The Annals of Statistics
2012, Vol. 40, No. 2, 1132–1170
DOI: 10.1214/12-AOS995

NONPARAMETRIC REGRESSION WITH NONPARAMETRICALLY GENERATED COVARIATES

BY ENNO MAMMEN, CHRISTOPH ROTHE AND MELANIE SCHIENLE

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

We analyze the statistical properties of nonparametric regression estimators using covariates which are not directly observable, but have be estimated from data in a preliminary step. These so-called generated covariates appear in numerous applications, including two-stage nonparametric regression, estimation of simultaneous equation models or censored regression models. Yet so far there seems to be no general theory for their impact on the final estimator's statistical properties. Our paper provides such results. We derive a stochastic expansion that characterizes the influence of the generation step on the final estimator, and use it to derive rates of consistency and asymptotic distributions accounting for the presence of generated covariates.

1. Introduction. A wide range of statistical applications requires nonparametric estimation of a regression function when some of the covariates are not directly observed, but have themselves only been estimated in a (possibly nonparametric) preliminary step. Examples include triangular simultaneous equation models [e.g., Newey, Powell and Vella (1999), Blundell and Powell (2004), Imbens and Newey (2009)], sample selection models [Das, Newey and Vella (2003)], treatment effect models [Heckman, Ichimura and Todd (1998), Heckman and Vytlacil (2005)], censored regression models [Lewbel and Linton (2002)], generalized Roy models [d'Haultfoeuille and Maurel (2009)], stochastic volatility models [Kanaya and Kristensen (2009)] and GARCH-in-Mean models [Conrad and Mammen (2009)], amongst many others. In contrast to fully parametric settings [Pagan (1984)], there seems to be no general theoretical results on how to derive the statistical properties of such nonparametric two-step estimators. Instead, most available results in the literature typically exploit peculiarities of a specific model, and can thus not easily be transferred to other applications.

In this paper, we study the statistical properties of a nonparametric estimator $\hat{m}_{LL}$ of a conditional mean function $m_0(x) = \mathbb{E}(Y|r_0(S) = x)$ when the function r_0 is unknown, but can be estimated from data. While we are specific about estimating m_0 by local linear regression [Fan and Gijbels (1996)] to simplify technical arguments, we neither require the generated regressors $\hat{R} = \hat{r}(S)$ to emerge from a

Received November 2011.
MSC2010 subject classifications. 62G08, 62G20.
Key words and phrases. Nonparametric regression, two-stage estimators, simultaneous equation models, empirical process.

specific type of model, nor do we require a specific procedure to estimate them. We only impose high-level conditions on the accuracy and complexity of the first step estimate. In particular, our main result holds irrespectively of whether the function r_0 is, for example, a density, a conditional mean function or a quantile regression function, or whether it is estimated by kernel methods, orthogonal series or sieves. Moreover, our results are not confined to nonparametrically generated covariates, but also apply in settings where r_0 is estimated using parametric or semiparametric restrictions.

Our main result uses techniques from empirical process theory to show that the presence of generated covariates affects the first-order asymptotic properties of $\hat{m}_{LL}$ only through a *smoothed* version of the estimation error $\hat{r}(s) - r_0(s)$. This additional smoothing typically improves the rate of convergence of the estimator's stochastic part, reducing the "curse of dimensionality" from estimating r_0 to a secondary concern in this context. It does not, however, affect the order of magnitude of the deterministic component. Still, the estimator $\hat{m}_{LL}$ can have a faster overall rate of convergence than the first step estimator $\hat{r}$ if the latter has a sufficiently small bias.

We extensively illustrate the implications of our main result for the important special case that r_0 is the conditional mean function in an auxiliary nonparametric regression. For this setting, we derive simple and explicit stochastic expansions that can not only be used to establish asymptotic normality or the rate of consistency of the estimated regression function itself, but also study the properties of more complex estimators, in which estimation of a regression function merely constitutes an intermediate step, such as structured nonparametric models imposing additive separability [Stone (1985)]. Our results thus cover a wide range of models, and should therefore be of general interest. We use our techniques to study two such examples in greater detail: nonparametric estimation of a simultaneous equation model and nonparametric estimation of a censored regression model.

To the best of our knowledge, there are only few papers on nonparametric regression with estimated covariates not tailored to a specific application. Andrews (1995) derives some results for generated covariates converging at a parametric rate. Sperlich (2009) uses restrictive assumptions which lead to asymptotic results that are different from the ones obtained in the present paper. Song (2008) considers series estimation of the functional $g(x, r) = \mathbb{E}(Y|r(X) = x)$ indexed by $x \in \mathcal{X} \subset \mathbb{R}$ and $r \in \Lambda$, where Λ is a function space with finite integral bracketing entropy, and derives a rate of consistency uniformly over $(x, r) \in \mathcal{X} \times \Lambda$; see also Einmahl and Mason (2000) for a related problem.

Our paper is also related to a recent literature on semiparametric estimation problems with generated covariates. Li and Wooldridge (2002) consider a partial linear model with generated covariates. Hahn and Ridder (2011) use pathwise derivatives to derive the influence function of semiparametric linear GMM-type estimators. Escanciano, Jacho-Chávez and Lewbel (2011) provide stochastic expansions for sample means of weighted semiparametric regression residuals with potentially generated regressors, and study their application to certain index models. Compared to the nonparametric problems studied in this paper, semiparametric applications typically exhibit several additional technical issues. In particular, different techniques are needed to control the magnitude of certain remainder terms. Addressing these issues would require substantial refinements our results, which are not needed for the class of nonparametric problems we are focusing on. To keep the present paper more readable, we study semiparametric estimators with generated covariates separately in Mammen, Rothe and Schienle (2011).

The outline of this paper is as follows. In the next section, we describe our setup in detail. Section 3 gives some motivating examples. Section 4 establishes the asymptotic theory and states the main results. In Section 5, we apply our results to some of the examples given in Section 3, thus illustrating their application in practice. Finally, Section 6 concludes. All proofs are collected in the Appendix.

2. Nonparametric regression with generated covariates. The nonparametric regression model with generated regressors can be written as

$$Y = m_0(r_0(S)) + \varepsilon \qquad \text{with } \mathbb{E}(\varepsilon | r_0(S)) = 0, \tag{2.1}$$

where Y is the dependent variable, S is a p-dimensional vector of covariates, $m_0 : \mathbb{R}^d \to \mathbb{R}$ and $r_0 : \mathbb{R}^p \to \mathbb{R}^d$ are unknown functions and ε is an error term that has mean zero conditional on the true value of covariates to covariates $r_0(S)$.[1] We assume that there is additional information available outside of the basic model (2.1) such that the function r_0 is identified. For example, r_0 could be (some known transformation of) the mean function in an auxiliary nonparametric regression, which might involve another random vector, say T, in addition to Y and S.

Our aim is to estimate the function $m_0(x) = \mathbb{E}(Y | r_0(S) = x)$. Since r_0 is unobserved, obtaining a direct estimator based on a nonparametric regression of Y on $R = r_0(S)$ is clearly not feasible. We therefore consider the following two-stage procedure. In the first stage, an estimate $\hat{r}$ of r_0 is obtained. We do not require a specific estimator for this step. Instead, we only impose the high-level restrictions that the estimator $\hat{r}$ is uniformly consistent, converging at a rate specified below, and takes on values in a function class that is not too complex. Depending on the nature of the function r_0, these kind of regularity conditions are typically satisfied by various common nonparametric estimators, such as kernel-based procedures or series estimators, under suitable smoothness restrictions. In the second step, we then obtain our estimate $\hat{m}_{LL}$ of m_0 through a nonparametric regression of Y on

[1] Note that in contrast to an earlier working paper version of this paper, we do no longer assume that the "index" $r_0(S)$ is a sufficient statistic for the covariates S, which would imply that $\mathbb{E}(Y | r_0(S)) = \mathbb{E}(Y | S)$.

the generated covariates $\hat{R} = \hat{r}(S)$, using local linear smoothing. That is, our estimator is given by $\hat{m}_{LL}(x) = \hat{\alpha}$ obtained from

$$(\hat{\alpha}, \hat{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \bigl(Y_i - \alpha - \beta^T (\hat{R}_i - x)\bigr)^2 K_h(\hat{R}_i - x),$$

where $K_h(u) = \prod_{j=1}^{d} \mathcal{K}(u_j/h_j)/h_j$ is a d-dimensional product kernel with univariate kernel function $\mathcal{K}$, and $h = (h_1, \dots, h_d)$ is a vector of bandwidths that tend to zero as the sample size n increases to infinity.

For the later asymptotic analysis, it will also be useful to compare $\hat{m}_{LL}$ to an infeasible estimator $\tilde{m}_{LL}$ that uses the true function r_0 instead of an estimate $\hat{r}$. Such an estimator can be obtained by local linear smoothing of Y versus $R = r_0(S)$, that is, it is given by $\tilde{m}_{LL}(x) = \tilde{\alpha}$, where

$$(\tilde{\alpha}, \tilde{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \bigl(Y_i - \alpha - \beta^T (R_i - x)\bigr)^2 K_h(R_i - x).$$

In order to distinguish these two estimators, we refer to $\hat{m}_{LL}$ in the following as the *real* estimator, and to $\tilde{m}_{LL}$ as the *oracle* estimator.

Our use of local linear estimators in this paper is based on the following considerations. First, in a classical setting with fully observed covariates, estimators based on local linear regression are known to have attractive properties with regard to boundary bias and design adaptivity [see Fan and Gijbels (1996) for an extensive discussion], and they allow a complete asymptotic description of their distributional properties. In the present setting with generated covariates, these properties simplify the asymptotic treatment. The design adaptivity leads to a discussion of bias terms that does not require regular densities for the randomly perturbed covariates, and the complete asymptotic theory allows a clear description of how the final estimator is affected by the estimation of the covariates. On the other hand, our assumptions on the estimation of the covariates are rather general and can be verified for a broad class of smoothing methods, including sieves and orthogonal series estimators.

3. Motivating examples. There are many statistical applications which involve nonparametric estimation of a regression function using nonparametrically generated covariates. In this section, we give an overview of some of the most popular examples and explain how they fit into our framework. In Section 4, we revisit the first three of these examples, studying their asymptotic properties in detail. A thorough treatment of the remaining examples involves several additional technical issues beyond dealing with the presence of estimated covariates, such as boundary problems, and is thus omitted for brevity. See also Mammen, Rothe and Schienle (2011) for an extensive discussion of semiparametric problems with generated covariates.

3.1. *The generic example*: *Nonparametric two-stage regression*. In many applications, the unknown function r_0 is a conditional expectation function from an auxiliary nonparametric regression. As a first motivating example, we therefore consider a "two-stage" nonparametric regression model given by

$$Y = m_0(r_0(S)) + \varepsilon,$$
$$T = r_0(S) + \zeta,$$

where ζ is an unobserved error term that satisfies $E[\zeta|S] = E[\varepsilon|r_0(S)] = 0$. As the structure of this example is particularly simple, it is used extensively in Section 4 below to illustrate the application of our main result. Proceeding like this is instructive, as the types of technical difficulties encountered in this example are representative for those in a wide range of other statistical applications.

3.2. *Nonparametric censored regression*. Consider a nonparametric regression model with fixed censoring, that is,

$$Y = \max\bigl(0, \mu_0(X) - U\bigr), \tag{3.1}$$

where U is an unobserved mean zero error term that is assumed to be independent of the covariates X. Fixed censoring is a common phenomenon in many applications, for example, the analysis of wage data. Note that the censoring threshold could be different from zero, as long as it is known. Lewbel and Linton (2002) establish identification of the function μ_0 under the tail condition $\lim_{u\to-\infty} uF_U(u) = 0$ on the distribution function F_U of U. In particular, they show that the function μ_0 can be written as

$$\mu_0(x) = \lambda_0 - \int_{r_0(x)}^{\lambda_0} \frac{1}{q_0(r)}\, dr, \tag{3.2}$$

where $r_0(x) = \mathbb{E}(Y|X = x)$, $q_0(r) = \mathbb{E}(\mathbb{I}\{Y > 0\}|r_0(X) = r)$, and λ_0 is some suitably chosen constant. An estimate of the function μ_0 can then be obtained from a sample analog of (3.2), that is, through numerical integration of a nonparametric estimate of the function $q_0(r)^{-1}$. Nonparametric estimation of q_0 involves nonparametrically generated regressors, and thus fits into our framework with $(Y, S) = (\mathbb{I}\{Y > 0\}, X)$ and $r_0(S) = r_0(X)$.

3.3. *Nonparametric triangular simultaneous equation models*. Covariates that are correlated with disturbance terms appear in many economic models and are denoted as endogenous. When, for example, analyzing the relationship between wages and schooling, unobserved individual characteristics like ability or motivation might affect both the outcome and the explanatory variable. A common approach is to model these quantities jointly, achieving identification by using so-called instrumental variables, that are independent of unobservables, affect the endogenous variable, but exert no direct influence on the outcome. Consider, for example, the nonparametric triangular simultaneous equation model discussed in Newey, Powell and Vella (1999), which is of the form

$$Y = \mu_1(X_1, Z_1) + U, \tag{3.3}$$

$$X_1 = \mu_2(Z_1, Z_2) + V. \tag{3.4}$$

Here the interest is in estimating the function μ_1. To achieve identification, one imposes the restrictions $\mathbb{E}(V|Z_1, Z_2) = 0$, $\mathbb{E}(U) = 0$ and $\mathbb{E}(U|Z_1, Z_2, V) = \mathbb{E}(U|V)$, which follow, for example, if the vector of exogenous covariates and instruments $Z = (Z_1, Z_2)$ is jointly independent of the disturbances (U, V). Now let $m(x_1, z_1, v) = \mathbb{E}(Y|X_1 = x_1, Z_1 = z_1, V = v)$. Under the above assumptions, it is straightforward to show that

$$m(x_1, z_1, v) = \mu_1(x_1, z_1) + \lambda(v),$$

where $\lambda(v) = \mathbb{E}(U|V = v)$. The first component of this additive model could, for example, be estimated by marginal integration [Newey (1994a), Linton and Nielsen (1995)], which relies on the fact that

$$\int m(x_1, z_1, v) f_V(v)\, dv = \mu_1(x_1, z_1), \tag{3.5}$$

where f_V is the probability density function of V. Implementing a sample version of (3.5) requires estimating the function m. Since the residuals V are not directly observed but must be estimated by some nonparametric method, this fits into our framework with $(Y, S) = (Y, (X_1, Z_1, Z_2), X_1)$ and $r_0(S) = (X_1, Z_1, X_1 - \mu_2(Z_1, Z_2))$.

REMARK 1. An alternative to marginal integration would be an approach based on smooth backfitting [Mammen, Linton and Nielsen (1999)]. Smooth backfitting estimators avoid several problems encountered by marginal integration in case of covariates with moderate or high dimension, but involves a more involved statistical analysis which is beyond the scope of the present paper. We are going to study smooth backfitting with nonparametrically generated covariates in a separate paper.

3.4. *Generalized Roy model.* D'Hautfoeuille and Maurel (2009) consider a generalized Roy model of occupational choice that is related to the previous example in the sense that it also leads to an additive regression model. Let Y_k denote the individual's potential earnings in sector $k \in \{0, 1\}$ of an economy, $X = (X_0, X_1, X_c)$ a vector of covariates, and assume that $\mathbb{E}(Y_k|X, \eta_1, \eta_2) = \psi_k(X_k, X_c) + \eta_k$, where (η_0, η_1) are sector-specific productivity terms known by the agent but unobserved by the analyst. Expected utility from working in sector k is assumed to be $U_k = \mathbb{E}(Y_k|X, \eta_1, \eta_2) + G_k(X)$, the sum of sector-specific expected earnings and a nonpecuniary component that depends on X. Along with X, the analyst observes the chosen sector D, which satisfies $D = \mathbb{I}\{U_1 > U_0\}$, and the realized earnings $Y = DY_1 + (1 - D)Y_0$.

One object of interest in this context is the pair of functions (ψ_1, ψ_0). Under some weak additional conditions, d'Haultfoeuille and Maurel (2009) show that

$$\mathbb{E}(Y|D = d, X) = \psi_d(X_d, X_c) + \lambda_d\big(\Pr(D = d|X)\big)$$

for $d \in \{0, 1\}$, which is again an additive model involving unobserved covariates, namely the conditional probabilities $\Pr(D = d|X)$ of choosing sector d. This setting fits into our framework in the same way as the previous example.

3.5. *Nonparametric nonseparable triangular simultaneous equation models.* Imbens and Newey (2009) consider a generalized version of the above-mentioned triangular simultaneous equation model with nonadditive disturbances:

$$Y = \mu_1(X_1, Z_1, U), \tag{3.6}$$

$$X_1 = \mu_2(Z_1, Z_2, V). \tag{3.7}$$

Nonseparable models have become popular in the recent econometric literature, as they allow for substantially more general forms of unobserved heterogeneity than specifications in which the disturbance terms enter additively. The focus here is typically on averages of the function μ_1, such as the average structural function,

$$\mathrm{ASF}(x_1, z_1) = \mathbb{E}_U(\mu_1(x_1, z_1, U)).$$

To achieve identification, assume that the function μ_2 is strictly monotone in its last argument, that V is continuously distributed, and that the unobserved disturbances (U, V) are jointly independent of Z. Then it can be shown that U and (X_1, Z_1) are independently conditional on the so-called control variable $W = F_{X_1|Z}(X_1, Z)$, where $F_{X_1|Z}$ denotes the distribution function of X_1 given Z. Under an additional support condition, this result implies that the ASF is identified through the relationship

$$\mathrm{ASF}(x_1, z_1) = \int m(x_1, z_1, w)\, dF_W, \tag{3.8}$$

where $m(x_1, z_1, w) = \mathbb{E}(Y|X_1 = x_1, Z_1 = z_1, W = w)$. Since the control variable W is unobserved and has to be estimated in order to implement a sample analog estimator of (3.8), this setting also fits into the framework of this paper. In particular, nonparametric estimation of m is covered with $(Y, S) = (Y, (X_1, Z_1, Z_2), X_1)$ and $r_0(S) = (X_1, Z_1, F_{X_1|Z}(X_1, Z))$.

4. Asymptotic properties. It is straightforward to show that $\hat{m}_{LL}$ consistently estimates the function m_0 under standard conditions. Obtaining refined asymptotic properties, however, requires more involved arguments. In this section, we derive a stochastic expansion of the difference between the real and the oracle estimator, in which the leading terms are kernel-weighted averages of the first stage estimation error. This is our main result. It can be used, for example, to obtain uniform rates of consistency for the real estimator, or to prove its asymptotic normality. We demonstrate this in the next section for specific forms of r_0 and $\hat{r}$.

Throughout this section, we use the notation that for any vector $a \in \mathbb{R}^d$ the value $a_{\min} = \min_{1 \le j \le d} a_j$ denotes the smallest of its elements, $a_+ = \sum_{j=1}^d a_j$ 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 $b \in \mathbb{R}^d$. For ease of presentation in the following, we avoid logarithmic terms in rates of convergence; that is, we state assumptions and results in the form $o_P(n^{\xi})$ instead of $O_P(\log n^{\gamma})$ with $\xi, \gamma > 0$.

4.1. *Assumptions*. In order to analyze the asymptotic properties of the local linear estimator with nonparametrically generated regressors, we make the following assumptions.

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

(i) The sample observations (Y_i, S_i) are i.i.d.

(ii) The random vector $R = r_0(S)$ is continuously distributed with compact support I_R. Its density function f_R is twice continuously differentiable and bounded away from zero on I_R.

(iii) The function m_0 is twice continuously differentiable on I_R.

(iv) $E[\exp(l|\varepsilon|)|S] \le C$ almost surely for a constant $C > 0$ and $l > 0$ small enough.

(v) The kernel function $\mathcal{K}$ is a twice continuously differentiable, symmetric density function with compact support, say $[-1, 1]$.

(vi) The bandwidths $h = (h_1, \ldots, h_d)$ satisfies $h_j \sim n^{-\eta_j}$ for $j = 1, \ldots, d$ and $\eta_+ < 1$.

Most conditions in Assumption 1 are standard regularity and smoothness conditions for kernel-type nonparametric regression, with the exception of Assumption 1(iv). The subexponential tails of ε conditional on S assumed there are needed to apply certain results from empirical process theory in our proofs. Such a condition is not very restrictive though.

ASSUMPTION 2 (Accuracy). The components $\hat{r}_j$ and $r_{0,j}$ of $\hat{r}$ and r_0, respectively, satisfy

$$\sup_s |\hat{r}_j(s) - r_{0,j}(s)| = o_P(n^{-\delta_j})$$

for some $\delta_j > \eta_j$ and all $j = 1, \ldots, d$.

Assumption 2 is a "high-level" restriction on the accuracy of the estimator $\hat{r}$. It requires each component of the estimate of the function r_0 to be uniformly consistent, converging at rate at least as fast as the corresponding bandwidth in the second stage of the estimation procedure. This is typically not a restrictive condition, and it allows for estimators $\hat{r}$ that converge at a rate slower than the oracle estimator $\tilde{m}_{LL}$. Uniform rates of consistency are widely available for all common nonparametric estimators; see, for example, Masry (1996) for results on the Nadaraya–Watson, local linear and local polynomial estimators, or Newey (1997) for series estimators.

ASSUMPTION 3 (Complexity). There exist sequences of sets $\mathcal{M}_{n,j}$ such that:

(i) $\Pr(\hat{r}_j \in \mathcal{M}_{n,j}) \to 1$ as $n \to \infty$ for all $j = 1, \dots, d$.

(ii) For a constant $C_M > 0$ and a function $r_{n,j}$ with $\|r_{n,j} - r_{0,j}\|_\infty = o(n^{-\delta_j})$, the set $\overline{\mathcal{M}}_{n,j} = \mathcal{M}_{n,j} \cap \{r_j : \|r_j - r_{n,j}\|_\infty \le n^{-\delta_j}\}$ can be covered by at most $C_M \exp(\lambda^{-\alpha_j} n^{\xi_j})$ balls with $\|\cdot\|_\infty$-radius λ for all $\lambda \le n^{-\delta_j}$, where $0 < \alpha_j \le 2$, $\xi_j \in \mathbb{R}$ and $\|\cdot\|_\infty$ denotes the supremum norm.

Assumption 3 requires the first-stage estimator $\hat{r}$ to take values in a function space $\mathcal{M}_{n,j}$ that is not too complex, with probability approaching 1. Here the complexity of the function space is measured by the cardinality of the covering sets. This is a typical requirement for many results from empirical process theory; see van der Vaart and Wellner (1996). The second part of Assumption 3 is typically fulfilled under suitable smoothness restrictions. For example, suppose that $\mathcal{M}_{n,j}$ is the set of functions defined on some compact set $I_S \subset \mathbb{R}^p$ whose partial derivatives up to order k exist and are uniformly bounded by some multiple of $n^{\xi_j^*}$ for some $\xi_j^* \ge 0$. Then Assumption 3(ii) holds with $\alpha_j = p/k$ and $\xi_j = \xi_j^* \alpha_j$ [van der Vaart and Wellner (1996), Corollary 2.7.2]. For kernel-based estimators of r_0, one can then verify part (i) of Assumption 3 by explicitly calculating the derivatives. Consider, for example, the one-dimensional Nadaraya–Watson estimator $\hat{r}_{n,j}$ 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 $\hat{r}_{n,j} - r_{n,j}$ is absolutely bounded by $O_P(\sqrt{\log n}) = o_P(n^{\xi_j^*})$ for all $\xi_j^* > 0$. 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 (2000).

ASSUMPTION 4 (Continuity). For any $r \in \mathcal{M}_n = \mathcal{M}_{n,1} \times \cdots \times \mathcal{M}_{n,d}$ the conditional expectation $\tau^B(x, r) = \mathbb{E}(\rho(S) | r(S) = x)$ with $\rho(S) = \mathbb{E}(Y|S) - \mathbb{E}(Y|r_0(S))$ exists and is twice differentiable with respect to its first argument, with derivatives that are uniformly bounded in absolute value, and satisfies

$$\|\tau^B(x, r_1) - \tau^B(x, r_2)\| \le C_B^* \|r_1 - r_2\|_\infty \qquad \text{a.s.}$$

for all $r_1, r_2 \in \mathcal{M}_n$ and a constant $C_B^* > 0$.

Assumption 4 imposes certain smoothness restrictions on the conditional expectation of $\rho(S)$. The term $\rho(S)$ can be thought of as capturing the influence of the underlying covariates S on the outcome variable Y that is not excreted through the "index" $r_0(S)$. In certain applications, the "index" $r_0(S)$ is a sufficient statistic for the function m_0, and thus $\rho(S) = 0$ with probability 1. In this case, Assumption 4 is trivially satisfied. Note that $\rho(S) = \mathbb{E}(\varepsilon|S)$, and that $\tau^B(\cdot, r_0) \equiv 0$ by construction.

4.2. *The key stochastic expansion.* With the assumptions given in the previous section, we are now ready to state our main result, which is a stochastic expansion of the real estimator $\hat{m}_{LL}(x)$ around the oracle estimator $\tilde{m}_{LL}(x)$. Our aim is to derive an explicit characterization of the influence of the presence of generated regressors on the final estimator of the function m_0. To this end, we define $w(x, r) = (1, (r_1(S) - x_1)/h_1, \ldots, (r_d(S) - x_d)/h_d)$, and set $N_h(x) = \mathbb{E}(w(x, r)w(x, r)^T K_h(r(S) - x))$. Next, we define

$$\Delta(x, r) = e_1^\top N_h(x)^{-1} \mathbb{E}\big(K_h\big(r_0(S) - x\big) w(x, r)\big(r(S) - r_0(S)\big)\big),$$
$$\Gamma(x, r) = e_1^\top N_h(x)^{-1} \mathbb{E}\big(K_h'\big(r_0(S) - x\big)^\top w(x, r)\big(r(S) - r_0(S)\big)\rho(S)\big)$$

for any $r \in \mathcal{M}_n$, where $K_h'(u) = (\mathcal{K}_{h,j}'(u) : j = 1, \ldots, d)^\mathsf{T}$ is a vector with elements $\mathcal{K}_{h,j}'(u) = \mathcal{K}'(u_j/h_j)/h_j^2 \prod_{j^* \neq j} \mathcal{K}(u_{j^*}/h_{j^*})/h_{j^*}$. Finally, we put $\hat{\Delta}(x) = \Delta(x, \hat{r})$ and $\hat{\Gamma}(x) = \Gamma(x, \hat{r})$. With this notation, we can now state our main theorem.

THEOREM 1. *Suppose Assumptions 1–4 hold. Then*

$$\sup_{x \in \mathrm{I_R}} |\hat{m}_{LL}(x) - \tilde{m}_{LL}(x) + m_0'(x)\hat{\Delta}(x) - \hat{\Gamma}(x)| = O_P(n^{-\kappa}),$$

where $\kappa = \min\{\kappa_1, \ldots, \kappa_3\}$ *with*

$$\kappa_1 < \frac{1}{2}(1 - \eta_+) + (\delta - \eta)_{\min} - \frac{1}{2} \max_{1 \le j \le d} (\delta_j \alpha_j + \xi_j),$$
$$\kappa_2 < 2\eta_{\min} + (\delta - \eta)_{\min},$$
$$\kappa_3 < \delta_{\min} + (\delta - \eta)_{\min}.$$

The two leading terms in our stochastic expansion of the real estimator $\hat{m}_{LL}(x)$ around the oracle estimator $\tilde{m}_{LL}(x)$, which are accounting for the presence of generated covariates, are both smoothed versions of the first-stage estimation error $\hat{r}(s) - r_0(s)$. To see this more clearly, note that it follows from standard arguments for local polynomial smoothing that

$$\Delta(x,r) = \frac{\mathbb{E}(K_h(r_0(S)-x)(r(S)-r_0(S)))}{f_R(x)} + O_P(n^{-\kappa}) \qquad \text{and}$$

$$\Gamma(x,r) = \frac{\mathbb{E}(K_h'(r_0(S)-x)^\top(r(S)-r_0(S))\rho(S_i))}{f_R(x)} + O_P(n^{-\kappa}),$$

uniformly over $x \in I_{R,n}^- = \{x \in I_R : \text{the support of } K_h(\cdot - x) \text{ is a subset of } I_R\}$. In order to achieve a certain rate of convergence for the real estimator, it is thus not necessary to have an estimator of r_0 that converges with the same rate or a faster one, since the asymptotic properties of the estimator using nonparametrically generated regressors only depend on a smoothed version of the first-stage estimation error. While smoothing does not affect the order of the estimator's deterministic part, it typically reduces the variance and thus allows for less precise first-stage estimators. Note that the first adjustment term is negligible in regions where the regression function is flat, since $m_0'(x) = 0$ in this case. Conversely, the impact of generated covariates is accentuated when the true regression function is steep. Also note that $\hat{\Gamma}(x) = 0$ when $\mathbb{E}(\varepsilon|S) = 0$, as the latter implies that $\rho(s) \equiv 0$. This is a natural condition in certain empirical applications.

REMARK 2. In Theorem 1 no assumptions are made about the process generating the data for estimation of r_0. In particular, nothing is assumed about dependencies between the errors in the pilot estimation and the regression errors ε_i. We conjecture that better rates than $n^{-\kappa}$ can be proven under such additional assumptions, but the results would only be specific to the respective full model under consideration. One way to extend our approach to such a setting would be to use our empirical process methods to bound the remainder term of higher order differences between $\hat{m}$ and $\tilde{m}$, and to treat the leading terms of the resulting higher order expansion by other, more direct methods.

5. Examples revisited. In this section, we apply our high-level results from Section 4 to some of the motivating examples presented in Section 3, which are representative for the others in terms of employed techniques. Assuming a specific nature of the function r_0 and a specific method to estimate it, explicit forms of the adjustment terms $\hat{\Delta}(x)$ and $\hat{\Gamma}(x)$ in Theorem 1 can be derived in order to account for the presence of generated covariates. Our focus in this section is on the practically most important case that r_0 is the conditional mean function in an auxiliary nonparametric regression. Many other applications can be treated along the same lines.

5.1. *Generic example*: *Two-stage nonparametric regression*. The main setting in which we illustrate the application of the stochastic expansion from Theorem 1 is the "two-stage" nonparametric regression model given by

$$Y = m_0(r_0(S)) + \varepsilon,$$
$$T = r_0(S) + \zeta,$$

where ζ is an unobserved error term that satisfies $E[\zeta|S] = E[\varepsilon|r_0(S)] = 0$. For simplicity, we focus on the case that $R = r_0(S)$ is a one-dimensional covariate, but generalizations to multiple generated covariates or the presence of additional observed covariates are immediate.

Our strategy for deriving asymptotic properties of $\hat{m}_{LL}$ in this framework is to first provide an explicit representation for the adjustment terms $\hat{\Delta}(x)$ and $\hat{\Gamma}(x)$ from Theorem 1, which are then combined with standard results about the oracle estimator $\tilde{m}_{LL}$. For this approach it is convenient to use a kernel-based smoother to estimate r_0. Since the bias of both $\hat{\Delta}(x)$ and $\hat{\Gamma}(x)$ is of the same order as of this first-stage estimator, we propose to estimate the function r_0 via qth order local polynomial smoothing, which includes the local linear estimator as the special case $q = 1$. Formally, the estimator is given by $\hat{r}(s) = \hat{\alpha}$, where

$$(5.1)\qquad (\hat{\alpha}, \hat{\beta}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \biggl(T_i - \alpha - \sum_{1 \le u_+ \le q} \beta_r^T (S_i - s)^u \biggr)^2 L_g(S_i - s)$$

and $L_g(s) = \prod_{j=1}^{p} \mathcal{L}(s_j/g)/g$ is a p-dimensional product kernel built from the univariate kernel $\mathcal{L}$, g is a vector of bandwidths, whose components are assumed to be the same for simplicity, and $\sum_{1 \le u_+ \le q}$ denotes the summation over all $u = (u_1, \dots, u_p)$ with $1 \le u_+ \le q$. When r_0 is sufficiently smooth, the asymptotic bias of local polynomial estimators of order q is well known to be $O(g^{q+1})$ uniformly over $x \in I_R$ (if q is uneven), and can thus be controlled. A further technical advantage of using local polynomials is that the corresponding estimator admits a certain stochastic expansion under general conditions, which is useful for our proofs. We make the following assumption, which is essentially analogous to Assumption 1, except for Assumption 4(iii). This additional assumption requires higher order smoothness of the kernel, necessary to bound the kth derivative of the estimator $\hat{r}$. This allows us to verify Complexity Assumption 3 for $\hat{r}$.

ASSUMPTION 5. We assume the following properties for the data distribution, the bandwidth and kernel function $\mathcal{L}$:

(i) The observations (S_i, Y_i, T_i) are i.i.d., and the random vector S is continuously distributed with compact support I_S. Its density function f_S is bounded and bounded away from zero on I_S. It is also differentiable with a bounded derivative. The residuals ζ satisfy $\mathbb{E}|\zeta|^{\epsilon} < \infty$ for some $\epsilon > 2$.

(ii) The function r_0 is $q + 1$ times continuously differentiable on I_S.

(iii) The kernel function $\mathcal{L}$ is a k-times continuously differentiable, symmetric density function with compact support, say $[-1, 1]$, for some natural number $k \ge \max\{2, p/2\}$.

(iv) The bandwidth satisfies $g \sim n^{-\theta}$ for some $0 < \theta < 1/p$.

To simplify the presentation, we also assume that the function $r_0(s)$ is strictly monotone in at least one of its arguments, which can be taken to be the last one without loss of generality. This assumption could be easily removed at the cost of a substantially more involved notation in the following results.

ASSUMPTION 6. The function $r_0(u_{-p}, u_p)$ is strictly monotone in u_p, and we have that $r_0(u_{-p}, \varphi(u_{-p}, x)) = x$ for some twice continuously differentiable function φ.

The following proposition shows that in the present context the function $\hat{\Delta}(x)$ can be written as the sum of a smoothed version of the first stage estimator's bias function, a kernel-weighted average of the first-stage residuals $\zeta_1, \dots, \zeta_n$, and some higher order remainder terms. For a concise presentation of the result, we introduce some particular kernel functions. Let L^* denote the p-dimensional equivalent kernel of the local polynomial regression estimator, given in (A.27) in the Appendix, and define the one-dimensional kernel functions

$$J_h(x, s) = \int K_h\bigl(r_0(s) - x - \partial_s r_0(s) u h\bigr) L^*(u)\, du,$$

$$H_g^{\Delta}(x, v) = \frac{\partial_x \varphi(v_{-p}, x)}{g} \int L^*\biggl(s_{-p}, \frac{\varphi(v_{-p}, x) - v_p}{g} + s_p \partial_{-p} \varphi(v_{-p}, x)\biggr)\, ds.$$

Then, with this notation, we obtain the following proposition.

PROPOSITION 1. *Suppose that Assumptions 1 and 4–6 hold. Then we have for the correction factor $\hat{\Delta}$ in Theorem 1 that*

$$\sup_{x \in I_R} |\hat{\Delta}(x) - \hat{\Delta}_A(x) - \hat{\Delta}_B(x)| = O_p\biggl(\frac{\log(n)}{n g^p}\biggr),$$

where the terms $\hat{\Delta}_A(x)$ and $\hat{\Delta}_B(x)$ satisfy

$$\sup_{x \in I_R} |\hat{\Delta}_A(x)| = O_p\bigl(\bigl(\log(n)/(n \max\{g, h\})\bigr)^{1/2}\bigr) \quad \textit{and}$$

$$\sup_{x \in I_R} |\hat{\Delta}_B(x)| = O_p(g^{q+1}).$$

Moreover, uniformly over $x \in I_{R,n}^-$, it is $\hat{\Delta}_B(x) = g^{q+1} E[b(S) | r_0(S) = x] + o_p(g^{q+1})$ with a bounded function $b(s)$ given in (A.25) in the Appendix, and the term $\hat{\Delta}_A(x)$ allows for the following expansions uniformly over $x \in I_{R,n}^-$, depending on the limit of g/h:

(a) *If* $g/h \to 0$, *then*

$$\hat{\Delta}_A(x) = \frac{1}{nf_R(x)} \sum_{i=1}^{n} K_h\big(r_0(S_i) - x\big)\zeta_i + O_p\biggl(\biggl(\frac{g^2}{h^2} + \frac{g^{3/2}}{h}\biggr)\biggl(\frac{\log(n)}{nh}\biggr)^{1/2}\biggr).$$

(b) *If* $h = g$, *then*

$$\hat{\Delta}_A(x) = \frac{1}{nf_R(x)} \sum_{i=1}^{n} J_h(x, S_i)\zeta_i + O_p\biggl(\biggl(\frac{\log(n)}{n}\biggr)^{1/2}\biggr).$$

(c) *If* $g/h \to \infty$, *then*

$$\hat{\Delta}_A(x) = \frac{1}{nf_R(x)} \sum_{i=1}^{n} H_g^{\Delta}(x, S_i)\zeta_i + O_p\biggl(\frac{g^2}{h^2}\biggl(\frac{\log(n)}{ng}\biggr)^{1/2} + \biggl(\frac{\log(n)}{n}\biggr)^{1/2}\biggr).$$

It should be emphasized that in all three cases of the above proposition the leading term in the expression for $\hat{\Delta}_A(x)$ is equal to an average of the error terms ζ_i weighted by a *one-dimensional* kernel function, irrespective of $p = \dim(S)$. The dimension of the covariates thus affects the properties of $\hat{\Delta}(x)$ only through higher-order terms. Furthermore, it should be noted that one can also derive expressions of $\hat{\Delta}(x)$ similar to the ones above for values of x close to the boundary of the support. Likewise these take the form of a one-dimensional kernel weighted average of the error terms ζ_i plus a higher-order term. The corresponding kernel function, however, has a more complicated closed form varying with the point of evaluation.

The following proposition establishes a result similar to Proposition 1 for the second adjustment term $\hat{\Gamma}(x)$. We again introduce a particular one-dimensional kernel function, defined as

$$H_g^{\Gamma}(x, v) = \int g^{-1} L^*\biggl(s_{-p}, \frac{\varphi(v_{-p}, x) - v_p}{g} + s_p\, \partial_p \varphi(v_{-p}, x)\biggr)\, ds \lambda(v_{-p}, x)$$

with

$$\lambda(v_{-p}, x) = \frac{\partial_{v_p}(\rho(v_{-p}, \varphi(v_{-p}, x) f_S(v_{-p}, \varphi(v_{-p}, x))\det(\partial_{v_{-p}}\varphi(v_{-p}, x))}{f_S(v_{-p}, \varphi(v_{-p}, x))\, \partial_{v_p} r_0(v_{-p}, \varphi(v_{-p}, x))},$$

where L^* still denotes the p-dimensional equivalent kernel of the local polynomial regression estimator, given in (A.27) in the Appendix.

PROPOSITION 2. *Suppose that Assumptions* 1 *and* 4–6 *hold. Then we have that*

$$\sup_{x \in I_R} |\hat{\Gamma}(x) - \hat{\Gamma}_A(x) - \hat{\Gamma}_B(x)| = O_p\biggl(\frac{\log(n)}{ng^p}\biggr),$$

where the terms $\hat{\Gamma}_A(x)$ *and* $\hat{\Gamma}_B(x)$ *satisfy*

$$\sup_{x\in I_R} |\hat{\Gamma}_A(x)| = O_p\big((\log(n)/(ng))^{1/2}\big) \quad \textit{and} \quad \sup_{x\in I_R} |\hat{\Gamma}_B(x)| = O_p(g^{q+1}).$$

Moreover, uniformly over $x \in I_{R,n}^-$, *it is* $\hat{\Gamma}_B(x) = g^{q+1}\,\partial_x E[b(S)\rho(S)|r_0(S) = x] + o_p(g^{q+1})$ *with a bounded function* $b(s)$ *given in* (A.25) *in the Appendix, and the term* $\hat{\Gamma}_A(x)$ *allows for the following expansion uniformly over* $x \in I_{R,n}^-$:

$$\hat{\Gamma}(x) = \frac{1}{nf_R(x)}\sum_{i=1}^{n} H_g^{\Gamma}(x, S_i)\zeta_i + o_P\left(\sqrt{\frac{\log(n)}{ng}}\right). \tag{5.2}$$

Again, the leading term in the expression for $\hat{\Gamma}_A(x)$ is equal to an average of the error terms ζ_i weighted by a *one-dimensional* kernel function, and thus behaves similarly to one-dimensional nonparametric regression estimator. A similar result could be established for regions close to the boundary of the support. Note that in contrast to Proposition 1, the details of the result in Proposition 2 do not depend on the relative magnitude of the bandwidths used in the first and second stage of the estimation procedure.

Combining Theorem 1 and Propositions 1–2 with well-known results about the oracle estimator $\tilde{m}_{LL}$, various asymptotic properties of the real estimator $\hat{m}_{LL}$ can be derived. In the following corollaries we present results for the most relevant scenarios, addressing uniform rates of consistency and stochastic expansions of order $o_P(n^{-2/5})$ for proving pointwise asymptotic normality. More refined expansions of higher orders such as $o_P(n^{-1/2})$, which are useful for the analysis of semiparametric problems in which m_0 plays the role of an infinite dimensional nuisance parameter [e.g., Newey (1994b), Andrews (1994), Chen, Linton and Van Keilegom (2003)], would also be possible. We do not present such results here as they would require strong smoothness restrictions that are unattractive in applications. See Mammen, Rothe and Schienle (2011) for an alternative approach to controlling the influence of generated covariates in semiparametric models.

Starting with considering the uniform rate of consistency, it is well known [Masry (1996)] that under Assumption 1 the oracle estimator satisfies

$$\sup_{x\in I_R} |\tilde{m}_{LL}(x) - m(x)| = O_p\big((\log(n)/nh)^{1/2} + h^2\big).$$

This implies the following result.

COROLLARY 1. *Suppose that Assumptions* 1, 4 *and* 5 *hold. Then*

$$\sup_{x\in I_R} |\hat{m}_{LL}(x) - m(x)| = O_p\left(\frac{\log(n)^{1/2}}{(n\max\{h,g\})^{1/2}} + h^2 + \frac{\log(n)}{ng^p} + g^{q+1} + n^{-\kappa}\right).$$

Straightforward calculations show that, under appropriate smoothness restrictions, it is possible to recover the oracle rate for the real estimator given suitable choice of η and θ, even if the first-stage estimator converges at a strictly slower rate. Note that the rate in Corollary 1 improves upon a bound on the uniform rate of convergence of a two-stage regression estimator derived in Ahn (1995) for a similar setting.

Next, we derive stochastic expansions of $\hat{m}_{LL}$ of order $o_P(n^{-2/5})$ for the case that $\eta = 1/5$. Such expansions immediately imply results on pointwise asymptotic normality of the real estimator. We start with the case that $\theta = \eta$, in which the stochastic terms $\hat{\Gamma}_A(x)$ and $\hat{\Delta}_A(x)$ are of the same order of magnitude (other bandwidth choices will be discussed below). During the analysis of this setting, it becomes clear that applying Theorem 1 requires $p\theta < 3/10$. Thus in order to use the expansion in Proposition 1(b), only $p = 1$ is admissible; that is, S must be one-dimensional for the choice $\theta = \eta$ to be feasible. In this setting, the notation for the kernel functions appearing in the stochastic expansions can be somewhat simplified. We define

$$\tilde{J}(v,x) = \int K\big(v - r_0'(r_0^{-1}(x))u\big)L^*(u)\,du,$$

$$\tilde{H}^{\Gamma}(v,x) = \int L^*\big(v + s\,\partial_x r_0^{-1}(x)\big)\,ds\tilde{\lambda}(x),$$

where

$$\tilde{\lambda}(x) = \frac{\partial_v(\rho(r_0^{-1}(x))f_S(r_0^{-1}(x)))}{f_S(r_0^{-1}(x))r_0'(r_0^{-1}(x))},$$

where r_0^{-1} is the inverse function of r_0, which exists by Assumption 6.

COROLLARY 2. *Suppose that Assumptions 1 and 4–6 hold with $\eta = \theta = 1/5$ and $p = q = 1$. Then the following expansions hold uniformly over $x \in I_{R,n}^-$:*

$$\begin{aligned}
&\hat{m}_{LL}(x) - m_0(x)\\
&\quad= \frac{1}{nf_R(x)}\sum_{i=1}^n K_h\big(r_0(S_i) - x\big)\varepsilon_i\\
&\qquad - \frac{1}{nf_R(x)}\sum_{i=1}^n\big(m_0'(x)\tilde{J}_h\big(r_0(S_i) - x, x\big) - \tilde{H}_h^{\Gamma}\big(S_i - r_0^{-1}(x), x\big)\big)\zeta_i\\
&\qquad + \frac{1}{2}\beta(x)h^2 + o_p(n^{-2/5}),
\end{aligned}$$

where the bias is given by

$$\begin{aligned}
\beta(x) &= \int u^2K(u)\,du m_0''(x)\\
&\quad - \int u^2L(u)\,du\big(r_0''(r_0^{-1}(x))m_0'(x) - \partial_x[r_0''(r_0^{-1}(x))\rho(r_0^{-1}(x))]\big).
\end{aligned}$$

In particular, we have

$$(nh)^{1/2}\big(\hat{m}_{LL}(x) - m_0(x) - \beta(x)h^2\big) \xrightarrow{d} N(0, \sigma_m^2(x)),$$

where $\sigma_m^2(x) = [\mathrm{Var}(\varepsilon|R = x)\int K(t)^2\,dt - 2E(\varepsilon\zeta|R = x)\int K(t)(\tilde{J}(t,x)m_0'(x) - \tilde{H}^\Gamma(t,x))\,dt\,\mathrm{Var}(\zeta|R = x)\int(m_0'(x)\tilde{J}(t,x) - \tilde{H}^\Gamma(t,x))^2\,dt]/f_R(x)$ *is the asymptotic variance.*

Under the conditions of the corollary, the limiting distribution of $\hat{m}_{LL}(x)$ is generally affected by the pilot estimation step, although a qualitative description of the impact seems difficult. Depending on the curvature of m_0 and the covariance of ε and ζ, the asymptotic variance of the estimator using generated regressors can be bigger or smaller than that of the oracle estimator $\tilde{m}_{LL}$. There thus exist settings where in practice it would be preferable to base inference on the real estimator even if one was actually able to compute the oracle estimator.

The next corollary considers the case that $\theta > \eta$, and thus $g/h \to 0$. Again, applying Theorem 1 requires $p\theta < 3/10$ in this setting, and thus only $p = 1$ is admissible when using Proposition 1(a) for such a choice of bandwidths. The corollary also focuses on the special case that $\rho(S) := \mathbb{E}(Y|R) - \mathbb{E}(Y|S) = 0$, which implies that $\hat{\Gamma}(x) = 0$ with probability 1. This condition is satisfied for certain empirical applications, such as, for example, models IV models. Without this additional restriction, an expansion of the difference $\hat{m}_{LL}(x) - m_0(x)$ would be dominated by the term $\hat{\Gamma}_A(x)$, which is $O_p((\log(n)/(ng))^{1/2})$ and thus converges at a *slower* rate than the oracle estimator.

COROLLARY 3. *Suppose that Assumptions* 1, 4 *and* 5 *hold with* $\eta = 1/5$, $1/5 < \theta < 3/10$ *and* $p = q = 1$, *and that* $\rho(S) = 0$ *with probability* 1. *Then the following expansion holds uniformly over* $x \in I_{R,n}^-$:

$$\hat{m}_{LL}(x) - m_0(x) = \frac{1}{nf_R(x)}\sum_{i=1}^n K_h\big(r_0(S_i) - x\big)\big(\varepsilon_i - m_0'(x)\zeta_i\big) + \frac{1}{2}h^2\int u^2K(u)\,du\,m_0''(x) + o_p(n^{-2/5}).$$

In particular, we have

$$(nh)^{1/2}\left(\hat{m}_{LL}(x) - m_0(x) - \frac{1}{2}h^2\int u^2K(u)\,du\,m_0''(x)\right) \xrightarrow{d} N(0, \sigma_m^2(x)),$$

where $\sigma_m^2(x) = \mathrm{Var}(\varepsilon - m_0'(R)\zeta|R = x)\int K(t)^2\,dt/f_R(x)$ *is the asymptotic variance.*

The limiting distribution of $\hat{m}_{LL}(x)$ is again affected by the use of generated covariates under the conditions of the corollary. In this particular case, the form of the asymptotic variance has an intuitive interpretation: the estimator $\hat{m}_{LL}(x)$ has the same limiting distribution as the local linear oracle estimator in the hypothetical regression model

$$Y = m_0(r_0(S)) + \varepsilon^*,$$

where $\varepsilon^* = \varepsilon - m_0'(r_0(S))\zeta$. As in Corollary 2 above, depending on the curvature of m_0 and the covariance of ε and ζ, the asymptotic variance of the estimator using generated regressors can be bigger or smaller than that of the oracle estimator $\tilde{m}_{LL}$.

The next corollary discusses the case when $\theta < \eta$. For such a choice of bandwidth, applying Theorem 1 requires no restrictions on the dimensionality of S. It turns out that in this case $\hat{m}_{LL}(x) = \tilde{m}_{LL}(x) + o_p(n^{-2/5})$, and thus the limit distribution of $\hat{m}_{LL}$ is the same as for the oracle estimator $\tilde{m}_{LL}$. The effect exerted by the presence of nonparametrically generated regressors is thus first-order asymptotically negligible for conducting inference on m_0 in this case.

COROLLARY 4. *Suppose that Assumptions 1, 4 and 5 hold with $\theta < \eta = 1/5$. Then the following expansion holds uniformly over $x \in I_{R,n}^-$ if $\frac{2}{5}(q+1)^{-1} < \theta < \frac{3}{10}p^{-1}$:*

$$\begin{aligned}\hat{m}_{LL}(x) - m_0(x) = {} & \frac{1}{nf_R(x)} \sum_{i=1}^n K_h\big(r_0(S_i) - x\big)\varepsilon_i \\ & + \frac{1}{2}h^2 \int u^2 K(u)\, du\, m_0''(x) + o_p(n^{-2/5}).\end{aligned}$$

In particular, we have

$$(nh)^{1/2}\left(\hat{m}_{LL}(x) - m_0(x) - \frac{1}{2}h^2 \int u^2 K(u)\, du\, m_0''(x)\right) \xrightarrow{d} N(0, \sigma_m^2(x)),$$

where $\sigma_m^2(x) = \operatorname{Var}(\varepsilon|R = x) \int K(t)^2\, dt / f_R(x)$ is the asymptotic variance.

5.2. *Nonparametric censored regression.* Consider estimation of the censored regression model in (3.1). Let $\hat{r}(x)$ be the qth order local polynomial estimator of the conditional mean $r_0(x) = \mathbb{E}(Y|X = x)$, and let $\hat{q}(r)$ be the local linear estimator of $q_0(r)$ using the generated covariates $\hat{r}(X_i)$. Then an estimate of μ_0 is given by

$$\hat{\mu}(x) = \lambda + \int_{\hat{r}(x)}^{\lambda} \frac{1}{\hat{q}(u)}\, du, \tag{5.3}$$

where the constant λ is chosen large enough to satisfy $\lambda > \max_{i=1,\dots,n} \hat{r}(X_i)$ with probability tending to one. Generalizing Lewbel and Linton (2002), we consider the use of higher-order local polynomials for the first stage estimator, and allow the bandwidth used for the computation of $\hat{r}$ and $\hat{q}$ to be different. For presenting the asymptotic properties of $\hat{\mu}$, let $s_0(x) = \mathbb{E}(\mathbb{I}\{Y > 0\}|X = x)$ be the proportion of uncensored observations conditional on $X = x$, and assume that this function is continuously differentiable and bounded away from zero on the support of X. We then obtain the following result.

COROLLARY 5. *Suppose that Assumptions 1 and 5 hold with* $(Y, S, T) = (\mathbb{I}\{Y > 0\}, X, Y)$ *and* $R = r_0(S) = r_0(X)$. *Furthermore, suppose that* $\theta \in (\underline{\theta}, \bar{\theta})$ *where* $\underline{\theta}$ *and* $\bar{\theta}$ *are constants depending on* η, q *and* p *as follows*:

$$\bar{\theta} = \frac{1 - 3\eta}{p} \quad \text{and} \quad \underline{\theta} = \max\left\{\frac{1 - 4\eta}{p}, \frac{1}{2(q + 1) + p}\right\}.$$

Under these conditions, we have that

$$\sqrt{ng^p}\big(\hat{\mu}(x) - \mu_0(x)\big) \xrightarrow{d} N\left(0, \frac{\sigma_r^2(x)}{f_S(x)s_0^2(x)} \int L(t)^2\,dt\right),$$

where $\sigma_r^2(x) = \mathrm{Var}(Y|X = x)$.

The corollary is analogous to Theorem 5 in Lewbel and Linton (2002). However, using our results, substantially simplifies the proof and provides insights on admissible choices of bandwidths. Note that the lower bound $\underline{\theta}$ is chosen such that both the bias of $\hat{r}$ and $\hat{q}$ tends to zero at a rate faster than $(ng^p)^{-1/2}$. Due to this undersmoothing, the limiting distribution of $\hat{\mu} - \mu$ is centered at zero. Note that the final estimator converges at the same rate as the generated regressors. This is due to the fact that the function $\hat{r}$ is not only used to compute $\hat{q}$, but also determines the limits of integration in (5.3). The "direct" influence of the generated regressors in the estimation of q is asymptotically negligible in this particular application.

5.3. *Nonparametric triangular simultaneous equation models*. Now consider nonparametric estimation of the structural function μ_1 in the triangular simultaneous equation model (3.3)–(3.4) using a marginal integration estimator. In order to keep the notation simple, we restrict our attention to the arguably most relevant case with a single endogenous regressor, but allow for an arbitrary number of exogenous regressors and instruments. Let $\hat{\mu}_2(z)$ be the qth order local polynomial estimator of $\mu_2(z) = \mathbb{E}(X_1|Z = z)$, and let $\hat{m}(x_1, z_1, v)$ be the local linear estimator of $m(x_1, z_1, v) = \mathbb{E}(Y|X_1 = x_1, Z_1 = z_1, V = v)$. The latter is computed using the generated covariates $\hat{V}_i = X_{1i} - \hat{\mu}_2(Z_i)$ instead of the true residuals V_i from equation (3.4). For simplicity, we use the same bandwidth for all components of $\hat{m}$; that is, we put $\eta_j \equiv \eta$ for all $j = 1, \ldots, (2 + d_1)$. The marginal integration estimator of $\mu_1(x_1, z_1)$ is then given by the following sample version of (3.5):

$$\hat{\mu}_1(x_1, z_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(x_1, z_1, \hat{V}_i). \tag{5.4}$$

The following result establishes the estimator's asymptotic normality.

COROLLARY 6. *Suppose that Assumption 1 holds with* $(Y, S, T) = (Y, (X_1, Z_1, Z_2), X_1)$ *and* $R = r_0(S) = (X_1, Z_1, X_1 - \mu_2(Z_1, Z_2))$, *and that Assumption 5 holds with* $r_0(S) = \mu_2(Z_1, Z_2)$. *Furthermore, suppose that* $\eta \in (\max\{1/(5+d_1), 1/(2p+3)\}, 1/(1+d_1))$, *and that* $\theta \in (\underline{\theta}, \bar{\theta})$, *where* $\underline{\theta}$ *and* $\bar{\theta}$ *are constants depending on* η, q *and* $d_j = \dim(Z_j)$ *as follows*:

$$\bar{\theta} = \frac{1-3\eta}{2p} \quad \text{and} \quad \underline{\theta} = \frac{1-\eta(d_1+1)}{2(q+1)},$$

where $p = d_1 + d_2$. *Under these conditions, we have that*

$$\sqrt{nh^{1+d_1}}(\hat{\mu}_1(x_1, z_1) - \mu_1(x_1, z_1)) \xrightarrow{d} N\left(0, \mathbb{E}\left(\frac{\sigma_\varepsilon^2(x_1, z_1, V)}{f_{XZ|V}(x_1, z_1, V)}\right) \int \tilde{K}(t)^2\, dt\right),$$

where $\tilde{K}(t) = \prod_{i=1}^{1+d_1} \mathcal{K}(t_i)$ *is a* $(1+d_1)$*-dimensional product kernel, and* $\sigma_\varepsilon^2(x_1, z_1, v) = \operatorname{Var}(Y - m(R)|R = (x_1, z_1, v))$.

Under the conditions of the corollary, the asymptotic variance of $\hat{\mu}_1(x_1, z_1)$ is not influenced by the presence of generated regressors: If $\hat{m}$ was replaced in (5.4) with an oracle estimator $\tilde{m}$ using the actual disturbances V_i instead of the reconstructed ones, the result would not change. Also, note that the exclusion restrictions on the instruments imply that $\mathbb{E}(Y|X_1, Z_1, V) = \mathbb{E}(Y|X_1, Z_1, Z_2)$. Therefore Assumption 4 is automatically satisfied, and the adjustment term $\hat{\Gamma}(x)$ from Theorem 1 is equal to zero and does not have to be considered for the proof.

6. Conclusions. In this paper, we analyze the properties of nonparametric estimators of a regression function, when some the covariates are not directly observable, but have been estimated by a nonparametric first-stage procedure. We derive a stochastic expansion showing that the presence of generated regressors affects the limit behavior of the estimator only through a smoothed version of the first-stage estimation error. We apply our results to a number of practically relevant statistical applications.

APPENDIX: PROOFS

Throughout the Appendix, C and c denote generic constants chosen sufficiently large or sufficiently small, respectively, which may have different values at each appearance. Furthermore, define $\bar{\mathcal{M}}_n = \bar{\mathcal{M}}_{n,1} \times \cdots \times \bar{\mathcal{M}}_{n,d}$.

A.1. Proof of Theorem 1. In order to prove the statement of the theorem, we have to introduce some notation. Throughout the proof of this and the following statements, we denote the unit vector $(1, 0, \ldots, 0)^T$ in $\mathbb{R}^{p+1}$ by e_1. We also write $w_i(x, r) = (1, (r_1(S_i) - x_1)/h_1, \ldots, (r_d(S_i) - x_d)/h_d)$, and put $w_i(x) = w_i(x, r_0)$, $\hat{w}_i(x) = w_i(x, \hat{r})$ and $\tilde{w}_i(x) = w_i(x, \tilde{r})$. We also define $M_h(x, r) = n^{-1}\sum_{i=1}^n w_i(x, r) w_i(x, r)^T K_h(r(S_i) - x)$, and put $M_h(x) = M_h(x, r_0)$, $\hat{M}_h(x) = M_h(x, \hat{r})$ and $\tilde{M}_h(x) = M_h(x, \tilde{r})$ and set $N_h(x) = \mathbb{E}(M_h(x, r_0))$. Furthermore, define $\varepsilon^* = \varepsilon - \rho(S)$ and note that we have $\mathbb{E}(\varepsilon^*|S) = 0$ by construction. It also holds that

$$Y_i = m_0(r_0(S_i)) + \varepsilon_i^* + \rho(S_i).$$

Next, it follows from standard calculations that the real estimator $\hat{m}_{LL}$ can be written as

$$\hat{m}_{LL}(x) = m_0(x) + \hat{m}_{LL,A}(x) + \hat{m}_{LL,B}(x) + \hat{m}_{LL,C}(x) + \hat{m}_{LL,D}(x) + \hat{m}_{LL,E}(x),$$

where $\hat{m}_{LL,j}(x) = \hat{\alpha}_j$ for $j \in \{A, B, C, D, E\}$, and

$$(\hat{\alpha}_A, \hat{\beta}_A) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^n \bigl(\varepsilon_i^* - \alpha - \beta^T \bigl(\hat{r}(S_i) - x\bigr)\bigr)^2 K_h\bigl(\hat{r}(S_i) - x\bigr),$$

$$\begin{aligned}(\hat{\alpha}_B, \hat{\beta}_B) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^n \bigl(&m_0(r_0(S_i)) - m_0(x) - m_0'(x)^T \bigl(r_0(S_i) - x\bigr) \\ &- \alpha - \beta^T \bigl(\hat{r}(S_i) - x\bigr)\bigr)^2 \\ &\times K_h\bigl(\hat{r}(S_i) - x\bigr),\end{aligned}$$

$$\begin{aligned}(\hat{\alpha}_C, \hat{\beta}_C) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^n &\bigl(-m_0'(x)^T \bigl(\hat{r}(S_i) - r_0(S_i)\bigr) - \alpha - \beta^T \bigl(\hat{r}(S_i) - x\bigr)\bigr)^2 \\ &\times K_h\bigl(\hat{r}(S_i) - x\bigr),\end{aligned}$$

$$\begin{aligned}(\hat{\alpha}_D, \hat{\beta}_D) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^n &\bigl(m_0'(x)^T \bigl(\hat{r}(S_i) - x\bigr) - \alpha - \beta^T \bigl(\hat{r}(S_i) - x\bigr)\bigr)^2 \\ &\times K_h\bigl(\hat{r}(S_i) - x\bigr),\end{aligned}$$

$$(\hat{\alpha}_E, \hat{\beta}_E) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^n \bigl(\rho(S_i) - \alpha - \beta^{\mathsf{T}} \bigl(\hat{r}(S_i) - x\bigr)\bigr)^2 K_h\bigl(\hat{r}(S_i) - x\bigr).$$

Similarly, the oracle estimator $\tilde{m}_{LL}$ can be represented as

$$\tilde{m}_{LL}(x) = m_0(x) + \tilde{m}_{LL,A}(x) + \tilde{m}_{LL,B}(x) + \tilde{m}_{LL,C}(x) + \tilde{m}_{LL,D}(x) + \tilde{m}_{LL,E}(x),$$

where $\tilde{m}_{LL,j}(x) = \tilde{\alpha}_j$ for $j \in \{A, B, C, D, E\}$, and

$$(\tilde{\alpha}_A, \tilde{\beta}_A) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^n \bigl(\varepsilon_i - \alpha - \beta^T \bigl(r_0(S_i) - x\bigr)\bigr)^2 K_h\bigl(r_0(S_i) - x\bigr),$$

$$\begin{aligned}(\tilde{\alpha}_B, \tilde{\beta}_B) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^n \bigl(&m_0(r_0(S_i)) - m_0(x) - m_0'(x)^T \bigl(r_0(S_i) - x\bigr) \\ &- \alpha - \beta^T \bigl(r_0(S_i) - x\bigr)\bigr)^2 \\ &\times K_h\bigl(r_0(S_i) - x\bigr),\end{aligned}$$

$$(\tilde{\alpha}_C, \tilde{\beta}_C) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \bigl(-m_0'(x)^T \bigl(\hat{r}(S_i) - r_0(S_i)\bigr) - \alpha - \beta^T \bigl(r_0(S_i) - x\bigr)\bigr)^2 \times K_h\bigl(r_0(S_i) - x\bigr)$$

$$(\tilde{\alpha}_D, \tilde{\beta}_D) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \bigl(m_0'(x)^T \bigl(r_0(S_i) - x\bigr) - \alpha - \beta^T \bigl(r_0(S_i) - x\bigr)\bigr)^2 \times K_h\bigl(r_0(S_i) - x\bigr).$$

$$(\tilde{\alpha}_E, \tilde{\beta}_E) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \bigl(\rho(S_i) - \alpha - \beta^{\top}\bigl(r(S_i) - x\bigr)\bigr)^2 K_h\bigl(r(S_i) - x\bigr).$$

Note that by construction,

$$\hat{m}_{LL,D}(x) \equiv \tilde{m}_{LL,D}(x) \equiv 0. \tag{A.1}$$

We now argue that

$$\sup_{x \in I_R} |\hat{m}_{LL,A}(x) - \tilde{m}_{LL,A}(x)| = O_p(n^{-\kappa_1}). \tag{A.2}$$

For a proof of (A.2) note that $\hat{m}_{LL,A}(x)$ and $\tilde{m}_{LL,A}(x)$ are given by the first elements of the vectors $\hat{M}(x)^{-1} n^{-1} \sum_{i=1}^n K_h(\hat{r}(S_i) - x)\varepsilon_i \hat{w}_i(x)$ and $M(x)^{-1} \times n^{-1} \sum_{i=1}^n K_h(r_0(S_i) - x)\varepsilon_i \tilde{w}_i(x)$, respectively. Using these representations, one sees that (A.2) follows from Lemmas 1 and 2 below.

As a second step, we now show that

$$\sup_{x \in I_R} |\hat{m}_{LL,E}(x) - \tilde{m}_{LL,E}(x) - \hat{\Gamma}(x)| = O_p(n^{-\kappa_1} + n^{-\kappa_2} + n^{-\kappa_3}). \tag{A.3}$$

To prove (A.3), put $\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^n K_h(\hat{r}(S_i) - x)\hat{w}_i(x)\rho(S_i)$ and $\mu(x) = \frac{1}{n} \sum_{i=1}^n K_h(r_0(S_i) - x) w_i(x)\rho(S_i)$, and write $G(x) = e_1^{\top}(N_h(x))^{-1}\mathbb{E}(\hat{\mu}(x) - \mu(x))$. With this notation, $\hat{m}_{LL,E}(x) = e_1^{\top}\hat{M}_h(x)^{-1}\hat{\mu}(x)$ and $\tilde{m}_{LL,E}(x) = e_1^{\top} \times M_h(x)^{-1}\mu(x)$. Using Lemma 4 and some results of Lemma 3, we then find that

$$\begin{aligned}
&\hat{m}_{LL,E}(x) - \tilde{m}_{LL,E}(x) - G(x) \\
&\quad = e_1^{\top}\bigl(\hat{M}_h(x)^{-1}\hat{\mu}(x) - M_h(x)^{-1}\mu(x) - \mathbb{E}(M_h(x))^{-1}\mathbb{E}\bigl(\hat{\mu}(x) - \mu(x)\bigr)\bigr) \\
&\quad = O_P\bigl(n^{-((1/2)(1-\eta_+)+(\delta-\eta)_{\min})} + n^{-((1/2)(1-\eta_+)+\delta_{\min})} + n^{-\kappa_1}\bigr) = O_P(n^{-\kappa_1})
\end{aligned}$$

uniformly over $x \in I_R$. Using standard smoothing arguments, we also get that

$$\begin{aligned}
G(x) &= e_1^{\top} N_h(x)^{-1}\mathbb{E}\bigl(\hat{\mu}(x) - \mu(x)\bigr) \\
&= \frac{1}{f_R(x)} \int \bigl(K_h\bigl(\hat{r}(u) - x\bigr) - K_h\bigl(r_0(u) - x\bigr)\bigr)\rho(u) f_S(u)\, dx\, du \\
&\quad + O_P\bigl(n^{-2\eta_{\min} - (\delta-\eta)_{\min}}\bigr)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{f_R(x)} \int K_h'\big(r_0(u) - x\big)\big(\hat{r}(u) - r_0(u)\big)\rho(u) f_S(u)\,dx\,du \\
&\quad + O_P\big(n^{-\delta_{\min} - (\delta-\eta)_{\min}}\big) + O_P(n^{-\kappa_2}) \\
&= \hat{\Gamma}(x) + O_P(n^{-\kappa_2}) + O_P(n^{-\kappa_3})
\end{aligned}
$$

uniformly over $x \in I_R$. This shows the claim in (A.3).

Finally, from Lemmas 2 and 3 we get that

$$\sup_{x \in I_R} |\hat{m}_{LL,B}(x) - \tilde{m}_{LL,B}(x)| = O_p(n^{-\kappa_2}), \tag{A.4}$$

$$\sup_{x \in I_R} |\hat{m}_{LL,C}(x) - \tilde{m}_{LL,C}(x)| = O_p(n^{-\kappa_3}), \tag{A.5}$$

and it is easy to see that

$$\sup_{x \in I_R} |\tilde{m}_{LL,C}(x) - m_0'(x)\hat{\Delta}(x)| = O_p(n^{-\kappa}). \tag{A.6}$$

Taken together, the results in (A.1)–(A.6) imply the statement of the theorem.

LEMMA 1. *Suppose that the conditions of Theorem 1 hold. Then*

$$
\begin{aligned}
&\sup_{x \in I_R, r_1, r_2 \in \bar{\mathcal{M}}_n} \left| \frac{1}{n}\sum_{i=1}^n K_h\big(r_1(S_i) - x\big)\varepsilon_i - \frac{1}{n}\sum_{i=1}^n K_h\big(r_2(S_i) - x\big)\varepsilon_i \right| \\
&\quad = O_p(n^{-\kappa_1}), \\
&\sup_{x \in I_R, r_1, r_2 \in \bar{\mathcal{M}}_n} \left| \frac{1}{n}\sum_{i=1}^n K_h\big(r_1(S_i) - x\big)\frac{r_{1,j}(S_i) - x_j}{h_j}\varepsilon_i \right. \\
&\qquad\qquad \left. - \frac{1}{n}\sum_{i=1}^n K_h\big(r_2(S_i) - x\big)\frac{r_{2,j}(S_i) - x_j}{h_j}\varepsilon_i \right| \\
&\quad = O_p(n^{-\kappa_1}).
\end{aligned}
$$

PROOF. We only prove the first statement of the lemma. The second claim can be shown using essentially the same arguments. Without loss of generality, we also assume that

$$\kappa_1 > (\delta - \eta)_{\min}. \tag{A.7}$$

If $\kappa_1 \le (\delta - \eta)_{\min}$ the statement of the lemma follows from a direct bound. For $C_1, C_2 > 0$ large enough (see below) we choose C_ε such that

$$\Pr\Big(\max_i |\varepsilon_i| > C_\varepsilon \log(n)\Big) \le n^{-C_1}, \tag{A.8}$$

$$\big|\mathbb{E}\varepsilon_i \mathbb{I}\{|\varepsilon_i| \le C_\varepsilon \log(n)\}\big| \le n^{-C_2}. \tag{A.9}$$

With this choice of C_ε we define

$$\Delta_i(r_1, r_2) = \big(K_h\big(r_1(S_i) - x\big) - K_h\big(r_2(S_i) - x\big)\big)\varepsilon_i^*$$

with

$$\varepsilon_i^* = \varepsilon_i \mathbb{I}\{|\varepsilon_i| \le C_{\varepsilon_i}\log(n)\} - \mathbb{E}\big(\varepsilon_i \mathbb{I}\{|\varepsilon_i| \le C\log(n)\}\big).$$

For the proof of the lemma we apply a chaining argument; compare, for example, the proof of Theorem 9.1 in van de Geer (2000). Now for $s \ge 0$, let $\bar{\mathcal{M}}^*_{s,n,j}$ be a set of functions chosen such that for each $r \in \bar{\mathcal{M}}_{n,j}$ there exists $r^* \in \bar{\mathcal{M}}^*_{s,n,j}$ such that $\|r - r^*\|_\infty \le 2^{-s}n^{-\delta_j}$. That is, the functions in $\bar{\mathcal{M}}^*_{s,n,j}$ are the midpoints of a $(2^{-s}n^{-\delta_j})$-covering of $\bar{\mathcal{M}}_{n,j}$. By Assumption 3, the set $\bar{\mathcal{M}}^*_{s,n,j}$ can be chosen such that its cardinality $\#\bar{\mathcal{M}}^*_{s,n,j}$ is at most $C\exp((2^{-s}n^{-\delta_j})^{-\alpha_j}n^{\xi_j})$. Furthermore, define $\bar{\mathcal{M}}^*_{s,n} = \bar{\mathcal{M}}^*_{s,n,1} \times \cdots \times \bar{\mathcal{M}}^*_{s,n,d}$.

For $r_1, r_2 \in \bar{\mathcal{M}}_n$ we now choose $r_1^s, r_2^s \in \bar{\mathcal{M}}^*_{s,n}$ such that $\|r^s_{1,j} - r_{1,j}\|_\infty \le 2^{-s}n^{-\delta_j}$ and $\|r^s_{2,j} - r_{2,j}\|_\infty \le C2^{-s}n^{-\delta_j}$, for all j. We then consider the chain

$$\begin{aligned}\Delta_i(r_1, r_2) = \Delta_i(r_1^0, r_2^0) - \sum_{s=1}^{G_n}\Delta_i(r_1^{s-1}, r_1^s) + \sum_{s=1}^{G_n}\Delta_i(r_2^{s-1}, r_2^s)\\ - \Delta_i(r_1^{G_n}, r_1) + \Delta_i(r_2^{G_n}, r_2),\end{aligned}$$

where G_n is the smallest integer that satisfies $G_n > (1 + c_G)(\kappa_1 - (\delta - \eta)_{\min}) \times \log(n)/\log(2)$ for a constant $c_G > 0$. With this choice of G_n, we obtain that for $l = 1, 2$

$$T_1 = \left|\frac{1}{n}\sum_{i=1}^n \Delta_i(r_l^{G_n}, r_l)\right| \le C\log(n)2^{-G_n}n^{-(\delta-\eta)_{\min}} \le Cn^{-\kappa_1}. \tag{A.10}$$

Now for any $a > c_G$ define the constant $c_a = (\sum_{s=1}^\infty 2^{-as})^{-1}$. It then follows that

$$\begin{aligned}&\Pr\left(\sup_{r_1\in\bar{\mathcal{M}}_n}\left|\frac{1}{n}\sum_{i=1}^n\sum_{s=1}^{G_n}\Delta_i(r_1^{s-1}, r_1^s)\right| > n^{-\kappa_1}\right)\\
&\quad\le \sum_{s=1}^{G_n}\Pr\left(\sup_{r_1\in\bar{\mathcal{M}}_n}\left|\frac{1}{n}\sum_{i=1}^n\Delta_i(r_1^{s-1}, r_1^s)\right| > c_a2^{-as}n^{-\kappa_1}\right)\\
&\quad\le \sum_{s=1}^{G_n}\#\bar{\mathcal{M}}^*_{s-1,n}\#\bar{\mathcal{M}}^*_{s,n}\Pr\left(\frac{1}{n}\sum_{i=1}^n\Delta_i(r_1^{*,s}, r_1^{**,s}) > c_a2^{-as}n^{-\kappa_1}\right)\\
&\qquad + \sum_{s=1}^{G_n}\#\bar{\mathcal{M}}^*_{s-1,n}\#\bar{\mathcal{M}}^*_{s,n}\Pr\left(\frac{1}{n}\sum_{i=1}^n\Delta_i(\tilde{r}_1^{*,s}, \tilde{r}_1^{**,s}) < c_a2^{-as}n^{-\kappa_1}\right)\\
&\quad= T_2 + T_3,\end{aligned}$$

where the functions $r_1^{*,s}, \tilde{r}_1^{*,s} \in \bar{\mathcal{M}}_{s-1,n}^*$ and $r_1^{**,s}, \tilde{r}_1^{**,s} \in \bar{\mathcal{M}}_{s,n}^*$ are chosen such that

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n \Delta_i(r_1^{*,s}, r_1^{**,s}) > c_a 2^{-as} n^{-\kappa_1}\right)$$
$$= \max_{r_1^{s-1}, r_1^s} \Pr\left(\frac{1}{n}\sum_{i=1}^n \Delta_i(r_1^{s-1}, r_1^s) > c_a 2^{-as} n^{-\kappa_1}\right),$$
$$\Pr\left(\frac{1}{n}\sum_{i=1}^n \Delta_i(\tilde{r}_1^{*,s}, \tilde{r}_1^{**,s}) < c_a 2^{-as} n^{-\kappa_1}\right)$$
$$= \max_{r_1^{s-1}, r_1^s} \Pr\left(\frac{1}{n}\sum_{i=1}^n \Delta_i(r_1^{s-1}, r_1^s) > c_a 2^{-as} n^{-\kappa_1}\right).$$

We now show that both T_2 and T_3 tend to zero at an exponential rate:

$$T_2 \le \exp(-cn^c), \tag{A.11}$$

$$T_3 \le \exp(-cn^c). \tag{A.12}$$

We only show (A.11), as the statement (A.12) follows by essentially the same arguments. Using Assumption 3, we obtain by application of the Markov inequality that

$$\begin{aligned} T_2 &\le C\sum_{s=1}^{G_n}\prod_j \exp\bigl((2^{-s}n^{-\delta_j})^{-\alpha_j}n^{\xi_j}\bigr) \\ &\qquad \times \mathbb{E}\left(\exp\left(\gamma_{n,s}\frac{1}{n}\sum_{i=1}^n \Delta_i(r_1^{*,s}, r_1^{**,s}) - \gamma_{n,s}c_a 2^{-as}n^{-\kappa_1}\right)\right) \\ &\le C\sum_{s=1}^{G_n}\exp\left(\sum_j 2^{s\alpha_j}n^{\delta_j\alpha_j+\xi_j} - \gamma_{n,s}c_a 2^{-as}n^{-\kappa_1}\right) \\ &\qquad \times \prod_{i=1}^n \mathbb{E}\left(\exp\left(\gamma_{n,s}\frac{1}{n}\Delta_i(r_1^{*,s}, r_1^{**,s})\right)\right), \end{aligned} \tag{A.13}$$

where $\gamma_{n,s} = c_\gamma 2^{(2-a)s} n^{-\kappa_1+1-\eta_+ +2(\delta-\eta)_{\min}}$ with a constant $c_\gamma > 0$, small enough. Now the last term on the right-hand side of (A.13) can be bounded as follows:

$$\begin{aligned} \mathbb{E}\left(\exp\left(\gamma_{n,s}\frac{1}{n}\Delta_i(r_1^{*,s}, r_1^{**,s})\right)\right) &\le 1 + C\mathbb{E}(\gamma_{n,s}^2 n^{-2}\Delta_i^2(r_1^{*,s}, r_1^{**,s})) \\ &\le \exp(C\gamma_{n,s}^2 n^{-2} n^{\eta_+ -2(\delta-\eta)_{\min}} 2^{-2s}), \end{aligned} \tag{A.14}$$

where we have used that

$$\begin{aligned}\left|\gamma_{n,s}\frac{1}{n}\Delta_i(r_1^{*,s},r_1^{**,s})\right| &\le C\gamma_{n,s}\frac{1}{n}\log(n)n^{\eta_+}n^{-(\delta-\eta)_{\min}}2^{-s}\\ &\le C\log(n)n^{(\delta-\eta)_{\min}-\kappa_1}2^{-as+s}\\ &\le C\log(n)n^{(c_G-a)(\kappa_1-(\delta-\eta)_{\min})}\\ &\le C\end{aligned}$$

for n large enough because of (A.7). Inserting (A.14) into (A.13), we obtain, if a and c_γ were chosen sufficiently small, that

$$\begin{aligned}T_2 &\le C\sum_{s=1}^{G_n}\exp\biggl(\sum_j 2^{s\alpha_j}n^{\delta_j\alpha_j+\xi_j}-c2^{2(1-a)s}n^{1-2\kappa_1-\eta_++2(\delta-\eta)_{\min}}\biggr)\\ &\le C\sum_{s=1}^{G_n}\exp(-c^s n^c)\\ &\le \exp(-cn^c).\end{aligned}$$

Finally, it follows from a simple argument that

$$T_4=\Pr\biggl(\sup_{r_1,r_2\in\bar{\mathcal{M}}_n}\left|\frac{1}{n}\sum_{i=1}^n\Delta_i(r_1^0,r_2^0)\right|>n^{-\kappa_1}\biggr)\le\exp(-cn^c) \tag{A.15}$$

because the set $\bar{\mathcal{M}}_{0,n}^*$ can always be chosen such that it contains only a single element.

From (A.10), (A.11), (A.12) and (A.15), we thus obtain that

$$\begin{aligned}\sup_{x\in I_R}\Pr\biggl(\sup_{r_1,r_2\in\bar{\mathcal{M}}_n}&\left|\frac{1}{n}\sum_{i=1}^n K_h\bigl(r_1(S_i)-x\bigr)\varepsilon_i^*\right.\\ &\left.-\frac{1}{n}\sum_{i=1}^n K_h\bigl(r_2(S_i)-x\bigr)\varepsilon_i^*\right|>Cn^{-\kappa_1}\biggr)\le\exp(-cn^c).\end{aligned} \tag{A.16}$$

Now for $C_I>0$ choose a grid $I_{R,n}$ of I_R with $O(n^{C_I})$ points, such that for each $x\in I_R$ there exists a grid point $x^*=x^*(x)\in I_{R,n}$ such that $\|x-x^*\|\le n^{-cC_I}$. If C_I is chosen large enough, this implies that

$$\sup_{x\in I_R}\sup_{r\in\bar{\mathcal{M}}_n}\left|\frac{1}{n}\sum_{i=1}^n K_h\bigl(r(S_i)-x\bigr)\varepsilon_i-\frac{1}{n}\sum_{i=1}^n K_h\bigl(r(S_i)-x^*\bigr)\varepsilon_i\right|\le n^{-\kappa_1} \tag{A.17}$$

for large enough n, with probability tending to one. Furthermore, it follows from (A.16) that

$$\sup_{x\in I_{R,n}}\sup_{r_1,r_2\in\bar{\mathcal{M}}_n}\left|\frac{1}{n}\sum_{i=1}^n K_h\bigl(r_1(S_i)-x\bigr)\varepsilon_i-\frac{1}{n}\sum_{i=1}^n K_h\bigl(r_2(S_i)-x\bigr)\varepsilon_i\right|\le n^{-\kappa_1}. \tag{A.18}$$

The statement of the lemma then follows from (A.8)–(A.9) and (A.17)–(A.18), if the constants C_1 and C_2 were chosen large enough. $\square$

LEMMA 2. *Suppose that the conditions of Theorem 1 hold. Then*

$$\begin{aligned}\sup_{x\in I_R, r_1, r_2\in\bar{\mathcal{M}}_n}&\left|\frac{1}{n}\sum_{i=1}^n K_h\big(r_1(S_i)-x\big)\left(\frac{r_{1,j}(S_i)-x_j}{h_j}\right)^a\left(\frac{r_{1,l}(S_i)-x_l}{h_l}\right)^b\right.\\&\qquad\left.-\frac{1}{n}\sum_{i=1}^n K_h\big(r_2(S_i)-x\big)\left(\frac{r_{2,j}(S_i)-x_j}{h_j}\right)^a\left(\frac{r_{2,l}(S_i)-x_l}{h_l}\right)^b\right|\\&=O_p\big(n^{-(\delta-\eta)_{\min}}\big)\end{aligned}$$

for $j,l=1,\dots,q$ $j\neq l$ *and* $0\le a+b\le 2$, $0\le a,b$.

PROOF. The lemma follows from

$$\sup_{x,s}\big|K_h\big(r_1(s)-x\big)-K_h\big(r_2(s)-x\big)\big|\le Cn^{-(\delta-\eta)_{\min}+\eta_+}$$

for $r_1, r_2\in\bar{\mathcal{M}}_n$ and from

$$\begin{aligned}\sup_{x\in I_R, r\in\bar{\mathcal{M}}}&\left|\frac{1}{n}\sum_{i=1}^n K_h\big(r(S_i)-x\big)\right|\\&\le Cn^{-1+\eta_+}\sup_{x\in I_R}\#\{i:|r_{0,j}(S_i)-x_j|\le Cn^{-\eta_j}\text{ for } j=1,\dots,d\}\\&=O_p(1),\end{aligned}$$

which follows from a simple calculation. $\square$

LEMMA 3. *Suppose that the assumptions of Theorem 1 hold. For a random variable $R_n=O_p(1)$ that neither depends on x nor i, it holds that*

$$\begin{aligned}\sup_{x\in I_R, 1\le i\le n}&\big|\big[m_0(r_0(S_i))-m_0(x)-m_0'(x)^T\big(r_0(S_i)-x\big)\big]I_i(x)\big|\\&\le R_n n^{-2\eta_{\min}},\end{aligned}\tag{A.19}$$

$$\begin{aligned}\sup_{x\in I_R}&\left\|\frac{1}{n}\sum_{i=1}^n K_h\big(\hat{r}(S_i)-x\big)\hat{w}_i(x)\hat{w}_i(x)^T\right.\\&\qquad\left.-\frac{1}{n}\sum_{i=1}^n K_h\big(r_0(S_i)-x\big)\tilde{w}_i(x)\tilde{w}_i(x)^T\right\|\\&\le R_n n^{-(\delta-\eta)_{\min}},\end{aligned}\tag{A.20}$$

$$\sup_{x\in I_R}\left\|\frac{1}{n}\sum_{i=1}^{n}K_h\big(r_0(S_i)-x\big)\tilde{w}_i(x)\tilde{w}_i(x)^T - f_R(x)B_K\right\| \\ \leq R_n\big(n^{-\eta_{\min}}+n^{-(1-\eta_+)/2}\sqrt{\log n}\big), \tag{A.21}$$

where $I_i(x)=\mathbb{I}\{\|(\hat{r}(S_i)-x)/h\|_1\leq 1\}$ *is an equals one if* $\hat{r}(S_i)-x$ *lies in the support of the kernel function* K_h *and zero otherwise, and* $B_K=\mathrm{diag}(1,\int u^2K(u)\,du,\dots,\int u^2K(u)\,du)$ *is a* $(d+1)\times(d+1)$ *diagonal matrix.*

PROOF. Claim (A.19) follows by a simple calculation. Claim (A.20) is a direct consequence of Lemma 2, and (A.21) follows from standard arguments from kernel smoothing theory. For the stochastic part, one makes use of Lemma 5, given in Appendix A.7, below. □

LEMMA 4. *Suppose that the assumptions of Theorem 1 hold. Then it holds that*

$$\sup_{x\in I_R, r_1,r_2\in\bar{\mathcal{M}}}\|\mu(x,r_1)-\mu(x,r_2)-\mathbb{E}[\mu(x,r_1)-\mu(x,r_2)]\|=O_p(n^{-\kappa_1}), \tag{A.22}$$

$$\sup_{x\in I_R}|\hat{\mu}(x)|=O_p\big(\sqrt{\log n}n^{-(1-\eta_+)/2}\big), \tag{A.23}$$

where

$$\hat{\mu}(x)=n^{-1}\sum_{i=1}^{n}K_h\big(\hat{r}(S_i)-x\big)\hat{w}_i(x)\rho(S_i)$$

and

$$\mu(x)=n^{-1}\sum_{i=1}^{n}K_h\big(r_0(S_i)-x\big)w_i(x)\rho(S_i).$$

PROOF. For a proof of (A.22) one proceeds as in Lemma 1. Claim (A.23) follows by classical smoothing arguments. Note that we have that $\mathbb{E}(\hat{\mu}(x,r_0))=0$. □

A.2. Proof of Proposition 1. In order to prove Proposition 1, we use the fact that the local polynomial estimator satisfies a certain uniform stochastic expansion if Assumption 4 holds. In order to present this result, we first have to introduce a substantial amount of further notation. For simplicity we assume $g_1=\cdots=g_p$, and we write g for this joint value and for the vector $g=(g,\dots,g)$.

Let $N_i=\binom{i+q-1}{q-1}$ be the number of distinct q-tuples u with $u_+=i$. Arrange these q-tuples as a sequence in a lexicographical order (with the highest priority given to the last position so that $(0,\dots,0,i)$ is the first element in the sequence, and $(i, 0, \dots, 0)$ the last element). Let τ_i denote this one-to-one mapping, that is, $\tau_i(1) = (0, \dots, 0, i)$, $\dots$, $\tau_i(N_i) = (i, 0 \dots, 0)$. For each $i = 1, \dots, q$, define a $N_i \times 1$ vector $\mu_i(x)$ with its kth element given by $x^{\tau_i(k)}$, and write $\mu(x) = (1, \mu_1(x)^T, \dots, \mu_q(x)^T)^T$, which is a column vector of length $N = \sum_{i=1}^q N_i$. Let $\nu_i = \int L(u)u^i\,du$ and define $\nu_{ni}(x) = \int L(u)u^i f_S(x + gu)\,du$. For $0 \le j, k \le q$, let $M_{j,k}$ and $M_{n,j,k}(x)$ be two $N_j \times N_k$ matrices with their (l, m) elements, respectively, given by

$$[M_{j,k}]_{l,m} = \nu_{\tau_j(l)+\tau_k(m)} \quad \text{and} \quad [M_{nj,k}(x)]_{l,m} = \nu_{n,\tau_j(l)+\tau_k(m)}(x).$$

Now define the $N \times N$ matrices M_q and $M_{n,q}(x)$ by

$$M_q = \begin{pmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,q} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,q} \\ \vdots & \vdots & \ddots & \vdots \\ M_{q,0} & M_{q,1} & \cdots & M_{q,q} \end{pmatrix},$$

$$M_{n,q}(x) = \begin{pmatrix} M_{n,0,0}(x) & M_{n,0,1}(x) & \cdots & M_{n,0,q}(x) \\ M_{n,1,0}(x) & M_{n,1,1}(x) & \cdots & M_{n,1,q}(x) \\ \vdots & \vdots & \ddots & \vdots \\ M_{n,q,0}(x) & M_{n,q,1}(x) & \cdots & M_{n,q,q}(x) \end{pmatrix}.$$

Finally, denote the first unit q-vector by $e_1 = (1, 0, \dots, 0)$. With this notation, it can be shown along classical lines that the local polynomial estimator $\hat{r}$ admits the following stochastic expansion:

$$\begin{aligned} \hat{r}(s) = r_0(s) &+ \frac{1}{n}\sum_{i=1}^n e_1 M_{nq}^{-1}(s)\mu\big((S_i - s)/g\big)L_g(S_i - s)\zeta_i \\ &+ g^{q+1}B_n(s) + R_n(s), \end{aligned} \tag{A.24}$$

where $\sup_{s\in I_S} \|R_n(s)\| = O_p((\log(n)/ng^p)^{1/2})$, and B_n is a bias term that satisfies

$$B_n(s) = \frac{1}{(q+1)!} e_1 M_q^{-1} A_q r_0^{(q+1)}(s) + o_p(1) \equiv b(s) + o_p(1). \tag{A.25}$$

To prove the proposition, define the stochastic component and the bias term of the expansion (A.24) as $\hat{r}_A(s) = n^{-1}\sum_{i=1}^n e_1 M_{nq}^{-1}(s)\mu((S_i - s)/g)L_g(S_i - s)\zeta_i$ and $\hat{r}_B(s) = g^{q+1}B_n(s)$, respectively. Now the function $\hat{\Delta}$ can be written as

$$\begin{aligned} \hat{\Delta}(x) &= e_1^T N_h(x)^{-1}\mathbb{E}\big(K_h\big(r_0(S) - x\big)w(x, r)\hat{r}_A(S)\big) \\ &\quad + e_1^T N_h(x)^{-1}\mathbb{E}\big(K_h\big(r_0(S) - x\big)w(x, r)\hat{r}_B(S)\big) + O_p\left(\frac{\log(n)}{ng^p}\right) \\ &\equiv \hat{\Delta}_A(x) + \hat{\Delta}_B(x) + O_p\left(\frac{\log(n)}{ng^p}\right), \end{aligned}$$

uniformly over $x \in I_R$. We first analyze the term $\hat{\Delta}_B(x)$. Through the usual arguments from kernel smoothing theory, one can show for $x \in I_{R,n}^-$ that

$$\begin{aligned}\hat{\Delta}_B(x) &= g^{q+1} e_1^T N_h(x)^{-1} \mathbb{E}\big(K_h\big(r_0(S) - x\big) w(x, r) b(S)\big) + o_p(g^{q+1}) \\ &= g^{q+1} \mathbb{E}\big(b(S) | r_0(S) = x\big) + o_p(g^{q+1} + n^{-2\eta})\end{aligned}$$

since the function $\mathbb{E}(b(S)|r_0(S) = x)$ is continuous with respect to x because of Assumptions 5 and 6. Explicitly, we have

$$\begin{aligned}&\mathbb{E}\big(b(S) | r_0(S) = x\big) \\ &\quad = \frac{\int b(s_{-p}, \varphi(s_{-p}, x)) f_S(s_{-p}, \varphi(s_{-p}, x))\, \partial_{s_{-p}} \varphi(s_{-p}, x)\, ds_{-p}}{\int f_S(s_{-p}, \varphi(s_{-p}, x))\, \partial_{s_{-p}} \varphi(s_{-p}, x)\, ds_{-p}}.\end{aligned}$$

Next, consider the term $\hat{\Delta}_A(x)$. Note that for $x \in I_{R,n}^-$ we have that

$$\hat{\Delta}_A(x) = \frac{1}{n f_R(x)} \sum_{j=1}^{n} \psi_n(x, S_j) \zeta_j \tag{A.26}$$

with

$$\begin{aligned}\psi_n(x, s) &= \int_{I_S} \big(K_h\big(r_0(u) - x\big) e_1 \bar{M}_{nq}^{-1}(u) \mu\big((s - u)/g\big) L_g(s - u)\big) f_S(u)\, du \\ &= \int K_h\big(r_0(u) - x\big) L_{n,g}^*(s, u - s)\, du,\end{aligned}$$

where $L_{n,g}^*(s, t) = f_S(s - t) e_1 \bar{M}_{nq}^{-1}(s - t) \mu(t/g) L_g(t)$. Define $I_{S,n}^-$ as the set that contains all $s \in I_S$ that do not lie in a g-neighborhood of the boundary of I_S. Uniformly over $s \in I_{S,n}^-$, we have that $M_{n,q}(s) - f_S(s) M_q = O(g)$. Thus for $s \in I_{S,n}^-$, we have that $\psi_n(x, s) = (1 + O(g)) \psi(x, s)$ where the function ψ is equal to $\psi(x, s) = \int K_h(r_0(u) - x) L_g^*(u - s)\, du$ with modified kernel L^* defined as

$$L^*(t) = e_1 M_q^{-1} \mu(t) L(t). \tag{A.27}$$

Note that L^* is the *equivalent kernel* of the local polynomial regression estimator; see Fan and Gijbels (1996), Section 3.2.2. For $q = 0, 1$ the equivalent kernel is in fact equal to the original one, whereas $L^*(t)$ is equal to $L(t)$ times a polynomial in t of order q for $q \geq 2$, with coefficients such that its moments up to the order q are equal to zero. The kernel $L_{n,g}^*(u, t)$ has the same moment conditions in t as L_g^* but depends on u.

We now derive explicit expressions for the leading term in equation (A.26) for the cases (a)–(c) of the proposition. Starting with case (a), in which $g/h \to 0$, it follows by substitution and Taylor expansion arguments that with $K_h'(v) =$

$h^{-1}K'(h^{-1}v)$ and $K_h''(v) = h^{-1}K''(h^{-1}v)$

$$\begin{aligned}
\psi_n(x,v) &= \int K_h\bigl(r_0(s)-x\bigr)L_{n,g}^*(s,s-v)\,ds \\
&= \int K_h\bigl(r_0(v-tg)-x\bigr)L_n^*(v-tg,t)\,dt \\
&= \int \biggl(K_h\bigl(r_0(v)-x\bigr) + K_h'\bigl(r_0(v)-x\bigr)\frac{r_0(v-tg)-r_0(v)}{h} \\
&\qquad\qquad + K_h''(\chi_1-x)\frac{1}{2}\biggl(\frac{r_0(v-tg)-r_0(v)}{h}\biggr)^2\biggr) \\
&\qquad \times L_n^*(v-tg,t)\,dt \\
&= K_h\bigl(r_0(v)-x\bigr) \\
&\quad + K_h'\bigl(r_0(v)-x\bigr)\int\biggl(-\partial_s r_0(v)\frac{tg}{h} + \partial_s^2 r_0(\chi_2)\frac{t^2g^2}{2h}\biggr)L_n^*(v-tg,t)\,dt \\
&\quad - \int K_h''(\chi_1-x)\frac{1}{2}\biggl(\frac{\partial_s r_0(\chi_3)tg}{h}\biggr)^2 L_n^*(v-tg,t)\,dt,
\end{aligned}$$

where χ_1, χ_2 and χ_3 are intermediate values between $r_0(v)$ and $r_0(v-tg)$, v and $v-tg$, and v and $v-tg$, respectively. This gives an expansion for $\psi_n(x,v)$ of order $(g/h)^2$. For $v \notin I_{S,n}^-$ one gets an expansion of order g/h. Put $k_n(v) = -\partial_s r_0(v)\int tL_n^*(v-tg,t)\,dt$. Together with Lemma 5 in Appendix A.7, we thus obtain that

$$\begin{aligned}
&\frac{1}{nf_R(x)}\sum_{j=1}^n \psi_n(x,S_j)\zeta_j \\
&\quad = \frac{1}{nf_R(x)}\sum_{i=1}^n\biggl(K_h\bigl(r_0(S_i)-x\bigr) + \frac{g}{h}K_h'\bigl(r_0(S_i)-x\bigr)k_n(S_i)\biggr)\zeta_i \\
&\qquad + O_p\biggl(\biggl(\frac{g}{h}\biggr)^2\biggl(\frac{\log(n)}{nh}\biggr)^{1/2}\biggr) \\
&\quad = \frac{1}{nf_R(x)}\sum_{i=1}^n K_h\bigl(r_0(S_i)-x\bigr)\zeta_i + O_p\biggl(\biggl(\frac{g^2}{h^2}+\sqrt{\frac{g^3}{h^2}}\biggr)\sqrt{\frac{\log(n)}{nh}}\biggr),
\end{aligned}$$

as claimed. To show statement (b) of the proposition, we rewrite the function ψ_n as follows:

$$\begin{aligned}
\psi_n(x,v) &= \int\biggl(K_h\bigl(r_0(v)-x+\partial_s r_0(v)th\bigr) + K'\biggl(\frac{\chi_1}{h}\biggr)\partial_s^2 r_0(\chi_2)\frac{1}{2}t^2\biggr) \\
&\qquad \times L_n^*(v-th,t)\,dt \\
&\quad = J_{n,h}(x,v) + h\int K_h'(\chi_1)\partial_s^2 r_0(\chi_2)\frac{1}{2}t^2 L_n^*(v-th,t)\,dt,
\end{aligned}$$

where $J_{n,h}(x,s) = \int K_h(r_0(s) - x - \partial_s r_0(s)uh) L_n^*(s - uh, u)\,du$, and χ_1 is an intermediate value between $r_0(v+gt)$ and $r_0(v) + \partial_s r_0(v)tg$, and χ_2 is an intermediate value between v and $v+gt$. As in the proof of part (a), it follows from Lemma 5 in Appendix A.7 that

$$\begin{aligned}\frac{1}{nf_R(x)}\sum_{j=1}^n \psi_n(x,S_j)\zeta_j &= \frac{1}{nf_R(x)}\sum_{j=1}^n J_{n,h}(x,S_j)\zeta_j + O_p\left(h\sqrt{\frac{\log(n)}{nh}}\right)\\ &= \frac{1}{nf_R(x)}\sum_{j=1}^n J_h(x,S_j)\zeta_j + O_p\left(\sqrt{\frac{\log(n)}{n}}\right),\end{aligned}$$

where J_h uses the location independent form of the equivalent kernel L^* as defined in the text in front of Proposition 1. This implies the desired result.

Now consider statement (c) of the proposition. In this case, where $g/h \to \infty$, we can rewrite the function ψ_n as follows:

$$\begin{aligned}\psi_n(x,v) = \int & K_h(w_p - x)\\ &\times L_{n,g}^*\big((w_{-p}, \varphi(w))^T, (w_{-p} - v_{-p}, \varphi(w) - v_p)^T\big)\,\partial_x\varphi(w)\,dw.\end{aligned}$$

From tedious but conceptually simple Taylor expansion arguments similar to the ones employed for case (a), and from Lemma 5, one gets that

$$\frac{1}{nf_R(x)}\sum_{j=1}^n \psi_n(x,S_j)\zeta_j = \frac{1}{nf_R(x)}\sum_{j=1}^n H_{n,g}(x,S_j)\zeta_j + O_p\left(\frac{h^2}{g^2}\sqrt{\frac{\log(n)}{ng}}\right),$$

where

$$\begin{aligned}H_{n,g}(x,v) = \int K(t) L_{n,g}^*\big(&(v_{-p} + gs_{-p}, G_n(v_{-p}, x; s_{-p}, t)),\\ &(s_{-p}, G_n(v_{-p}, x; s_{-p}, t) - v_p)\big)\\ \partial_x\varphi(v_{-p}, x)&\,ds_{-p}\,dt\end{aligned} \tag{A.28}$$

and $G_n(v_{-p}, x; s_{-p}, t) = \varphi(v_{-p}, x) + gs_{-p}\,\partial_{-p}\varphi(v_{-p}, x) + ht\,\partial_x\varphi(v_{-p}, x)$. With H_n^Δ as defined in the text, we find

$$\begin{aligned}\frac{1}{nf_R(x)}\sum_{j=1}^n \psi_n(x,S_j)\zeta_j &= \frac{1}{nf_R(x)}\sum_{j=1}^n H_n^\Delta(x,S_j)\zeta_j\\ &\quad + O_p\left(\left(1 + \sqrt{\frac{h}{g}}\right)\sqrt{\frac{\log(n)}{n}} + \frac{h^2}{g^2}\sqrt{\frac{\log(n)}{ng}}\right).\end{aligned}$$

Since $O(h/g) = o(1)$, this completes our proof.

A.3. Proof of Proposition 2. To show the result, note that

$$\begin{aligned}\Gamma(x,r) &= e_1^T N_h(x)^{-1}\mathbb{E}\big(\big(K_h\big(r(S)-x\big)-K_h\big(r_0(S)-x\big)\big)w(x)\rho(S)\big)\\ &\quad + O_p\big(n^{-((1/2)(1-\eta_+)+2\delta-\eta)}\big)\\ &= \mathbb{E}\big(\rho(S)|r(S)=x\big)-\mathbb{E}\big(\rho(S)|r_0(S)=x\big)\\ &\quad + O_p\big(n^{-2\eta}+n^{-((1/2)(1-\eta_+)+2\delta-\eta)}\big)\end{aligned}$$

uniformly over $x \in I_R$ and $r \in \mathcal{M}_n$. Since $\mathbb{E}(\rho(S)|r_0(S)) \equiv 0$ by construction, it suffices to consider the term $\mathbb{E}(\rho(S)|r(S)=x)$. To simplify the exposition, we strengthen Assumption 6 and suppose that in addition to r_0 all functions $r \in \mathcal{M}_n$ are strictly monotone with respect to their last argument, and write φ_r for corresponding the inverse function that satisfies $r(u_{-p}, \varphi_r(u_{-p}, x)) = x$ (without this condition, the notation would be much more involved, as we would have to consider all regions where the functions $r \in \mathcal{M}_n$ are piecewise monotone with respect to the last component separately). Using rules for integrals on manifolds, we derive the following explicit expression for $\mathbb{E}(\rho(S)|r(S)=x)$:

$$\begin{aligned}&\mathbb{E}\big(\rho(S)|r(S)=x\big)\\ &\qquad = \frac{\int \rho(s_{-p},\varphi_r(s_{-p},x))f_S(s_{-p},\varphi_r(s_{-p},x))\,\partial_{-p}\varphi_r(s_{-p},x)\,ds_{-p}}{\int f_S(s_{-p},\varphi_r(s_{-p},x))\,\partial_{-p}\varphi_r(s_{-p},x)\,ds_{-p}}.\end{aligned}$$

Set the numerator of the above expression as $\gamma_1(x,r)$ and the denominator as $\gamma_2(x,r)$. Then clearly $\gamma_2(x,\hat{r}) = f_R(x) + o_p(1)$ uniformly over $x \in I_R$. Moreover, note that the mapping

$$r \mapsto \rho(s_{-p},\varphi_r(s_{-p},x))f_S(s_{-p},\varphi_r(s_{-p},x))$$

is Hadamard differentiable at r_0, with derivative

$$r \mapsto \frac{\partial_p\lambda(s_{-p},\varphi(s_{-p},x))}{\partial_p r_0(s_{-p},\varphi(s_{-p},x))} r(s_{-p},\varphi(s_{-p},x)).$$

It follows with $\gamma_1(x,r_0)=0$ that

$$\begin{aligned}\gamma_1(x,r) = &\int \frac{\partial_p\lambda(s_{-p},\varphi(s_{-p},x))}{\partial_p r_0(s_{-p},\varphi(s_{-p},x))}\big(r(s_{-p},\varphi(s_{-p},x))-r_0(s_{-p},\varphi(s_{-p},x))\big)\\ &\quad\times(\partial_{-p}\varphi_r(s_{-p},x))\,ds_{-p}\\ &+O_p(\|r-r_0\|_\infty^2).\end{aligned}$$

We evaluate the term $\gamma_1(x,\hat{r})$, substitute the uniform expansion (A.24) for $\hat{r}(s) - r_0(s)$ into the explicit expression derived above, and use standard arguments from kernel smoothing theory. This gives the desired expansion for $\hat{\Gamma}_A$. The form of $\hat{\Gamma}_B$ follows from the same arguments used to derive the form of $\hat{\Delta}_B$ in the proof of Proposition 1.

A.4. Proofs of Corollaries 1–4. The statements of these corollaries follow by direct application of Proposition 1–2 and Theorem 1. The statement of Corollary 1 is immediate. For Corollaries 2–4, we only have to check that the error bounds in Theorem 1 and Proposition 1–2 are of the desired order. We only discuss how the constants α, δ and ξ can be chosen. Note that all these constants have no subindex because we only consider the case $d = 1$. We apply Theorem 1 conditionally on the values of $S_1, \dots, S_n$. Then the only randomness in the pilot estimation comes from $\zeta_1, \dots, \zeta_n$. We can decompose $\hat{r}$ into $\hat{r}_A + \hat{r}_B$, where $\hat{r}_A$ is the local polynomial fit to (S_i, ζ_i), and $\hat{r}_B$ is the local polynomial fit to $(S_i, r_0(S_i))$. Conditionally given $S_1, \dots, S_n$, the value of $\hat{r}_B$ is fixed, and for checking Assumption 3, we only have to consider entropy conditions for sets of possible outcomes of $\hat{r}_A$. We will show that with $\alpha = p/k$ one can choose for δ and ξ any value that is larger than $(1 - p\theta)/2$ or $-pk^{-1}(1 - p\theta)/2 + p\theta$, respectively. Note that then $\alpha \le 2$ because of Assumption 4(iii). It can be easily checked that we get the desired expansions in Corollaries 1 and 2 with this choices of $\alpha = p/k$, δ and ξ (with δ and ξ small enough). In particular note that we can make $\delta\alpha + \xi$ as close to $p\theta$ as we like.

It is clear that Assumption 2 holds for this choice of δ. This follows by standard smoothing theory for local polynomials. Compare also Lemma 5 and the proof of Proposition 1. It remains to check Assumption 3. It suffices to check the entropy conditions for the tuple of functions $(n^{-1} \sum_{i=1}^n L_h(S_i - s)[(S_i - s)/g]^\pi \zeta_i : 0 \le \pi_+ \le q, \pi_j \ge 0 \text{ for } j = 1, \dots, p)$. This follows because we get $\hat{r}_A$ by multiplying this tuple of functions with a (stochastically) bounded vector. We now argue that all derivatives of order k of the functions $n^{-1} \sum_{i=1}^n L_h(S_i - s)[(S_i - s)/g]^\pi \zeta_i$ can be bounded by a variable B_n that fulfills $B_n \le b_n = n^{\xi^{**}})$ with probability tending to one. Here ξ^{**} is a number with $\xi^{**} > -\frac{1}{2}(1 - p\theta) + k\theta$. This bound holds uniformly in s and π. Furthermore, the functions $n^{-1} \sum_{i=1}^n L_h(S_i - s)[(S_i - s)/g]^\pi \zeta_i$ can be bounded by a variable A_n that fulfills $A_n \le a_n = n^{\xi^*})$ with probability tending to one. Here ξ^* is a number with $\xi^* > -\frac{1}{2}(1 - p\theta)$. Again, this bound holds uniformly in s and π. We now consider the set of functions on I_S that are absolutely bounded by a_n and that have all partial derivatives of order k absolutely bounded by b_n. We argue that this set can be covered by $C \exp(\lambda^{-p/k} b_n^{p/k})$ balls with $\|\cdot\|_\infty$-radius λ for $\lambda \le a_n$. Here the constant C does not depend on a_n and b_n. This entropy bound shows that Assumption 3 holds with these choices of α, δ and ξ. For the proof of the entropy bound one applies an entropy bound for the set of functions on I_S that are absolutely bounded by 1 and that have all partial derivatives of order k absolutely bounded by 1. This set can be covered by $C \exp(\lambda^{-p/k})$ balls with $\|\cdot\|_\infty$-radius λ for $\lambda \le 1$. The desired entropy bound follows by rescaling of the functions. Note that we have that $b_n^{-1} a_n \to 0$.

A.5. Proof of Corollary 5. Our proof has the same structure as the one provided by Lewbel and Linton (2002), but making use of Theorem 1 considerably simplifies some of their arguments. First, note that the restriction that $\underline{\theta} < \theta < \overline{\theta}$

implies that $(ng^p)^{1/2}h^2 \to 0$ and $(ng^p)^{1/2}g^{q+1} \to 0$. From a second-order Taylor expansion, we furthermore obtain that

$$\begin{aligned}\hat{\mu}(x) - \mu_0(x) &= \frac{1}{q_0(r_0(x))}\big(\hat{r}(x) - r_0(x)\big) \\ &\quad + \int_{r_0(x)}^{\lambda} \frac{\hat{q}(s) - q_0(s)}{q_0(s)^2}\,ds - \frac{\hat{q}'(\bar{r}(x))}{2\hat{q}(\bar{r}(x))^2}\big(\hat{r}(x) - r(x)\big)^2 \\ &\quad - \int_{r(x)}^{\lambda} \frac{(\hat{q}(s) - q_0(s))^2}{\hat{q}(s)q_0(s)^2}\,ds \\ &\quad + \frac{(\hat{q}(\check{r}(x)) - q_0(\check{r}(x)))^2}{\hat{q}(\check{r}(x))q_0(\check{r}(x))}\big(\hat{r}(x) - r_0(x)\big) \\ &\equiv T_1 + T_2 + T_3 + T_4 + T_5,\end{aligned}$$

where $\hat{r}(x)$ and $\check{r}(x)$ are intermediate values between $r(x)$ and $\hat{r}(x)$. Now it follows from standard arguments for local linear estimators that

$$\sqrt{ng^p}T_1 \xrightarrow{d} N\left(0, \frac{\sigma_r^2(x)}{f_S(x)s_0^2(x)} \int L^2(t)\,dt\right),$$

since $s_0(x) = q_0(r_0(x))$. To prove the corollary, it thus only remains to be shown that the remaining four terms in the above expansion are of smaller order than T_1. Under the conditions of the corollary, it is easy to show with straightforward rough arguments that $\inf q(s) > 0$, $\sup \hat{q}'(s) = O_p(1)$ and $\sup |\hat{q}(s) - q_0(s)|^2 = o_p((ng^p)^{-1/2})$ where the supremum and infimum are taken over $s \in (r_o(x) - \epsilon, \lambda_0 + \epsilon)$ for some $\epsilon > 0$, respectively. This directly implies that $T_3 + T_4 + T_5 = o_p((ng^p)^{-1/2})$. Now consider the term T_2. From Theorem 1, we obtain that

$$T_2 = \int_{r_0(x)}^{\lambda} \frac{\tilde{q}(s) - q_0(s)}{q_0(s)^2}\,ds - \int_{r_0(x)}^{\lambda} \frac{q_0'(s)\hat{\Delta}(s) - \hat{\Gamma}(s)}{q_0(s)^2}\,ds + O_p(n^{-\kappa}),$$

where $\tilde{q}(x)$ is the oracle estimator of the function q obtained via local linear regression of $\mathbb{I}\{Y > 0\}$ on $r_0(X)$, and $\hat{\Delta}(s)$ and $\hat{\Gamma}(x)$ are the adjustment terms that appear in the main expansion in Theorem 1, with the necessary adjustments to the notation. Using similar arguments as in the proof of Proposition 1–2 and Corollaries 2–4, and the restriction that $\underline{\theta} < \theta < \bar{\theta}$, we obtain that

$$\begin{aligned}&\int_{r(x)}^{\lambda} \frac{\tilde{q}(s) - q(s)}{q^2(s)}\,ds \\ &\quad = \frac{1}{n}\sum_{i=1}^{n} \frac{\varepsilon_i}{f_R(r_0(X_i))} + O_p(h^2) \\ &\quad = O_p(n^{-1/2}) + O_p(h^2) = o_p((ng^p)^{-1/2})\end{aligned}$$

for $\varepsilon_i = \mathbb{I}\{Y_i > 0\} - q_0(X_i)$, and similarly that

$$\int_{r(x)}^{\lambda} \frac{q_0'(s)\hat{\Delta}(s) - \hat{\Gamma}(s)}{q_0(s)^2}\,ds = O_p(n^{-1/2}) + O_p\left(\frac{\log n}{ng^p}\right) + O_p(g^{q+1})$$
$$= o_p((ng^p)^{-1/2}).$$

Thus $T_2 = o_p((ng^p)^{-1/2})$. Finally, straightforward calculations show that $\underline{\theta} < \theta < \bar{\theta}$ also implies that $O_p(n^{-\kappa}) = o_p((ng^p)^{-1/2})$. This completes the proof.

A.6. Proof of Corollary 6. Let $\hat{f} = (\hat{m}, \hat{\mu}_2)$ and $\bar{f} = (m, \mu_2)$, define the functional $S_n(f)$ as

$$S_n(f) = \frac{1}{n}\sum_{i=1}^{n} f_1\bigl(x_1, z_1, X_{1i} - f_2(Z_i)\bigr) - \mu_1(x_1, z_1),$$

and let $\dot{S}_n(f)[h] = \lim_{t\to 0}(S_n(f + th) - S_n(f))/t$ denote its directional derivative. One then obtains through direct calculations that for any $f = (f_{1,A} + f_{1,B}, f_2)$ with bounded second derivatives we have that

$$\begin{aligned}&\|S_n(f) - S_n(\bar{f}) - \dot{S}_n(\bar{f})[f - \bar{f}]\|_\infty \\ &\quad = O(\|f_2 - \bar{f}_2\|_\infty^2) + O\bigl(\|f_2 - \bar{f}_2\|_\infty \|f_{1,A}^{(v)} - \bar{f}_1^{(v)}\|_\infty\bigr) + O(\|f_{1,B}\|_\infty),\end{aligned}$$

where $f_{1,A}^{(v)}(x_1, z_1, v) = \partial_v f_{1,A}(x_1, z_1, v)$. Using the same kind of arguments as in the proof of Proposition 1, under the conditions of the corollary one can derive the following stochastic expansion of $\hat{m}$ up to order $o_p((nh^{1+d_1})^{-1/2})$, uniformly over (x_1, z_1, v) in the h-interior of the support of (X_1, Z_1, V):

$$\begin{aligned}&\hat{m}(x_1, z_1, v) - m(x_1, z_1, v) \\ &\quad = \frac{1}{nf_R(x_1, z_1, v)}\sum_{i=1}^{n} K_h\bigl((X_{1i}, Z_{1i}, V_i) - (x_1, z_1, v)\bigr)\varepsilon_i \\ &\qquad + o_p((nh^{1+d_1})^{-1/2}),\end{aligned} \tag{A.29}$$

where $\varepsilon_i = Y - m(X_{1i}, Z_{1i}, V_i)$. A similar, but notationally more involved expansion can be derived for values of (x_1, z_1, v) in the proximity of the boundary. Note that since exclusion restriction on the instruments that $\mathbb{E}(U|Z_1, Z_2, V) = \mathbb{E}(U|V)$ implies that $\mathbb{E}(\varepsilon|Z_1, Z_2, V) = 0$. In the notation of Theorem 1, this means that $\rho(s) \equiv 0$, and hence the term corresponding to $\hat{\Gamma}(x)$ is equal to zero and does not need to be considered.

Now let $\hat{f}_{1,A}$ denote the sum of the function m and the leading term of the expansion (A.29), and denote the remainder term by $\hat{f}_{1,B}$. Then it follows from, for example, Masry (1996) and the conditions on η and θ, that

$$\|\hat{f}_2 - \bar{f}_2\|_\infty = O_P\bigl((\log(n)/(ng^{d_1+d_2}))^{1/2}\bigr) = o_p((nh^{1+d_1})^{-1/4}),$$

and it follows from the same result together with Lemma 5 in Appendix A.7 that

$$\|\hat{f}_2 - \bar{f}_2\|_\infty \|\hat{f}_{1,A}^{(v)} - \bar{f}_1^{(v)}\|_\infty = O_P\big(\log(n)/(n^2 h^{3+d_1} g^{d_1+d_2})^{1/2}\big)$$
$$= o_p((nh^{1+d_1})^{-1/2}).$$

For any fixed values (x_1, z_1) we thus have that

$$\hat{\mu}_1(x_1, z_1) - \mu_1(x_1, z_1) = S_n(\hat{f}) = S_n(\bar{f}) + T_{1,n} + T_{2,n} + o_p((nh^{1+d_1})^{-1/2}),$$

where

$$T_{1,n} = -\frac{1}{n}\sum_{i=1}^{n} m^{(v)}(x_1, z_1, V_i)\big(\hat{\mu}_2(Z_i) - \mu_2(Z_i)\big),$$
$$T_{2,n} = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{m}(x_1, z_1, V_i) - m(x_1, z_1, V_i)\big).$$

Being a simple sample average of i.i.d. mean zero random variables, one can directly see that $S_n(f_0) = O_p(n^{-1/2}) = o_p((nh^{1+d_1})^{-1/2})$. Using a stochastic expansion for $\hat{\mu}_2$ as in the proof of Proposition 1, and applying projection arguments for U-statistics, one also finds that $T_{1,n} = O_p(n^{-1/2}) = o_p((nh^{1+d_1})^{-1/2})$. Now consider the term $T_{2,n}$. From the expansion in (A.29), it follows that for any fixed values (x_1, z_1) we have that

$$T_{2,n} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{n f_R(x_1, z_1, V_j)} \sum_{i=1}^{n} K_h\big((X_{1i}, Z_{1i}, V_i) - (x_1, z_1, V_j)\big)\varepsilon_i + o_p((nh^{1+d_1})^{-1/2}). \tag{A.30}$$

This in turn implies that

$$\sqrt{nh^{1+d_1}}\, T_{2,n} \xrightarrow{d} N\left(0, \mathbb{E}\left(\frac{\sigma_\varepsilon^2(x_1, z_1, V)}{f_{XZ_1|V}(x_1, z_1, V)}\right) \int \tilde{K}(t)^2\, dt\right)$$

using again projection arguments for U-statistics.

A.7. Uniform rates for generalized kernels. The following auxiliary lemma states uniform rates for averages of i.i.d. mean zero random variables weighted by "kernel-type" expressions. It is used in the proofs of several of our results. Modifications of the lemma are well known in the smoothing literature; see, for example, Härdle, Janssen and Serfling (1988). The lemma can be proved by standard smoothing arguments. One can proceed by using a Markov inequality as in the proof of Lemma 1, but without making use of a chaining argument.

LEMMA 5. *Assume that $D \subset \mathbb{R}^{d_x}$ is a compact set, and $W_{n,h}$ is a kernel-type function that satisfies $W_{n,h}(u, z) = 0$ for $\|u - t(z)\| > b_n h$ for some deterministic sequence $0 < b \le |b_n| \le B < \infty$, and $t : \mathbb{R}^{d_S} \to \mathbb{R}^{d_x}$ a continuously* *differentiable function, for any* $u \in D$ *and* $z \in \mathbb{R}^{d_S}$. *Furthermore, assume that* $|W_{n,h}(u,z) - W_{n,h}(v,z)| \le l \frac{\|u-t(z)\|}{h} h^{-d_x} \tilde{W}_n(v,t(z))$ *with* $\sup_n \tilde{W}_n$ *bounded, and that* $\mathbb{E}[\exp(\rho|\varepsilon|)|S] < C$ *a.s. for a constant* $C > 0$ *and* $\rho > 0$ *small enough. Then we have that*

$$\sup_{x\in D}\left|\frac{1}{n}\sum_{i=1}^{n} a_n W_{n,h}(x,S_i)\varepsilon_i\right| = O_p\left(\sqrt{\frac{\log(n)}{nh^{d_x}}}\right)$$

for any deterministic sequence a_n *with* $|a_n| \le A$.

Acknowledgments. We would like to thank the Associate Editor and three anonymous referees for their comments.

REFERENCES

AHN, H. (1995). Nonparametric two-stage estimation of conditional choice probabilities in a binary choice model under uncertainty. *J. Econometrics* **67** 337–378. MR1333107

ANDREWS, D. W. K. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* **62** 43–72. MR1258665

ANDREWS, D. W. K. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric Theory* **11** 560–596. MR1349935

BLUNDELL, R. W. and POWELL, J. L. (2004). Endogeneity in semiparametric binary response models. *Rev. Econom. Stud.* **71** 655–679. MR2062893

CHEN, X., LINTON, O. and VAN KEILEGOM, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71** 1591–1608. MR2000259

CONRAD, C. and MAMMEN, E. (2009). Nonparametric regression on a generated covariate with an application to semiparametric GARCH-in-Mean models. Unpublished manuscript.

DAS, M., NEWEY, W. K. and VELLA, F. (2003). Nonparametric estimation of sample selection models. *Rev. Econom. Stud.* **70** 33–58. MR1952565

D'HAULTFOEUILLE, X. and MAUREL, A. (2009). Inference on a generalized Roy model, with an application to schooling decisions in France. Unpublished manuscript.

EINMAHL, U. and MASON, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.* **13** 1–37. MR1744994

ESCANCIANO, J. C., JACHO-CHÁVEZ, D. and LEWBEL, A. (2011). Uniform convergence for semiparametric two step estimators and tests. Unpublished manuscript.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. CRC Press, New York.

HAHN, J. and RIDDER, G. (2011). The asymptotic variance of semiparametric estimators with generated regressors. Unpublished manuscript.

HÄRDLE, W., JANSSEN, P. and SERFLING, R. (1988). Strong uniform consistency rates for estimators of conditional functionals. *Ann. Statist.* **16** 1428–1449. MR0964932

HECKMAN, J. J., ICHIMURA, H. and TODD, P. (1998). Matching as an econometric evaluation estimator. *Rev. Econom. Stud.* **65** 261–294. MR1623713

HECKMAN, J. J. and VYTLACIL, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* **73** 669–738. MR2135141

IMBENS, G. W. and NEWEY, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* **77** 1481–1512. MR2561069

KANAYA, S. and KRISTENSEN, D. (2009). Estimation of stochastic volatility models by nonparametric filtering. Unpublished manuscript.

LEWBEL, A. and LINTON, O. (2002). Nonparametric censored and truncated regression. *Econometrica* **70** 765–779. MR1913830

LI, Q. and WOOLDRIDGE, J. M. (2002). Semiparametric estimation of partially linear models for dependent data with generated regressors. *Econometric Theory* **18** 625–645. MR1906328

LINTON, O. and NIELSEN, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93–100. MR1332841

MAMMEN, E., LINTON, O. and NIELSEN, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490. MR1742496

MAMMEN, E., ROTHE, C. and SCHIENLE, M. (2011). Semiparametric estimation with generated covariates. Unpublished manuscript.

MASRY, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Ser. Anal.* **17** 571–599. MR1424907

NEWEY, W. K. (1994a). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* **10** 233–253. MR1293201

NEWEY, W. K. (1994b). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382. MR1303237

NEWEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147–168. MR1457700

NEWEY, W. K., POWELL, J. L. and VELLA, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica* **67** 565–603. MR1685723

PAGAN, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *Internat. Econom. Rev.* **25** 221–247. MR0741926

SONG, K. (2008). Uniform convergence of series estimators over function spaces. *Econometric Theory* **24** 1463–1499. MR2456535

SPERLICH, S. (2009). A note on non-parametric estimation with predicted variables. *Econom. J.* **12** 382–395. MR2562393

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

E. MAMMEN
DEPARTMENT OF ECONOMICS
UNIVERSITY OF MANNHEIM
D-68131 MANNHEIM
GERMANY
E-MAIL: emammen@rumms.uni-mannheim.de

C. ROTHE
TOULOUSE SCHOOL OF ECONOMICS
21 ALLEE DE BRIENNE
F-31000 TOULOUSE
FRANCE
E-MAIL: rothe@cict.fr
URL: http://www.christophrothe.net

M. SCHIENLE
SCHOOL OF BUSINESS AND ECONOMICS
HUMBOLDT UNIVERSITY BERLIN
SPANDAUER STR. 1
D-10178 BERLIN
GERMANY
E-MAIL: melanie.schienle@wiwi.hu-berlin.de