# Multivariate Local Polynomial Estimators: Uniform Boundary Properties and Asymptotic Linear Representation*

Yanqin Fan
Department of Economics
University of Washington
Box 353330
Seattle, WA 98195

Emmanuel Guerre
School of Economics and Finance
Queen Mary, University of London
Mile End Road
London E1 4NS, United Kingdom

This version: July 2015

**Abstract**

The asymptotic bias and variance of a general class of local polynomial estimators of M-regression functions are studied over the whole *compact* support of the *multivariate* covariate under a minimal assumption on the support. The support assumption ensures that the vicinity of the boundary of the support will be visited by the multivariate covariate. The results show that like in the univariate case, multivariate local polynomial estimators have good bias and variance properties near the boundary. For the local polynomial regression estimator, we establish its asymptotic normality near the boundary and the usual optimal uniform convergence rate over the whole support. For local polynomial quantile regression, we establish a uniform linearization result which allows us to obtain similar results to the local polynomial regression. We demonstrate both theoretically and numerically that with our uniform results, the common practice of trimming local polynomial regression or quantile estimators to avoid 'the boundary effect' is not needed.

*Keywords*: Compact support; Boundary effect; Pseudo true value; Newton-Kantorovich Theorem; Regression discontinuity design; Trimming.

*We thank Ruixuan Liu for drawing our attention to Chen and Wu (2013) and Shih-Tang Hwu for pointing out an error in a preliminary draft. All remaining errors are ours.

# 1 Introduction

Recent work illustrate the practical relevance of correcting the boundary bias of kernel estimators. For example, Hickman and Hubbard (2015) demonstrate that standard kernel procedures used in the estimation of auction models may not be able to uncover important features due to boundary bias. In the context of regression discontinuity design (RDD), estimating the parameters of interest involves the estimation of two conditional expectations or conditional quantiles at the discontinuity point(s), so suffers from the boundary effect if kernel estimators are used. It is well known in the literature that theoretically one must deal with either the small denominator problem or the boundary bias for semiparametric estimators and test statistics involving averages of kernel estimators over all sample points. One way to address boundary effects is to use boundary kernels as applied in Hickman and Hubbard (2015). This solution is simple to implement in the univariate case but can become burdensome in the multivariate case where it may require estimation of the support or the support is of complicated form. See for instance Müller and Stadtmüller (1999) or Bouezmarni and Rombouts (2010) for the case of a known support.

Local polynomial estimators are known to have better boundary properties than the popular kernel estimators—their bias order is the same for interior and boundary points. This could alleviate the afore-mentioned technical and practical problems with kernel estimators. Work in both RDD and semiparametric estimation and inference using local polynomial estimators have started to appear, see e.g., Altonji, Ichimura, and Otsu (2012), Aryal, Gabrielli, and Vuong (2014), Bravo and Jacho-Chávez (2011), Hoderlein, Su, White, and Yang (2015), Qu and Yoon (2014), Su and Ullah (2008), and Su and White (2012) for estimation and inference in semiparametric and nonparametric models; and Hahn, Todd, and Van der Klaauw (2001), Porter (2003), Frandsen, Frölich, and Melly (2010), Imbens and Zajonc (2011), and Oka (2009) for the estimation in RDD. However, these work are either limited to the univariate case or circumvent potential boundary issues by trimming or focusing interest on an inner subset of the covariate support. Notable exceptions are Banerjee (2007) and Kong, Linton, and Xia (2010) who average over the whole support (unit hypercube) of a multivariate covariate to estimate, respectively, regression average derivatives and an additive quantile specification. More theoretical work in this direction are Ruppert and Wand (1994), Gu, Li, and Yang (2015), and Chen and Wu (2013) who deal with pointwise bias and variance expressions for local polynomial regression estimators.

Two important properties of local polynomial estimators are crucial to their successful applications in Econometrics: a precise characterization of their boundary properties including the bias and variance and a uniform asymptotic linear representation. For univariate covariate, boundary bias and variance expressions are well known, see Fan and Gijbels (1996). For multivariate covariate, they become complicated. The only general paper that deals with boundary bias and variance for multivariate covariates is Ruppert and Wand (1994). Under the boundary assumption (A4) of their paper, Ruppert and Wand (1994) establish expressions of bias and variance of local polynomial regression estimators at the boundary points. Specifically, Ruppert and Wand (1994) assumption (A4) considers a boundary point $x_\partial$ belonging to a convex neighborhood $\mathcal{C}$ such that $\inf_{x\in\mathcal{C}} f(x) > 0$, where $f(\cdot)$ is the probability density function (p.d.f.) of the covariate. Imbens and Zajonc (2011) apply results in Ruppert and Wand (1994) in the context of RDD with multiple forcing variables. The local assumption (A4) of Ruppert and Wand (1994) seems difficult to extend to allow for uniform estimation in the neighborhood of the support boundary. The uniform results of Kong et al. (2010) are specific to hypercube supports which, as the half spaces support considered in Gu et al. (2015), may be too restrictive in practice. We propose instead a uniform version of Chen and Wu (2013) which accounts for general boundaries.

As seen from Guerre (2000), what matters for consistent estimation is to have "enough" observations near the estimation location, a condition which involves the geometry of the support in a more subtle way and is flexible enough to cover uniform estimation. The first contribution of this paper is to provide expressions for bias and variance of a general class of multivariate local polynomial estimators of M-regression functions, including both regression and conditional quantile functions under weaker support assumptions than (A4) in Ruppert and Wand (1994). Specifically, our support condition does not require the support of the covariate to be connected and allows it to have holes. Under the new support condition, we show that the asymptotic order of the bias and variance of the local polynomial estimators of M-regression functions are not affected by boundary.

The weaker support condition is made possible by our novel application of the Newton-Kantorovich Theorem[1] to local polynomial estimation. The new approach is in line with White's (1982) approach and uses a pseudo true value to center the local polynomial estimator. Because the pseudo-true value minimizes the expectation of the objective function and satisfies a key centered score condition, it provides a more natural centering than the partial derivatives of the function to be estimated, as in Fan and Gijbels (1996) or Fan, Heckman, and Wand (1995) and the vast majority of the local polynomial literature. The Newton-Kantorovich Theorem is then used to study the bias, which is defined here as the difference between the pseudo true value and the partial derivatives. The first-order variance of the estimator can be defined using the usual sandwich formula taken at the pseudo true value. An interesting finding is that the bias expression may not depend on the estimation method, being for instance identical for regression or quantile local polynomial estimators.

Our results for the bias and variance of multivariate local polynomial estimators have immediate applications in RDD with multiple forcing variables. First, the result for regression could be used

[1]We refer interested readers to Gragg and Tapia (1974) for a complete statement of the Newton-Kantorovich Theorem and to Lemma 7.2 in Section 7 of this paper for the part used in this paper.

to relax Assumption (A4) in Ruppert and Wand (1994) adopted in Imbens and Zajonc (2011) in the context of RDD with multiple forcing variables. In addition, our result for the conditional quantile could be used to extend the estimator of quantile treatment effect in RDD with a univariate forcing variable (see, e.g., Frandsen, Frölich, and Melly, 2010; Oka, 2009), to allow for multiple forcing variables.

The second contribution of this paper is to establish asymptotic normality and uniform consistency over the whole support for local polynomial estimators of both regression and conditional quantile functions. A simple consistent estimator of the asymptotic variance is also proposed for both models. This requires to establish a new uniform linearization result for local polynomial quantile regression which holds over the whole support. In sharp contrast to Corollary 2, ii) in Masry (1996) which is concerned with multivariate covariates whose support is the entire Euclidean space and is uniformly valid over a compact subset of the support, our results deal directly with multivariate covariates with compact supports and are uniformly valid over the whole supports.

Although the order of bias and variance are not affected by boundary, a qualitative conclusion of our results is that the variance may nevertheless significantly increase near the boundary as well documented in the univariate case. The intuition is that the small denominator problem of the Kernel estimation method is not specific and should affect any nonparametric methods. Estimating a function in a narrow area of the support can only be based on few observations so that a high variance should be effected. Hence trimming to avoid high estimation variance may make sense in practice. This issue is investigated through a small simulation experiment which considers testing, additive and single index specifications. It suggests that although the small denominator problem cannot be completely ruled out for a simple unit square support, the impact of trimming is mostly negative with small potential improvements. This is especially true for testing problems where trimming may decrease the power against some standard alternatives, as well as against more specific boundary alternative which will be poorly detected.

The rest of this paper is organized as follows. The next section introduces the set-up, a general class of local polynomial estimators, and the motivating examples. Section 3 presents our results for the asymptotic bias and variance for the class of local polynomial estimators in Section 2. Section 4 establishes asymptotic normality and uniform convergence of the local polynomial regression over the entire support. A uniform linearization result for the local polynomial quantile regression is also derived in Section 4. Section 5 presents results from our simulation experiment. Section 6 offers concluding remarks and Section 7 collects technical proofs.

To close this section, we introduce some notations that will be used throughout the rest of this paper. Let $\lfloor s \rfloor$ be the lower integer part of the real number $s$, that is the unique integer number such that $\lfloor s \rfloor < s \leq \lfloor s \rfloor + 1$. In what follows, $\|\cdot\|$ stands for the Euclidean or a vector norm and $\mathcal{V}(x, r) = \{z; \|z - x\| \leq r\}$ is the closed ball with center $x$ and radius $r$. When $M$ is a matrix,

$\|M\| = \text{Tr}^{1/2}(M'M)$ is the Frobenius norm of $M$. For symmetric matrices $A$ and $B$, $A \succeq B$ means that $A - B$ is a positive matrix. The indicator function $\mathbb{I}(X \in A)$ takes value 1 when $X$ lies in $A$ and 0 otherwise.

## 2 M-Regression, Multivariate Local Polynomial Estimation, and Motivating Examples

This section first introduces a general multivariate M-regression and its local polynomial estimator. It then reviews several examples including the RDD with multiple forcing variables in Imbens and Zajonc (2011), consistent model specification testing, additive models, and average derivative estimation, where estimation of a M-regression either at the boundary or the entire support is needed.

Consider a univariate dependent variable $Y$ and a $d$-dimensional covariate $X$. We assume that the support of $X$, denoted as $\mathcal{X}$, is a compact set with a boundary $\mathcal{B}$. Let $\rho(\cdot)$ be a loss function and define the associated M-regression of $Y$ on $X$ as

$$\mu(X) = \arg\min_{\mu \in \mathbb{R}} \mathbb{E}\left[\rho(Y - \mu) | X\right]. \tag{1}$$

It is assumed that $\mu(X)$ is the unique minimizer of $\mathbb{E}[\rho(Y - \cdot)|X]$. When $\rho(t) = t^2$, $\mu(X)$ is the regression function $\mathbb{E}[Y|X]$ while when $\rho(t) = (1-\alpha)t\mathbb{I}(t \leq 0) - \alpha t\mathbb{I}(t > 0)$ for some $\alpha \in (0,1)$, $\mu(X)$ is the $\alpha$th quantile of the conditional distribution of $Y$ given $X$. The $\alpha$th expectile of Newey and Powell (1987) corresponds to $\rho(t) = (1-\alpha)t^2 I(t \leq 0) + \alpha t^2 I(t > 0)$.

The $p$th order local polynomial estimator of $\mu(\cdot)$ in (1) is defined as follows. Let $U(x)$ be the vector which groups the power $x^\pi = x_1^{\pi_1} \times \cdots \times x^{\pi_d}$ for all non negative integer numbers $\pi_1, \ldots,$ $\pi_d$ with $|\pi| = \pi_1 + \cdots + \pi_d \leq p$ according to the lexicographic order. Let $K(\cdot)$ be a non negative kernel function and $h > 0$ a bandwidth. The local polynomial estimator of $\mu(x) = \beta_0(x)$ and its partial derivatives

$$\mu^{(\pi)}(x) = \frac{\partial^{|\pi|}\mu(x)}{\partial x_1^{\pi_1} \cdots \partial x^{\pi_d}} = \frac{|\pi|!}{\pi_1! \times \cdots \times \pi_d!}\beta_\pi(x) \tag{2}$$

are denoted as $\widehat{\beta}(x;h) = \left(\widehat{\beta}_\pi(x;h), |\pi| \leq p\right)'$ with

$$\widehat{\beta}(x;h) = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(Y_i - U(X_i - x)'\beta\right) K\left(\frac{X_i - x}{h}\right), \tag{3}$$

where a suitable convention is used to break possible ties when the minimizer $\widehat{\beta}(x;h)$ is not unique as in the case of local polynomial quantile regression.

M-regressions are not only of interest in their own right but also play important roles in other contexts including estimation of average treatment effect parameters, semiparametric models, and consistent model specification testing. Throughout this paper, we use the RDD in Example 1 below to illustrate the usefulness of the bias and variance expressions established in Section 3 and Examples 2-4 below to demonstrate the usefulness of the uniform results of the type established in this paper in the context of consistent model specification testing and semiparametric estimation involving (weighted) averages of multivariate local polynomial estimators by avoiding fixed trimming of the boundary commonly adopted in existing work.

**Example 1: Regression discontinuity design.** Suppose that the support $\mathcal{X}$ is partitioned into $\mathcal{X}_0$ and $\mathcal{X}_1$, with boundaries $\mathcal{B}_0$ and $\mathcal{B}_1$. Let $\mathcal{B}_{01} = \mathcal{B}_0 \cap \mathcal{B}_1$ be the frontier between $\mathcal{X}_0$ and $\mathcal{X}_1$. Suppose that individuals with $X = X_1$ in $\mathcal{X}_1$ receive a treatment and let $Y_1$ be the associated response. Denote $(X_0, Y_0)$ in $\mathcal{X}_0 \times \mathbb{R}$ the control group variables, and

$$\mu_j (x) = \arg\min_{\mu\in\mathbb{R}} \mathbb{E} [\rho (Y_j - \mu) | X_j = x], \quad x \in \mathcal{X}_j \text{ for } j = 0, 1. \tag{4}$$

Extending the regression setup of Imbens and Zajonc (2011), we define the conditional average treatment effect as

$$\tau (x) = \mu_1 (x) - \mu_0 (x), \quad x \in \mathcal{B}_{01}. \tag{5}$$

When the treatment has no effect on the conditional parameter $\mu(\cdot)$, $\tau(\cdot) = 0$ and $\tau(\cdot) \neq 0$ otherwise. A potential difficulty with this multivariate setup is that the conditional average treatment effect is a function, as $\mathcal{B}_{01}$ is in general not a singleton. Let $x_k \in \mathcal{B}_{01}$ for $k = 1, ..., K$, where $K$ could be fixed or grow with the sample size. Simple summaries of the average treatment effect on the boundary $\mathcal{B}_{01}$ such as

$$\tau_A = \frac{1}{K} \sum_{k=1}^{K} \tau (x_k), \quad \tau_M = \max_{k=1,\ldots,K} \tau (x_k), \quad \tau_m = \min_{k=1,\ldots,K} \tau (x_k), \tag{6}$$

may be of interest. It is also possible to change the discrete set $\{x_1, \ldots, x_K\}$ to the entire frontier $\mathcal{B}_{01}$ but $\tau_A$ should be redefined using an integral instead of a discrete sum.

**Example 2: Significance testing.** Consider the null and alternative hypotheses:

$$H_0 : \mathbb{P} (\mu (X) = 0\ ) = 1,\ H_1 : \mathbb{P} (\mu (X) = 0\ ) < 1.$$

To avoid boundary issues, trimming is often used to construct test statistics for $H_0$ versus $H_1$. Let $\mathcal{X}_c$ be an inner subset of the support $\mathcal{X}$ of $X$, which is for instance obtained by selecting those $x$ in

$\mathcal{X}$ at a distance $c$ from the boundary $\mathcal{B}$, where $c$ is a trimming parameter. The null and alternative hypotheses are

$$H_{0c} : \mathbb{P}\left(\mu\left(X\right)\mathbb{I}\left[X \in \mathcal{X}_c\right] = 0\ \right) = 1,$$
$$H_{1c} : \mathbb{P}\left(\mu\left(X\right)\mathbb{I}\left[X \in \mathcal{X}_c\right] = 0\ \right) < 1.$$

Note that the alternative $H_{1,0} = H_1$ contains all the alternatives $H_{1c}$ with $c > 0$, so that trimming may give a test which is not consistent against the alternatives in $H_1$ but not in $H_{1c}$. A possible test statistic is

$$\widehat{t}_c^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}^2\left(X_i\right)\mathbb{I}\left[X_i \in \mathcal{X}_c\right], \tag{7}$$

where $\widehat{\mu}\left(X_i\right)$ is a local polynomial estimator of $\mu\left(\cdot\right)$ in (1). As is well documented in the testing literature, such test statistics can also be applied to residuals to test for more general model specification. The statistic $\widehat{t}_c^2$ is an average version of the integral test statistic of Härdle and Mammen (1993) for testing specification of the regression model. A study of the asymptotic behavior of $\widehat{t}_0^2$ for a regression null hypothesis with a univariate covariate can be found in Li (2005). A similar test was developed for a linear regression null hypothesis with dependent data by Hjellvik, Yao and Tjøstheim (1998). It follows from these authors that for some $\Delta_c$ and $\sigma_c^2$ which depend upon the distribution of the observations,

$$nh^{d/2}\widehat{t}_c^2 - h^{-d/2}\Delta_c \xrightarrow{d} \mathcal{N}\left(0, \sigma_c^2\right), \tag{8}$$

under the null. The asymptotic mean $h^{-d/2}\Delta_c$ and variance $\sigma_c^2$ can be consistently estimated provided that $h^{-d/2}$ does not diverge too fast, so that (8) leads to a rejection region: $nh^{d/2}\widehat{t}_c^2 - h^{-d/2}\widehat{\Delta}_c \geq \widehat{\sigma}_c^2\Phi\left(1-\alpha\right)$, where $\Phi\left(1-\alpha\right)$ is the standard normal $(1-\alpha)$th quantile.

An alternative approach is Fan and Li (1996) who propose a test statistic which is asymptotically centered, so that $\widehat{\Delta}_c$ is not needed. In our context, this would lead to a test statistic of the form:

$$\widehat{T}_c = \sum_{i=1}^{n} Y_i\widehat{\mu}_{-i}\left(X_i\right)\mathbb{I}\left[X_i \in \mathcal{X}_c\right], \tag{9}$$

where $\widehat{\mu}_{-i}\left(X_i\right)$ is a local polynomial leave-one-out estimation of $\mu\left(X_i\right)$. But $\widehat{T}_c$ is asymptotically normal under the null $E\left[Y\mu\left(X\right)\mathbb{I}\left[X \in \mathcal{X}_c\right]\right] = 0$, which is equivalent to $H_{0c}$ in the regression case but not necessarily for alternative choices of $\rho\left(\cdot\right)$, so that (7) should be preferred for general $\rho\left(\cdot\right)$.

**Example 3: Additive specification estimation.** A useful dimension reduction technique to estimate a function $\mu\left(\cdot\right)$ depending upon a high-dimensional covariate is to impose an additive structure on $\mu$. Suppose that $X = (X_1', X_2')'$ and that the function $\mu(\cdot)$ of (1) has an additive decomposition,

$$\mu(X) = m_1(X_1) + m_2(X_2).$$

A popular method for estimating $m_1(\cdot)$ is the marginal integration method of Linton and Nielsen (1995). Let $\mathcal{X}_{2c}$ be an inner subset of the support $\mathcal{X}_2$ of $X_2$. An estimator of

$$\mu_{1,c}(x_1) = m_1(x_1) + \mathbb{E}\left[m_2(X_2)\,\mathbb{I}(X_2 \in \mathcal{X}_{2c})\right]$$

is

$$\widehat{\mu}_{1,c}(x_1) = \frac{\sum_{i=1}^n \mathbb{I}(X_{2i} \in \mathcal{X}_{2c})\,\widehat{\mu}_{-i}(x_1, X_{2i})}{\sum_{i=1}^n \mathbb{I}(X_{2i} \in \mathcal{X}_{2c})}. \tag{10}$$

In a regression setup, Linton and Nielsen (1995) consider an integral version of (10) where $\widehat{\mu}_{-i}(\cdot,\cdot) = \widehat{\mu}(\cdot,\cdot)$ is a standard kernel regression estimator. The role of the trimming set $\mathcal{X}_{2c}$ is to avoid boundary effects. In a quantile setup and for the support $\mathcal{X}_2 = [0,1]^{d_2}$, Kong et al. (2010) show that trimming is not needed when using local polynomial estimation. In both papers, the convergence rate of the additive component is shown to be faster than the ones usually obtained for the estimation of an unrestricted function $\mu(\cdot)$.

**Example 4: Average derivative estimation.** Another popular dimension reduction approach is based on the single-index specification

$$\mu(X) = g\left(X'\beta\right),$$

where $\beta$ is a $d$ dimensional vector and $g(\cdot)$ a real-valued function. The slope parameter $\beta$ can be identified, up to a scaling coefficient, using the average derivatives of $\mu(X)$ since

$$\mathbb{E}\left[\frac{\partial \mu(X)}{\partial X}\right] = \beta \mathbb{E}\left[g^{(1)}\left(X'\beta\right)\right],$$

where $g^{(1)}(\cdot)$ is the derivative of $g(\cdot)$. See Härdle and Stocker (1989) or Powell, Stock, and Stocker (1989) for the regression case and Chaudhuri, Doksum, and Samarov (1997) for quantile functions. As seen from Altonji et al. (2012), average derivatives are also of independent interest in microeconometric issues. Direct methods average an estimator of $\partial\mu(X)/\partial X$ over $\mathcal{X}_c$ as in

$$\widehat{M}_c^{(1)} = \frac{1}{\sum_{i=1}^n \mathbb{I}[X_i \in \mathcal{X}_c]} \sum_{i=1}^n \mathbb{I}[X_i \in \mathcal{X}_c]\,\widehat{\frac{\partial \mu}{\partial X}}(X_i). \tag{11}$$

Chaudhuri, Doksum and Samarov (1997) consider local polynomial estimators of $\partial\mu(X)/\partial X$ in a quantile setup while Li, Lu and Ullah (2003), Banerjee (2007), and Altonji et al (2012) implement local polynomial for regressions. Chaudhuri et al. (1997) use a weighting function instead of trimming and Altonji et al (2012) consider $\mathcal{X}_c = [0.5, 3.5]$ for a support $\mathcal{X} = [0,4]$.

# 3 The Bias and Variance of Multivariate Local Polynomial Estimators

In this section, we establish results for the asymptotic bias and variance of multivariate local polynomial estimators of a M-regression that are valid uniformly over $\mathcal{X}$. First, we introduce the main assumptions including our new support condition and contrast it with Assumption (A4) in Ruppert and Wand (1994) who use it to develop pointwise results.

## 3.1 Main assumptions

A key issue that we have to deal with is that $\widehat{\beta}(x;h)$ is an estimator of the pseudo true value $\overline{\beta}(x;h)$ defined as

$$\overline{\beta}(x;h) = \arg\min_{\beta} \mathbb{E}\left[\rho\left(Y - U(X-x)'\beta\right) K\left(\frac{X-x}{h}\right)\right] \tag{12}$$

and that $\overline{\beta}(x;h)$ may differ from $\beta(x)$, the vector with entries

$$\beta^*_\pi(x) = \begin{cases} 0 & \text{if the partial derivative } \mu^{(\pi)}(x) \text{ does not exist} \\ \beta_\pi(x) \text{ defined in (2)} & \text{otherwise} \end{cases}.$$

The first goal of this paper is to study the bias term $\left[\overline{\beta}(x;h) - \beta(x)\right]$ over the support $\mathcal{X}$ of $X$. This will be done under the Assumptions introduced below.

**Assumption R** *(Loss function $\rho(\cdot)$). (i) For each $x \in \mathcal{X}$, $\mu \in \mathbb{R} \mapsto \mathbb{E}[\rho(Y-\mu)|X=x] = R(\mu|x)$ is twice continuously differentiable with respect to $\mu$. The second order derivative $R^{(2)}(\cdot|\cdot)$ satisfies, for some $\epsilon > 0$,*

$$\inf_{x\in\mathcal{X}} \inf_{\mu\in[\mu(x)-\epsilon,\mu(x)+\epsilon]} R^{(2)}(\mu|x) > 0,\ \sup_{x\in\mathcal{X}} \left|R^{(2)}(\mu(x)|x)\right| < \infty, \text{ and}$$
$$\left|R^{(2)}(\mu|x) - R^{(2)}(\mu'|x)\right| < C\left|\mu - \mu'\right|$$

*for all real numbers $\mu$ and $\mu'$ in $[\mu(x)-\epsilon, \mu(x)+\epsilon]$. (ii) $\rho(\cdot)$ is continuous and there is a finite collection of intervals $(a_j, a_{j+1})$ with $\bigcup_{j=0}^{J}(a_j, a_{j+1})$ such that $\rho(\cdot)$* is *continuously differentiable with derivative $\rho^{(1)}(\cdot)$ over each $(a_j, a_{j+1})$. Moreover*

$$\sup_{x\in\mathcal{X}} \mathbb{E}\left[\left(\rho^{(1)}(Y-\mu(X))\right)^2 |X=x\right] < \infty.$$

*The functions $(\mu, x) \in \mathbb{R}\times\mathcal{X} \mapsto R^{(2)}(\mu|x)$ and $\mathbb{E}\left[\left(\rho^{(1)}(Y-\mu)\right)^2 |X=x\right]$ are continuous.*

**Assumption S** *(Smoothness of $\mu(\cdot)$). There is a smoothness index $s > 0$ such that either S1 or S2 below holds:*

**S1:** *$s > 1$, $\mu(\cdot)$ is $\lfloor s \rfloor$ times differentiable and for some $L > 0$, the partial derivatives of order $\pi$ with $|\pi| = \lfloor s \rfloor$ satisfy: for all $x$, $x' \in \mathcal{X}$,*

$$\left|\mu^{(\pi)}\left(x'\right) - \mu^{(\pi)}(x)\right| \leq L \left\|x' - x\right\|^{s - \lfloor s \rfloor}.$$

**S2:** *$s$ is an integer and $\mu(\cdot)$ is $s$ times continuously differentiable.*

**Assumption X** *($\mathcal{X}$-boundary). (i) The marginal probability density function $f(\cdot)$ of the d-dimensional $X$ is continuously differentiable and bounded away from 0 on its support $\mathcal{X}$. (ii) There are some $\kappa_0, \kappa_1$ in $(0,1]$ such that for any $x \in \mathcal{X}$ and all $\epsilon$ in $(0, \kappa_0]$, there is a $x' \in \mathcal{X}$ satisfying*

$$\mathcal{V}\left(x', \kappa_1 \epsilon\right) \subset \mathcal{V}(x, \epsilon) \cap \mathcal{X}. \tag{13}$$

*(iii) $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$.*

**Assumption K** *(Kernel function). The kernel function $K(\cdot)$ is non negative and Lipschitz, i.e. $|K(x) - K(x')| \leq L \|x - x'\|$ for any $x$, $x' \in \mathbb{R}^d$. The kernel function $K(\cdot)$ has a compact support and is bounded away from 0 over the unit ball $\mathcal{V}(0,1)$. The bandwidth $h = h_n \to 0$ as $n \to \infty$.*

A brief discussion of the assumptions is in order. Assumption R-(i) is important to ensure that $\mu(x)$ in (1) and the pseudo true value $\overline{\beta}(x;h)$ in (12) are unique as implied by $R^{(2)}(\cdot|x) > 0$ and Assumption K. Indeed this implies that the Hessian matrix

$$\frac{\partial}{\partial \beta \partial \beta'} \mathbb{E}\left[\rho\left(Y - U(X-x)' \beta\right) K\left(\frac{X-x}{h}\right)\right]$$

is strictly positive and remains so when $h \to 0$. The additional Lipshitz condition on $R^{(2)}(\cdot|x)$ is used to study $\overline{\beta}(x;h)$ when $h \to 0$. Assumption R-(i) clearly holds when $\rho(t) = t^2$. In the quantile case with $\alpha$ in $(0,1)$, $\rho(t) = (1-\alpha) t \mathbb{I}(t \leq 0) - \alpha t \mathbb{I}(t > 0)$ and $R^{(2)}(\mu|x) = f(\mu|x)$, where $f(\cdot|x)$ denotes the conditional pdf of $Y$ given $X = x$, so that Assumption R-(i) holds when $\inf_{x \in \mathcal{X}} f(\mu(x)|x) > 0$ as standard in quantile estimation and if $\mu \mapsto f(\mu|x)$ is differentiable or Lipschitz in the vicinity of $\mu(x)$. Assumption R-(ii) is used to study the variance of the local polynomial estimator.

Assumption S describes some smoothness conditions for $\mu(\cdot)$. Assumption S1 is from Chaudhuri (1991) or Masry (1996). It allows for a non integer smoothness index $s$ as necessary when considering some power functions like $|x|^s$ whose derivatives of order $\lfloor s \rfloor$ can have a singular behavior in the vicinity of the origin. Assumption S2 is slightly stronger and implies Assumption S1 when the support $\mathcal{X}$ of the covariate is compact. Note that it is not assumed that $p < s$ as it is also clear from the definition of the pseudo true value $\beta_\pi^*(x)$ which allows for $p \geq s$. Indeed, non existence of the partial derivative $\mu^{(\pi)}(x)$ in the definition of $\beta_\pi^*(x)$ should be understood as $|\pi| > \lfloor s \rfloor$ under Assumption S1 and as $|\pi| > p$ under Assumption S2. Assumption K is rather standard in local polynomial estimation.

Assumption X is our key support condition. Assumption X-(i, iii) are standard but Assumption X-(ii) seems to be new to the best of our knowledge. Assumption X-(ii) holds for hypercubes $[a, b]^d$ or hyperrectangles, spheres, or supports delimited by smooth boundaries but also for more irregular support shapes. It relaxes the usual intuition of a connected support with no hole and delimited by a smooth boundary. In particular, the support $\mathcal{X}$ of the covariate $X$ does not need to be connected and can have holes. The latter possibility extends Assumption (A4) in Ruppert and Wand (1994) who assume that there is a non trivial convex set $\mathcal{C} \subset \mathcal{X}$ with non empty interior containing $x$. This restricts the shape of $\mathcal{X}$ since there cannot be a sequence of holes with vanishing size converging to $x$. In sharp contrast, under Assumption X-(ii), the boundary of $\mathcal{X}$ can be very irregular in the vicinity of $x$ with many peaks, as illustrated by an example below. In addition, Assumption X-(ii) is suitable for uniform or global studies of local polynomial estimation, while Ruppert and Wand (1994) only consider pointwise estimation of a regression function $\mu(x)$ for a given $x$.

The intuition behind Assumption X-(ii) is that a local polynomial estimator performs well provided that there are many observations close to $x$, say up to a distance $h \to 0$ given by a bandwidth, to estimate $\mu(x)$. This will hold if it is possible to find a sequence of balls in the support $\mathcal{X}$ with a radius proportional to $h$ which will converge to $x$. The key point here is that these balls do not need to be centered at $x$, which would be impossible when $x$ is on the boundary of $\mathcal{X}$, but can be centered at an interior point $x' \neq x$ of $\mathcal{X}$. This is the intuitive content of (13), where $\epsilon$ plays the role of a bandwidth. For $\epsilon > 0$ small enough, (13) means that there is a ball $\mathcal{V}(x', \kappa_1 \epsilon)$ in the support $\mathcal{X}$ which is also in the vicinity set $\mathcal{V}(x, \epsilon)$ of $x$. The fact that the constants $\kappa_0$ and $\kappa_1$ in (13) do not depend upon $x$ is essential to establish the uniformity results in this paper.

To compare and contrast with Assumption (A4) in Ruppert and Wand (1994), consider the case $d = 2$ for the sake of brevity. Like Assumption (A4) in Ruppert and Wand (1994), Assumption X-(ii) also excludes the support defined below:

$$\mathcal{X}_b = \left\{ x = (x_1, x_2) : 0 \leq x_1 \leq 1,\ 0 \leq x_2 \leq x_1^2 \right\}$$

due to the origin $o = (0,0)'$ and the fast decrease of $x_2$ when $x_1 \to 0$. Indeed balls in $\mathcal{V}(o,\epsilon) \cap \mathcal{X}_b$ have a radius which is of order $\epsilon^2$ and therefore not compatible with (13). On the other hand, a support like

$$\mathcal{X}_c = \{x = (x_1, x_2) : 0 \le x_1 \le 1,\ 0 \le x_2 \le b(x_1)\}, \quad \text{with } b(x_1) \ge \underline{b}(x_1),$$
$$\text{where } \underline{b}(x_1) = C \sum_{k=1}^{\infty} |x_1| \, \mathbb{I}\left(x_1 \in \left[\frac{1}{k+1/2}, \frac{1}{k}\right]\right) \text{ for some } C > 0$$

will satisfy Assumption X-(ii) due to the restriction $b(x_1) \ge \underline{b}(x_1)$ with $\underline{b}(x_1) = C|x_1|$ on $\left[(k+1/2)^{-1}, k^{-1}\right]$. Indeed, it is sufficient to consider, for all $k \ge 1$, those $x'$ with $x'_1 \in \left[(k+1/2)^{-1}, k^{-1}\right]$ and $x'_2 \in \left[0, Ck^{-1}\right]$ and balls $\mathcal{V}(x', \epsilon)$ in the rectangle $\left[(k+1/2)^{-1}, k^{-1}\right] \times \left[0, Ck^{-1}\right]$ to show that (13) holds. On the other hand, when $b(x_1) = \underline{b}(x_1)$, $\mathcal{X}_c$ does not satisfy Assumption (A4) in Ruppert and Wand (1994), due to the irregular behavior of $\underline{b}(x_1)$ when $x_1 \to 0$. Indeed since $\underline{b}(x_1) = 0$ on all $\left((k+1)^{-1}, (k+1/2)^{-1}\right)$, there is no non trivial convex set in $\mathcal{X}_c$ which contains the origin $o$.

## 3.2 The bias

Under the assumptions introduced in Section 3.1, it is possible to obtain the orders of the bias terms: $\left[\overline{\beta}_\pi(x,h) - \beta_\pi(x)\right]$ uniformly over $x \in \mathcal{X}$, including the boundary of $\mathcal{X}$.

**THEOREM 3.1** *Suppose Assumptions K, R, S1, and X hold with $p \ge \lfloor s \rfloor$. Then for all $\pi \in \mathbb{N}^d$ with $|\pi| \le \lfloor s \rfloor$ and $h$ small enough,*

$$\sup_{x \in \mathcal{X}} \left|\overline{\beta}_\pi(x,h) - \beta_\pi(x)\right| \le CLh^{s-|\pi|}.$$

Theorem 3.1 extends existing bias results for $x$ in inner subsets of $\mathcal{X}$. See for instance Chaudhuri (1991) and Guerre and Sabbah (2012) for the quantile case. Theorem 3.1 therefore shows that, thanks to Assumption X, the order of the bias of the local polynomial estimator is not affected by boundary. This contrasts with Nadaraya-Watson Kernel estimators and is a key reason for preferring local polynomial methods as argued by Fan and Gijbels (1996) for univariate local polynomial regression. The proof of Theorem 3.1 works by checking that the tentative limit $\beta_\pi(x)$ approximately satisfies the first-order condition of the minimization (12). Since Assumption R ensures that the Hessian of the objective function of (12) is full-rank, the Newton-Kantorovich Theorem is then used to show that $\overline{\beta}_\pi(x,h) - \beta_\pi(x) = O\left(Lh^{s-|\pi|}\right)$ uniformly over $\mathcal{X}$ as stated in the Theorem. As far as we know, this approach is new in the context of local polynomial estimation.

Theorem 3.2 below gives a uniform expansion of $\left[\overline{\beta}_\pi(x,h) - \beta_\pi(x)\right]$ under the stronger smoothness Assumption S2 which allows for a better description of the bias. Theorem 3.2 also completes

Theorem 3.1 by considering the case where $p \leq s-1 = \lfloor s \rfloor$. Define

$$\Omega_1(x,h) = \int \mathbb{I}(x+hz \in \mathcal{X})\, U(z)\, U(z)'\, K(z)\, dz, \tag{14}$$

$$\mathrm{Sp}\left(\Omega_1^{-1}(x,h)\right) = \text{ the largest eigenvalue of } \Omega_1^{-1}(x,h),$$

$$b_{p+1}(x,h) = \sum_{\pi \in \mathbb{N}^d : |\pi| = p+1} \kappa_\pi(x,h)\, \beta_\pi(x), \text{ and} \tag{15}$$

$$\kappa_\pi(x,h) = \int \mathbb{I}(x+hz \in \mathcal{X})\, z^\pi U(z)\, K(z)\, dz.$$

Let $e_\pi$ be the $\pi$th element of the canonical basis, i.e. the vector with a 1 in the $\pi$th lexicographic position and 0 elsewhere.

**THEOREM 3.2** *Suppose Assumptions K, R, S2, and X hold and that $p$ satisfies $p \leq \lfloor s \rfloor = s-1$. Then for all $\pi \in \mathbb{N}^d$ with $|\pi| \leq p$,*

$$\overline{\beta}_\pi(x,h) = \beta_\pi(x) + h^{p+1-|\pi|} e_\pi' \Omega_1(x,h)^{-1} b_{p+1}(x,h) + o\left(h^{p+1-|\pi|}\right),$$

*uniformly in $x \in \mathcal{X}$ with* $\max_{x\in\mathcal{X}} \mathrm{Sp}\left(\Omega_1(x,h)^{-1}\right) < \infty$ *and* $\max_{x\in\mathcal{X}} \|b_{p+1}(x,h)\| < \infty$ *for $h$ small enough.*

A first noticeable fact is that the leading term in the bias expansion in Theorem 3.2 is independent of the loss function $\rho(\cdot)$. The term $e_\pi' \Omega_1(x,h)^{-1} b_{p+1}(x,h)$ multiplying $h^{p+1-|\pi|}$ is identical to the one obtained for regression local polynomial estimators which has been already studied by Ruppert and Wand (1994) and more recently by Gu et al. (2015). Second, the bias boundary effect is captured through the matrix $\Omega_1(x,h)$ and the vector $\kappa_\pi(x,h)$. When $x$ is an inner point of $\mathcal{X}$, or, more precisely, when $x+hz$ lies in $\mathcal{X}$ for all $z$ in the support of the kernel function, $\Omega_1(x,h) = \Omega_1$ and $\kappa_\pi(x,h) = \kappa_\pi$ with

$$\Omega_1 = \int U(z)\, U(z)'\, K(z)\, dz \text{ and } \kappa_\pi = \int z^\pi U(z)\, K(z)\, dz, \tag{16}$$

in which case $b_{p+1}(x,h) = b_{p+1}(x)$ with,

$$b_{p+1}(x) = \sum_{\pi \in \mathbb{N}^d : |\pi| = p+1} \kappa_\pi \beta_\pi(x).$$

An important issue is whether the term $e_\pi' \Omega_1(x,h)^{-1} b_{p+1}(x,h)$ in front of $h^{p+1-|\pi|}$ vanishes or not. This has been recently discussed for the regression case and a symmetric kernel $K(\cdot)$ by Gu et al. (2015) and, as noted above, their results can also be applied to a more general $\rho(\cdot)$. When $p+1-|\pi|$ is odd, the exact order $h^{p+1-|\pi|}$ holds over the whole support except for those unlikely $x$ such that $b_{p+1}(x,h) = 0$. In this case, Theorem 3.2 can be used in conjunction with Theorem 3.4

below to propose an optimal bandwidth for inner $x$. The situation differs when $p+1-|\pi|$ is even. When $K(\cdot)$ is symmetric, $\int z^{\pi_1} z^{\pi_2} K(z)\, dz = 0$ when $|\pi_1|$ is even and $|\pi_2|$ odd, so that reorganizing the entries of $U(z)$ gives a block-diagonal $\Omega$. As shown by Gu et al. (2015), this implies that $e'_\pi \Omega_1^{-1} b(x) = 0$ when $p+1-|\pi|$ is even. Because $\Omega_1(x,h)$ is not similar to a block-diagonal matrix when $x$ lies on the boundary $\mathcal{B}$, it is unlikely to have $e'_\pi \Omega_1(x,h)^{-1} b(x,h) = 0$ unless the partial derivatives $\mu^{(\pi)}(x)$ take some very specific values. Hence for inner $x$ the bias has a smaller order $o\left(h^{p+1-|\pi|}\right)$ while for $x$ on the boundary the bias will be $O\left(h^{p+1-|\pi|}\right)$. This shows that the bias can be slightly larger near the boundary. Under some additional regularity conditions, it is possible to show that the bias is $O\left(h^{p+2-|\pi|}\right)$ for inner $x$.

Let

$$\boldsymbol{R}^{(2)}_{f,\mu}(x) = R^{(2)}(\mu(x)|x) f(x), \ \kappa_{\pi,j} = \int z_j z^\pi U(z) K(z)\, dz \text{ and}$$
$$\boldsymbol{b}_{p+1}(x) = \sum_{j=1}^{d} \sum_{\pi \in \mathbb{N}^d : |\pi| = p+1} \kappa_{\pi,j} \frac{1}{\boldsymbol{R}^{(2)}_{f,\mu}(x)} \frac{\partial \boldsymbol{R}^{(2)}_{f,\mu}(x)}{\partial x_j} \beta_\pi(x).$$

**Proposition 3.3** *Suppose Assumptions K, R, S2, and X hold and that $p$ satisfies $p \leq \lfloor s \rfloor = s-1$. Assume in addition that $K(\cdot)$ is symmetric and that $\boldsymbol{R}^{(2)}_{f,\mu}(x)$, $\mu^{(p+1)}(x)$ are continuously differentiable over $\mathcal{X}$. Then if $p+1-|\pi|$ is even and if $x$ is in the interior of $\mathcal{X}$,*

$$\overline{\beta}_\pi(x,h) = \beta_\pi(x) + h^{p+2-|\pi|} e'_\pi \Omega_1^{-1} \left(b_{p+2}(x) + \boldsymbol{b}_{p+1}(x)\right) + o\left(h^{p+2-|\pi|}\right).$$

Proposition 3.3 suggests that for an even $p+1-|\pi|$, the bias grows from $O\left(h^{p+2-|\pi|}\right)$ for those $x$ at a distance $O(h)$ to the higher order $O\left(h^{p+1-|\pi|}\right)$ when $x$ is a boundary point. As in Gu et al. (2015), Ruppert and Wand (1994) or when $d=1$, it is possible to study the bias for a sequence $x_h = x + hc$ in the interior of $\mathcal{X}$ and $x$ in $\mathcal{B}$ under additional simplifying assumptions on the boundary shape. Note also that the higher-order expansion of Proposition 3.3 now depends on the choice of $\rho(\cdot)$ through the partial derivatives of $R^{(2)}_{f,\mu}(x)$ in the term $b_{p+1}(x)$ of the expansion.

### 3.3 The variance

Consider now the variance of the local polynomial estimator. As is well-known, the partial derivative estimators $\widehat{\beta}_\pi(x;h)$ converge with a different rate and should be first standardized with the diagonal matrix

$$H = \text{Diag}\left(h^{|\pi|}, |\pi| \leq p\right).$$

Standard arguments as in Huber (1967) then suggest that the variance of $H\left(\widehat{\beta}(x,h) - \overline{\beta}(x,h)\right)$ is close to $\overline{V}(x,h)/nh^d$, where $\overline{V}(x,h)$ has the usual sandwich form

$$\overline{V}(x,h) = \overline{\boldsymbol{R}}^{(2)}(x,h)^{-1}\, \overline{\boldsymbol{S}}(x,h)\, \overline{\boldsymbol{R}}^{(2)}(x,h)^{-1}$$

with

$$\overline{\boldsymbol{R}}^{(2)}(x,h) = h^{-d}\mathbb{E}\left[R^{(2)}\left(U(X-x)'\overline{\beta}(x,h)\,|X\right)U\left(\frac{X-x}{h}\right)U\left(\frac{X-x}{h}\right)'K\left(\frac{X-x}{h}\right)\right],$$

$$\overline{\boldsymbol{S}}(x,h) = h^{-d}\mathbb{E}\left[\left(\rho^{(1)}\left(Y-U(X-x)'\overline{\beta}(x,h)\right)\right)^2 U\left(\frac{X-x}{h}\right)U\left(\frac{X-x}{h}\right)'K^2\left(\frac{X-x}{h}\right)\right],$$

where the superscript "–" indicates dependence on $\overline{\beta}(x,h)$. Since Assumptions R-(i) and X-(i) give, for $\Omega_1(x,h)$ in (14),

$$\begin{aligned}\overline{\boldsymbol{R}}^{(2)}(x,h) &\succeq Ch^{-d}\mathbb{E}\left[U\left(\frac{X-x}{h}\right)U\left(\frac{X-x}{h}\right)'K\left(\frac{X-x}{h}\right)\right]\\ &= C\int U(z)\,U(z)'\,K(z)\,f(x+hz)\,dz \succeq C\Omega_1(x,h),\end{aligned}$$

and because $\max_{(x,h)\in\mathcal{X}\times[0,\infty)}\operatorname{Sp}\left(\Omega_1(x,h)^{-1}\right) < \infty$ as stated in Theorem 3.2, $\overline{\boldsymbol{R}}^{(2)}(x,h)$ has an inverse for all $x \in \mathcal{X}$ and $h \geq 0$, so that $\overline{V}(x,h)$ is well-defined.

Consider the following approximations for $\overline{V}(x,h)$,

$$\begin{aligned}V(x,h) &= \frac{\mathbb{E}\left[\left(\rho^{(1)}(Y-\mu(X))\right)^2 |X=x\right]}{\left(R^{(2)}(\mu(x)\,|x)\right)^2 f(x)}\Omega(x,h)^{-1} \quad \text{and}\\ V(x) &= \frac{\mathbb{E}\left[\left(\rho^{(1)}(Y-\mu(X))\right)^2 |X=x\right]}{\left(R^{(2)}(\mu(x)\,|x)\right)^2 f(x)}\Omega^{-1},\end{aligned}$$

where

$$\begin{aligned}\Omega_2(x,h) &= \int \mathbb{I}(x+hz\in\mathcal{X})\,U(z)\,U(z)'\,K^2(z)\,dz,\\ \Omega(x,h)^{-1} &= \Omega_1(x,h)^{-1}\,\Omega_2(x,h)\,\Omega_1(x,h)^{-1},\\ \Omega_2 &= \int U(z)\,U(z)'\,K^2(z)\,dz, \quad \Omega^{-1} = \Omega_1^{-1}\Omega_2\Omega_1^{-1},\end{aligned}$$

and $\Omega_1(x,h)$ and $\Omega_1$ are as in (14) and (16) respectively. The next Theorem shows that $V(x,h)$ is a suitable approximation for $\overline{V}(x,h)$ over $\mathcal{X}$, while $V(x)$ is a suitable approximation for $\overline{V}(x,h)$ over interior subsets of $\mathcal{X}$.

**THEOREM 3.4** *Suppose Assumptions K, R, X hold together with Assumptions S1 or S2. Then* $\max_{x\in\mathcal{X}}\operatorname{Sp}(V(x,h)) = O(1)$ *and*

$$\sup_{x\in\mathcal{X}}\left\|\overline{V}(x,h) - V(x,h)\right\| = o(1).$$

*Consider a subset* $\mathcal{X}_0$ *of* $\mathcal{X}$ *such that* $\bigcup_{x\in\mathcal{X}_0}\mathcal{V}(x,\epsilon)\subset\mathcal{X}$ *for some* $\epsilon>0$*. Then*

$$\sup_{x\in\mathcal{X}_0}\left\|\overline{V}(x,h) - V(x)\right\| = o(1).$$

Theorem 3.4 shows that the asymptotic variance $\overline{V}(x,h)$ of $\left(nh^d\right)^{1/2} H\left(\widehat{\beta}(x,h)-\overline{\beta}(x,h)\right)$ stays bounded over $\mathcal{X}$ so that the order of $H\left(\widehat{\beta}(x,h)-\overline{\beta}(x,h)\right)$ should be $\left(nh^d\right)^{-1/2}$ for all $x\in\mathcal{X}$. Combining this result with Theorem 3.1 gives that $\left[\widehat{\beta}_\pi(x,h)-\beta_\pi(x)\right]$ should be of order $\left(\left(nh^d\right)^{-1/2}+h^s\right)/h^{-|\pi|}$ for all $\pi$ with $|\pi|\le s$ and for all $x\in\mathcal{X}$ including the boundary, that is, there is no boundary effect for the consistency rate of $\widehat{\beta}_\pi(x,h)$.

As shown as Theorem 3.4, the variance boundary effect arises because the limit of $\Omega(x,h)^{-1}$ may differ from $\Omega^{-1}$ when $x$ is close to the boundary as when $x=x_\partial+hc$ where $x_\partial$ lies on the boundary. In the univariate case, Ruppert and Wand (1994) mentioned an increase by a factor 4 when $x$ goes to the boundary and $K(\cdot)$ is a uniform kernel over $[-1,1]$. The situation can be much worse in a multidimensional case since the variance increase due to boundary is not bounded and can be made arbitrarily large by considering a support which is very narrow in the vicinity of some estimation point $x$. For instance, consider a bivariate covariate $x=(x_1,x_2)'$ and the ray of the unit disk determined by angle $\overline{\theta}>0$, i.e.,

$$\mathcal{X}_{\overline{\theta}}=\left\{(x_1,x_2)'\,;x_1=r\cos\theta,\ x_2=r\sin\theta,\ \ (r,\theta)\in[0,1]\times\left[0,\overline{\theta}\right]\right\}$$

and estimation at the vertex $o=(0,0)$. Suppose that $K(x_1,x_2)=K(r)$ with $\int_0^1 K(r)\,rdr=1/(2\pi)$ and $U(x)=[1,x_1,x_2]'$. Then for $h$ small enough and $\overline{\theta}$ small, the symmetric $\Omega_j(o,h)$, $j=1,2$ satisfy

$$\Omega_j(o,h)=\begin{bmatrix}\frac{\overline{\theta}}{2\pi} & \left(\overline{\theta}-\frac{\overline{\theta}^3}{6}\right)\int_0^1 rK^j(r)\,dr & \frac{\overline{\theta}^2}{2}\int_0^1 rK^j(r)\,dr\\ \times & \left(\overline{\theta}-\frac{\overline{\theta}^3}{3}\right)\int_0^1 r^2K^j(r)\,dr & \frac{\overline{\theta}^2}{2}\int_0^1 r^2K^j(r)\,dr\\ \times & \times & \frac{\overline{\theta}^3}{3}\int_0^1 r^2K^j(r)\,dr\end{bmatrix}+o\left(\overline{\theta}^3\right)$$

which is such $\Omega_j(o,h)$ goes to 0 when $\overline{\theta}\to 0$. Further calculations yield that all the diagonal entries of

$$\Omega(o,h)=\Omega_1(o,h)^{-1}\,\Omega_2(o,h)\,\Omega_1(o,h)^{-1}$$

**diverge as $1/\overline{\theta}$ diverges** suggesting estimation of all derivatives becomes imprecise. In other words, although Assumption $X$ implies that $\Omega(x,h)^{-1}$ stays bounded when $x$ varies over $\mathcal{X}$, this example suggests that $\Omega(x,h)^{-1}$ can be large especially when the boundary takes the shape of such small angle ray with center $x$. **By contrast, the leading bias term $\Omega_1(x,h)^{-1}b_{p+1}(x,h)$ from Theorem 3.2 is probably less affected by such small denominator problems. Indeed, by (14) and (15),**

$$\Omega_1(x,h)^{-1}b_{p+1}(x,h)=\sum_{\pi\in\mathbb{N}^d:|\pi|=p+1}\Omega_1(x,h)^{-1}\kappa_\pi(x,h)\,\beta_\pi(x),$$

**and a large $\Omega_1(x,h)^{-1}$ can be compensated by a small $\kappa_\pi(x,h)$. (Fan: are we sure about this? as a large $\Omega_1(x,h)^{-1}$ in $\Omega(o,h)$ is not compensated by small $\Omega_2(o,h)$ in variance! I changed $\Omega(x,h)$ to $\Omega_1(x,h)$)**

# 4 Local Polynomial Regression and Quantile Regression

In this section, we focus on two specific M-regressions, the local polynomial regression and local polynomial quantile regression. For each case, we establish asymptotic normality and uniform convergence rate. For local polynomial quantile, we first derive a uniform Bahadur representation valid over the whole support of the covariate $X$. We illustrate its usefulness via Examples 2-4.

## 4.1 Local polynomial regression

Consider a regression model with a heteroscedastic error term,

$$Y_i = m(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i|X_i] = 0, \quad \mathrm{Var}(\varepsilon_i|X_i) = \sigma^2(X_i), \tag{17}$$

and the quadratic loss function $\rho(t) = t^2$. We shall assume that

**Assumption E**. *The variables $(Y_i, X_i')'$, $i = 1, ..., n$, are i.i.d. The conditional p.d.f. of $\varepsilon_i$ given $X_i$ denoted as $f(e|x)$ is continuous with respect to $e$ and $x$. The variance function $\sigma^2(\cdot) = \mathrm{Var}(\varepsilon_i|X_i = \cdot)$ is bounded away from 0 and continuous over $\mathcal{X}$.* Moreover, $\sup_{x\in\mathcal{X}} \mathbb{E}\left[|\varepsilon_i|^{2+\nu}|X_i = x\right] < \infty$ *for some $\nu > 0$.*

For a quadratic loss function $\rho(\cdot)$, the local polynomial regression estimator in (3) is

$$\begin{aligned}\widehat{\beta}(x,h) &= \left(\sum_{i=1}^n U(X_i - x)U(X_i - x)' K\left(\frac{X_i - x}{h}\right)\right)^{-1} \sum_{i=1}^n U(X_i - x) Y_i K\left(\frac{X_i - x}{h}\right) \\ &= H^{-1}\left(\sum_{i=1}^n U\left(\frac{X_i - x}{h}\right) U\left(\frac{X_i - x}{h}\right)' K\left(\frac{X_i - x}{h}\right)\right)^{-1} \sum_{i=1}^n U\left(\frac{X_i - x}{h}\right) Y_i K\left(\frac{X_i - x}{h}\right),\end{aligned} \tag{18}$$

where $H$ is the diagonal matrix with entries $h^{|\pi|}$. The first entry of $\widehat{\beta}(x,h)$, say $\widehat{m}_h(x)$, is an estimator of the regression function $m(x)$, whereas the other entries estimate its partial derivatives. The pseudo-true value $\overline{\beta}(x,h)$ from (12) can be computed explicitly as

$$\overline{\beta}(x,h) = \left(\mathbb{E}\left[U\left(\frac{X-x}{h}\right) U\left(\frac{X-x}{h}\right)' K\left(\frac{X-x}{h}\right)\right]\right)^{-1} \mathbb{E}\left[\mu(X) U\left(\frac{X-x}{h}\right) K\left(\frac{X-x}{h}\right)\right].$$

The asymptotic variance $V(x,h)$ is

$$V(x,h) = \frac{\sigma^2(x)}{f(x)}\Omega(x,h)^{-1}$$

which can be estimated using

$$\widehat{V}(x,h) = \widehat{\sigma}^2(x)\,\widehat{\Omega}_1(x,h)^{-1}\,\widehat{\Omega}_2(x,h)\,\widehat{\Omega}_1(x,h)^{-1},$$

where

$$\widehat{\Omega}_j(x,h) = \frac{1}{nh^d}\sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U\left(\frac{X_i - x}{h}\right) K^j\left(\frac{X_i - x}{h}\right), \quad j = 1,2,$$

$$\widehat{\sigma}^2(x) = \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_0(x,h)\right)^2 K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)},$$

which is consistent but with a large $O(h)$ boundary bias. Note that both $\widehat{\beta}(x,h)$ and $\widehat{V}(x,h)$ depend upon an inverse matrix which may not exist. The next lemma shows that $\widehat{\beta}(x,h)$ and $\widehat{V}(x,h)$ are well defined with a probability tending to 1.

**Lemma 4.1** *Suppose that Assumptions K and X hold and that $h = h_n \to 0$ with $\log n/\left(nh^d\right) = o(1)$. Then, for $j = 1,2$,*

$$\sup_{x\in\mathcal{X}}\left\|\frac{1}{nh^d}\sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U\left(\frac{X_i - x}{h}\right)' K^j\left(\frac{X_i - x}{h}\right) - \int U(z)\,U'(z)\,K^j(z)\,f(x+hz)\,dz\right\|$$
$$= O_{\mathbb{P}}\left(\left(\frac{\log n}{nh^d}\right)^{1/2}\right) = o_{\mathbb{P}}(1). \quad (19)$$

*Moreover, for $j = 1$, the limit matrix has an inverse and*

$$\max_{x\in\mathcal{X}} \mathrm{Sp}\left(\left(\int U(z)\,U'(z)\,K(z)\,f(x+hz)\,dz\right)^{-1}\right) = O(1).$$

The next two propositions show that standard asymptotic normality and uniform convergence results hold over the whole support. Because, as is well-known, the process $\left(nh^d\right)^{1/2} H\left(\widehat{\beta}(\cdot;h) - \overline{\beta}(\cdot;h)\right)$ is not tight so that convergence in distribution to a Gaussian process cannot hold in usual functional sense, Proposition 4.2 considers the asymptotic normality of $\left(nh^d\right)^{1/2}\widehat{V}(x_n,h)^{-1/2} H\left(\widehat{\beta}(x_n;h) - \overline{\beta}(x_n;h)\right)$, where the sequence $\{x_n\} \subset \mathcal{X}$ can go to the boundary of the support $\mathcal{X}$. Note that Proposition 4.2 also establishes the consistency of the variance estimator $\widehat{V}(x_n,h)$ near the boundary. These asymptotic normality and consistency results imply that each

$$\left(nh^d\right)^{1/2} h^{|\pi|}\left(\widehat{\beta}_\pi(x_n;h) - \overline{\beta}_\pi(x_n;h)\right)$$

is asymptotically normal and gives an estimate of its asymptotic variance.

**Proposition 4.2** *Suppose that Assumptions E, K, X, and S1 or S2 hold, that* $h = h_n \to 0$ *with* $\log n / \left(nh^d\right) = o(1)$ *and that* $\{x_n\} \subset \mathcal{X}$ *is a deterministic sequence. Then*

$$\left(nh^d\right)^{-1/2} \widehat{V}(x_n, h)^{-1/2} H\left(\widehat{\beta}(x_n; h) - \overline{\beta}(x_n; h)\right)$$

*converges in distribution to a standard multivariate normal with*

$$\widehat{V}(x_n, h) = V(x_n, h) + o_p(1).$$

**Proposition 4.3** *Suppose that Assumptions E, K, X and S1 or S2 hold, that* $h = h_n \to 0$ *with* $h^{-d} = O\left(n^{(\nu-2)/\nu} / \log n\right)$*. Then*

$$\sup_{x \in \mathcal{X}} \left\| H\left(\widehat{\beta}(x; h) - \beta(x)\right) \right\| = O_p\left(\left(\frac{\log n}{nh^d}\right)^{1/2} + h^s\right).$$

**Example 1 (Cont'd).** *In this example, the two different regression functions are estimated using two different and independent samples of observations* $(X_{ji}, Y_{ji})$, $j = 0, 1$, *with the same sample size* $n$ *and the same bandwidth for the sake of simplicity. The estimator of the average treatment effect in (5) is then*

$$\widehat{\tau}_h(x) = \widehat{\mu}_{1h}(x) - \widehat{\mu}_{2h}(x), \quad x \in \mathcal{B}_{01}.$$

*Proposition 4.3 implies that* $\widehat{\tau}_h(x)$ *converges uniformly to* $\tau(x)$ *over* $B_{01}$ *with a rate* $\left[\left(\log n / \left(nh^d\right)\right)^{1/2} + h^s\right]$ *extending the pointwise result in Imbens and Zajonc (2011). Proposition 4.2 easily extends to a vector* $(\widehat{\tau}_h(x_1), \ldots, \widehat{\tau}_h(x_K))'$ *which will be asymptotically independent. It follows that the estimation* $\widehat{\tau}_{Ah}$ *of the average treatment effect mean* $\tau_A$ *as in (6) is asymptotically normal with a bias which can be derived from Theorems 3.1 and 3.2 and an asymptotic variance of order* $1/\left(nh^d\right)$ *obtained by averaging the ones from Proposition 4.2. Approximation for the asymptotic distribution of* $\widehat{\tau}_{Mh}$ *and* $\widehat{\tau}_{mh}$ *can also be easily obtained under the additional condition that the maximum and minimum of the* $\tau(x_k)$*'s are achieved for one* $x_k$*, say* $x_M$ *and* $x_m$*. In this case,* $\left(nh^d\right)^{1/2}(\widehat{\tau}_{Mh} - \widehat{\tau}_h(x_M))$ *and* $\left(nh^d\right)^{1/2}(\widehat{\tau}_{mh} - \widehat{\tau}_h(x_m))$ *are both* $o_p(1)$ *so that the asymptotic distribution of* $\left(nh^d\right)^{1/2}(\widehat{\tau}_{Mh} - \tau_M)$ *and* $\left(nh^d\right)^{1/2}(\widehat{\tau}_{mh} - \tau_m)$ *are the ones of* $\left(nh^d\right)^{1/2}(\widehat{\tau}_h(x_M) - \tau(x_M))$ *and* $\left(nh^d\right)^{1/2}(\widehat{\tau}_h(x_m) - \tau(x_m))$*. Interestingly, under the assumption of no treatment effect, the estimator* $\widehat{\tau}_h(x)$ *is asymptotically unbiased because the same bandwidth is used for the treated and control samples. This may considerably simplify testing.*

**Examples 2, 3 and 4 (Cont'd)** *Although Propositions 4.2 and 4.3 do not apply to these Examples, the proof of these results suggests that our bias and variance results are sufficient to extend existing results which involve trimming to the bounded support case under Assumption X. For the full support test statistic* $\widehat{t}_0$ *in (7), establishing that* $nh^{d/2}\widehat{t}_0$ *converges in distribution to a centered normal distribution can be done with minor modifications of the arguments of Hjellvik et al. (1998). Studying the full support marginal integration estimator* $\widehat{\mu}_{1,0}(x_1)$ *in (10) easily follows from Kong et al. (2010) while the full support average derivatives estimator* $\widehat{M}_0^{(1)}$ *in (11) can be studied following Chaudhuri et al. (1997), Li et al. (2003) or Banerjee (2007).*

## 4.2 Local polynomial quantile regression

Consider the family of loss functions

$$\rho_\alpha(t) = t\left[\alpha - \mathbb{I}(t \leq 0)\right], \quad \alpha \in [\underline{\alpha}, \overline{a}] \subset (0,1),$$

so that

$$\rho_\alpha^{(1)}(t) = \alpha - \mathbb{I}(t \leq 0).$$

For this choice of loss functions, $\widehat{\beta}_0(\alpha|x,h) = \widehat{\beta}_0(x,h)$ is an estimator of the conditional quantile function $Q(\alpha|x)$ of $Y_i$ given $X_i = x$ and $\widehat{\beta}_\pi(\alpha|x,h) = \widehat{\beta}_\pi(x,h)$ estimates the partial derivative $\partial^{|\pi|}Q(\alpha|x)/\partial x^\pi$. We shall use the following standard assumption.

> **Assumption F**. *The variables* $(Y_i, X_i')_{i=1}^n$ *are i.i.d. The conditional p.d.f. of* $Y$ *given* $X = x$ *denoted as* $f(y|x)$ *is continuous and differentiable with respect to* $y$ such that $f(y|x) > 0$ *for all* $(x,y) \in \mathcal{X}\times\mathbb{R}$, $\sup_{(x,y)\in\mathcal{X}\times\mathbb{R}}\left|\frac{\partial f(y|x)}{\partial y}\right| < \infty$, *and* $\sup_{(x,y)\in\mathcal{X}\times\mathbb{R}}\left|\frac{\partial f(y|x)}{\partial x}\right| < \infty$.

Under Assumption F, the minimizers $Q(\alpha|x) = \mu(x)$ and $\overline{\beta}(\alpha|x,h) = \overline{\beta}(x,h)$ of (1) and (12) are unique. Let $F(\cdot|\cdot)$ be the cumulative distribution function of $Y$ given $X$. The functions $R_\alpha^{(1)}(\mu|x)$, $\mathbb{E}\left[\left(\rho_\alpha^{(1)}(Y-\mu)\right)^2 |X = x\right]$, $\mathbb{E}\left[\left(\rho_\alpha^{(1)}(Y-Q(\alpha|x))\right)^2 |X = x\right]$, $R_\alpha^{(2)}(\mu|x)$, $V(\alpha|x,h)$, and $V(\alpha|x)$ are:

$$R_\alpha^{(1)}(\mu|x) = F(\mu|x) - \alpha, \quad R_\alpha^{(2)}(\mu|x) = f(\mu|x),$$

$$\mathbb{E}\left[\left(\rho_\alpha^{(1)}(Y-\mu)\right)^2 |X = x\right] = \mathbb{E}\left[\left(\mathbb{I}(Y \leq \mu) - \alpha\right)^2 |X = x\right],$$

$$\mathbb{E}\left[\left(\rho_\alpha^{(1)}(Y-Q(\alpha|x))\right)^2 |X = x\right] = \alpha(1-\alpha),$$

$$V(\alpha|x,h) = \frac{\alpha(1-\alpha)}{f^2(Q(\alpha|x)|x)f(x)}\Omega(x,h)^{-1}, \text{ and}$$

$$V(\alpha|x) = \frac{\alpha(1-\alpha)}{f^2(Q(\alpha|x)|x)f(x)}\Omega^{-1}.$$

Hence Assumption R follows from Assumption F. Since

$$\frac{1}{f\left(Q\left(\alpha|x\right)|x\right)}=\frac{\partial Q\left(\alpha|x\right)}{\partial\alpha},$$

a possible estimator of the asymptotic variance $V\left(\alpha|x,h\right)$ is

$$\widehat{V}\left(\alpha|x,h\right)=\alpha\left(1-\alpha\right)\left(\frac{\partial\widehat{Q}\left(\alpha|x\right)}{\partial\alpha}\right)^{2}\Omega\left(x,h\right)^{-1},$$

$$\text{where }\frac{\partial\widehat{Q}\left(\alpha|x\right)}{\partial\alpha}=\frac{\widehat{\beta}_{0}\left(\alpha+\eta|x,h\right)-\widehat{\beta}_{0}\left(\alpha-\eta|x,h\right)}{2\eta},\quad\eta=\eta_{n}\to0.$$

An important difference between conditional regression and quantile estimation is that the local polynomial quantile regression is neither explicit nor linear with respect to the $Y_i$. An additional step is needed to show that $\widehat{\beta}\left(\alpha|x,h\right)$ is asymptotically linear, see e.g. Chaudhuri (1991), Su and Xiao (2009) or Guerre and Sabbah (2012). To this aim, define

$$\widehat{S}\left(\alpha|x,h\right)=\frac{1}{\left(nh^{d}\right)^{1/2}}\sum_{i=1}^{n}\left(\mathbb{I}\left(Y_{i}\leq U\left(X_{i}-x\right)'\overline{\beta}\left(\alpha|x,h\right)\right)-\alpha\right)U\left(\frac{X_{i}-x}{h}\right)K\left(\frac{X_{i}-x}{h}\right),$$

$$\widehat{J}\left(\alpha|x,h\right)=\frac{1}{nh^{d}}\sum_{i=1}^{n}f\left(U\left(X_{i}-x\right)'\overline{\beta}\left(\alpha|x,h\right)|X_{i}\right)U\left(\frac{X_{i}-x}{h}\right)U'\left(\frac{X_{i}-x}{h}\right)K\left(\frac{X_{i}-x}{h}\right),\text{ and}$$

$$\overline{J}\left(\alpha|x,h\right)=\mathbb{E}\left[\widehat{J}\left(\alpha|x,h\right)\right].$$

Lemma 4.1 and Assumption F ensure that $\widehat{J}\left(\alpha|x,h\right)$ has an inverse asymptotically. The next linearization proposition is an extension of Guerre and Sabbah (2012, Theorem 2) which allows for estimation location $x$ close or on the boundary of $\mathcal{X}$. Note that under Assumption F, the smoothness index $s$ in Assumption S can be taken greater than 1 as assumed in all the results below.

**Proposition 4.4** *Suppose that Assumptions K, F, X, and S1 or S2 hold for some $s\geq1$, that $h=h_n\to0$ with $\log n/\left(nh^{d}\right)=o\left(1\right)$. Then*

$$\sup_{(\alpha,x)\in[\underline{\alpha},\overline{\alpha}]\times\mathcal{X}}\left\|H\left(\widehat{\beta}\left(\alpha|x,h\right)-\overline{\beta}\left(\alpha|x,h\right)\right)+\frac{\widehat{J}\left(\alpha|x,h\right)^{-1}\widehat{S}\left(\alpha|x,h\right)}{\left(nh^{d}\right)^{1/2}}\right\|=O_{p}\left(\left(\frac{\log n}{nh^{d}}\right)^{\frac{3}{4}}\right)\text{ and}$$

$$\sup_{(\alpha,x)\in[\underline{\alpha},\overline{\alpha}]\times\mathcal{X}}\left\|H\left(\widehat{\beta}\left(\alpha|x,h\right)-\overline{\beta}\left(\alpha|x,h\right)\right)+\frac{\overline{J}\left(\alpha|x,h\right)^{-1}\widehat{S}\left(\alpha|x,h\right)}{\left(nh^{d}\right)^{1/2}}\right\|=O_{p}\left(\left(\frac{\log n}{nh^{d}}\right)^{\frac{3}{4}}\right).$$

Approximating $H\left(\widehat{\beta}\left(\alpha|x,h\right)-\overline{\beta}\left(\alpha|x,h\right)\right)$ with $\widehat{J}\left(\alpha|x,h\right)^{-1}\widehat{S}\left(\alpha|x,h\right)$ as in the first equation gives an approximation similar to that for a local polynomial regression estimator. Using the leading term $\overline{J}(\alpha|x,h)^{-1}\widehat{S}(\alpha|x,h)$ shows that the local polynomial quantile estimator is asymptotically a sum which can be handled with standard limit theorems. A similar result can be established for a larger class of loss functions $\rho(\cdot)$ such as the one used in Powell and Newey (1987). However, the rate $\left(\log n/\left(nh^d\right)\right)^{3/4}$ is typical of the quantile check function which is not twice continuously differentiable and better rates hold for smoother $\rho(\cdot)$.

This linearization result is the key tool to establish a Central Limit Theorem and uniform consistency for local polynomial quantile regression. The next two propositions parallel Propositions 4.2 and 4.3 for local polynomial regression. Like Theorems 3.1, 3.2, and 3.4, these results show that the boundary effect can be weak for local polynomial quantile regression estimators.

**Proposition 4.5** *Suppose that Assumptions K, F, X, and S1 or S2 hold with* $s \geq 1$, *that* $h = h_n \rightarrow 0$ *and* $\eta = \eta_n \rightarrow 0$ *with* $\log^3 n/\left(nh^d\right) = o(1)$, $h + \left(\log n/\left(nh^d\right)\right)^{1/2} = o(\eta)$ *and that* $\{x_n\} \subset \mathcal{X}$ *is a deterministic sequence. Then*

$$\left(nh^d\right)^{-1/2}\widehat{V}(\alpha|x_n,h)^{-1/2}H\left(\widehat{\beta}(\alpha|x_n;h) - \overline{\beta}(\alpha|x_n;h)\right)$$

*converges in distribution to a multivariate normal with*

$$\widehat{V}(\alpha|x_n,h) = V(\alpha|x_n,h) + o_p(1).$$

**Proposition 4.6** *Suppose that Assumptions K, F, X, and S1 or S2 hold, that* $h = h_n \rightarrow 0$ *with* $\log^3 n/\left(nh^d\right) = o(1)$. *Then*

$$\sup_{(\alpha,x)\in[\underline{\alpha},\overline{\alpha}]\times\mathcal{X}}\left\|H\left(\widehat{\beta}(\alpha|x;h) - \beta(\alpha|x;h)\right)\right\| = O_p\left(\left(\frac{\log n}{nh^d}\right)^{1/2} + h^s\right).$$

**Example 1 (Cont'd).** In the quantile setup, the estimation of the average treatment effect in (5) becomes

$$\widehat{\tau}_h(\alpha|x) = \widehat{Q}_{1h}(\alpha|x) - \widehat{m}_{2h}(\alpha|x), \quad \alpha \in [0,1], \quad x \in \mathcal{B}_{01}.$$

Proposition 4.6 yields that $\widehat{\tau}_h(\alpha|x)$ converges uniformly to $\tau(\alpha|x)$ over $[\underline{\alpha},\overline{\alpha}] \times \mathcal{B}_{01}$ with a rate $\left(\log n/\left(nh^d\right)\right)^{1/2} + h^s$ extending Imbens and Zajonc (2011) to quantile set-up. The indicators in (6) can be computed for each quantile levels. For a given $\alpha$, they behave as in the regression case. When $\alpha$ varies in $[\underline{\alpha},\overline{\alpha}]$, these indicators should be considered as stochastic processes whose asymptotic distribution can be derived using the asymptotic expansion stated in Proposition 4.4.

**Example 2 (Cont'd).** *Proposition 4.4 is useful to obtain a suitable approximation of the test statistic* $\widehat{t}_c^2$*. Let* $\widetilde{\mu}(x) = \widetilde{Q}(\alpha|x)$ *be the leading term of the conditional quantile local polynomial estimator* $\widehat{\mu}(x) = \widehat{Q}(\alpha|x)$

$$\widetilde{\mu}(x) = \overline{Q}_h(\alpha|x) - e_1' \frac{\overline{J}(\alpha|x,h)^{-1}\widehat{S}(\alpha|x,h)}{(nh^d)^{1/2}}, \tag{20}$$

*where* $e_1' = [1,0,\ldots,0]$ *and* $\overline{Q}_h(\alpha|x) = e_1'\overline{\beta}(\alpha|x;h)$*. The "linearized" version of* $\widehat{t}_c^2$ *is*

$$\widetilde{t}_c^2 = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mu}^2(X_i)\,\mathbb{I}[X_i \in \mathcal{X}_c].$$

*Proposition 4.4 and the triangular inequality yield that*

$$\begin{aligned} nh^{d/2}\left(\widehat{t}_c^2 - \widetilde{t}_c^2\right) &= \left(\left(nh^{d/2}\right)^{1/2}\left[\widehat{t}_c - \widetilde{t}_c\right]\right) \times \left(\left(nh^{d/2}\right)^{1/2}\left[\widehat{t}_c + \widetilde{t}_c\right]\right) \\ &= \left(nh^{d/2}\right)^{1/2} O_p\left(\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right) \times \left(\left(nh^{d/2}\right)^{1/2} 2\widehat{t}_c + \left(nh^{d/2}\right)^{1/2} O_p\left(\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right)\right) \\ &= O_p\left(\left(\frac{\log^3 n}{nh^{2d}}\right)^{\frac{1}{4}}\right) \times \left(\left(nh^{d/2}\right)^{1/2} 2\widehat{t}_c + O_p\left(\left(\frac{\log^3 n}{nh^{2d}}\right)^{\frac{1}{4}}\right)\right). \end{aligned}$$

*Consider the null hypothesis. Solving the first-order condition for* $\overline{\beta}(\alpha|x;h)$ *gives* $\overline{\beta}(\alpha|x;h) = 0$ *when* $Q(\alpha|x) = 0$ *for all* $x$ *in* $\mathcal{X}_c$*. As a consequence,* $\widetilde{t}_c^2$ *is a quadratic form similar for the one obtained in the regression framework, but with the centered variables* $\mathbb{I}(Y_i \le 0) - \alpha$ *instead of the regression error terms. It follows that* $\widetilde{t}_c^2$ *will satisfy (8) with a proper choice of standardizing constants. This gives,*

$$nh^{d/2}\left(\widehat{t}_c^2 - \widetilde{t}_c^2\right) = O_p\left(\left(\frac{\log^3 n}{nh^{3d}}\right)^{\frac{1}{4}}\right),$$

*which goes to* 0 *when* $\left(\log^3 n\right)/\left(nh^{3d}\right) \to 0$*, a condition which will also ensure that* $\widehat{t}_c^2$ *is asymptotically normal as in (8) and that normal critical values can be used to perform the test.*

**Example 3 (Cont'd).** *Defining a leave-one-out version of the linear* $\widetilde{\mu}(x_1,x_2)$ *as in (20) yields for the marginal integration estimator (10) of* $\mu_1(x_1)$ *and applying Proposition 4.4 suggests,*

$$\widehat{\mu}_{1,c}(x_1) = \frac{\sum_{i=1}^{n}\mathbb{I}(X_{2i}\in\mathcal{X}_{2c})\,\widetilde{\mu}_i(x_1,X_{2i})}{\sum_{i=1}^{n}\mathbb{I}(X_{2i}\in\mathcal{X}_{2c})} + O_p\left(\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right),$$

*with a leading term which can be studied as in Kong et al. (2010), for any* $c \ge 0$*. The estimator error* $\left(nh^{d_1}\right)^{1/2}\left(\widehat{\mu}_{1,c}(x_1) - \mu_{1,c}(x_1)\right)$ *will satisfies a CLT if the contribution of the linearization remainder term is negligible, yielding the necessary condition,*

$$\left(nh^{d_1}\right)^{1/2}\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}} = \left(\frac{\log^3 n}{nh^{3d-2d_1}}\right)^{\frac{1}{4}} \to 0.$$

**Example 4 (Cont'd).** *The case of the average derivatives estimator is similar to Example 3. However, the contribution of the linearization remainder term must take into account an additional* $1/h$ *due to the estimation of a derivative. The* $n^{1/2}$ *asymptotic normality of the estimator requests a more drastic bandwidth condition,*

$$n^{1/2}\frac{1}{h}\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}} = \left(\frac{\log^3 n}{nh^{3d+4}}\right)^{\frac{1}{4}} \to 0.$$

*The asymptotic bias and variance of the estimator can be obtained from Chaudhuri et al. (1997) and our bias and variance results to account for the good boundary properties of the conditional quantile local polynomial estimator.*

## 5 Simulation Experiments

Our theoretical results have illustrated the good bias and variance boundary properties of local polynomial estimation for general loss functions. The boundary bias has the same order as the minimax bias obtained for worst case specifications. The behavior of the variance is similar, with an order which is not affected by the boundary. However, for the variance, the constant in front of the order may be larger for points close to the boundary than for points away from the boundary. Preventing poor estimation induced by such areas may justify trimming in Examples 2, 3, and 4. The purpose of this section is to use a small simulation experiment.to illustrate how trimming influences inference in various settings.

We will use the setup of Examples 2, 3, and 4. To avoid nonlinearity issues, we will consider the regression model,

$$Y_i = \mu\left(X_{1i}, X_{2i}\right) + \varepsilon_i, \quad i = 1, \ldots, 1,000,$$

where $\left\{[X_{1i}, X_{2i}]'\right\}_{i=1}^n$ are i.i.d. Uniform over $[0,1]^2$ and $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. centered normal with standard deviation 0.1. The choice of the regression function $\mu(\cdot)$ will vary across examples. The number of replications is 5,000. The regression local polynomial estimators of order 1 (linear) and order 2 (quadratic) will be considered with the kernel:

$$K(x) = \left(1 - x^2\right)\mathbb{I}\left(x \in [-1,1]\right).$$

Three trimming values will be investigated, see the following table, where trimming 1 corresponds to no trimming.

| Trimming 1 ($c = 0$) | Trimming 2 ($c = 0.1$) | Trimming 3 ($c = 0.2$) |
|---|---|---|
| $(X_{1i}, X_{2i}) \in [0,1]^2$ | $(X_{1i}, X_{2i}) \in [0.1, 0.9]^2$ | $(X_{1i}, X_{2i}) \in [0.2, 0.8]^2$ |

**Table 1: Trimming values used in the experiment**

### 5.1 Example 2: significance testing

Instead of (7), the more popular test statistic $\widehat{T}_c$ of (9) is used. This test statistic can be written as a quadratic form $Y_n'WY_n$, where the symmetric matrix $W$ depends on the local polynomial estimator, the bandwidth $h$, and the trimming parameter $c$. Its variance can be estimated using

$$\widehat{\sigma}_c^2 = 2 \sum_{i \leq i \neq j \leq n} W_{ij}^2 \widehat{\varepsilon}_i^2 \widehat{\varepsilon}_j^2,$$

where the residuals $\widehat{\varepsilon}_i$ are computed from the local polynomial estimation. The considered bandwidths are

$$h^2 \in \{0.01, 0.04, 0.07, 0.10, 0.13, 0.16, 0.19\}.$$

We estimated the 90%, 95% and 99% critical values by computing the test statistic $\widehat{T}_c / \widehat{\sigma}_c$ for each level of trimming and each bandwidth over $5,000$ replications of the null model: $\mu(\cdot) = 0$. This did not show specific impact of trimming with simulated critical values reasonably close to their norminal counterparts.

The simulated 90% critical values were used to study the power of the tests using $1,000$ replications of the model for the two regression functions below:

$$m_1(x_1, x_2) = 0.1K\left(\frac{x_1}{0.1}\right) - 0.1K\left(\frac{1 - x_2}{0.1}\right) \text{ and}$$
$$m_2(x_1, x_2) = -0.05\cos(6\pi x_1) + 0.05\cos(6\pi x_2).$$

The alternative $m_1(\cdot)$ consists of two bumps along the boundaries $x_1 = 0$ and $x_2 = 1$ which are difficult to detect with the trimming tests. On the contrary, the alternative $m_2(\cdot)$ violates the null for most $x$ in the support $[0, 1]^2$. Figure 1 reports the results of the simulation experiment. As expected the power of the tests against alternative $m_1(\cdot)$ deteriorates with the level of trimming and the test without trimming clearly dominates. The evidence is less clear for alternative $m_2(\cdot)$ which periodicity induces an irregular bandwidth behavior. However, only the test without trimming achieves a power close to 1. Considering a quadratic local polynomial estimator gives a much less powerful test for the considered bandwidths. This surprising finding is however in line with the theoretical results of Guerre and Lavergne (2002) which shows that lower order methods can have good power properties in a minimax framework.

### 5.2 Example 3: additive specification estimation

In this experiment, the regression function is set to $m(x_1) + m(x_2)$ *with* $m(x_1) = \sin(2\pi x_1)$. The local linear smoother $\widehat{\mu}_{1,c}(x_1)$ from (10) is an estimator of

$$\mu_{1,c}(x_1) = m(x_1) + \mathbb{E}[m(X_2)\mathbb{I}(X_2 \in [c, 1 - c])]$$

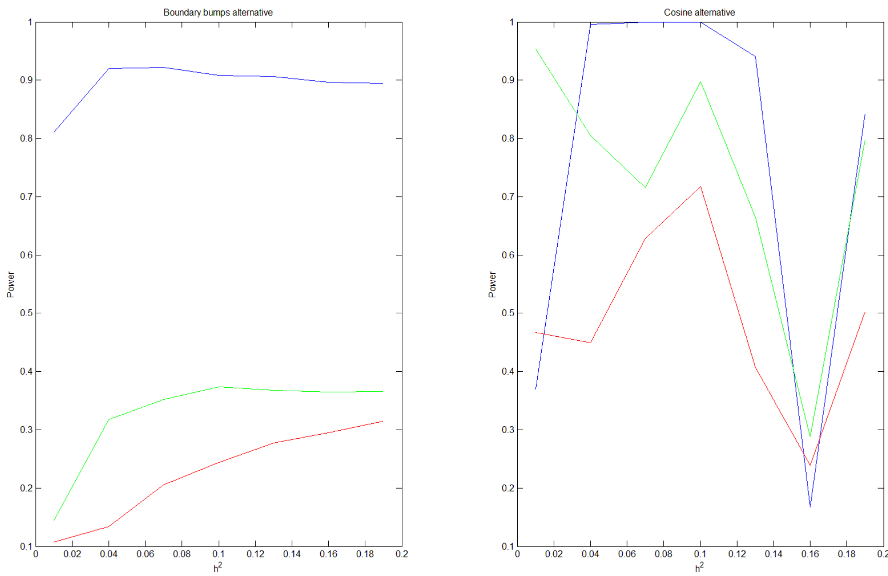


Figure 1: Power of the three trimming 90% tests as a function of $h^2$. Left: Alternative $m_1(\cdot)$. Right: Alternative $m_2(\cdot)$. Blue: No trimming. Green: Trimming 2. Red: Trimming 3

which includes an expectation term. This term can be removed by imposing an identification restriction such as $m(1/2) = 0$. The proposed estimator of $m(x_1)$ is therefore[2]

$$\widehat{m}_c(x_1) = \widehat{\mu}_{1,c}(x_1) - \widehat{\mu}_{1,c}(1/2).$$

The local linear estimator performs poorly and the reported results are for the local quadratic estimator. The considered bandwidths are smaller than the ones used in Example 1. They are:

$$h^2 \in \{0.02, 0.03, \ldots, 0.06\}.$$

The performance of the estimator is evaluated using the square root average mean squared error (RAMSE):

$$RAMSE = \left[\frac{1}{11}\sum_{j=0}^{10}\left(\widehat{m}_c\left(\frac{j}{10}\right) - m_c\left(\frac{j}{10}\right)\right)^2\right]^{\frac{1}{2}}.$$

The next Figure shows that the estimator without trimming clearly dominates. RAMSE is around 0.40 at best, which is quite big but not surprising since the best bandwidth is $h^2 = 0.04$ implying that only 40 observations are used to estimate each $\mu_{1,c}(j/10)$.

[2]When $m(x_1) = \sin(2\pi x_1)$, $\mathbb{E}[m(X_2)\mathbb{I}(X_2 \in [c, 1-c])] = 0$ so that $\widehat{\mu}_{1,c}(x_1)$ could also be used. However identifying restrictions such as $m(1/2) = 0$ seems to be more popular and does not involve the trimming parameter. Unreported simulation results suggests however that $\widehat{\mu}_{1,c}(\cdot)$ may have better performances than $\widehat{m}_c(\cdot)$.

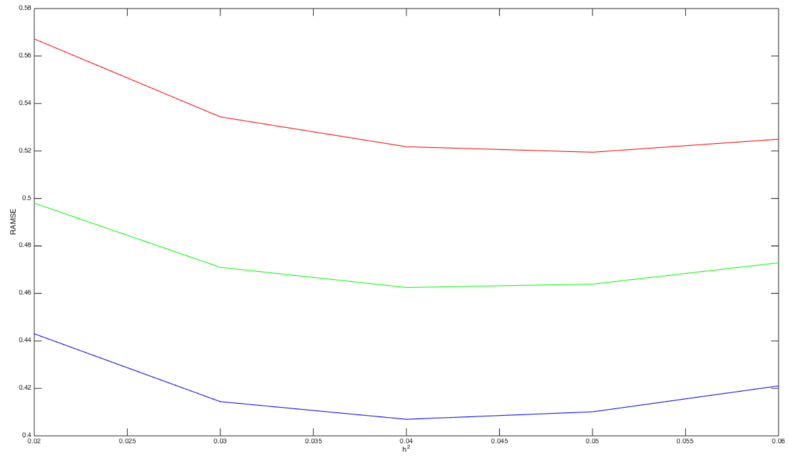


Figure 2: Square root average mean squared error of $\widehat{m}_c(\cdot)$ as a function of $h^2$. Blue: No trimming. Green: Trimming 2. Red: Trimming 3.

### 5.3 Example 4: average derivative estimation

For this example, a single-index specification is used and the regression function is

$$\mu(x) = \sin^3\left(\frac{\pi}{2}\left(x' - [1/2, 1/2]\right)\beta_0\right), \quad \beta_0' = [1, 1/2].$$

The parameter of interest is $\delta_0 = \beta_{02}/\beta_{01} = 1/2$ which is estimated using a ratio of average derivatives

$$\widehat{\delta}_c = \frac{\sum_{i=1}^n \widehat{\partial\mu_i/\partial x_2}(X_i)\,\mathbb{I}[X_i \in \mathcal{X}_c]}{\sum_{i=1}^n \widehat{\partial\mu_i/\partial x_1}(X_i)\,\mathbb{I}[X_i \in \mathcal{X}_c]}.$$

The performance of $\widehat{\delta}_c$ is measured using the square root mean squared error (RMSE). Because estimation of derivatives yields a bigger variance of order $1/\left(nh^3\right)$ compared to estimation of a regression for which the variance order is $1/\left(nh^2\right)$, the RMSE of $\widehat{\delta}_c$ could be more sensitive than the RAMSE of Example 3 to an increase of the variance near the boundary. The local linear and quadratic regression estimators perform similarly and both are reported here. The considered bandwidths are

$$h^2 \in \{0.015, 0.035, \ldots, 0.135\} \text{ for linear local estimation and}$$
$$h^2 \in \{0.02, 0.07, \ldots, 0.32\} \text{ for local quadratic estimation.}$$

The highest trimming estimator is dominated by lower trimming ones, for both local linear and quadratic estimation. The trimming 1 and 2 estimators behave similarly but trimming 2 seems slightly better than no trimming when looking at the optimal performance. This holds for both local linear and quadratic estimators. Since the bias of the local quadratic estimator is smaller, this

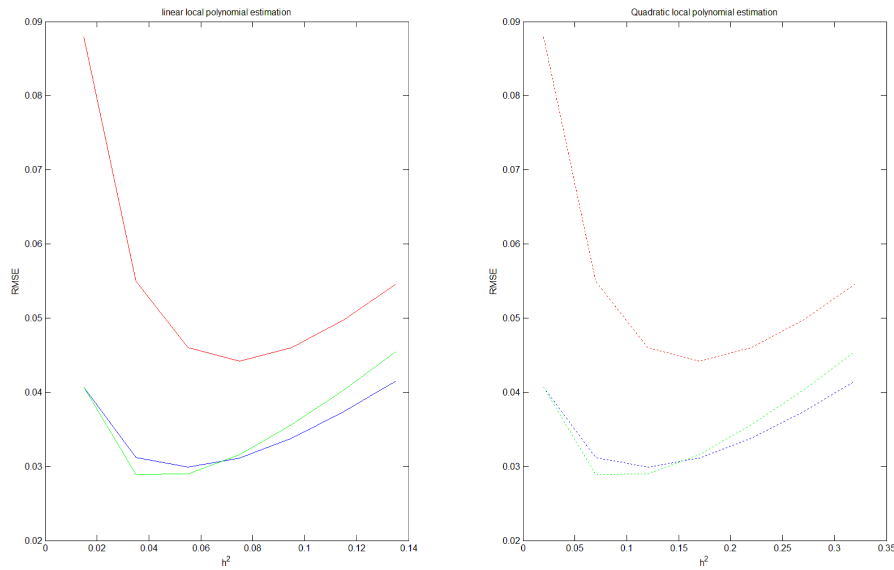


Figure 3: Square root mean squared error for the three trimming estimators as a function of $h^2$. Left: Linear local polynomial. Right: Quadratic local polynomial. Blue: No trimming. Green: Trimming 2. Red: Trimming 3.

behavior suggests that there is an optimal level of trimming possibly due to an increase of variance near the boundary. However the potential gain seems very small in this experiment.

## 6 Conclusion

In this paper, we have investigated the boundary and uniform asymptotic properties of multivariate local polynomial estimators of M-regression functions under a weak condition on the *compact* support of the multivariate covariate. This is made possible by a pseudo true value approach based on a novel application of the Newton-Kantorovich Theorem in our context. Compared with Assumption (A4) in Ruppert and Wand (1994) who use it to establish pointwise boundary properties of local polynomial regression estimators, our support condition allows for more general support shapes, in particular, it allows the support of the covariate to be non-connected and have holes. Compared with the uniform result in Corollary 2, ii) in Masry (1996), our results deal directly with multivariate covariates with compact support and are uniformly valid over the entire support. As such they should be useful in contexts where estimation or testing require (weighted) averages of multivariate nonparametric estimators with compactly supported covariate as in Examples 2-4.

# 7 Proof section

Note that Assumption X ensures that for any $h > 0$ small enough, any $x \in \mathcal{X}$, there is a $x_h \in \mathcal{X}$ such that $\mathcal{V}(x_h, \kappa_1 h) \subset \mathcal{V}(x, h) \cap \mathcal{X}$. We will use this equivalent statement of (13) throughout this section. Also we will use $C$ to denote a generic positive constant whose value may differ in different places.

We start with Lemma 7.1 which will be used in many proofs below.

**Lemma 7.1** *Under Assumptions K and X, there is a $C > 1$ such that the eigenvalues of $\int U(z) U'(z) K^j(z) f(x+hz)\,dz$ and $\Omega_j(x,h)$, $j = 1, 2$, are in $[1/C, C]$ for all $x \in \mathcal{X}$ and $h \geq 0$ small enough.*

**Proof of Lemma 7.1**. It is sufficient to consider $j = 1$. Since

$$\int U(z) U'(z) K(z) f(x+hz)\,dz \preceq C\Omega_j(x,h) \preceq C \int U(z) U'(z) K(z)\,dz,$$

the eigenvalues are in $[0, C]$ for all $x \in \mathcal{X}$ and $h \geq 0$. We now show that the eigenvalues can be bounded from below by $1/C$. Assumptions K and X-(i) give, uniformly in $x \in \mathcal{X}$,

$$\begin{aligned}
&\int U(z) U'(z) K(z) f(x+hz)\,dz \succeq C\Omega(x,h) \\
&\succeq C \int U(z) U(z)' \mathbb{I}(x+hz \in \mathcal{X}, z \in \mathcal{V}(0,1))\,dz.
\end{aligned}$$

Assume that $h$ is small enough. Assumption X-(ii) gives, with $(t-x)/h = z$,

$$\begin{aligned}
&\int U(z) U(z)' \mathbb{I}(x+hz \in \mathcal{X}, z \in \mathcal{V}(0,1))\,dz \\
&= \int U\left(\frac{t-x}{h}\right) U\left(\frac{t-x}{h}\right)' \mathbb{I}(t \in \mathcal{V}(x,h) \cap \mathcal{X}) \frac{dt}{h^d} \\
&\succeq \int U\left(\frac{t-x}{h}\right) U\left(\frac{t-x}{h}\right)' \mathbb{I}(t \in \mathcal{V}(x_h, \kappa_1 h)) \frac{dt}{h^d} \\
&\succeq \int U(z) U(z)' \mathbb{I}\left(z \in \mathcal{V}\left(\frac{x_h - x}{h}, \kappa_1\right)\right) dz,
\end{aligned}$$

where $\|(x_h - x)/h\| \leq 1$. Hence the eigenvalues of $\int U(z) U'(z) K(z) f(x+hz)\,dz$ and $\Omega_j(x,h)$ are larger than

$$\inf_{y \in \mathcal{V}(0,1)} \min_{b: b'b=1} b' \left( \int U(z) U(z)' \mathbb{I}(z \in \mathcal{V}(y, \kappa_1))\,dz \right) b.$$

Suppose now that this lower bound is equal to 0. This implies that there is a sequence $y_n \in \mathcal{V}(0,1)$ and $b_n$ with $b_n' b_n = 1$ such that

$$\int \left(U(z)' b_n\right)^2 \mathbb{I}(z \in \mathcal{V}(y_n, \kappa_1))\,dz \to 0.$$

By compacity and continuity of $(b,y) \mapsto \int \left(U(z)' b\right)^2 \mathbb{I}(z \in \mathcal{V}(y,\kappa_1))\, dz$, this implies that there is a $y \in \mathcal{V}(0,1)$ and $b$ with $b'b = 1$ such that

$$\int \left(U(z)' b\right)^2 \mathbb{I}(z \in \mathcal{V}(y,\kappa_1))\, dz = 0.$$

Hence $U(z)' b = 0$, but this is impossible since $b \neq 0$. Hence the eigenvalues of $\int U(z) U'(z) K(z) f(x+hz)\, dz$ are in $[1/C, C]$ for all $x \in \mathcal{X}$ and $h$ small enough. □

## 7.1 Theorems 3.1, 3.2, and Proposition 3.3

A technical challenge comes from the fact that $\overline{\beta}(x,h)$ is not explicit but defined through (12). The next lemma is the key tool to study the bias term $\left[\overline{\beta}(x,h) - \beta(x)\right]$ when using the first-order condition which characterizes $\overline{\beta}(x,h)$. In this lemma, $D$ is an integer number and $\|\cdot\|$ stands for the Euclidean norm over $\mathbb{R}^D$ or for the associated operator norm.

**Lemma 7.2 (Newton-Kantorovich)** *Let $\mathcal{F}(\cdot) : \mathbb{R}^D \to \mathbb{R}$ be a twice continuously differentiable convex function with a unique minimizer $\overline{\mathbf{b}}$. Suppose that*

1. *There is $\mathbf{b}^* \in \mathbb{R}^D$ such that $\left\|\mathcal{F}^{(1)}(\mathbf{b}^*)\right\| \leq \eta$ and $\left\|\left[\mathcal{F}^{(2)}(\mathbf{b}^*)\right]^{-1}\right\| \leq C_0$;*
2. *$\left\|\mathcal{F}^{(2)}(b) - \mathcal{F}^{(2)}(b')\right\| \leq C_1 \|b - b'\|$ for all $b, b' \in \mathbb{R}^D$;*
3. *$C_0^2 C_1 \eta \leq 1/4$.*

*Then $\left\|\mathbf{b}^* - \overline{\mathbf{b}}\right\| \leq 2C_0\eta$.*

**Proof of Lemma 7.2**. This follows from conclusion 3 in the Newton-Kantorovich Theorem stated in Gragg and Tapia (1974). □

It is convenient to rescale $\beta$ with $H^{-1}$ and to set $b = H^{-1}\beta$, where $H$ is a diagonal matrix with entries $h^{|\pi|}$. This gives in particular

$$U(X_i - x)' \beta = U\left(\frac{X_i - x}{h}\right)' H\beta = U\left(\frac{X_i - x}{h}\right)' b.$$

Lemma 7.2 will be applied for $\overline{\mathbf{b}} = H\overline{\beta}(x;h) = \overline{b}(x;h)$, where $\overline{\beta}(x;h)$ is as in (12) and we now define a candidate $\mathbf{b}^*$. Define

$$s_p = \begin{cases} s & \text{for the proof of Theorem 3.1} \\ p & \text{for the proof of Theorem 3.2} \end{cases},$$

and $b^*(x,h) = (b^*_\pi(x,h), |\pi| \leq p)'$ with

$$b^*_\pi(x,h) = \begin{cases} \frac{h^{|\pi|}\pi_1! \cdots \pi_d! \mu^{(\pi)}(x)}{|\pi|!} & \text{for } |\pi| \leq \lfloor s_p \rfloor, \\ 0 & \text{for } \lfloor s_p \rfloor < |\pi| \leq p \end{cases}.$$

In other words, when proving Theorem 3.1, $b^*(x,h)$ completes the entries $h^{|\pi|}\mu^{(\pi)}(x)/(\pi_1!\cdots\pi_d!)$, $|\pi|\leq\lfloor s\rfloor$, with entries equal to 0 whereas in the proof of Theorem 3.2 the entries of $b^*(x,h)$ are all the $h^{|\pi|}\mu^{(\pi)}(x)/(\pi_1!\cdots\pi_d!)$, $|\pi|\leq p$. The Taylor formula implies, under Assumptions S and X which implies that $\mathcal{X}$ has no isolated points,

$$\max_{(x,z)\in\mathcal{X}\times\operatorname{Supp}K, x+hz\in\mathcal{X}}\left|\mu(x+hz)-U(z)'b^*(x,h)\right|\leq Ch^{s_p}, \tag{21}$$

with $C=C'L$ under Assumption S1.

Define now

$$\boldsymbol{R}_h(b|x)=\int R\left(U(z)'b|x+hz\right)f(x+hz)K(z)\,dz,$$

where $x+hz$ stands for $X$. The Lebesgue Dominated Convergence Theorem gives under Assumptions R, X and K that $b\mapsto\boldsymbol{R}_h(b|x)$ is twice continuously differentiable with first and second derivatives

$$\boldsymbol{R}_h^{(1)}(b|x)=\int R^{(1)}\left(U(z)'b|x+hz\right)U(z)f(x+hz)K(z)\,dz \text{ and}$$
$$\boldsymbol{R}_h^{(2)}(b|x)=\int R^{(2)}\left(U(z)'b|x+hz\right)U(z)U(z)'f(x+hz)K(z)\,dz.$$

The next lemma shows that the matrix $\boldsymbol{R}_h^{(2)}(b|x)$ satisfies some of the conditions of Lemma 7.2.

**Lemma 7.3** *Under Assumptions K, R, and X and for $h\leq\kappa_0/\kappa_1$, there is a $C>1$ such that the eigenvalues of $\boldsymbol{R}_h^{(2)}(b|x)$ are in $[1/C,\infty)$ for all $b$ and all $x\in\mathcal{X}$ and, for all $b$, $b'$ and all $x\in\mathcal{X}$, $\left\|\boldsymbol{R}_h^{(2)}(b|x)-\boldsymbol{R}_h^{(2)}(b'|x)\right\|\leq C\|b-b'\|$.*

**Proof of Lemma 7.3**. The bound for $\left\|\boldsymbol{R}_h^{(2)}(b|x)-\boldsymbol{R}_h^{(2)}(b'|x)\right\|$ follows from Assumption R:

$$\begin{aligned}
&\left\|\boldsymbol{R}_h^{(2)}(b|x)-\boldsymbol{R}_h^{(2)}(b'|x)\right\|\\
&\leq\int\left|R^{(2)}\left(U(z)'b|x+hz\right)-R^{(2)}\left(U(z)'b'|x+hz\right)\right|\left\|U(z)U(z)'\right\|f(x+hz)K(z)\,dz\\
&\leq C\int\left|U(z)'\left(b-b'\right)\right|\|U(z)\|^2f(x+hz)K(z)\,dz\leq C\int\left\|b-b'\right\|\|U(z)\|^3f(x+hz)K(z)\,dz\\
&\leq C\left\|b-b'\right\|,
\end{aligned}$$

since the support of $K(\cdot)$ is compact and $f(\cdot)$ is bounded. For the lower bound of the eigenvalues of $\boldsymbol{R}_h^{(2)}(b|x)$, observe

$$\boldsymbol{R}_h^{(2)}(b|x)\succeq C\int U(z)U(z)'K(z)f(x+hz)\,dz,$$

so that the result follows from Lemma 7.1. $\square$

Let us now return to the proof of Theorems 3.1 and 3.2. (12) gives $\overline{\beta}(x;h) = \arg\min_\beta \boldsymbol{R}_h\left(H^{-1}\beta|x\right)$ so that $\bar{b}(x;h)$ satisfies the first order condition:

$$\boldsymbol{R}_h^{(1)}\left(\bar{b}(x;h)|x\right) = 0 \text{ for all } x \in \mathcal{X}. \tag{22}$$

We now study $\boldsymbol{R}_h^{(1)}(b^*(x;h)|x)$ which satisfies:

$$\begin{aligned}\boldsymbol{R}_h^{(1)}(b^*(x;h)|x) &= \int U(z)\, R^{(1)}\left(U(z)' b^*(x;h)|x+hz\right) f(x+hz)\, K(z)\, dz \\ &= \int U(z)\, R^{(1)}\left(\mu(x+hz) + U(z)' b^*(x;h) - \mu(x+hz)|x+hz\right) f(x+hz)\, K(z)\, dz.\end{aligned}$$

Hence (21), Assumption R and (1) give, uniformly in $x \in \mathcal{X}$

$$\begin{aligned}\left\|\boldsymbol{R}_h^{(1)}(b^*(x;h)|x)\right\| &\leq \left\|\int U(z)\, R^{(1)}(\mu(x+hz)|x+hz)\, f(x+hz)\, K(z)\, dz\right\| \\ &\quad + C \max_{(x,x+hz,z)\in\mathcal{X}^2\times\operatorname{Supp} K} \left|\mu(x+hz) - U(z)' b^*(x,h)\right| \\ &\leq Ch^{s_p}.\end{aligned}$$

Then Lemma 7.2 shows that Theorem 3.1 is proved.

For Theorem 3.2, recall that $s_p = p$. A Taylor expansion of order $p+1$ gives that uniformly in $x \in \mathcal{X}$

$$\mu(x+hz) - U(z)' b^*(x,h) = h^{p+1} \sum_{|\pi|=p+1} \frac{\pi_1! \cdots \pi_d!}{(p+1)!} z^\pi \mu^{(\pi)}(x) + o\left(h^{p+1}\right).$$

Recall that $\boldsymbol{R}_{f,\mu}^{(2)}(x) = R^{(2)}(\mu(x)|x)\, f(x)$ is bounded away from 0. Hence (22), (1) and standard uniform expansions give, for $\mathbb{I}_{\mathcal{X}}(x+hz) = \mathbb{I}(x+hz \in \mathcal{X})$,

$$\begin{aligned}0 &= \int U(z)\, R^{(1)}\left(\mu(x+hz) + U(z)' \bar{b}(x;h) - \mu(x+hz)|x+hz\right) f(x+hz)\, K(z)\, dz \\ &= \underbrace{\int U(z)\, R^{(1)}(\mu(x+hz)|x+hz)\, f(x+hz)\, K(z)\, dz}_{=0} \\ &\quad + \int \left(\boldsymbol{R}_{f,\mu}^{(2)}(x+hz) + o(1)\right) U(z) \left(U(z)' \bar{b}(x;h) - \mu(x+hz)\right) \mathbb{I}_{\mathcal{X}}(x+hz)\, K(z)\, dz \\ &= \left(\boldsymbol{R}_{f,\mu}^{(2)}(x) + o(1)\right) \int U(z)\, U(z)' \left(\bar{b}(x;h) - b^*(x;h)\right) \mathbb{I}_{\mathcal{X}}(x+hz)\, K(z)\, dz \\ &\quad + \left(\boldsymbol{R}_{f,\mu}^{(2)}(x) + o(1)\right) \int U(z) \left\{h^{p+1} \sum_{|\pi|=p+1} \frac{\pi_1! \cdots \pi_d!}{(p+1)!} z^\pi \mu^{(\pi)}(x)\right\} \mathbb{I}_{\mathcal{X}}(x+hz)\, K(z)\, dz \\ &\quad + o\left(h^{p+1}\right).\end{aligned}$$

Rearranging gives,

$$\begin{aligned}\bar{b}(x;h) - b^*(x;h) = h^{p+1}&\left(\int U(z)\,U(z)'\,K(z)\,\mathbb{I}_{\mathcal{X}}(x+hz)\,dz\right)^{-1}\\&\times\int U(z)\left\{\sum_{|\pi|=p+1}\frac{\pi_1!\cdots\pi_d!}{(p+1)!}z^{\pi}\mu^{(\pi)}(x)\right\}\mathbb{I}_{\mathcal{X}}(x+hz)\,K(z)\,dz\\&+o\left(h^{p+1}\right),\end{aligned}$$

showing that Theorem 3.2 is proved.

For Proposition 3.3, performing a Taylor expansion of order $p+2$ and arguing as above gives, for inner $x$,

$$\bar{b}(x;h) - b^*(x;h) = \frac{(1+o(1))}{\boldsymbol{R}^{(2)}_{f,\mu}(x)}\Omega^{-1}\boldsymbol{B}(x;h) + o\left(h^{p+2}\right),$$

where

$$\begin{aligned}\boldsymbol{B}(x;h) &= \int U(z)\left\{h^{p+1}\boldsymbol{R}^{(2)}_{f,\mu}(x+hz)\sum_{|\pi|=p+1}\frac{\pi_1!\cdots\pi_d!}{(p+1)!}z^{\pi}\mu^{(\pi)}(x)\right\}K(z)\,dz\\&+\int U(z)\left\{h^{p+2}\boldsymbol{R}^{(2)}_{f,\mu}(x)\sum_{|\pi|=p+2}\frac{\pi_1!\cdots\pi_d!}{(p+1)!}z^{\pi}\mu^{(\pi)}(x)\right\}K(z)\,dz\\&= h^{p+1}\boldsymbol{R}^{(2)}_{f,\mu}(x)\,b_{p+1}(x) + h^{p+2}\sum_{j=1}^{d}\frac{\partial\boldsymbol{R}^{(2)}_{f,\mu}(x)}{\partial x_j}\mu^{(\pi)}(x)\int z_j z^{\pi}U(z)\,K(z)\,dz + o\left(h^{p+2}\right)\\&+ h^{p+2}\boldsymbol{R}^{(2)}_{f,\mu}(x)\,b_{p+2}(x),\end{aligned}$$

This gives the expansion of the Proposition since $e_{\pi}'\Omega_1^{-1}b_{p+1}(x) = 0$ by Gu et al. (2015). □

## 7.2 Theorem 3.4

Let

$$\begin{aligned}\boldsymbol{R}^{(2)}(x,h) &= h^{-d}\mathbb{E}\left[R^{(2)}(\mu(x)\,|X)\,U\left(\frac{X-x}{h}\right)U\left(\frac{X-x}{h}\right)'K\left(\frac{X-x}{h}\right)\right],\\\boldsymbol{S}(x,h) &= h^{-d}\mathbb{E}\left[\left(\rho^{(1)}(Y-\mu(X))\right)^2 U\left(\frac{X-x}{h}\right)U\left(\frac{X-x}{h}\right)'K^2\left(\frac{X-x}{h}\right)\right],\end{aligned}$$

Theorems 3.1 and 3.2, (21), give that

$$\max_{(x,z)\in\mathcal{X}\times\operatorname{Supp}K,x+hz\in\mathcal{X}}\left|U(hz)'\,\beta(x,h) - \mu(x+hz)\right| = o(1).$$

Therefore $\max_{x,h}\operatorname{Sp}\left(h^{-d}\mathbb{E}\left[U\left(\frac{X-x}{h}\right)U\left(\frac{X-x}{h}\right)'K^j\left(\frac{X-x}{h}\right)\right]\right) < \infty$, $j=1,2$, Assumption R-(ii), and uniform continuity of $\mu(\cdot)$ over $\mathcal{X}$ give when $h\to 0$,

$$\max_{x\in\mathcal{X}}\left\|\overline{\boldsymbol{R}}^{(2)}(x,h) - \boldsymbol{R}^{(2)}(x,h)\right\| = o(1),\quad \max_{x\in\mathcal{X}}\left\|\overline{\boldsymbol{S}}(x,h) - \boldsymbol{S}(x,h)\right\| = o(1).$$

A standard change of variable gives

$$\begin{aligned}\boldsymbol{R}^{(2)}(x,h) &= \int R^{(2)}(\mu(x)|x+hz)\,U(z)\,U(z)'\,K(z)\,f(x+hz)\,dx\\ &= R^{(2)}(\mu(x)|x)\,f(x)\,\Omega_1(x,h)+o(1),\\ \boldsymbol{S}(x,h) &= \int \mathbb{E}\left[\left(\rho^{(1)}(Y-\mu(X))\right)^2|X=x+hz\right]U(z)\,U(z)'\,K^2(z)\,f(x+hz)\,dx\\ &= \mathbb{E}\left[\left(\rho^{(1)}(Y-\mu(X))\right)^2|X=x\right]f(x)\,\Omega_2(x,h)+o(1),\end{aligned}$$

uniformly over $\mathcal{X}$ by Assumption R-(ii) and X-(i). Hence Lemma 7.1 yields that $\overline{V}(x,h)$ is equal to

$$\begin{aligned}&\left(R^{(2)}(\mu(x)|x)\,f(x)\,\Omega_1(x,h)\right)^{-1}\mathbb{E}\left[\left(\rho^{(1)}(Y-\mu(X))\right)^2|X=x\right]\\ &\qquad\times f(x)\,\Omega_2(x,h)\left(R^{(2)}(\mu(x)|x)\,f(x)\,\Omega_1(x,h)\right)^{-1}+o(1)\\ &=\frac{\mathbb{E}\left[\left(\rho^{(1)}(Y-\mu(X))\right)^2|X=x\right]}{\left(R^{(2)}(\mu(x)|x)\right)^2 f(x)}\Omega(x,h)^{-1}+o(1),\end{aligned}$$

uniformly over $\mathcal{X}$, that is the first approximation in the Theorem. The second approximation follows since $\Omega(x,h)=\Omega$ for all $x$ in subset $\mathcal{X}_0$ of $\mathcal{X}$ as in the theorem provided $h$ is small enough. □

### 7.3 Lemma 4.1, Propositions 4.2 and 4.3

The proof of these results makes use of the Bernstein inequality, which states that for independent centered real random variables $Z_i$ with $|Z_i|\le M$,

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\right|\ge t\right)\le 2\exp\left(-\frac{t^2/2}{\frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}(Z_i)+\frac{1}{3}\frac{Mt}{\sqrt{n}}}\right)\text{ for any } t\ge 0. \tag{23}$$

**Proof of Lemma 4.1**. It is sufficient to consider $j=1$. To prove (19), it is sufficient to show that for any $\epsilon>0$ large enough, all $\pi$ with $|\pi|\le 2p$, $K_\pi(z)=z^\pi K(z)$ and $r_n=\left(\log n/\left(nh^d\right)\right)^{1/2}$,

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}\left|\frac{1}{nh^d}\sum_{i=1}^{n}K_\pi\left(\frac{X_i-x}{h}\right)-\int K_\pi(z)\,f(x+hz)\,dz\right|\ge r_n\epsilon\right)\to 0. \tag{24}$$

By Assumptions K, X and $1/h^d=O(n/\log n)$, there is a $\delta=\delta_n=n^{-a}$ such that

1. There is an integer number $J_n=O\left(n^b\right)$, $b>0$, and some $x_j\in\mathcal{X}$ such that $\mathcal{X}=\bigcup_{j=1}^{J_n}\mathcal{V}_X(x_j,\delta_n)$, where $\mathcal{V}_X(x_j,\delta_n)=\mathcal{V}(x_j,\delta_n)\cap\mathcal{X}$;
2. For all $x$, $x'$ with $\|x-x'\|\le\delta_n$ and all $i$ and $n$, $\left|K_\pi\left(\frac{X_i-x}{h}\right)-K_\pi\left(\frac{X_i-x'}{h}\right)\right|\le h^d r_n\epsilon/3$;
3. For all $x$, $x'\in\mathcal{X}$ with $\|x-x'\|\le\delta_n$ and all $n$, $\left|\int K_\pi(z)\,f(x+hz)\,dz-\int K_\pi(z)\,f(x'+hz)\,dz\right|\le r_n\epsilon/3$.

This gives

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\frac{1}{nh^d}\sum_{i=1}^{n}K_\pi\left(\frac{X_i-x}{h}\right)-\int K_\pi\left(z\right)f\left(x+hz\right)dz\right| \\
&\leq \max_{j=1,\ldots,J_n}\left|\frac{1}{nh^d}\sum_{i=1}^{n}K_\pi\left(\frac{X_i-x_j}{h}\right)-\int K_\pi\left(z\right)f\left(x_j+hz\right)dz\right| \\
&+\max_{j=1,\ldots,J_n}\sup_{x\in\mathcal{V}_X(x_j,\delta_n)}\left|\frac{1}{nh^d}\sum_{i=1}^{n}\left(K_\pi\left(\frac{X_i-x}{h}\right)-K_\pi\left(\frac{X_i-x_j}{h}\right)\right)\right| \\
&+\max_{j=1,\ldots,J_n}\sup_{x\in\mathcal{V}_X(x_j,\delta_n)}\left|\int K_\pi\left(z\right)\left(f\left(x\right)-f\left(x_j+hz\right)\right)dz\right| \\
&\leq \max_{j=1,\ldots,J_n}\left|\frac{1}{nh^d}\sum_{i=1}^{n}K_\pi\left(\frac{X_i-x_j}{h}\right)-\int K_\pi\left(z\right)f\left(x_j+hz\right)dz\right|+\frac{2}{3}r_n\epsilon.
\end{aligned}$$

Hence (24) holds if

$$\mathbb{P}\left(\max_{j=1,\ldots,J_n}\left|\frac{1}{nh}\sum_{i=1}^{n}K_\pi\left(\frac{X_i-x_j}{h}\right)-\int K_\pi\left(z\right)f\left(x_j+hz\right)dz\right|\geq r_n\frac{\epsilon}{3}\right)\to 0.$$

Since an elementary change of variables gives

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{h^d}K_\pi\left(\frac{X_i-x_j}{h}\right)\right] &= \int K_\pi\left(z\right)f\left(x_j+hz\right)dz \text{ and} \\
\mathrm{Var}\left(\frac{1}{h^{d/2}}K_\pi\left(\frac{X_i-x_j}{h}\right)\right) &\leq \mathbb{E}\left[\frac{1}{h^d}K_\pi^2\left(\frac{X_i-x_j}{h}\right)\right]\leq\frac{C}{h^d},
\end{aligned}$$

(24) follows from the Bonferoni inequality and (23) which give

$$\begin{aligned}
&\mathbb{P}\left(\max_{j=1,\ldots,J_n}\left|\frac{1}{nh^d}\sum_{i=1}^{n}K_\pi\left(\frac{X_i-x_j}{h}\right)-\int K_\pi\left(z\right)f\left(x_j+hz\right)dz\right|\geq r_n\frac{\epsilon}{3}\right) \\
&\leq\sum_{j=1}^{J_n}\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}h^{-\frac{d}{2}}K_\pi\left(\frac{X_i-x_j}{h}\right)-\int K_\pi\left(z\right)f\left(x_j+hz\right)dz\right|\geq\log^{1/2}n\frac{\epsilon}{3}\right) \\
&\leq 2J_n\exp\left(-\frac{\epsilon^2\log n}{C+C/\left(nh^d\right)^{1/2}}\right)\leq 2\exp\left(-\frac{\epsilon^2\log n-b\log n}{C}\right)\to 0,
\end{aligned}$$

for $\epsilon^2>b$. Hence (19) is proved. The existence of the inverse matrix stated in the Lemma and the uniform bound for its spectral radius follow from Lemma 7.1. □

**Proof of Proposition 4.2.** Let

$$\begin{aligned}
\widehat{M}_n\left(x,h\right) &= \frac{1}{nh^d}\sum_{i=1}^{n}U\left(\frac{X_i-x}{h}\right)U'\left(\frac{X_i-x}{h}\right)K\left(\frac{X_i-x}{h}\right) \text{ and} \\
M\left(x,h\right) &= h^{-d}\mathbb{E}\left[U\left(\frac{X_i-x}{h}\right)U'\left(\frac{X_i-x}{h}\right)K\left(\frac{X_i-x}{h}\right)\right].
\end{aligned}$$

We first show that $\left(nh^{d}\right)^{1/2} V\left(x_n,h\right)^{-1/2} H\left(\widehat{\beta}\left(x_n;h\right)-\overline{\beta}\left(x_n;h\right)\right)$ is asymptotically a standard normal. Observe that

$$\begin{aligned}
&\left(nh^{d}\right)^{1/2} V\left(x_n,h\right)^{-1/2} H\left(\widehat{\beta}\left(x_n;h\right)-\overline{\beta}\left(x_n;h\right)\right) \\
&= V\left(x_n,h\right)^{-1/2} \widehat{M}_n\left(x_n,h\right)^{-1} \frac{\sum_{i=1}^{n} U\left(\frac{X_i-x_n}{h}\right)\xi_i\left(x_n;h\right) K\left(\frac{X_i-x_n}{h}\right)}{\left(nh^{d}\right)^{1/2}} \\
&+ V\left(x_n,h\right)^{-1/2} \widehat{M}_n\left(x_n,h\right)^{-1} \frac{\sum_{i=1}^{n} U\left(\frac{X_i-x_n}{h}\right)\varepsilon_i K\left(\frac{X_i-x_n}{h}\right)}{\left(nh^{d}\right)^{1/2}},
\end{aligned}$$

where

$$\xi_i\left(x;h\right) = m\left(X_i\right) - U\left(X_i-x\right)'\overline{\beta}\left(x;h\right).$$

Since, for all $x\in\mathcal{X}$, $\overline{\beta}\left(x;h\right)$ satisfies the first-order condition

$$\mathbb{E}\left[U\left(X_i-x\right)\left(m\left(X_i\right)-U\left(X_i-x\right)'\overline{\beta}\left(x;h\right)\right)K\left(\frac{X_i-x}{h}\right)\right]=0, \tag{25}$$

the variables $U\left(\frac{X_i-x_n}{h}\right)\xi_i\left(x_n;h\right)K\left(\frac{X_i-x_n}{h}\right)$ are centered. Moreover, Theorems 3.1 and 3.2 together with (21), Assumptions X-(iii) and K and Lemma 4.1 give that

$$\begin{aligned}
&\operatorname{Var}\left(h^{-d/2}U\left(X_i-x\right)\left(m\left(X_i\right)-U\left(X_i-x\right)'\overline{\beta}\left(x;h\right)\right)K\left(\frac{X_i-x}{h}\right)\right) \\
&= O\left(h^{2s}\right)\int U\left(z\right)U\left(z\right)'K^2\left(z\right)f\left(x+hz\right)dz = o\left(1\right),
\end{aligned}$$

uniformly in $x\in\mathcal{X}$. Hence

$$V\left(x_n,h\right)^{-1/2} \widehat{M}_n\left(x_n,h\right)^{-1} \frac{\sum_{i=1}^{n} U\left(\frac{X_i-x_n}{h}\right)\xi_i\left(x_n;h\right) K\left(\frac{X_i-x_n}{h}\right)}{\left(nh^{d}\right)^{1/2}} = o_p\left(1\right),$$

and it is sufficient to show that

$$V\left(x_n,h\right)^{-1/2} \widehat{M}_n\left(x_n,h\right)^{-1} \frac{\sum_{i=1}^{n} U\left(\frac{X_i-x_n}{h}\right)\varepsilon_i K\left(\frac{X_i-x_n}{h}\right)}{\left(nh^{d}\right)^{1/2}} \xrightarrow{d} N\left(0,I_d\right).$$

This follows from Theorem 3.4, Assumption E and the Lindeberg Central Limit Theorem for triangular arrays. To complete the proof of the Proposition, it is now sufficient to show that $\widehat{V}\left(x_n,h\right)=V\left(x_n,h\right)+o_p\left(1\right)$. This follows from Lemma 4.1 and $\widehat{\sigma}^2\left(x_n\right)=\sigma^2\left(x_n\right)+o_p\left(1\right)$ as established now. Let

$$\widetilde{\sigma}^2\left(x\right)=\frac{\sum_{i=1}^{n}\left(Y_i-\mu\left(x\right)\right)^2 K\left(\frac{X_i-x}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)} \text{ and } \overline{\sigma}^2\left(x\right)=\frac{\sum_{i=1}^{n}\sigma^2\left(X_i\right)K\left(\frac{X_i-x}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)}.$$

The asymptotic normality above, Theorems 3.1 and 3.2, yield that

$$\begin{aligned}
\left|\widehat{\sigma}\left(x_n\right)-\widetilde{\sigma}\left(x_n\right)\right| &\leq \left(\frac{\sum_{i=1}^n\left(\widehat{\beta}_0\left(x_n, h\right)-\mu\left(x_n\right)\right)^2 K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)}\right)^{1/2} \\
&= \left|\widehat{\beta}_0\left(x_n, h\right)-\mu\left(x_n\right)\right| = O_p\left(\frac{1}{\left(nh^d\right)^{1/2}}+h^s\right) = o_p(1).
\end{aligned}$$

Let $\delta_i=\varepsilon_i^2-\sigma^2\left(X_i\right)$. Observe that

$$\widetilde{\sigma}^2\left(x_n\right)=\overline{\sigma}^2\left(x_n\right)+O_p\left(\frac{\sum_{i=1}^n\left(\mu\left(X_i\right)-\mu\left(x_n\right)\right)^2 K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)}\right)+O_p\left(\frac{\sum_{i=1}^n \delta_i K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)}\right),$$

with, by uniform continuity of $\mu(\cdot)$ over the compact $\mathcal{X}$ and since $h \rightarrow 0$,

$$\frac{\sum_{i=1}^n\left(\mu\left(X_i\right)-\mu\left(x_n\right)\right)^2 K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)}=o(1).$$

Let $\mathbb{E}_n[\cdot]$ be the conditional expectation given $X_1, \ldots, X_n$ and $\boldsymbol{i}=\sqrt{-1}$. Then, under Assumption E and assuming w.l.o.g. that $\nu \leq 2$, Assumptions X-(i) and K, and by standard manipulations involving uniform $O(\cdot)$ terms, we have for any $t$,

$$\begin{aligned}
&\left|\mathbb{E}_n\left[\exp\left(\boldsymbol{i}t\frac{\sum_{i=1}^n \delta_i K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)}\right)\right]\right| = \left|\prod_{i=1}^n \mathbb{E}_n\left[\exp\left(\boldsymbol{i}t\frac{\delta_i K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)}\right)\right]\right| \\
&= \left|\prod_{i=1}^n \exp\left\{\ln\left(1-\boldsymbol{i}t\mathbb{E}_n\left[\frac{\delta_i K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)}\right]+|t|^{1+\nu/2} O\left(\mathbb{E}_n\left[\left|\frac{\delta_i K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)}\right|^{1+\nu/2}\right]\right)\right)\right\}\right| \\
&\leq \exp\left(-C|t|^{1+\nu/2}\sum_{i=1}^n\left(\frac{K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)}\right)^{1+\nu/2}\right) = \exp\left(O_p\left((nh)^{-\nu/2}\right)\right) \stackrel{p}{\rightarrow} 1.
\end{aligned}$$

This implies that

$$\frac{\sum_{i=1}^n \delta_i K\left(\frac{X_i-x_n}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x_n}{h}\right)} \stackrel{d}{\rightarrow} 0 \text{ and then } \stackrel{p}{\rightarrow} 0,$$

so that $\widehat{\sigma}^2\left(x_n\right)=\overline{\sigma}^2\left(x_n\right)+o_p(1)$, with $\overline{\sigma}^2\left(x_n\right)=\sigma^2\left(x_n\right)+o_p(1)$ by uniform continuity of $\sigma(\cdot)$ over the compact $\mathcal{X}$. Hence $\widehat{\sigma}^2\left(x_n\right)=\sigma^2\left(x_n\right)+o_p(1)$. □

**Proof of Proposition 4.3.** Let $K_\pi(z)=z^\pi K(z)$,

$$S_n(x;h)=\frac{1}{\left(nh^d\right)^{1/2}}\sum_{i=1}^n\left(\xi_i(x;h)+\varepsilon_i\right)K_\pi\left(\frac{X_i-x}{h}\right), \quad \xi_i(x;h)=m\left(X_i\right)-U\left(\frac{X_i-x}{h}\right)^{\prime} H\overline{\beta}(x;h).$$

Lemma 4.1 implies that it is sufficient to show that

$$\sup_{x\in\mathcal{X}} |S_n(x;h)| = O_p\left(\log^{1/2} n\right). \tag{26}$$

Define, for $\eta_i = \varepsilon_i \mathbb{I}(|\varepsilon_i| < \tau_n) - \mathbb{E}[\varepsilon_i \mathbb{I}(|\varepsilon_i| < \tau_n)|X_i]$,

$$s_n(x;h) = \frac{1}{(nh^d)^{1/2}} \sum_{i=1}^n (\xi_i(x;h) + \eta_i) K_\pi\left(\frac{X_i - x}{h}\right).$$

The Chebychev inequality gives, for $\tau_n = C_\tau n^{1/\nu}$,

$$\mathbb{P}\left(\max_{i=1,\dots,n} |\varepsilon_i| \geq \tau_n\right) \leq \sum_{i=1}^n \mathbb{P}(|\varepsilon_i| \geq \tau_n) \leq \frac{\mathbb{E}[|\varepsilon_i|^\nu]}{C_\tau^\nu},$$

which can be made arbitrarily small by increasing $C_\tau$. Arguing as in the proof of Lemma 4.1 gives, when $\max_{i=1,\dots,n}|\varepsilon_i| \geq \tau_n$ and since $-\mathbb{E}[\varepsilon_i \mathbb{I}(|\varepsilon_i| < \tau_n)|X_i] = \mathbb{E}[\varepsilon_i \mathbb{I}(|\varepsilon_i| \geq \tau_n)|X_i]$,

$$\begin{aligned}
\sup_{x\in\mathcal{X}} |S_n(x;h) - s_n(x;h)| &= \sup_{x\in\mathcal{X}} \left| \frac{1}{(nh^d)^{1/2}} \sum_{i=1}^n \mathbb{E}[\varepsilon_i \mathbb{I}(|\varepsilon_i| \geq \tau_n)|X_i] K_\pi\left(\frac{X_i - x}{h}\right) \right| \\
&\leq \sup_{x\in\mathcal{X}} \frac{1}{(nh^d)^{1/2}} \sum_{i=1}^n \mathbb{E}\left[\frac{|\varepsilon_i|^\nu \mathbb{I}(|\varepsilon_i| \geq \tau_n)}{\tau_n^{\nu-1}} |X_i\right] \left|K_\pi\left(\frac{X_i - x}{h}\right)\right| \\
&= O_p\left(\frac{(nh^d)^{1/2}}{n^{1-1/\nu}}\right) = o_p\left(\frac{1}{n^{1/2-1/\nu}}\right) = o_p(1),
\end{aligned}$$

since $\nu > 2$ by Assumption E. Therefore it is sufficient to show that

$$\sup_{x\in\mathcal{X}} |s_n(x;h)| = O_p\left(\log^{1/2} n\right) \tag{27}$$

to prove (26). Observe that the expectation of the summands in $s_n(x;h)$ is 0 by (25) and the definition of $\eta_i$. (25) also gives

$$\begin{aligned}
H\overline{\beta}(x;h) = &\left(\mathbb{E}\left[U\left(\frac{X_i - x}{h}\right) U\left(\frac{X_i - x}{h}\right)' K\left(\frac{X_i - x}{h}\right)\right]\right)^{-1} \\
&\times \mathbb{E}\left[U\left(\frac{X_i - x}{h}\right) m(X_i) K\left(\frac{X_i - x}{h}\right)\right]
\end{aligned}$$

so that Assumption K gives that $\left\|H\left(\overline{\beta}(x;h) - \overline{\beta}(x';h)\right)\right\| \leq Ch^{-1}\|x - x'\|$ for all $x$, $x' \in \mathcal{X}$. This also gives, by (21) which ensures that $\max_{x\in\mathcal{X}} |\xi_i(x;h)| \leq C$,

$$\left\|(\xi_i(x;h) + \eta_i) K_\pi\left(\frac{X_i - x}{h}\right) - (\xi_i(x';h) + \eta_i) K_\pi\left(\frac{X_i - x'}{h}\right)\right\| \leq C\frac{\tau_n}{h}\|x - x'\|,$$

for all $i = 1, \dots, n$ and all $n$, all $x$, $x' \in \mathcal{X}$,

$$|s_n(x;h) - s_n(x';h)| \leq C\frac{n^{1/2}\tau_n}{h^{d/2+1}}\|x - x'\|.$$

This together Assumptions K, X and $1/h^d = O(n/\log n)$, there is a $\delta = \delta_n = n^{-a}$ such that

1. There is an integer number $J_n = O\left(n^b\right)$, $b > 0$, and some $x_j \in \mathcal{X}$ such that $\mathcal{X} = \bigcup_{j=1}^{J_n} \mathcal{V}_X\left(x_j, \delta_n\right)$, where $\mathcal{V}_X\left(x_j, \delta_n\right) = \mathcal{V}\left(x_j, \delta_n\right) \cap \mathcal{X}$;

2. For all $x$, $x'$ with $\|x - x'\| \leq \delta_n$ and all $i$ and $n$, $|s_n(x;h) - s_n(x';h)| \leq 1$.

Hence

$$\begin{aligned}\sup_{x\in\mathcal{X}} |s_n(x;h)| &\leq \max_{j=1,\ldots,J_n} |s_n(x_j;h)| + \max_{j=1,\ldots,J_n} \sup_{x\in\mathcal{V}_X(x_j,\delta_n)} |s_n(x;h) - s_n(x_j;h)| \\ &\leq \max_{j=1,\ldots,J_n} |s_n(x_j;h)| + 1.\end{aligned}$$

As a consequence (27) holds if $\max_{j=1,\ldots,J_n} |s_n(x_j;h)| = O_p\left(\log^{1/2} n\right)$. The Bonferoni inequality and (23) give, using $\max_{x\in\mathcal{X}} |\xi_i(x;h)| \leq C$ and $h^{-d} = O\left(n^{1-2/\nu}\log n\right)$,

$$\begin{aligned}&\mathbb{P}\left(\max_{j=1,\ldots,J_n} |s_n(x_j;h)| \geq t\log^{1/2} n\right) \leq \sum_{j=1}^{n} \mathbb{P}\left(|s_n(x_j;h)| \geq t\log^{1/2} n\right) \\ &\leq 2J_n \exp\left(-\frac{t\log n}{C + \left(\frac{h^{-d}\tau_n^2\log n}{n}\right)^{1/2} t}\right) = \exp\left(-\frac{(t-C)\log n}{C'}\right) \to 0\end{aligned}$$

for $t > C$. This ends the proof of Proposition 4.3. □

## 7.4 Propositions 4.4, 4.5 and 4.6

**Proof of Proposition 4.4.** Lemma 4.1, Theorems 3.1 and 3.2 which give (21), imply that the conclusions of Lemmas A.1, A.2 and A.3 in Guerre and Sabbah (2012) are true when $x \in \mathcal{X}$. Hence the first equation in Proposition 4.4 follows from minor modifications of the proof of Theorem 2 in Guerre and Sabbah (2012). The second equation follows from $\max_{(\alpha,x)\in[\underline{\alpha},\overline{\alpha}]\times\mathcal{X}} \left\|\widehat{J}(\alpha|x,h) - \overline{J}(\alpha|x,h)\right\| = O_{\mathbb{P}}\left(\left(\log n/(nh^d)\right)^{1/2}\right)$ and $\max_{(\alpha,x)\in[\underline{\alpha},\overline{\alpha}]\times\mathcal{X}} \left\|\widehat{S}(\alpha|x,h)\right\| = O_p\left(\log^{1/2} n\right)$ as established below. □

**Proof of Proposition 4.6**. Theorems 3.1 and 3.2 which give (21) are sufficient to show that the conclusion of Lemma A.3 in Guerre and Sabbah (2012) holds for $x \in \mathcal{X}$, that is

$$\max_{(\alpha,x)\in[\underline{\alpha},\overline{\alpha}]\times\mathcal{X}} \left\|\widehat{S}(\alpha|x,h)\right\| = O_p\left(\log^{1/2} n\right).$$

Hence Proposition 4.4 and Lemma 4.1 give

$$\begin{aligned}&\sup_{(\alpha,x)\in[\underline{\alpha},\overline{\alpha}]\times\mathcal{X}} \left\|\left(nh^d\right)^{1/2} H\left(\widehat{\beta}(\alpha|x,h) - \overline{\beta}(\alpha|x,h)\right)\right\| \\ &= O_p\left(\max_{(\alpha,x)\in[\underline{\alpha},\overline{\alpha}]\times\mathcal{X}} \left\|\frac{\widehat{S}(\alpha|x,h)}{(nh^d)^{1/2}}\right\|\right) + O_p\left(\left(\frac{\log n}{nh^d}\right)^{3/4}\right) = O_p\left(\left(\frac{\log n}{nh^d}\right)^{1/2}\right).\end{aligned}$$

Hence Theorems 3.1 and 3.2 show that the Proposition is proved. □

**Proof of Proposition 4.5**. As in the proof of Proposition 4.2, the key issue here is to show that $\widehat{V}(\alpha|x_n,h) = V(\alpha|x_n,h) + o_p(1)$, and Lemma 4.1 shows that is sufficient to show that $\partial\widehat{Q}(\alpha|x_n)/\partial\alpha$ is consistent. This follows from Proposition 4.6, the fact that $s \geq 1$ and the choice of $\eta$ which gives under Assumption F which ensures that $(\alpha, x) \mapsto \partial Q(\alpha|x)/\partial\alpha$ is continuous,

$$\frac{\partial\widehat{Q}(\alpha|x_n)}{\partial\alpha} = \frac{\widehat{Q}(\alpha+\eta|x_n) - \widehat{Q}(\alpha+\eta|x_n)}{2\eta}$$
$$= \frac{Q(\alpha+\eta|x_n) - Q(\alpha+\eta|x_n)}{2\eta} + O_p\left(\frac{1}{\eta}\left(h + \left(\frac{\log n}{nh^d}\right)^{1/2}\right)\right)$$
$$= \frac{\partial Q(\alpha|x_n)}{\partial\alpha} + o_p(1). \qquad \square$$

## References


[1] Altonji, J.G., Ichimura, H., & Otsu, T. (2012). Estimating Derivatives in Nonseparable Models with Limited Dependent Variables. *Econometrica*, 80, 1701–1719.

[2] Aryal, G., Gabrielli, M.F., & Vuong, Q. (2014). Semiparametric Estimation of First-Price Auction Models. Working Paper.

[3] Banerjee, A.N. (2007). A Method of Estimating the Average Derivative. *Journal of Econometrics*, 136, 65–88.

[4] Bouezmarni, T., & Rombouts, J. (2010). Nonparametric Density Estimation for Multivariate Bounded Data. *Journal of Statistical Planning and Inference*, 140, 139–152.

[5] Bravo, F., & Jacho-Chávez, D.T. (2011). Empirical Likelihood for Efficient Semiparametric Average Treatment Effects. *Econometric Reviews*, 30, 1–24.

[6] Chaudhuri, P. (1991). Nonparametric Estimates of Regression Quantiles and their Local Bahadur Representation. *The Annals of Statistics*, 11, 760–777.

[7] Chaudhuri, P., Doksum, K., & Samarov, A. (1997). On Average Derivative Quantile Regression. *The Annals of Statistics*, 25, 715–744.

[8] Chen, M.-Y., & Wu, H.-T. (2013). Local Linear Regression on Manifolds and Its Geometric Interpretation. *Journal of the American Statistical Association*, 108, 1421–1434.

[9] Fan, J., & Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman & Hall/CRC.

[10] Fan, J., Heckman, N.E., & Wand, M.P. (1995). Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions. *Journal of the American Statistical Association*, 90, 141–150.

[11] Fan, Y., & Li, Q. (1996). Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms. *Econometrica*, 64, 865–890.

[12] Frandsen, B.R., Frölich, M., & Melly, B. (2010). Quantile Treatment Effects in the Regression Discontinuity Design. Working Paper.

[13] Gu, J., Li, Q., & Yang, J.-C. (2015). Multivariate Local Polynomial Kernel Estimators: Leading Bias and Asymptotic Distribution. *Econometric Reviews*, 34, 979–1010.

[14] Guerre, E. (2000). Design Adpative Nearest Neighbor Regression Estimation. *Journal of Multivariate Analysis*, 75, 219–244.

[15] Guerre, E., & Lavergne, P. (2002). Optimal Minimax Rates for Nonparametric Specification Testing in Regression Models. *Econometric Theory*, 18, 1139–1171.

[16] Guerre, E., & Sabbah, C. (2012). Uniform Bias Study and Bahadur Representation for Local Polynomial Estimators of the Conditional Quantile Function. *Econometric Theory*, 28, 87–129.

[17] Gragg, W.B., & Tapia, R.A. (1974). Optimal Error Bounds for the Newton-Kantorovich Theorem. *SIAM Journal on Numerical Analysis*, 11, 10–13.

[18] Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69, 201–209.

[19] Härdle, W., & Mammen, E. (1993). Comparing Nonparametric versus Parametric Regression Fits. *The Annals of Statistics*, 21, 1926–1947.

[20] Härdle, W., & Stoker, T.H. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association*, 84, 986–995.

[21] Hickman, B.R., & Hubbard, T.P. (2015). Replacing Sample Trimming with Boundary Correction in Nonparametric Estimation of First-Price Auctions. *Journal of Applied Econometrics*, 30, 739–762.

[22] Hjellvik, J., Yao, Q., & Tjøstheim, D. (1998). Linearity Testing Using Local Polynomial Approximation. *Journal of Statistical Planning and Inference*, 68, 295–321.

[23] Hoderlein, S., Su, L., White, H., & Yang, T.T. (2015). Testing for Monotonicity in Unobservables under Unconfoundedness. Working Paper.

[24] Huber, P.J. (1967). The Behavior of Maximum Likelihood Estimation under Nonstandard Conditions. *Proceedings of the Fith Berkeley Symposium on Mathematical Statistics and Probability*, Vol 1. University of California Press.

[25] Imbens, G.W., & Lemieux, T. (2008). Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics*, 142, 615–635.

[26] Imbens, G.W., & Zajonc, T. (2011). Regression Discontinuity Design with Multiple Forcing Variables. Working Paper.

[27] Kong, E., Linton, O., & Xia, Y. (2010). Uniform Bahadur Representation for Local Polynomial Estimates of M-regression and its Application to the Additive Model. *Econometric Theory*, 26, 1529–1564.

[28] Li, C.-S. (2005). Using Local Linear Kernel Smoothers to Test the Lack of Fit of Nonlinear Regression Models. *Statistical Methodology*, 2, 267–284.

[29] Li, Q., Lu, X., & Ullah, A. (2003). Multivariate Local Polynomial Regression for Estimating Average Derivatives. *Journal of Nonparametric Statistics*, 15, 607–624.

[30] Linton, O., & Nielsen, J.P. (1995). A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration. *Biometrika*, 82, 93–100.

[31] Masry, E. (1996). Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis*, 17, 571–599.

[32] Müller, H.G., & Stadtmüller, U. (1999). Multivariate Boundary Kernels and a Continuous Least Squares Principle. *Journal of the Royal Statistical Society B*, 61, 439–458.

[33] Newey, W.K., & Powell, J.L. (1987). Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55, 819–847.

[34] Oka, T. (2009). Nonparametric Quantile Regression with Discontinuities. Working Paper.

[35] Porter, J. (2003). Estimation in the Regression Discontinuity Model. Manuscript, Harvard University.

[36] Newey, W., & Powell, J.L. (1987). Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55, 819–847.

[37] Powell, J.L., Stock, J.H., & Stoker, T.M. (1989). Semiparametric Estimation of Index Coefficients. *Econometrica*, 57, 1403–1430.

[38] Qu, Z., & Yoon, J. (2014). Nonparametric Estimation and Inference on Conditional Quantile Processes. Working Paper.

[39] Ruppert, D., & Wand, M.P. (1994). Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics*, 22, 1346–1370.

[40] Su, L., & Ullah, A. (2008). Local Polynomial Estimation of Nonparametric Simultaneous Equations Models. *Journal of Econometrics*, 144, 193–218.

[41] Su, L., & Xiao, Z. (2009). Testing Structural Change in Conditional Distribution. Unpublished Manuscript.

[42] Su, L., & White, H. (2012). Conditional Independence Specification Testing for Dependent Processes with Local Polynomial Quantile Regression. *Advances in Econometrics*, 29, 355–434.

[43] White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50, 1–25.