'(v_j/h_j)/h_j^2) \prod_{j^* \neq j} \mathcal{K}(v_{j^*}/h_{j^*})/h_{j^*},$$

and $\mathcal{K}'$ derivative of $\mathcal{K}$, and recall that $\rho(X, \theta) = \mathbb{E}(Y|X) - \mathbb{E}(Y|T(\theta))$. With this notation, we can then define the approximating function $\widehat{m}_\Delta$ by

$$\widehat{m}_\Delta(t, \theta) = \widetilde{m}(t, \theta) + \varphi_n^A(t, \theta, \widehat{r}) + \varphi_n^B(t, \theta, \widehat{r}), \tag{3.1}$$

where

$$\varphi_n^A(t, \theta, r) = e_1^\top N_h(t, \theta)^{-1} \mathbb{E}\left(w_i(t, \theta, r) K_h(T_i(\theta) - t) m'_{pol}(T_i(r), t, \theta)(T_i(\theta, r) - T_i(\theta))\right),$$
$$\varphi_n^B(t, \theta, r) = e_1^\top N_h(t, \theta)^{-1} \mathbb{E}\left(w_i(t, \theta, r) K'_h(T_i(\theta) - t)^\top (T_i(\theta, r) - T_i(\theta)) \rho(X_i, \theta)\right)$$

for any $r \in \mathcal{R}_n^*$.

The function $\widehat{m}_\Delta$ consists of two components: the term $\widetilde{m}(\cdot, \theta)$ is the oracle estimator of $m_0(\cdot, \theta)$ introduced above, whereas $\varphi_n^A(t, \theta, \widehat{r}) + \varphi_n^B(t, \theta, \widehat{r})$ is an adjustment term that captures the additional uncertainty due to the presence of generated covariates. Note that the generated covariates enter the expansion only through *smoothed* versions of the estimation error $T(\theta, \widehat{r}) - T(\theta, r_0)$. Since this additional smoothing typically improves the rate of convergence of the stochastic part of the first-step estimator (although it does not improve the order of the bias component), we generally expect the adjustment term to have a faster rate of convergence. Hence the dimensionality of the generation step should play a less pronounced role in this context. Our main result concerns the accuracy of using $\widehat{m}_\Delta$ as an approximation of $\widehat{m}$.

Theorem 1. *Suppose that Assumption 1, 2(ii), 3 and 4 hold. Then uniformly for $\theta \in \Theta$, we have*

$$\int (\widehat{m}(t,\theta) - \widehat{m}_\Delta(t,\theta))\omega(t)dt = o_P(n^{-\kappa^*}) \tag{3.2}$$

for any weight function $\omega : \mathbb{R}^{d_T} \to \mathbb{R}$ whose partial derivatives of order one are uniformly absolutely bounded, and that satisfies $\omega(x) = 0$ for all $x \notin I_T^*$, and $\kappa^* < \min\{\kappa_1^*, \ldots, \kappa_4^*\}$ with

$$\kappa_1^* = \frac{1}{2} + (1 - \frac{\alpha_{max}}{2})\delta_{min} - \frac{(\alpha\eta + \chi)_{max}}{2}, \ \kappa_2^* = (p+1)\eta_{min} + (\delta - \eta)_{min},$$
$$\kappa_3^* = (2 - \frac{\alpha_{max}}{2})\delta_{min} + \frac{1}{2}(1 - \eta_+) - \frac{(\alpha\eta + \chi)_{max}}{2}, \ \kappa_4^* = 2\delta_{min}.$$

The theorem provides a bound on weighted averages of the approximation error $\widehat{m}(t,\theta) - \widehat{m}_\Delta(t,\theta)$. We focus on such weighted averages of the approximation error because they are helpful when it comes to verifying conditions of the type (2.7). In particular, they can be shown to vanish faster than $n^{-1/2}$ under reasonable conditions on the primitives of the model. On the other hand, bounds on the supremum norm of the approximation error, as studied in Mammen, Rothe, and Schienle (2012), typically vanish at a rate slower than $n^{-1/2}$, and are thus not useful to establish the "asymptotic normality" condition. They can however, with some adaptation, be employed to verify the conditions (2.5) and (2.6), as explained below. For this purpose, we state the following theorem, which is a variation of an earlier result in Mammen, Rothe, and Schienle (2012) that gives a uniform rate of consistency of the estimator $\widehat{m}(t,\theta)$. See also Escanciano, Jacho-Chávez, and Lewbel (2014, Appendix B) for a related result.

Theorem 2 (Uniform Consistency)**.** *Suppose Assumption 1, 2(ii), 3 and 4(i)–(ii) hold. Then*

$$\sup_{t \in I_T^*, \theta \in \Theta} |\widehat{m}(t,\theta) - m_0(t,\theta)| = O_P\left(n^{-(p+1)\eta_{min}} + \sqrt{\log(n) n^{-(1-\eta_+)}} + n^{-\delta_{min}} + n^{-\kappa}\right),$$

where $\kappa < \min\{\kappa_1, ..., \kappa_3\}$ *with*

$$\kappa_1 = \frac{1}{2}(1 - \eta_+) + (\delta - \eta)_{min} - \frac{1}{2}(\delta\alpha + \chi)_{max}, \ \kappa_2 = (p+1)\eta_{min} + (\delta - \eta)_{min},$$
$$\kappa_3 = \delta_{min} + (\delta - \eta)_{min}.$$

Note that the first two terms in the error bound on the right hand side follow from a standard uniform consistency result of the oracle estimator $\widetilde{m}$ (Masry, 1996), whereas the remaining two terms are due to the presence of generated covariates. We remark that the rates given in Theorem 2 could be improved under additional restrictions on the form of the estimator $\widehat{r}$, such as those given in Assumption 5 below. See the remark at the end of the proof of Theorem 2 in Appendix A for details.

4. Application to Semiparametric Estimation

In this section, we show how to use the technical results of the previous section to establish $\sqrt{n}$-consistency and asymptotic normality of the semiparametric estimator $\widehat{\theta}$ defined in (2.3). In particular, we show how to verify the uniform consistency conditions given in (2.5) and (2.6), and the asymptotic normality condition in (2.7). We begin with the former two uniform consistency conditions, as they are conceptually simpler to establish. Recall that our aim is to show that

$$\|\widehat{\xi} - \xi_0\|_\Xi := \sup_{t \in I_T^*, \theta \in \Theta} |\widehat{m}(t,\theta) - m_0(t,\theta)| + \sup_{s \in I_R^*} |\widehat{r}(s) - r_0(s)| = o_P(a_n), \tag{4.1}$$

with either $a_n = 1$ (as in (2.5)) or $a_n = n^{-1/4}$ (as in (2.6)). While a bound on the second supremum term in the above equation is standard, a bound on the first one follows from our Theorem 2.

Theorem 3. *(i) Suppose that Assumptions 1–3 and Assumption 4(i)–(ii) hold, and that* $\eta_+ < \min(1, 1 + 2(\delta - \eta)_{min} - (\delta\alpha + \chi)_{max})$*. Then condition* (2.5) *holds. (ii) Suppose that the conditions of part (i) hold with* $\delta_j^*, \delta_j > 1/4$ *for all* j*, that* $\eta_{\min} > 1/(4(p+1))$*, and that* $\eta_+ < \min(1/2, 1/2 + 2(\delta - \eta)_{\min} - (\delta\alpha + \chi)_{\max})$*. Then condition* (2.6) *holds.*

We remark that part (i) of the theorem could also be shown if one would weaken Assumption 2 and allow for $0 < \delta_j < \eta_j$, $j = 1, \ldots, d_T$, if one in turn ensures that $-(\delta - \eta)_{\min} < \min(\delta_{\min}, (p+1)\eta_{\min})$. The various conditions of the theorem involve a tradeoff between the complexity of the first and second estimation step for the nonparametric component. They can be shown to be satisfied when r_0 is "sufficiently regular" (i.e. the α_j and χ_j are small) and $m_0(\cdot, \theta)$ is "sufficiently smooth" (i.e. p is large and thus the η_j can be chosen small). Exact conditions are difficult to give in general, but are easy to check for a specific application, where specific values for the α_j and χ_j are available. See the discussion after Assumption 3 above for an example.

To verify the asymptotic normality condition (2.7), we make use of the stochastic expansion derived in Theorem 1. Recall that our aim is to show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} q(Z_i, \theta_0, \xi_0) + \sqrt{n} Q_0^{\xi}[\widehat{\xi} - \xi_0] \xrightarrow{d} N(0, V)$$

for some positive definite variance matrix V. Given a specific estimator $\widehat{r}$ of r_0, the term $\widehat{m}_\Delta(t, \theta)$ defined in (3.1) can usually be calculated more explicitly, and then be used to verify this condition, and to obtain a general formula for the variance matrix V. To illustrate this idea in a general setting, suppose that the estimator used to generate the covariates satisfies the following asymptotically linear representation, which is similar to conditions used e.g. in Rothe (2009) or Ichimura and Lee (2010). The assumption can be shown to be satisfied for a wide range of nonparametric, semiparametric, and fully parametric estimation procedures (we also discuss two representative examples below).[4]

Assumption 5 (Linear Representation). *The estimator $\widehat{r}$ of r_0 satisfies*

$$\widehat{r}(s) - r_0(s) = \frac{1}{n} \sum_{i=1}^{n} \varphi_{ni}^{\widehat{r}}(s) + R_n^r(s) \tag{4.2}$$

with $\varphi_{ni}^{\widehat{r}}(s) = \mathcal{H}_n(S_i, s)\nu(W_i)$ for some S_i, a random subvector of W_i, and $\sup_{s \in I_R^*} |R_n^r(s)| = o_P(n^{-1/2})$. The term $\nu(W_i)$ satisfies $\mathbb{E}(\nu(W_i)|S_i) = 0$ and $\mathbb{E}(\nu(W_i)\nu(W_i)^\top) < \infty$, and $\mathcal{H}_n$ is a weighting function satisfying $\mathbb{E}(\|\mathcal{H}_n(S_i, S_j)\|^2) = o(n)$ for $i \neq j$.

To see how this additional structure can be utilized for our purposes, recall that it follows from elementary rules for pathwise derivatives that

$$Q_0^{\xi}[\widehat{\xi} - \xi_0] = Q^m(\theta_0, \xi_0)[\widehat{m} - m_0] + Q^r(\theta_0, \xi_0)[\widehat{r} - r_0],$$

[4] Note that Assumption 5 is typically not satisfied for estimators that are not asymptotically Gaussian, such as the Maximum Score estimator of a single-index binary choice model, or other estimators that follow so-called cube-root asymptotics. See Song (2013) for a further discussion of this point.

where for any (θ, r) the functional $Q^m(\theta, \xi)[\bar{m}]$ is the pathwise derivative of $Q(\theta, (m, r))$ at m in the direction $\bar{m}$, and similarly for Q^r. As noted in Section 2, the notation is such that computing Q^r does *not* involve computing the pathwise derivative of the functional $r \mapsto E(Y|T(\theta, r) = \cdot)$. In most applications, the structure of the criterion function $Q(\theta, \xi) = \mathbb{E}(q(Z, \theta, m, r))$ is such that (with some abuse of notation) we have $q(Z, \theta_0, m_0, r_0) = q(Z, \theta, m_0(Z_m), r_0(Z_r))$. That is, the term $q(Z, \theta, m, r)$ only depends on the functions m and r through their value when evaluated at some random vectors Z_m and Z_r. Here Z_m and Z_r could be subvectors of the data Z, or known transformations thereof that might even involve m, r and θ (for example, we could have $Z_r = X_r$ and $Z_m = T(X, r(X_r), \theta)$). All econometric applications we consider in Section 6 below exhibit this structure. Its most important implication is that the pathwise derivatives of the criterion function are of the form

$$Q^m(\theta_0, \xi_0)[\widehat{m} - m_0] = \int \lambda_m(z)(\widehat{m}(z) - m_0(z))dF_{Z_m}(z), \tag{4.3}$$

$$Q^r(\theta_0, \xi_0)[\widehat{r} - r_0] = \int \lambda_r(z)(\widehat{r}(z) - r_0(z))dF_{Z_r}(z). \tag{4.4}$$

with

$$\lambda_m(z_m) = \mathbb{E}(\partial q(Z, \theta, m_0, r_0)/\partial m_0(Z_m, \theta_0)|Z_m = z_m)$$

$$\lambda_r(z_r) = \mathbb{E}(\partial q(Z, \theta, m_0, r_0)/\partial r_0(Z_r)|Z_r = z_r).$$

Note that a representation like (4.3)–(4.4) with square integrable functions λ_m and λ_r also follows from the Riesz representation theorem under more general conditions (e.g. Newey, 1994).

If λ_m and λ_r are sufficiently smooth, one can use Assumption 5 together with our main stochastic expansion to show that there exist fixed functions ψ_j with $\mathbb{E}(\psi_j(Z)) = 0$ and $\mathbb{E}(\psi_j(Z)\psi_j(Z)^\top) < \infty$ for $j = 1, 2, 3$ such that

$$\int \lambda_m(z)\widetilde{m}(z, \theta_0)dF_{Z_m}(z) = \frac{1}{n}\sum_{i=1}^n \psi_1(Z_i) + o_P(n^{-1/2})$$

$$\int \lambda_m(z)\left(\varphi_n^A(z, \theta_0, \widehat{r}) + \varphi_n^B(z, \theta_0, \widehat{r})\right) dF_{Z_m}(z) = \frac{1}{n}\sum_{i=1}^n \psi_2(Z_i) + o_P(n^{-1/2}),$$

$$\int \lambda_r(z)\frac{1}{n}\sum_{i=1}^n \varphi_{ni}^{\widehat{r}}(z)dF_{Z_r}(z) = \frac{1}{n}\sum_{i=1}^n \psi_3(Z_i) + o_P(n^{-1/2}).$$

Moreover, the properties of the remainder term $R_n^m(t) = \widehat{m}(t,\theta_0) - \widehat{m}_\Delta(t,\theta_0)$ established in Theorem 1 ensure, under suitable regularity conditions, that

$$\int \lambda_m(z) R_n^m(z) dF_{Z_m}(z) = o_P(n^{-1/2}).$$

If we now put $\psi_0(Z_i) = q(Z_i,\theta_0,\xi_0)$ and $\psi(z) = \sum_{j=0}^3 \psi_j(z)$, the above statements imply that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n q(Z_i,\theta_0,\xi_0) + \sqrt{n} Q_0^\xi[\widehat{\xi} - \xi_0] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(Z_i) + o_P(1)$$

Together with the Central Limit Theorem, the previous equation then implies that condition (2.7) holds with $V = \mathbb{E}(\psi(Z)\psi(Z)^\top)$. The following theorem formalizes this argument, and provides a general formula to compute the variance matrix V (we study the question how to explicitly compute V in the following section). To state the result, define $G(t) = \lambda_m(t) f_{Z_m}(t)/f_T(t)$, $G'(t) = \partial G(t)/\partial t$, $T^{(r)}(x) = \partial t(x,\theta_0,r_0(x_r))/\partial r_0(x_r)$ and $\lambda_m^*(x_r) = \mathbb{E}(T^{(r)}(X)(\rho(X)G'(T) + m_0'(T)G(T))|X_r = x_r)$.

Theorem 4. *Suppose Assumptions 1– 5 hold with $p+1 > d_T$,*

$$\frac{(\alpha\eta+\chi)_{max}}{2} < \min\{(1-\frac{\alpha_{max}}{2})\delta_{min}, (2-\frac{\alpha_{max}}{2})\delta_{min} + \frac{1}{2}(1-\eta_+)\}, \tag{4.5}$$

the criterion function satisfies (4.3)– (4.4) with $\lambda_m(\cdot)$ and $\lambda_r(\cdot)$ being $(p+1)$-times continuously differentiable, $(2p+2)^{-1} < \eta_j < (2d_T)^{-1}$ for $j = 1,\ldots,d_T$, and let

$$\begin{aligned}
\psi_0(Z_i) &= q(Z_i,\theta_0,\xi_0)\\
\psi_1(Z_i) &= \varepsilon_i \lambda_m(T_i) f_{Z_m}(T_i)/f_T(T_i)\\
\psi_2(Z_i) &= -\nu(W_i)\lim_{n\to\infty}\mathbb{E}(\lambda_m^*(X_r)\mathcal{H}_n(S_i,X_r)|S_i)\\
\psi_3(Z_i) &= \nu(W_i)\lim_{n\to\infty}\mathbb{E}(\lambda_r(Z_r)\mathcal{H}_n(S_i,Z_r)|S_i),
\end{aligned}$$

Then condition (2.7) holds with $V = \mathbb{E}(\psi(Z)\psi(Z)^\top)$, where $\psi(z) = \sum_{j=0}^3 \psi_j(z)$.

Restriction (4.5) involves a tradeoff between the complexity of the first and second estimation step for the nonparametric component that is analogous to the one discussed after the statement of Theorem 3.

With the results of this section, and a result from Chen, Linton, and Van Keilegom (2003), we are now ready to state the following theorem, which formally states the asymptotic properties of our semiparametric two-step estimator with generated covariates.

Theorem 5. *(i) Suppose that the conditions of Theorem 3(i) and Assumption C.1 in Appendix C hold. Then* $\widehat{\theta} \xrightarrow{p} \theta_0$. *(b) Suppose that the conditions of Theorems 3(ii) and 4 and Assumption C.2 in Appendix C hold. Then* $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega)$, *where* $\Omega = (Q_0^{\theta\top} A Q_0^{\theta})^{-1} Q_0^{\theta\top} A V A Q_0^{\theta} (Q_0^{\theta\top} A Q_0^{\theta})^{-1}$.

5. The Asymptotic Variance and the Bootstrap

In this section, we first provide some intuition for the form of the asymptotic variance of the estimator $\widehat{\theta}$, and illustrate how to evaluate the general formulas in Theorem 4 and 5 for several settings. We then discuss conditions under which valid inference on θ_0 can be conducted via the bootstrap.

5.1. The Asymptotic Variance. The argument in the previous subsection conveys some important intuition for the form of the asymptotic variance of $\widehat{\theta}$. Recall that under the conditions of Theorem 1 this variance is given by

$$\Omega = (Q_0^{\theta\top} A Q_0^{\theta})^{-1} Q_0^{\theta\top} A V A Q_0^{\theta} (Q_0^{\theta\top} A Q_0^{\theta})^{-1}$$

with $V = \mathbb{E}(\psi(Z)\psi(Z)^\top)$ and $\psi(z) = \sum_{j=0}^{3} \psi_j(z)$ as described in Theorem 4. In contrast, the asymptotic variance of the oracle estimator $\widetilde{\theta}$ can be shown to be

$$\widetilde{\Omega} = (Q_0^{\theta\top} A Q_0^{\theta})^{-1} Q_0^{\theta\top} A \widetilde{V} A Q_0^{\theta} (Q_0^{\theta\top} A Q_0^{\theta})^{-1}$$

with $\widetilde{V} = \mathbb{E}((\psi_0(Z)+\psi_1(Z))(\psi_0(Z)+\psi_1(Z))^\top)$, by simply setting $\widehat{r} = r_0$. The presence of generated covariates thus affects the asymptotic variance only through the additional summands $\psi_2(Z)$ and $\psi_3(Z)$ used to calculate V, as the weight matrix A is chosen by the econometrician and Q_0^{θ} is simply a population quantity. In particular, the term $\psi_2(Z)$ captures the additional uncertainty due to using generated covariates when *estimating* the function m_0, whereas the term $\psi_3(Z)$ accounts for *directly using* the generated covariates in other parts of the model, e.g. as a point of evaluation of an estimated function. A simple condition for the presence of generated covariates to be asymptotically negligible, i.e. that $\Omega = \widetilde{\Omega}$, is then of course that $\psi_2(Z) = -\psi_3(Z)$ with probability one. See Hahn and Ridder (2013) for a similar result based on different arguments.

An important practical issue is how to explicitly calculate V in the context of a concrete semiparametric model. It seems difficult to construct an estimator of V based on the formula in

Theorem 4 alone due to its high level of generality. However, in the context of a specific model a more explicit formula can usually be derived, and then used to construct a consistent sample analogue estimator of V (and thus of Ω). This requires explicit expressions for the various terms introduced in Assumption 5. We now give two examples for which this assumption is satisfied: the case where r_0 is a conditional expectation function estimated by nonparametric regression, and the case where $r_0(x_r) = \bar{r}(x_r, \vartheta_0)$ is a function known up to a finite dimensional parameter ϑ_0, for which there exists a regular asymptotically linear estimator. These are arguably the most important cases from an applied point of view. We refer to Kong, Linton, and Xia (2010) for general results on kernel-based M-estimators.

Example 1 (Nonparametric Regression)**.** Suppose that W is partitioned as $W = (D, S)$, and we have that $D = r_0(S) + \zeta$ with $\mathbb{E}(\zeta|S) = 0$. Consider a kernel-based nonparametric regression estimator $\widehat{r}$ of r_0, such as the Nadaraya-Watson or a local polynomial estimator. Then one can show that Assumption 5 holds under suitable smoothness conditions and choice of I_R^* with $\nu(W_i) = \zeta_i$ and $\mathcal{H}_n(S_i, s) = f_{S,n}(s)^{-1} L_g(S_i - s)$, where L is a kernel function and g is a bandwidth that tends to zero at an appropriate rate, and some $f_{S,n}(s) \to f_S(s)$ as $n \to \infty$. We then find that

$$\psi_2(Z_i) = -\zeta_i \lambda_m^*(S_i) \frac{f_{X_r}(S_i)}{f_S(S_i)} \quad \text{and} \quad \psi_3(Z_i) = \zeta_i \lambda_r(S_i) \frac{f_{Z_r}(S_i)}{f_S(S_i)}.$$

The form of $\psi_0(\cdot)$ and $\psi_1(\cdot)$ remains unchanged. □

Example 2 (Nonlinear Parametric Estimation)**.** Assume that $r_0(s) = \bar{r}(s, \vartheta_0)$ is a parametrically specified function (not necessarily a conditional expectation) known up to the finite dimensional parameter ϑ_0. Suppose there exists an estimator $\widehat{\vartheta}$ of ϑ_0 that satisfies

$$\widehat{\vartheta} - \vartheta_0 = \frac{1}{n} \sum_{i=1}^{n} \varphi^{\widehat{\vartheta}}(W_i) + o_P(n^{-1/2}),$$

where $\mathbb{E}(\varphi^{\widehat{\vartheta}}(W)) = 0$, $\mathbb{E}(\varphi^{\widehat{\vartheta}}(W)\varphi^{\widehat{\vartheta}}(W)^\top) < \infty$, that $\bar{r}(x_r, \vartheta)$ is continuously differentiable in its second argument with derivative $\bar{r}'(x_r, \vartheta) = \partial \bar{r}(x_r, \vartheta)/\partial \vartheta$. Then Assumption 5 is satisfied with $\nu(W_i) = \varphi^{\widehat{\vartheta}}(W_i)$ and $\mathcal{H}_n(S_i, s) = \bar{r}'(s, \vartheta_0)$, and thus

$$\psi_2(Z_i) = -\nu(W_i)\mathbb{E}(T^{(r)}(X)\bar{r}'(X_r, \vartheta_0)(\rho(X)G'(T) + m_0'(T)G(T)))$$

$$\psi_3(Z_i) = \nu(W_i)\mathbb{E}(\lambda_r(Z_r)\bar{r}'(Z_r, \vartheta_0)),$$

with $G(t) = \lambda_m(t) f_{Z_m}(t)/f_T(t)$ and $G'(t) = \partial G(t)/\partial t$. An important special case of this setting is the one where W is partitioned as $W = (D, S)$, we have that $D = \bar{r}(S, \vartheta_0) + \zeta$ with $\mathbb{E}(\zeta|S) = 0$ and $\mathbb{E}(\zeta^2|S) < \infty$, and $\widehat{\vartheta}$ is the nonlinear least squares estimator of ϑ_0. In such a setting, we would have that $\nu(W_i) = \mathbb{E}(\bar{r}'(S, \vartheta_0)\bar{r}'(S, \vartheta_0)^\top)^{-1}\bar{r}'(S_i, \vartheta_0)(D_i - r_0(S_i))$, under the usual regularity conditions. □

Using results like those in Example 1–2, one can then derive the asymptotic variance Ω of a wide range of semiparametric estimators by calculating the functions λ_m, λ_m^*, and λ_r. The following two examples give an explicit formula for Ω in particular classes of semiparametric estimators. These examples illustrate two important issues. First, they give some insight under which conditions the presence of generated covariates can be asymptotically negligible. Second, they show that the "index bias" $\rho(X) = \mathbb{E}(Y|X) - \mathbb{E}(Y|T)$ appears explicitly in the asymptotic variance of a large class of estimators, and thus assuming that $\rho(X) = 0$ as in Escanciano, Jacho-Chávez, and Lewbel (2014) can be restrictive.

Example 3 (Linear Estimator)**.** Consider a setup where $T(X, \theta, r) = (X_1, r(X_r))$ and the parameter of interest is $\theta_0 = \mathbb{E}(s(m_0(T)))$ for some known function s, and thus the criterion function is of the form $Q_n(\theta, m, r) = n^{-1}\sum_{i=1}^n s(m((X_{1i}, r(X_{ri})))) - \theta$. This setting is also considered in Hahn and Ridder (2013, Theorem 3). Suppose that r_0 is a nonparametric regression function satisfying $D = r_0(X_r) + \zeta$ with $\mathbb{E}(\zeta|X_r) = 0$. Applying Theorem 4 as in Example 1 above, we find that the asymptotic variance of the estimator $\widehat{\theta}$ is given by

$$\Omega = \mathbb{E}((\Psi_1 + \Psi_2)(\Psi_1 + \Psi_2)^\top)$$

where, writing $T = (X_1, r_0(X_r))$,

$$\Psi_1 = s(m_0(T)) - \theta + s'(m_0(T))\varepsilon,$$
$$\Psi_2 = -\zeta\mathbb{E}(s''(m_0(T))m_0^{(2)}(T)(Y - E(Y|T))|X_r)$$

with $m_0^{(2)}(t)$ the partial derivative of $m_0(t)$ with respect to the second component of t. In this simple setting, it is easy to give intuitive conditions under which the presence of generated covariates is asymptotically negligible. Note that the term $\Psi_2 = \psi_2(Z) + \psi_3(Z)$ accounts for the estimation error from using an estimate of r_0 instead of the actual function. This term is easily seen to be equal to zero if either $s(\cdot)$ is a linear function or if the index restriction $\mathbb{E}(Y|X) = \mathbb{E}(Y|T)$ holds.

Example 4 (Semiparametric Regression)**.** Consider a setup where the objective function is of the form $Q_n(\theta, m, r) = n^{-1}\sum_{i=1}^{n}(Y_i - m(T(X_i, \theta, r), \theta))s(X_i)$ for some known function s. This type of objective function occurs in many semiparametric regression problems, such as the estimation of single- or multi-index models with generated covariates by semiparametric maximum likelihood or semiparametric least squares (e.g. Rothe, 2009). Suppose again that the function r_0 is a nonparametric regression function that satisfies $D = r_0(X_r) + \zeta$ with $\mathbb{E}(\zeta|X_r) = 0$ and $\mathbb{E}(\zeta^2|X_r) < \infty$. Applying Theorem 4 as in Example 1, we find that the asymptotic variance of the estimator $\widehat{\theta}$ is equal to

$$\Omega = (Q_0^\theta)^{-1}\mathbb{E}((\Psi_1 + \Psi_2 + \Psi_3)(\Psi_1 + \Psi_2 + \Psi_3)^\top)(Q_0^\theta)^{-1},$$

where, writing $u(t) = \mathbb{E}(s(X)|T = t)$ and $u'(t) = \partial u(t)/\partial t$,

$$\begin{aligned}
\Psi_1 &= \varepsilon(s(X) - \mathbb{E}(s(X)|T)) \\
\Psi_2 &= -\zeta\mathbb{E}((s(X) - \mathbb{E}(s(X)|T))m_0'(T)T^{(r)}(X)|X_r) \\
\Psi_3 &= \zeta\mathbb{E}(u'(T)T^{(r)}(X)(\mathbb{E}(Y|X) - E(Y|T))|X_r).
\end{aligned}$$

The terms Ψ_2 and Ψ_3 account for the estimation error from using an estimate of r_0 instead of the actual function. In this setting there are generally no simple conditions under which the presence of generated covariates is asymptotically negligible. Still, the form of the asymptotic variance simplifies considerably if the index restriction $\mathbb{E}(Y|X) = \mathbb{E}(Y|T)$ holds, as $\Psi_3 = 0$ in this case.

5.2. Validity of the Bootstrap. In practice, inference based on combining an asymptotic normality result with an estimate of the asymptotic variance can be very complicated. Both V and Ω could be difficult to estimate since they depend on the nonparametrically estimated components of the model in a potentially nontrivial fashion. In such cases, resampling techniques like the ordinary nonparametric bootstrap can be used for tasks like obtaining confidence regions for the parameters of interest or critical values for certain hypothesis tests. Using results from Chen, Linton, and Van Keilegom (2003), our techniques can be used to establish the validity of such an approach. Consider for example a setting where the sample and population objective function are of the form $Q_n(\theta,\xi) = n^{-1}\sum_{i=1}^n q(Z_i,\theta,m(Z_{m,i},\theta),r(Z_{r,i}))$ and $Q(\theta,\xi) = \mathbb{E}(q(Z,\theta,m(Z_m,\theta),r(Z_r)))$, respectively. Let $\{Z_1^*,\ldots,Z_n^*\}$ be drawn with replacement from the original sample $\{Z_1,\ldots,Z_n\}$, let $\widehat{\xi}^*$ be the same estimator as $\widehat{\xi}$ but based on the bootstrap data, and put $Q_n^*(\theta,\xi) = n^{-1}\sum_{i=1}^n q(Z_i^*,\theta,m(Z_{m,i}^*,\theta),r(Z_{r,i}^*))$. Next, define the bootstrap estimator $\widehat{\theta}^*$ as any sequence that minimizes a GMM-type criterion function based on a recentered moment condition:

$$\|Q_n^*(\widehat{\theta}^*,\widehat{\xi}^*) - Q_n(\widehat{\theta},\widehat{\xi})\| = \inf_{\theta\in\Theta}\|Q_n^*(\theta,\widehat{\xi}^*) - Q_n(\widehat{\theta},\widehat{\xi})\| + o_{P^*}(1/\sqrt{n}).$$

Sufficient conditions for the asymptotic validity of this bootstrap procedures were studied by Chen, Linton, and Van Keilegom (2003). These conditions are mostly minor strengthenings of those in Appendix C, that can be verified irrespective of the presence of generated covariates. However, there are also two conditions that are affected by the presence of generated covariates, which are the following variants of (2.6) and (2.7), respectively:

$$\|\widehat{\xi}^* - \widehat{\xi}\|_\Xi = o_{P^*}(n^{-1/4}) \tag{5.1}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(q(Z_i^*,\widehat{\theta},\widehat{\xi}) - q(Z_i,\widehat{\theta},\widehat{\xi})\Big) + \sqrt{n}Q_0^\xi[\widehat{\xi}^* - \xi_0] \xrightarrow{d} N(0,V) \tag{5.2}$$

under the probability measure P^* implied by bootstrap sampling.[5] By adapting the discussion after Theorem B in Chen, Linton, and Van Keilegom (2003) in an obvious fashion, and applying a result from Giné and Zinn (1990), these two conditions can be verified in the same way we establish (2.6) and (2.7) above, and are thus immediate for a wide range of applications. We thus obtain the following result.

Theorem 6. *(a) Suppose that* (5.1)–(5.2) *and Assumption C.3 in Appendix C hold. Then* $\sqrt{n}(\widehat{\theta}^* -$

[5] We remark that any term that is of the order $o_{P^*}(a_n)$ is also automatically of the order $o_P(a_n)$. This is because for a generic positive statistic U_n the statement that $U_n = o_{P^*}(a_n)$ is equivalent to $P^*(U_n > ca_n) = o_P(1)$ for any $c > 0$, which by the boundedness of $P^*(U_n > ca_n)$ is in turn is equivalent to $\mathbb{E}(P^*(U_n > ca_n)) = o(1)$, which means that $U_n = o_P(a_n)$. The notation o_{P^*} is thus only used to make the intention of the statement more clear.

$\widehat{\theta})$ *converges in distribution to* $N(0,\Omega)$ *under the probability measure* P^* *implied by bootstrap sampling. (b) Under the conditions of Theorem 3(ii) and 4, the conditions* (5.1) *and* (5.2) *are fulfilled.*

6. Application to Treatment Effect Estimation

Semiparametric estimation problems with generated covariates occur in various fields of econometrics. In this subsection, we discuss one of these applications in greater detail, namely the estimation of average treatment effects under unconfoundedness via regression on the propensity score. To save space, we only sketch the construction of the estimator, and refer to Appendix B for details and regularity conditions. We also restrict attention to deriving asymptotic normality results, as showing consistency of the respective estimators only requires arguments that are very similar to those that would be used in the absence of generated covariates.

6.1. Model. Consider the potential outcomes framework, which is commonly used in the extensive literature on program evaluation (Imbens, 2004): Let Y_1 and Y_0 be the potential outcomes with and without program participation, respectively, $D \in \{0,1\}$ an indicator of program participation, $Y = Y_1 D + Y_0(1-D)$ be the observed outcome, X a vector of exogenous covariates, and let $\Pi(x) = \Pr(D=1|X=x)$ be the propensity score. A typical object of interest in this context is the average treatment effect (ATE), defined as

$$\theta_0 = \mathbb{E}(Y_1 - Y_0).$$

Since selection into the program may be nonrandom, this object cannot be obtained by simply comparing the average outcomes of treated and untreated individuals. However, when selection into the treatment is unconfounded, biases due to nonrandom selection into the program can be removed by conditioning on the propensity score (Rosenbaum and Rubin, 1983). That is, the condition that $Y_1, Y_0 \bot D|X$ implies that $Y_1, Y_0 \bot D|\Pi(X)$. Moreover, writing $\nu_d(\pi) = \mathbb{E}(Y|D=d, \Pi(X)=\pi)$, we have that $\nu_d(\pi) = \mathbb{E}(Y_d|\Pi(X)=\pi)$, and thus by the law of iterated expectations, the ATE is identified through the relationship

$$\theta_0 = \mathbb{E}(\nu_1(\Pi(X)) - \nu_0(\Pi(X))). \tag{6.1}$$

A similar argument can be made for other measures of program effectiveness (e.g. Heckman, Ichimura, and Todd, 1998). Estimating the ATE by a sample analogue of (6.1) requires nonparametric estimation of the functions $\nu_1(\pi)$ and $\nu_0(\pi)$. Since the propensity score is generally unknown and has to be estimated in a first stage, this fits into our framework with $Z \equiv (Y, X, (D, X))$, $r_0(X_r) \equiv \Pi(X)$, $t(X, r_0(X_r), \theta) \equiv (D, \Pi(X))$, $m_0(z_1) \equiv \nu_d(p)$ and $q(z, \theta, m_0, r_0) \equiv \nu_1(\Pi(x)) - \nu_0(\Pi(x)) - \theta$.

6.2. Estimator and Asymptotic Properties. Using the path-derivative approach of Newey (1994), Hahn and Ridder (2013) derived the form of the influence function of a hypothetical estimator in this problem that is assumed to satisfy an asymptotic linearity condition. Here we complement their result by giving explicit conditions for root-n consistency and asymptotic normality of a concrete estimator, which were thus far not known (Imbens, 2004). In particular, we consider the following sample version of (6.1) as a natural estimate of the ATE:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\nu}_1(\widehat{\Pi}(X_i)) - \widehat{\nu}_0(\widehat{\Pi}(X_i))),$$

where $\widehat{\Pi}(x)$ is the q-th order local polynomial estimator of $\Pi(x)$, and $\widehat{\nu}_d(\pi)$ is the local linear estimator of $\nu_d(\pi)$, computed using the first-stage estimates of the propensity score (alternatively, we could consider a parametric estimator for the propensity score, such as Probit). Here the binary covariate D is accommodated via the usual frequency method, i.e. the estimate $\widehat{\nu}_d$ is computed by local linear regression of Y_i on $\widehat{\Pi}(X_i)$ using the $n_d = \sum_{i=1}^{n} \mathbb{I}\{D_i = d\}$ observations with $D = d$ only. The following proposition gives the asymptotic properties of the estimator.

Proposition 1. *Suppose that the regularity conditions given in Appendix B hold. Then we have that* $\widehat{\theta} \xrightarrow{p} \theta_0$ *and* $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathbb{E}(\Psi(Y, D, X)^2)$*, where*

$$\Psi(Y, D, X) = \mu_1(X) - \mu_0(X) + \frac{D(Y - \mu_1(X))}{\Pi(X)} - \frac{(1 - D)(Y - \mu_0(X))}{1 - \Pi(X)} - \theta_0$$

is the influence function, and $\mu_d(x) = \mathbb{E}(Y|D = d, X = x)$ *for* $d = 0, 1$.

Under the conditions of the proposition the asymptotic variance of $\widehat{\theta}$ equals the semiparametric efficiency bound for estimating θ_0, which was obtained by Hahn (1998). The estimator obtained via regression on the estimated propensity score thus has the same first-order limit properties as other popular efficient estimators of the ATE under unconfoundedness, such as the propensity score reweighting estimator of Hirano, Imbens, and Ridder (2003) or the estimator in Hahn (1998). Note that the flexibility of our Assumption 4 plays an important role for deriving this result. If we were to assume that the "index bias" is equal to zero in this application, we would in fact impose the restriction that $\nu_d(x) = \mu_d(x)$, and thus restrict the distribution of potential outcomes.

7. Conclusions

In this paper, we have derived new tools for the analysis of semiparametric methods that require the use of generated covariates to estimate the nonparametric component. Our main technical results are two new stochastic expansions that characterize the influence of the generation step on the estimator of the nonparametric part. We show how these expansions can be used to verify classical conditions for the $\sqrt{n}$-consistency and asymptotic normality of semiparametric two-step estimators. Our results should be useful for researchers that wish to establish such results for the estimation procedures in their concrete applications.

A. Proofs of Main Results

A.1. Proof of Theorem 1. To simplify notation, we give a detailed proof only for the special case $d_T = 1$, i.e. $T = T(X, \theta, r)$ is a univariate random variable, but we will shortly comment how rates change for $d_T > 1$. The proof for higher-dimensional T is conceptually similar. The following notation is used throughout our proofs (some of it is simply a restatement of notation that we introduced before for the special case $d_T = 1$). The unit vector $(1, 0, \ldots, 0)^\top$ in $\mathbb{R}^N$, where $N = \sum_{s=0}^p n_s$, is denoted by e_1. We write

$$\begin{aligned} w_i(t,\theta,r) &= (1, (T_i(r,\theta) - t)/h, ..., (T_i(r,\theta) - t)^p/h^p)^\top, \\ M_h(t,\theta,r) &= \frac{1}{n}\sum_{i=1}^n w_i(t,r,\theta) w_i(t,r,\theta)^\top K_h(T_i(r,\theta) - t), \\ m_0^*(t,\theta) &= (m_0(t,\theta), hm_0'(t,\theta)/2, ..., h^p m_0^p(t,\theta)/p!)^\top. \end{aligned}$$

We also set $w_i(t,\theta) = w_i(t,\theta,r_0)$ and $\widehat{w}_i(t,\theta) = w_i(t,\theta,\widehat{r})$, and define $M_h(t,\theta)$ and $\widehat{M}_h(t,\theta)$ analogously. Finally, we put $N_h(t,\theta) = \mathbb{E}(M_h(t,\theta))$. Using $\varepsilon^*(\theta) = \varepsilon(\theta) - \rho(X,\theta)$, we can write

$$Y_i = m_0(T_i(\theta), \theta) + \varepsilon_i^*(\theta) + \rho(X_i, \theta)\,.$$

Note that $\mathbb{E}(\varepsilon^*(\theta)|X) = 0$ for any $\theta \in \Theta$. With this representation of the dependent variable, we define the following decompositions of both the real and the oracle estimator:

$$\widehat{m}(t,\theta) = \widehat{m}_A(t,\theta) + \widehat{m}_B(t,\theta) + \widehat{m}_C(t,\theta) + \widehat{m}_D(t,\theta) + \widehat{m}_E(t,\theta)$$
$$\widetilde{m}(t,\theta) = \widetilde{m}_A(t,\theta) + \widetilde{m}_B(t,\theta) + \widetilde{m}_C(t,\theta) + \widetilde{m}_D(t,\theta) + \widetilde{m}_E(t,\theta),$$

with respective components $\widehat{m}_j(t,\theta) = e_1^\top \beta_j(\theta,\widehat{r})$ and $\widetilde{m}_j(t,\theta) = e_1^\top \beta_j(\theta,r_0)$ defined for $j \in \{A,B,C,D,E\}$ as follows:

$$\begin{aligned}
\beta_A(t,\theta,r) &= \operatorname*{argmin}_\beta \sum_{i=1}^n (\varepsilon_i^*(\theta) - \beta^\top w_i(t,\theta,r))^2 K_h(T_i(\theta,r) - t),\\
\beta_B(t,\theta,r) &= \operatorname*{argmin}_\beta \sum_{i=1}^n (m_0(T_i(\theta,r_0),\theta) - m_0^*(t,\theta)^\top w_i(t,\theta,r_0) - \beta^\top w_i(t,\theta,r))^2 K_h(T_i(\theta,r) - t),\\
\beta_C(t,\theta,r) &= \operatorname*{argmin}_\beta \sum_{i=1}^n (m_0^*(t,\theta)^\top w_i(t,\theta,r_0) - m_0^*(t,\theta)^\top w_i(t,\theta,r) - \beta^\top w_i(t,\theta,r))^2 K_h(T_i(\theta,r) - t),\\
\beta_D(t,\theta,r) &= \operatorname*{argmin}_\beta \sum_{i=1}^n (m_0^*(t,\theta)^\top w_i(t,\theta,r) - \beta^\top w_i(t,\theta,r))^2 K_h(T_i(\theta,r) - t),\\
\beta_E(t,\theta,r) &= \operatorname*{argmin}_\beta \sum_{i=1}^n (\rho(X_i,\theta) - \beta^\top w_i(t,\theta,r))^2 K_h(T_i(\theta,r) - t).
\end{aligned}$$

We also denote the component-wise differences between the real and the oracle estimator by

$$R_{j,n}(t,\theta) = \widehat{m}_j(t,\theta) - \widetilde{m}_j(t,\theta) \text{ for } j \in \{A,B,C,D,E\}. \tag{A.1}$$

Finally, recall the definition that $\widehat{m}_\Delta(t,\theta) = \widetilde{m}(t,\theta) + \varphi_n^A(t,\theta,\widehat{r}) + \varphi_n^B(t,\theta,\widehat{r})$ given in (3.1). The statement of the theorem follows if for any $\theta \in \Theta$ the term $R_n(t,\theta) = \widehat{m}(t,\theta) - \widehat{m}_\Delta(t,\theta)$ satisfies

$$\int R_n(t,\theta)\omega(t)\,\mathrm{d}t = o_P(n^{-\kappa^*}) .$$

This is what we show in the following. To simplify the notation, we fix $\theta = \theta_0$ for the rest of the proof and we omit θ as an argument of functions. The proof can be easily extended to show that the results hold uniformly over $\theta \in \Theta$. To see this note that we show at several places that expansions hold uniformly over function classes. This can be easily extended to uniformity over the function class and $\theta \in \Theta$.

We will show that

$$\int R_{A,n}(t)\omega(t)\,\mathrm{d}t = O_P(n^{-\kappa_1^*}), \tag{A.2}$$
$$\int R_{B,n}(t)\omega(t)\,\mathrm{d}t = O_P(n^{-\kappa_2^*}), \tag{A.3}$$
$$\int R_{C,n}(t)\omega(t)\,\mathrm{d}t = \int \varphi_n^A(t,\widehat{r})\omega(t)\,\mathrm{d}t + O_P(n^{-\kappa_3^*} + n^{-\kappa_4^*}), \tag{A.4}$$
$$\int R_{E,n}(t)\omega(t)\,\mathrm{d}t = \int \varphi_n^B(t,\widehat{r})\omega(t)\,\mathrm{d}t + O_P(n^{-\kappa_1^*} + n^{-\kappa_2^*}). \tag{A.5}$$

where the terms $R_{j,n}$ are defined in (A.1) above. This directly implies the statement of the theorem since

$$\int (\widehat{m}(t) - \widetilde{m}(t))\omega(t)\,\mathrm{d}t = \sum_{j\in\{A,\ldots,E\}} \int R_{j,n}(t)\omega(t)\,\mathrm{d}t, \tag{A.6}$$

and $R_{D,n}(t) \equiv 0$ by construction.

We start with the proof of (A.2). In the following, let $c > 0$ be some generic constant which can take different values at each appearance. Furthermore V_n is a generic sequence of stochastically bounded random variables, again with different meaning at different appearances. Write $\Phi_i(t,r) = e_1^\top M_h(t,r)^{-1} w_i(t,r) K_h(T_i(r) - t)$ and $\Phi_i(r) = \int \Phi_i(t,r)\omega(t)\,\mathrm{d}t$. Furthermore let $L_h(T_i(r) - t) = K_h(T_i(r) - t)w_i(t,r)$ be a vector-valued kernel type function. Then it holds that

$$R_{A,n}(t) = \frac{1}{n}\sum_{i=1}^{n} \left(\Phi_i(t, r_0) - \Phi_i(t, \widehat{r})\right) \varepsilon_i^*.$$

Using elementary arguments, one can show that

$$M_h(T_i(r_1), r_1) - M_h(T_i(r_2), r_2) = V_n n^\eta |T_i(r_1) - T_i(r_2)|$$

for $r_1, r_2 \in \mathcal{R}_n^*$ and $1 \le i \le n$. With the help of this bound, we find that, for $r_1, r_2 \in \mathcal{R}_n^*$ and $1 \le i \le n$,

$$\begin{aligned}
&|\Phi_i(r_1) - \Phi_i(r_2)| \\
&\le \left| \int \left[e_1^\top M_h(t, r_1)^{-1} L_h(T_i(r_1) - t) - e_1^\top M_h(t, r_2)^{-1} L_h(T_i(r_2) - t) \right] \omega(t) dt \right| \\
&= \left| \int \left[e_1^\top M_h(T_i(r_1) - hu, r_1)^{-1} \omega(T_i(r_1) - hu) \right.\right. \\
&\qquad \left.\left. -\ e_1^\top M_h(T_i(r_2) - hu, r_2)^{-1} \omega(T_i(r_2) - hu) \right] L(u) du \right|, \\
&\le V_n n^\eta |T_i(r_1) - T_i(r_2)|.
\end{aligned}$$

For the case $d_T > 1$, one gets by similar arguments that

$$|\Phi_i(r_1) - \Phi_i(r_2)| \le V_n \max_{1\le j\le d_T} n^{\eta_j} |T_{i,j}(r_1) - T_{i,j}(r_2)|. \tag{A.7}$$

This last bound can be used to calculate a rough bound on the entropy $H_n(\lambda)$ of the class of functions $X_i \mapsto \Phi_i(r)$. Here, $\exp(H_n(\lambda))$ denotes the number of balls with radius λ that are necessary to cover the functions $X_i \mapsto \Phi_i(r)$. Using Assumption 3, the class of functions $T_j(\cdot, r)$ can be covered by $c\exp((\lambda V_n^{-1} n^{-\eta_j})^{-\alpha_j} n^{\chi_j})$ balls of radius $\lambda V_n^{-1} n^{-\eta_j}$. Thus we find that the entropy $H_n(\lambda) \le c\sum_{j=1}^{d_T} \lambda^{-\alpha_j} V_n^{\alpha_j} n^{\eta_j\alpha_j + \chi_j}$ $\le c\max_{1\le j\le d_T} \lambda^{-\alpha_j} n^{\eta_j\alpha_j+\chi_j}$ for some constant $c > 0$. This implies

$$\int_0^{C_n} H_n^{1/2}(\lambda) d\lambda \le V_n n^{-(1-\alpha_{max}/2)\delta_{min} + (\eta\alpha+\chi)_{max}/2}$$

for $C_n = n^{-\delta_{min}}$.

We now apply Theorem 8.13 in van de Geer (2009) with $\bar{Z}_\theta = n^{-1}\sum_{i=1}^n Z_{i,\theta}$, $Z_{i,\theta} = \Phi_i(r)\varepsilon_i^*$, $\theta = r$, $R = C_n = n^{-\delta_{min}}$, and a is the entropy bound above. Conditional on observations $X_1, ..., X_n$, we obtain an exponential bound for $\bar{Z}_\theta$ uniformly in $\mathcal{R}_n^*$ since $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\exp(\ell^*|\varepsilon_i^*|)|T_i] \leq C^*$ with probability tending to one, for some constants $C^*, \ell^* > 0$ due to Assumption 1 (iv). With standard arguments this yields

$$\sup_{r_1, r_2 \in \mathcal{R}_n^*} \frac{1}{n}\sum_{i=1}^n (\Phi_i(r_1) - \Phi_i(r_2))\varepsilon_i^* = O_P\left(n^{-(1/2)-(1-\alpha_{max}/2)\delta_{min}+(\eta\alpha+\chi)_{max}/2}\right). \tag{A.8}$$

Equation (A.8) then implies the desired result (A.2) because $\hat{r}_n \in \mathcal{R}_n^*$ with probability tending to one and thus

$$P\left(\left|\int R_{A,n}(t)\omega(t)dt\right| \leq \left|\sup_{r_1, r_2 \in \mathcal{R}_n^*} \frac{1}{n}\sum_{i=1}^n (\Phi_i(r_1) - \Phi_i(r_2))\varepsilon_i^*\right|\right) \to 1 \text{ as } n \to \infty.$$

For the proof of (A.3), note that for some non-negative integers a, b and constants $C_1, C_2 > 0$ it holds that $\left|m_0(T_i(r)) - m_0^*(t)^\top w_i(t, r)\right| \leq C_1 n^{-(p+1)\eta_{min}}$ and

$$\left|\frac{1}{n}\sum_{i=1}^n K_h(T_i(r_1) - t)w_{i,k}^a(t, r_1)w_{i,l}^b(t, r_1) - K_h(T_i(r_2) - t)w_{i,k}^a(t, r_2)w_{i,l}^b(t, r_2)\right| \leq C_2 n^{-(\delta-\eta)_{min}}$$

for components l, k and all $t \in I_T^*$ and $r, r_1, r_2 \in \mathcal{R}_n^*$. These two statements directly imply (A.3).

For the proof of (A.4), note that uniformly over $1 \leq i \leq n$, $t \in I_T^*$ and $r \in \mathcal{R}_n^*$ it holds that

$$m_0^*(t)^\top w_i(t, r_0) - m_0^*(t)^\top w_i(t, r) = m'_{pol}(T_i(r), t)(T_i(r) - T_i(r_0)) + O_P(n^{-2\delta_{min}}).$$

Substituting this expression into $R_{C,n}$, we find that

$$\int R_{C,n}(t)\omega(t)dt = \frac{1}{n}\sum_{i=1}^n \Phi_i^*(\widehat{r})(T_i(\widehat{r}) - T_i(r_0)) + O_P(n^{-2\delta_{min}}),$$

where

$$\Phi_i^*(r) = \int e_1^\top M_h(t, r)^{-1} L_h(T_i(r) - t) m'_{pol}(T_i(r), t)\omega(t)dt.$$

Furthermore, we have that

$$\int \varphi_n^A(t, \widehat{r})\omega(t)dt = \frac{1}{n}\sum_{i=1}^n \Phi_i^*(r_0)(T_i(\widehat{r}) - T_i(r_0)) + o_P(n^{-1/2}).$$

Thus, for (A.4) we have to show that

$$\frac{1}{n}\sum_{i=1}^n (\Phi_i^*(\widehat{r}) - \Phi_i^*(r_0))(T_i(\widehat{r}) - T_i(r_0)) = O_P(n^{-\kappa_3^*} + n^{-\kappa_4^*}). \tag{A.9}$$

Since $|T_i(r) - T_i(r_0)| = O_P(n^{-\delta_{min}})$ uniformly over $r \in \mathcal{R}_n^*$ and $1 \leq i \leq n$, for (A.9) one only has to prove that

$$|\Phi_i^*(r) - \Phi_i^*(r_0)| = O_P(n^{\delta_{min}-\kappa_3^*} + n^{\delta_{min}-\kappa_4^*})$$

uniformly for $r \in \mathcal{R}_n^*$ and $1 \le i \le n$. To see why the last claim holds, note that we can write:

$$\begin{aligned}
\Phi_i^*(r) - \Phi_i^*(r_0) = \int & e_1^\top [M_h(t,r)^{-1} L_h(T_i(r) - t) m'_{pol}(T_i(r), t) \\
& - M_h(t,r_0)^{-1} L_h(T_i(r_0) - t) m'_{pol}(T_i(r_0), t)] \omega(t) dt \\
= \int & e_1^\top [M_h(T_i(r) - hu, r)^{-1} \omega(T_i(r) - hu) m'_{pol}(T_i(r), T_i(r) - hu) \\
& - M_h(T_i(r_0) - hu, r_0)^{-1} \omega(T_i(r_0) - hu) m'_{pol}(T_i(r_0), T_i(r_0) - hu)] L(u) du.
\end{aligned}$$

First, it is easy to see that

$$\max_{1\le i\le n} \sup_{r\in\mathcal{R}_n^*} \sup_{t\in I_T^*} |\omega(T_i(r) - t) - \omega(T_i(r_0) - t)| = O_P(n^{-\delta_{min}}) \qquad \text{and}$$

$$\max_{1\le i\le n} \sup_{r\in\mathcal{R}_n^*} \sup_{t\in I_T^*} |m'_{pol}(T_i(r), T_i(r) - t) - m'_{pol}(T_i(r_0), T_i(r_0) - t)| = O_P(n^{-\delta_{min}})$$

due to the smoothness of the functions involved. It thus remains to consider the elements of the matrix $M_h(T_i(r) - t, r) - M_h(T_i(r_0) - t, r_0)$. Any such element is of the form

$$\frac{1}{n}\sum_{i=1}^n \left[(T_i(r) - t)^u h^{-u} K_h(T_i(r) - t)\right] - \left[(T_i(r_0) - t)^u h^{-u} K_h(T_i(r_0) - t)\right]$$

for some $0 \le u_+ \le p$. We thus show that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n & \left[(T_i(r) - t)^u h^{-u} K_h(T_i(r) - t)\right] \\
& - \left[(T_i(r_0) - t)^u h^{-u} K_h(T_i(r_0) - t)\right] = O_P(n^{\delta_{min} - \kappa_3^*} + n^{\delta_{min} - \kappa_4^*}).
\end{aligned} \tag{A.10}$$

uniformly over $r \in \mathcal{R}_n^*$. Because of Assumption 4(iii), we have that

$$\mathbb{E}\left[(T_i(r) - t)^u h^{-u} K_h(T_i(r) - t)\right] - \mathbb{E}\left[(T_i(r_0) - t)^u h^{-u} K_h(T_i(r_0) - t)\right] = O_P(n^{-\delta_{min}})$$

uniformly over $r \in \mathcal{R}_n^*$. Thus, for a proof of (A.10) it suffices to establish that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n & \left[(T_i(r) - t)^u h^{-u} K_h(T_i(r) - t)\right] - \mathbb{E}\left[(T_i(r) - t)^u h^{-u} K_h(T_i(r) - t)\right] \\
& - \left[(T_i(r_0) - t)^u h^{-u} K_h(T_i(r_0) - t)\right] - \mathbb{E}\left[(T_i(r_0) - t)^u h^{-u} K_h(T_i(r_0) - t)\right] \\
= {} & O_P(n^{\delta_{min} - \kappa_3^*} + n^{\delta_{min} - \kappa_4^*}).
\end{aligned}$$

The last claim follows from the same type of arguments used in the treatment of $R_{A,n}(t)$. Taken together, the above derivation shows that

$$\int R_{C,n}(t)\omega(t)\,\mathrm{d}t = \int \varphi_n^A(t, \widehat{r})\omega(t)\,\mathrm{d}t + o_P(n^{-\kappa_3^*} + n^{-\kappa_4^*}),$$

as claimed. It remains to show (A.5). Note that

$$\int R_{E,n}(t)\omega(t)\,\mathrm{d}t = \frac{1}{n}\sum_{i=1}^{n}[\Phi_i(\widehat{r}) - \Phi_i(r_0)]\rho(X_i).$$

Using the same reasoning as in the treatment of $R_{A,n}(t)$, and Assumption 4(i)–(ii), we find that

$$\frac{1}{n}\sum_{i=1}^{n}\Phi_i(r)(\rho(X_i) - \mathbb{E}[\rho(X_i)|T_i(r)]) - \Phi_i(r_0)(\rho(X_i) - \mathbb{E}[\rho(X_i)|T_i(r_0)]) = O_P(n^{-\kappa_1^*})$$

uniformly for $r \in \mathcal{R}_n^*$. Note that $\mathbb{E}[\rho(X_i)|T_i(r_0)] = 0$. We now use that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Phi_i(r)\mathbb{E}[\rho(X_i)|T_i(r)] &= \frac{1}{n}\sum_{i=1}^{n}\int e_1^\top M_h(t,r)^{-1}L_h(T_i(r)-t)\mathbb{E}[\rho(X_i)|T_i(r)]\omega(t)dt\\ &= \int \varphi_n^B(t)\omega(t)dt + O_P(n^{-\kappa_2^*})\end{aligned}$$

uniformly over $r \in \mathcal{R}_n^*$, and thus (A.5) holds. This concludes the proof of Theorem 1. □

A.2. Proof of Theorem 2. First, standard results in e.g. Masry (1996), imply that the oracle estimator $\widetilde{m}$ satisfies

$$\sup_{t\in I_T^*,\theta\in\Theta}|\widetilde{m}(t,\theta) - m_0(t,\theta)| = O_P\left(n^{-(p+1)\eta_{min}} + \sqrt{\log(n)n^{-(1-\eta_+)}}\right).$$

under the conditions of the theorem. Second, one can show that

$$\sup_{t\in I_T^*,\theta\in\Theta}|\widehat{m}(t,\theta) - \widehat{m}_\Delta(t,\theta)| = o_P(n^{-\kappa}). \tag{A.11}$$

The statement (A.11) is an extension of Theorem 1 in Mammen, Rothe, and Schienle (2012), which gives a stochastic expansion of a local linear estimator regression estimator with generated covariates, and the special case that $T(x,r,\theta) = r(x_r)$. Generalizing this result to higher order local polynomials and more general forms of T is conceptually straightforward, and thus a proof is omitted. With (A.11), the statement of the Theorem follows from a trivial bound on $\widehat{m}_\Delta(t,\theta) - \widetilde{m}(t,\theta)$. □

Remark 1. One could use the additional structure implied by Assumption 5 to prove a somewhat better uniform rate of consistency under some minor additional regularity conditions. In particular, one can show that

$$\sup_{t\in I_T^*,\theta\in\Theta}|\widehat{m}_\Delta(t,\theta) - \widetilde{m}(t,\theta)| = O_P(n^{-\delta_{min}}\sqrt{n^{-(1-\eta_+)}\log n} + n^{-2\delta_{min}}), \tag{A.12}$$

which is better than the rate of $O_P(n^{-\delta_{min}})$ obtained from a crude bound that appears in Theorem 2.

A.3. Proof of Theorem 3 We only show part b) explicitly. Calculations for part a) are conceptually the same. By assumption, we have that $||\widehat{r}(s) - r_0(s)||_\infty = o_P(n^{-1/4})$. Thus, it only remains to be shown that

$$||\widehat{m}(t, \theta) - m_0(t, \theta)||_\infty = o_P(n^{-1/4}).$$

This is fulfilled if all three remaining terms in the error bound in Theorem 2 are of smaller order than $n^{-1/4}$. For the two terms corresponding to the rate of convergence of the oracle estimator $\widetilde{m}$, this is directly achieved for bandwidths larger than the stated lower bound, and such that $\eta_+ < 1/2$. Such a bandwidth exists under sufficient smoothness conditions. The restriction that $\kappa_1 > 1/4$ imposes a binding restriction on the complexity of the sets $\mathcal{T}_{n,j}$. It can be satisfied if $\eta_+ < 1/2 + 2(\delta - \eta)_{min} - (\delta\alpha + \chi)_{max}$.

A.4. Proof of Theorem 4. To prove this result, we first establish a linear stochastic expansion for the oracle estimator $\widetilde{m}$. Using arguments in Masry (1996), Kong, Linton, and Xia (2010) or Ichimura and Lee (2010), one can show that

$$\widetilde{m}(t, \theta) = \frac{1}{n} \sum_{i=1}^{n} \varphi_{ni}^{\widetilde{m}}(t, \theta) + O(n^{-(p+1)\eta_{min}}) + O_P(\log(n) n^{-(1-\eta_+)}),$$

uniformly over $t \in I_T^*$ and $\theta \in \Theta$, where

$$\varphi_{ni}^{\widetilde{m}}(t, \theta) = e_1^\top N_h(t)^{-1} w(T_i(\theta) - t) K_h(T_i(\theta) - t) \varepsilon_i(\theta).$$

with $w(t) = (1, t, ..., t^p)^\top$ and $N_h(t, \theta) = \mathbb{E}(w((T_i(\theta) - t)/h, \theta) w((T(\theta) - t)/h, \theta)^\top K_h(T(\theta) - t))$. Next, note that the conditions of the Theorem imply that that $O(n^{-(p+1)\eta_{min}}) = o(n^{-1/2})$ and $O_P(\log(n) n^{-(1-\eta_+)}) = o_P(n^{-1/2})$ and $O(n^{-2\delta_{min}}) = o_P(n^{-1/2})$. Applying Theorem 1, we therefore find that Q_0^ξ can be decomposed as follows:

$$Q_0^\xi[\widehat{\xi} - \xi_0] = A_1 + A_2 + A_3 + A_4 + o_P(n^{-1/2}),$$

where

$$\begin{aligned}
A_1 &= \int \lambda_m(z_m) \frac{1}{n} \sum_{i=1}^{n} \varphi_{ni}^{\widetilde{m}}(z_m, \theta_0) f_{Z_m}(z_m) dz_m, \\
A_2 &= \int \lambda_m(z_m) \varphi_n^A(z_m, \theta_0, \widehat{r}) f_{Z_m}(z_m) dz_m, \\
A_3 &= \int \lambda_m(z_m) \varphi_n^B(z_m, \theta_0, \widehat{r}) f_{Z_m}(z_m) dz_m \\
A_4 &= \int \lambda_r(z_r) \varphi_{ni}^{\widehat{r}}(z_r, \theta_0, \widehat{r}) f_{Z_r}(z_r) dz_r,
\end{aligned}$$

We deal with each of these four terms separately. First, applying standard arguments from kernel smoothing theory, we find that

$$\begin{aligned}
A_1 &= \frac{1}{n}\sum_{i=1}^n \varepsilon_i \int e_1^\top N_h(z_m)^{-1} w(T_i - z_m) K_h(T_i - z_m)\lambda_m(z_m) f_{Z_m}(z_m) dz_m \\
&= \frac{1}{n}\sum_{i=1}^n \varepsilon_i \int e_1^\top N_h(T_i - th)^{-1} w(t) K(t) \lambda_m(T_i - th) f_{Z_m}(T_i - th) dt \\
&= \frac{1}{n}\sum_{i=1}^n \varepsilon_i \lambda_m(T_i) f_{Z_m}(T_i)/f_T(T_i) + O(n^{-(p+1)\eta_{min}}) \\
&= \frac{1}{n}\sum_{i=1}^n \psi_1(Z_i) + o_P(n^{-1/2})
\end{aligned}$$

For the second term, first note that it follows from standard bias calculations for kernel-type estimators that

$$\begin{aligned}
&\int \lambda_m(z_m)\varphi_n^A(z_m,\theta_0,r) f_{Z_m}(z_m) dz_m \\
&= -\mathbb{E}\left(T_i^{(r)}(X)(r(X_{ri}) - r_0(X_{ri}))\lambda_m(T_i) m_0'(T_i) \frac{f_{Z_m}(T_i)}{f_T(T_i)} \right) + O_P(h^{p+1})
\end{aligned}$$

uniformly for fixed functions $r \in \mathcal{R}_n^*$. Substituting the expansion for $\widehat{r} - r_0$ from Assumption 5 we then directly find that

$$\begin{aligned}
A_2 &= -\frac{1}{n}\sum_{i=1}^n \nu(W_i) \lim_{n\to\infty} \mathbb{E}\left(T^{(r)}(X)\lambda_m(T) m_0'(T) \frac{f_{Z_m}(T)}{f_T(T)} \mathcal{H}_n(S_i, X_r) \middle| S_i \right) \\
&\quad + O_P(n^{-(p+1)\eta_{min}} + n^{-2\delta_{min}}) + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^n \psi_2^A(Z_i) + o_P(n^{-1/2}).
\end{aligned}$$

Concerning the term A_3, we have that

$$\begin{aligned}
A_3 &= \iint \frac{\lambda_m(z_m)}{f_T(z_m)} K_h'(T(x) - z_m)(\widehat{T}(x) - T(x))\rho(x) f_{Z_m}(z_m) f_X(x)\,\mathrm{d}x dz_m \\
&= \int \frac{1}{h}\int K'(t) G(T(x) + th)\,\mathrm{d}t (\widehat{T}(x) - T(x))\rho(x) f_X(x)\,\mathrm{d}x \\
&= \int G'(T(x))(\widehat{T}(x) - T(x))\rho(x) f_X(x)\,\mathrm{d}x + O(h^{p+1}) \\
&= \int G'(T(x)) T^{(r)}(x) \left(\frac{1}{n}\sum_{i=1}^n \mathcal{H}_n(S_i, x_r)\nu(W_i) \right) \rho(x) f_X(x)\,\mathrm{d}x + O_P(h^{p+1} + n^{-2\delta_{min}}) \\
&= \frac{1}{n}\sum_{i=1}^n \nu(W_i) \lim_{n\to\infty} \mathbb{E}(G'(T) T^{(r)}(X) \mathcal{H}_n(S_i, X_r)\rho(X) | S_i) + O_P(n^{-(p+1)\eta_{min}} + n^{-2\delta_{min}}) \\
&= \frac{1}{n}\sum_{i=1}^n \psi_2^B(Z_i) + o_P(n^{-1/2})
\end{aligned}$$

with $G(t) = \lambda_m(t) f_{Z_m}(t) f_T(t)^{-1}$ and $G'(t) = \partial_t G(t)$ using integration by parts to obtain the fourth equality. Finally, we have

$$\begin{aligned} A_4 &= \nu(W_i) \lim_{n\to\infty} \mathbb{E}(\lambda_r(X_r)\mathcal{H}_n(S_i, X_r)|S_i) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \psi_3(Z_i) + o_P(n^{-1/2}) \end{aligned}$$

using the same type of arguments as the ones applied above. The statement of the Theorem then follows since $\psi_2 = \psi_2^A + \psi_2^B$.

A.5. Proof of Theorem 5. The statement of the theorem follows from the results on consistency and asymptotic normality of semiparametric two-step estimators in Chen, Linton, and Van Keilegom (2003).

A.6. Derivation of Example 1. Suppose that r_0 is a $q+1$-times continuously differentiable regression function estimated by qth order local polynomial regression using a bandwidth g and a kernel function L. Assume that S is continuously distributed with compact support I_S, and that the corresponding density f_S is q-times continuously differentiable, bounded, and bounded away from zero on I_S. Then it follows under some further standard regularity conditions (e.g. Kong, Linton, and Xia, 2010) that

$$\widehat{r}(s) - r_0(s) = \frac{1}{n}\sum_{i=1}^{n} e_1^\top N_g^S(s)^{-1} w(S_i - s) L_g(S_i - s)\zeta_i + O_P(g^{q+1} + \log(n)/(ng^{d_S}))$$

uniformly over $s \in I_S$, $w(t) = (1, t, ..., t^q)^\top$ as above and $N_g^S(s) = \mathbb{E}(w((S_i - s)/g)w((S_i - s)/g)^\top L_g(S_i - s))$. The remainder term in the last equation can be made as small as $o_P(n^{-1/2})$ by choosing an appropriate bandwidth if q is sufficiently large. It follows that Assumption 5 is satisfied with $\nu(W_i) = \zeta_i$ and $\mathcal{H}_n(S_i, s) = e_1^\top N_g^S(s)^{-1} w(S_i - s) L_g(S_i - s)$. The condition that $\mathbb{E}(\|\mathcal{H}_n(S_i, S_j)\|^2) = o(n)$ holds if $ng^{d_S} \to \infty$. To obtain the explicit expressions for ψ_2 and ψ_3, we insert the above relation into the expression from Theorem 4 and apply standard U-Statistics arguments (e.g. Powell, Stock, and Stoker, 1989). □

A.7. Derivation of Example 2. It easy to see that Assumption 5 is satisfied with $\nu(W_i) = \varphi^{\widehat{\vartheta}}(W_i)$ and $\mathcal{H}_n(S_i, s) = r'(s, \vartheta_0)$ under the conditions given in the example. By substituting these expression into the general formulas in Theorem 4, one directly obtains the specific expressions for ψ_2 and ψ_3 given in the main text. □

B. Details on Application to Treatment Effect Estimation

In this section, we give details on the construction of the estimator $\widehat{\theta}$, and the regularity conditions under which Proposition 1 is valid. The data consist of a sample $\{(Y_i, D_i, X_i), i = 1, \ldots, n\}$ from the distribution of (Y, D, X). The estimator of the propensity score $\Pi(x) = \mathbb{E}(D|X = x)$ is given by $\widehat{\Pi}(x) = \widehat{\alpha}$, where

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} (D_i - \alpha - \sum_{1 \leq u_+ \leq q} \beta_u^\top (X_i - x)^u)^2 L_g(X_i - x)$$

and $L_g(s) = \prod_{j=1}^{d_X} \mathcal{L}(s_j/g)/g$ is a d_X-dimensional product kernel built from the univariate kernel $\mathcal{L}$, g is a bandwidth, which for simplicity is assumed to be the same for all components, and $\sum_{1 \leq u_+ \leq q}$ denotes the summation over all $u = (u_1, \dots, u_q)$ with $1 \leq u_+ \leq q$. Next, for $d \in \{0, 1\}$ the estimate of $\nu_d(\pi) = \mathbb{E}(Y|D = d, \Pi(X) = \pi)$ is given by the third-order local polynomial estimator: we set $\widehat{\nu}_d(\pi) = \widehat{\alpha}_d$, where

$$(\widehat{\alpha}_d, \widehat{\beta}_d) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \mathbb{I}\{D_i = d\}(Y_i - \alpha - \sum_{1 \leq v \leq 3} \beta_v^\top (\widehat{\Pi}(X_i) - \pi)^v)^2 K_h(\widehat{\Pi}(X_i) - \pi) ,$$

with $K_h(u) = K(u/h)/h$, K a one-dimensional kernel function and h a bandwidth that tends to zero as the sample size n tends to infinity. The final estimator of θ_0 is then given by

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{\nu}_1(\widehat{\Pi}(X_i)) - \widehat{\nu}_0(\widehat{\Pi}(X_i))).$$

To prove Proposition 1, we make the following assumptions.

Assumption 6. *The sample observations $\{(Y_i, D_i, X_i), i = 1, \dots, n\}$ are i.i.d.*

Assumption 7. *(i) The random vector X is continuously distributed with compact support I_X. Its density function f_X is bounded and bounded away from zero on I_X, and is also $q+1$-times continuously differentiable for some uneven number $q \geq d_X$. (ii) The function $\Pi(x)$ is bounded away from zero and one on I_X, and is also $q + 1$-times continuously differentiable. (iii) For any $d \in \{0, 1\}$, the random variable $\Pi(X)$ is continuously distributed conditional on $D = d$, with compact support I_Π. Its conditional density function $f_{\Pi|D}(\cdot, d)$ is bounded and bounded away from zero on I_Π, and is also four times continuously differentiable. (iv) For any $d \in \{0, 1\}$, the function $\nu_d(\pi)$ is four times continuously differentiable on I_Π.*

Assumption 8. *The residual $\varepsilon = Y - \mathbb{E}(Y|\Pi(X))$ satisfies $E[\exp(l|\varepsilon|)|X] \leq C$ almost surely for a constant $C > 0$ and $l > 0$ small enough.*

Assumption 9. *(i) The function K is twice continuously differentiable and satisfies the following conditions: $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int |u^2 K(u)|du < \infty$, and $K(u) = 0$ for values of u not contained in some compact interval, say $[-1, 1]$. (ii) The function $\mathcal{L}$ is k-times continuously differentiable for some natural number $k \geq \max\{2, d_{X/2}\}$, and satisfies the following conditions: $\int \mathcal{L}(u)du = 1$, $\int u\mathcal{L}(u)du = 1$, and $\mathcal{L}(u) = 0$ for values of u not contained in some compact interval, say $[-1, 1]$.*

Assumption 10. *The bandwidths satisfy $h \sim n^{-\eta}$ and $g \sim n^{-\gamma}$ with $\gamma = 1/(2q + 1)$ and $1/8 < \eta < (q + 2)/(8q + 4)$.*

Proof of Proposition 1. The proof uses the same arguments as that of Theorem 4 and Example 1, and thus the details are omitted. The only issue is to show that we can choose $\kappa^* > 1/2$. To see this, note that the conditions of the Proposition imply that Assumption 2 holds with $\delta = (q+1)/(4q+2) > 1/4$, and that Assumption 3 holds with $\alpha \leq q/(q+1)$ and $\chi = 0$. The restrictions on η then ensure that $\delta - \eta > (1/2)(\delta\alpha + \chi)$ and $(1-\eta)/2 - \eta > (1/2)(\delta\alpha + \chi)$. We then easily see that $\kappa^* > 1/2$ can be chosen. □

C. Additional Assumptions

In this section, we state Assumption C.1–C.3, which collect those conditions of Theorem 5 and Theorem 6 that can be verified irrespective of the question whether the function m_0 is estimated using generated covariates or not. The assumptions are all minor variations of those given in Chen, Linton, and Van Keilegom (2003), and could be replaced by similar conditions considered in other papers studying $\sqrt{n}$-consistency and asymptotic normality of semiparametric "plug-in" estimators, such as Newey (1994).

Throughout the section, we use the following notation. For some small $\delta > 0$, we define $\Theta_\delta = \{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta\}$ and $\Xi_\delta = \{\xi \in \Xi : \|\xi - \xi_0\|_\Xi \leq \delta\}$. For any $(\theta, \xi) \in \Theta \times \Xi$, we also denote the ordinary derivative of $Q(\theta,\xi)$ with respect to θ by $Q^\theta(\theta,\xi)$. For any $\theta \in \Theta$, we say that $Q(\theta,\xi)$ is pathwise differentiable at $\xi \in \Xi$ in the direction $\bar{\xi}$ if there exists a continuous linear functional $Q^\xi(\theta,\xi) : \Theta \times \Xi \to \mathbb{R}^l$ such that $Q^\xi(\theta,\xi)[\bar{\xi}] = \lim_{\tau \to 0}(Q(\theta, \xi + \tau\bar{\xi}) - Q(\theta,\xi))/\tau$. The functional $Q^\xi(\theta,\xi)$ is called the pathwise derivative of $Q(\theta,\xi)$.

Assumption C.1. *Suppose that:*

(C1) For all $\delta > 0$, there exists an $\epsilon > 0$ such that $\inf_{\|\theta-\theta_0\|>\delta} \|Q(\theta,\xi_0)\| \geq \epsilon$.

(C2) Uniformly over $\theta \in \Theta$, $Q(\theta,\xi)$ is continuous in ξ at $\xi = \xi_0$ with respect to the metric $\|\cdot\|_\Xi$.

(C3) It holds that

$$\sup_{\theta\in\Theta, \|\xi-\xi_0\|_\Xi \leq \delta_n} \frac{\|Q_n(\theta,\xi) - Q(\theta,\xi) - Q_n(\theta_0,\xi_0)\|}{1 + \sqrt{n}(\|Q_n(\theta,\xi)\| + \|Q(\theta,\xi)\|)} = o_P(1)$$

for all positive sequences $\delta_n = o(1)$.

Assumption C.2. *Suppose that:*

(N1) $\theta_0 \in int(\Theta)$ satisfies $Q(\theta_0,\xi_0) = 0$.

(N2) (i) the ordinary derivative $Q^\theta(\theta,\xi_0)$ of $Q(\theta,\xi_0)$ in θ exists for $\theta \in \Theta_\delta$ and is continuous at $\theta = \theta_0$; (ii) the matrix $Q^\theta_0 = Q^\theta(\theta_0,\xi_0)$ is of full rank.

(N3) For all $\theta \in \Theta_\delta$ the pathwise derivative $Q^\xi(\theta, \xi_0)[\xi - \xi_0]$ of $Q(\theta, \xi_0)$ exists in all directions $(\xi - \xi_0) \in \Xi$; and for all $(\theta, \xi) \in \Theta_{\delta_n} \times \Xi_{\delta_n}$ with a positive sequence $\delta_n = o(1)$: (i) $\|Q(\theta, \xi) - Q(\theta, \xi_0) - Q^\xi(\theta, \xi_0)[\xi - \xi_0]\| \leq c\|\xi - \xi_0\|_\Xi^2$ for a constant $c \geq 0$; (ii) $\|Q^\xi(\theta, \xi_0)[\xi - \xi_0] - Q_0^\xi[\xi - \xi_0]\| \leq o(1)\delta_n$, where $Q_0^\xi[\xi - \xi_0] = Q^\xi(\theta_0, \xi_0)[\xi - \xi_0]$.

(N4) $\widehat{\xi} \in \Xi$ with probability tending to one.

(N5) It holds that

$$\sup_{\|\theta - \theta_0\| \leq \delta_n, \|\xi - \xi_0\|_\Xi \leq \delta_n} \frac{\sqrt{n}\|Q_n(\theta, \xi) - Q(\theta, \xi)\|}{1 + \|Q_n(\theta, \xi)\| + \|Q(\theta, \xi)\|} = o_P(1)$$

for any positive sequence $\delta_n = o(1)$.

Assumption C.3. *Suppose that:*

(B1) $\theta_0 \in int(\Theta)$ satisfies $Q(\theta_0, \xi_0) = 0$, and $\widehat{\theta} \xrightarrow{a.s.} \theta_0$.

(B2) $\|Q_n(\widehat{\theta}, \widehat{\xi})\| = \inf_{\theta \in \Theta} \|Q_n(\theta, \widehat{\xi})\| + o_{a.s.}(1/\sqrt{n})$

(B3) (i) $\widehat{\xi} \in \Xi$ almost surely, (ii) $\widehat{\xi}^* \in \Xi$ with P^* probability tending to one, and (iii) $\|\widehat{\xi} - \xi_0\|_\Xi = o_{a.s.}(n^{-1/4})$.

(B4) (i) For all $\xi \in \Xi_\delta$, the ordinary derivative $Q^\theta(\theta, \xi)$ of $Q(\theta, \xi)$ in θ exists for $\theta \in \Theta_\delta$ and is continuous at $\theta = \theta_0$; (ii) the matrix $Q_0^\theta = Q^\theta(\theta_0, \xi)$ is of full rank.

(B5) For all $\theta \in \Theta_\delta$ and $\xi \in \Xi_{\delta_n}$ with a positive sequence $\delta_n = o(1)$, the pathwise derivative $Q^\xi(\theta, \xi)[\bar{\xi} - \xi]$ of $Q(\theta, \xi)$ exists in all directions $(\bar{\xi} - \xi) \in \Xi$; and for all $(\theta, \bar{\xi}) \in \Theta_{\delta_n} \times \Xi_{\delta_n}$: (i) $\|Q(\theta, \bar{\xi}) - Q(\theta, \xi) - Q^\xi(\theta, \xi)[\bar{\xi} - \xi]\| \leq c\|\bar{\xi} - \xi\|_\Xi^2$ for a constant $c \geq 0$; (ii) $\|Q^\xi(\theta, \xi)[\bar{\xi} - \xi] - Q^\xi(\theta_0, \xi)[\bar{\xi} - \xi]\| \leq o(1)\delta_n$.

(B6) It holds that $\sup_{\|\theta - \theta_0\| \leq \delta_n, \|\xi - \xi_0\|_\Xi \leq \delta_n} \|Q_n(\theta, \xi) - Q(\theta, \xi) - Q_n(\theta_0, \xi_0)\| = o_{a.s.}(n^{-1/2})$ for any positive sequence $\delta_n = o(1)$.

(B7) It holds that $\sup_{\|\theta - \theta_0\| \leq \delta_n, \|\xi - \xi_0\|_\Xi \leq \delta_n} \|Q_n^*(\theta, \xi) - Q_n(\theta, \xi) - (Q_n^*(\theta_0, \xi_0) - Q_n(\theta_0, \xi_0))\| = o_{P^*}(n^{-1/2})$ for any positive sequence $\delta_n = o(1)$.

References

Ai, C., and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

——— (2007): "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables," *Journal of Econometrics*, 141(1), 5–43.

Andrews, D. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica*, 62(1), 43–72.

——— (1995): "Nonparametric kernel estimation for semiparametric models," *Econometric Theory*, 11(03), 560–586.

Blundell, R., and J. Powell (2004): "Endogeneity in semiparametric binary response models," *The Review of Economic Studies*, 71(3), 655–679.

Caetano, C., C. Rothe, and N. Yildiz (2014): "A Discontinuity Test for Identification in Triangular Nonseparable Models," *Working Paper*.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71(5), 1591–1608.

Chen, X., and D. Pouzo (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152(1), 46–60.

Chen, X., and X. Shen (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, 66(2), 289–314.

Einmahl, U., and D. Mason (2005): "Uniform in bandwidth consistency of kernel-type function estimators," *Annals of Statistics*, 33(3), 1380–1403.

Escanciano, J., D. Jacho-Chávez, and A. Lewbel (2012): "Identification and Estimation of Semiparametric Two Step Models," *Unpublished manuscript*.

——— (2014): "Uniform Convergence of Weighted Sums of Non- and Semi-parametric Residuals for Estimation and Testing," *Journal of Econometrics*, 178, 426–443.

Giné, E., and J. Zinn (1990): "Bootstrapping general empirical measures," *The Annals of Probability*, pp. 851–869.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

Hahn, J., and G. Ridder (2013): "Asymptotic Variance of Semiparametric Estimators With Generated Regressors," *Econometrica*, 81(1), 315–340.

Heckman, J., H. Ichimura, and P. Todd (1998): "Matching as an econometric evaluation estimator," *Review of Economic Studies*, 65(2), 261–294.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71(4), 1161–1189.

Ichimura, H., and S. Lee (2010): "Characterization of the asymptotic distribution of semiparametric M-estimators," *Journal of Econometrics*, 159(2), 252–266.

Imbens, G. (2004): "Nonparametric estimation of average treatment effects under exogeneity: A review," *Review of Economics and Statistics*, 86(1), 4–29.

Kong, E., O. Linton, and Y. Xia (2010): "Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model," *Econometric Theory*, 26(05), 1529–1564.

Levinsohn, J., and A. Petrin (2003): "Estimating production functions using inputs to control for unobservables," *Review of Economic Studies*, 70(2), 317–341.

Li, Q., and J. Wooldridge (2002): "Semiparametric estimation of partially linear models for dependent data with generated regressors," *Econometric Theory*, 18(03), 625–645.

Linton, O., S. Sperlich, and I. Van Keilegom (2008): "Estimation of a semiparametric transformation model," *Annals of Statistics*, 36(2), 686–718.

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric Regression with Nonparametrically Generated Covariates," *Annals of Statistics*, 40, 1132–1170.

Masry, E. (1996): "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17(6), 571–599.

Murphy, K. M., and R. H. Topel (1985): "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics*, 3, 370–379.

Newey, W. (1984): "A method of moments interpretation of sequential estimators," *Economics Letters*, 14(2-3), 201–206.

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

Newey, W. (1997): “Convergence rates and asymptotic normality for series estimators,” *Journal of Econometrics*, 79(1), 147–168.

Olley, G., and A. Pakes (1996): “The dynamics of productivity in the telecommunications equipment industry,” *Econometrica*, 64(6), 1263–1297.

Oxley, L., and M. McAleer (1993): “Econometric issues in macroeconomic models with generated regressors,” *Journal of Economic Surveys*, 7(1), 1–40.

Pagan, A. (1984): “Econometric issues in the analysis of regressions with generated regressors,” *International Economic Review*, 25(1), 221–247.

Powell, J., J. Stock, and T. Stoker (1989): “Semiparametric estimation of index coefficients,” *Econometrica*, 57(6), 1403–1430.

Rosenbaum, P., and D. Rubin (1983): “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70(1), 41–55.

Rothe, C. (2009): “Semiparametric estimation of binary response models with endogenous regressors,” *Journal of Econometrics*, 153(1), 51–64.

Song, K. (2008): “Uniform convergence of series estimators over function spaces,” *Econometric Theory*, 24(6), 1463–1499.

Song, K. (2012): “On the smoothness of conditional expectation functionals,” *Statistics & Probability Letters*, 82(5), 1028–1034.

Song, K. (2013): “Semiparametric models with single-index nuisance parameters,” *Working Paper*.

Sperlich, S. (2009): “A note on non-parametric estimation with predicted variables,” *Econometrics Journal*, 12(2), 382–395.

van de Geer, S. (2009): *Empirical Processes in M-Estimation*. Cambridge University Press.

Van der Vaart, A., and J. Wellner (1996): *Weak convergence and empirical processes: with applications to statistics*. Springer Verlag.