Semiparametric Estimation with Generated Covariates

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

University of Mannheim, Toulouse School of Economics, and Humboldt University Berlin

Abstract

In this paper, we study a general class of semiparametric optimization 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*

[*] This Version: May 13, 2012. Enno Mammen, Department of Economics, University of Mannheim, D-68131 Mannheim, Germany. E-mail: emammen@rumms.uni-mannheim.de. Christoph Rothe, Toulouse School of Economics, 21 Allée de Brienne, F-31000 Toulouse, France. E-mail: rothe@cict.fr. Melanie Schienle, School of Business and Economics, Humboldt University Berlin, Spandauer Str. 1, D-10178 Berlin, Germany. E-mail: melanie.schienle@wiwi.hu-berlin.de.

1. Introduction

In this paper, we study the theoretical properties of semiparametric estimators with generated covariates. Such estimators are used frequently to e.g. 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 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 obtain the generated covariates in the first place. 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. Deriving asymptotic properties of estimators in this class is a non-standard problem due to the presence of generated covariates. We give conditions on the primitives of the model 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. These results have important implications for econometric practice in a wide range of applications. In this paper, we apply our methods to two substantial examples: estimation of average treatment effects via regression on the propensity score (Rosenbaum and Rubin, 1983), and estimation of production functions in the presence of serially correlated technology shocks (Olley and Pakes, 1996; Levinsohn and Petrin, 2003). In both cases, our results contribute new insights to the respective extensive literature.

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. Here we argue that in order to account for such a structure it is not necessary to derive a completely new theory. Perhaps surprisingly, the "high-level" conditions given in the aforementioned papers are sufficiently general to encompass the generation step. What needs to be adapted substantially, however, are the methods used to verify these condition. Compared to a standard analysis, 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 contribution of our paper is to provide a connection between the extensive literature on estimation and inference in semiparametric models and the one on applications with generated covariates. We derive a new stochastic expansion that characterizes the influence of generated covariates in the model's nonparametric component on the asymptotic properties of the final estimator. We then show how to use this expansion 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 expansion, which is proven using techniques from empirical process theory (e.g. Van der Vaart and Wellner, 1996; van de Geer, 2009), is related to a result 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 sharp bounds on weighted integrals of the remainder term instead of controlling its supremum norm. This requires substantially different mathematical methods. The new 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 further contribution, we provide an explicit formula for the asymptotic variance of semiparametric estimators contained in the general class we consider. This formula is essentially a byproduct of the verification of the asymptotic normality condition mentioned above. 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) and Escanciano, Jacho-Chávez, and Lewbel (2010), among many others. Hahn and Ridder (2011) study the form of the influence function of semiparametric linear, just-identified GMM-type estimators with generated covariates using Newey's (1994) path-derivative method, and point out some mistakes in asymptotic variance calculations in the earlier literature. In contrast, the focus of this paper is on giving explicit conditions that ensure the estimators' root-n consistency and asymptotic normality, 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 (2011) provide stochastic expansions for sample means of weighted residuals of semiparametric regressions with generated covariates. 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 particular "index" condition, which is undesirable in many applications such as e.g. the estimation of average treatment effects that we study below, as it can imply strong restrictions on the underlying economic model and affect the form of the asymptotic variance. Our results, which were obtained independently, do not require such restrictions the underlying model, and apply to a substantially more general class of estimation procedures.

The remainder of the paper is structured as follows: In Section 2, we describe the class of models we consider. In Section 3, we present our main technical result, a stochastic expansion that characterizes the influence of generated covariates in the model's nonparametric component. Section 4 shows how this expansion can be used to verify classic conditions for $\sqrt{n}$-consistency and asymptotic normality of semiparametric estimators, derives a general formula for the asymptotic variance, and shows how to establish validity of the bootstrap. In Section 5, we discuss two econometric applications that make use of our results. All proofs and further details on the applications are collected in Appendix A and B, respectively.

2. Generated Covariates in Semiparametric Models

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. No specific estimation procedure is required for the latter object. Our results cover both parametrically and nonparametrically generated covariates, as well as intermediate cases. The setting and notation is otherwise similar to Chen, Linton, and Van Keilegom (2003), and thus allows for nonsmooth criterion functions and profiled estimation of the nonparametric components.

2.1. Model and Estimation Procedure. Let $Z = (Y, X, W) \in \mathbb{R}^{d_Z}$ be a random variable distributed according to some probability measure P_0 that is contained in a semiparametric model $\mathcal{P} = \{P_{\theta,\xi} : \theta \in \Theta, \xi \in \Xi\}$, where $\Theta \subset \mathbb{R}^{d_\theta}$ denotes a finite dimensional parameter space with generic element θ, and $\Xi = \mathcal{M} \times \mathcal{R}$ is 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, which implies that $P_0 = P_{\theta_0, \xi_0(\cdot, \theta_0)}$. We assume that there exists a nonrandom function $q : \text{supp}(Z) \times \Theta \times \Xi \to \mathbb{R}^{d_q}$ such that $Q(\theta, \xi_0(\cdot, \theta)) = \mathbb{E}(q(Z, \theta, \xi_0(\cdot, \theta))) = 0$ 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 value 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 is identified from the distribution of $W \subset Z$, and that this distribution does not depend on the true value of the other parameters in the model. This allows for a consistent estimate of r_0 to be computed without knowledge of θ_0 and m_0. 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 $m_0(\cdot, \theta) = \mathbb{E}(Y|T(X, \theta, r_0) = \cdot)$ where $T(X, \theta, r) = t(X, r(X_r), \theta)$ is a random vector of dimension d_T, $X_r \subset X$ are the covariates that enter 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 role of r_0 is thus to generate (some of) the covariates used to compute the function m_0. By allowing 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.

To make the notation more compact, we usually suppress the arguments of the infinite dimensional 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 assume that Ξ is a class of continuous and bounded functions endowed with the pseudo-norm $\|\cdot\|_\Xi$ induced by the sup-norm, i.e. we have $\|\xi\|_\Xi = \sup_\theta \sup_x |m(x, \theta)| + \sup_{x_r} |r(x_r)|$. We also write $\|B\| = (\mathrm{tr}(B'AB))^{1/2}$ for any matrix B, where we suppress the dependence of the norm on the fixed symmetric positive definite matrix A for notational convenience.

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. In the second step, for every $\theta \in \Theta$ we obtain an estimate $\widehat{m}(\cdot, \theta)$ of $m_0(\cdot, \theta)$ through a nonparametric regression of Y on the generated covariates $\widehat{T}(\theta) = T(\theta, \widehat{r})$. We discuss how to implement these two estimation procedures in detail below. 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:

$$\|Q_n(\widehat{\theta},\widehat{\xi})\| = \inf_{\theta\in\Theta} \|Q_n(\theta,\widehat{\xi})\| + o_P(1/\sqrt{n}), \tag{2.1}$$

where $Q_n(\theta,\widehat{\xi}) = \frac{1}{n}\sum_{i=1}^n q(Z_i,\theta,\widehat{\xi})$. Here, we avoid evaluating $\widehat{\xi}$ in areas where it is imprecisely estimated by restricting the influence of the nuisance parameter to be exceeded through its value over some compact set $I_T^* \times I_R^*$ introduced above. Such "fixed trimming" procedures are commonly used to derive properties of profiled semiparametric estimators. Allowing for "vanishing" trimming schemes where $I_T^* \times I_R^*$ increases with the sample size would be possible at the cost of tedious calculations that are unrelated to the issues caused by the presence of generated covariates. For the sake of clarity and brevity in exposition, these are therefore omitted here.

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 $m_0(\cdot,\theta)$. Note that because $\widehat{r}$ is also allowed to appear as a separate argument in the objective function Q_n, it does not only determine the shape of the function $\widehat{m}$, but could also exert a direct influence. This flexibility is useful for all examples we consider below. For instance, the objective function could depend on $\widehat{m}$ through its value at (some transformation of)end on $\widehat{m}$ through its value at (some transformation of) the generated covariates. However, it is important to stress that this is not required in our setting. Suppose for example that $\widehat{m}$ does not depend on θ, and that $\widehat{\theta} = n^{-1}\sum_{i=1}^n \widehat{m}(Z_i)$. Such an estimator, where $\widehat{r}$ is only used to compute $\widehat{m}$, can easily be analyzed in our framework.[1]

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 $\theta \in \Theta$, and then finding an approximate minimizer

[1] We remark that this is a simple example of an estimator that could not be analyzed using the results in Escanciano, Jacho-Chávez, and Lewbel (2011). They derive stochastic expansions for terms of the form $n^{-1}\sum_{i=1}^n (Y_i - \widehat{m}(\widehat{T}_i))s(X_i)$, where $s(X_i)$ is some weighting term (their results are somewhat more general, as they allow for estimated weights, the presence of vanishing trimming terms, and data-dependent choices of the bandwidth). Such terms typically appear in expansions of $\widehat{\theta}$ only if this estimator depends on $\widehat{m}$ through its values at $\widehat{T}_i$ only, which is not the case in this example.

of an infeasible version of the objective function:

$$\|Q_n(\widetilde{\theta}, \widehat{\xi})\| = \inf_{\theta \in \Theta} \|Q_n(\theta, \widetilde{\xi})\| + o_P(1/\sqrt{n}) \tag{2.2}$$

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.

2.2. A Framework for Asymptotic Analysis. It is straightforward to show that $\widehat{\theta}$ is a consistent estimate of the true value θ_0 under standard conditions. We therefore focus on the more interesting problem of establishing its asymptotic distribution. A number of papers have given "high level" conditions for semiparametric estimators to be root-n consistent and asymptotically normal in models that *do not* involve generated covariates. 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). It turns out that these conditions are generally sufficient to establish the same type of asymptotic properties for semiparametric estimators in models *with* generated covariates. What needs to be adjusted, however, are the arguments to verify some of them.

To illustrate how previous results in the literature on semiparametric estimation can be adapted to our context, we consider the main theorem from Chen, Linton, and Van Keilegom (2003). We choose this setting because it allows for a wide range of semiparametric estimators, including those that are based on a nonsmooth criterion function, or require profiled estimation of the nonparametric components. However, the arguments we are going to present are by no means specific to this setup, and apply analogously to all similar theoretical analyses of semiparametric estimators based on linearization arguments.

Before we state the main result from Chen, Linton, and Van Keilegom (2003), we have to introduce some further notation. Since we assume that $\widehat{\theta}$ is consistent, we can work with small subsets of the parameter spaces. For some small $\delta > 0$, define $\Theta_\delta = \{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta\}$ and $\Xi_\delta = \{\xi \in \Xi : \|\xi - \xi_0\|_\Xi \leq \delta\}$. Furthermore, for any $(\theta, \xi) \in \Theta \times \Xi$, we 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 $Q(\theta,\xi)$.

Theorem 1 (Chen, Linton, and Van Keilegom (2003))**.** *Suppose that* $\theta_0 \in int(\Theta)$ *satisfies* $Q(\theta_0,\xi_0)=0$, *that* $\widehat{\theta}=\theta_0+o_P(1)$, *and that:*

(N1) $\|Q_n(\widehat{\theta},\widehat{\xi})\| = \inf_{\theta\in\Theta}\|Q_n(\theta,\widehat{\xi})\| + o_P(1/\sqrt{n})$.

(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* $Q_0^{\theta}=Q^{\theta}(\theta_0,\xi_0)$ *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; and* $\|\widehat{\xi}-\xi_0\|_{\Xi}=o_P(n^{-1/4})$

(N5) For any positive sequence $\delta_n=o(1)$.

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

(N6) $\sqrt{n}(Q_n(\theta_0,\xi_0)+Q_0^{\xi}[\widehat{\xi}-\xi_0])\xrightarrow{d}N(0,V)$ *for some finite matrix* V.

Then $\sqrt{n}(\widehat{\theta}-\theta_0)\xrightarrow{d}N(0,\Omega)$*, where* $\Omega=(Q_0^{\theta\top}AQ_0^{\theta})^{-1}Q_0^{\theta\top}AVAQ_0^{\theta}(Q_0^{\theta\top}AQ_0^{\theta})^{-1}$.

Chen, Linton, and Van Keilegom (2003) provide an extensive discussion of the conditions of Theorem 1, arguing that they are fairly general and thus satisfied in a wide range of semiparametric models. Moreover, the result is sufficiently flexible to apply in our setting. Neither its conditions nor a single step in its proof do rule out the type of semiparametric estimation problems with generated covariates we consider in this paper. Asymptotic normality of the real estimator of $\widehat{\theta}$ can thus simply be established by checking (N1)–(N6). There is no need to develop a completely new theory.[2]

[2] To the best of our knowledge, this point has not been made explicitly in the literature on semiparametric estimation. However, it has at least implicitly been noted for a special case in Linton, Sperlich, and Van Keilegom (2008).

This does not imply that the presence of generated covariates does not affect the asymptotic properties of our estimator. Verification of the "uniform convergence" condition (N4) and the "asymptotic normality" condition (N6) are substantially more complicated, and the asymptotic variance V in (N6) will generally be different from the one we would have obtained if the true value r_0 had been used in the estimation procedure instead of the estimate $\widehat{r}$. In the following section, we therefore derive new and general methods to check conditions like (N4) and (N6), which also appear in many other papers.

On the other hand, note that the remaining conditions of Theorem 1 are not affected by the presence of generated covariates, and can thus be verified by standard arguments: (N1) simply states that $\widehat{\theta}$ is an approximate minimizer of the objective function, which we assumed in the first place; (N2) and (N3) are smoothness conditions on the population moment function, and (N5) is a stochastic equicontinuity condition. Neither involves estimates of the nonparametric components of our model, and thus they can be verified independently of the issue of generated covariates.

3. Controlling the Influence of Generated Covariates

This section contains our main technical result. In particular, we consider a stochastic expansion of nonparametrically estimated regression functions under very general conditions, deriving a sharp bound on weighted averages of the respective remainder terms. This is the key ingredient for showing condition (N6). Throughout this section, 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, and $a^b = (a_1^{b_1}, \ldots, a_d^{b_d})$ for any vector $b \in \mathbb{R}^d$.

3.1. Estimating the Nonparametric Component. To derive our main result, we need to be more specific about the estimation procedures for the infinite-dimensional nuisance parameters. We do not require a specific procedure for the estimator $\widehat{r}$ of r_0, but only impose certain "high-level" restrictions that cover a wide range of methods. Given an estimate of r_0, for every $\theta \in \Theta$ we then obtain an estimate of $m_0(\cdot, \theta)$ 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 $\widehat{m}(x,\theta) = \widehat{\alpha}$, where

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \sum_{1\le u_+ \le p} \beta_u^{\intercal} (\widehat{T}_i(\theta) - x)^u)^2 K_h(\widehat{T}_i(\theta) - x) \ , \tag{3.1}$$

where $K_h(v) = \prod_{j=1}^{d_T} \mathcal{K}(v_j/h_j)/h_j$ 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 $\sum_{1\le u_+\le p}$ denotes the summation over all $u = (u_1, \ldots, u_p)$ with $1 \le u_+ \le d_T$. For $p = 1$, we get the usual local linear estimator. We allow for uneven orders $p > 1$ for the purpose of bias control.[3]

To present our results later, it will also be useful to introduce the infeasible oracle estimate $\widetilde{m}(\cdot, \theta)$, which is obtained via local linear smoothing of Y versus $T(\theta)$ for every $\theta \in \Theta$, i.e. it is given by $\widetilde{m}(x, \theta) = \widetilde{\alpha}$, where

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

We focus on local polynomial 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 e.g. orthogonal series estimators, we conjecture that this would require substantially more involved technical arguments.

3.2. Assumptions. We now state our assumptions on the data generating process and the preliminary estimator $\widehat{r}$ of r_0. To this end, we define the generalized regression residual $\varepsilon(\theta) = Y - \mathbb{E}(Y|T(\theta))$, which allows us to write the dependent variable Y as $Y = m_0(T(\theta), \theta) + \varepsilon(\theta)$ with $\mathbb{E}(\varepsilon(\theta)|T(\theta)) = 0$.

[3] Note that the definition of the estimator $\widehat{m}(\cdot, \theta)$ in (3.1) implicitly requires $\widehat{T}(\theta)$ to be continuously distributed (see also Assumption 1(ii) below). This is not a restriction, however, as it would be straightforward to modify the estimator $\widehat{m}(\cdot, \theta)$ by the usual frequency method if some components of $\widehat{T}(\theta)$ 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 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 (3.1).

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

(i) The sample observations Z_i are independent and identically distributed.

(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 satisfying $I_T^* \subset int(I_T)$ with I_T^* compact. The corresponding density function $f_T(\cdot, \theta)$ is continuously differentiable for every $\theta \in \Theta$, and $\inf_{\theta\in\Theta, x\in I_T^*} f_T(x, \theta) > 0$.

(iii) For every $\theta \in \Theta$, the functions $m_0(\cdot, \theta)$ and $t(\cdot, \theta)$ are $(p + 1)$-times continuously differentiable on their respective domains.

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

(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 $\int |u^2\mathcal{K}(u)|du < \infty$, and $\mathcal{K}(u) = 0$ for values of u not contained in some compact interval, say $[-1, 1]$.

(vi) The bandwidth $h = (h_1, \dots, h_{d_T})$ satisfies $h_j \sim n^{-\eta_j}$ for all $j = 1, \dots, d_T$, 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. Part (ii) states that the covariates $T(\theta)$ are continuously distributed, and that the density is bounded away from zero over the set I_T^*, thus ensuring a stable estimate $\widehat{m}(\cdot, \theta)$ at the points of evaluation. 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 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 (2011), 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.[4]

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* $\delta_j^* > 1/4$ *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 $\widehat{r}$. Part (i) implies the classic condition that that $\sup |\widehat{r}_j(s) - r_{0,j}(s)| = o_P(n^{-1/4})$, which is necessary for condition (N4) of Theorem 1 to hold. This condition is required because we allow $\widehat{r}$ to appear as a separate argument in the objective function Q_n. It thus does potentially not only determine the shape of the function $\widehat{m}$, but could also exert a direct influence. 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 2 is also important from a practical point of view, as it gives some (admittedly rough) guidance for bandwidth choice in the presence of generated covariates.

Assumption 3 (Complexity)**.** *For every $j = 1, \ldots, d_T$, there exist a sequence of sets of functions $\mathcal{T}_{n,j}$ such that*

[4] 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.

(i) $\Pr(T_j(\cdot, \widehat{r}) \in \mathcal{T}_{n,j}) \to 1$ *as* $n \to \infty$.

(ii) For a constant $C_T > 0$ *and a function* r_n *with* $\|T_j(x, \theta, r_n) - T_j(x, \theta, r_0)\|_\infty = o_P(n^{-\delta_j})$, *the set* $\mathcal{T}^*_{n,j} = \mathcal{T}_{n,j} \cap \{T_j(\cdot, r) : \|T_j(x, \theta, r) - T_j(x, \theta, r_n)\|_\infty \leq n^{-\delta_j}\}$ *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, 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 $j = 1, \ldots, d_T\}$.

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 compact 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 $\chi^*_j \geq 0$, and that $|T_j(x, r(x_r), \theta) - T_j(x, r(x_r), \theta^*)| \leq C\|\theta - \theta^*\|$ for every θ, θ^* and every value of x and r. Then the set $\mathcal{T}_{n,j}$ satisfies Assumption 3(ii) with $\alpha_j = d_{X_r}/k$ and $\chi_j = \chi^*_j \alpha_j$ (Van der Vaart and Wellner, 1996, Corollary 2.7.2). 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 $\widehat{r}$. Thus it is not necessary that the bias term is a smooth function.

For kernel-based estimators of r_0, one can then verify Assumption 3(i) by explicitly calculating the derivatives. Consider e.g. the one-dimensional Nadaraya-Watson estimator $\widehat{r}_{n,j}$ with bandwidth of order $n^{-1/5}$. Choose $r_{n,j}$ equal to $r_{0,j}$ plus 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 $\chi^*_j > 0$. For sieve and orthogonal series estimators, Assumption 3(i) immediately holds when the set $\mathcal{M}_{n,j}$ is chosen as the sieve set or as a subset of the linear span of an increasing number of basis functions, respectively. For a discussion of entropy bounds and further references we refer to van de Geer (2009). 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). *We assume that the elements of $\mathcal{R}_n^* = \{r \in \mathcal{R} : T_j(\cdot, r) \in \mathcal{T}_{n,j}^*$ for all $j = 1, \dots, d_T\}$ satisfy the following properties:*

(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.

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

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

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

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

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

Assumption 4(i)–(ii) are technical conditions which ensure that the conditional expectation of the "index bias" $\rho(X, \theta)$ satisfies certain smoothness restrictions. These conditions trivially hold if $\rho(X, \theta) = 0$, as assumed in Escanciano, Jacho-Chávez, and Lewbel (2011). Generally speaking, the index bias could be equal to zero if the economic model implies certain exclusion restrictions on the relationship between the underlying covariates X and the variable Y. Such exclusion restrictions are e.g. be available in some 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 5.1

below for an application where this flexibility is important. Assumption 4(iii) is a further smoothness condition. If the random vector $r(X_r)$ is continuously distributed, this condition holds if $\|f_1 - f_2\|_\infty \leq C_B \|r_1 - r_2\|_\infty$ for all $r_1, r_2 \in \mathcal{R}_n^*$, where f_j denotes the density function of $r_j(X_r)$ for $j = 1, 2$. See Escanciano, Jacho-Chávez, and Lewbel (2011, Assumption 10) for a similar restriction on the densities of the generated covariates.

3.3. Stochastic Expansions of the Nonparametric Component. Using the assumptions outlined above, we can now derive a sharp stochastic approximation of the nonparametric estimator $\widehat{m}$. To state the result, we denote the unit vector $(1, 0, \ldots, 0)^\top$ in $\mathbb{R}^{p+1}$ by e_1, the derivative of $m_0(t, \theta)$ with respect to t by $m_0'(t, \theta)$, and write $w_i(t, \theta, r) = (1, (T_i(r, \theta) - t)/h, ..., (T_i(r, \theta) - t)^p / h^p)^\top$ and $N_h(x, \theta) = \mathbb{E}(w_i(t, \theta, r) w_i(t, \theta, r)^\top K_h(T_i(r, \theta) - t))$. Recalling that $\rho(X, \theta) = \mathbb{E}(Y|X) - \mathbb{E}(Y|T(\theta))$, we then define the approximating function $\widehat{m}_\Delta$ by

$$\widehat{m}_\Delta(t, \theta) = \widetilde{m}(t, \theta) + \varphi_n^A(t, \theta, \widehat{r}) + \varphi_n^B(t, \theta, \widehat{r}), \tag{3.2}$$

where

$$\varphi_n^A(t, \theta, r) = -m_0'(t, \theta) e_1^\top N_h(x, \theta)^{-1} \mathbb{E}(K_h(T_i(\theta) - t) w_i(x, \theta)(T_i(r, \theta) - T_i(\theta)))$$

in case of local linear regression with $p = 1$ (a general, notationally much more involved definition for higher order local polynomials is given in (A.2) in Appendix A), and

$$\varphi_n^B(t, \theta, r) = e_1^\top N_h(x, \theta)^{-1} \mathbb{E}(K_h'(T_i(\theta) - t)^\top w_i(x, \theta)(T_i(r, \theta) - T_i(\theta)) \rho(X, \theta))$$

for any $r \in \mathcal{R}_n^*$. Here we use the notation $K_h'(v) = (\mathcal{K}_{h,j}'(v) : j = 1, ..., d_T)^\top$ with elements $\mathcal{K}_{h,j}'(v) = \mathcal{K}'(v_j/h_j)/h_j^2 \prod_{j^* \neq j} \mathcal{K}(v_{j^*}/h_{j^*})/h_{j^*}$. Our main result concerns the accuracy when using $\widehat{m}_\Delta$ as an approximation of $\widehat{m}$.

Theorem 2. *Suppose that Assumption 1–4 hold. Then for any $\theta \in \Theta$, it is*

$$\int (\widehat{m}(t, \theta) - \widehat{m}_\Delta(t, \theta)) \omega(t) dt = o_P(n^{-\kappa^*}) \tag{3.3}$$

for some weight function $\omega : \mathbb{R}^d \to \mathbb{R}$ 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 sharp bound on weighted averages the the approximation error $\widehat{m}(t,\theta) - \widehat{m}_\Delta(t,\theta)$. We focus on this class of distance measures because they are particularly suitable to verify conditions of the type (N6) in Theorem 1. Bounds on the supremum norm of the approximation error, as studied 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 "uniform consistency" condition (N4), as explained below.

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.

4. Application to Semiparametric Estimation

In this section, we show how to verify conditions of the type (N4) and (N6) in Theorem 1. We also derive a general formula for the asymptotic variance of the estimator $\widehat{\theta}$. Throughout the section, we assume that the smoothness conditions (N2)–(N3) on the criterion function Q hold.

4.1. Verifying "Uniform Consistency". To verify the "Uniform Consistency" condition (N4), we use a variation of an earlier result in Mammen, Rothe, and Schienle (2012) to derive the uniform rate of consistency of the estimator $\widehat{m}(t,\theta)$.

Theorem 3 (Uniform Consistency)**.** *Suppose Assumption 1–3 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*

$$\kappa_1 < \frac{1}{2}(1-\eta_+) + (\delta-\eta)_{min} - \frac{1}{2}(\delta\alpha+\chi)_{max}, \; \kappa_2 < (p+1)\eta_{min} + (\delta-\eta)_{min},$$

$$\kappa_3 < \delta_{min} + (\delta-\eta)_{min}.$$

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. In order for condition (N4) to hold, these terms have to be of smaller order than $n^{-1/4}$. For the oracle part, this can easily be achieved by choosing an appropriate bandwidth under sufficient smoothness conditions. For the remaining terms, Assumption 2(i) and Assumption 1(iii) jointly imply that $\delta_{min} > 1/4$. It then follows from simple calculations that $O_P\left(n^{-\delta_{min}} + n^{-\kappa}\right) = o_P(n^{-1/4})$ under appropriate restrictions on the sets $\mathcal{T}_{n,j}$.[5]

4.2. Verifying "Asymptotic Normality". Given a specific estimator $\widehat{r}$ of r_0, the expansion $\widehat{m}_\Delta(t,\theta)$ in (3.2) can usually be calculated more explicitly, and can then be used to verify (N6). To illustrate this idea in a general setting, suppose that the estimator used to generate the covariates satisfies the following asymptotically linear representation, which can be shown to be satisfied for a wide range nonparametric, semiparametric, and fully parametric estimation procedures (we also discuss two representative examples below).[6]

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

$$\widehat{r}(s) - r_0(s) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{ni}^{\widehat{r}}(s) + R_n(s) \tag{4.1}$$

[5] Note that when studying the "asymptotic normality" condition (N6) in the next subsection, we will introduce some additional structure on the estimator $\widehat{r}$ of r_0 in Assumption 5. Using this additional structure, it would be possible to derive better rates than the one given in Theorem 3. See the remark at the end of the proof of Theorem 3 in Appendix A for details.

[6] Note that Assumption 5 is typically not satisfied for estimators that are not asymptotically Gaussian, such as e.g. the Maximum Score estimator of a single-index binary choice model, or other estimators that follow so-called cube-root asymptotics. See Song (2011) for a further discussion of this point.

with $\varphi_{ni}^{\widehat{r}}(s) = \mathcal{H}_n(S_i, s)\nu(W_i)$ *for some* $S_i \subset W_i$ *and* $\sup_{s \in I_R^*} |R_n(s)| = o_P(n^{-1/2})$*. 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

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

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. In most applications, m and r are square integrable functions of random vectors Z_m and Z_r, respectively, and it follows from the Riesz representation theorem that there exists unique square integrable functions λ_m and λ_r such that

$$Q^m(\theta_0, \xi_0)[\widehat{m} - m_0] = \int \lambda_m(z)(\widehat{m}(z) - m_0(z))dF_{Z_m}(z), \tag{4.2}$$

$$Q^r(\theta_0, \xi_0)[\widehat{r} - r_0] = \int \lambda_r(z)(\widehat{r}(z) - r_0(z))dF_{Z_r}(z). \tag{4.3}$$

See e.g. Newey (1994). The form of λ_m and λ_r depends on the particular application. For example, if the criterion function $Q(\theta, \xi) = \mathbb{E}(q(Z, \theta, m, r))$ is such that the term $q(Z, \theta, m, r)$ only depends on the functions m and r smoothly through their value when evaluated at some random vectors Z_m and Z_r, respectively, we have that

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

All econometric applications we consider in Section 5 below exhibit this structure.

When λ_m and λ_r are sufficiently smooth, one can use Assumption 5 together with the representation in (3.2) 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 $R_n(t) = \widehat{m}(t, \theta_0) - \widehat{m}_\Delta(t, \theta_0)$ established in Theorem 2 ensure, under suitable regularity conditions, that

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

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

$$\sqrt{n}(Q_n(\theta_0, \xi_0) + Q_0^\xi[\widehat{\xi} - \xi_0]) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Z_i) + o_P(1) \xrightarrow{d} N(0, \mathbb{E}(\psi(Z)\psi(Z)^\top)) \tag{4.4}$$

by the Central Limit Theorem, and thus condition (N6) holds with $V = \mathbb{E}(\psi(Z)\psi(Z)^\top)$. The following Corollary formalizes this argument, and provides a general formula to compute the variance matrix V.

Corollary 1 (Normality). *Suppose Assumption 1– 5 holds 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 (4.2)– (4.3) *with* $\lambda_m(\cdot)$ *and* $\lambda_r(\cdot)$ *being* $(p+1)$*-times continuously differentiable, and* $1/2(p+1) < \eta_j < 1/2d_T$ *for* $j = 1, \ldots, d_T$*. Then equation* (4.4) *holds with*

$$\psi_1(Z_i) = \varepsilon_i \lambda_m(T_i) f_{Z_m}(T_i) f_T(T_i)^{-1}$$
$$\psi_2(Z_i) = -\nu(W_i)\mathbb{E}(\lambda_m^*(X_r)\mathcal{H}_n(S_i, X_r)|S_i)$$
$$\psi_3(Z_i) = \nu(W_i)\mathbb{E}(\lambda_r(Z_r)\mathcal{H}_n(S_i, Z_r)|S_i),$$

where

$$\lambda_m^*(x_r) = \mathbb{E}(T^{(r)}(X)(\rho(X)G'(T) + m_0'(T)G(T))|X_r = x_r)$$

and $G(t) = \lambda_m(t) f_{Z_m}(t) f_T(t)^{-1}$ *and* $G'(t) = \partial_t G(t)$ *and* $T^{(r)}(x) = \partial T(x, \theta_0, r_0)/\partial r_0(x_r)$*.*

Restriction (4.5) involves a tradeoff between the complexity of the first and second estimation step for the nonparametric component: It 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.

Assumption 5 is similar to conditions used e.g. in Rothe (2009) or Ichimura and Lee (2010). We now give two examples for which it 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 with $\nu(W_i) = \zeta_i$ and $\mathcal{H}_n(S_i, s) = f_S(s)^{-1} L_g(S_i - s)$, where L is a kernel function and g is a bandwidth that tends to zero at an appropriate rate. 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 $\mathbb{E}(\varphi^{\widehat{\vartheta}}(W)) = 0$, $\mathbb{E}(\varphi^{\widehat{\vartheta}}(W)\varphi^{\widehat{\vartheta}}(W)^\top) < \infty$, and that $r(x_r, \mu)$ is continuously differentiable in its second argument. Then Assumption 5 is satisfied with $\nu(W_i) = \varphi^{\widehat{\vartheta}}(W_i)$ and $\mathcal{H}_n(S_i, s) = \partial_\vartheta r(s, \vartheta_0)$, and thus

$$\psi_2(Z_i) = -\nu(W_i)\mathbb{E}(T^r(X)\partial_\vartheta r(X_r, \vartheta_0)(\rho(X)g(T) + \lambda_m(T)m_0'(T)f_{Z_m}(T)f_T(T)^{-1}))$$
$$\psi_3(Z_i) = \nu(W_i)\mathbb{E}(\lambda_r(Z_r)\partial_\vartheta r(Z_r, \vartheta_0)).$$

In case that W is partitioned as $W = (D, S)$, and we have that $D = \bar{r}(S, \vartheta_0) + \zeta$ with $\mathbb{E}(\zeta|S) = 0$, and that $\widehat{\vartheta}$ is the nonlinear least squares estimator of ϑ_0. In such a setting, we would have that $\nu(W_i) = \mathbb{E}(\partial_\vartheta r(S, \vartheta_0)\partial_\vartheta r(S, \vartheta_0)^\top)^{-1}\partial_\vartheta r(S_i, \vartheta_0)(D_i - r_0(S_i))$, under the usual regularity conditions. □

4.3. 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 $\psi(z) = \sum_{j=0}^3 \psi_j(z)$. 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. This finding complements and generalizes results in Hahn and Ridder (2011), who were the first to derive the influence function for a class of semiparametric estimators with generated covariates.

The following two examples give an explicit formula for the asymptotic variance of 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 (2011) 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 (2011, 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 Corollary 1 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 $T = (X_1, r_0(X_r))$,

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

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 may semiparametric regression problems, such as e.g. 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 $\mathbb{E}(\zeta|X_r) = 0$. Applying Corollary 1 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 $u(t) = \mathbb{E}(s(X)|T = t)$,

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

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.

4.4. Validity of the Bootstrap. In some applications, the asymptotic variance matrix V could be difficult to estimate since it depends 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 useful to compute confidence regions for the parameters of interest. Our results 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 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 which imply that $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ converges in distribution to $N(0, \Omega)$ under the probability measure P^* implied by the bootstrap are given in Theorem B in Chen, Linton, and Van Keilegom (2003). In the presence of generated covariates, the central requirements to be checked are the following variants of (N4) and (N6), respectively:

(B4) $\widehat{\xi} \in \Xi$ with P^*-probability tending to one; and $\|\widehat{\xi}^* - \widehat{\xi}\|_\Xi = o_{P^*}(n^{-1/4})$.

(B6) $\sqrt{n}(Q_n^*(\theta_0, \xi_0) + Q_0^\xi[\widehat{\xi}^* - \xi_0]) \xrightarrow{d} N(0, V)$ under P^*.

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 (N4) and (N6) above, and are thus immediate for a wide range of applications. We thus obtain the following Corollary.

Corollary 2. *Under the conditions of Corollary 1, (B4) and (B6) are fulfilled.*

The remaining conditions for the validity of the bootstrap given by Chen, Linton, and Van Keilegom (2003) are mostly minor strengthenings of those in Theorem 1, that can be verified irrespective of the presence of generated covariates.

5. Econometric Applications

Semiparametric estimation problems with generated covariates occur in various fields of econometrics. In this subsection, we discuss two applications in greater detail: estimation of average treatment effects via regression on the propensity score, and estimation of production functions in the presence of serially correlated technology shocks. To save space, we only sketch the construction of estimators, and refer to Appendix B for details and regularity conditions.

5.1. Regression on the Propensity Score. Consider the potential outcomes framework, which is commonly used in the 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 depends on observable covariates X only, 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 $Y_1, Y_0 \perp D|X$ implies that $Y_1, Y_0 \perp D|\Pi(X)$. 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

$$\theta_0 = \mathbb{E}(\nu_1(\Pi(X)) - \nu_0(\Pi(X))). \tag{5.1}$$

Similar arguments can be made for other measures of program effectiveness (e.g. Heckman, Ichimura, and Todd, 1998). Estimating the ATE by a sample analogue of (5.1) 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 $q(z, \theta, m_0, r_0) \equiv \nu_1(\Pi(x)) - \nu_0(\Pi(x)) - \theta$.

Using the path-derivative approach of Newey (1994), Hahn and Ridder (2011) were the first to derive the form of the influence function for this estimation problem. 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 available. In particular, we consider the following sample version of (5.1) 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 e.g. 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.1 hold. Then we have that* $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathbb{E}(\Psi(Y, D, X)^2)$*, 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 corresponding semiparametric efficiency bound 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 e.g. 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.

5.2. Estimation of Production Functions. When estimating the parameters of production functions, a simultaneity problem arises if there is contemporaneous correlation between a firm's inputs and shocks to productivity. In a highly influential paper, Olley and Pakes (1996) propose a methodology to address this issue, which can be seen as a control function approach. Here we consider a simplified version of their method, as described in Levinsohn and Petrin (2003). This setting assumes that firms do not age and cannot be closed. The Cobb-Douglas model for log output Y_t of a firm in period t is given by

$$Y_t = \beta_0 + \beta_L L_t + \beta_K K_t + \omega_t + \eta_t, \tag{5.2}$$

where L_t and K_t are labor and capital inputs, respectively, ω_t is a productivity index that follows a first-order Markov process, and η_t is an i.i.d. productivity shock. Here ω_t and η_t are both unobserved. The main difference is that ω_t is a state variable, and hence impacts the firm's input choices, while η_t has no impact on firm behavior. In particular, the firms' investment I_t in the capital stock is a function of ω_t and K_t: $I_t = \iota_t(\omega_t, K_t)$. Under suitable conditions, firms that choose to invest have investment functions that are strictly increasing in the unobserved productivity index, and hence by invertability ω_t can be written as function of capital and investment

$$\omega_t = \omega(K_t, I_t).$$

Substituting this relationship into (5.2), we find that

$$Y_t = \beta_L L_t + \phi_t + \eta_t, \tag{5.3}$$

where $\phi_t = \phi(K_t, I_t) = \beta_K K_t + \omega(K_t, I_t)$. Equation (5.3) is a standard partially linear model, and thus β_L and the function $\phi(\cdot)$ can be identified and estimated as in Robinson (1988) through the usual least squares arguments. To identify the coefficient β_K, it is assumed that capital does not immediately respond to innovations in the productivity index ω_t, which together with the Markov assumption implies that

$$\omega_t = \Pi(\omega_{t-1}) + \xi_t \text{ with } \mathbb{E}(\xi_t | \omega_{t-1}, K_t) = 0.$$

We can thus rewrite the output net of labor's contribution $Y_t^* = Y_t - \beta_L L_t$ as

$$Y_t^* = \beta_K K_t + \Pi^*(\phi_{t-1} - \beta_K K_{t-1}) + \eta_t^*, \tag{5.4}$$

with $\Pi^*(x) = \Pi(x) + \beta_0$ and $\eta_t^* = \eta_t + \xi_t$. Note that while equation (5.4) resembles a partially linear model (given knowledge of β_L and $\phi(\cdot)$), its structure is actually somewhat different, as the coefficient β_K appears both in the linear part and inside the unknown function Π^*. Still, the parameter β_K can be characterized as the solution to a profiled nonlinear least squares problem:

$$\beta_K = \operatorname*{argmin}_b \mathbb{E}(Y_t - \beta_L L_t - bK_t - \pi(\phi_{t-1} - bK_{t-1}|b))^2, \tag{5.5}$$

where $\pi(c|b) = \mathbb{E}(Y_t - \beta_L L_t - bK_t | \phi_{t-1} - bK_{t-1} = c)$ for any $b \in \mathbb{R}$. Implementing a sample analogue of (5.5) to estimate β_K requires nonparametric estimation of the function $\pi(\cdot|b)$ using an estimates of the coefficient β_L and the function $\phi(\cdot)$, both obtained by estimating (5.3) in a first stage. This problem fits into our framework with $Z \equiv (Y_t, L_t, K_t, I_t, K_{t-1}, I_{t-1})$, $\theta_0 \equiv \beta_K$, $r_0(X_r) \equiv (\beta_L, \phi_{t-1})$, $T(X, \theta, r_0) \equiv \phi_{t-1} - bK_{t-1}$, $m_0(\cdot, \theta) \equiv \pi(\cdot|b)$ and $q(Z, \theta, m_0, r_0) \equiv (Y_t - \beta_L L_t - bK_t - \pi(\phi_{t-1} - bK_{t-1}|b))(K_t - \partial_b \pi(\phi_{t-1} - bK_{t-1}|b)K_{t-1})$.

To give an explicit expression for an estimator $\widehat{\beta}_K$ of β_K, let $\widehat{\beta}_L$ and $\widehat{\phi}(\cdot)$ be estimates of β_L and $\phi(\cdot)$, respectively, obtained via the method in Robinson (1988). For every $b \in \mathbb{R}$, let $\widehat{\pi}(\cdot|b)$ be an estimate of $\widehat{\pi}(\cdot|b)$, computed by local linear regression of $Y_{it} - \widehat{\beta}_L L_{it} - bK_{it}$ on $\widehat{\phi}_{i,t-1} - bK_{i,t-1}$. Then we can define the final estimator as

$$\widehat{\beta}_K = \operatorname*{argmin}_b \frac{1}{n} \sum_{i=1}^n (Y_{it} - \widehat{\beta}_L L_{it} - bK_{it} - \widehat{\pi}(\widehat{\phi}_{i,t-1} - bK_{it-1}|b))^2. \tag{5.6}$$

Note that computing $\widehat{\pi}(\cdot|b)$ and $\phi(\cdot)$ involves the use of a generated *dependent* variable. However, compared to the problems arising from the presence of generated covariates, this issue is straightforward to address for linear smoothers like local linear regression. To simplify the expression for the influence function, we introduce the following notation:

Let $\pi^{(b)}(c|b) = \partial_a \pi(a|b)|_{a=c} + \partial_a \pi(c|a)|_{a=b}$ be the total derivative of $\pi(b|b)$ with respect to b, and $\pi'(c|b) = \partial_c \pi(c|b)$ the ordinary derivative with respect to the first component. We also define $G_{it} = K_{it} - \pi^{(b)}(\phi_{i,t-1} - \beta_K K_{i,t-1}|\beta_K)K_{i,t-1}$ and the "projection residuals" $G_t^{\perp} = G_t - \mathbb{E}(G_t|\phi_{t-1} - \beta_K K_{t-1})$ and $L_t^{\perp} = L_t - \mathbb{E}(L_t|\phi_{t-1} - \beta_K K_{t-1})$.

Proposition 2. *Suppose that the regularity conditions given in Appendix B.2 hold. Then we have that* $\sqrt{n}(\widehat{\beta}_K - \beta_K) \xrightarrow{d} N(0, \Omega)$ *with*

$$\Omega = Q_0^{\theta-1}\mathbb{E}\left[(\Psi_0 + \Psi_1 + \Psi_2)(\Psi_0 + \Psi_1 + \Psi_2)^\top\right] Q_0^{\theta-1},$$

where

$$\begin{aligned}
\Psi_0 &= G_t^{\perp}\eta_t^* \\
\Psi_1 &= -\mathbb{E}(G_t^{\perp}|K_{t-1}, I_{t-1})\pi'(\phi_{t-1} - \beta_K K_{t-1}|\beta_K)\eta_{t-1} \\
\Psi_2 &= -\mathbb{E}(G_t(L_t^{\perp} - \mathbb{E}(L_t|K_{t-1}, I_{t-1})\pi'(\phi_{t-1} - \beta_K K_{t-1}|\beta_K))) \\
&\qquad \times \mathbb{E}((L_t - \mathbb{E}(L_t|K_t, I_t))^2)^{-1}(L_t - \mathbb{E}(L_t|K_t, I_t))\eta_t.
\end{aligned}$$

Asymptotic properties of a somewhat more general version the above estimation procedure were first studied in Pakes and Olley (1995). Our expression for the influence function given in Proposition 2 differs from their result, even when taking into account that we only consider a simplified version of their model. The reason is that our derivation does account for the estimation error from using an estimate of $\phi(\cdot)$ when *estimating* $\widehat{\pi}(\cdot|b)$, and not only for the estimation error resulting from using an estimate of $\phi(\cdot)$ when *evaluating* $\widehat{\pi}(\cdot|b)$. In our Proposition 2, both contributions are collected in the term Ψ_1. The estimation problem was also mentioned in an early working paper version of Hahn and Ridder (2011), but to the best of our knowledge they did not derive an explicit expression for the influence function, or make any comparison with Pakes and Olley (1995).

6. Concluding Remarks

In this paper, we have derived a general asymptotic theory for a large class of semiparametric optimization estimators when the infinite-dimensional component is estimated using generated covariates. Using our Theorems 2–3, we have shown how general "high-level" conditions for root-n consistency, asymptotic normality, and the validity of the bootstrap, given in Chen, Linton, and Van Keilegom (2003) can be verified in such a context. However, it is important to stress that our arguments are not specific to the setting in Chen, Linton, and Van Keilegom (2003), but can easily be combined with the results from other papers on semiparametric estimation that allow for kernel-based estimators, such as e.g. Newey (1994), Andrews (1994), or Ichimura and Lee (2010). In certain simple settings for which the conditions in the just mentioned papers are unnecessarily general, it is straightforward to derive asymptotic properties by applying our Theorems 2–3 directly to an expansion of the estimator of interest, and mimicking the arguments in Section 4.1–4.2. Moreover, our results appear to be sufficiently flexible to allow extending our analysis to semiparametric estimators that are asymptotically normal but do not satisfy an asymptotic linearity condition, as studied e.g. by Cattaneo, Crump, and Jansson (2011). We leave the details of the last point for future research.

A. Proofs of Main Results

A.1. Proof of Theorem 2. To simplify notation, we provide the proof only for the special case $d_T = 1$, i.e. $T = T(X, \theta, r)$ is a univariate random variable, but calculated rates are stated in general form. The proof for higher-dimensional T is conceptually similar. The following notation is used throughout all our proofs. The unit vector $(1, 0, \ldots, 0)^\top$ in $\mathbb{R}^{p+1}$ is denoted by e_1. We write

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

and $N_h(t,\theta) = \mathbb{E}(M_h(t,\theta))$. Furthermore, we 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. 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:

$$\widehat{m}(t,\theta) = m_0(t,\theta) + \widehat{m}_A(t,\theta) + \widehat{m}_B(t,\theta) + \widehat{m}_C(t,\theta) + \widehat{m}_D(t,\theta) + \widehat{m}_E(t,\theta)$$
$$\widetilde{m}(t,\theta) = m_0(t,\theta) + \widetilde{m}_A(t,\theta) + \widetilde{m}_B(t,\theta) + \widetilde{m}_C(t,\theta) + \widetilde{m}_D(t,\theta) + \widetilde{m}_E(t,\theta),$$

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:

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

Finally, we denote the component-wise differences between the real and the oracle estimator by

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

The statement of the theorem follows if for any $\theta \in \Theta$ the remainder term $R_n(t,\theta) = \widehat{m}(t,\theta) - \widehat{m}_\Delta(t,\theta)$ satisfies

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

Here $\widehat{m}_\Delta(t,\theta) = \widetilde{m}(t,\theta) + \varphi_n^A(t,\theta,\widehat{r}) + \varphi_n^B(t,\theta,\widehat{r})$. The term $\varphi_n^B(t,\theta,r)$ is as defined in (3.2) , and for $p=1$ the term $\varphi_n^A(t,\theta,r)$ is also as defined in (3.2). More generally, for uneven $p>1$ we set

$$\varphi_n^A(t,\theta,r) = e_1^\top N_h(\theta)^{-1}\mathbb{E}(K_h(T_i(r)-t)w_i(t,\theta,r)m'_{pol}(T_i(r),t,\theta)(T_i(r,\theta)-T_i(\theta)), \tag{A.2}$$

where $m'_{pol}(u,t,\theta)$ is the derivative of $m_{pol}(u,t,\theta)$ with respect to its first argument and $m_{pol}(u,t,\theta)$ is the following polynomial approximation of $m_0(u,\theta)$ in a neighborhood of t:

$$m_{pol}(u,t,\theta) = m_0^*(t,\theta)^\top(1,(u-t)/h,...,(u-t)^p/(p!h^p))^\top.$$

To simplify the notation, we fix $\theta = \theta_0$ for the rest of the proof and we omit θ as an argument of functions. To prove Theorem 2, we will then show that

$$\int R_{A,n}(t)\omega(t)\,\mathrm{d}t = O_P(n^{-\kappa_1^*}), \tag{A.3}$$

$$\int R_{B,n}(t)\omega(t)\,\mathrm{d}t = O_P(n^{-\kappa_2^*}), \tag{A.4}$$

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

$$\int R_{E,n}(t)\omega(t)\,\mathrm{d}t = \int \varphi_n^B(t,\widehat{r})\omega(t)\,\mathrm{d}t + O_P(n^{-\kappa_1^*} + n^{-\kappa_2^*}). \tag{A.6}$$

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

$$\int (\widehat{m}(t) - \widetilde{m}(t))\omega(t)\,\mathrm{d}t = \sum_{j\in\{A,\ldots,E\}} \int R_{n,j}(t)\omega(t)\,\mathrm{d}t, \tag{A.7}$$

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

We start with the proof of (A.3). Denote $\Phi_i(t,r) = e_1^\top M_h(t,r)^{-1} w_i(t,r) K_h(T_i(r) - t)$ and write $\Phi_i(r) = \int \Phi_i(t,r)\omega(t)\,\mathrm{d}t$. 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

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

Using elementary arguments, one can show that

$$M_h(T_i(r_1), r_1) - M_h(T_i(r_2), r_2) = O_P(n^{\eta_{max}})\|r_1 - r_2\|_\infty.$$

uniformly for $r_1, r_2 \in \mathcal{R}_n^*$ and $1 \le i \le n$. With the help of this bound, we find that, uniformly for $r_1, r_2 \in \mathcal{R}_n^*$ and $1 \le i \le n$ and some generic constant $c > 0$ which can take different values at each appearance

$$\begin{aligned}
&|\Phi_i(r_1) - \Phi_i(r_2)| \\
&\le \left| \int \left[e_1^\top M_h(t,r_1)^{-1} L_h(T_i(r_1) - t) - e_1^\top M_h(t,r_2)^{-1} L_h(T_i(r_2) - t) \right] \omega(t) dt \right| \\
&= \left| \int \left[e_1^\top M_h(T_i(r_1) - hu, r_1)^{-1} \omega(T_i(r_1) - hu) \right.\right. \\
&\qquad \left.\left. -\ e_1^\top M_h(T_i(r_2) - hu, r_2)^{-1} \omega(T_i(r_2) - hu) \right] L(u) du \right| \\
&\le \max_{1\le j\le d_T} c n^{\eta_j} |T_j(r_1) - T_j(r_2)|.
\end{aligned} \tag{A.8}$$

This last bound can be used to calculate a rough bound on the entropy $H_n(\lambda)$ of the class of functions $i \to \Phi_i(r)$. Using Assumption 3, this class of functions can be covered by $c\exp((\lambda n^{-\eta_j})^{-\alpha} n^{\chi})$

balls of radius $\lambda n^{-\eta_j}$. Thus we find that the entropy $H_n(\lambda) \leq c \max_{1\leq j\leq d_t} \lambda^{-\alpha_j} n^{\eta_j\alpha_j+\chi_j}$ for some constant $c > 0$. This implies

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

for $C_n = n^{-\delta_{min}}$. We now apply Theorem 8.13 in van de Geer (2009) with $\bar{Z}_\theta = n^{-1}\sum_{i=1}^n Z_{i,\theta}$, $Z_{i,\theta} = \Phi_i(r)\epsilon_i^*$, $\theta = r$, $R = C_n = n^{-\delta_{min}}$, 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 $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\exp(\ell^*|\epsilon_i^*|)|X_i] \leq C^*$ with probability tending to one, for some constants $C^*, \ell^* > 0$ due to Assumption 1 (iv). With standard arguments this yields

$$\sup_{r_1,r_2\in\mathcal{R}_n^*} \frac{1}{n}\sum_{i=1}^n (\Phi_i(r_1) - \Phi_i(r_2))\epsilon_i^* = o_P\left(n^{-(1/2)-(1-\alpha_{max}/2)\delta_{min}+(\eta\alpha+\chi)_{max}/2}\right). \tag{A.9}$$

Equation (A.9) provides the desired result (A.3) for R_A.

For the proof of (A.4), note that for some nonnegative 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.4).

For the proof of (A.5), note that uniformly over $1 \leq i \leq n$ and $r \in \mathcal{R}_n^*$ it holds that

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

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.5) we have to show that

$$\frac{1}{n}\sum_{i=1}^n (\Phi_i^*(\widehat{r}) - \Phi_i^*(r_0))(T_i(\widehat{r}) - T_i(r_0)) = O_P(n^{-\kappa_3^*} + n^{-\kappa_4^*}). \tag{A.10}$$

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

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

that uniformly for $r \in \mathcal{R}_n^*$ and $1 \leq i \leq n$ in order to establish (A.10). 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

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

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

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

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

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

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n & \big[(T_i(r) - t)^u h^{-u} K_h(T_i(r) - t)\big] \\
& - \big[(T_i(r_0) - t)^u h^{-u} K_h(T_i(r_0) - t)\big] = O_P(n^{-\delta_{min}} + n^{-\kappa_4^* + \delta_{min}}). \qquad \text{(A.11)}
\end{aligned}$$

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

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

uniformly over $r \in \mathcal{R}_n^*$. Thus, for a proof of (A.11) it suffices to establish that

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

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

$$\int R_{C,n}(t)\omega(t)\,\mathrm{d}t = \int \varphi_n^A(t,\widehat{r})\omega(t)\,\mathrm{d}t + O_P(n^{-\kappa_4^*} + n^{-\kappa_5^*}),$$

as claimed

It remains to show (A.6). 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 $R_{A,n}$ 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.6) holds. This concludes the proof of Theorem 2. □

A.2. Proof of Theorem 3. First, standard results in e.g. Masry (1996), imply that the oracle estimator $\widetilde{m}$ satisfies

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

uniformly over $t \in I_T^*$ and $\theta \in \Theta$ under the conditions of the theorem. Second, one can show that

$$\sup_{t\in I_T^*,\theta\in\Theta} |\widehat{m}(t,\theta) - \widehat{m}_\Delta(t,\theta)| = o_P(n^{-\kappa}). \tag{A.12}$$

The statement (A.12) 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.12), the statement of the Theorem follows from a trivial bound on the leading terms of the expansion $\widehat{m}_\Delta$. □

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

$$\sup_{t\in I_T^*,\theta\in\Theta} |\widehat{m}_\Delta(t,\theta) - \widetilde{m}(t,\theta)| = O_P(n^{-\delta_{min}}\sqrt{n^{-(1-\eta_+)}\log n} + n^{-2\delta_{min}}), \tag{A.13}$$

which is better than the rate of $O_P(n^{-\delta_{min}})$ obtained from a crude bound that appears in Theorem 3.

A.3. Proof of Corollary 1. 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

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

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 corollary imply that that $O(n^{-p\eta_{min}}) = o(n^{-1/2})$ 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 2, we therefore we 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

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

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

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

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

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

Concerning the term A_3, we have that

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

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

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

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

A.4. Derivation of Example 1. Suppose that r_0 is a q-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

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

uniformly over $s \in I_S$, $w(t) = (1, t, ..., t^p)^\top$ as above and $N_h^S(t) = \mathbb{E}(w((S_i - s)/g, \theta)w((S_i - s)/g, \theta)^\top L_g(S_i - s)$. 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 $\mathcal{H}_n(S_i, s) = e_1^\top N_h^S(s)^{-1}w(S_i - s)L_g(S_i - s)$. 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 Corollary 1 and apply standard U-Statistics arguments (e.g. Powell, Stock, and Stoker, 1989). □

A.5. Derivation of Example 2. This derivation is trivial and thus omitted. □

B. Details on Econometric Applications

B.1. Regression on the Propensity Score. 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 $\{(Y_i, D_i, X_i), i = 1, \dots, n\}$ 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 $L_g(s) = \prod_{j=1}^{p} \mathcal{L}(s_j/g)/g$ 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 $u = (u_1, \ldots, u_p)$ 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

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

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 $\{(Y_i, D_i, X_i), i = 1, \ldots, n\}$ 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 Corollary 1 and Example 1, and thus the details are omitted. The only issue is to show that $\kappa^* > 1/2$. 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 $\kappa^* > 1/2$. □

B.2. Estimation of Production Functions. In this section, we give details on the construction of the estimator $\widehat{\theta}$, and the regularity conditions under which Proposition 2 is valid. The data consist of a sample $\{(Y_{it}, L_{it}, K_{it}, I_{it}, K_{it-1}, I_{it-1}), i = 1, \ldots, n\}$ from the distribution of $(Y_t, L_t, K_t, I_t, K_{t-1}, I_{t-1})$. As a first step, we obtain an estimator $\widehat{\beta}_L$ of β_L using the method in Robinson (1988). Under regularity conditions given in that paper,

$$\sqrt{n}(\widehat{\beta}_L - \beta_L) = \mathbb{E}((L_t - \mathbb{E}(L_t|K_t, I_t))^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (L_{it} - \mathbb{E}(L_{it}|K_{it}, I_{it}))\eta_{it} + o_P(1).$$

Next, the estimator of $\phi(\cdot)$ is given by $\widehat{\phi}(a,b) = \widehat{\alpha}$, where

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} ((Y_{it} - \widehat{\beta}_L L_{it}) - \alpha - \sum_{1 \leq u_+ \leq q} \beta_u^T ((K_{it}, I_{it}) - (a,b))^u)^2 L_g((K_{it}, I_{it}) - (a,b)),$$

and $L_g(s) = \prod_{j=1}^{p} \mathcal{L}(s_j/g)/g$ 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 $u = (u_1, \ldots, u_p)$ with $1 \leq u_+ \leq q$. To simplify the exposition below, we also define an infeasible estimator of $\phi(\cdot)$ that uses the true value of the dependent variable. We set $\widehat{\phi}^*(a,b) = \widehat{\alpha}$, where

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} ((Y_{it} - \beta_L L_{it}) - \alpha - \sum_{1 \leq u_+ \leq q} \beta_r^T ((K_{it}, I_{it}) - (a,b))^u)^2 L_g((K_{it}, I_{it}) - (a,b)).$$

We also define $\widehat{\phi}_t = \widehat{\phi}(K_t, L_t)$. Next, for every b the estimator of $\pi(\cdot|b)$ is given by the third-order local polynomial estimator $\widehat{\pi}(c|b) = \widehat{\alpha}$, where

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} ((Y_{it} - \widehat{\beta}_L L_{it} - bK_{it}) - \alpha - \sum_{1 \leq v \leq 3} \beta_v^\top (\widehat{\phi}_{it-1} - bK_{it-1} - c)^v)^2 K_h(\widehat{\phi}_{it-1} - bK_{it-1} - c)\ ,$$

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. Again, we also define an infeasible estimator that uses the true value of the dependent variable. We set $\widehat{\pi}^*(c|b) = \widehat{\alpha}$, where

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} ((Y_{it} - \beta_L L_{it} - bK_{it}) - \alpha - \sum_{1 \leq v \leq 3} \beta_v^\top (\widehat{\phi}_{it-1} - bK_{it-1} - c)^v)^2 K_h(\widehat{\phi}_{it-1} - bK_{it-1} - c)\ ,$$

Our final estimator is then given as a solution to an empirical moment condition. Let

$$M_n(b) = \frac{1}{n}\sum_{i=1}^{n}(Y_{it} - \widehat{\beta}_L L_{it} - bK_{it} - \widehat{\pi}(\phi_{it-1} - bK_{it-1}|b))(K_{it} - \partial_b\widehat{\pi}(\phi_{it-1} - bK_{it-1}|b)K_{it-1})$$

Then the final estimator $\widehat{\beta}_K$ satisfies $M_n(\widehat{\beta}_K) = 0$.

To prove Proposition 2, we make the following assumptions. We remark that Assumption 12 follows under standard regularity conditions for the estimation of partially linear models, see e.g. Robinson (1988).

Assumption 11. *The sample observations $\{(Y_{it}, L_{it}, K_{it}, I_{it}, K_{it-1}, I_{it-1}), i = 1, \ldots, n\}$ are i.i.d.*

Assumption 12. *The estimator $\widehat{\beta}_L$ satisfies*

$$\sqrt{n}(\widehat{\beta}_L - \beta_L) = \mathbb{E}((L_t - \mathbb{E}(L_t|K_t, I_t))^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(L_{it} - \mathbb{E}(L_{it}|K_{it}, I_{it}))\eta_{it} + o_P(1).$$

Assumption 13. *(i) The random vector $S_{t-1} = (K_{t-1}, I_{t-1})$ is continuously distributed with compact support I_S. Its density function f_S is bounded and bounded away from zero on I_S, and is also $q+1$-times continuously differentiable for some uneven number $q \geq 3$. (ii) The function $\phi(s)$ is $q+1$-times continuously differentiable. (iii) Suppose that $\beta_K \in \int(B)$ for some known compact set B. For any $b \in B$, the random variable $T_{t-1}(b) = \phi(S_{t-1}) - bK_{t-1}$ is continuously distributed with compact support I_T. Its density function $f_T(\cdot, b)$ is bounded and bounded away from zero on I_T, uniformly over $b \in B$. The density is also four times continuously differentiable. (iv) For any $b \in B$, the function $\pi(\cdot, b)$ is four times continuously differentiable on I_T.*

Assumption 14. *For any $b \in B$, the residual $\varepsilon(b) = (Y_t - \beta_L L_t - bK_t) - \pi(T_{t-1}(b)|b)$ satisfies $E[\exp(l|\varepsilon(b)|)|S_{t-1}] \leq C$ almost surely for a constant $C > 0$ and $l > 0$ small enough.*

Assumption 15. *(i) The function K is two times continuously differentiable and satisfies the following conditions: $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int |u^2K(u)|du < \infty$, 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 $k \geq 2$, 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 16. *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 2. Again, we can use the same arguments as that of Corollary 1 and Example 1 to show this result. To show that $\kappa^* > 1/2$ under the conditions of the proposition, we proceed as in the proof of Proposition 1. To derive the influence function, it is useful to note that (4.2)–(4.3) hold with

$$\lambda_m(c) = -\mathbb{E}(G_t|T_{t-1} = c)$$
$$\lambda_r(c_1, c_2) = -(\mathbb{E}(\pi'(T_{t-1})G_t|S_{t-1} = c_1), \mathbb{E}(G_t|L_t = c_2))^\top.$$

Moreover, the proof uses that

$$\widehat{\phi}_{t-1} = \widehat{\phi}^*_{t-1} - \mathbb{E}(L_{t-1}|K_{t-1}, I_{t-1})(\widehat{\beta}_L - \beta_0) + o_P(n^{-1/2})$$
$$\widehat{\pi}(c|b) = \widehat{\pi}^*(c|b) - (\widehat{\beta}_L - \beta_L)\mathbb{E}(L_t|\phi_{t-1} - b_{K,t-1} = c) + o_P(n^{-1/2}).$$

This follows directly from the linearity of the local polynomial smoothing operator. □

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.

Cattaneo, M., R. Crump, and M. Jansson (2011): "Generalized Jackknife Estimators of Weighted Average Derivatives," *CREATES Research Papers*.

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 (2010): "Identification and Estimation of Semiparametric Two Step Models," *Unpublished manuscript.*

——— (2011): "Uniform Convergence for Semiparametric Two Step Estimators and Tests," *Unpublished manuscript.*

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 (2011): "The Asymptotic Variance of Semiparametric Estimators with Generated Regressors," *Unpublished manuscript.*

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," *forthcoming in the Annals of Statistics.*

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.

Pakes, A., and S. Olley (1995): "A limit theorem for a smooth class of semiparametric estimators," *Journal of Econometrics*, 65(1), 295–332.

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

Robinson, P. (1988): “Root-N-consistent semiparametric regression,” *Econometrica*, 56(4), 931–954.

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.

——— (2011): “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.