Nonparametric Regression with Nonparametrically Generated Covariates

Enno Mammen, Christoph Rothe, and Melanie Schienle*

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

This Version: March 11, 2010

Abstract

In this paper, we analyze the properties of nonparametric estimators of a regression function when some covariates are not directly observed, but have only been estimated by some nonparametric procedure. We provide general results that can be used to establish rates of consistency or asymptotic normality in numerous econometric applications, including nonparametric estimation of simultaneous equation models, sample selection models, treatment effect models, and censored regression models.

JEL Classification: C14, C31

Keywords: *Nonparametric regression, Generated regressors, Propensity score, Control variable methods*

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

1 Introduction

This paper considers the properties of nonparametric estimators of a regression function when some of the covariates are not directly observed, but have only been estimated nonparametrically. This type of "nonparametric generated regressors" problem is a common occurrence in many econometric applications. Examples include the nonparametric estimation of simultaneous equation models (Imbens and Newey 2009), sample selection models (Das, Newey, and Vella 2003), treatment effect models (Heckman, Ichimura, and Todd 1998), and censored regression models (Linton and Lewbel 2002), amongst many others. While parametric regression problems with generated regressors have been studied extensively, few general results concerning nonparametric settings are available in the literature.

This paper provides general methods to derive asymptotic properties, such as asymptotic normality or rates of consistency, for a wide class of estimators in models involving regression with nonparametrically generated covariates. Our results cannot only be used to study the limit behavior of the estimated regression function itself, but also to analyse more complex estimators, in which estimation of a regression function constitutes only an intermediate step. Examples for the latter case include structured nonparametric models imposing e.g. additive separability (Stone 1985), and semiparametric M-estimators involving infinite dimensional nuissance parameters (e.g. Andrews 1994, Newey 1994b, Chen, Linton, and Van Keilegom 2003).

Our paper considers nonparametric estimation of a regression function $m_0(x) = \mathbb{E}(Y|r_0(S) = x)$ when the function r_0 is unknown, but can be estimated from the data. In particular, we study the properties of the estimator $\widehat{m}_{LL}$ obtained through local linear regression (Fan and Gijbels 1996) of the dependent variable Y on the generated covariates $\hat{R} = \hat{r}(S)$, where $\hat{r}$ is some nonparametric estimate of r_0 from a first stage. Using results from empirical process theory, we show that the presence of generated covariates affects the first-order asymptotic properties of $\widehat{m}_{LL}$ only through a *smoothed* version of the estimation error $\hat{r}(s) - r_0(s)$. This additional smoothing typically improves the rate of convergence. In order to achieve a certain rate of convergence of the estimator $\widehat{m}_{LL}$ it is thus not necessary that the estimator $\hat{r}$ converges with the same rate or a faster one. This result constitutes the main contribution of this paper.

Our main result can e.g. be used to establish asymptotic normality or uniform rates of consistency of the estimate of m_0. This is demonstrated in the present paper for the important special case that r_0 is the conditional mean function in an auxiliary nonparametric regression. Extensions to other settings are immediate. Moreover, our result provides a convenient way

to analyze the properties of more complex estimation procedures, in which estimation of m_0 constitutes an intermediate step. In this paper, we consider three important econometric applications exhibiting such a structure in greater detail: nonparametric estimation of a simultaneous equation model, nonparametric estimation of a censored regression model, and estimation of average treatment effects via regression on the nonparametrically estimated propensity score. The types of technical difficulties encountered in these examples are representative for those in a wide range of econometric applications.

It should be stressed that our main result does neither require the generated regressors to emerge from a specific type of model, nor do we require a specific nonparametric procedure to estimate them. In particular, our main result holds independent of whether the function r_0 is a regression function or a density, or whether it is estimated by kernel methods, orthogonal series or sieves. We only require certain "high-level" conditions, which are straightforward to verify in practice. Our main result, however, is specific to using a local linear smoother for obtaining the final estimate of m_0. In particular, our proofs make use of certain technical properties of this estimator that are not shared by other common methods. While we do not rule out that one could derive a result similar to our main finding for other methods such as orthogonal series or sieves, we conjecture that this would require a substantially more involved technical argument.

We now discuss some of the related literature. As noted above, parametric estimation of models with generated regressors has a long tradition in econometrics. We refer to Pagan (1984) or Oxley and McAleer (1993) for extensive surveys of the literature. More recently, a number of paper has studied models with nonparametrically generated regressors. Imbens and Newey (2009) use nonparametric estimates of control variables to correct for endogeneity in triangular structural equation models with nonseparable disturbances. Similar techniques are used by Newey, Powell, and Vella (1999) for simultaneous equation models with additive disturbances, Blundell and Powell (2004) and Rothe (2009) for single-index binary choice models with endogenous regressors, and Ahn and Powell (1993) and Das, Newey, and Vella (2003) for the estimation of sample selection models with a nonparametrically specified selection mechanism. Linton and Lewbel (2002) face nonparametrically generated covariates when estimating a regression function under fixed censoring of the dependent variable. Lewbel and Linton (2007) consider estimation of a homothetically separable functions. Rilstone (1996) uses generated regressors to reduce the dimensionality of certain nonparametric regression problems. In the literature on program evaluation, Heckman, Ichimura, and Todd (1998) consider estimating the average treatment effect on the treated through regression on the estimated propensity score.

Conditioning on an estimate of a propensity score is also required for computing the Marginal Treatment Effect discussed in Heckman and Vytlacil (2005, 2007) and Carneiro, Heckman, and Vytlacil (2009a, 2009b). Similar issues also appear for the estimation of a generalized Roy model in d'Haultfoeuille and Maurel (2009). Conrad and Mammen (2009) consider non- and semiparametric specifications of GARCH-in-Mean models where generated covariates are plugged iteratively into a nonparametric mean equation. They make use of empirical process methods that are related to the approach of this paper. The aforementioned papers typically rely on restrictions implied by the respective application for their asymptotic analysis. In contrast, our paper provides general results on nonparametric regression with nonparametrically generated regressors, requiring only certain "high-level" conditions not tied to a specific model.

In a recent contribution, Hahn and Ridder (2010) consider the asymptotic variance of semiparametric M-estimators based on nonparametrically generated covariates, generalizing classic results by Newey (1994b). Their approach is to derive the influence function of the estimator of the finite dimensional parameter vector heuristically, i.e. without giving explicit regularity conditions on the estimators involved. In contrast, our paper provides a complete asymptotic theory for nonparametric regression with generated covariates, that would be needed to implement the results in Hahn and Ridder (2010) for a specific estimator. Furthermore, whereas Hahn and Ridder (2010) focus on the estimation of finite dimensional parameters in a certain semiparametric settings, our paper deals with the properties of nonparametric regression with generated covariates in general.

The outline of this paper is as follows. In the next section, we describe our setup in detail and give some motivating examples. Section 3 establishes the asymptotic theory and states the main results. Section 4 provides a number of useful extensions. In Section 5, we apply our results to the examples given in Section 2, thus illustrating their application in practice. Finally, Section 6 concludes.

2 Nonparametric Regression with Generated Covariates

2.1 Model and Estimation Procedure

The nonparametric regression model with generated regressors can be written as

$$Y = m_0(r_0(S)) + \varepsilon, \tag{2.1}$$

where Y is the dependent variable, S is a p-dimensional vector of covariates, $m_0 : \mathbb{R}^d \to \mathbb{R}$ and $r_0 : \mathbb{R}^p \to \mathbb{R}^d$ is an unknown function and ε is an error term that satisfies $\mathbb{E}(\varepsilon|S) = 0$. 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 may involve another random vector T in addition to Y and S.

Our aim is to estimate the function $m_0(r) = \mathbb{E}(Y|r_0(S) = r)$. 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 prescribe 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 $\widehat{m}_{LL}$ of m_0 through a nonparametric regression of Y on the generated covariates $\hat{R} = \hat{r}(S)$, using local linear smoothing. That is, our estimator is given by $\widehat{m}_{LL}(x) = \widehat{\alpha}$, where

$$(\widehat{\alpha}, \widehat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \alpha - \beta^T(\widehat{R}_i - x))^2 K_h(\widehat{R}_i - x),$$

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

For the later asymptotic analysis, it will be useful to compare $\widehat{m}_{LL}$ to an infeasible estimator $\widetilde{m}_{LL}$ 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 $R = r_0(S)$, i.e. it is given by $\widetilde{m}_{LL}(x) = \widetilde{\alpha}$, where

$$(\widetilde{\alpha}, \widetilde{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \alpha - \beta^T(R_i - x))^2 K_h(R_i - x).$$

In order to distinguish these two estimators, we refer to $\widehat{m}_{LL}$ in the following as the *real* estimator, and to $\widetilde{m}_{LL}$ 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 do not require regular densities for the randomly perturbed covariates, and the complete asymptotic theory allows a clear description 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.

2.2 Motivating Examples

There are many econometric applications which involve nonparametric estimation of a regression function using nonparametrically generated covariates. Here we focus on three motivating examples. In this section we state their setup and explain how they fit into our framework. In Section 5, we show how our general high-level results given in the following section can be used to study their asymptotic properties in detail.

2.2.1 Regression on the Propensity Score

Propensity score methods are widely used in the program evaluation literature (see e.g. Imbens (2004) for an extensive review). Their popularity is due to Rosenbaum and Rubin's (1983) famous result that when all confounders are observable, biases due to nonrandom selection into the program can be removed by conditioning on the propensity score, which is defined as the probability of selection into the program given the confounders. To be specific, let Y_1, Y_0 be the potential outcomes with and without program participation, respectively, $D \in \{0,1\}$ an indicator of program participation, $Y = Y_1 D + Y_0(1-D)$ be the observed outcome, and X the vector of confounders, i.e. exogenous covariates. Furthermore, let $\Pi(x) = \Pr(D=1|X=x)$ be the propensity score and define $\nu_d(\pi) = \mathbb{E}(Y_d|\Pi(X)=\pi)$. A typical object of interest in this context is the average treatment effect (ATE), defined as

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

Under unconfoundedness $Y_1, Y_0 \perp D|X$ and full overlap, i.e. $0 < \Pi(X) < 1$ a.s., Rosenbaum and Rubin (1983) showed that $\nu_d(\pi) = \mathbb{E}(Y|D=d, \Pi(X)=\pi)$, which implies that the ATE is identified through the relationship

$$\gamma_{ATE} = \mathbb{E}(\nu_1(\Pi(X)) - \nu_0(\Pi(X))). \tag{2.2}$$

Similar arguments can be made for other measures of program effectiveness (e.g. Heckman, Ichimura, and Todd 1998). Estimating the ATE by a sample analogue of (2.2) requires nonpara-

metric estimation of $\nu_1(\pi)$ and $\nu_0(\pi)$, which fits into our framework with $(Y, S) = (Y, (D, X))$ and $r_0(S) = (D, \Pi(X))$.

2.2.2 Nonparametric Simultaneous Equation Models

Another field of application for our results is the analysis of nonparametric estimators that use control variable techniques to account for endogeneity. The key idea of this approach is to introduce additional conditioning variables which fully capture the dependence between covariates and the unobserved heterogeneity. Such control variables appear naturally in many settings, but are often not directly observable and have to be estimated from the data. Consider for example the estimation of nonparametric simultaneous equation models with additive disturbances discussed in Newey, Powell, and Vella (1999). These authors study a triangular system of equations of the form

$$Y = \mu_1(X_1, Z_1) + U \tag{2.3}$$

$$X_1 = \mu_2(Z_1, Z_2) + V, \tag{2.4}$$

imposing the restrictions that $\mathbb{E}(V|Z_1, Z_2) = 0$, $\mathbb{E}(U) = 0$ and $\mathbb{E}(U|Z_1, Z_2, V) = \mathbb{E}(U|V)$. The last conditions follows e.g. if the instruments $Z = (Z_1, Z_2)$ are jointly independent of the disturbances (U, V) and if the disturbances have mean zero. Now let $m(x_1, z_1, v) = \mathbb{E}(Y|X_1 = x_1, Z_1 = z_1, V = v)$. An implication of this model is that

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

where $\lambda(v) = \mathbb{E}(U|V = v)$. Newey, Powell, and Vella (1999) proposed a series estimator of the structural function μ_1 that exploits this additive separability. An alternative approach to estimating μ_1, which we pursue in this paper, is to use the method of marginal integration (Newey 1994a, Linton and Nielsen 1995). This method relies on the fact that

$$\int m(x_1, z_1, v) f_V(v) dv = \mu_1(x_1, z_1), \tag{2.5}$$

where f_V is the probability density function of V. An estimate of μ_1 can thus be obtained from a sample version of (2.5). Since the residuals V are not directly observed but have themselves to be estimated by some nonparametric method, estimation of the function m 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. Imbens and Newey (2009) consider a generalized version of the above simultaneous equation model where the disturbances may not enter the equations additively. This model fits

into the framework of this paper but requires a careful analysis of additional boundary problems that go beyond the scope of this paper. We will therefore study this model in a separate paper.

Remark 2. 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. Results on smooth backfitting with nonparametrically generated covariates will be presented in a separate paper.

2.2.3 Nonparametric Censored Regression

As a final example, consider the nonparametric estimator of a regression function in the presence of fixed censoring proposed by Linton and Lewbel (2002). Consider the model

$$Y = \max(0, \mu_0(X) - U), \tag{2.6}$$

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 economic applications, e.g. the analysis of wage data. Note that the censoring threshold could be different from zero, as long as it is known. Linton and Lewbel 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

$$\mu_0(x) = \lambda_0 - \int_{r_0(x)}^{\lambda_0} \frac{1}{q_0(r)} dr, \tag{2.7}$$

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 analogue of (2.7), i.e. 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 Asymptotic Properties

It is straightforward to show that $\widehat{m}_{LL}$ consistently estimates the function m_0 under standard conditions. Obtaining refined asymptotic properties, however, requires more involved arguments. Our main result, derived in this section, is a stochastic expansion of the difference

between the real and the oracle estimator, in which the leading term turns out to be a kernel-weighted average of the first stage estimation error. This important finding can e.g. be used to obtain uniform rates of consistency for the real estimator, or to prove its asymptotic normality. This is demonstrated explicitly for the case that $\hat{r}$ results from a local polynomial conditional mean regression.

Throughout this section, we use the notation that for any vector $a \in \mathbb{R}^d$ the value $a_{min} = \min_{1 \leq j \leq d} a_j$ 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 $b \in \mathbb{R}^d$.

3.1 Assumptions

In order to analyse the asymptotic properties of the local linear estimator with nonparametrically generated regressors, we make the following assumptions.[1]

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), $i = 1, \ldots, n$ are independent and identically distributed.

(ii) The random vector $R = r_0(S)$ is continuously distributed with compact support $I_R = I_{R,1} \times ... \times I_{R,d}$. 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) $E[\exp(\rho|\varepsilon|)|S] \leq C$ almost surely for a constant $C > 0$ and $\rho > 0$ 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 $\eta_+ < 1$.

Assumption 1 contains mostly standard conditions from the literature on kernel-based nonparametric regression, with the exception of Assumption 1 (iv). This assumption restricts the

[1] At various points in this section, we will impose assumptions on the rates at which certain quantities tend to zero. We prefer to formulate these assumption without including (various powers of) logarithmic terms. This simplifies the notation for the theorems and proofs at the cost of only a minor loss in generality.

distribution of the error term ε to have subexponential tails conditional on S. This is a technical condition that will be needed to apply certain results from empirical process theory in our proofs.

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

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

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. Such results are widely available for all common nonparametric estimators. See e.g. 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) $\Pr(\widehat{r}_j \in \mathcal{M}_{n,j}) \to 1$ *as* $n \to \infty$ *for all* $j = 1, \ldots, d$.

(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 $\overline{\mathcal{M}}_{n,j} = \mathcal{M}_{n,j} \cap \{r_j : \|r_j - r_{n,j}\|_\infty \leq n^{-\delta_j}\}$ can be covered by at most $C_M \exp(\lambda^{-\alpha_j} n^{\xi_j})$ balls with $\|\cdot\|_\infty$-radius λ for all $\lambda \leq n^{-\delta_j}$, where $0 < \alpha_j \leq 2$, $\xi_j \in \mathbb{R}$ 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) for details. 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 $\xi_j^* \geq 0$. 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 e.g. the one-dimensional Nadaraya-Watson estimator $\widehat{r}_{n,j}$ with bandwidth of order $n^{-1/5}$. Choose $r_{n,j}$ equal to $r_{0,j}$ plus asymptotic bias term. Then one can check that the second derivative of $\widehat{r}_{n,j} - r_{n,j}$ is absolutely

bounded by $O_P(\sqrt{\log n}) = o_P(n^{\xi_j^*})$ 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.

3.2 The Key Stochastic Expansion

With the assumptions described in the previous section, we are now ready to state our main result, a stochastic expansion of our real estimator $\widehat{m}_{LL}(x)$ around the oracle estimator $\widetilde{m}_{LL}(x)$. The results explicitly characterizes the influence of the presence of nonparametrically generated regressors on the final estimator of the regression function m_0. To state the theorem, let $\hat{\Delta}(x) = \overline{\alpha}$, where

$$(\overline{\alpha}, \overline{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} ((\widehat{r}(S_i) - r_0(S_i)) - \alpha - \beta^T (r_0(S_i) - x))^2 K_h(r_0(S_i) - x),$$

and define the set $I_{R,n}^- = \{x \in I_R : \text{the support of } K_h(\cdot - x) \text{ is a subset of } I_R\}$.

Theorem 1. *Suppose Assumptions 1–3 hold. Then*

$$\sup_{x \in I_R} \left| \widehat{m}_{LL}(x) - \widetilde{m}_{LL}(x) + m_0'(x)\hat{\Delta}(x) \right| = O_P(n^{-\kappa})$$

where $\kappa = \min\{\kappa_1, ..., \kappa_4\}$ *with*

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

$$\kappa_3 < \delta_{min} + \eta_{min}, \;\; \kappa_4 < \delta_{min} + (\delta - \eta)_{min}.$$

Uniformly over $x \in I_{R,n}^-$ *we have that*

$$\hat{\Delta}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)(\widehat{r}(S_i) - r_0(S_i))}{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)} + O_P(n^{-\kappa}). \tag{3.1}$$

Proof. See the Appendix. □

The leading term in the above expansion of the real estimator $\widehat{m}_{LL}(x)$ around the oracle estimator $\widetilde{m}_{LL}(x)$ is given by the product of the derivative of m_0 and a smoothed version of the first-stage estimation error $\widehat{r}(s) - r_0(s)$. 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 bias, it typically reduces the variance and thus

allows for less precise first-stage estimators. Another implication of the theorem is that using generated regressors has asymptotically negligible consequences in regions where the regression function is flat, since $m_0'(x) = 0$ in this case.

Remark 3. 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 $\widehat{m}$ and $\widetilde{m}$, and to treat the leading terms of the resulting higher order expansion by other more direct methods.

Remark 4. One could also derive an explicit representation of the term $\hat{\Delta}(x)$ for values of x near the boundary of the support of R. This would be similar to the one given in (3.1), but involve weighting by more complicated kernel functions.

3.3 Two-Stage Nonparametric Regression

Theorem 1 can be used to derive asymptotic properties of the real estimator $\widehat{m}_{LL}$, such as uniform rates of consistency or pointwise asymptotic normality in various econometric models. In this subsection, we demonstrate how explicit forms of the results in Theorem 1 can be obtained in the specific case that r_0 is the conditional expectation function in an auxiliary nonparametric regression. Then we show how these can be employed to derive desired asymptotic properties. The chosen setting is arguably the most common way nonparametrically generated covariates appear in practice, and all the applications we consider in detail in this paper are either of this or a very closely related form.

We 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 $E[\zeta|S] = E[\varepsilon|S] = 0$. 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 $\widehat{m}_{LL}$ in this framework is as follows: We first derive an explicit representation for the adjustment term $\hat{\Delta}(x)$ from Theorem 1, which can

then be combined with standard results about the oracle estimator $\widetilde{m}_{LL}$. In order to obtain such a result, it is convenient to use a kernel-based smoother in the first stage to estimate r_0. Since the bias of $\hat{\Delta}(x)$ is of the same order as of this first-stage estimator, we propose to estimate the function r_0 via q-th 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

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(T_i - \alpha - \sum_{1 \leq u_+ \leq q} \beta_r^T (S_i - s)^u \right)^2 L_g(S_i - s) \tag{3.2}$$

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 bandwidth, which for simplicity is assumed to be the same for all components, and $\sum_{1 \leq u_+ \leq q}$ denotes the summation over all $u = (u_1, \ldots, u_p)$ with $1 \leq u_+ \leq q$. When r_0 is sufficiently smooth, the asymptotic bias of local polynomial estimators of order q is well-known to be of order $O(g^{q+1})$ 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 k-th derivative of the estimator $\widehat{r}$. This allows to verify the Complexity Assumption 3 for $\widehat{r}$.

Assumption 4. *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 $I_S = I_{S,1} \times ... \times I_{S,p}$. Its density function f_S is bounded and bounded away from zero on I_S. It is also differentiable with a bounded derivative.

(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 $k \geq \max\{2, p/2\}$.

(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 5. *The function $r_0(s_{-p}, u)$ is strictly monotone in u, and $r_0(s_{-p}, \varphi(s_{-p}, x)) = x$ for some twice continuously differentiable function φ.*

The following Lemma 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 $\zeta_1, \ldots, \zeta_n$, 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.22) in the Appendix, and define the one-dimensional kernel functions

$$H_g(x,s) = \int \frac{1}{g} L^* \left(u_1, \frac{\varphi(s_{-p}, x) - s_p}{g} - \partial_1 \varphi(s_{-p}, x) u_1 \right) du_1, \tag{3.3}$$

$$M_h(x,s) = \int K_h \left(r_0(s) - x - r_0'(s) uh \right) L^*(u) du. \tag{3.4}$$

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

Lemma 1. *Suppose that Assumptions 1, 4 and 5 hold. Then we have that, uniformly over $x \in I_R$,*

$$\hat{\Delta}(x) = \hat{\Delta}_A(x) + \hat{\Delta}_B(x) + O_p\left(\frac{\log(n)}{ng^p}\right) + O_p\left(\frac{\log(n)}{(ng^p)^{1/2}(nh)^{1/2}}\right),$$

where $\hat{\Delta}_B(x) = O_p(g^{q+1})$ and $\hat{\Delta}_A(x) = O_p((\log(n)/(n \max\{g,h\}))^{1/2})$. 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.21) 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

$$\hat{\Delta}_A(x) = \frac{1}{n f_R(x)} \sum_{i=1}^n K_h(r_0(S_i) - x)\zeta_i + O_p\left(\left(\frac{g}{h}\right)^2 \left(\frac{\log(n)}{nh}\right)^{1/2} \right).$$

b) If $h = g$ then

$$\hat{\Delta}_A(x) = \frac{1}{n f_R(x)} \sum_{i=1}^n M_h(x, S_i)\zeta_i + o_p(n^{-1/2}).$$

c) If $g/h \to \infty$ then

$$\hat{\Delta}_A(x) = \frac{1}{n f_R(x)} \sum_{i=1}^n H_g(x, S_i) \partial_x \varphi(S_{-p,i}, x)\zeta_i + O_p\left(\left(\frac{h}{g}\right)^2 \left(\frac{\log(n)}{ng}\right)^{1/2} \right).$$

Proof. See the Appendix. □

It should be emphasized that in all three cases of the Lemma 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.

Remark 5. The previous lemma can easily be modified in two directions. First, if the second-order kernel function K is replaced with a kernel function of order k, the order of the remainder term in the representation of $\hat{\Delta}_A(x)$ can be strengthened to $O_p((g/h)^k(nh/\log(n))^{-1/2})$ in case a) of the Lemma, and to $O_p((h/g)^k(nh/\log(n))^{-1/2})$ for case c), under appropriate smoothness conditions. The expansions in Lemma 1 also continue to hold if the local polynomial estimator of r_0 is replaced by a Nadaraya-Watson estimator with a higher order kernel function whose moments up to order q equal zero.

Combining Theorem 1 and Lemma 1 with well-known results about the oracle estimator $\widetilde{m}_{LL}$, various asymptotic properties of the real estimator $\widehat{m}_{LL}$ can be derived. In the following theorems we present results in the most relevant scenarios, addressing uniform rates of consistency, stochastic expansions of order $o_P(n^{-2/5})$ for proving pointwise asymptotic normality, and a more refined expansion of order $o_P(n^{-1/2})$ that is useful when m_0 is estimated as an intermediate step in a semiparametric problem.

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

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

This implies the following result.

Theorem 2. *Suppose that Assumptions 1, 4 and 5 hold. Then*

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

Straightforward calculations show that the term of order $O_P(n^{-\kappa})$ is dominated by the other remainder terms if $\theta < \max\{(1/2-\eta)/p, (1-7\eta/2)/p, (1-3\eta/2)/(p+q+1)\}$. Similarly, under appropriate smoothness restrictions, all of the last four terms on the right-hand side of the last

equation can be made strictly smaller than the first two ones given an appropriate choice of η and θ. One can thus recover the oracle rate for the real estimator, even if the first-stage estimator converges at a strictly slower rate.

Next, we derive stochastic expansions of $\widehat{m}_{LL}$ 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. It turns out that applying Theorem 1 requires $p\theta < 3/10$ in this case. Therefore, in order to use expansions a) and b) of Lemma 1, only $p = 1$ is admissible, i.e. S must be one-dimensional in order for choices of θ with $\theta \geq \eta$ to be feasible. We will consider this case in the next theorem. The case of oversmoothed pilot estimation with $\theta < \eta$ will be discussed in Theorem 4.

Theorem 3. *Suppose that Assumptions 1, 4 and 5 hold with $\eta = 1/5$ and $p = q = 1$ Then the following expansions hold uniformly over $x \in I^-_{R,n}$:*

a) If $1/5 < \theta < 3/10$ then

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

In particular, we have

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

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

b) If $\theta = 1/5$ then

$$\widehat{m}_{LL}(x) - m_0(x) = \frac{1}{nf_R(x)} \sum_{i=1}^{n} K_h(r_0(S_i) - x)\varepsilon_i - K_h^x(r_0(S_i) - x)\zeta_i + \beta(x)h^2 + o_p\left(n^{-2/5}\right),$$

where $K^x(v) = \int K(v - r'(r^{-1}(x))u)L^*(u)du$ is a kernel that depends on x and the bias is given by $\beta(x) = \frac{1}{2}\int u^2 K(u)du \; m_0''(x) - \frac{1}{2}h^2 \int u^2 L(u)du \; r_0''(r_0^{-1}(x))m_0'(x)$. In particular, we have

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

where now $\sigma_m^2(x) = [\,Var(\varepsilon | R = x) \int K(t)^2 dt - 2m_0'(x)E(\varepsilon\zeta | R = x) \int K(t)K^x(t)dt + m_0'(x)^2 \, Var(\zeta | R = x) \int K^x(t)^2 dt]/f_R(x)$ is the asymptotic variance.

We can see that under the conditions of the theorem the limiting distribution of $\widehat{m}_{LL}(x)$ is affected by the pilot estimation step. In particular, if $\theta > \eta$ the estimator $\widehat{m}_{LL}(x)$ has the same limiting distribution as the local linear estimator in the hypothetical regression model

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

where $\varepsilon^* = \varepsilon - m_0'(r_0(S))\zeta$. 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 $\widetilde{m}_{LL}$.

The next theorem discusses the case when $\theta < \eta$. For such a choice of bandwidth, the limit distribution of $\widehat{m}_{LL}$ is the same as for the oracle estimator $\widetilde{m}_{LL}$. The effect exerted by the presence of nonparametrically generated regressors is thus asymptotically negligible for conducting inference on m_0 in this case.

Theorem 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 $\frac{2}{5}(q+1)^{-1} < \theta < \frac{3}{10}p^{-1}$:*

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

In particular, we have

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

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

When the bandwidth parameters are chosen such that $\theta < \eta$, i.e we have that $g/h \to \infty$, we can also derive stochastic expansions of $\widehat{m}_{LL}$ of order $o_P(n^{-1/2})$ for choices of $\eta > 1/4$. This type of expansion is often needed for the analysis of semiparametric problems in which m_0 plays the role of an infinite dimensional nuisance parameter. Examples include estimation of weighted averages or weighted average derivatives of m_0, or more generally the class of semiparametric M-estimators (e.g. Newey (1994b), Andrews (1994) or Chen, Linton, and Van Keilegom (2003)).

Compared to the expansion of order $o_P(n^{-2/5})$ in the previous Theorem, expansions of order $o_P(n^{-1/2})$ contain an additional higher order term that accounts for estimation errors in the pilot estimation step.

Theorem 5. *Suppose that Assumptions 1, 4 and 5 hold with $\eta > \theta$. Under these conditions, the following expansions hold uniformly over $x \in I_{R,n}^-$ if $\eta > 1/4$ and $\frac{1}{2}(q+1)^{-1} < \theta < \frac{1}{2}(1-3\eta)p^{-1}$:*

$$\widehat{m}_{LL}(x) - m_0(x) = \frac{1}{nf_R(x)} \sum_{i=1}^{n} K_h(r_0(S_i) - x)\varepsilon_i \\ - m_0'(x)\frac{1}{nf_R(x)} \sum_{i=1}^{n} H_g(x, S_i)\partial_x \varphi(S_{-p,i}, x)\zeta_i + o_p\left(n^{-1/2}\right).$$

Note that the conditions of the last two theorems impose restrictions on the smoothness of the function r_0. To obtain the expansion of order $o_P(n^{-2/5})$ in Theorem 4 we need that $q + 1 > \frac{10}{3}\frac{2}{5}p = \frac{4}{3}p$. For the expansion of order $o_P(n^{-1/2})$ in Theorem 5 it is necessary that $q + 1 > (1 - 3\eta)^{-1}p > 4p$. Thus, in both cases the required number of derivatives q has to increase linearly with the dimension of the respective covariates p. In Section 4.3, we discuss a modified version of the real estimators that requires weaker smoothness conditions.

4 Extensions

4.1 Estimation of Derivatives

In certain applications, it is necessary to estimate the derivatives of the regression function m_0, instead of the function itself. One example from the literature on program evaluation is the estimation of the Marginal Treatment Effects (MTE), which is defined as the derivative of the conditional expectation of an outcome variable given the (usually unobserved) propensity score. See e.g. Heckman and Vytlacil (2005, 2007) or Carneiro, Heckman, and Vytlacil (2009a, 2009b) for details. In this section, we therefore discuss extensions of the results in the last section to the estimation of derivatives of m_0. Therefore we consider an estimator based on local quadratic fits. The theory of the last section could also be extended to higher order derivatives (by using higher order local polynomials), but we restrict our analysis to first order derivatives because of their importance in econometrics. We define the real estimator of the derivative as $\widehat{m}_{LQ}^*(x) = \widehat{\beta}$, where with $\widehat{R}_i = \widehat{r}(S_i)$

$$(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}) = \underset{\alpha,\beta,\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \alpha - \beta^T(\widehat{R}_i - x) - (\widehat{R}_i - x)^T\gamma(\widehat{R}_i - x)\right)^2 K_h(\widehat{R}_i - x).$$

Furthermore, the oracle estimator is defined as $\widetilde{m}_{LQ}^*(x) = \widetilde{\beta}$ with

$$(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\gamma}) = \underset{\alpha,\beta,\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \alpha - \beta^T(R_i - x) - (R_i - x)^T\gamma(R_i - x)\right)^2 K_h(R_i - x),$$

where $R_i = r_0(S_i)$. We also define $\widehat{\Delta}^*(x) = \overline{\beta}$ by

$$(\overline{\alpha}, \overline{\beta}, \overline{\gamma}) = \underset{\alpha,\beta,\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} (-m_0'(R_i)(\widehat{R}_i - R_i) - \alpha - \beta^T(R_i - x) - (R_i - x)^T\gamma(R_i - x))^2 K_h(R_i - x).$$

With this notation, we can state a result analogous to Theorem 1.

Theorem 6. *Suppose Assumptions 1–3 hold and assume additionally that the function m_0 is three-times continuously differentiable on I_R. Then it holds for $1 \leq j \leq d$ with $\kappa_1, ..., \kappa_4$ as in Theorem 1 that*

$$\sup_{x \in I_R} \left| \widehat{m}^*_{LQ,j}(x) + \widetilde{m}^*_{LQ,j}(x) - \widehat{\Delta}^*(x) \right| = O_P(n^{\eta_j}(n^{-\kappa_1} + n^{-\kappa_4}) + n^{\eta_j - \eta_{min}}(n^{-\kappa_2} + n^{-\kappa_3})).$$

Furthermore, uniformly over $x \in I^-_{R,n}$ we have that

$$\widehat{\Delta}^*(x) = \left[\frac{1}{n} \sum_{i=1}^{n} K_h(r_0(S_i) - x)(r_0(S_i) - x)(r_0(S_i) - x)^T \right]^{-1}$$
$$\frac{1}{n} \sum_{i=1}^{n} K_h(r_0(S_i) - x) m_0'(r_0(S_i))(r_0(S_i) - x)(\widehat{r}(S_i) - r_0(S_i)) + O_P(n^{\eta_j - \kappa_4}). \tag{4.1}$$

Proof. See the Appendix. □

For the important special case that r_0 is a conditional expectation function estimated by local polynomials, one can derive results analogous to those obtained in Section 3.3 by using the same type of arguments. These are omitted here for the sake of brevity.

4.2 Design densities with unbounded support

One of the assumptions used to derive the stochastic expansion in Theorem 1 is that the covariates $R = r_0(S)$ have bounded support. In this subsection, we relax this condition, allowing R to be supported on an arbitrary subset of $\mathbb{R}^p$. This result might be helpful in settings involving unbounded covariates, or more generally covariates whose density tends zero in certain areas. We make the following assumption.

Assumption 6. *The variable $R = r_0(S)$ is continuously distributed with support $I_R \subset \mathbb{R}^q$. Its density has a bounded continuous derivative.*

We obtain the following generalization of Theorem 1, a stochastic expansion that holds uniformly over an increasing sequence of subsets of the support I_R where the density f_R is sufficiently large. Note that when the support is unbounded the density cannot be strictly positive everywhere.

Theorem 7. *Suppose Assumptions 1(i),(iii)-(vi),2,3 and 6 hold. Then for $C_S > 0$ large enough it holds that*

$$\sup_{x \in I^*_{R,n}} \gamma_n(x)^{-1} \left| \widehat{m}_{LL}(x) - \widetilde{m}_{LL}(x) + m_0'(x)\hat{\Delta}(x) \right| = O_P(n^{-\kappa}) \tag{4.2}$$

where κ is defined as in Theorem 1 and $\gamma_n(x) = (\inf_{u \in S_h(x)} f_R(u))^{1/2} (\sup_{u \in S_h(x)} f_R(u))^{-1}$, where $S_h(x)$ is the support of $K_h(x - \cdot)$ and where

$$\hat{\Delta}(x) = \frac{\frac{1}{n}\sum_{i=1}^n K_h(r_0(S_i) - x)(\widehat{r}(S_i) - r_0(S_i))}{\frac{1}{n}\sum_{i=1}^n K_h(r_0(S_i) - x)} + O_P(n^{-\kappa}). \tag{4.3}$$

The supremum in (4.2) runs over the set $I^*_{R,n} = \{x \in I_R : \inf_{u \in S_h(x)} f_R(u) > C_S(nh)^{-1} \log n\}$ for a constant C_S that is large enough.

4.3 Avoiding Entropy Conditions via Crossvalidation

In this subsection, we consider a slightly modified version of our estimator of m_0, obtained through L-fold crossvalidation. We show that using such an estimator can improve the result of Theorem 1 in two directions. First, an analogous result can be established without imposing an entropy condition such as Assumption 3, and second, one can obtain a faster rate for the remainder term. The improvements are asymptotic. For finite samples, cross validation may be affected by using smaller subsamples in the estimation steps. This may cause instabilities that are not reflected in a first order asymptotic analysis.

Our following theoretical treatment contains crossvalidation as a leading example, but the framework is slightly more general. Nevertheless, we call the resulting estimator crossvalidation estimator and denote it by $\widehat{m}^{CV}_{LL}$. The estimator works as follows. Let N_l, $l = 1, \ldots, L$ be a partition of $N = \{1, ..., n\}$, and denote the number of elements in the l-th set by $\#N_l$. Assume that for every $l \in \{1, \ldots, L\}$ there exists an estimator $\widehat{r}^{[l]}$ of r_0 that is independent of $(Y_i, S_i) : i \in N_l$. In the two-stage regression model discussed in Section 3.3, a possible approach would be to compute $\widehat{r}^{[l]}$ in the same way as $\widehat{r}$ before, but only using the data points (Y_i, S_i, T_i) with $i \notin N_l$. For each $l \in \{1, \ldots, L\}$ we then define the estimators $\widehat{m}^{[l]}_{LL}$ where $\widehat{m}^{[l]}_{LL}(x) = \widehat{\alpha}^{[l]}$, and

$$(\widehat{\alpha}^{[l]}, \widehat{\beta}^{[l]}) = \operatorname*{argmin}_{\alpha,\beta} \sum_{i \in N_l} (Y_i - \alpha - \beta^T(\widehat{r}^{[l]}(S_i) - x))^2 K_h(\widehat{r}^{[l]}(S_i) - x).$$

Finally, we define the crossvalidation estimator $\widehat{m}^{CV}_{LL}$ of the function m_0 as a weighted average of the $\widehat{m}^{[l]}_{LL}$, with weights given by the proportion of data points used in the second stage. That is, we put $\widehat{m}^{CV}_{LL}(x) = \sum_{l=1}^L \omega_l \widehat{m}^{[l]}_{LL}(x)$ with $\omega_l = \#N_l/n$. For this estimator, a result similar to Theorem 1 can be established under the following assumption.

Assumption 7. *We impose the following restriction about the accuracy of the first stage estimators and the number of partitions L.*

(i) For $1 \leq l \leq L$ there exist estimators $\widehat{r}_l$ of the functions r_l, that are independent of $\{(S_i, Y_i) : i \in N_l\}$. The components $\widehat{r}_j^{[l]}$ and $r_{0,j}$ of $\widehat{r}^{[l]}$ and r_0, respectively, satisfy

$$\sup_s \max_{1 \leq l \leq L} |\widehat{r}_j^{[l]}(s) - r_{0,j}(s)| = O_P(n^{-\delta_j})$$

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

(ii) It holds that $cn^\beta \leq \#N_l \leq Cn^\beta$, for some constants $0 < c < C$ and $0 < \beta \leq 1$.

This first part of this assumption is a slight modification of Assumption 2, requiring a certain uniform rate of consistency for the first-stage estimators calculated from the different subsamples. Again, such results are straightforward to verify for many common nonparametric estimation procedures. The second part imposes a restriction on the size of the crossvalidation sets.

Theorem 8. *Suppose that Assumptions 1 and 7 hold. Then*

$$\sup_{x \in I_R} \left| \widehat{m}_{LL}^{CV}(x) - \widetilde{m}_{LL}(x) + m'(x)\hat{\Delta}_{CV}(x) \right| = O_P(n^{-\kappa_{CV}}).$$

Here $\kappa_{CV} = \min\{\kappa_{CV,1}, ..., \kappa_{CV,4}\}$ with

$$\kappa_{CV,1} < \frac{1}{2}(\beta - \eta_+) + (\delta - \eta)_{min}, \ \kappa_{CV,2} < 2\eta_{min} + (\delta - \eta)_{min},$$
$$\kappa_{CV,3} < \delta_{min} + \eta_{min}, \ \kappa_{CV,4} < \delta_{min} + (\delta - \eta)_{min}.$$

Furthermore,

$$\hat{\Delta}_{CV}(x) = \sum_{l=1}^{L} \omega_l \hat{\Delta}_{CV}^{[l]}(x)$$

with $\hat{\Delta}_{CV}^{[l]}(x)_{LL}^{[l]}(x) = \widehat{\alpha}^{[l]}$, where

$$(\widehat{\alpha}^{[l]}, \widehat{\beta}^{[l]}) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i \in N_l} ((\widehat{r}^{[l]}(S_i) - r_0(S_i)) - \alpha - \beta^T(\widehat{r}^{[l]}(S_i) - x))^2 K_h(\widehat{r}^{[l]}(S_i) - x).$$

For $x \in I_{R,n}^-$ we have that

$$\hat{\Delta}_{CV}(x) = \sum_{l=1}^{L} \omega_l \frac{n^{-1} \sum_{i \in N_l} K_h(r_0(S_i) - x)(\widehat{r}^{[l]}(S_i) - r_0(S_i))}{n^{-1} \sum_{i \in N_l} K_h(r_0(S_i) - x)} + O_P(n^{-\kappa_{CV}}).$$

The result in Theorem 8 provides an improvement over Theorem 1 because it holds without imposing a restriction on the complexity of the function r_0, such as the entropy condition in Assumption 3. Of course, some kind of smoothness restrictions are still usually needed to verify Assumption 5 for a specific estimator. A further refinement compared to Theorem 1 is that the stochastic expansion is typically more precise, in the sense that the rate at which the remainder term converges to zero is weakly faster, i.e. we have that $\kappa_{CV} \geq \kappa$ because $\kappa_{CV,1} > \kappa_1$.

4.3.1 Crossvalidation for estimating averages of m_0

We now discuss a cross validation approach for the estimation of a weighted average $\vartheta = \int m_0(x)w(x)dx$ of the regression function m_0. The advantage of this method is that it requires somewhat weaker regularity conditions than direct approaches based on Theorem 1. The framework is as above. Again we divide the sample into L subsets $N_1, ..., N_L \subset \{1, ..., n\}$, $\bigcup_{l=1}^{L} N_l = \{1, ..., n\}$ but now we assume that L is fixed. We rewrite ϑ as $\vartheta = \int m^*(x)w^*(x)dx$ with $m^*(x) = m(x)f_R(x)$ and $w^*(x) = w(x)/f_R(x)$. Now, we assume that there exist estimators $\widehat{w}_l^*$ and $\widehat{r}_l$ of the functions w^* and r_l, that are independent of $\{(S_i, Y_i) : i \in N_l\}$ and we consider the following estimator of ϑ:

$$\hat{\vartheta} = \sum_{l=1}^{L} \frac{n_l}{n} \int \widehat{m}_l^*(x)\widehat{w}_l^*(x)dx,$$

where

$$\widehat{m}_l^*(x) = \frac{1}{n_l} \sum_{i \in N_l} K_h(\widehat{r}_l(S_i) - x)Y_i.$$

Our next theorem states that this estimator is $n^{1/2}$-consistent. For the theorem we make the following assumptions.

Assumption 8. *(i) The observations (S_i, Y_i), $i = 1, ..., n$ are i.i.d. and it holds that $Y_i = m(r(S_i)) + \varepsilon_i$ with $E[\varepsilon_i|S_i] = 0$ and $E[\varepsilon_i^2|S_i] < C_\varepsilon$, almost surely, for a constant $C_\varepsilon < \infty$.*

(ii) The function m_0 is bounded. It holds $\int w^*(r(s))^2 f_S(s)ds < \infty$ and $c \leq n_l/n \leq C$ for some constants $0 < c < C$.

(iii) For $1 \leq l \leq L$ there exist estimators $\widehat{w}_l^*$ and $\widehat{r}_l$ of the functions w^* and r_l, that are independent of $\{(S_i, Y_i) : i \in N_l\}$ with the properties:

$$\int [\widehat{w}_{l,h}^*(\widehat{r}_l(s)) - w^*(r(s))]^2 f_S(s)ds = o_P(1),$$

$$\int \widehat{w}_{l,h}^*(\widehat{r}_l(s))m(r(s))f_S(s)ds - \int w^*(r(s))m(r(s))f_S(s)ds = O_P(n^{-1/2}), \tag{4.4}$$

where $\widehat{w}_{l,h}^*(u) = \int K_h(u - x)\widehat{w}_l^*(x)dx$.

Theorem 9. *Suppose that Assumption 8 holds. Then we have that* $\hat{\vartheta} = \vartheta + O_P(n^{-1/2})$.

For a derivation of the asymptotic distribution of $\hat{\vartheta} - \vartheta$ one would need more information about the construction of the estimators $\widehat{w}_l^*$ and $\widehat{r}_l$. In particular one would need a linear expansion of the left hand side of (4.4).

5 Applications

5.1 Regression on the Propensity Score

As our first application, consider estimation of the Average Treatment Effect (ATE) via regression on the (estimated) propensity score. Recall that the parameter of interest is given by

$$\gamma_{ATE} = \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \mathbb{E}(\nu_1(p(X))) - \mathbb{E}(\nu_1(p(X))), \tag{5.1}$$

where $\Pi(x) = \mathbb{E}(D|X = x)$ is the propensity score and $\nu_d(\pi) = \mathbb{E}(Y|D = d, \Pi(X) = \pi)$ for $d = 0, 1$. A natural estimate of the ATE is thus the following sample version of (5.1):

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

where $\hat{\Pi}(x)$ is the q-th order local polynomial estimator of $\Pi(x)$, and $\hat{\nu}_d(\pi)$ is the local linear estimator of $\nu_d(\pi)$, computed using the first-stage estimates of the propensity score. Here the binary covariate D is accommodated via the usual frequency method, i.e. the estimate $\hat{\nu}_d$ is computed by local linear regression of Y_i on $\hat{\Pi}(X_i)$ using the $n_d = \sum_{i=1}^{n}\mathbb{I}\{D_i = d\}$ observations with $D = d$ only. While Hahn (1998) conjectures that conditioning on the propensity score will not lead to an efficient estimator of the ATE, to the best of our knowledge the asymptotic properties of $\hat{\gamma}$ have not been derived before in the literature.[2]

Proposition 1. *Assume that Assumptions 1 holds with* $(Y, S, T) = (Y, (D, X), D)$, $r_0(S) = (D, \Pi(X))$, $m_0(d, \pi) = \nu_d(\pi)$, *and the obvious modifications to accommodate the binary covariate* D, *and that Assumption 4 holds with* $r_0(S) = \Pi(X)$. *Also suppose that* $\eta \in (1/4, 1/3)$ *and* $(1/2)(q+1)^{-1} < \theta < (1-3\eta)p^{-1}$. *Under these conditions, we have that*

$$\sqrt{n}(\hat{\gamma} - \gamma_{ATE}) \xrightarrow{d} N(0, \mathbb{E}(\psi(Y, D, X)^2))$$

[2]Heckman, Ichimura, and Todd (1998) consider estimating a closely related parameter, average treatment effect on the treated, by conditioning on the estimated propensity score.

where

$$\psi(Y,D,X) = \frac{DY}{\Pi(X)} - \frac{(1-D)Y}{1-\Pi(X)} - (D-\Pi(X))\left(\frac{\nu_1(\Pi(X))}{\Pi(X)} + \frac{\nu_0(\Pi(X))}{1-\Pi(X)}\right) - \gamma_{ATE}$$

is the influence function.

One can show that asymptotic variance of $\hat{\gamma}$ is the same as that of an infeasible estimator based on regression on the true instead of the estimated propensity score. Interestingly, the proof of Proposition 1 reveals that when deriving the influence function ψ, the terms accounting for using the estimated propensity score to *compute* $\hat{\nu}_d$, and those accounting for *evaluating* $\hat{\nu}_d$ at the estimated propensity score, exactly cancel each other. Note that the asymptotic variance of $\hat{\gamma}$ turns out to be different from the corresponding semiparametric efficiency bound obtained by Hahn (1998), confirming the conjecture that the estimator is not fully efficient. We also remark that the conditions of the proposition imply that both $\hat{p}$ and $\hat{\nu}_d$ are uniformly consistent for their respective population counterparts at a rate faster than the well-known minimal convergence rate of $n^{-1/4}$ given by Newey (1994b) for semiparametric two-stage procedures.

5.2 Nonparametric Simultaneous Equation Models

We now consider nonparametric estimation of the structural function μ_1 in the triangular simultaneous equation model (2.3)–(2.4) using the method of marginal integration. 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 (2.4). For simplicity, we use the same bandwidth for all components of $\widehat{m}$, i.e 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 (2.5):

$$\hat{\mu}_1(x_1,z_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(x_1,z_1,\hat{V}_i). \tag{5.2}$$

Using similar arguments as in the proof of Theorem 3, the following proposition established the estimator's asymptotic normality.

Proposition 2. *Suppose that Assumptions 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 4 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:*

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

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* $\sigma_\varepsilon^2(x_1,z_1,v) = Var(Y - m(R)|R = (x_1,z_1,v))$*.*

Under the conditions of the proposition, 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 (5.2) with an oracle estimator $\widetilde{m}$ using the actual disturbances V_i instead of the reconstructed ones, the result would not change.

5.3 Nonparametric Censored Regression

We now consider estimation of the censored regression model given in (2.6). 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 the estimate μ_0 is given by

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

where the constant λ is chosen large enough to satisfy $\lambda > \max_{i=1,\ldots,n} \hat{r}(X_i)$ with probability tending to one. Generalizing Linton and Lewbel (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 $\widehat{\mu}$, 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 proposition.

Proposition 3. *Suppose that Assumptions 1 and 4 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

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

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

The proposition is analogous to Theorem 5 in Linton and Lewbel (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. In contrast to the other examples, here 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 again asymptotically negligible.

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 econometric applications, thus illustrating their usefulness.

A Mathematical Appendix

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 $\bar{\mathcal{M}}_n = \bar{\mathcal{M}}_{n,1} \times \ldots \times \bar{\mathcal{M}}_{n,d}$.

A.1 Proof of Theorem 1

In order to prove the statement of the theorem, we have to introduce some notation. First, the real estimator $\widehat{m}_{LL}$ can be written as

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

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

$$
\begin{aligned}
(\widehat{\alpha}_A, \widehat{\beta}_A) &= \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (\varepsilon_i - \alpha - \beta^T(\widehat{r}(S_i) - x))^2 K_h(\widehat{r}(S_i) - x), \\
(\widehat{\alpha}_B, \widehat{\beta}_B) &= \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (m_0(r_0(S_i)) - m_0(x) - m_0'(x)^T(r_0(S_i) - x) \\
&\quad - \alpha - \beta^T(\widehat{r}(S_i) - x))^2 K_h(\widehat{r}(S_i) - x), \\
(\widehat{\alpha}_C, \widehat{\beta}_C) &= \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (-m_0'(x)^T(\hat{r}(S_i) - r_0(S_i)) - \alpha - \beta^T(\widehat{r}(S_i) - x))^2 K_h(\widehat{r}(S_i) - x), \\
(\widehat{\alpha}_D, \widehat{\beta}_D) &= \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (m_0'(x)^T(\widehat{r}(S_i) - x) - \alpha - \beta^T(\widehat{r}(S_i) - x))^2 K_h(\widehat{r}(S_i) - x).
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
(\widetilde{\alpha}_A, \widetilde{\beta}_A) &= \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (\varepsilon_i - \alpha - \beta^T(r_0(S_i) - x))^2 K_h(r_0(S_i) - x), \\
(\widetilde{\alpha}_B, \widetilde{\beta}_B) &= \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (m_0(r_0(S_i)) - m_0(x) - m_0'(x)^T(r_0(S_i) - x) \\
&\quad - \alpha - \beta^T(r_0(S_i) - x))^2 K_h(r_0(S_i) - x), \\
(\widetilde{\alpha}_D, \widetilde{\beta}_D) &= \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (m_0'(x)^T(r_0(S_i) - x) - \alpha - \beta^T(r_0(S_i) - x))^2 K_h(r_0(S_i) - x).
\end{aligned}
$$

Finally, we set $\widetilde{m}_{LL,C}(x) = m_0'(x)\hat{\Delta}(x)$. Note that by construction

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

We now argue that

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

For a proof of (A.2) note that $\widehat{m}_{LL,A}(x)$ and $\widehat{m}_{LL,A}(x)$ are given by the first elements of the vectors $\widehat{M}(x)^{-1} n^{-1} \sum_{i=1}^{n} K_h(\widehat{r}(S_i) - x)\varepsilon_i \widehat{w}_i(x)$ or $\widetilde{M}(x)^{-1} n^{-1} \sum_{i=1}^{n} K_h(r_0(S_i) - x)\varepsilon_i \widetilde{w}_i(x)$, respectively, where $\widehat{w}_i(x)$ and $\widetilde{w}_i(x)$ are the vectors with elements $1, (\widehat{r}_1(S_i) - x_1)/h_1, ..., (\widehat{r}_d(S_i) - x_d)/h_d$ or $1, ..., (r_{0,d}(S_i) - x_d)/h_d$, respectively. Furthermore, we have put $\widehat{M}(x) = n^{-1} \sum_{i=1}^{n} \widehat{w}_i(x)\widehat{w}_i(x)^T \; K_h(\widehat{r}(S_i) - x)$ and $\widetilde{M}(x) = n^{-1} \sum_{i=1}^{n} \widetilde{w}_i(x)\widetilde{w}_i(x)^T K_h(r_0(S_i) - x)$. Using these representations of $\widehat{m}_{LL,A}(x)$ and $\widehat{m}_{LL,A}(x)$ one sees that (A.2) follows from Lemma 2 and 3 below.

From Lemmas 3 and 4 we get that

$$
\sup_{x \in I_R} |\widehat{m}_{LL,B}(x) - \widetilde{m}_{LL,B}(x)| = O_p(n^{-\kappa_2}), \tag{A.3}
$$

$$
\sup_{x \in I_R} |\widehat{m}_{LL,C}(x) - \widetilde{m}_{LL,C}(x)| = O_p(n^{-\kappa_3} + n^{-\kappa_1}). \tag{A.4}
$$

Taken together, the results in (A.1)–(A.4) imply the statement of the theorem. □

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

$$\sup_{x\in I_R, r_1, r_2\in\bar{\mathcal{M}}_n} |\frac{1}{n}\sum_{i=1}^n K_h(r_1(S_i)-x)\varepsilon_i - \frac{1}{n}\sum_{i=1}^n K_h(r_2(S_i)-x)\varepsilon_i| = O_p(n^{-\kappa_1})$$

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

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

$$\kappa_1 > (\delta-\eta)_{\min}. \tag{A.5}$$

If $\kappa_1 \leq (\delta-\eta)_{\min}$ 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

$$\Pr(\max_i |\varepsilon_i| > C_\varepsilon \log(n)) \leq n^{-C_1}, \tag{A.6}$$

$$|\mathbb{E}\varepsilon_i \mathbb{I}\{|\varepsilon| \leq C_\varepsilon \log(n)\}| \leq n^{-C_2}. \tag{A.7}$$

With this choice of C_ε we define

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

with

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

Now for $s \geq 0$, 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 $\|r - r^*\|_\infty \leq 2^{-s}n^{-\delta_j}$. 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 $\bar{\mathcal{M}}^*_{s,n} = \bar{\mathcal{M}}^*_{s,n,1} \times \ldots \times \bar{\mathcal{M}}^*_{s,n,d}$.

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 $\|r^s_{1,j} - r_{1,j}\|_\infty \leq 2^{-s}n^{-\delta_j}$ and $\|r^s_{2,j} - r_{2,j}\|_\infty \leq C2^{-s}n^{-\delta_j}$, for all j. We then consider the chain

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

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

$$T_1 = |\frac{1}{n}\sum_{i=1}^n \Delta_i(r_l^{G_n}, r_l)| \leq C\log(n)2^{-G_n}n^{-(\delta-\eta)_{\min}} \leq Cn^{-\kappa_1}. \tag{A.8}$$

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

$$\begin{aligned}
&\Pr(\sup_{r_1 \in \bar{\mathcal{M}}_n} |\frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{G_n} \Delta_i(r_1^{s-1}, r_1^s)| > n^{-\kappa_1}) \\
&\leq \sum_{s=1}^{G_n} \Pr(\sup_{r_1 \in \bar{\mathcal{M}}_n} |\frac{1}{n}\sum_{i=1}^{n} \Delta_i(r_1^{s-1}, r_1^s)| > c_a 2^{-as} n^{-\kappa_1}) \\
&\leq \sum_{s=1}^{G_n} \#\bar{\mathcal{M}}^*_{s-1,n} \#\bar{\mathcal{M}}^*_{s,n} \Pr(\frac{1}{n}\sum_{i=1}^{n} \Delta_i(r_1^{*,s}, r_1^{**,s}) > c_a 2^{-as} n^{-\kappa_1}) \\
&\quad + \sum_{s=1}^{G_n} \#\bar{\mathcal{M}}^*_{s-1,n} \#\bar{\mathcal{M}}^*_{s,n} \Pr(\frac{1}{n}\sum_{i=1}^{n} \Delta_i(\tilde{r}_1^{*,s}, \tilde{r}_1^{**,s}) < c_a 2^{-as} n^{-\kappa_1}) \\
&= T_2 + T_3
\end{aligned}$$

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

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

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

$$T_2 \leq \exp(-cn^c), \tag{A.9}$$

$$T_3 \leq \exp(-cn^c). \tag{A.10}$$

We only show (A.9), as the statement (A.10) follows by essentially the same arguments. Using Assumption 3, we obtain by application of the Markov inequality that

$$\begin{aligned}
T_2 &\leq C \sum_{s=1}^{G_n} \prod_j \exp((2^{-s} n^{-\delta_j})^{-\alpha_j} n^{\xi_j}) \mathbb{E}(\exp(\gamma_{n,s} \frac{1}{n}\sum_{i=1}^{n} \Delta_i(r_1^{*,s}, r_1^{**,s}) - \gamma_{n,s} c_a 2^{-as} n^{-\kappa_1})) \\
&\leq C \sum_{s=1}^{G_n} \exp(\sum_j 2^{s\alpha_j} n^{\delta_j \alpha_j + \xi_j} - \gamma_{n,s} c_a 2^{-as} n^{-\kappa_1}) \prod_{i=1}^{n} \mathbb{E}(\exp(\gamma_{n,s} \frac{1}{n} \Delta_i(r_1^*, r_1^{**}))) \qquad \text{(A.11)}
\end{aligned}$$

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.11) can be bounded as follows:

$$\begin{aligned}
\mathbb{E}(\exp(\gamma_{n,s} \frac{1}{n} \Delta_i(r_1^*, r_1^{**}))) &\leq 1 + C\mathbb{E}(\gamma_{n,s}^2 n^{-2} \Delta_i^2(r_1^*, r_1^{**})) \\
&\leq \exp(C\gamma_{n,s}^2 n^{-2} n^{\eta_+ - 2(\delta-\eta)_{\min}} 2^{-2s}), \qquad \text{(A.12)}
\end{aligned}$$

where we have used that

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

for n large enough because of (A.5). Inserting (A.12) into (A.11), we obtain, if a and c_γ were chosen sufficiently small, that

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

Finally, it follows from a simple argument that

$$T_4 = \Pr(\sup_{r_1, r_2 \in \bar{\mathcal{M}}_n} |\frac{1}{n} \sum_{i=1}^n \Delta_i(r_1^0, r_2^0)| > n^{-\kappa_1}) \leq \exp(-cn^c) \tag{A.13}$$

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

From (A.8), (A.9), (A.10) and (A.13), we thus obtain that

$$\sup_{x \in I_R} \Pr(\sup_{r_1, r_2 \in \bar{\mathcal{M}}_n} |\frac{1}{n} \sum_{i=1}^n K_h(r_1(S_i) - x)\varepsilon_i^* - \frac{1}{n} \sum_{i=1}^n K_h(r_2(S_i) - x)\varepsilon_i^*| > Cn^{-\kappa_1}) \leq \exp(-cn^c) \tag{A.14}$$

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 $\|x - x^*\| \leq n^{-cC_I}$. If C_I is chosen large enough, this implies that

$$\sup_{x \in I_R} \sup_{r \in \bar{\mathcal{M}}_n} |\frac{1}{n} \sum_{i=1}^n K_h(r(S_i) - x)\varepsilon_i - \frac{1}{n} \sum_{i=1}^n K_h(r(S_i) - x^*)\varepsilon_i| \leq n^{-\kappa_1} \tag{A.15}$$

for large enough n, with probability tending to one. Furthermore, it follows from (A.14) that

$$\sup_{x \in I_{R,n}} \sup_{r_1, r_2 \in \bar{\mathcal{M}}_n} |\frac{1}{n} \sum_{i=1}^n K_h(r_1(S_i) - x)\varepsilon_i - \frac{1}{n} \sum_{i=1}^n K_h(r_2(S_i) - x)\varepsilon_i| \leq n^{-\kappa_1}. \tag{A.16}$$

The statement of the lemma then follows from (A.6)–(A.7) and (A.15) – (A.16), if the constants C_1 and C_2 were chosen large enough. □

Lemma 3. *Suppose that the conditions of Theorem 1 hold. Then*

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

for $j, l = 1, \ldots, q$ $j \neq l$ and $0 \leq a + b \leq 2$, $0 \leq a, b$.

Proof. The lemma follows from

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

for $r_1, r_2 \in \bar{\mathcal{M}}_n$ and the fact that

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

□

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

Lemma 4. *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*

$$\sup_{x \in I_R, 1 \leq i \leq n} |[m_0(r_0(S_i)) - m_0(x) - m_0'(x)^T (r_0(S_i) - x)] I_i(x)| \leq R_n n^{-2\eta_{\min}}, \tag{A.17}$$

$$\sup_{x \in I_R} \|\frac{1}{n} \sum_{i=1}^{n} K_h(\hat{r}(S_i) - x) \widehat{w}_i(x) \widehat{w}_i(x)^T - \frac{1}{n} \sum_{i=1}^{n} K_h(r_0(S_i) - x) \widetilde{w}_i(x) \widetilde{w}_i(x)^T \| \leq R_n n^{-(\delta - \eta)_{\min}}, \tag{A.18}$$

$$\sup_{x \in I_R} \|\frac{1}{n} \sum_{i=1}^{n} K_h(r_0(S_i) - x) \widetilde{w}_i(x) \widetilde{w}_i(x)^T - B_K\| \leq R_n (n^{-\eta_{min}} + n^{-(1-\eta_+)/2} \sqrt{\log n}). \tag{A.19}$$

Proof. Claim (A.17) follows by a simple calculation. Claim (A.18) is a direct consequence of Lemma 3. And (A.19) follows from standard arguments from kernel smoothing theory. For the stochastic part one makes use of Lemma 5. □

A.2 Proof of Lemma 1

In order to prove Lemma 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 $g_1 = ... = g_p$ and we write g for this joint value and for the vector $g = (g, ..., g)$.

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 $(0, \ldots, 0, i)$ is the first element in the sequence and $(i, 0, \ldots, 0)$ the last element). Let τ_i denote this one-to-one mapping, i.e. $\tau_i(1) = (0, \ldots, 0, i)$, ..., $\tau_i(N_i) = (i, 0 \ldots, 0)$. For each $i = 1, \ldots, q$, define a $N_i \times 1$ vector $\mu_i(x)$ with its kth element given by $x^{\tau_i(k)}$, and write $\mu(x) = (1, \mu_1(x)^T, \ldots, \mu_q(x)^T)^T$, 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 $0 \leq j, k \leq q$, 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

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

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

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

Finally, denote the first unit q-vector by $e_1 = (1, 0, \ldots, 0)$. With this notation, it can be shown along classical lines (e.g. Masry 1996) that the local polynomial estimator $\hat{r}$ admits the following stochastic

expansion:

$$\hat{r}(s) - r_0(s) = \frac{1}{n}\sum_{i=1}^{n} e_1 M_{nq}^{-1}(s)\mu((S_i - s)/g)L_g(S_i - s)\zeta_i + g^{q+1}B_n(s) + R_n(s), \tag{A.20}$$

where, B_n is a bias term that satisfies

$$B_n(s) = \frac{1}{(q+1)!}e_1 M_q^{-1} A_q r_0^{(q+1)}(s) + o_p(1) \equiv b(s) + o_p(1), \tag{A.21}$$

with $A_q = [M_{0,q+1}, M_{1,q+1}, \ldots, M_{q,q+1}]^T$, and R_n is a remainder term which satisfies

$$\sup_{s\in I_S} |R_n(s)| = O_p\left(\log(n)/ng^p\right).$$

The value of the expansion (A.20) is that this remainder term be made to be as small as $o_p(n^{-1/2})$ by using an appropriate bandwidth g. When the function r_0 is sufficiently smooth, and a local polynomial of appropriate order is used, the corresponding bias term is of smaller order than the remainder, and thus asymptotically negligible. We remark that Kong, Linton, and Xia (2009) have recently shown the validity of expansions analogous to the one presented in (A.20) for more general local polynomial M-regressions and certain time series frameworks.

To prove the lemma, define the stochastic component and the bias term of the expansion (A.20) as $\widehat{r}_A(s) = n^{-1}\sum_{i=1}^{n} e_1 M_{nq}^{-1}(s)\mu((S_i - s)/g)L_g(S_i - s)\zeta_i$ and $\widehat{r}_B(s) = g^{2k}B_n(s)$, respectively. Now the function $\hat{\Delta}$ can be written as

$$\begin{aligned}\hat{\Delta}(x) &= \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)\widehat{r}_A(S_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)} + \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)\widehat{r}_B(S_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)} + O_p\left(\frac{\log(n)}{ng^p}\right) \\ &\equiv \hat{\Delta}_A(x) + \hat{\Delta}_B(x) + O_p\left(\frac{\log(n)}{ng^p}\right),\end{aligned}$$

uniformly over $x \in I_{R,n}^{-}$. 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

$$\begin{aligned}\hat{\Delta}_B(x) &= g^{2k}\frac{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)b(S_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)} + o_p(g^{2k}) \\ &= g^{2k}\mathbb{E}(b(S)|r_0(S) = x) + o_p(g^{2k})\end{aligned}$$

because the function $\mathbb{E}(b(S)|r_0(S) = x)$ is continuous with respect to x, see Assumption (A4) (ii).

Next, consider the term $\hat{\Delta}_A(x)$. Using standard arguments from e.g. Masry (1996), we obtain that

$$\begin{aligned}\hat{\Delta}_A(x) &= \frac{1}{n^2\hat{f}_R(x)}\sum_{j=1}^{n}\zeta_j\sum_{i=1}^{n} K_h(r_0(S_i) - x)e_1 M_{nq}^{-1}(S_i)\mu((S_j - S_i)/g)L_g(S_j - S_i) \\ &= \frac{1}{nf_R(x)}\sum_{j=1}^{n}\psi(x, S_j)\zeta_j + O_p\left(\frac{\log(n)}{(n^2hg^p)^{1/2}}\right)\end{aligned}$$

with $\psi(x,s) = \mathbb{E}\left(f_S(S)^{-1}K_h(r_0(S) - x)e_1 M_q^{-1}\mu((S_j - S)/g)L_g(S_j - S)|S_j = s\right)$ for $s \in I_{S,n}^{-}$, where the set $I_{S,n}^{-}$ contains all $s \in I_S$ with the property that their k-th element s_k does not lie in a g_k-neighborhood of the boundary of $I_{S,k}$ for $k = 1, ..., p$. This holds since $M_{n,q}(s)$ converges to $f_S(s)M_q$ uniformly for s

in $I_{S,n}^-$. For $s \in I_{S,n}^-$ this can be written as $\psi(x,s) = \int K_h(r_0(u)-x)L_g^*(u-s)du$ the modified kernel L^* is defined by

$$L^*(t) = e_1 M_q^{-1} \mu(t) L(t). \tag{A.22}$$

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. For $s \notin I_{S,n}^-$ we get that $\psi(x,s) = \int K_h(r_0(u)-x)L_g^*(u,u-s)du$ with a kernel $L^*(u,t)$ that has the same moment conditions in t but depends on u. We thus obtain that

$$\hat{\Delta}_A(x) = \frac{1}{nf_R(x)} \sum_{j=1}^{n} \psi(x, S_j)\zeta_j + O_p(\log(n)/(n^2 h g^p)^{1/2}). \tag{A.23}$$

We now derive explicit expressions for the leading term in equation (A.23) for the cases a)–c) of the Lemma. Starting with case a), for which $g/h \to 0$, it follows by substitution and Taylor expansion arguments that for $v \in I_{S,n}^-$ with $K_h'(v) = h^{-1}K'(h^{-1}v)$ and $K_h''(v) = h^{-1}K''(h^{-1}v)$

$$\begin{aligned}
\psi(x,v) &= \int K_h(r_0(s)-x)L_g^*(s-v)ds \\
&= \int K_h(r_0(v+tg)-x)L^*(t)dt \\
&= \int (K_h(r_0(v)-x) + K_h'(r_0(v)-x)\frac{r_0(v+tg)-r_0(v)}{h} \\
&\quad + K_h''(\chi_1 - x)\frac{1}{2}\left(\frac{r_0(v+tg)-r_0(v)}{h}\right)^2)L^*(t)dt \\
&= K_h(r_0(v)-x) + K_h'(r_0(v)-x)\int (r_0'(v)\frac{tg}{h} + r_0''(\chi_2)\frac{t^2g^2}{2h})L^*(t)dt \\
&\quad + \int K_h''(\chi_1 - x)\frac{1}{2}\left(\frac{r_0'(\chi_3)tg}{h}\right)^2 L^*(t)dt,
\end{aligned}$$

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 $\psi(x,v)$ of order $(h/g)^2$. For $v \notin I_{S,n}^-$ one gets an expansion of order h/g. Together with Lemma 5 in Appendix B, we thus obtain that

$$\frac{1}{nf_R(x)}\sum_{j=1}^{n}\psi(x,S_j)\zeta_j = \frac{1}{nf_R(x)}\sum_{j=1}^{n}K_h(r_0(S_j)-x)\zeta_j + O_p((\frac{g}{h})^2\sqrt{\frac{\log(n)}{nh}}),$$

as claimed. To show statement b) of the Lemma, we rewrite the function ψ for $v \in I_{S,n}^-$ as follows:

$$\begin{aligned}
\psi(x,v) &= \int (K_h(r_0(v)-x+r_0'(v)tg) + K'(\frac{\chi_1}{g})r_0''(\chi_2)\frac{1}{2}t^2)L^*(t)dt \\
&= M_h(x,v) + g\int K_g'(\zeta_1)r_0''(\chi_2)\frac{1}{2}t^2L^*(t)dt
\end{aligned}$$

where χ_1 is an intermediate value between $r_0(v+gt)$ and $r_0(v)+r_0'(v)tg$, 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 B that

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

which implies the desired result. Now consider statement c) of the Lemma. In this case, where $g/h \to \infty$, we can rewrite the function $\psi(\cdot)$ for $v \in I_{S,n}^-$ as follows:

$$\begin{aligned}\psi(x,v) &= \int K_h(r_0(s)-x)L_g^*(s-v)ds \\ &= \frac{1}{g}\int K(t)L^*\left(u,(\varphi(v_1+ug,x+th)-v_2)/g\right)\partial_x\varphi(v_1+ug,x+th)dtdu\end{aligned}$$

The statement of the Lemma then follows from tedious but conceptionally simple Taylor expansion arguments similar to the ones employed for case b), and Lemma 5.

A.3 Proofs of Theorems 2–5.

The statements of these theorems follow by direct application of Lemma 1 and Theorem 1. The statement of Theorem 2 is immediate. For Theorem 3–5, we only have to check that the error bounds in Theorem 1 and Lemma 1 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,...,S_n. Then the only randomness in the pilot estimation comes from $\zeta_1, ..., \zeta_n$. We can decompose $\widehat{r}$ into $\widehat{r}_A + \widehat{r}_B$, where $\widehat{r}_A$ is the local polynomial fit to (S_i, ζ_i) and $\widehat{r}_B$ is the local polynomial fit to $(S_i, r_0(S_i))$. Conditionally given S_1,...,S_n, the value of $\widehat{r}_B$ is fixed and for checking Assumption 3 we only have to consider entropy conditions for sets of possible outcomes of $\widehat{r}_A$. 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 $\alpha \leq 2$ because of Assumption 4(iii). It can be easily checked that we get the desired expansions in Theorems 2 and 3 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 Lemma 1. It remains to check Assumption 3. It suffices to check the entropy conditions for the tuple of functions $(n^{-1}\sum_{i=1}^n L_h(S_i - s)[(S_i - s)/g]^\pi \zeta_i : 0 \leq \pi_+ \leq q, \pi_j \geq 0 \text{ for } j = 1, ..., p)$. This follows because we get $\widehat{r}_A$ 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 $B_n \leq b_n = n^{\xi^{**}})$ 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 $A_n \leq a_n = n^{\xi^*})$ 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 $\lambda \leq a_n$. 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 $\lambda \leq 1$. The desired entropy bound follows by rescaling of the functions. Note that we have that $b_n^{-1}a_n \to 0$.

A.4 Proof of Theorem 6

For $x \in I_R$ we can decompose $Y_i = Y_{i,A}+Y_{i,B}(x)+...+Y_{i,G}(x)$, where $Y_{i,A} = \varepsilon_i$, $Y_{i,B}(x) = m_0'(x)^T(\widehat{r}(S_i)-x)$, $Y_{i,C}(x) = m_0(x) + \frac{1}{2}(\widehat{r}(S_i)-x)^T m_0''(x)(\widehat{r}(S_i)-x)$, $Y_{i,D}(x) = m_0(r_0(S_i)) - m_0(x) - m_0'(x)^T(r_0(S_i)-x) - \frac{1}{2}(r_0(S_i)-x)^T m_0''(x)(r_0(S_i)-x)$, $Y_{i,E}(x) = -m_0'(r_0(S_i))^T(\widehat{r}(S_i)-r_0(S_i))$, $Y_{i,F}(x) = (m_0'(r_0(S_i)) - m_0'(x) - m_0''(x)(r_0(S_i)-x))^T(\widehat{r}(S_i)-r_0(S_i))$, and $Y_{i,G}(x) = -\frac{1}{2}(\widehat{r}(S_i)-r_0(S_i))^T m_0''(x)(\widehat{r}(S_i)-r_0(S_i))$. The decomposition of Y_i defines an additive decomposition of $\widehat{m}^*_{LQ}(x)$ into $\widehat{m}^*_{LQ,A}(x) + ... + \widehat{m}^*_{LQ,G}(x)$. Similarly, by decomposing $Y_i = Y^\#_{i,A}+Y^\#_{i,B}(x)+...+Y^\#_{i,D}(x)$ we get $\widetilde{m}^*_{LQ}(x) = \widetilde{m}^*_{LQ,A}(x)+...+\widetilde{m}^*_{LQ,D}(x)$. In the latter decomposition we have chosen $Y^\#_{i,A} = Y_{i,A}$, $Y^\#_{i,B} = m_0'(x)^T(r_0(S_i)-x)$, $Y^\#_{i,C}(x) = m_0(x) + \frac{1}{2}(r_0(S_i)-x)^T m_0''(x)(r_0(S_i)-x)$, and $Y^\#_{i,D}(x) = Y_{i,D}(x)$.

Now, we compare these two additive decompositions. The difference $\widehat{m}^*_{LQ,A}(x) - \widetilde{m}^*_{LQ,A}(x)$ can be treated as in the first part of the proof of Theorem 1 by application of empirical process methods. It is helpful to multiply the j-th element of $\widehat{m}^*_{LQ,A}(x)$ and $\widetilde{m}^*_{LQ,A}(x)$ by $n^{-\eta_j}$. Then, for these new vectors the whole analysis of the first part of Theorem 1 goes through without changing any exponential constants.

It remains to compare the other additive components. First, we have $\widehat{m}^*_{LQ,B}(x) = \widetilde{m}^*_{LQ,B}(x) = m_0'(x)$ and $\widehat{m}^*_{LQ,C}(x) = \widetilde{m}^*_{LQ,C}(x) = 0$ by definition. Furthermore, one can easily check that $Y^\#_{i,D}(x) = Y_{i,D}(x)$ is uniformly in x bounded by $O(n^{-3\eta_{min}})$. By some algebra this results in a uniform bound for $\widehat{m}^*_{j,LQ,D}(x) - \widetilde{m}^*_{j,LQ,D}(x)$ of the order $O(n^{-3\eta_{min}+\eta_j-(\delta-\eta)_{min}})$. The terms $\widehat{m}^*_{LQ,F}(x)$ and $\widehat{m}^*_{LQ,G}(x)$ can be bounded by using uniform bounds on $Y_{i,F}(x)$ and $Y_{i,G}(x)$. Making use of all these results we get that (4.1) follows from the fact that $\Delta^*(x) = \widehat{m}^*_{LQ,E}(x)$. Equation (4.1) follows with a classical smoothing argument. □

A.5 Proof of Theorem 7

The proof is analogous to Theorem 1 over increasing subsets. Direct calculations show that $I^*_{R,n}$ is appropriately chosen.

A.6 Proof of Theorem 8

The proof is similar to the one of Theorem 1, but uses more direct arguments to show a result analogous to Lemma 2.

A.7 Proof of Theorem 9

We can write $\widehat{\theta} = \widehat{\theta}_A + \widehat{\theta}_B$ with

$$\begin{aligned}\widehat{\theta}_A &= \sum_{l=1}^{L} \frac{n_l}{n} \frac{1}{n_l} \sum_{i \in N_l} \widehat{w}_l^*(\widehat{r}_l(S_i))\varepsilon_i, \\ \widehat{\theta}_B &= \sum_{l=1}^{L} \frac{n_l}{n} \frac{1}{n_l} \sum_{i \in N_l} \widehat{w}_l^*(\widehat{r}_l(S_i))m(r(S_i)).\end{aligned}$$

We first show that $\widehat{\theta}_A = O_P(n^{-1/2})$. This claim immediately follows from

$$\widehat{\theta}_A = n^{-1} \sum_{i=1}^{n} w(r(S_i))\varepsilon_i + o_P(n^{-1/2}). \tag{A.24}$$

For a proof of (A.24) we consider the conditional variance of

$$\sqrt{n_l}\left(\frac{1}{n_l} \sum_{i \in N_l} \widehat{w}_l^*(\widehat{r}_l(S_i))\varepsilon_i - \frac{1}{n_l} \sum_{i \in N_l} w^*(r(S_i))\varepsilon_i\right),$$

given the functions $\widehat{w}_l^*$ and $\widehat{r}_l$ and the values of S_i for $i \in N_l$. This conditional variance is bounded by $C_\varepsilon n_l^{-1} \sum_{i \in N_l} [\widehat{w}_{l,h}^*(\widehat{r}_l(S_i)) - w^*(r(S_i))]^2$. Because of Assumption 8(iii) this bound is of order $o_P(1)$. This shows (A.24).

It remains to show $\widehat{\theta}_B - \theta = O_P(n^{-1/2})$. This claim can be shown by calculating the conditional variance and expectation of $\frac{1}{n_l} \sum_{i \in N_l} \widehat{w}_l^*(\widehat{r}_l(S_i))m(r(S_i))$, given the functions $\widehat{w}_l^*$ and $\widehat{r}_l$.

A.8 Proof of Proposition 1

Let $\hat{f} = (\hat{\nu}_1, \hat{\nu}_0, \hat{\Pi})$ and $\bar{f} = (\nu_1, \nu_0, \Pi)$, define the functional $S_n(f)$ as

$$S_n(f) = \frac{1}{n} \sum_{i=1}^{n} f_1(f_3(X_i)) - f_2(f_3(X_i)) - \gamma_{ATE},$$

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 $f = (f_{1,A} + f_{1,B}, f_2, f_3)$ we have that

$$\begin{aligned}&\|S_n(f) - S_n(\bar{f}) - \dot{S}_n(\bar{f})[f - \bar{f}] - ((f'_{1,A} - \bar{f}'_1) + (f'_{2,A} - \bar{f}'_2))(f_3 - \bar{f}_3)\|_\infty \\ &\quad = O(\|f_3 - \bar{f}_3\|_\infty^2(\|f''_{1,A}\|_\infty + \|f''_{2,A}\|_\infty)) + O(\|f_3 - \bar{f}_3\|_\infty^2) + O(\|f_{1,B}\|_\infty + \|f_{2,B}\|_\infty).\end{aligned}$$

Now set $\hat{f}_{1,A}$ equal to the leading terms of a stochastic expansion of $\hat{\nu}_1$ up to order $o_p(n^{-1/2})$ (analogous to the one given in Theorem 5, but acommodating the presence of the indicator variable D), let $\hat{f}_{1,B} = \hat{f}_1 - \hat{f}_{1,A} = o_p(n^{-1/2})$ be the corresponding remainder term, and define $\hat{f}_{2,A}, \hat{f}_{2,B}$ analogously. Since the conditions of the proposition imply that $\|\hat{f}_3 - \bar{f}_3\|_\infty = o_p(n^{-1/4})$ and $\|\hat{f}''_{j,A}\|_\infty = O_p(1)$ for $j = 1, 2$, we have that

$$\hat{\gamma}_{ATE} - \gamma_{ATE} = S_n(\hat{f}) = S_n(\bar{f}) + T_{1,n} + T_{2,n} + T_{3,n} + T_{4,n} + o_p(n^{-1/2}),$$

where

$$T_{1,n} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\nu}_1(\Pi(X_i)) - \nu_1(\Pi(X_i))),$$
$$T_{2,n} = -\frac{1}{n}\sum_{i=1}^{n}(\hat{\nu}_0(\Pi(X_i)) - \nu_0(\Pi(X_i))),$$
$$T_{3,n} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\Pi}(X_i) - \Pi(X_i))(\nu_1'(\Pi(X_i)) - \nu_0'(\Pi(X_i))),$$
$$T_{4,n} = \frac{1}{n}\sum_{i=1}^{n}((\hat{f}_{1,A}'(\Pi(X_i)) - \nu_1(\Pi(X_i))) + (\hat{f}_{2,A}'(\Pi(X_i)) - \nu_0(\Pi(X_i)))(\hat{\Pi}(X_i) - \Pi(X_i))$$

To prove the asymptotic normality result, we show that

$$\sqrt{n}(S_n(\bar{f}) + T_{1,n} + T_{2,n} + T_{3,n} + T_{4,n}) \xrightarrow{d} N(0, \mathbb{E}(\psi(Y, D, X)^2)).$$

First, note that the term $S_n(\bar{f})$ is simply the sample average of i.i.d. mean zero random variables, and thus easy to handle. Now consider the term $T_{1,n}$, which can be rewritten as $T_{1,n} = T_{1,n}^A + T_{1,n}^B$, where

$$T_{1,n}^A = \frac{1}{n}\sum_{i=1}^{n}(\hat{\nu}_1(\Pi(X_i)) - \nu_1(\Pi(X_i)))\mathbb{I}\{\Pi(X_i) \in I_{P,n}^-\}$$
$$T_{1,n}^B = \frac{1}{n}\sum_{i=1}^{n}(\hat{\nu}_1(\Pi(X_i)) - \nu_1(\Pi(X_i)))\mathbb{I}\{\Pi(X_i) \notin I_{P,n}^-\}$$

Using the stochastic expansion in Theorem 5 and projection arguments for U-Statistics (Powell, Stock, and Stoker 1989, Lemma 3.1), it follows that

$$T_{1,n}^A = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \nu_1(\Pi(X_i)))}{\Pi(X_i)} - \nu_1'(\Pi(X_i))(D_i - \Pi(X_i))\right) + o_p(n^{-1/2}).$$

By deriving an expansion analogous to the one obtained in Theorem 5 for the boundary regions, one can furthermore show that $T_{1,n}^B = o_P(n^{-1/2})$. Taken together, this shows that

$$T_{1,n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \nu_1(\Pi(X_i)))}{\Pi(X_i)} - \nu_1'(\Pi(X_i))(D_i - \Pi(X_i))\right) + o_p(n^{-1/2}). \tag{A.25}$$

Using the same line of reasoning, we also find that

$$T_{2,n} = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1 - D_i)(Y_i - \nu_0(\Pi(X_i))}{1 - \Pi(X_i)} - \nu_0'(\Pi(X_i))(D_i - \Pi(X_i))\right) + o_p(n^{-1/2}). \tag{A.26}$$

Now consider the term $T_{3,n}$. Using again standard projection arguments for U-Statistics and a stochastic expansion for the estimated propensity score $\hat{\Pi}(x)$ analogous to the one used in the proof of Lemma 1, one can show that

$$T_{3,n} = \frac{1}{n}\sum_{i=1}^{n}(D_i - \Pi(X_i))(\nu_1'(\Pi(X_i)) - \nu_0'(\Pi(X_i))) + o_p(n^{-1/2}). \tag{A.27}$$

Finally, by using again the stochastic expansions from Theorem 5 and the stochastic expansion of the estimated propensity score $\hat{\Pi}(x)$ mentioned before, one can show that $T_{4,n}$ is equal third order U-statistic

up to terms of order $o_P(n^{-1/2})$. This leading U-Statistic turns out to be degenerate, and we thus find that

$$T_{4,n} = o_P(n^{-1/2}) \tag{A.28}$$

by applying Lemma A.3 in Ahn and Powell (1993). The statement of the proposition then follows from (A.25)–(A.28) and an application of the central limit theorem.

A.9 Proof of Proposition 2

Let $\hat{f} = (\hat{m}, \hat{\mu}_2)$ and $\bar{f} = (m, \mu_2)$, define the functional $S_n(f)$ as

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

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 $f = (f_{1,A} + f_{1,B}, f_2, f_3)$ with bounded second derivatives we have that

$$\|S_n(f) - S_n(\bar{f}) - \dot{S}_n(\bar{f})[f - \bar{f}]\|_\infty \leq O(\|f_2 - \bar{f}_2\|_\infty^2) + O(\|f_2 - \bar{f}_2\|_\infty \|f_{1,A}^{(v)} - \bar{f}_1^{(v)}\|_\infty) + O(\|f_{1,B}\|_\infty)$$

where $f_{1,A}^{(v)}(x_1, z_1, v) = df_{1,A}(x_1, z_1, v)/dv$. Using the same kind of arguments as in the proof of Lemma 1, under the conditions of the proposition one can derive the following uniform stochastic expansion of $\hat{m}$ up to order $o_p((nh^{1+d_1})^{-1/2})$:

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

where $\varepsilon_i = Y - m(X_{1i}, Z_{1i}, V_i)$. Let $\hat{f}_{1,A}$ denote the two leading terms of this expansion, and denote the remainder term by $\hat{f}_{1,B}$. Now it follows from e.g. Masry (1996) and the conditions on η and θ that

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

and it follows from the same result together with Lemma 5 in Appendix B that

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

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

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

$$T_{2,n} = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}(x_1, z_1, V_i) - m(x_1, z_1, V_i))$$

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 Lemma 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

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

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 projection arguments for U-Statistics. □

A.10 Proof of Proposition 3

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

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

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

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

since $s_0(x) = q_0(r_0(x))$. To prove the proposition, 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 Proposition, 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 sup and inf are taken over $s \in (r_o(x) - \epsilon, \lambda_0 + \epsilon)$ for some $\epsilon > 0$. 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

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

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 $w(s) = \sum_{i=1}^n K_h(r_0(X_i) - s)(\hat{r}(X_i) - r_0(X_i))/\sum_{i=1}^n K_h(r_0(X_i) - s)$. Using similar arguments

as in the proof of Lemma 1 and the other propositions, and the restriction that $\underline{\theta} < \theta < \bar{\theta}$, we obtain that

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

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

$$\begin{aligned}\int_{r(x)}^{\lambda} \frac{q_0'(s)w(s)}{q_0(s)^2} ds &= \frac{1}{n} \sum_{i=1}^{n} \zeta_i \frac{q_0'(r_0(X_i))}{q_0(r_0(X_i))^2 f_R(r_0(X_i)) f_X(X_i)} + O_p\left(\frac{\log n}{ng^p}\right) + O_p(g^{q+1}) \\ &= o_p((ng^p)^{-1/2}),\end{aligned}$$

for $\zeta_i = Y_i - r_0(X_i)$. 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. □

B Additional Results

B.1 Uniform Rates for Generalized Kernels

The following 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 e.g. (Härdle, Jansen, 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 2 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 $0 < b \leq |b_n| \leq B < \infty$, 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 $|W_{n,h}(u,z) - W_{n,h}(v,z)| \leq l\frac{||u-t(z)||}{h} h^{-d_x} \widetilde{W}_n(v, t(z))$ with $\sup_n \widetilde{W}_n$ bounded, and that $\mathbb{E}[\exp(\rho|\varepsilon|)|S] < C$ a.s. for a constant $C > 0$ and $\rho > 0$ small enough. Then with a deterministic sequence a_n with $|a_n| \leq A$ we have that*

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

References

Ahn, H., and J. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58(1-2), 3–29.

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

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

Carneiro, P., J. Heckman, and E. Vytlacil (2009a): "Estimating marginal and average returns to education," *Unpublished manuscript, University of Chicago*.

Carneiro, P., J. Heckman, and E. Vytlacil (2009b): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*.

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

Conrad, C., and E. Mammen (2009): "Nonparametric regression on a generated covariate with an application to semiparametric GARCH-in-Mean models," *Unpublished manuscript, University of Mannheim*.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *The Review of Economic Studies*, 70(1), 33–58.

d'Haultfoeuille, X., and A. Maurel (2009): "Inference on a Generalized Roy Model, with an Application to Schooling Decisions in France," *Unpublished manuscript, CREST-INSEE, Paris*.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications*. CRC Press.

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

Hahn, J., and G. Ridder (2010): "The Asymptotic Variance of Semiparametric Estimators with Generated Regressors," .

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

Heckman, J., and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73(3), 669–738.

Heckman, J. J., and E. J. Vytlacil (2007): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 71. Elsevier.

Härdle, W., P. Jansen, and R. Serfling (1988): "Strong Uniform Consistency Rates for Estimators of Conditional Functionals," *Annals of Statistics*, 16, 1428–1449.

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

Imbens, G., and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), 1481–1512.

Kong, E., O. Linton, and Y. Xia (2009): "Uniform bahadur representation for local polynomial estimates of M-regression and its application to the additive model," *Econometric Theory*.

Lewbel, A., and O. Linton (2007): "Nonparametric matching and efficient estimators of homothetically separable functions," *Econometrica*, 75(4), 1209–1227.

Linton, O., and A. Lewbel (2002): "Nonparametric censored and truncated regression," *Econometrica*, 70(2), 765–779.

Linton, O., and J. Nielsen (1995): "A kernel method of estimating structured nonparametric regression based on marginal integration," *Biometrika*, 82(1), 93–100.

Mammen, E., O. Linton, and J. Nielsen (1999): "The existence and asymptotic properties of a backfitting algorithm under weak conditions," *Annals of Statistics*, 27, 1443–1490.

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

Newey, W. (1994a): "Kernel estimation of partial means and a general variance estimator," *Econometric Theory*, 10(2), 233–253.

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

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

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67(3), 565–603.

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

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

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

Rilstone, P. (1996): "Nonparametric estimation of models with generated regressors," *International Economic Review*, 37(2), 299–313.

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

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

Stone, C. (1985): "Additive regression and other nonparametric models," *Annals of Statistics*, 13(2), 689–705.

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