2 Nonparametric regression

2.1 Introduction

The last chapter considered nonparametric estimation of densities and distributions. We now turn to the estimation of conditional expectations, commonly referred to as regressions. We start with a suitable definition of conditional expectation. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $D \in \mathbb{N}$.

Recall that if $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D) := \mathcal{B}^D)$ is a random vector, then $X^{-1}(\mathcal{B}^D) \subset \mathcal{F}$ is a σ-algebra and we write $X^{-1}(\mathcal{B}^D) := \sigma(X)$, the σ-algebra generated by X. Now, consider a random variable $Y : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$. It is legitimate to ask when Y is a random variable with respect to $\sigma(X)$.[1] The following theorem provides a useful characterization.

Theorem 1. *Let $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}^D, \mathcal{B}^D)$ be a random vector and $Y : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ be a random variable. Y is $\sigma(X) - \mathcal{B}$ measurable if, and only if, there exists $m : (\mathbb{R}^D, \mathcal{B}^D) \to (\mathbb{R}, \mathcal{B})$ such that $Y = m(X)$, and m is $\mathcal{B}^D - \mathcal{B}$ measurable.*

Proof. ($\Longleftarrow$) We want to show that for every $B \in \mathcal{B}$ we have $Y^{-1}(B) \in \sigma(X)$. But $Y^{-1}(B) = X^{-1}(m^{-1}(B))$ and by measurability of m, $m^{-1}(B) \in \mathcal{B}^D$ and since X is a random vector $X^{-1}(m^{-1}(B)) \in \sigma(X)$. Thus, Y is $\sigma(X) - \mathcal{B}$-measurable.
($\Longrightarrow$) Suppose $Y^{-1}(B) \in \sigma(X)$ for all $B \in \mathcal{B}$. First, assume that Y is a simple function. Then, for $k \in \mathbb{N}$ we have $Y(\omega) = \sum_{i=1}^{k} a_i I_{A_i}(\omega)$ for a_i distinct and $\{A_i\}_{i=1}^{k}$ pairwise-disjoint. In this case, $Y^{-1}(\{a_i\}) = A_i$ and by assumption $A_i \in \sigma(X)$. Hence, there exists $B_i \in \mathcal{B}^D$ such that $X^{-1}(B_i) = A_i$ by definition of $\sigma(X)$. Hence,

$$Y(\omega) = \sum_{i=1}^{k} a_i I_{X^{-1}(B_i)}(\omega) = \sum_{i=1}^{k} a_i I_{X \in B_i}(\omega).$$

Define $m(x) = \sum_{i=1}^{k} a_i I_{B_i}(x)$, then $Y = m(X)$, m is $\mathcal{B}^D$-measurable. Thus, the implication is proved for every Y simple that is $\sigma(X)$-measurable.

If $Y : (\Omega, \mathcal{F}, P) \to [0, \infty)$ then, by Lemma 8, there exist $Y_n(\omega)$ simple such that

$$Y(\omega) = \lim_{n \to \infty} Y_n(\omega), \; 0 \leq Y_n(\omega) \leq Y_{n+1}(\omega).$$

If each Y_n is $\sigma(X) - \mathcal{B}$ measurable and $Y_n := m_n(X)$ as in the proof when Y is a simple function, set $m(x) = \limsup_{n \to \infty} m_n(x)$ and note $Y = \lim_{n \to \infty} Y_n = \lim_{n \to \infty} m_n(X)$.

Given that $m(X) := (\limsup_{n \to \infty} m_n)(X) = \limsup_{n \to \infty} m_n(X)$, $m(x)$ is $\mathcal{B}^D - \mathcal{B}$-measurable. For general Y, write $Y = Y^+ - Y^-$ which reduces to the preceding case. □

[1] More generally, if $\mathcal{G}$ is a σ-algebra and $\mathcal{G} \subset \mathcal{F}$, we say that X is $\mathcal{G} - \mathcal{B}$ measurable if for all $B \in \mathcal{B}$, $X^{-1}(B) \in \mathcal{G}$. If G denotes the collection of all such σ-algebras, then it can be easily shown that $\sigma(X) = \bigcap_{\mathcal{G} \in G} \mathcal{G}$.

Let $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ be the space of random variables $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ such that $\int X^2(\omega) dP < \infty$. Let X be a random vector taking values in $\mathbb{R}^D$ defined in the same probability space with $\sigma(X) \subset \mathcal{F}$. Then, $\mathcal{L}^2(\Omega, \sigma(X), P) \subset \mathcal{L}^2(\Omega, \mathcal{F}, P)$ is a (Hilbert) space with the same inner product ($\int X(\omega)X'(\omega)dP$) and $\mathcal{L}^2(\Omega, \sigma(X), P)$ is a closed subspace of $\mathcal{L}^2(\Omega, \mathcal{F}, P)$. By the Projection Theorem, we next define conditional expectation.

Definition 1. *Let $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$. The conditional expectation of Y given X is the unique element $\hat{Y} \in \mathcal{L}^2(\Omega, \sigma(X), P)$ such that*

$$E((Y - \hat{Y})s) = 0, \text{ for all } s \in \mathcal{L}^2(\Omega, \sigma(X), P).$$

We write $\hat{Y} = E(Y|X)$ or $\hat{Y} = E(Y|\sigma(X))$.

Remark 1. *1. Again, by the Projection Theorem,*

$$\|Y - E(Y|X)\|_2 := \left(\int_\Omega |Y - E(Y|X)|^2 dP \right)^{1/2} \leq \|Y - s\|_2 \text{ for all } s \in \mathcal{L}^2(\Omega, \sigma(X), P).$$

Hence,

$$E(Y|X) := arg \inf_{s \in \mathcal{L}^2(\Omega, \sigma(X), P)} \|Y - s\| := arg \inf_{m \in \mathcal{M}} \|Y - m(X)\|.$$

where $\mathcal{M}$ is the set of Borel measurable functions from $\mathbb{R}^D$ to $\mathbb{R}$.

2. Since $E(Y|X)$ is $\sigma(X)$-measurable, by Theorem 1, there exists $m : \mathbb{R}^D \to \mathbb{R}$ which is Borel measurable such that $E(Y|X) = m(X)$ and m is unique. Hence, we can write $E[(Y - m(X))g(X)] = 0$, for all $g : \mathbb{R}^D \to \mathbb{R}$ Borel measurable such that $\int g^2 dP < \infty$.

3. The requirement that Y be square integrable is not necessary for the definition of conditional expectation. It is convenient here because the Projection Theorem can be used to obtain existence and uniqueness of $E(Y|X)$. However, $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$ is sufficient to obtain existence and uniqueness of $E(Y|X)$.

4. We may write

$$Y = E(Y|X) + U = m(X) + U \tag{1}$$

where $U = Y - E(Y|X) = Y - m(X)$ and by definition of conditional expectation $E(Ug(X)) = 0$ for any g that is $\mathcal{B}^D - \mathcal{B}$ measurable. Note, in particular, that $E(UX) = 0$ and by definition $E(U|X) = 0$.

Let $Z = \begin{pmatrix} Y & X^T \end{pmatrix}^T$ be a $(1 + D)$-dimensional random vector defined on $(\Omega, \mathcal{F}, P)$ having distribution F. Assume that we observe a random sample $\{Z_i = (Y_i \;\; X_i^T)^T\}_{i=1}^n$

where $X_i^T = \begin{pmatrix} X_{i1} & \cdots & X_{iD} \end{pmatrix}$ *with* $D \in \mathbb{N}$. *We call* Y *the regressand* and the components of X the *regressors*. Our objective is to estimate the regression $m(x)$ for some $x \in S \subset \mathbb{R}^D$. We write $U_i = Y_i - m(X_i)$ and call it the regression error. As $\{U_i\}_{i=1}^n$ is a sequence of measurable functions of $\{Z_i\}_{i=1}^n$, an IID sequence of random vectors, it is itself an IID sequence of (unobserved) random variables.

2.2 The Nadaraya-Watson estimator

Assuming that Z has a density denoted by $f_Z(z) := f(y,x)$ and the existence of $f_X(x) = \int f(y,x)dy$ with $f_X(x) \neq 0$, the conditional density of Y given $X = x$ exists and is denoted by $f_{Y|X=x}(y) = \frac{f(y,x)}{f_X(x)}$. Thus, we write

$$m(x) := E(Y|X=x) = \int y f_{Y|X=x}(y)dy = \frac{1}{f_X(x)} \int y f(y,x)dy. \tag{2}$$

Hence, it is natural to propose the following estimator for $m(x)$.

Definition 2. *The Nadaraya-Watson estimator for $m(x)$ is defined as*

$$\hat{m}(x) = \frac{1}{\hat{f}_X(x)} \int y \hat{f}(y,x)dy, \tag{3}$$

where $\hat{f}$ and $\hat{f}_X$ are Rosenblatt kernel density estimators. That is,

$$\hat{f}(y,x) = \frac{1}{n\, det(G_n)} \sum_{i=1}^n \mathcal{K}\left(G_n^{-1}(Z_i - (y,x')')\right) \text{ and } \hat{f}_X(x) = \frac{1}{n\, det(H_n)} \sum_{t=1}^n K\left(H_n^{-1}(X_i - x)\right)$$

where G_n and H_n are bandwidth matrices and $\mathcal{K}$ and K are $D+1$ and D dimensional kernels.

Remark 2. *1. Let $h_{y,n}$ be the bandwidth associated with the first component of Z_i and $h_{1,n}, \cdots, h_{D,n}$ be the bandwidths associated with the regressors. We note that*

$$\begin{aligned}
\int y\hat{f}(y,x)dy &= \frac{1}{n\ det(G_n)} \int y \sum_{j=1}^n \mathcal{K}\left(G_n^{-1}\begin{pmatrix} Y_j - y \\ X_j - x \end{pmatrix}\right) dy \\
&= \frac{1}{nh_{y,n}h_{1,n}\cdots h_{D,n}} \sum_{j=1}^n \int y\, \mathcal{K}\left(G_n^{-1}\begin{pmatrix} Y_j - y \\ X_j - x \end{pmatrix}\right) dy \\
&= \frac{1}{n} \sum_{j=1}^n \int (Y_j - uh_{y,n})(h_{y,n}\ldots h_{D,n})^{-1} \mathcal{K}\left(u, H_n^{-1}(X_j - x)\right) h_{y,n} du \\
&= (nh_{1,n}\ldots h_{D,n})^{-1} \sum_{j=1}^n \int (Y_j - uh_{y,n}) \mathcal{K}\left(u, \frac{X_{j1} - x_1}{h_{1,n}}, \ldots, \frac{X_{jD} - x_D}{h_{D,n}}\right) du \\
&= (nh_{1,n}\ldots h_{D,n})^{-1} \sum_{j=1}^n \left(Y_j \int \mathcal{K}\left(u, \frac{X_{j1} - x_1}{h_{1,n}}, \ldots, \frac{X_{jD} - x_D}{h_{D,n}}\right) du \right. \\
&- \left. h_{y,n} \int u\mathcal{K}\left(u, \frac{X_{j1} - x_1}{h_{1,n}}, \ldots, \frac{X_{jD} - x_D}{h_{D,n}}\right) du\right).
\end{aligned}$$

If $\int \mathcal{K}(u, x_1, \cdots, x_D)du = K(x_1, \cdots, x_D)$ *and if* $\int u\mathcal{K}(u, x_1, \cdots, x_D)du = 0$*, then*

$$\int y\hat{f}(y, x)dy = (nh_{1,n} \dots h_{D,n})^{-1} \sum_{j=1}^{n} Y_j K\left(H_n^{-1}(X_j - x)\right)$$

and the Nadaraya-Watson estimator can be written as

$$\hat{m}(x) = \frac{\sum_{j=1}^{n} K\left(H_n^{-1}(X_j - x)\right) Y_j}{\sum_{j=1}^{n} K\left(H_n^{-1}(X_j - x)\right)} = \sum_{j=1}^{n} w_{jn}(x) Y_j,$$

where $w_{jn}(x; X_1, \cdots, X_n) = \frac{K\left(H_n^{-1}(X_j - x)\right)}{\sum_{i=1}^{n} K\left(H_n^{-1}(X_i - x)\right)}$*. Thus,* $\hat{m}(x)$ *can be written as a linear combination of the observations* $Y_1, \cdots, Y_n$ *on the regressand* Y *with coefficients* $w_{jn}(x; X_1, \cdots, X_n)$*. In addition,* $\sum_{j=1}^{n} w_{jn}(x; X_1, \cdots, X_n) = 1$*. Hence, the Nadaraya-Watson estimator is a weighted* average of the observations on the regressand.

2. In this context it is convenient to work with product kernels, where $\mathcal{K}(x_1, \dots, x_D) = \prod_{j=1}^{D} k(x_j)$ and k is a univariate kernel centered at zero.

The Nadaraya-Watson estimator can also be obtained from a different perspective that will motivate an entire new class of estimators. Hence, consider the following estimator.

Definition 3. *The local constant estimator for* $m(x)$ *is denoted by* $\hat{m}_0(x)$ *and given by,*

$$\begin{aligned} \hat{m}_0(x) &:= \underset{c}{arg\,min} \sum_{j=1}^{n} (Y_j - c)^2 K\left(H_n^{-1}(X_j - x)\right) := \underset{c}{arg\,min}\, (\mathbf{Y} - 1_n c)^T P(x)(\mathbf{Y} - 1_n c) \\ &:= \underset{c}{arg\,min}\, S_n(c) \end{aligned}$$

where $P(x) = diag\{K(H_n^{-1}(X_j - x))\}_{j=1}^{n}$*,* $1_n = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}^T$*,* $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ *and* c *is a scalar.*

Remark 3. *Taking the first derivative of* $S_n(c)$ *with respect to* c *and finding* $\hat{c} := \hat{m}_0(x)$ *that sets it equal to zero gives* $-21_n' P(x)\mathbf{Y} + 2\hat{m}_0(x)1_n' P(x)1_n = 0$ *and*

$$\hat{m}_0(x) = \frac{\sum_{j=1}^{n} K\left(H_n^{-1}(X_j - x)\right) Y_j}{\sum_{j=1}^{n} K\left(H_n^{-1}(X_j - x)\right)}.$$

The second derivative of $S_n(c)$ *with respect to* c *is* $1_n' P(x)1_n = \sum_{i=1}^{n} K(H_n^{-1}(X_i - x))$*. For a minimum, we need* $\sum_{i=1}^{n} K(H_n^{-1}(X_i - x)) > 0$*, which is true if* $K(H_n^{-1}(X_i - x)) > 0$ *for all* i*.*

2.3 Local Polynomial Estimation

The optimization in the definition of the local constant estimator suggests a class of estimators that uses more flexible functions for the local approximation of $m(x)$. Taylor's Theorem suggests the following estimator, known in the literature as the local polynomial estimator.

Definition 4. *Let $\alpha \in \mathbb{R}^D$ with $\alpha_i \in \{0, 1, \cdots\}$, $|\alpha| = \sum_{i=1}^{D} \alpha_i$ and for $x \in \mathbb{R}^D$, $x^\alpha = x_1^{\alpha_1} \cdots x_D^{\alpha_D}(x)$. In addition, let $\beta^\alpha := \partial^\alpha m(x) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_D^{\alpha_D}} m(x)$ and define the vector $\beta_{p,D} \in \mathbb{R}^c$ where $c = \begin{pmatrix} D+p \\ p \end{pmatrix}$ such that*

$$\beta_{p,D} = \begin{pmatrix} \beta^{(0,\cdots,0)} & \beta^{(0,1,\cdots,0)} & \cdots & \beta^{(0,0,\cdots,1)} & \cdots \end{pmatrix}^T.$$

The local polynomial estimator of order p for $m(x)$ is given by $\hat{m}_p(x) := \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \hat{\beta}_{p,D}$, where

$$\hat{\beta}_{p,D} := \underset{\beta}{arg\min} \sum_{i=1}^{n} \left(Y_i - \sum_{|\alpha| \le p} \frac{1}{\alpha!} (X_i - x)^\alpha \beta^\alpha \right)^2 K\left(H_n^{-1}(X_i - x)\right) := \underset{\beta}{arg\min}\, S_n(\beta). \tag{4}$$

Remark 4. *1. A much studied version of this estimator takes $p = 1$. $\hat{m}_1(x)$ is called the local linear estimator and in this case*

$$\beta_{1,D} = \begin{pmatrix} \beta^{(0,\cdots,0)} & \beta^{(0,1,\cdots,0)} & \cdots & \beta^{(0,0,\cdots,1)} \end{pmatrix}^T := \begin{pmatrix} \beta^{(0,\cdots,0)} & \beta_1^T \end{pmatrix}^T,$$

where $\beta_1 \in \mathbb{R}^D$. Then, if $\mathbf{X} := \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}^T$ and $\otimes$ denotes the Kronecker product

$$\begin{aligned} S_n(\beta) &:= \sum_{i=1}^{n} \left(Y_i - \sum_{|\alpha| \le 1} \frac{1}{\alpha!} (X_i - x)^\alpha \beta^\alpha \right)^2 K\left(H_n^{-1}(X_i - x)\right) \\ &= \sum_{i=1}^{n} \left(Y_i - \beta^{(0,0,\cdots,0)} - (X_i - x)^T \beta_1\right)^2 K\left(H_n^{-1}(X_i - x)\right) \\ &= \left(\mathbf{Y} - \begin{pmatrix} 1_n & \mathbf{X} - (1_n \otimes x^T) \end{pmatrix} \begin{pmatrix} \beta^{(0,0,\cdots,0)} \\ \beta_1 \end{pmatrix}\right)^T P(x) \\ &\times \left(\mathbf{Y} - \begin{pmatrix} 1_n & \mathbf{X} - (1_n \otimes x^T) \end{pmatrix} \begin{pmatrix} \beta^{(0,0,\cdots,0)} \\ \beta_1 \end{pmatrix}\right) \end{aligned}$$

If $R(x) = \begin{pmatrix} 1_n & \mathbf{X} - (1_n \otimes x^T) \end{pmatrix}$ the vector $\hat{\beta}_{1,D} := \begin{pmatrix} \hat{\beta}^{(0,\cdots,0)} \\ \hat{\beta}_1 \end{pmatrix}$ that minimizes $S_n(\beta) = (\mathbf{Y} - R(x)\beta)^T P(x) (\mathbf{Y} - R(x)\beta)$ is given by

$$\hat{\beta}_{1,D} = \begin{pmatrix} \hat{\beta}^{(0,\cdots,0)} \\ \hat{\beta}_1 \end{pmatrix} = \left(R(x)^T P(x) R(x)\right)^{-1} R(x)^T P(x) \mathbf{Y} \tag{5}$$

whenever $\det\left(R(x)^T P(x) R(x)\right)) \neq 0$. *Hence, the local polynomial estimator of order* 1 *for* $m(x)$ *is therefore given by*

$$\hat{\beta}^{(0,\cdots,0)} := \hat{m}_1(x) = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \left(R(x)^T P(x) R(x)\right)^{-1} R(x)^T P(x) \mathbf{Y} := W(x)\mathbf{Y}.$$

where $W(x) := \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \left(R(x)^T P(x) R(x)\right)^{-1} R(x)^T P(x)$.

2. *Note that estimators for* $\frac{\partial^1}{\partial x_1^{\alpha_1} \cdots \partial x_D^{\alpha_D}} m(x) = \frac{\partial}{\partial x_d} m(x)$ *for* $d = 1, \cdots, D$ *can be obtained directly from equation (5) by selecting the appropriate element from the vector* $\begin{pmatrix} \hat{\beta}^{(0,\cdots,0)} \\ \hat{\beta}_1 \end{pmatrix}$. *Hence, we can define the estimator*

$$\widehat{\frac{\partial}{\partial_d} m(x)} := \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta}^{(0,\cdots,0)} \\ \hat{\beta}_1 \end{pmatrix},$$

where the number 1 appears on the $d+1$ *position of the vector* $\begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}$.

3. *Let* I_N *denote the identity matrix of dimension* N *and note that* $R(x) \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 1_n$.

Then, since $(R(x)^T P(x) R(x))^{-1} R(x)^T P(x) R(x) = I_{D+1}$ *and*

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} I_{D+1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 1$$

we have

$$\begin{aligned} 1 &= \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} (R(x)^T P(x) R(x))^{-1} R(x)^T P(x) R(x) \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} (R(x)^T P(x) R(x))^{-1} R(x)^T P(x) 1_n = W(x) 1_n \end{aligned}$$

Hence, $\hat{m}_1(x)$ *can be expressed as a weighted average of the observations on* Y. *The conclusion easily generalizes to any polynomial estimator of order* p.

3 Asymptotic properties of the Nadaraya-Watson estimator

3.1 The case where $D = 1$

We start with consistency.

Theorem 2. *Let* $\left\{ \begin{pmatrix} Y_i & X_i \end{pmatrix}^T \right\}_{i=1}^n$ *be a sequence of IID random vectors distributed as* $\begin{pmatrix} Y & X \end{pmatrix}'$ *and assume that a)* $|K(u)| \leq C$ *for all* $|u| \in \mathbb{R}$*; b)* $\int |K(u)|du < \infty$ *and* $\int K(u)du = 1$*; c)* $|u||K(u)| \to 0$ *as* $u \to \infty$*. In addition, assume that the marginal density of* X*,* $f_X(x) > 0$ *and* x *is a point of continuity of* f_X *and* m *with* $|m(x)|, f(x) < C$*. If* $V(Y|X) = \sigma^2$*,* $h_n \to 0$ *and* $nh_n \to \infty$ *as* $n \to \infty$*,*

$$\hat{m}(x) - m(x) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Proof. By definition

$$\begin{aligned}\hat{m}(x) &= \frac{\sum_{j=1}^n K\left(\frac{X_j - x}{h_n}\right) Y_j}{\sum_{j=1}^n K\left(\frac{X_j - x}{h_n}\right)} = \frac{1}{\frac{1}{nh_n}\sum_{j=1}^n K\left(\frac{X_j - x}{h_n}\right)} \frac{1}{nh_n} \sum_{j=1}^n K\left(\frac{X_j - x}{h_n}\right) Y_j \\ &= \frac{1}{\hat{f}_X(x)} \frac{1}{nh_n} \sum_{j=1}^n K\left(\frac{X_j - x}{h_n}\right) Y_j,\end{aligned}$$

where $\hat{f}_X(x) = \frac{1}{nh_n}\sum_{j=1}^n K\left(\frac{X_j - x}{h_n}\right)$. From chapter 1, under the conditions in this theorem $\hat{f}_X(x) \xrightarrow{p} f_X(x)$. Hence, it suffices to show that

$$\hat{g}(x) = \frac{1}{nh_n} \sum_{j=1}^n K\left(\frac{X_j - x}{h_n}\right) Y_j \xrightarrow{p} m(x) f_X(x).$$

Here, we show that $E(\hat{g}(x)) \to m(x)f_X(x)$ and $V(\hat{g}(x)) \to 0$ as $n \to \infty$. By the Law of Iterated Expectations and the IID assumption

$$\begin{aligned}E(\hat{g}(x)) &= \frac{1}{h_n} E\left(K\left(\frac{X - x}{h_n}\right) Y\right) = \frac{1}{h_n} E_X\left(K\left(\frac{X - x}{h_n}\right) E(Y|X)\right) \\ &= \frac{1}{h_n} E_X\left(K\left(\frac{X - x}{h_n}\right) m(X)\right) \\ &= \frac{1}{h_n} \int K\left(\frac{u - x}{h_n}\right) m(u) f_X(u) du = \int K(v) m(x + h_n v) f_X(x + h_n v) dv.\end{aligned}$$

Now, given that $K \in \mathcal{L}^1$ and provided $|m(x)|, f_X(x) < C$ and x is a point of continuity of f_X and m, we we conclude by dominated convergence that

$$E(\hat{g}(x)) \to f_X(x)m(x) \int K(u)du = f_X(x)m(x).$$

provided $\int K(u)du = 1$. Now, given that $V(Y|X) = \sigma^2$, by the Law of Iterated Expectations

$$V(\hat{g}(x)) = \sigma^2 \frac{1}{nh_n^2} \int K^2\left(\frac{u - x}{h_n}\right) f_X(u)du + \frac{1}{nh_n^2} V\left(K\left(\frac{X - x}{h_n}\right) m(X)\right).$$

Note that $V\left(K\left(\frac{X-x}{h_n}\right)m(X)\right) = E\left(K^2\left(\frac{X-x}{h_n}\right)m^2(X)\right) - \left(E\left(K\left(\frac{X-x}{h_n}\right)m(X)\right)\right)^2$. Furthermore,

$$E\left(K^2\left(\frac{X-x}{h_n}\right)m^2(X)\right) = h_n\int K^2(u)m^2(x-uh)f_X(x-uh)du$$

and

$$E\left(K\left(\frac{X-x}{h_n}\right)m(X)\right) = h_n\int K(u)m(x-uh)f_X(x-uh)du.$$

Then,

$$V(\hat{g}(x)) = \sigma^2\frac{1}{nh_n}\int K^2(u)\, f_X(x-uh)du + \frac{1}{nh_n}\int K^2(u)m^2(x-uh)f_X(x-uh)du + \frac{1}{n}\left(\int K(u)m(x-uh)f_X(x-uh)du\right)^2.$$

Given the bounds on $|m|$ and f_X we have by dominated convergence that as $n\to\infty$ the integrals in the previous equation display converge. Hence, if $nh_n\to\infty$ we have that $V(\hat{g}(x))\to 0$. □

We now give sufficient conditions for asymptotic normality of $\hat{m}(x)$ under suitable centering and normalization.

Theorem 3. *Let* $\left\{\begin{pmatrix} Y_i & X_i \end{pmatrix}^T\right\}_{i=1}^n$ *be a sequence of IID random vectors distributed as* $\begin{pmatrix} Y & X \end{pmatrix}'$ *and assume that a)* $|K(u)|\le C$ *for all* $u\in\mathbb{R}$*; b)* $\int|K(u)|du<\infty$ *and* $\int K(u)du=1$*; c)* $|u||K(u)|\to 0$ *as* $|u|\to\infty$*. In addition, assume that the marginal density of* X*,* $f_X(x)>0$ *and* x *is a point of continuity of* f_X *and* m *with* $|m(x)|, f(x)<C$*. If* $V(Y|X)=\sigma^2$*,* $E(|Y-m(X)|^{2+\delta}|X)<C<\infty$ *for some* $\delta>0$*,* $h_n\to 0$ *and* $nh_n\to\infty$ *as* $n\to\infty$*,*

$$(nh_n)^{1/2}(\hat{m}(x)-E(\hat{m}(x)|X)) \xrightarrow{d} N\left(0, \frac{\sigma^2}{f_X(x)}\int K^2(u)du\right)$$

Proof. Since $\hat{m}(x) = \frac{(nh_n)^{-1}\sum_{j=1}^n K(h_n^{-1}(X_j-x))Y_j}{(nh_n)^{-1}\sum_{j=1}^n K(h_n^{-1}(X_j-x))}$ and the denominator converges in probability to $f_X(x)>0$, we are mainly concerned with the convergence in distribution of the numerator, denoted by $\hat{g}(x)$. Let $\mathcal{X}=\{X_1,\cdots X_n\}$ and note that

$$\begin{aligned} E(\hat{g}(x)|\mathcal{X}) &= \frac{1}{nh_n}\sum_{j=1}^n K\left(\frac{X_j-x}{h_n}\right)E(Y_j|\mathcal{X}) \\ &= \frac{1}{nh_n}\sum_{j=1}^n K\left(\frac{X_j-x}{h_n}\right)m(X_j) \text{ by the fact that } \{(Y_i,X_i)\}_{i=1}^n \text{ is IID.} \end{aligned}$$

Then,

$$\hat{g}(x) - E(\hat{g}(x)|\mathcal{X}) = \frac{1}{nh_n}\sum_{j=1}^{n} K\left(\frac{X_j - x}{h_n}\right)(Y_j - m(X_j)).$$

Let $Z_{jn} = \frac{K\left(\frac{X_j - x}{h_n}\right)(Y_j - m(X_j))}{nh_n}$ and note that $E(Z_{jn}) = 0$ and

$$\begin{aligned} V(Z_{jn}) &= E(Z_{jn}^2) = \frac{\sigma^2}{n^2 h_n^2} E\left(K^2\left(\frac{X_j - x}{h_n}\right)\right) \\ &= \frac{\sigma^2}{n^2 h_n^2}\int K^2(h_n^{-1}(u-x)) f_X(u)du \end{aligned}$$

Now, let $s_n^2 = \sum_{j=1}^{n} E(Z_{jn}^2) = \frac{\sigma^2}{nh_n}\frac{1}{h_n}\int K^2(h_n^{-1}(u-x))f_X(u)du$ and

$$\begin{aligned} X_{jn} = \frac{Z_{jn}}{s_n} &= \frac{Z_{jn}}{(nh_n)^{-1/2}\left(\frac{\sigma^2}{h_n}\int K^2(h_n^{-1}(\alpha - x))f_X(\alpha)d\alpha\right)^{1/2}} \\ &= \frac{K(h_n^{-1}(X_j - x))(Y_j - m(X_j))}{(nh_n)^{1/2}\left(\frac{\sigma^2}{h_n}\int K^2(h_n^{-1}(\alpha - x))f_X(\alpha)d\alpha\right)^{1/2}} \end{aligned}$$

By Liapounov's CLT $\sum_{j=1}^{n} X_{jn} \xrightarrow{d} N(0,1)$ provided that $\lim_{n\to\infty}\sum_{j=1}^{n} E\left(|X_{jn}|^{2+\delta}\right) = 0$ for some $\delta > 0$. Note that $|X_{jn}| = \frac{|K(h_n^{-1}(X_j - x))||Y_j - m(X_j)|}{(nh_n)^{1/2}c(n)^{1/2}}$ with $c(n) = \frac{\sigma^2}{h_n}\int K^2(h_n^{-1}(\alpha - x))f_X(\alpha)d\alpha$ and $|X_{jn}|^{2+\delta} = \frac{|K(h_n^{-1}(X_j - x))|^{2+\delta}|Y_j - m(X_j)|^{2+\delta}}{(nh_n)^{1+\delta/2}c(n)^{1+\delta/2}}$. $c(n)$ is nonstochastic, therefore

$$E(|X_{jn}|^{2+\delta}) = (nh_n c(n))^{-1-\delta/2} E\left(|K(h_n^{-1}(X_j - x))|^{2+\delta}|Y_j - m(X_j)|^{2+\delta}\right)$$

and

$$\sum_{j=1}^{n} E(|X_{jn}|^{2+\delta}) = (nh_n c(n))^{-1-\delta/2}\sum_{j=1}^{n} E\left(|K(h_n^{-1}(X_j - x))|^{2+\delta}|Y_j - m(X_j)|^{2+\delta}\right).$$

Now, if $E(|Y_j - m(X_j)|^{2+\delta}|X_j) < C$

$$\begin{aligned} E\left(|K(h_n^{-1}(X_j - x))|^{2+\delta}|Y_j - m(X_j)|^{2+\delta}\right) &= E\left(|K(h_n^{-1}(X_j - x))|^{2+\delta}E(|Y_j - m(X_j)|^{2+\delta}|X_j)\right) \\ &\leq C\int\left|K\left(\frac{\alpha - x}{h_n}\right)\right|^{2+\delta} f_X(\alpha)d\alpha. \end{aligned}$$

Consequently,

$$\begin{aligned} \sum_{j=1}^{n} E(|X_{jn}|^{2+\delta}) &\leq (nh_n c(n))^{-1-\delta/2} nC\int\left|K\left(\frac{\alpha - x}{h_n}\right)\right|^{2+\delta} f_X(\alpha)d\alpha \\ &= (nh_n)^{-\delta/2}(c(n))^{-1-\delta/2}C\int|K(u)|^{2+\delta}f_X(x + h_n u)du \\ &\leq (nh_n)^{-\delta/2}(c(n))^{-1-\delta/2}C\,C^{1+\delta}\int|K(u)|f_X(x + h_n u)du. \end{aligned}$$

By dominated convergence, as $n \to \infty$, $c(n) \to \sigma^2 f_X(x) \int K^2(u)du$ and $\int |K(u)| f_X(x + h_n u)du \to f_X(x) \int |K(u)|du < C$. Since $nh_n \to \infty$ we have that $\lim_{n\to\infty} \sum_{j=1}^n E(|X_{jn}|^{2+\delta}) = 0$. Hence,

$$\frac{\sum_{j=1}^n K(h_n^{-1}(X_j - x))(Y_j - m(X_j))}{(nh_n)^{1/2}\left(\sigma^2 h_n^{-1} \int K^2(h_n^{-1}(\alpha - x))f(\alpha)d\alpha\right)^{1/2}} = \frac{(nh_n)^{1/2}(\hat{g}(x) - E(\hat{g}(x)|\mathcal{X}))}{c(n)^{1/2}} \xrightarrow{d} N(0,1),$$

which implies that $(nh_n)^{1/2}(\hat{g}(x) - E(\hat{g}(x)|\mathcal{X})) \xrightarrow{d} N\left(0, \sigma^2 f_X(x) \int K^2(u)du\right)$.

Now, since

$$\hat{m}(x) - E(\hat{m}(x)|\mathcal{X}) = \frac{\hat{g}(x) - E(\hat{g}(x)|\mathcal{X})}{\hat{f}_X(x)},$$

we have that

$$(nh_n)^{1/2}(\hat{m}(x) - E(\hat{m}(x)|\mathcal{X})) = \frac{(nh_n)^{1/2}(\hat{g}(x) - E(\hat{g}(x)|\mathcal{X}))}{\hat{f}_X(x)} \xrightarrow{d} \frac{1}{f_X(x)} Z_n,$$

where $Z_n \sim N\left(0, \sigma^2 f_X(x) \int K^2(u)du\right)$. Consequently,

$$(nh_n)^{1/2}(\hat{m}(x) - E(\hat{m}(x)|\mathcal{X})) \xrightarrow{d} N\left(0, \frac{\sigma^2}{f_X(x)} \int K^2(u)du\right).$$

□

Remark 5. *We write*

$$\sqrt{nh_n}(\hat{m}(x) - m(x)) = \sqrt{nh_n}(\hat{m}(x) - E(\hat{m}(x)|\mathcal{X})) + \sqrt{nh_n}(E(\hat{m}(x)|\mathcal{X}) - m(x))$$

and observe that

$$\begin{aligned} E(\hat{m}(x)|\mathcal{X}) - m(x) &= \frac{1}{\sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right)} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right)(m(X_i) - m(x)) \\ &= \frac{1}{\hat{f}_X(x)} \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right)(m(X_i) - m(x)). \end{aligned}$$

Hence, to establish the asymptotic distribution of $\sqrt{nh_n}(\hat{m}(x) - m(x))$ it is necessary to investigate the order (in probability) of the conditional bias of the Nadaraya-Watson estimator, which will be done in subsection 3.3. First we state our results for consistency and asymptotic normality for the case where $D > 1$.

3.2 The case where $D > 1$

Theorems 2 and 7 extend easily to the case where $D > 1$. Hence, we state the next two theorems without proof.

Theorem 4. *Let* $\left\{ \begin{pmatrix} Y_i & X_i^T \end{pmatrix}^T \right\}_{i=1}^n$ *be a sequence of IID random vectors distributed as* $\begin{pmatrix} Y & X \end{pmatrix}'$ *and assume that a)* $|K(u)| \leq C$ *for all* $u \in \mathbb{R}^D$*; b)* $\int |K(u)|du < \infty$ *and* $\int K(u)du = 1$*; c)* $||u||_1|K(u)| \to 0$ *as* $||u||_1 \to \infty$*. In addition, assume that the marginal density of* X*,* $f_X(x) > 0$ *and* x *is a point of continuity of* f_X *and* m *with* $|m(x)|, f(x) < C$*. If* $V(Y|X) = \sigma^2$*,* $h_{d,n} > 0$ *for all* n *and* $h_{d,n} \to 0$ *as* $n \to \infty$ *and* $n\,det(H_n) \to \infty$ *as* $n \to \infty$*,*

$$\hat{m}(x) - m(x) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Theorem 5. *Let* $\left\{ \begin{pmatrix} Y_i & X_i^T \end{pmatrix}^T \right\}_{i=1}^n$ *be a sequence of IID random vectors distributed as* $\begin{pmatrix} Y & X \end{pmatrix}'$ *and assume that a)* $|K(u)| \leq C$ *for all* $u \in \mathbb{R}$*; b)* $\int |K(u)|du < \infty$ *and* $\int K(u)du = 1$*; c)* $||u||_1|K(u)| \to 0$ *as* $||u||_1 \to \infty$*. In addition, assume that the marginal density of* X*,* $f_X(x) > 0$ *and* x *is a point of continuity of* f_X *and* m *with* $|m(x)|, f(x) < C$*. If* $V(Y|X) = \sigma^2$*,* $E(|Y - m(X)|^{2+\delta}|X) < C < \infty$ *for some* $\delta > 0$*,* $h_{d,n} \to 0$ *and* $n\,det(H_n) \to \infty$ *as* $n \to \infty$*,*

$$(n\,det(H_n))^{1/2}(\hat{m}(x) - E(\hat{m}(x)|X)) \xrightarrow{d} N\left(0, \frac{\sigma^2}{f_X(x)} \int K^2(u)du\right).$$

Remark 6. *As in the case of density estimation, the multivariate Nadaraya-Watson estimator converges in distribution at a rate that is slower than the parametric rate. In addition, the speed of convergence decays exponentially with* D*. Hence, the curse of dimensionality manifests itself again in the case of the Nadaraya-Watson estimator.*

3.3 Bias of the Nadaraya Watson Estimator

Recall that a sequence of random variables $\{Z_n\}_{n\in\mathbb{N}}$ is said to be bounded in probability if for all $\epsilon > 0$ there exists $C_\epsilon > 0$ such that for all $n \geq N_\epsilon$

$$P(|Z_n| > C_\epsilon) \leq \epsilon$$

and we write $Z_n = O_p(1)$. If $\{a_n\}_{n\in\mathbb{N}}$ is a nonstochastic sequence sequence of positive real numbers, we say that $Z_n = O_p(a_n)$ if $\frac{Z_n}{a_n} = O_p(1)$. It is a direct consequence of Markov's Inequality that: a) if $E(|Z_n|) = O(a_n)$ then $Z_n = O_p(a_n)$ and b) $E(Z_n^2) = O(a_n)$ then $Z_n = O_p(a_n^{1/2})$.

Theorem 6. *Let* $\left\{ \begin{pmatrix} Y_i & X_i \end{pmatrix}^T \right\}_{i=1}^n$ *be a sequence of IID random vectors distributed as* $\begin{pmatrix} Y & X \end{pmatrix}^T$ *and assume that a)* $|K(u)| \leq C$ *for all* $u \in \mathbb{R}$*; b)* $\int |K(u)|du < \infty$ *and* $\int K(u)du = 1$*; c)* $|u||K(u)| \to 0$ *as* $|u| \to \infty$*; d)* $\int uK(u)du = 0$*,* $\int u^2K(u)du := \mu_2$*. In addition, assume that the marginal density of* X*,* $f_X(x) > 0$*,* x *is a point of continuity of*

f_X and m with $|m(x)|, f(x) < C$ and $f, m \in \mathcal{C}_b^2$. If $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, then for $I_n(x) = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right)(m(X_i) - m(x))$ we have

$$a)\ E(I_n(x)) = \frac{h_n^2}{2}\mu_2 f_X(x)\left(m^{(2)}(x) + 2\,\frac{m^{(1)}(x)f_X^{(1)}(x)}{f_X(x)}\right) + o(h_n^2) \text{ and } V(I_n(x)) = O(h_n/n),$$

$$b)\ I_n - \frac{h_n^2}{2}\mu_2 f_X(x)\left(m^{(2)}(x) + 2\,\frac{m^{(1)}(x)f_X^{(1)}(x)}{f_X(x)}\right) = O_p\left(h_n^2 + \frac{h_n^{1/2}}{n^{1/2}}\right).$$

Proof. a) First, note that

$$\begin{aligned}
E(I_n(x)) &= \int K(u)(m(x+hu) - m(x))f_X(x+hu)du \\
&= \int K(u)(m(x+hu)f_X(x+hu) - m(x)f_X(x) + m(x)f_X(x) - m(x)f_X(x+hu))du \\
&= \int K(u)(m(x+hu)f_X(x+hu) - m(x)f_X(x))du \\
&\quad - m(x)\int K(u)(f_X(x+hu) - f_X(x))du \\
&= \int K(u)\left([mf_X]^{(1)}(x)h_n u + \frac{1}{2}[mf_X]^{(2)}(x+\lambda h_n u)h_n^2 u^2\right)du \\
&\quad - m(x)\int K(u)\left(f_X^{(1)}(x)h_n u + \frac{1}{2}f_X^{(2)}(x+\lambda h_n u)h_n^2 u^2\right)du \\
&\quad \text{for } \lambda \in (0,1), \text{ since } m, f \in \mathcal{C}_b^2 \\
&= \frac{h_n^2}{2}\left(\int K(u)[mf_X]^{(2)}(x+\lambda h_n u)u^2 du - m(x)\int K(u)f_X^{(2)}(x+\lambda h_n u)u^2 du\right) \\
&\quad \text{since } \textstyle\int uK(u)du = 0.
\end{aligned}$$

Also, since $m, f \in \mathcal{C}_b^2$ and $\int u^2|K(u)|du = \mu_2$, by Lebesgue's Dominated Convergence Theorem we have

$$\begin{aligned}
\frac{1}{h_n^2}E(I_n(x)) &= \frac{1}{2}\left(\mu_2[mf_X]^{(2)}(x) + o(1) - m(x)(\mu_2 f_X^{(2)}(x) + o(1))\right) \\
&= \frac{1}{2}\mu_2 f_X(x)\left(m^{(2)}(x) + 2\frac{m^{(1)}(x)f_X^{(1)}(x)}{f_X(x)}\right) + o(1).
\end{aligned}$$

Second, given that $V(I_n) = E(I_n^2) - E(I_n)^2$ we investigate $E(I_n^2)$.

$$\begin{aligned}
E(I_n^2) = \frac{1}{n^2h_n^2}\Bigg(&\sum_{i=1}^n E\left(K^2\left(\frac{X_i - x}{h_n}\right)(m(X_i) - m(x))^2\right) \\
&+ \sum_{i=1}^n\sum_{j=1, j\neq i}^n E\left(K\left(\frac{X_i - x}{h_n}\right)(m(X_i) - m(x))K\left(\frac{X_j - x}{h_n}\right)(m(X_j) - m(x))\right)\Bigg)
\end{aligned}$$

where

$$E\left(K^2\left(\frac{X_i-x}{h_n}\right)(m(X_i)-m(x))^2\right)=h_n\int K^2(u)(m(x+hu)-m(x))^2 f_X(x+hu)du.$$

By the fact that the sequence $\{X_i\}_{i\in\mathbb{N}}$ is IID

$$\begin{aligned}&E\left(K\left(\frac{X_i-x}{h_n}\right)(m(X_i)-m(x))K\left(\frac{X_j-x}{h_n}\right)(m(X_j)-m(x))\right)\\&=h_n^2\left(\int K(u)(m(x+hu)-m(x))f_X(x+hu)du\right)^2.\end{aligned}$$

Consequently,

$$\begin{aligned}E(I_n^2)=&\frac{1}{nh_n}\int K^2(u)(m(x+hu)-m(x))^2 f_X(x+hu)du\\&+\left(1-\frac{1}{n}\right)\left(\int K(u)(m(x+hu)-m(x))f_X(x+hu)du\right)^2,\end{aligned}$$

and

$$\begin{aligned}V(I_n)=&\frac{1}{nh_n}\int K^2(u)(m(x+hu)-m(x))^2 f_X(x+hu)du\\&-\frac{1}{n}\left(\int K(u)(m(x+hu)-m(x))f_X(x+hu)du\right)^2=I_{n1}-I_{n2}\end{aligned}$$

where $I_{n2}=\frac{h_n^4}{n}\left(\frac{1}{2}\mu_2 f_X(x)\left(m^{(2)}(x)+2\frac{m^{(1)}(x)f_X^{(1)}(x)}{f_X(x)}\right)+o(1)\right)^2$. By the Mean Value Theorem, for $\lambda\in(0,1)$, $m(x+hu)-m(x)=m^{(1)}(x+\lambda h_n u)h_n u$. Hence,

$$I_{n1}=\frac{h_n^2}{nh_n}\int u^2K^2(u)(m^{(1)}(x+\lambda h_n u))^2 f_X(x+h_n u)du,$$

and given that $m, f\in\mathcal{C}_b^2$, by Lebesgue's Dominated Convergence Theorem we have

$$\int u^2K^2(u)(m^{(1)}(x+\lambda h_n u))^2 f_X(x+h_n u)du\to(m^{(1)}(x))^2 f_X(x)\int u^2K^2(u)du$$

and $\frac{nh_n}{h_n^2}V(I_n)=O(1)$.

b) Let $B(x)=\frac{1}{2}\mu_2 f_X(x)\left(m^{(2)}(x)+2\,\frac{m^{(1)}(x)f_X^{(1)}(x)}{f_X(x)}\right)$ and $\mathcal{I}_n=I_n-h_n^2B(x)$. Then, $\mathcal{I}_n^2=(I_n-E(I_n)+E(I_n)-h_n^2B(x))^2$ and

$$\begin{aligned}E(\mathcal{I}_n^2)&=V(I_n)+(E(I_n)-h_n^2B(x))^2\\&=O\left(\frac{h_n^2}{nh_n}\right)+o(h_n^4).\end{aligned}$$

Thus, by Markov's Inequality $\mathcal{I}_n=O_p\left(\left(\frac{h_n^2}{nh_n}+h_n^4\right)^{1/2}\right)=O_p\left(\frac{h_n^{1/2}}{n^{1/2}}+h_n^2\right)$. □

Remark 7. *Note that,*

$$\begin{aligned}
E(\hat{m}(x)|\mathcal{X}) &= m(x) + \frac{1}{\hat{f}_X(x)} I_n = m(x) + \frac{I_n}{\hat{f}_X(x)} - \frac{I_n}{f_X(x)} + \frac{I_n}{f_X(x)} \\
&= m(x) + \left(\frac{1}{\hat{f}_X(x)} - \frac{1}{f_X(x)} \right) I_n + \frac{I_n}{f_X(x)} \\
&= m(x) + (1 + o_p(1)) \frac{I_n}{f_X(x)} \\
&= m(x) + \frac{1}{f_X(x)} (1 + o_p(1))(I_n - h_n^2 B(x) + h_n^2 B(x)) \\
&= m(x) + \frac{h_n^2 B(x)}{f_X(x)} + o_p(h_n^2) + (1 + o_p(1)) \frac{\mathcal{I}_n}{f_X(x)}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\sqrt{nh_n} \left(\hat{m}(x) - E(\hat{m}(x)|\mathcal{X}) \right) = \sqrt{nh_n} \left(\hat{m}(x) - \left(m(x) + \frac{h_n^2 B(x)}{f_X(x)} + o_p(h_n^2) \right) \right) \\
- \sqrt{nh_n} \, (1 + o_p(1)) \frac{\mathcal{I}_n}{f_X(x)}.
\end{aligned}$$

Since $\mathcal{I}_n = O_p\left(\frac{h_n^{1/2}}{n^{1/2}} + h_n^2 \right)$ *and* $(1 + o_p(1))O_p(1) = O_p(1)$*, we have*

$$\begin{aligned}
\sqrt{nh_n} \left(\hat{m}(x) - E(\hat{m}(x)|\mathcal{X}) \right) = \sqrt{nh_n} \left(\hat{m}(x) - \left(m(x) + \frac{h_n^2 B(x)}{f_X(x)} + o_p(h_n^2) \right) \right) \\
- \left(h_n + (nh_n^5)^{1/2} \right) \frac{1}{f_X(x)} O_p(1).
\end{aligned}$$

If $nh_n^5 \to 0$ *then* $h_n + (nh_n^5)^{1/2} \to 0$ *and*

$$\begin{aligned}
\sqrt{nh_n} \left(\hat{m}(x) - E(\hat{m}(x)|\mathcal{X}) \right) = \sqrt{nh_n} \left(\hat{m}(x) - \left(m(x) + \frac{h_n^2 B(x)}{f_X(x)} + o_p(h_n^2) \right) \right) \\
+ o_p(1).
\end{aligned}$$

We immediately have the following Theorem.

Theorem 7. *Let* $\left\{ \begin{pmatrix} Y_i & X_i \end{pmatrix}^T \right\}_{i=1}^n$ *be a sequence of IID random vectors distributed as* $\begin{pmatrix} Y & X \end{pmatrix}^T$ *and assume that a)* $|K(u)| \leq C$ *for all* $u \in \mathbb{R}$*; b)* $\int |K(u)|du < \infty$ *and* $\int K(u)du = 1$*; c)* $|u||K(u)| \to 0$ *as* $|u| \to \infty$*; d)* $\int uK(u)du = 0$*,* $\int u^2 K(u)du = \mu_2$*. In addition, assume that the marginal density of* X*,* $f_X(x) > 0$ *and* x *is a point of continuity of* f_X *and* m *with* $|m(x)|, f(x) < C$*, and* $m, f_X \in \mathcal{C}_b^2$*. If* $V(Y|X) = \sigma^2$*,* $E(|Y - m(X)|^{2+\delta}|X) < C < \infty$ *for some* $\delta > 0$*,* $h_n \to 0$*,* $nh_n \to \infty$ *and* $nh_n^5 \to 0$ *as* $n \to \infty$*,*

$$\sqrt{nh_n} \left(\hat{m}(x) - \left(m(x) + \frac{h_n^2 B(x)}{f_X(x)} + o_p(h_n^2) \right) \right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{f_X(x)} \int K^2(\psi)d\psi \right).$$

and, equivalently,

$$\sqrt{nh_n}\left(\hat{m}(x) - m(x)\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{f_X(x)}\int K^2(\psi)d\psi\right).$$

4 Asymptotic properties of the Local Linear Estimator

4.1 Algebraic results

Before we study the asymptotic behavior of the local linear estimator we give some useful algebraic facts. We start by noting that since

$$R(x) = \begin{pmatrix} 1 & X_{11} - x_1 & \cdots & X_{1D} - x_D \\ 1 & X_{21} - x_1 & \cdots & X_{2D} - x_D \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} - x_1 & \cdots & X_{nD} - x_D \end{pmatrix} \text{ and } P(x) = \operatorname{diag}\left\{K\left(H_n^{-1}(X_t - x)\right)\right\}_{t=1}^n$$

matrix multiplication gives

$$\begin{aligned}\frac{1}{n\,\det(H_n)}R(x)^T P(x)R(x) := S(x) &= \begin{pmatrix} s_0(x) & s_1(x) & \cdots & s_D(x) \\ s_1(x) & s_{(1,1)}(x) & \cdots & s_{(1,D)}(x) \\ \vdots & \vdots & \vdots & \vdots \\ s_D(x) & s_{(D,1)}(x) & \cdots & s_{(D,D)}(x) \end{pmatrix} \\ &= \begin{pmatrix} s_0(x) & S_1(x)^T \\ S_1(x) & S_2(x) \end{pmatrix}\end{aligned}$$

where

$$s_0(x) = \frac{1}{n\,\det(H_n)}\sum_{t=1}^n K\left(H_n^{-1}(X_t - x)\right),\ s_j(x) = \frac{1}{n\,\det(H_n)}\sum_{t=1}^n K\left(H_n^{-1}(X_t - x)\right)(X_{tj} - x_j)$$

and

$$s_{(i,j)}(x) = \frac{1}{n\,\det(H_n)}\sum_{t=1}^n K\left(H_n^{-1}(X_t - x)\right)(X_{ti} - x_i)(X_{tj} - x_j) \text{ for } i,j = 1,\cdots,D.$$

Note that $s_{(i,j)}(x) = s_{(j,i)}(x)$. Furthermore,

$$\frac{1}{n\,\det(H_n)}R^T(x)P(x)\mathbf{Y} := \tau(x) = \begin{pmatrix} \tau_0(x) \\ \tau_1(x) \\ \vdots \\ \tau_D(x) \end{pmatrix}$$

where $\tau_0(x) = \frac{1}{n\,\det(H_n)}\sum_{t=1}^n K\left(H_n^{-1}(X_t - x)\right)Y_t$, $\tau_j(x) = \frac{1}{n\,\det(H_n)}\sum_{t=1}^n K\left(H_n^{-1}(X_t - x)\right)(X_{tj} - x_j)Y_t$. Hence, if $S(x)$ is invertible we may write the local linear estimator as

$$\hat{\beta}_{1,D}(x) = S(x)^{-1}\tau(x).$$

Lemma 1. *Assume that all partial derivatives of order 1 of $m(x)$ exist for $x \in \mathbb{R}^D$. Denote these partial derivatives by $\frac{\partial}{\partial_d} m(x)$ for $d = 1, \cdots, D$ and define*

$$m^{(1)}(x) := \left(\begin{array}{cccc} \frac{\partial}{\partial_1} m(x) & \frac{\partial}{\partial_2} m(x) & \cdots & \frac{\partial}{\partial_D} m(x) \end{array} \right)^T.$$

In addition, define $\mathbf{m} := \left(\begin{array}{cccc} m(X_1) & m(X_2) & \cdots & m(X_n) \end{array} \right)^T$, $\mathbf{U} := \left(\begin{array}{cccc} U_1 & U_2 & \cdots & U_n \end{array} \right)^T$ *where $U_t = Y_t - m(X_t)$. Also, put*

$$G_n = \mathrm{diag}\{1, \mathrm{diag}\{1/h_d^2\}_{d=1}^D\}$$

and

$$N_n = \mathrm{diag}\{\sqrt{n \ \det(H_n)}, \mathrm{diag}\{\sqrt{n \ \det(H_n)} h_d\}_{d=1}^D\},$$

both $D+1$-diagonal matrices. Then, letting $\beta_{1,D}(x) = \left(\begin{array}{cccc} m(x) & \cdots & m^{(1)}(x)^T \end{array} \right)^T$

$$\begin{aligned} & N_n \left(\hat{\beta}_{1,D}(x) - \beta_{1,D}(x) - (G_n S(x))^{-1} \frac{1}{n \ \det(H_n)} G_n R^T(x) P(x) \left(\mathbf{m} - R(x) \beta(x) \right) \right) \\ & = N_n (G_n S(x))^{-1} \frac{1}{n \ \det(H_n)} G_n R^T(x) P(x) \mathbf{U} \end{aligned}$$

Proof. We can immediately write

$$\begin{aligned} \hat{\beta}_{1,D}(x) - \beta_{1,D}(x) &= S(x)^{-1} \frac{1}{n \ \det(H_n)} R^T(x) P(x) \mathbf{Y} - S(x)^{-1} \frac{1}{n \ \det(H_n)} R^T(x) P(x) R(x) \beta_{1,D}(x) \\ &= (G_n S(x))^{-1} \frac{1}{n \ \det(H_n)} G_n R^T(x) P(x) \left(\mathbf{m} + \mathbf{U} - R(x) \beta_{1,D}(x) \right) \\ &= (G_n S(x))^{-1} \frac{1}{n \ \det(H_n)} G_n R^T(x) P(x) \left(\mathbf{m} - R(x) \beta_{1,D}(x) \right) \\ &+ (G_n S(x))^{-1} \frac{1}{n \ \det(H_n)} G_n R^T(x) P(x) \mathbf{U}. \end{aligned}$$

Consequently,

$$\begin{aligned} & N_n \left(\hat{\beta}_{1,D}(x) - \beta_{1,D}(x) - (G_n S(x))^{-1} \frac{1}{n \ \det(H_n)} G_n R^T(x) P(x) \left(\mathbf{m} - R(x) \beta_{1,D}(x) \right) \right) \\ & = N_n (G_n S(x))^{-1} \frac{1}{n \ \det(H_n)} G_n R^T(x) P(x) \mathbf{U}. \end{aligned}$$

□

4.2 The case where $D = 1$

If $D = 1$, we have from the previous subsection that

$$R(x) = \left(\begin{array}{cc} 1 & X_1 - x \\ 1 & X_2 - x \\ \vdots & \\ 1 & X_n - x \end{array} \right), \ P(x) = \mathrm{diag} \left\{ K \left(h_n^{-1} (X_t - x) \right) \right\}_{t=1}^n$$

and matrix multiplication gives

$$\frac{1}{nh_n}R(x)^T P(x)R(x) := S(x) = \begin{pmatrix} s_0(x) & s_1(x) \\ s_1(x) & s_{(1,1)}(x) \end{pmatrix} = \begin{pmatrix} s_0(x) & S_1(x)^T \\ S_1(x) & S_2(x) \end{pmatrix}$$

where

$$s_0(x) = \frac{1}{nh_n}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h_n}\right),\ s_1(x) = \frac{1}{nh_n}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h_n}\right)(X_t - x),\ \text{and}$$

$$s_{(1,1)}(x) := s_2(x) = \frac{1}{nh_n}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h_n}\right)(X_t - x)^2.$$

Furthermore, $G_n = \text{diag}\{1, 1/h_n^2\}$, $N_n = \text{diag}\{\sqrt{nh_n}, \sqrt{nh_n}h_n\}$ and

$$G_n S(x) = \begin{pmatrix} s_0(x) & s_1(x) \\ s_1(x)/h_n^2 & s_{(1,1)}(x)/h_n^2 \end{pmatrix}.$$

From chapter 1 we know that $s_0(x) = f_X(x) + o_p(1)$. The next lemma gives the order of the remaining terms.

Lemma 2. *Assume $f_X \in \mathcal{C}_b^2$, $K \geq 0$, $\int K(u)du = 1$, $\int uK(u)du = 0$, $\int u^2K(u)du = \mu_2$, $\int u^2K^2(u)du < \infty$ and $\int u^4K^2(u)du < \infty$. If $h_n \to 0$, $nh_n^3 \to \infty$ we have*

$$G_n S(x) = \begin{pmatrix} f_X(x) & 0 \\ \mu_2 f_X^{(1)}(x) & \mu_2 f_X(x) \end{pmatrix} + \begin{pmatrix} o_p(1) & O_p(h_n^2) \\ o_p(1) & o_p(1) \end{pmatrix} = Q + \begin{pmatrix} o_p(1) & O_p(h_n^2) \\ o_p(1) & o_p(1) \end{pmatrix}.$$

Proof. Note that $E(s_1(x)) = h_n \int uK(u)f_X(x + h_n u)du$ and

$$V(s_1(x)) = \frac{h_n}{n}\int K^2(u)u^2 f_X(x + h_n u)du - \frac{h_n^2}{n}\left(\int K(u)uf_X(x + h_n u)du\right)^2.$$

Hence, $V(s_1(x)) = O\left(\frac{h_n}{n}\right)$. Also, for $\lambda \in (0, 1)$

$$\begin{aligned}\frac{1}{h_n^2}E(s_1(x)) &= \frac{1}{h_n}\int uK(u)f_X(x + h_n u)du = \frac{1}{h_n}\int uK(u)[f_X(x) + h_n u f_X^{(1)}(x + \lambda h_n u)]du \\ &= \int u^2 K(u) f_X^{(1)}(x + \lambda h_n u)du \to f_X^{(1)}(x)\int u^2K(u)du := \mu_2 f_X^{(1)}(x) \text{ as } n \to \infty.\end{aligned}$$

Now,

$$\begin{aligned}V\left(\frac{1}{h_n^2}s_1(x)\right) &= \frac{1}{nh_n^3}\int K^2(u)u^2 f_X(x + h_n u)du - \frac{1}{nh_n^2}\left(\int K(u)uf_X(x + h_n u)du\right)^2 \\ &\to 0 \text{ if } nh_n^3 \to \infty.\end{aligned}$$

Hence, $s_1(x) = O_p(h_n^2)$ and $h_n^{-2}s_1(x) - \mu_2 f_X^{(1)}(x) = o_p(1)$.

Lastly, $E(s_{(1,1)}(x)) = h_n^2 \int K(u)u^2 f_X(x+h_n u)du$ and

$$V(s_{(1,1)}(x)) = \frac{h_n^3}{n}\int K^2(u)u^4 f_X(x+h_n u)du - \frac{h_n^4}{n}\left(\int K(u)u^2 f_X(x+h_n u)du\right)^2.$$

Hence, $\frac{1}{h_n^2}E(s_{(1,1)}(x)) = \int K(u)u^2 f_X(x+h_n u)du \to f_X(x)\mu_2$ as $n\to\infty$ and

$$V(h_n^{-2}s_{(1,1)}(x)) = \frac{1}{nh_n}\int K^2(u)u^4 f_X(x+h_n u)du - \frac{1}{n}\left(\int K(u)u^2 f_X(x+h_n u)du\right)^2.$$

Then, $h_n^{-2}s_{(1,1)}(x) - f_X(x)\mu_2 = o_p(1)$. □

Remark 8. *1. If $f_X(x) > 0$ and $\mu_2 > 0$, Q is invertible.*

2. It follows immediately from Slutsky Theorem and invertibility of Q that $(G_n S(x))^{-1} - Q^{-1} = o_p(1)$.

3. Recall the formula for the inverse of a partitioned matrix: if A_{11}, A_{22} and the Schur complements $A_{11} - A_{12}A_{22}^{-1}A_{21}$ and $A_{22} - A_{21}A_{11}^{-1}A_{12}$ are invertible, then

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} (A_{11}-A_{12}A_{22}^{-1}A_{21})^{-1} & -A_{11}^{-1}A_{12}(A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1} \\ -A_{22}^{-1}A_{21}(A_{11}-A_{12}A_{22}^{-1}A_{21})^{-1} & (A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1} \end{pmatrix}.$$

Applied to Q, this gives

$$Q^{-1} = \begin{pmatrix} \frac{1}{f_X(x)} & 0 \\ -\frac{f_X^{(1)}(x)}{f_X(x)^2} & \frac{1}{\mu_2 f_X(x)} \end{pmatrix}.$$

Applied to $G_n S(x)$, with $A_{11} = s_0(x)$, $A_{12} = s_1(x)$, $A_{21} = s_1(x)/h_n^2$ and $A_{22} = s_{(1,1)}(x)/h_n^2$ (all conditions hold with probability approaching one), and using Lemma 2 ,

$$A_{11} - A_{12}A_{22}^{-1}A_{21} = s_0(x) - \frac{s_1(x)^2}{s_{(1,1)}(x)} = s_0(x) + O_p(h_n^2) = f_X(x) + o_p(1),$$

$$A_{22} - A_{21}A_{11}^{-1}A_{12} = \frac{s_{(1,1)}(x)}{h_n^2} - \frac{s_1(x)}{h_n^2}\frac{s_1(x)}{s_0(x)} = \mu_2 f_X(x) + o_p(1),$$

$$A_{11}^{-1}A_{12} = \frac{s_1(x)}{s_0(x)} = O_p(h_n^2), \qquad A_{22}^{-1}A_{21} = \frac{s_1(x)}{s_{(1,1)}(x)} = \frac{f_X^{(1)}(x)}{f_X(x)} + o_p(1).$$

Therefore,

$$(G_n S(x))^{-1} = Q^{-1} + \begin{pmatrix} o_p(1) & O_p(h_n^2) \\ o_p(1) & o_p(1) \end{pmatrix}.$$

The $(1,1)$ error term is only $o_p(1)$ because the Lemma 2 gives $s_0(x) - f_X(x) = o_p(1)$, not $O_p(h_n^2)$.

4. *Note that if* $m \in \mathcal{C}_b^2$, *the* i^{th} *element of* $\mathbf{m} - R(x)\beta_{1,1}(x)$ *is given by* $m(X_t) - m(x) - m^{(1)}(x)(X_t - x) = \frac{1}{2}m^{(2)}(x + \lambda(X_t - x))(X_t - x)^2$

Theorem 8. *Let* $\{(Y_i, X_i)^T\}_{i=1}^n$ *be a sequence of IID random vectors distributed as* $(Y, X)^T$, *with* $m(x) := E(Y|X = x)$, $U := Y - m(X)$ *and* $\beta_{1,1}(x) := \left(m(x),\, m^{(1)}(x)\right)^T$. *Assume that*

a) K *is symmetric about zero, has compact support and* $|K(u)| \leq C$ *for all* $u \in \mathbb{R}$;

b) $\int K(u)\,du = 1$, $\mu_2 := \int u^2 K(u)\,du > 0$ *and put* $\nu_2 := \int u^2 K^2(u)\,du$;

c) $m, f_X \in \mathcal{C}_b^2$ *and* $f_X(x) > 0$, *where* f_X *is the marginal density of* X;

d) $V(Y|X) = \sigma^2$ *and* $E\left(|U|^{2+\delta} \,\middle|\, X\right) \leq C < \infty$ *for some* $\delta > 0$;

e) $h_n \to 0$, $nh_n^3 \to \infty$ *and* $nh_n^5 = O(1)$ *as* $n \to \infty$.

Then, with $N_n = \operatorname{diag}\left\{\sqrt{nh_n},\, \sqrt{nh_n}\, h_n\right\}$,

$$N_n\left(\hat{\beta}_{1,1}(x) - \beta_{1,1}(x) - \begin{pmatrix} \frac{1}{2}h_n^2\mu_2 m^{(2)}(x) \\ 0 \end{pmatrix}\right) \xrightarrow{d} N\left(\mathbf{0}, \frac{\sigma^2}{f_X(x)}\begin{pmatrix} \int K^2(u)\,du & 0 \\ 0 & \frac{\nu_2}{\mu_2^2} \end{pmatrix}\right).$$

Proof. Write the last equality in Lemma 1 (with $D = 1$) as

$$T_n := N_n\left(\hat{\beta}_{1,1}(x) - \beta_{1,1}(x) - B_n\right) = N_n(G_n S(x))^{-1} G_n v_n,$$

where $B_n := (G_n S(x))^{-1} \frac{1}{nh_n} G_n R(x)^T P(x)\left(\mathbf{m} - R(x)\beta_{1,1}(x)\right)$ and

$$v_n := \frac{1}{nh_n} R(x)^T P(x)\mathbf{U} = \begin{pmatrix} \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right) U_i \\ \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right)(X_i - x)U_i \end{pmatrix} := \begin{pmatrix} v_{n1} \\ v_{n2} \end{pmatrix}.$$

Step 1: moments of v_n. By assumption, $E(v_n) = \mathbf{0}$, and the change of variables $X_i = x + h_n u$ gives

$$V(v_{n1}) = \frac{\sigma^2}{nh_n}\int K^2(u) f_X(x + h_n u)\,du, \qquad V(v_{n2}) = \frac{\sigma^2 h_n}{n}\int u^2 K^2(u) f_X(x + h_n u)\,du,$$
$$E(v_{n1}v_{n2}) = \frac{\sigma^2}{n}\int u K^2(u) f_X(x + h_n u)\,du.$$

Hence $\sqrt{nh_n}\, v_{n1} = O_p(1)$ and $v_{n2} = O_p\left((h_n/n)^{1/2}\right)$, that is, $\sqrt{nh_n}\, v_{n2} = O_p(h_n)$ and $\sqrt{nh_n}\, v_{n2}/h_n = O_p(1)$.

Step 2: decomposition of T_n. Let $E_n := (G_n S(x))^{-1} - Q^{-1}$. By Remark 8.3, $E_n = \begin{pmatrix} o_p(1) & O_p(h_n^2) \\ o_p(1) & o_p(1) \end{pmatrix}$, and $G_n v_n = (v_{n1},\, v_{n2}/h_n^2)^T$. Then

$$T_n = N_n E_n G_n v_n + N_n Q^{-1} G_n v_n := T_{n1} + T_{n2},$$

with

$$T_{n1} = \begin{pmatrix} o_p(1)\sqrt{nh_n}\, v_{n1} + O_p(1)\sqrt{nh_n}\, v_{n2} \\ o_p(1)\, h_n\sqrt{nh_n}\, v_{n1} + o_p(1)\sqrt{nh_n}\, v_{n2}/h_n \end{pmatrix} = o_p(1)$$

by Step 1. Furthermore,

$$T_{n2} = \begin{pmatrix} \frac{1}{f_X(x)}\sqrt{nh_n}\, v_{n1} \\ -h_n \frac{f_X^{(1)}(x)}{f_X^2(x)}\sqrt{nh_n}\, v_{n1} + \frac{1}{\mu_2 f_X(x)}\frac{\sqrt{nh_n}}{h_n} v_{n2} \end{pmatrix} = T_{n2}^* + \begin{pmatrix} 0 \\ o_p(1) \end{pmatrix}, \qquad T_{n2}^* := \begin{pmatrix} \frac{1}{f_X(x)}\sqrt{nh_n}\, v_{n1} \\ \frac{1}{\mu_2 f_X(x)}\frac{\sqrt{nh_n}}{h_n} v_{n2} \end{pmatrix},$$

since $h_n\sqrt{nh_n}\, v_{n1} = O_p(h_n) = o_p(1)$. Hence $T_n = T_{n2}^* + o_p(1)$.

Step 3: asymptotic normality of T_{n2}^*. Let

$$\Sigma := \frac{\sigma^2}{f_X(x)} \begin{pmatrix} \int K^2(u)\, du & 0 \\ 0 & \frac{\nu_2}{\mu_2^2} \end{pmatrix}.$$

By the Cramér–Wold device it suffices to show that $a^T T_{n2}^* \xrightarrow{d} N(0, a^T\Sigma a)$ for every $a = (a_1, a_2)^T \in \mathbb{R}^2$, $a \neq \mathbf{0}$. Write $a^T T_{n2}^* = \sum_{i=1}^n \xi_{ni}$, where

$$\xi_{ni} := \frac{1}{\sqrt{nh_n}} K\left(\frac{X_i - x}{h_n}\right)\left(\frac{a_1}{f_X(x)} + \frac{a_2}{\mu_2 f_X(x)}\frac{X_i - x}{h_n}\right) U_i.$$

For each n the ξ_{ni} are i.i.d. with $E(\xi_{ni}) = 0$. By Step 1,

$$\begin{aligned} V(a^T T_{n2}^*) &= \frac{a_1^2}{f_X^2(x)}\, nh_n V(v_{n1}) + \frac{2a_1a_2}{\mu_2 f_X^2(x)}\, nE(v_{n1}v_{n2}) + \frac{a_2^2}{\mu_2^2 f_X^2(x)}\frac{n}{h_n} V(v_{n2}) \\ &\to \frac{a_1^2}{f_X^2(x)}\sigma^2 f_X(x)\int K^2(u)\, du + 0 + \frac{a_2^2}{\mu_2^2 f_X^2(x)}\sigma^2 f_X(x)\,\nu_2 = a^T\Sigma a, \end{aligned}$$

where the cross term vanishes because $\int uK^2(u)\, du = 0$ by the symmetry of K. Moreover, by assumption,

$$\begin{aligned} \sum_{i=1}^n E|\xi_{ni}|^{2+\delta} &\le \frac{Cn\, h_n}{(nh_n)^{1+\delta/2}} \int |K(u)|^{2+\delta} \left|\frac{a_1}{f_X(x)} + \frac{a_2 u}{\mu_2 f_X(x)}\right|^{2+\delta} f_X(x + h_n u)\, du \\ &= O\left((nh_n)^{-\delta/2}\right) \to 0, \end{aligned}$$

so Liapunov's condition holds. Since $a^T\Sigma a > 0$, Liapunov's CLT gives $a^T T_{n2}^* \xrightarrow{d} N(0, a^T\Sigma a)$. Therefore $T_{n2}^* \xrightarrow{d} N(\mathbf{0}, \Sigma)$ and, by Step 2, $T_n \xrightarrow{d} N(\mathbf{0}, \Sigma)$.

Step 4: the bias term. By Remark 8.4, the i^{th} element of $\mathbf{m} - R(x)\beta_{1,1}(x)$ is $\frac{1}{2}g(X_i)(X_i - x)^2$, where

$$g(X_i) := \frac{2\left[m(X_i) - m(x) - m^{(1)}(x)(X_i - x)\right]}{(X_i - x)^2} = m^{(2)}\big(x + \lambda_i(X_i - x)\big), \quad \lambda_i \in (0, 1).$$

Thus $|g| \le \sup|m^{(2)}| < \infty$ and $g(x + h_n u) \to m^{(2)}(x)$ for every u as $h_n \to 0$. Hence $\frac{1}{nh_n} G_n R(x)^T P(x)\,(\mathbf{m} - R(x)\beta_{1,1}(x)) = \frac{1}{2}(z_{n1}, z_{n2})^T$ with

$$z_{n1} = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right) g(X_i)(X_i - x)^2, \qquad z_{n2} = \frac{1}{nh_n^3}\sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right) g(X_i)(X_i - x)^3.$$

By dominated convergence,

$$E(z_{n1}) = h_n^2 \int u^2 K(u) g(x + h_n u) f_X(x + h_n u)\, du = h_n^2 \left(\mu_2 m^{(2)}(x) f_X(x) + o(1)\right),$$
$$E(z_{n2}) = h_n \int u^3 K(u) g(x + h_n u) f_X(x + h_n u)\, du = h_n \left(m^{(2)}(x) f_X(x) \int u^3 K(u)\, du + o(1) \right)$$
$$= o(h_n),$$

since $\int u^3 K(u)\, du = 0$. Also,

$$V(z_{n1}) \le \frac{1}{nh_n^2} E\left[K^2\left(\tfrac{X_i - x}{h_n}\right) g^2(X_i)(X_i - x)^4\right] = O\left(\frac{h_n^3}{n}\right) = o(h_n^4),\ V(z_{n2}) = O\left(\frac{h_n}{n}\right) = o(h_n^2),$$

because $nh_n \to \infty$. Consequently, $z_{n1} = h_n^2 \left(\mu_2 m^{(2)}(x) f_X(x) + o_p(1)\right)$ and $z_{n2} = o_p(h_n)$.

Now write

$$N_n B_n = \frac{1}{2} N_n E_n \begin{pmatrix} z_{n1} \\ z_{n2} \end{pmatrix} + \frac{1}{2} N_n Q^{-1} \begin{pmatrix} z_{n1} \\ z_{n2} \end{pmatrix} := B_{n1} + B_{n2}.$$

Using the orders of E_n, z_{n1} and z_{n2},

$$B_{n1} = \frac{1}{2} \begin{pmatrix} \sqrt{nh_n}\left[o_p(1) z_{n1} + O_p(h_n^2) z_{n2}\right] \\ h_n \sqrt{nh_n}\left[o_p(1) z_{n1} + o_p(1) z_{n2}\right] \end{pmatrix} = \begin{pmatrix} o_p\big(\sqrt{nh_n^5}\big) \\ o_p\big(\sqrt{nh_n^5}\big) \end{pmatrix} = o_p(1),$$

and

$$B_{n2} = \frac{1}{2} \begin{pmatrix} \sqrt{nh_n}\, \frac{z_{n1}}{f_X(x)} \\ h_n \sqrt{nh_n} \left[-\frac{f_X^{(1)}(x)}{f_X^2(x)} z_{n1} + \frac{1}{\mu_2 f_X(x)} z_{n2} \right] \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\sqrt{nh_n}\, h_n^2 \mu_2 m^{(2)}(x) + o_p\big(\sqrt{nh_n^5}\big) \\ O_p\big(\sqrt{nh_n^7}\big) + o_p\big(\sqrt{nh_n^5}\big) \end{pmatrix}.$$

Since $nh_n^5 = O(1)$, letting $b_n := \left(\frac{1}{2} h_n^2 \mu_2 m^{(2)}(x),\, 0\right)^T$ we obtain $N_n B_n = N_n b_n + o_p(1)$.

Conclusion. Combining Steps 3 and 4,

$$N_n \left(\hat{\beta}_{1,1}(x) - \beta_{1,1}(x) - b_n\right) = T_n + N_n(B_n - b_n) = T_n + o_p(1) \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

that is,

$$\sqrt{nh_n}\left(\hat{m}(x) - m(x) - \tfrac{1}{2} h_n^2 \mu_2 m^{(2)}(x)\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{f_X(x)} \int K^2(u)\, du\right),$$
$$\sqrt{nh_n^3}\left(\hat{m}^{(1)}(x) - m^{(1)}(x)\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{f_X(x)} \frac{\nu_2}{\mu_2^2}\right),$$

jointly, with asymptotically independent components. □

4.3 The case where $D \geq 1$

Let $K : \mathbb{R}^D \to \mathbb{R}$ be a product kernel such $K(u) = k(u_1) \cdots k(u_D)$ and k is a univariate kernel.

Lemma 3. *Let $H_n = \mathrm{diag}\{h_1, \ldots, h_D\}$, $\bar{h}_n := \max_d h_d$ and $\eta_n := n \det(H_n)$. Assume that $\{X_i\}_{i=1}^n$ is IID with density $f_X \in \mathcal{C}_b^2$; $K : \mathbb{R}^D \to \mathbb{R}$ is bounded, has compact support and satisfies $K(u_1, \ldots, -u_d, \ldots, u_D) = K(u)$ for every d; $\int K(u)\, du = 1$ and $\int u_d^2 K(u)\, du = \mu_2$ for $d = 1, \ldots, D$; $\bar{h}_n \to 0$, $\bar{h}_n / h_d \leq C$ for all d and n, and $\eta_n \bar{h}_n^2 \to \infty$. Then, with $\mathbf{0}$ a $D \times 1$ vector of zeros and I_D the $D \times D$ identity,*

$$G_n S(x) = Q + \begin{pmatrix} o_p(1) & O_p(\bar{h}_n^2) \\ o_p(1) & o_p(1) \end{pmatrix}, \qquad Q := \begin{pmatrix} f_X(x) & \mathbf{0}^T \\ \mu_2 f_X^{(1)}(x) & \mu_2 f_X(x) I_D \end{pmatrix},$$

where $f_X^{(1)}(x)$ is the gradient of f_X at x and each entry of a block of the second matrix has the indicated order.

Proof. Note first that the symmetry of K implies $\int u_d K(u)\, du = 0$ and $\int u_d u_j K(u)\, du = 0$ for $d \neq j$; the ratios $h_j / h_d \leq \bar{h}_n / h_d \leq C$ are bounded; and $\eta_n \to \infty$. Write $G_n S(x) = \begin{pmatrix} s_0(x) & S_1(x)^T \\ \tilde{G}_n S_1(x) & \tilde{G}_n S_2(x) \end{pmatrix}$ with $\tilde{G}_n := \mathrm{diag}\{1/h_d^2\}_{d=1}^D$. All expectations below follow from the change of variables $X_i = x + H_n u$, and all limits from dominated convergence, which applies because K has compact support and f_X and its first derivatives are bounded and continuous.

As for the Nadaraya–Watson estimator, $E(s_0(x)) = \int K(u) f_X(x + H_n u)\, du \to f_X(x)$ and $V(s_0(x)) \leq \frac{1}{\eta_n} \int K^2(u) f_X(x + H_n u)\, du \to 0$, so $s_0(x) = f_X(x) + o_p(1)$.

Next, $E(s_j(x)) = h_j \int u_j K(u) f_X(x + H_n u)\, du$. With

$$f_X(x + H_n u) = f_X(x) + \sum_{k=1}^D h_k u_k \frac{\partial}{\partial x_k} f_X(x + \lambda H_n u),$$

$\lambda = \lambda(u) \in (0, 1)$, and $\int u_j K(u)\, du = 0$,

$$\frac{1}{h_j^2} E(s_j(x)) = \sum_{k=1}^D \frac{h_k}{h_j} \int u_j u_k K(u) \frac{\partial}{\partial x_k} f_X(x + \lambda H_n u)\, du \to \mu_2 \frac{\partial}{\partial x_j} f_X(x),$$

since the term $k = j$ converges to $\mu_2 \frac{\partial}{\partial x_j} f_X(x)$, while for $k \neq j$ the integral converges to $\frac{\partial}{\partial x_k} f_X(x) \int u_j u_k K(u)\, du = 0$ and h_k / h_j is bounded. Also,

$$V\left(\frac{1}{h_j^2} s_j(x)\right) \leq \frac{1}{\eta_n h_j^4} h_j^2 \int u_j^2 K^2(u) f_X(x + H_n u)\, du = O\left(\frac{1}{\eta_n h_j^2}\right) \to 0,$$

because $\eta_n h_j^2 \geq \eta_n \bar{h}_n^2 / C^2 \to \infty$. Hence $h_j^{-2} s_j(x) - \mu_2 \frac{\partial}{\partial x_j} f_X(x) = o_p(1)$ and, in particular, $s_j(x) = O_p(h_j^2) = O_p(\bar{h}_n^2)$.

Lastly, $E(s_{(d,j)}(x)) = h_d h_j \int u_d u_j K(u) f_X(x + H_n u)\, du$, so that

$$\frac{1}{h_d^2} E(s_{(d,j)}(x)) = \frac{h_j}{h_d} \int u_d u_j K(u) f_X(x + H_n u)\, du \to \mu_2 f_X(x)\, \mathbb{1}\{d = j\},$$

again because h_j/h_d is bounded and $\int u_d u_j K(u)\, du = 0$ for $d \neq j$. Moreover,

$$V\left(\frac{1}{h_d^2} s_{(d,j)}(x)\right) \leq \frac{1}{\eta_n h_d^4} h_d^2 h_j^2 \int u_d^2 u_j^2 K^2(u) f_X(x + H_n u)\, du = O\left(\frac{1}{\eta_n}\right) \to 0.$$

Hence $h_d^{-2} s_{(d,j)}(x) - \mu_2 f_X(x) \mathbb{1}\{d = j\} = o_p(1)$, which completes the proof. □

Remark 9. *1. If $f_X(x) > 0$ and $\mu_2 > 0$, then $\det(Q) = \mu_2^D f_X(x)^{D+1} \neq 0$ and*

$$Q^{-1} = \begin{pmatrix} \frac{1}{f_X(x)} & \mathbf{0}^T \\ -\frac{f_X^{(1)}(x)}{f_X(x)^2} & \frac{1}{\mu_2 f_X(x)} I_D \end{pmatrix}.$$

2. By Lemma 3, $G_n S(x) \xrightarrow{p} Q$, so $G_n S(x)$ is invertible with probability approaching one and, by the continuous mapping theorem, $(G_n S(x))^{-1} - Q^{-1} = o_p(1)$.

3. Applying the partitioned inverse formula of Remark 8.3 with $A_{11} = s_0(x)$, $A_{12} = S_1(x)^T = O_p(\bar{h}_n^2)$, $A_{21} = \tilde{G}_n S_1(x) = \mu_2 f_X^{(1)}(x) + o_p(1)$ and $A_{22} = \tilde{G}_n S_2(x) = \mu_2 f_X(x) I_D + o_p(1)$ gives $A_{11} - A_{12} A_{22}^{-1} A_{21} = f_X(x) + o_p(1)$, $A_{22} - A_{21} A_{11}^{-1} A_{12} = \mu_2 f_X(x) I_D + o_p(1)$ and $A_{11}^{-1} A_{12} = O_p(\bar{h}_n^2)$. Therefore

$$E_n := (G_n S(x))^{-1} - Q^{-1} = \begin{pmatrix} o_p(1) & O_p(\bar{h}_n^2) \\ o_p(1) & o_p(1) \end{pmatrix},$$

with each entry of a block of the indicated order.

4. If $m \in \mathcal{C}_b^2$, with Hessian $m^{(2)}$, the i^{th} element of $\mathbf{m} - R(x)\beta_{1,D}(x)$ is

$$\begin{aligned} r(X_i) &:= m(X_i) - m(x) - m^{(1)}(x)^T (X_i - x) \\ &= \frac{1}{2}(X_i - x)^T m^{(2)}\big(x + \lambda_i (X_i - x)\big)(X_i - x), \quad \lambda_i \in (0, 1). \end{aligned}$$

Theorem 9. *Let $\{(Y_i, X_i^T)^T\}_{i=1}^n$ be a sequence of IID random vectors distributed as $(Y, X^T)^T$, where $X \in \mathbb{R}^D$, with $m(x) := E(Y|X = x)$, $U := Y - m(X)$ and $\beta_{1,D}(x) := \big(m(x), m^{(1)}(x)^T\big)^T$. Let $H_n = \mathrm{diag}\{h_1, \ldots, h_D\}$, $\bar{h}_n := \max_d h_d$ and $\eta_n := n \det(H_n)$. Assume that*

a) $K : \mathbb{R}^D \to \mathbb{R}$ has compact support, $|K(u)| \leq C$ for all $u \in \mathbb{R}^D$, and

$$K(u_1, \ldots, -u_d, \ldots, u_D) = K(u)$$

for every $d = 1, \ldots, D$;

b) $\int K(u)\,du = 1$, $\int u_d^2 K(u)\,du = \mu_2 > 0$ *and* $\int u_d^2 K^2(u)\,du = \nu_2$ *for every* $d = 1, \ldots, D$;

c) $m, f_X \in \mathcal{C}_b^2$ *and* $f_X(x) > 0$, *where* f_X *is the marginal density of* X;

d) $V(Y|X) = \sigma^2$ *and* $E\left(|U|^{2+\delta} \,\middle|\, X\right) \leq C < \infty$ *for some* $\delta > 0$;

e) $\bar{h}_n \to 0$, $\bar{h}_n / h_d \leq C$ *for all* d *and* n, $\eta_n \bar{h}_n^2 \to \infty$ *and* $\eta_n \bar{h}_n^4 = O(1)$ *as* $n \to \infty$.

Then, with $N_n = \operatorname{diag}\left\{\sqrt{\eta_n}, \operatorname{diag}\{\sqrt{\eta_n}\, h_d\}_{d=1}^D\right\}$ *and*

$$b_n := \begin{pmatrix} \frac{1}{2}\mu_2 \sum_{d=1}^D h_d^2 \frac{\partial^2}{\partial x_d^2} m(x) \\ \mathbf{0} \end{pmatrix},$$

$$N_n \left(\hat{\beta}_{1,D}(x) - \beta_{1,D}(x) - b_n\right) \xrightarrow{d} N\left(\mathbf{0}, \frac{\sigma^2}{f_X(x)} \begin{pmatrix} \int K^2(u)\,du & \mathbf{0}^T \\ \mathbf{0} & \frac{\nu_2}{\mu_2^2} I_D \end{pmatrix}\right).$$

Proof. Throughout, $K_i := K\left(H_n^{-1}(X_i - x)\right)$. Recall that the symmetry of K implies $\int u_d K(u)\,du = 0$, $\int u_d u_j K(u)\,du = \int u_d u_j K^2(u)\,du = 0$ for $d \neq j$, $\int u_d K^2(u)\,du = 0$ and $\int u_d u_j u_k K(u)\,du = 0$ for all d, j, k, since in each of these integrands some coordinate appears an odd number of times. Also, $h_j / h_d \leq C$ for all d, j.

Write the last equality in Lemma 1 as

$$T_n := N_n \left(\hat{\beta}_{1,D}(x) - \beta_{1,D}(x) - B_n\right) = N_n (G_n S(x))^{-1} G_n v_n,$$

where $B_n := (G_n S(x))^{-1} \frac{1}{\eta_n} G_n R(x)^T P(x) \left(\mathbf{m} - R(x)\beta_{1,D}(x)\right)$ and

$$v_n := \frac{1}{\eta_n} R(x)^T P(x) \mathbf{U} = \begin{pmatrix} \frac{1}{\eta_n}\sum_{i=1}^n K_i U_i \\ \frac{1}{\eta_n}\sum_{i=1}^n K_i (X_{i1} - x_1) U_i \\ \vdots \\ \frac{1}{\eta_n}\sum_{i=1}^n K_i (X_{iD} - x_D) U_i \end{pmatrix} := \begin{pmatrix} v_{n0} \\ v_{n1} \\ \vdots \\ v_{nD} \end{pmatrix}.$$

Step 1. By d), $E(v_n) = \mathbf{0}$, and the change of variables $X_i = x + H_n u$ gives

$$V(v_{n0}) = \frac{\sigma^2}{\eta_n} \int K^2(u) f_X(x + H_n u)\,du, \qquad V(v_{nd}) = \frac{\sigma^2 h_d^2}{\eta_n} \int u_d^2 K^2(u) f_X(x + H_n u)\,du.$$

Hence $\sqrt{\eta_n}\, v_{n0} = O_p(1)$ and $\sqrt{\eta_n}\, v_{nd} / h_d = O_p(1)$ for $d = 1, \ldots, D$.

Step 2. By Remark 9.3,

$$E_n := (G_n S(x))^{-1} - Q^{-1} = \begin{pmatrix} o_p(1) & O_p(\bar{h}_n^2) \\ o_p(1) & o_p(1) \end{pmatrix},$$

and $G_n v_n = \left(v_{n0},\, v_{n1}/h_1^2, \ldots, v_{nD}/h_D^2\right)^T$. Then

$$T_n = N_n E_n G_n v_n + N_n Q^{-1} G_n v_n := T_{n1} + T_{n2}.$$

The first component of T_{n1} is

$$o_p(1)\sqrt{\eta_n}\, v_{n0} + \sum_{j=1}^{D} O_p\left(\frac{\bar{h}_n^2}{h_j}\right) \frac{\sqrt{\eta_n}\, v_{nj}}{h_j} = o_p(1) + O_p(\bar{h}_n) = o_p(1),$$

and, for $d = 1, \ldots, D$, its $(d+1)^{th}$ component is

$$o_p(1)\, h_d \sqrt{\eta_n}\, v_{n0} + \sum_{j=1}^{D} o_p(1) \frac{h_d}{h_j} \frac{\sqrt{\eta_n}\, v_{nj}}{h_j} = o_p(1),$$

by Step 1 and the boundedness of $\bar{h}_n/h_j$ and h_d/h_j. Hence $T_{n1} = o_p(1)$. Furthermore, by Remark 9.1, the first component of T_{n2} is $\frac{1}{f_X(x)}\sqrt{\eta_n}\, v_{n0}$ and its $(d+1)^{th}$ component is

$$-h_d \frac{\frac{\partial}{\partial x_d} f_X(x)}{f_X^2(x)} \sqrt{\eta_n}\, v_{n0} + \frac{1}{\mu_2 f_X(x)} \frac{\sqrt{\eta_n}}{h_d} v_{nd} = \frac{1}{\mu_2 f_X(x)} \frac{\sqrt{\eta_n}}{h_d} v_{nd} + o_p(1),$$

since $h_d\sqrt{\eta_n}\, v_{n0} = O_p(h_d) = o_p(1)$. Hence $T_n = T_{n2}^* + o_p(1)$, where

$$T_{n2}^* := \begin{pmatrix} \frac{1}{f_X(x)}\sqrt{\eta_n}\, v_{n0} \\ \frac{1}{\mu_2 f_X(x)} \frac{\sqrt{\eta_n}}{h_1} v_{n1} \\ \vdots \\ \frac{1}{\mu_2 f_X(x)} \frac{\sqrt{\eta_n}}{h_D} v_{nD} \end{pmatrix}.$$

Step 3: asymptotic normality of T_{n2}^*. Let

$$\Sigma := \frac{\sigma^2}{f_X(x)} \begin{pmatrix} \int K^2(u)\, du & \mathbf{0}^T \\ \mathbf{0} & \frac{\nu_2}{\mu_2^2} I_D \end{pmatrix}.$$

By the Cramér–Wold device it suffices to show that $a^T T_{n2}^* \xrightarrow{d} N(0, a^T \Sigma a)$ for every $a = (a_0, a_1, \ldots, a_D)^T \in \mathbb{R}^{D+1}$, $a \neq \mathbf{0}$. Put $c_a(u) := \frac{a_0}{f_X(x)} + \frac{1}{\mu_2 f_X(x)} \sum_{d=1}^{D} a_d u_d$ and write $a^T T_{n2}^* = \sum_{i=1}^{n} \xi_{ni}$, where

$$\xi_{ni} := \frac{1}{\sqrt{\eta_n}} K_i\, c_a\left(H_n^{-1}(X_i - x)\right) U_i.$$

For each n the ξ_{ni} are IID with $E(\xi_{ni}) = 0$, and

$$\begin{aligned} V(a^T T_{n2}^*) = nE(\xi_{ni}^2) &= \sigma^2 \int K^2(u) c_a^2(u) f_X(x + H_n u)\, du \to \sigma^2 f_X(x) \int K^2(u) c_a^2(u)\, du \\ &= \frac{\sigma^2}{f_X(x)} \left(a_0^2 \int K^2(u)\, du + \frac{2a_0}{\mu_2} \sum_{d=1}^{D} a_d \int u_d K^2(u)\, du \right. \\ &\qquad \left. + \frac{1}{\mu_2^2} \sum_{d=1}^{D} \sum_{j=1}^{D} a_d a_j \int u_d u_j K^2(u)\, du \right) \\ &= \frac{\sigma^2}{f_X(x)} \left(a_0^2 \int K^2(u)\, du + \frac{\nu_2}{\mu_2^2} \sum_{d=1}^{D} a_d^2 \right) = a^T \Sigma a, \end{aligned}$$

where the cross terms vanish by the symmetry of K. Moreover, by a), c) and d),

$$\begin{aligned}\sum_{i=1}^{n} E|\xi_{ni}|^{2+\delta} &\le \frac{Cn\det(H_n)}{\eta_n^{1+\delta/2}} \int |K(u)|^{2+\delta}|c_a(u)|^{2+\delta} f_X(x+H_nu)\,du \\ &= O\left(\eta_n^{-\delta/2}\right) \to 0,\end{aligned}$$

so Liapunov's condition holds. Since $a^T\Sigma a > 0$, Liapunov's CLT and Slutsky's theorem give $a^T T_{n2}^* \xrightarrow{d} N(0, a^T\Sigma a)$. Therefore $T_{n2}^* \xrightarrow{d} N(\mathbf{0}, \Sigma)$ and, by Step 2, $T_n \xrightarrow{d} N(\mathbf{0}, \Sigma)$.
Step 4: the bias term. By Remark 9.4,

$$\frac{1}{\eta_n} G_n R(x)^T P(x)\left(\mathbf{m} - R(x)\beta_{1,D}(x)\right) = (z_{n0}, z_{n1}, \ldots, z_{nD})^T$$

with

$$z_{n0} = \frac{1}{\eta_n}\sum_{i=1}^{n} K_i\, r(X_i), \qquad z_{nd} = \frac{1}{\eta_n h_d^2}\sum_{i=1}^{n} K_i(X_{id} - x_d)\, r(X_i), \quad d = 1, \ldots, D.$$

Let $M := \sup \|m^{(2)}\|$. Then $|r(x+H_nu)| \le \frac{1}{2}M\bar{h}_n^2\|u\|^2$ and $r(x+H_nu) = \frac{1}{2}(H_nu)^T m^{(2)}(x)(H_nu) + \rho_n(u)$, where

$$|\rho_n(u)| \le \frac{1}{2}\bar{h}_n^2\|u\|^2 \sup_{\|z-x\|\le \bar{h}_n\|u\|} \left\|m^{(2)}(z) - m^{(2)}(x)\right\|,$$

so that $\sup_{u\in\operatorname{supp} K}|\rho_n(u)| = o(\bar{h}_n^2)$ by the continuity of $m^{(2)}$. By dominated convergence and the moment properties of K,

$$\begin{aligned}E(z_{n0}) &= \int K(u)\, r(x+H_nu) f_X(x+H_nu)\,du \\ &= \frac{1}{2}\sum_{d=1}^{D}\sum_{j=1}^{D} h_d h_j \frac{\partial^2 m(x)}{\partial x_d \partial x_j}\int u_d u_j K(u) f_X(x+H_nu)\,du + o(\bar{h}_n^2) \\ &= f_X(x)\,\frac{1}{2}\mu_2 \sum_{d=1}^{D} h_d^2 \frac{\partial^2}{\partial x_d^2} m(x) + o(\bar{h}_n^2), \\ E(z_{nd}) &= \frac{1}{h_d}\int u_d K(u)\, r(x+H_nu) f_X(x+H_nu)\,du \\ &= \frac{1}{2h_d}\sum_{j=1}^{D}\sum_{k=1}^{D} h_j h_k \frac{\partial^2 m(x)}{\partial x_j \partial x_k}\int u_d u_j u_k K(u) f_X(x+H_nu)\,du + o\left(\frac{\bar{h}_n^2}{h_d}\right) \\ &= o(\bar{h}_n),\end{aligned}$$

since $\int u_d u_j K(u) f_X(x+H_nu)\,du \to \mu_2 f_X(x)\mathbb{1}\{d=j\}$, $\int u_d u_j u_k K(u) f_X(x+H_nu)\,du \to f_X(x)\int u_d u_j u_k K(u)\,du = 0$ and $\bar{h}_n/h_d \le C$. Also,

$$\begin{aligned}V(z_{n0}) &\le \frac{1}{\eta_n}\int K^2(u)\, r^2(x+H_nu) f_X(x+H_nu)\,du = O\left(\frac{\bar{h}_n^4}{\eta_n}\right) = o(\bar{h}_n^4), \\ V(z_{nd}) &\le \frac{1}{\eta_n h_d^2}\int u_d^2 K^2(u)\, r^2(x+H_nu) f_X(x+H_nu)\,du = O\left(\frac{\bar{h}_n^4}{\eta_n h_d^2}\right) = o(\bar{h}_n^2),\end{aligned}$$

because $\eta_n \to \infty$ and $h_d \geq \bar{h}_n/C$ for every d. Consequently, with $b_{n0} := \frac{1}{2}\mu_2 \sum_{d=1}^{D} h_d^2 \frac{\partial^2}{\partial x_d^2} m(x) = O(\bar{h}_n^2)$,

$$z_{n0} = f_X(x) b_{n0} + o_p(\bar{h}_n^2) \quad \text{and} \quad z_{nd} = o_p(\bar{h}_n),\ d = 1, \ldots, D.$$

Now write $z_n := (z_{n0}, \ldots, z_{nD})^T$ and

$$N_n B_n = N_n E_n z_n + N_n Q^{-1} z_n := B_{n1} + B_{n2}.$$

Using the orders of E_n and z_n, the first component of B_{n1} is $\sqrt{\eta_n}\left[o_p(1) z_{n0} + \sum_j O_p(\bar{h}_n^2) z_{nj}\right] = o_p\big(\sqrt{\eta_n \bar{h}_n^4}\big)$ and its $(d+1)^{th}$ component is $\sqrt{\eta_n}\, h_d \left[o_p(1) z_{n0} + \sum_j o_p(1) z_{nj}\right] = o_p\big(\sqrt{\eta_n \bar{h}_n^4}\big)$, so $B_{n1} = o_p(1)$ by e). Similarly, the first component of B_{n2} is

$$\sqrt{\eta_n}\, \frac{z_{n0}}{f_X(x)} = \sqrt{\eta_n}\, b_{n0} + o_p\big(\sqrt{\eta_n \bar{h}_n^4}\big) = \sqrt{\eta_n}\, b_{n0} + o_p(1),$$

and its $(d+1)^{th}$ component is

$$\sqrt{\eta_n}\, h_d \left[-\frac{\frac{\partial}{\partial x_d} f_X(x)}{f_X^2(x)} z_{n0} + \frac{1}{\mu_2 f_X(x)} z_{nd} \right] = O_p\big(\sqrt{\eta_n \bar{h}_n^6}\big) + o_p\big(\sqrt{\eta_n \bar{h}_n^4}\big) = o_p(1).$$

Hence $N_n B_n = N_n b_n + o_p(1)$, with $b_n = (b_{n0}, \mathbf{0}^T)^T$ as in the statement.

Combining Steps 3 and 4,

$$N_n \left(\hat{\beta}_{1,D}(x) - \beta_{1,D}(x) - b_n \right) = T_n + N_n (B_n - b_n) = T_n + o_p(1) \xrightarrow{d} N(\mathbf{0}, \Sigma).$$

□

5 Uniform orders of the local linear coefficient vector

The objective in this subsection is to establish that for the vector

$$\beta_{1,D}(x) = \big(m(x), \nabla m(x)^T\big)^T, \qquad \sup_{x\in\mathcal{C}} \left\| \widehat{\beta}_{1,D}(x) - \beta_{1,D}(x) \right\| \xrightarrow{p} 0,$$

on a fixed nonempty compact set $\mathcal{C} \subset \mathbb{R}^D$. The dimension D is fixed. Vector norms are Euclidean; matrix norms are their induced operator norms.

5.1 Assumptions and main result

Write $U_i = Y_i - m(X_i)$ and, using the notation of the notes, put

$$H_n = \operatorname{diag}(h_{1,n}, \ldots, h_{D,n}), \quad \eta_n = n \det(H_n), \quad \underline{h}_n = \min_d h_{d,n}, \quad \overline{h}_n = \max_d h_{d,n}.$$

The following conditions are sufficient.

A1. The observations (Y_i, X_i) are IID, $m(X) = \mathbb{E}(Y \mid X)$, and for some $q = 2 + \delta > 2$,

$$\mathbb{E}(|U|^q \mid X) \le C_U < \infty \quad \text{almost surely.} \tag{6}$$

This is the error-moment condition already used in Theorem 9 of the notes.

A2. For some $\rho > 0$, let $\mathcal{C}_\rho = \{z : \operatorname{dist}(z, \mathcal{C}) \le \rho\}$. The density f_X is continuous and m is twice continuously differentiable on an open neighborhood of $\mathcal{C}_\rho$. Also $f_X(x) > 0$ for every $x \in \mathcal{C}$. In particular,

$$f_* := \inf_{x\in\mathcal{C}} f_X(x) > 0. \tag{7}$$

These local smoothness conditions follow from the notes' stronger $m, f_X \in \mathcal{C}_b^2$ assumptions, with positivity imposed throughout the set on which convergence is sought.

A3. The kernel $K : \mathbb{R}^D \to \mathbb{R}$ is bounded, supported in $\{u : \|u\| \le B\}$ for a finite B, symmetric separately in each coordinate, and satisfies

$$\int K(u)\, du = 1, \qquad \int u_d^2 K(u)\, du = \mu_2 > 0 \quad (d = 1, \ldots, D).$$

In addition, for some $\alpha \in (0, 1]$ and $L_K < \infty$,

$$|K(u) - K(w)| \le L_K \|u - w\|^\alpha \quad (u, w \in \mathbb{R}^D). \tag{8}$$

Only (8) is an additional kernel condition. A Lipschitz kernel has $\alpha = 1$. Nonnegativity and a product representation are not required.

A4. The bandwidths are deterministic and positive, $\overline{h}_n \to 0$, and $\overline{h}_n/\underline{h}_n \le C_H$. Replace the bandwidth conditions used for pointwise asymptotic normality by

$$\frac{\log n}{n \det(H_n) \underline{h}_n^2} \longrightarrow 0, \tag{9}$$

$$\frac{(\log n/n)^{1-1/q}}{\det(H_n)\underline{h}_n} \longrightarrow 0. \tag{10}$$

The second condition allows for unbounded errors under the existing finite qth moment. These are explicit sufficient conditions; no optimality of their logarithmic factors is claimed.

Define the deterministic sequence

$$a_n = \sqrt{\frac{\log n}{n \det(H_n)}} + \frac{(\log n/n)^{1-1/q}}{\det(H_n)}. \tag{11}$$

Conditions (9)–(10) say exactly that $a_n/\underline{h}_n \to 0$. For $R(x)$ and $P(x)$ as defined on page 15 of the notes, write

$$\widehat{\beta}_{1,D}(x) = \{R(x)^T P(x) R(x)\}^{-1} R(x)^T P(x) \mathbf{Y} = \left(\widehat{m}_1(x), \widehat{\beta}_1(x)^T\right)^T$$

on the event that the weighted Gram matrix is positive definite for every $x \in \mathcal{C}$. For definiteness, set the entire estimated function to zero on the complementary event. This gives a continuous random function on $\mathcal{C}$, and the theorem shows that this extension affects an event whose probability tends to zero.

Theorem 10. *Under **A1–A4**,*

$$\mathbb{P}\big(R(x)^T P(x) R(x) \text{ is positive definite for every } x \in \mathcal{C}\big) \longrightarrow 1. \tag{12}$$

Thus the local weighted least-squares minimizer exists uniquely throughout $\mathcal{C}$ with probability tending to one. Moreover,

$$\sup_{x \in \mathcal{C}} |\widehat{m}_1(x) - m(x)| = O_p(\overline{h}_n^2 + a_n), \tag{13}$$

$$\sup_{x \in \mathcal{C}} \left|\widehat{\beta}_{1,d}(x) - \partial_d m(x)\right| = O_p\left(\frac{\overline{h}_n^2 + a_n}{h_{d,n}}\right), \qquad d = 1, \dots, D. \tag{14}$$

In particular,

$$\sup_{x \in \mathcal{C}} \left\|\widehat{\beta}_{1,D}(x) - \beta_{1,D}(x)\right\| = O_p\left(\overline{h}_n + \sqrt{\frac{\log n}{n \det(H_n) \underline{h}_n^2}} + \frac{(\log n/n)^{1-1/q}}{\det(H_n)\underline{h}_n}\right) = o_p(1). \tag{15}$$

The proof follows after the supporting lemmas. It uses a symmetric rescaling of the Gram matrix, which isolates the factor $1/h_{d,n}$ required to recover each unscaled derivative coefficient.

5.2 A uniform bound for kernel sums

Lemma 4. *Let $\mathcal{C}$ be compact and let f_X be bounded on a fixed neighborhood of $\mathcal{C}$. Suppose $\overline{h}_n \to 0$ and (9) holds. Let (Z_i, X_i) be IID with $\mathbb{E}(|Z|^q \mid X) \le C$ for a fixed $q > 2$. For a bounded, compactly supported, globally α-Hölder function $L : \mathbb{R}^D \to \mathbb{R}$, put*

$$T_{n,L,Z}(x) = \frac{1}{n \det(H_n)} \sum_{i=1}^{n} Z_i L\big(H_n^{-1}(X_i - x)\big).$$

Then, with a_n in (11),

$$\sup_{x \in \mathcal{C}} |T_{n,L,Z}(x) - \mathbb{E}T_{n,L,Z}(x)| = O_p(a_n). \tag{16}$$

The result holds simultaneously for any fixed finite collection of such functions L. If Z is bounded, the right side improves to $O_p(\sqrt{\log n/(n \det(H_n))})$.

Proof. Constants below may depend on $\mathcal{C}, L, D, q$, but not on n or x. All kernel windows lie in the specified neighborhood for sufficiently large n.
Step 1: Let $\ell_n = \log n$, $\tau_n = (n/\ell_n)^{1/q}$ and $Z_i^{(n)} = Z_i I_{\{|Z_i| \le \tau_n\}}$. Put

$$F_n(x) := T_{n,L,Z}(x) - \mathbb{E}T_{n,L,Z}(x), \qquad S_n(x) := T_{n,L,Z^{(n)}}(x) - \mathbb{E}T_{n,L,Z^{(n)}}(x).$$

Note that $F_n(x) - S_n(x) = T_{n,L,Z-Z^{(n)}}(x) - \mathbb{E}T_{n,L,Z-Z^{(n)}}(x)$. Since L is bounded we have for $Z_i - Z_i^{(n)}$

$$\begin{aligned} |F_n(x) - S_n(x)| &= \left|T_{n,L,Z-Z^{(n)}}(x) - \mathbb{E}T_{n,L,Z-Z^{(n)}}(x)\right| \\ &\le \frac{\|L\|_\infty}{\det(H_n)} \left\{ \frac{1}{n} \sum_{i=1}^{n} |Z_i|\, I_{\{|Z_i| > \tau_n\}} + \mathbb{E}\big[|Z|\, I_{\{|Z| > \tau_n\}}\big] \right\}. \end{aligned}$$

For any $\tau > 0$ and $q > 1$, $|Z| > \tau$ is equivalent to $|Z|^{q-1} > \tau^{q-1}$ or $|Z| < \tau^{1-q}|Z|^q$. Hence, $|Z|\mathbf{I}_{\{|Z|>\tau\}} \le \tau^{1-q}|Z|^q$. Now, since $\mathbb{E}|Z|^q = \mathbb{E}[\mathbb{E}(|Z|^q \mid X)] \le C$ we have

$$\mathbb{E}\big[|Z|\mathbf{1}_{\{|Z|>\tau_n\}}\big] \le C\tau_n^{1-q}.$$

Let $B_n = \frac{\|L\|_\infty}{\det(H_n)} \left\{ \frac{1}{n} \sum_{i=1}^{n} |Z_i| I_{\{|Z_i|>\tau_n\}} + \mathbb{E}\big[|Z| I_{\{|Z|>\tau_n\}}\big] \right\}$, and because the Z_i have the same distribution as Z, taking expectations gives

$$\mathbb{E}B_n = \frac{2\|L\|_\infty}{\det(H_n)} \mathbb{E}\big[|Z|\mathbf{1}_{\{|Z|>\tau_n\}}\big] \le \frac{C'\tau_n^{1-q}}{\det(H_n)},$$

where $C' := 2\|L\|_\infty C$ is independent of n. Substituting $\tau_n = (n/\ell_n)^{1/q}$,

$$\tau_n^{1-q} = \left(\frac{n}{\ell_n}\right)^{(1-q)/q} = \left(\frac{\ell_n}{n}\right)^{1-1/q},$$

and $C'\tau_n^{1-q}/\det(H_n) = C(\ell_n/n)^{1-1/q}/\det(H_n)$. Markov's inequality gives

$$B_n = O_p((\ell_n/n)^{1-1/q}/\det(H_n)).$$

Hence, $\sup_{x\in\mathcal{C}} |F_n(x) - S_n(x)| \le B_n = O_p\Big(\frac{(\log n/n)^{1-1/q}}{\det(H_n)}\Big)$. By the triangle inequality

$$\sup_{x\in\mathcal{C}} |F_n(x)| \le \sup_{x\in\mathcal{C}} |S_n(x)| + \sup_{x\in\mathcal{C}} |F_n(x) - S_n(x)|.$$

Step 2: An r-net $\mathcal{G} \subset \mathcal{C}$ is such that for every $x \in \mathcal{C}$, there exists $x^\circ \in \mathcal{G}$ with $\|x - x^\circ\| \le r$. Because $\mathcal{C}$ is compact, we can choose $M > 0$ such that $\mathcal{C} \subset (-M, M)^D \subset [-M, M]^D$. Set

$$m = \left\lceil \frac{2M\sqrt{D}}{r} \right\rceil, \qquad \ell = \frac{2M}{m}.$$

Each coordinate interval $[-M, M]$ has length $2M$, so dividing it into m equal pieces gives intervals of length ℓ. Since $m \ge \frac{2M\sqrt{D}}{r}$, we have $\ell = \frac{2M}{m} \le \frac{r}{\sqrt{D}}$. More precisely, let

$$t_j = -M + j\ell, \qquad j = 0, \dots, m,$$

and use the intervals

$$I_j = [t_{j-1}, t_j) \quad (j < m), \qquad I_m = [t_{m-1}, t_m].$$

This convention assigns every boundary point to exactly one interval. The Cartesian products

$$Q_{j_1,\dots,j_D} = I_{j_1} \times \cdots \times I_{j_D}, \qquad j_d \in \{1, \dots, m\},$$

partition $[-M, M]^D$ into m^D cells. If $x, y \in Q_{j_1,\dots,j_D}$, then their dth coordinates belong to the same interval of length ℓ. Thus

$$|x_d - y_d| \le \ell, \qquad d = 1, \dots, D.$$

Consequently,

$$\|x - y\|^2 = \sum_{d=1}^{D} |x_d - y_d|^2 \le D\ell^2 \le D\left(\frac{r}{\sqrt{D}}\right)^2 = r^2.$$

Taking square roots gives $\|x - y\| \le r$. For each cell Q satisfying $Q \cap \mathcal{C} \ne \varnothing$, select $g_Q \in Q \cap \mathcal{C}$, and define $\mathcal{G} = \{g_Q : Q \cap \mathcal{C} \ne \varnothing\}$. The collection is finite, and $\mathcal{G} \subset \mathcal{C}$. Now take any $x \in \mathcal{C}$. It belongs to a cell Q, which necessarily intersects $\mathcal{C}$. Hence that cell has a selected representative g_Q. Both points lie in Q, so the preceding calculation gives

$$\|x - g_Q\| \le r.$$

Equivalently,

$$\mathcal{C} \subseteq \bigcup_{g \in \mathcal{G}} \overline{B}(g, r),$$

which is precisely the r-net property. The construction also gives a bound on the cardinality of $\mathcal{G}$

$$\#\mathcal{G} \leq m^D \leq \left(1 + \frac{2M\sqrt{D}}{r}\right)^D \leq C_0(1 + r^{-1})^D, \qquad C_0 = \max\{1, 2M\sqrt{D}\}^D.$$

Now, let $r := r_n = \underline{h}_n(\det(H_n)/n)^{1/\alpha}$ and choose an r_n-net $\mathcal{G}_n \subset \mathcal{C}$, with $\#\mathcal{G}_n \leq C(1 + r_n^{-1})^D$.

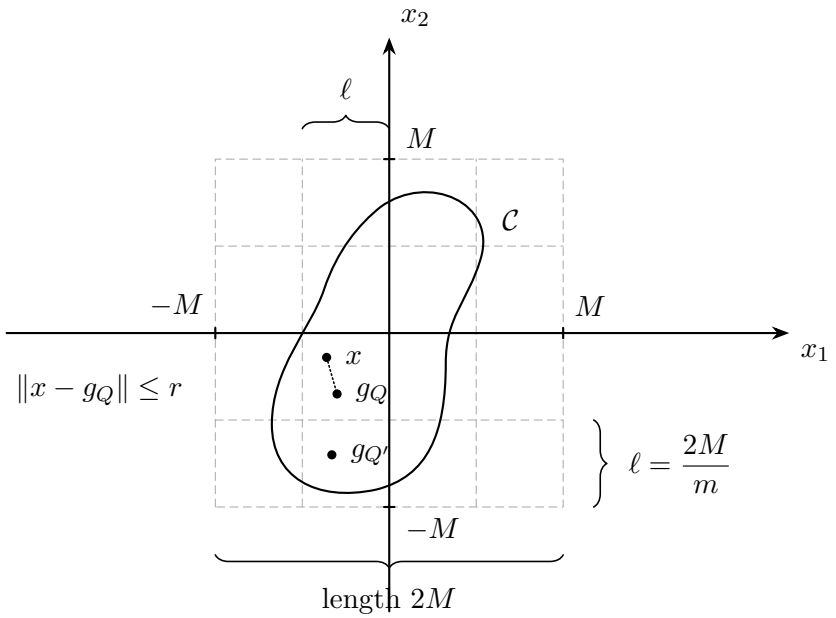


Figure 1: Constructing a grid associated with a r-net $\mathcal{G}$

Since we assume that

$$\frac{\log n}{n \det(H_n)\underline{h}_n^2} \longrightarrow 0$$

and $h_{d,n} \to 0$ for all d, for sufficiently large n, $\log n \geq 1$ and

$$\det(H_n)\underline{h}_n^2 \geq \frac{\log n}{n} \geq \frac{1}{n}.$$

Hence,

$$\frac{1}{n} \leq \det(H_n)\underline{h}_n^2 \leq \det(H_n) \quad \Longrightarrow \quad \frac{1}{\det(H_n)} \leq n,$$

and

$$\frac{1}{n} \leq \det(H_n)\underline{h}_n^2 \leq \underline{h}_n^2 \quad \Longrightarrow \quad \underline{h}_n^{-1} \leq n^{1/2}.$$

Finally, these bounds show that

$$r_n^{-1} = \underline{h}_n^{-1} \left(\frac{n}{\det(H_n)} \right)^{1/\alpha} \leq n^{1/2} (n^2)^{1/\alpha} = n^{1/2+2/\alpha}.$$

Therefore, $\#\mathcal{G}_n \leq C(1 + r_n^{-1})^D \leq C n^{D(1/2+2/\alpha)}$, and if $A := D\left(\frac{1}{2} + \frac{2}{\alpha}\right) < \infty$

$$\#\mathcal{G}_n \leq C n^A. \tag{17}$$

Now, recall that the truncated sum is

$$S_n(x) := \frac{1}{n \det(H_n)} \sum_{i=1}^{n} \left[Z_i^{(n)} L\big(H_n^{-1}(X_i - x)\big) - \mathbb{E}\Big\{ Z_i^{(n)} L\big(H_n^{-1}(X_i - x)\big) \Big\} \right],$$

where $Z_i^{(n)} = Z_i \mathbf{1}_{\{|Z_i| \leq \tau_n\}}$. For each $x \in \mathcal{C}$, choose $x^\circ \in \mathcal{G}_n$ with $\|x - x^\circ\| \leq r_n$, and put

$$\Delta_i(x) := L\big(H_n^{-1}(X_i - x)\big) - L\big(H_n^{-1}(X_i - x^\circ)\big).$$

Now, by definition,

$$\|H_n^{-1}\| := \sup_{\|z\|=1} \|H_n^{-1} z\| \leq \frac{1}{\underline{h}_n}.$$

Choose an index j with $h_{j,n} = \underline{h}_n$, and take the unit coordinate vector e_j. Then $\|H_n^{-1} e_j\| = \left\| \frac{e_j}{h_{j,n}} \right\| = \frac{1}{\underline{h}_n}$. Thus $\|H_n^{-1}\| = \max_d \frac{1}{h_{d,n}} = \frac{1}{\underline{h}_n}$. By Hölder continuity, $\|H_n^{-1}\| = 1/\underline{h}_n$ and the choice of r_n

$$|\Delta_i(x)| \leq C \left\| H_n^{-1}(x^\circ - x) \right\|^\alpha \leq C \left(\frac{r_n}{\underline{h}_n} \right)^\alpha = C \frac{\det(H_n)}{n}.$$

This bound holds for every i, every observation X_i, and every $x \in \mathcal{C}$. Now subtract the centered sums:

$$S_n(x) - S_n(x^\circ) = \frac{1}{n \det(H_n)} \sum_{i=1}^{n} \left[Z_i^{(n)} \Delta_i(x) - \mathbb{E}\Big\{ Z_i^{(n)} \Delta_i(x) \Big\} \right].$$

Using the triangle inequality, $|\mathbb{E}V| \leq \mathbb{E}|V|$, and $|Z_i^{(n)}| \leq |Z_i|$, we obtain

$$\begin{aligned}
|S_n(x) - S_n(x^\circ)| &\leq \frac{1}{n \det(H_n)} \sum_{i=1}^{n} \left[|Z_i^{(n)}| \, |\Delta_i(x)| + \mathbb{E}\Big\{ |Z_i^{(n)}| \, |\Delta_i(x)| \Big\} \right] \\
&\leq \frac{1}{n \det(H_n)} \frac{C \det(H_n)}{n} \left[\sum_{i=1}^{n} |Z_i| + n\mathbb{E}|Z| \right] \\
&= \frac{C}{n} \left[\frac{1}{n} \sum_{i=1}^{n} |Z_i| + \mathbb{E}|Z| \right].
\end{aligned}$$

The bound is uniform in x because its right-hand side does not depend on x. Thus it also bounds

$$R_n := \sup_{x \in \mathcal{C}} |S_n(x) - S_n(x^\circ)|.$$

Set

$$B_n := \frac{1}{n} \sum_{i=1}^{n} |Z_i| + \mathbb{E}|Z|, \qquad \mu := \mathbb{E}|Z| < \infty.$$

The moment assumption guarantees finiteness because

$$\mathbb{E}|Z| \leq \left(\mathbb{E}|Z|^q\right)^{1/q} < \infty.$$

Since the Z_i are identically distributed,

$$\mathbb{E}B_n = 2\mu.$$

Now, $R_n \leq \frac{C}{n} B_n$. If $R_n > M/n$, then its upper bound must also exceed M/n:

$$R_n > \frac{M}{n} \quad \Longrightarrow \quad \frac{CB_n}{n} > \frac{M}{n} \quad \Longrightarrow \quad CB_n > M.$$

Hence

$$\{R_n > M/n\} \subset \{CB_n > M\},$$

and taking probabilities gives

$$\mathbb{P}(R_n > M/n) \leq \mathbb{P}(CB_n > M).$$

Since

$$\mathbb{E}(CB_n) = C\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|Z_i| + \mu\right) = C(\mu + \mu) = 2C\mu.$$

By Markov's inequality,

$$\mathbb{P}(CB_n > M) \leq \frac{2C\mu}{M}.$$

Hence,

$$\mathbb{P}\left(R_n > \frac{M}{n}\right) \leq \mathbb{P}(CB_n > M) \leq \frac{2C\mu}{M} \text{ and } \sup_{x \in \mathcal{C}} |S_n(x) - S_n(x^\circ)| = O_p(n^{-1}).$$

Then,

$$\begin{aligned}\sup_{x \in \mathcal{C}} |F_n(x)| &\leq \sup_{x \in \mathcal{C}} |F_n(x) - S_n(x)| + \sup_{x \in \mathcal{C}} |S_n(x) - S_n(x^\circ)| + \max_{x^\circ \in \mathcal{G}_n} |S_n(x^\circ)| \\ &= O_p\left(\frac{(\log n/n)^{1-1/q}}{\det(H_n)}\right) + O_p(n^{-1}) + \max_{x^\circ \in \mathcal{G}_n} |S_n(x^\circ)|\end{aligned}$$

Step 3: Recall that $Z_i^{(n)} = Z_i \mathbf{1}_{\{|Z_i| \le \tau_n\}}$, and now define $W_i = A_i - \mathbb{E}A_i$, where $A_i = Z_i^{(n)} L(H_n^{-1}(X_i - x^\circ))$. First, note that since $|Z_i^{(n)}| \le \tau_n$, we have

$$|A_i| \le \tau_n \|L\|_\infty, \qquad |\mathbb{E}A_i| \le \mathbb{E}|A_i| \le \tau_n \|L\|_\infty.$$

Therefore, $|W_i| = |A_i - \mathbb{E}A_i| \le |A_i| + |\mathbb{E}A_i| \le 2\tau_n \|L\|_\infty$. Because W_i is centered its variance is

$$\mathbb{E}W_i^2 = \mathrm{Var}(A_i) = \mathbb{E}A_i^2 - (\mathbb{E}A_i)^2 \le \mathbb{E}A_i^2 \le \mathbb{E}\big[Z_i^2 L^2\big(H_n^{-1}(X_i - x^\circ)\big)\big],$$

where the last step uses $(Z_i^{(n)})^2 \le Z_i^2$. To bound this expectation, we use

$$\mathbb{E}(|Z_i|^q \mid X_i) \le C_q, \qquad q > 2.$$

Since $t \mapsto t^{2/q}$ is concave, conditional Jensen's inequality gives

$$\mathbb{E}(Z_i^2 \mid X_i) = \mathbb{E}\big[(|Z_i|^q)^{2/q} \mid X_i\big] \le [\mathbb{E}(|Z_i|^q \mid X_i)]^{2/q} \le C_q^{2/q}.$$

Thus the conditional second moment is bounded by a constant independent of X_i. Conditioning on X_i

$$\begin{aligned}\mathbb{E}\big[Z_i^2 L^2\big(H_n^{-1}(X_i - x^\circ)\big)\big] &= \mathbb{E}\big[\mathbb{E}(Z_i^2 \mid X_i) L^2\big(H_n^{-1}(X_i - x^\circ)\big)\big] \\ &\le C_q^{2/q} \int_{\mathbb{R}^D} L^2\big(H_n^{-1}(z - x^\circ)\big)\, f_X(z)\, dz.\end{aligned}$$

Now by a change of variables $z = x^\circ + H_n u,$, $dz = \det(H_n)\, du$. The integral becomes

$$\det(H_n) \int_{\mathbb{R}^D} L^2(u) f_X(x^\circ + H_n u)\, du.$$

To bound the density here, let $\mathrm{supp}(L) \subseteq \{u : \|u\| \le B_L\}$. Whenever $L(u) \ne 0$,

$$\|x^\circ + H_n u - x^\circ\| \le \overline{h}_n B_L \longrightarrow 0.$$

Since $x^\circ \in \mathcal{C}$, these points lie in the fixed neighborhood where f_X is bounded, for all sufficiently large n. Writing that density bound as M_f, we get

$$\mathbb{E}W_i^2 \le C_q^{2/q} M_f \det(H_n) \int_{\mathbb{R}^D} L^2(u)\, du = C' \det(H_n).$$

The integral is finite because L is bounded and compactly supported. The constant

$$C' = C_q^{2/q} M_f \int L^2(u)\, du$$

is independent of n, i, and x°. Recall that by Bernstein's inequality if $W_1, \ldots, W_n$ are independent, centered real random variables with $|W_i| \le b$ and $\sum_i \mathbb{E}W_i^2 \le s^2$, then for $z > 0$,

$$\mathbb{P}\left(|\sum_i W_i| > z\right) \le 2\exp\left\{-\frac{z^2}{2(s^2 + bz/3)}\right\}.$$

Now for $z = \sqrt{2s^2t} + \frac{2bt}{3}$ we have

$$z^2 - 2t(s^2 + bz/3) = \left(\sqrt{2s^2t} + \frac{2bt}{3}\right)^2 - 2s^2t - \frac{2bt}{3}\left(\sqrt{2s^2t} + \frac{2bt}{3}\right) = \frac{2bt}{3}\sqrt{2s^2t} \geq 0.$$

Therefore,

$$\frac{z^2}{2(s^2 + bz/3)} \geq t \quad \Longrightarrow \quad \exp\left\{-\frac{z^2}{2(s^2 + bz/3)}\right\} \leq e^{-t}.$$

Hence, we can write

$$\mathbb{P}\left(|\sum_i W_i| > \sqrt{2s^2t} + \frac{2bt}{3}\right) \leq 2\exp\{-t\}.$$

Now, letting $b = 2\|L\|_\infty \tau_n := C_b\tau_n$ and $s^2 = C'n\det(H_n)$ we have for any $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n W_i(x^\circ)\right| > \sqrt{2C'n\det(H_n)t} + \frac{2C_b\tau_n t}{3}\right) \leq 2e^{-t}.$$

Dividing the terms inside the probability by $n\det(H_n)$, and writing

$$S_n(x^\circ) := T_{n,L,Z^{(n)}}(x^\circ) - \mathbb{E}T_{n,L,Z^{(n)}}(x^\circ),$$

gives

$$\mathbb{P}\left(|S_n(x^\circ)| > \sqrt{\frac{2C't}{n\det(H_n)}} + \frac{2C_b\tau_n t}{3n\det(H_n)}\right) \leq 2e^{-t}.$$

Now choose $t = (A+2)\ell_n$. Since A is fixed, its contribution can be absorbed into a constant C. Thus

$$\mathbb{P}\left(|S_n(x^\circ)| > C\left\{\sqrt{\frac{\ell_n}{n\det(H_n)}} + \frac{\tau_n\ell_n}{n\det(H_n)}\right\}\right) \leq 2n^{-(A+2)}.$$

By Lemma 11 in the Appendix of Chapter 1,

$$\begin{aligned}\mathbb{P}&\left(\max_{x^\circ\in\mathcal{G}_n}|S_n(x^\circ)| > C\left\{\sqrt{\frac{\ell_n}{n\det(H_n)}} + \frac{\tau_n\ell_n}{n\det(H_n)}\right\}\right)\\ &\leq \sum_{x^\circ\in\mathcal{G}_n} 2n^{-(A+2)}\\ &\leq Cn^A\cdot 2n^{-(A+2)} = \frac{2C}{n^2} \longrightarrow 0.\end{aligned}$$

Now,

$$\frac{\tau_n\ell_n}{n} = \left(\frac{n}{\ell_n}\right)^{1/q}\frac{\ell_n}{n} = \left(\frac{\ell_n}{n}\right)^{1-1/q}.$$

Hence the expression in braces equals

$$a_n = \sqrt{\frac{\ell_n}{n\det(H_n)}} + \frac{(\ell_n/n)^{1-1/q}}{\det(H_n)},$$

and we have proved

$$\mathbb{P}\left(\max_{x^\circ\in\mathcal{G}_n} |S_n(x^\circ)| \le Ca_n\right) \ge 1 - \frac{2C}{n^2} \longrightarrow 1.$$

Finally,

$$\begin{aligned}
\sup_{x\in\mathcal{C}} |F_n(x)| &\le \sup_{x\in\mathcal{C}} |F_n(x) - S_n(x)| + \sup_{x\in\mathcal{C}} |S_n(x) - S_n(x^\circ)| + \max_{x^\circ\in\mathcal{G}_n} |S_n(x^\circ)| \\
&= O_p\left(\frac{(\log n/n)^{1-1/q}}{\det(H_n)}\right) + O_p(n^{-1}) + \max_{x^\circ\in\mathcal{G}_n} |S_n(x^\circ)| \\
&= O_p\left(\frac{(\log n/n)^{1-1/q}}{\det(H_n)}\right) + O_p(n^{-1}) + O_p\left(\sqrt{\frac{\log n}{n\det(H_n)}} + \frac{(\log n/n)^{1-1/q}}{\det(H_n)}\right) = O_p(a_n).
\end{aligned}$$

□

Remark 10. *1. Suppose $|Z| \le M$ almost surely, where $M < \infty$. Define*

$$W_i(x) = Z_i L\big(H_n^{-1}(X_i - x)\big) - \mathbb{E}\big[Z_i L\big(H_n^{-1}(X_i - x)\big)\big].$$

Then $|W_i(x)| \le 2M\|L\|_\infty$,, $\sum_{i=1}^n \mathbb{E}W_i(x)^2 \le Cn\det(H_n)$. Applying Bernstein's inequality gives

$$\sup_{x\in\mathcal{C}} |T_{n,L,Z}(x) - \mathbb{E}T_{n,L,Z}(x)| = O_p\left(\sqrt{\frac{\log n}{n\det(H_n)}} + \frac{\log n}{n\det(H_n)} + \frac{1}{n}\right).$$

Now put

$$s_n = \sqrt{\frac{\log n}{n\det(H_n)}}.$$

If $n\det(H_n)/\log n \to \infty$, then $s_n \to 0$, and

$$\frac{\log n}{n\det(H_n)} = s_n^2 = o(s_n).$$

Also, $1/n = o(s_n)$. Consequently, $\sup_{x\in\mathcal{C}} |T_{n,L,Z}(x) - \mathbb{E}T_{n,L,Z}(x)| = O_p(s_n)$.

2. Note that in the order of the bound we obtained the ratio of the second term to the first is

$$\frac{(\log n/n)^{1-1/q}/\det(H_n)}{\sqrt{\log n/(n\det(H_n))}} = \frac{(\log n/n)^{1-1/q}\det(H_n)^{-1}}{(\log n/n)^{1/2}\det(H_n)^{-1/2}} = \left(\frac{\log n}{n}\right)^{1-1/q-1/2} \det(H_n)^{-1+1/2}$$

$$= \frac{(\log n/n)^{1/2-1/q}}{\sqrt{\det(H_n)}} = \sqrt{\frac{(\log n/n)^{1-2/q}}{\det(H_n)}}.$$

Consequently, if $\left(\frac{\log n}{n}\right)^{1-2/q} = O\big(\det(H_n)\big)$ *then*

$$\sup_{x\in\mathcal{C}} |T_{n,L,Z}(x) - \mathbb{E}T_{n,L,Z}(x)| = O_p\left(\sqrt{\frac{\log n}{n\det(H_n)}}\right).$$

5.3 Uniform invertibility and the Taylor remainder

Use $u_{ni}(x) = H_n^{-1}(X_i - x)$, $r(u) = (1, u^T)^T$, and

$$D_n = \operatorname{diag}(1, H_n), \qquad A_n(x) = \frac{1}{n\det(H_n)} \sum_{i=1}^{n} K(u_{ni}(x)) r(u_{ni}(x)) r(u_{ni}(x))^T.$$

Here D_n is different from G_n. The exact relation to the matrix $S(x)$ previously defined is

$$A_n(x) = D_n^{-1} S(x) D_n^{-1}, \qquad R(x)^T P(x) R(x) = n\det(H_n) D_n A_n(x) D_n. \tag{18}$$

The kernel moment matrix is

$$M_K = \int K(u) r(u) r(u)^T \, du = \begin{pmatrix} 1 & 0^T \\ 0 & \mu_2 I_D \end{pmatrix}. \tag{19}$$

Coordinatewise symmetry supplies the zero first and mixed second moments.

Lemma 5. *Under* ***A2–A3****,* $\overline{h}_n \to 0$*, and* (9)*,*

$$\sup_{x\in\mathcal{C}} \|A_n(x) - f_X(x) M_K\| = o_p(1). \tag{20}$$

Let $c_* = f_* \min(1, \mu_2) > 0$ *and* $\mathcal{E}_n = \{\inf_{x\in\mathcal{C}} \lambda_{\min}(A_n(x)) \geq c_*/2\}$*. Then* $\mathbb{P}(\mathcal{E}_n) \to 1$ *and, on* $\mathcal{E}_n$*,*

$$\sup_{x\in\mathcal{C}} \left\|A_n(x)^{-1}\right\| \leq 2/c_*.$$

Also,

$$\sup_{x\in\mathcal{C}} \frac{1}{n\det(H_n)} \sum_{i=1}^{n} |K(u_{ni}(x))| = O_p(1). \tag{21}$$

Proof. The functions $K(u)$, $u_j K(u)$, $u_j u_k K(u)$ and $|K(u)|$ are bounded, compactly supported and globally α-Hölder. Consider $F(u) = p(u)K(u)$, $p(u) \in \{1, u_j, u_j u_k\}$. Since $F(u) = 0$ wherever $K(u) = 0$,

$$\operatorname{supp}(F) \subseteq \operatorname{supp}(K),$$

so F has compact support. Define

$$P = \sup_{\|z\|\leq B+1} |p(z)|, \qquad G = \sup_{\|z\|\leq B+1} \|\nabla p(z)\|.$$

Both P and G are finite because p and its derivatives are continuous on a compact ball. Also,

$$\|F\|_\infty \le P\|K\|_\infty.$$

To establish global Hölder continuity, consider two cases. 1. Nearby points: $0 < \|u-v\| \le 1$. If $K(u) = K(v) = 0$, then $F(u) = F(v) = 0$, and the desired inequality is immediate. Otherwise, at least one point lies in $\operatorname{supp}(K)$. Because the points are within distance one, both belong to the ball of radius $B + 1$. The segment joining them also lies in this ball, so the mean value inequality gives

$$|p(u) - p(v)| \le G\|u - v\|.$$

Now write

$$p(u)K(u) - p(v)K(v) = p(u)\big(K(u) - K(v)\big) + K(v)\big(p(u) - p(v)\big).$$

Consequently,

$$\begin{aligned} |F(u) - F(v)| &\le PL_K\|u - v\|^\alpha + \|K\|_\infty G\|u - v\| \\ &\le (PL_K + \|K\|_\infty G)\|u - v\|^\alpha. \end{aligned}$$

The last step uses

$$d \le d^\alpha \qquad \text{when } 0 \le d \le 1, \quad 0 < \alpha \le 1.$$

2. Distant points: $\|u - v\| > 1$. Boundedness alone gives

$$|F(u) - F(v)| \le |F(u)| + |F(v)| \le 2P\|K\|_\infty \le 2P\|K\|_\infty\|u - v\|^\alpha.$$

Combining the cases,

$$|F(u) - F(v)| \le C_p\|u - v\|^\alpha \quad \text{for all } u, v \in \mathbb{R}^D,$$

where

$$C_p = \max\{PL_K + \|K\|_\infty G,\ 2P\|K\|_\infty\} < \infty.$$

Finally, $|K|$ has the same support and supremum bound as K. The reverse triangle inequality gives

$$\big||K(u)| - |K(v)|\big| \le |K(u) - K(v)| \le L_K\|u - v\|^\alpha,$$

so $|K|$ is also globally α-Hölder.

Lemma 4 with $Z = 1$ therefore gives

$$\sup_{x\in\mathcal{C}} \|A_n(x) - \mathbb{E}A_n(x)\| = O_p\left(\sqrt{\frac{\log n}{n \det(H_n)}}\right) = o_p(1).$$

Now, since

$$\mathbb{E}A_n(x) = \int K(u)r(u)r(u)^T f_X(x + H_n u)\, du.$$

$$\mathbb{E}A_n(x) - f_X(x)M_K = \int K(u)r(u)r(u)^T \big[f_X(x + H_n u) - f_X(x)\big]\, du.$$

By the triangle inequality for matrix-valued integrals,

$$\begin{aligned}\|\mathbb{E}A_n(x) - f_X(x)M_K\| &\le \int |K(u)|\, \|r(u)r(u)^T\|\, |f_X(x + H_n u) - f_X(x)|\, du \\ &= \int |K(u)|\, \|r(u)\|^2\, |f_X(x + H_n u) - f_X(x)|\, du,\end{aligned}$$

where we used the identity $\|rr^T\| = \|r\|^2$ for the Euclidean norm. Now, whenever $K(u) \neq 0$, we have $\|u\| \le B$. Therefore,

$$\|(x + H_n u) - x\| = \|H_n u\| \le \|H_n\|\|u\| \le B\bar{h}_n.$$

Consequently, for every $x \in \mathcal{C}$ and every contributing u,

$$|f_X(x + H_n u) - f_X(x)| \le \sup_{\substack{y \in \mathcal{C} \\ \|z-y\| \le B\bar{h}_n}} |f_X(z) - f_X(y)|.$$

Hence, uniform continuity of f_X on $\mathcal{C}_\rho$ implies

$$\begin{aligned}&\sup_{x \in \mathcal{C}} \|\mathbb{E}A_n(x) - f_X(x)M_K\| \\ &\quad \le \sup_{\substack{x \in \mathcal{C} \\ \|z-x\| \le B\bar{h}_n}} |f_X(z) - f_X(x)| \int |K(u)|\, \|r(u)\|^2\, du \longrightarrow 0.\end{aligned}$$

This proves (20). Now,

Recall that

$$M_K = \mathrm{diag}(1, \mu_2, \ldots, \mu_2), \qquad c_* := f_* \min(1, \mu_2), \qquad f_* = \inf_{x \in \mathcal{C}} f_X(x) > 0.$$

For a unit vector $a = (a_0, \ldots, a_D)^T$,

$$\begin{aligned}a^T f_X(x) M_K a &= f_X(x) \left(a_0^2 + \mu_2 \sum_{d=1}^{D} a_d^2 \right) \\ &\ge f_* \min(1, \mu_2) \sum_{d=0}^{D} a_d^2 = c_*.\end{aligned}$$

Also, for any matrix B, Cauchy-Schwarz and the definition of the operator norm give

$$|a^T B a| \le \|a\|\, \|Ba\| \le \|B\|\, \|a\|^2 = \|B\|.$$

In particular,

$$a^T\big(A_n(x) - f_X(x)M_K\big)a \geq -\|A_n(x) - f_X(x)M_K\|.$$

Adding the two bounds,

$$\begin{aligned} a^T A_n(x)a &= a^T f_X(x)M_K a + a^T\big(A_n(x) - f_X(x)M_K\big)a \\ &\geq c_* - \|A_n(x) - f_X(x)M_K\|. \end{aligned}$$

Then, since $A_n(x)$ is symmetric, its smallest eigenvalue is the minimum of its quadratic form over unit vectors, i.e.,

$$\lambda_{\min}(A_n(x)) = \min_{\|a\|=1} a^T A_n(x)a.$$

Hence,

$$\lambda_{\min}(A_n(x)) \geq c_* - \|A_n(x) - f_X(x)M_K\|.$$

Now define the uniform error

$$\Delta_n := \sup_{x\in\mathcal{C}} \|A_n(x) - f_X(x)M_K\|.$$

We have already shown that $\Delta_n = o_p(1)$. Because $\Delta_n \geq 0$, this means that for every fixed $\varepsilon > 0$,

$$\mathbb{P}(\Delta_n > \varepsilon) \longrightarrow 0.$$

Setting $E_n := \{\Delta_n \leq c_*/2\}$ and taking the fixed threshold $\varepsilon = c_*/2 > 0$,

$$\mathbb{P}(E_n) = 1 - \mathbb{P}(\Delta_n > c_*/2) \longrightarrow 1.$$

In particular,

$$\mathbb{P}\left(\inf_{x\in\mathcal{C}} \lambda_{\min}(A_n(x)) \geq c_*/2\right) \geq \mathbb{P}(E_n) \longrightarrow 1.$$

This proves the uniform eigenvalue bound with probability tending to one. On that same event, every eigenvalue of every $A_n(x)$ is positive, so all these matrices are invertible. The eigenvalues of $A_n(x)^{-1}$ are the reciprocals of those of $A_n(x)$. For a symmetric positive-definite matrix, this implies

$$\|A_n(x)^{-1}\| = \max_j \frac{1}{\lambda_j(A_n(x))} = \frac{1}{\lambda_{\min}(A_n(x))} \leq \frac{2}{c_*}.$$

The bound holds simultaneously for all $x \in \mathcal{C}$, hence

$$\sup_{x\in\mathcal{C}} \|A_n(x)^{-1}\| \leq \frac{2}{c_*}$$

on an event whose probability tends to one.

Finally apply Lemma 4 with $Z = 1$, $L = |K|$; the expectations of these kernel sums are bounded by $\sup_{\mathcal{C}_\rho} f_X \int |K|$. This proves (21). □

Lemma 6. *Under* ***A1–A3****,* $\overline{h}_n \to 0$*, and* (9)*, define*

$$\varepsilon_{ni}(x) = m(X_i) - m(x) - \nabla m(x)^T (X_i - x),$$
$$b_n(x) = \frac{1}{n \det(H_n)} \sum_{i=1}^{n} K(u_{ni}(x)) r(u_{ni}(x)) \varepsilon_{ni}(x),$$
$$w_n(x) = \frac{1}{n \det(H_n)} \sum_{i=1}^{n} K(u_{ni}(x)) r(u_{ni}(x)) U_i.$$

Then

$$\sup_{x \in \mathcal{C}} \|b_n(x)\| = O_p(\overline{h}_n^2), \qquad \sup_{x \in \mathcal{C}} \|w_n(x)\| = O_p(a_n). \tag{22}$$

Proof. For sufficiently large n, $B\overline{h}_n < \rho$. If $K(u_{ni}(x)) \neq 0$, the segment joining x to X_i lies in $\mathcal{C}_\rho$. Taylor's theorem, as used in Remark 9.4 on page 23 of the notes, gives

$$|\varepsilon_{ni}(x)| \leq \frac{1}{2} \sup_{z \in \mathcal{C}_\rho} \left\|\nabla^2 m(z)\right\| \|X_i - x\|^2 \leq C\overline{h}_n^2.$$

Also $\|r(u_{ni}(x))\| \leq \sqrt{1 + B^2}$ on the kernel support, since by definition,

$$r(u) = \begin{pmatrix} 1 \\ u \end{pmatrix},$$

so its Euclidean norm satisfies

$$\|r(u)\|^2 = 1 + \sum_{d=1}^{D} u_d^2 = 1 + \|u\|^2.$$

The kernel is supported on $\{u : \|u\| \leq B\}$. Thus, whenever $u_{ni}(x) = H_n^{-1}(X_i - x)$ lies in the kernel support, we have $\|u_{ni}(x)\| \leq B$. Consequently,

$$\|r(u_{ni}(x))\| = \sqrt{1 + \|u_{ni}(x)\|^2} \leq \sqrt{1 + B^2}.$$

Consequently (21) gives the first assertion in (22). For the second assertion apply Lemma 4 with $Z = U$ and $L \in \{K, u_1 K, \ldots, u_D K\}$. Every expectation is zero because $\mathbb{E}(U \mid X) = 0$. □

5.4 Proof of the theorem and interpretation

Proof of Theorem 10. By (18) and Lemma 5, the weighted Gram matrix is positive definite for every $x \in \mathcal{C}$ on $\mathcal{E}_n$, proving (12). On this event the quadratic least-squares objective has a unique minimizer, even if K takes negative values.

The algebra of Lemma 1 on page 16 of the notes, expressed in the symmetric scaling (18), gives the exact identity

$$D_n\{\widehat{\beta}_{1,D}(x) - \beta_{1,D}(x)\} = A_n(x)^{-1}\{b_n(x) + w_n(x)\}. \tag{23}$$

Indeed, the ith row of $R(x)$ is $r(u_{ni}(x))^T D_n$, and $Y_i - r(u_{ni}(x))^T D_n \beta_{1,D}(x) = U_i + \varepsilon_{ni}(x)$. Lemmas 5 and 6 now imply

$$\sup_{x\in\mathcal{C}} \left\| D_n\{\widehat{\beta}_{1,D}(x) - \beta_{1,D}(x)\} \right\| = O_p(\overline{h}_n^2 + a_n).$$

The first coordinate proves (13). Dividing coordinate $d+1$ by $h_{d,n}$ proves (14). Since D is fixed, $\underline{h}_n \le 1$ eventually, and $\overline{h}_n/\underline{h}_n \le C_H$,

$$\sup_{x\in\mathcal{C}} \left\| \widehat{\beta}_{1,D}(x) - \beta_{1,D}(x) \right\| = O_p\left(\frac{\overline{h}_n^2 + a_n}{\underline{h}_n} \right) = O_p(\overline{h}_n + a_n/\underline{h}_n) = o_p(1).$$

All bounds extend to the specified estimator because $\mathbb{P}(\mathcal{E}_n^c) \to 0$. The suprema are measurable: the kernel sums are continuous in x, and on $\mathcal{E}_n$ so is their matrix inverse; a countable dense subset of the compact set suffices. □

Corollary 1 (A common bandwidth)**.** *If $h_{d,n} = h_n$ for all d, sufficient bandwidth conditions are*

$$h_n \to 0, \qquad \frac{n h_n^{D+2}}{\log n} \to \infty, \qquad \frac{n^{1-1/q} h_n^{D+1}}{(\log n)^{1-1/q}} \to \infty. \tag{24}$$

For example, $h_n = n^{-\gamma}$ satisfies these conditions whenever

$$0 < \gamma < \min\left\{ \frac{1}{D+2}, \frac{q-1}{q(D+1)} \right\}.$$

Thus a valid bandwidth exists for every fixed D and every $q > 2$ already allowed by the notes.

Proof. Substitute $\det(H_n) = h_n^D$ and $\underline{h}_n = \overline{h}_n = h_n$ into (9)–(10). Strict positivity of both powers of n in (24) gives the displayed range of γ. □

Corollary 2 (Improved uniform rates)**.** *Assume **A1**–**A3**. Let the bandwidths be such that $\overline{h}_n \to 0$, $\overline{h}_n/\underline{h}_n \le C_H$, and*

$$\frac{\log n}{n \det(H_n) \underline{h}_n^2} \longrightarrow 0.$$

Suppose, in addition, that

$$\left(\frac{\log n}{n} \right)^{1-2/q} = O\big(\det(H_n)\big), \qquad q = 2 + \delta > 2, \tag{25}$$

and put

$$s_n := \sqrt{\frac{\log n}{n \det(H_n)}}.$$

Then the local linear estimator satisfies

$$\sup_{x\in\mathcal{C}} |\widehat{m}_1(x) - m(x)| = O_p\left(\overline{h}_n^2 + \sqrt{\frac{\log n}{n\det(H_n)}}\right), \tag{26}$$

$$\sup_{x\in\mathcal{C}} \left|\widehat{\beta}_{1,d}(x) - \partial_d m(x)\right| = O_p\left(\frac{\overline{h}_n^2}{h_{d,n}} + \sqrt{\frac{\log n}{n\det(H_n)h_{d,n}^2}}\right), \quad d = 1,\dots,D. \tag{27}$$

Consequently, the entire coefficient vector obeys

$$\sup_{x\in\mathcal{C}} \left\|\widehat{\beta}_{1,D}(x) - \beta_{1,D}(x)\right\| = O_p\left(\overline{h}_n + \sqrt{\frac{\log n}{n\det(H_n)\underline{h}_n^2}}\right) = o_p(1). \tag{28}$$

For a common bandwidth $h_{d,n} = h_n$, these bounds become

$$\sup_{x\in\mathcal{C}} |\widehat{m}_1(x) - m(x)| = O_p\left(h_n^2 + \sqrt{\frac{\log n}{nh_n^D}}\right), \tag{29}$$

$$\sup_{x\in\mathcal{C}} \left\|\widehat{\beta}_1(x) - \nabla m(x)\right\| = O_p\left(h_n + \sqrt{\frac{\log n}{nh_n^{D+2}}}\right). \tag{30}$$

In this case, sufficient bandwidth conditions are

$$h_n \to 0, \qquad \frac{nh_n^{D+2}}{\log n} \to \infty, \qquad \left(\frac{\log n}{n}\right)^{1-2/q} = O(h_n^D).$$

For example, $h_n = n^{-\gamma}$ satisfies all three whenever

$$0 < \gamma < \min\left\{\frac{1}{D+2}, \frac{q-2}{qD}\right\}. \tag{31}$$

No stronger error-moment or smoothness assumption is required for these bounds.

Proof. First, (25) implies

$$\frac{(\log n/n)^{1-1/q}/\det(H_n)}{s_n} = \frac{(\log n/n)^{1/2-1/q}}{\sqrt{\det(H_n)}} = \sqrt{\frac{(\log n/n)^{1-2/q}}{\det(H_n)}} = O(1).$$

Thus $a_n = O(s_n)$ in (11). Moreover,

$$\frac{s_n}{\underline{h}_n} = \sqrt{\frac{\log n}{n\det(H_n)\underline{h}_n^2}} \longrightarrow 0.$$

It follows that

$$\frac{(\log n/n)^{1-1/q}}{\det(H_n)\underline{h}_n} = O(s_n/\underline{h}_n) \longrightarrow 0.$$

Hence (10) follows from the corollary's assumptions and need not be imposed separately; all conditions of Theorem 10 hold.

For completeness, the source of the different intercept and slope rates can be seen directly from (23). On the event $\mathcal{E}_n$ of Lemma 5, the normalized inverse is uniformly bounded, while Lemma 6 and $a_n = O(s_n)$ give

$$\sup_{x\in\mathcal{C}} \|b_n(x)\| = O_p(\overline{h}_n^2), \qquad \sup_{x\in\mathcal{C}} \|w_n(x)\| = O_p(s_n).$$

Therefore

$$\sup_{x\in\mathcal{C}} \left\| D_n\{\widehat{\beta}_{1,D}(x) - \beta_{1,D}(x)\} \right\| = O_p(\overline{h}_n^2 + s_n).$$

The first diagonal entry of D_n is one, yielding (26). Its $(d+1)$st diagonal entry is $h_{d,n}$, so dividing that coordinate by $h_{d,n}$ yields (27). Since D is fixed and $\overline{h}_n^2/\underline{h}_n \le C_H \overline{h}_n$, the same bounds imply (28). As in the main theorem, the extension of the estimator outside $\mathcal{E}_n$ does not affect these probability orders because $\mathbb{P}(\mathcal{E}_n^c) \to 0$.

Finally, substitute $\det(H_n) = h_n^D$ and $\underline{h}_n = \overline{h}_n = h_n$ to obtain (29)–(30). For $h_n = n^{-\gamma}$, the two bandwidth ratios are

$$\frac{nh_n^{D+2}}{\log n} = \frac{n^{1-\gamma(D+2)}}{\log n}, \qquad \frac{(\log n/n)^{1-2/q}}{h_n^D} = (\log n)^{1-2/q} n^{\gamma D-(1-2/q)}.$$

Under (31), the first tends to infinity and the second tends to zero, proving the claimed sufficient range. □

Remark 11. *The additional substantive conditions are kernel regularity (8), positivity of the density throughout the chosen compact set, and the two uniform bandwidth conditions. The positive lower bound (7) follows from continuity and compactness; it is not a separate assumption. The existing $q = 2 + \delta$ moment is retained, with no boundedness or exponential-tail assumption on U.*

6 Bandwidth choice

6.1 Preliminaries

As in the case of density estimation, the asymptotic properties of Nadaraya-Watson and local linear polynomial estimators depend on selecting a nonstochastic bandwith sequences h_n (or H_n) that converge to zero at specified speeds. Consequently, implementation of these estimators depends crucially on bandwidth selection. We focus initially on the case where $D = 1$.

Following the discussion of density estimation, there are two main criteria for bandwidth selection for an arbitrary nonparametric estimator $\tilde{m}(x;h)$ that depends on a bandwidth h. First, inspired by cross validation in Chapter 1, we consider the minimization of

$$ISE(\tilde{m};h) := \int_{\mathbb{R}} (\tilde{m}(x;h) - m(x))^2 \, dx \tag{32}$$

with respect to h. Second, we consider the minimization of

$$MISE(\tilde{m};h) = \mathbb{E}\left(\int (\tilde{m}(x;h) - m(x))^2 dx \mid \chi \right) \tag{33}$$

with respect to h. Here, χ is the σ-algebra generated by $X_1, \cdots, X_n$.

If w is a function defined on the set where X takes values, such that $\int w(x)dx = 1$, for example, $w = f_X$ the density of X, then we could consider different optimands,

$$WISE(\tilde{m};h) = \int (\tilde{m}(x;h) - m(x))^2 w(x) dx$$

and

$$WMISE(\tilde{m};h) = \mathbb{E}\left(\int (\tilde{m}(x;h) - m(x))^2 \, w(x) dx \mid \chi \right).$$

Note that by Fubini's Theorem

$$\begin{aligned}
MISE(\tilde{m};h) &= \mathbb{E}\left(\int (\tilde{m}(x) - \mathbb{E}(\tilde{m}(x) \mid \chi) + \mathbb{E}(\tilde{m}(x) \mid \chi) - m(x))^2 \, dx \mid \chi \right) \\
&= \mathbb{E}\left(\int (\tilde{m}(x) - \mathbb{E}(\tilde{m}(x) \mid \chi))^2 \, dx + \int (\mathbb{E}(\tilde{m}(x) \mid \chi) - m(x))^2 \, dx \right. \\
&\left. + 2 \int (\tilde{m}(x) - \mathbb{E}(\tilde{m}(x) \mid \chi)) \, (\mathbb{E}(\tilde{m}(x) \mid \chi) - m(x)) \, dx \mid \chi \right) \\
&= \int V(\tilde{m}(x) \mid \chi) dx + \int (\mathbb{E}(\tilde{m}(x) \mid \chi) - m(x))^2 \, dx.
\end{aligned}$$

Consider the an approximation for $ISE(\tilde{m};h)$ given by $\frac{1}{n}\sum_{i=1}^n (\tilde{m}(X_i) - m(X_i))^2$. Since $m(X_i) = \mathbb{E}(Y_i|X_i))$ we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (\tilde{m}(X_i;h) - m(X_i))^2 &= \frac{1}{n}\sum_{i=1}^n (\tilde{m}(X_i;h) - Y_i + (Y_i - \mathbb{E}(Y_i|X_i)))^2 \\
&= \frac{1}{n}\sum_{i=1}^n (Y_i - \tilde{m}(X_i;h))^2 + \frac{1}{n}\sum_{i=1}^n (Y_i - \mathbb{E}(Y_i|X_i))^2 \\
&- \frac{2}{n}\sum_{i=1}^n (Y_i - \tilde{m}(X_i;h))(Y_i - E(Y_i|X_i)).
\end{aligned}$$

Taking conditional expectations on both sides and letting $U_i := Y_i - E(Y_i|X_i)$

$$\mathbb{E}\left(n^{-1}\sum_{i=1}^{n}(\tilde{m}(X_i;h)-m(X_i))^2 \mid \chi\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left((Y_i-\tilde{m}(X_i;h))^2 \mid \chi\right)+\sigma^2$$
$$-\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Y_i-\tilde{m}(X_i;h))U_i) \mid \chi\right].$$

Now, note that in general

$$\begin{aligned}E[(Y_i-\tilde{m}(X_i;h))U_i \mid \chi] &= E[(m_i+U_i-\tilde{m}(X_i;h))U_i \mid \chi]\\ &= \sigma_i^2 - \mathrm{Cov}(\tilde{m}(X_i;h), Y_i \mid \chi) \neq 0.\end{aligned}$$

For example, for linear estimators, such as Nadaraya-Watson or local polynomial estimators,

$$\tilde{m}(X_i;h) = \sum_{j=1}^{n} w_{ij}(\chi;h)Y_j,$$

we have

$$\begin{aligned}\mathbb{E}(Y_jU_i \mid \chi) &= m(X_j)\mathbb{E}(U_i \mid \chi) + \mathbb{E}(U_jU_i \mid \chi)\\ &= \mathbb{E}(U_jU_i \mid \chi).\end{aligned}$$

For $j \neq i$, conditional independence and conditional mean zero imply

$$\mathbb{E}(U_jU_i \mid \chi) = \mathbb{E}(U_j \mid \chi)\mathbb{E}(U_i \mid \chi) = 0.$$

For $j = i$, this expectation is $\mathbb{E}(U_i^2 \mid \chi) = \sigma^2$. Thus

$$\mathbb{E}(Y_jU_i \mid \chi) = \begin{cases}\sigma^2, & j = i,\\ 0, & j \neq i.\end{cases}$$

Consequently, by linearity,

$$\begin{aligned}\mathbb{E}[(Y_i-\widetilde{m}(X_i;h))U_i \mid \chi] &= \mathbb{E}(Y_iU_i \mid \chi) - \sum_{j=1}^{n} w_{ij}\mathbb{E}(Y_jU_i \mid \chi)\\ &= \sigma_i^2 - w_{ii}\sigma_i^2 = (1-w_{ii})\sigma^2.\end{aligned}$$

Now let $\tilde{m}_{-i}(X_i;h)$ be computed without observation i. Conditional on χ, it is independent of U_i. Consequently,

$$\mathbb{E}[\tilde{m}_{-i}(X_i;h)U_i \mid \chi] = 0,$$

and hence

$$\mathbb{E}[(Y_i-\widetilde{m}_{-i}(X_i;h))U_i \mid \chi] = \sigma^2.$$

Then,

$$\mathbb{E}\left(n^{-1}\sum_{i=1}^{n}(\tilde{m}_{-i}(X_i;h)-m(X_i))^2 \mid \chi\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left([Y_i-\tilde{m}_{-i}(X_i;h)]^2 \mid \chi\right)-\sigma^2.$$

$\tilde{m}_{-i}(x;h)$ is called the jackknife estimator associated with $\tilde{m}(x;h)$. In the case of the Nadaraya-Watson estimator $\hat{m}(x)$ with $D=1$ the jackknife estimator is

$$\hat{m}_{-i}(X_i;h) = \frac{\sum_{j=1,j\neq i}^{n} K(h^{-1}(X_j - X_i))Y_j}{\sum_{j=1,j\neq i}^{n} K\left(\frac{X_j - X_i}{h}\right)}$$

and for the local linear estimator the jacknife estimator is,

$$\hat{m}_{1,-i}(X_i;h) = \frac{1}{nh}\sum_{j=1,i\neq j}^{n} \begin{pmatrix} 1 & 0 \end{pmatrix} S_J^{-1}(X_i) \begin{pmatrix} 1 \\ \frac{X_j - X_i}{h} \end{pmatrix} K\left(\frac{X_j - X_i}{h}\right) Y_j$$

where $S_{nJ}(X_i) = \begin{pmatrix} s_0(X_i) & h^{-1}s_1(X_i) \\ h^{-1}s_1(X_i) & h^{-2}s_{(1,1)}(X_i) \end{pmatrix}$ where s_0, s_1 and $s_{(1,1)}$ are defined as in section 3.1 with $D=1$. These developments lead to the following definition.

Definition 5. *For an arbitrary linear nonparametric estimator $\tilde{m}(x;h)$, the cross-validated bandwidth h_{CV} is given by*

$$h_{CV} := \underset{h}{argmin}\, \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{m}_{-i}(X_i;h))^2.$$

For $f, m \in \mathcal{C}_b^2$ and given our previous results on order for conditional bias and variance we have for the local linear estimator

$$\begin{aligned} MISE(\hat{m}_1;h) = &\frac{1}{4}\mu_2^2 h_n^4 \int (m^{(2)}(x))^2 dx \\ &+ \frac{1}{nh_n}\sigma^2 \int K^2(\psi)d\psi \int f_X(x)^{-1}dx + o_p(h_n^4) + o_p((nh_n)^{-1}) \end{aligned}$$

and for the Nadaraya-Watson estimator

$$\begin{aligned} MISE(\hat{m};h) = &h_n^4\mu_2^2 \int \lambda(x)^2 dx + \frac{1}{4}\mu_2^2 h_n^4 \int (m^{(2)}(x))^2 dx \\ &+ \frac{1}{nh_n}\sigma^2 \int K^2(\psi)d\psi \int f_X(x)^{-1}dx + o_p(h_n^4) + o_p((nh_n)^{-1}) \end{aligned}$$

where $\lambda(x) = \frac{m^{(1)}(x)f_X^{(1)}(x)}{f_X(x)}$. Concentrating on terms of order larger than h_n^4 and $(nh_n)^{-1}$ gives an asymptotic approximation for MISE which we label AMISE.

Definition 6. *The plug-in* bandwidth h_{PI} that minimizes the asymptotic mean integrated squared error for $\hat{m}_1(x)$ is given by

$$h_{PI} := \underset{h_n}{\text{argmin}} \left(\frac{1}{4}\mu_2^2 h_n^4 \int (m^{(2)}(x))^2 dx + \frac{1}{nh_n}\sigma^2 \int K^2(u)du \int f_X(x)^{-1}dx \right).$$

Remark 12. *1. It is straightforward to obtain*

$$h_{PI} = n^{-1/5} \left(\frac{\sigma^2 \int K^2(u)du \int f_X(x)^{-1}dx}{\mu_2^2 \int (m^{(2)}(x))^2 dx} \right)^{\frac{1}{5}}.$$

2. Note that if we minimize WMISE with $w(x) = f_X(x)$ we have

$$\begin{aligned} WMISE(\hat{m}_1) = \sigma^2 \frac{1}{nh_n} \int K^2(u)du \int dx \\ + h_n^4 \frac{\mu_2^2}{4} \int (m^{(2)}(x))^2 f_X(x)dx + o_p((nh)^{-1}) + o_p(h_n^4). \end{aligned}$$

In fact, for $\hat{m}_p(x)$ we have

$$\begin{aligned} MISE(\hat{m}_p(x)) = \sigma^2 \frac{1}{nh_n} \int K_p^2(u)du \int dx \\ + h^{2p+2} \left(\frac{\mu_{p+1}}{(p+1)!} \right)^2 \int (m^{(p+1)}(x))^2 f_X(x)dx \\ + o_p((nh)^{-1}) + o_p(h^{2p+2}) \end{aligned}$$

where $\mu_{p+1} = \int u^{p+1} K(u)du$ and $K_p(u) = K_{0,p}(u)$ with

$$K_{s,p}(u) = s! \frac{det(M_{s,p}(u))}{det(N_p)} K(u),$$

where N_p is a $(p+1)\times(p+1)$ matrix having (i,j) entry given by $\int \psi^{i+j-2} K(u)du$ and $M_{s,p}(u) = N_p$ except that the $(s+1)^{th}$ row is given by $(1, u, \ldots, u^p)$. Consequently, taking derivatives we obtain,

$$h_{PI} = \left(\frac{\sigma^2 (p+1)(p!)^2 \int K_p^2(u)du \int dx}{2\mu_{p+1}^2 \int (m^{p+1}(x))^2 f_X(x)dx} \right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}.$$

3. As in the case of density estimation, optimal bandwidth selection via minimization of AMISE precludes asymptotic normality of $(nh_n)^{1/2}(\hat{m}_1(x) - m(x))$. Note that,

$$\begin{aligned} (nh_n)^{1/2}(\hat{m}_1(x) - m(x)) = (nh_n)^{1/2}(\hat{m}_1(x) - E(\hat{m}_1(x)|X_1, \cdots, X_n)) \\ - (nh_n)^{1/2}(m(x) - E(\hat{m}_1(x)|X_1, \cdots, X_n)) \end{aligned}$$

and therefore for asymptotic normality of $(nh_n)^{1/2}(\hat{m}_1(x) - m(x))$, we need the last term to be $o_p(1)$. But given the order of the bias, for this to follow it is necessary that $nh_n^5 \to 0$, whereas with optimal smoothing $nh_n^5 \to C$.

6.2 Implementation and bandwidth estimation

The expression for h_{PI} involves $\int dx$, $\int \left(m^{(2)}(x)\right)^2 f_X(x)dx$ and σ^2, which are all unknown and need to be estimated. We start by discussing estimation procedures for $\int \left(m(x)^{(2)}\right)^2 f_X(x)dx$. We consider the more general case $\theta_{d,\delta} = \int m^{(d)}(x)m^{(\delta)}(x)f_X(x)dx$, where d, δ are positive integers and $d+\delta$ is an even number. A natural choice is to consider $\hat{\theta}_{d,\delta} = \frac{1}{n}\sum_{j=1}^{n} m^{(d)}(X_j)m^{(\delta)}(X_j)$ with $m^{(d)}(x)$ estimated non-parametrically. Hence, consider the estimator

$$\begin{aligned}\left(\begin{array}{cccc} \hat{m}(x) & \hat{m}^{(1)}(x) & \cdots & \frac{\hat{m}^{(p)}(x)}{p!} \end{array}\right) &:= \underset{\beta_0,\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1(X_i - x) - \beta_2(X_i - x)^2\right. \\ &\quad \left. \cdots \quad \beta_p(X_i - x)^p\right)^2 K\left(\frac{X_i - x}{h}\right)\end{aligned}$$

which can be written as

$$\left(\begin{array}{c} \hat{m}(x) \\ \hat{m}^{(1)}(x) \\ \vdots \\ \frac{\hat{m}^{(p)}(x)}{p!} \end{array}\right) = (R_p(x)^T P(x) R_p(x))^{-1} R_p(x)^T P(x) Y$$

where $R_p(x) = (1_n, (X - 1_n x), (X - 1_n x)^2, \cdots (X - 1_n x)^p)$, Y is the vector of observations on the regressand and $1_n^T = (1, \cdots, 1)$. Let e_k be a $p+1$ row vector with zeros for all of its components, except the k^{th} component which is set to 1. Then a natural estimator for the s^{th} derivative of $m(x)$ based on a p-degree polynomial is,

$$\hat{m}^{(s)}(x) = s! e_{s+1}(R_p(x)^T P(x) R_p(x))^{-1} R_p(x)^T P(x) Y$$

where $P(x) = \operatorname{diag}\left\{K\left(\frac{X_t - x}{g_n}\right)\right\}_{t=1}^{n}$ and g_n is a bandwidth. Hence a natural estimator for $\theta_{d,\delta}$ is

$$\hat{\theta}_{d,\delta} = \frac{1}{n}\sum_{j=1}^{n} \hat{m}^{(d)}(X_j)\hat{m}^{(\delta)}(X_j).$$

Since the estimation of $\hat{\theta}_{d,\delta}$ depends on g_n, an expression for g_n that emerges from the optimization of a suitably defined criterion is necessary. If $g_n \to 0$ and $n g_n^{d+\delta+1} \to \infty$ as $n \to \infty$, we have the following approximations

$$\begin{aligned} E(\hat{\theta}_{d,\delta} - \theta_{d,\delta} | X_1, \cdots, X_n) &\approx \frac{1}{(p+1)!}\left(\mu_{p+1}(K_{\delta,p}) g_n^{p-\delta+1}\theta_{d,p+1} + \mu_{p+1}(K_{d,p}) g_n^{p-d+1}\theta_{\delta,p+1}\right) \\ &+ n^{-1} g(n)^{-d-\delta-1}\left(\int K_{d,p}(\psi)K_{\delta,p}(\psi)d\psi\right)\sigma^2 \int dx \end{aligned}$$

where, $\mu_l(K_{d,p}) = \int \psi^l K_{d,p}(\psi)d\psi$, and

$$V(\hat{\theta}_{d,\delta}|X_1, \cdots, X_n) \approx 2n^{-2}g_n^{-2d-2\delta-1}R(K_{d,p} * K_{\delta,p}) \int \sigma^2(\sigma^2 + 2m(x)^2)dx$$
$$+ 4n^{-1}\sigma^2 \int ((m(x)f_X(x))^{(d+\delta)})^2 f_X(x)^{-1}dx$$

where $R(L) = \int L(\psi)^2 d\psi$ for an integrable function L.

These approximations for bias and variance allow the construction of an asymptotic expression for the conditional mean squared error, which we represent by $AMSE$. We are interested, in particular, in the situation where $d = \delta = 2$ and $p = 3$. In this case,

$$AMSE(\hat{\theta}_{2,2}|X) = \left(\frac{1}{12}\mu_4(K_{2,3})g_n^2\theta_{2,4} + n^{-1}g_n^{-5}R(K_{2,3})\sigma^2 \int dx\right)^2$$
$$+ 2\sigma^2 n^{-2}g_n^{-9}R(K_{2,3} * K_{2,3}) \int (\sigma^2 + 2m(x)^2)dx$$
$$+ \frac{4\sigma^2}{n} \int (m(x)f_X(x))^{(4)} f_X(x)^{-1}dx$$

which produces an optimal g_{AMSE}, given by

$$g_{AMSE} = C_2(K)\left(\frac{\sigma^2 \int dx}{|\theta_{2,4}|n}\right)^{\frac{1}{7}}, \text{ where, } C_2(K) = \begin{cases} C_2^I(K), & \theta_{2,4} < 0 \\ C_2^{II}(K), & \theta_{2,4} > 0 \end{cases},$$

with $C_2^I(K) = \left(\frac{12R(K_{2,3})}{\mu_4(K_{2,3})}\right)^{\frac{1}{7}}$, $C_2^{II}(K) = \left(\frac{30R(K_{2,3})}{\mu_4(K_{2,3})}\right)^{\frac{1}{7}}$.[2]

The expression for g_{AMSE} depends on $\theta_{2,4}$ and σ^2, which are not known. The way $\theta_{2,4}$ and σ^2 are estimated determines a variety of *plug in* methods. Note that σ^2 is an unknown in the bandwidth formula for g_{AMSE} and it appears also on the expressions for h_{PI}.

We now turn to the estimation of σ^2. Following Hall and Marron (1990) it can be verified that for a local polynomial estimator of order p, $\hat{m}_p(x;\lambda_n)$ constructed with a bandwidth $\lambda_n$we have that

$$E\left(\sum_{i=1}^{n}(Y_i - \hat{m}_p(X_i;\lambda_n))^2 \mid \chi\right) = \nu\sigma^2,$$

where $\nu = \text{trace}((I - S_X)(I - S_X)^T)$, where S_X is a matrix whose i^{th} row is

$$e_1(R_p^T(X_i)P(X_i)R_p(X_i))^{-1}R_p^T(X_i)P(X_i).$$

This motivates the p^{th} degree variance estimator

$$\hat{\sigma}_p^2(\lambda_n) = (n - \nu)^{-1}\sum_{i=1}^{n}(Y_i - \hat{m}_p(x;\lambda_n))^2.$$

[2]For a Gaussian kernel $C_2^I(K) = (\frac{3}{8\sqrt{\pi}})^{\frac{1}{7}}, C_2^{II}(K) = (\frac{15}{16\sqrt{\pi}})^{\frac{1}{7}}$.

Implementation of this estimator requires a choice for λ_n. Ruppert et al. (1995) propose basing such choice on the conditional $AMSE$ of σ_p^2. They show that, if $\lambda_n \to 0$, and $n\lambda_n \to \infty$ as $n \to \infty$, p is odd and m has p continuous derivatives,

$$E(\hat{\sigma}_p^2(\lambda_n) - \sigma^2|X) \approx \lambda_n^{2p+2}(\frac{\mu_{p+1}(K)}{(p+1)!})^2\theta_{p+1,p+1}$$

and

$$V(\hat{\sigma}_p^2(\lambda_n)|X) \approx n^{-1}\sigma^4 V(\epsilon^2) + 2(n^2\lambda_n)^{-1}\sigma^4 \int dx R(K_p * K_p - 2K_p).$$

It follows that the optimal value of λ_n (in the sense that it minimizes $AMSE$) is,

$$\lambda_{AMSE} = C_3(K)\left(\frac{\sigma^4 \int dx}{\theta_{p+1,p+1}^2 n^2}\right)^{\frac{1}{(4p+5)}},$$

where $C_3(K) = \left(\frac{((p+1)!)^4 R(K_p * K_p - 2K_p)}{2(p+1)\mu_{p+1}(K_p)^4}\right)^{\frac{1}{(4p+5)}}$. If $K(\psi)$ is a standard normal Kernel and $p = 1$, $C_3(K) = \{\frac{4(0.5+2\sqrt{2}-(\frac{4}{3})\sqrt{3}}{\sqrt{2\pi}}\}^{\frac{1}{9}}$.

We now provide some recommendations on how to proceed with estimation:
Step 1 : $m(x) \cong m(0) + m^{(1)}(0)x + \frac{1}{2!}m^{(2)}(0)x^2 + \frac{1}{3!}m^{(3)}(0)x^3 + \frac{1}{4!}m^{(4)}(0)x^4$, therefore let $\beta_0 = m(0)$, $\beta_1 = m^{(1)}(0), \ldots, \beta_4 = m^{(4)}(0)$ and define the following quartic regression,

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 \frac{X_t^2}{2} + \beta_3 \frac{X_t^3}{6} + \beta_4 \frac{X_t^4}{24} + \eta_t, \ t = 1, 2, ..., n$$

Obtain OLS estimators $\hat{\beta}_0, \ldots, \hat{\beta}_4$ and $\hat{\sigma}_Q^2 = (n-5)^{-1}\sum_{t=1}^n \hat{\eta}_t^2$. These could, of course, be used directly on the expression for h_{PI} to produce,

$$h_{ROT} = n^{-\frac{1}{5}}\left(\frac{\hat{\sigma}_Q^2 \int K^2(\psi)d\psi \int dx}{\mu_2^2\left(\frac{1}{n}\sum_{i=1}^n(\hat{\beta}_2 + \hat{\beta}_3 X_i + \hat{\beta}_4 \frac{X_i^2}{2})^2\right)}\right)^{\frac{1}{5}}$$

Step 2 : Use $\hat{\sigma}_Q^2$, $\hat{\beta}_4$, $\hat{\beta}_2$ and the expression for g_{AMSE} to produce

$$\hat{g}_{AMSE} = C_2(K)\left(\frac{\hat{\sigma}_Q^2 \int dx}{n|\frac{1}{n}\sum_{i=1}^n \hat{\beta}_4(\hat{\beta}_2 + \hat{\beta}_3 X_i + \hat{\beta}_4 \frac{X_i^2}{2})|}\right)^{\frac{1}{7}}$$

Now, use $\hat{g}_{AMSE}$ to estimate $\hat{m}^{(2)}(X_j)$, which is then used to produce $\frac{1}{n}\sum_{j=1}^n \hat{m}^{(2)}(X_j)^2$ and $\hat{\lambda}_{AMSE} = C_3(K)\left(\frac{(\hat{\sigma}_Q^2)^2 \int dx}{(n^{-1}\sum_{j=1}^n\{\hat{m}^{(2)}(X_j)\}^2)^2 n^2}\right)^{\frac{1}{9}}$. $\hat{\lambda}_{AMSE}$ can now be used to obtain $\hat{\sigma}_1^2$, which combined with $\frac{1}{n}\sum_{j=1}^n\{\hat{m}^{(2)}(X_j)\}^2$ gives,

$$\hat{h}_{DPI} = n^{-\frac{1}{5}}\left(\frac{\hat{\sigma}_1^2 \int K^2(\psi)d\psi \int dx}{\mu_2^2 n^{-1}\sum_{i=1}^n\{\hat{m}^{(2)}(X_i)\}^2}\right)^{\frac{1}{5}}.$$

Ruppert et al. (1995) have observed that for $p = 1$ we have that

$$g_{AMSE} = C_2(K)C_1(K)^{-\frac{5}{7}}\{\frac{\theta_{22}}{|\theta_{24}|}\}^{\frac{1}{7}}(h_{PI})^{\frac{5}{7}}.$$

By exploring this equation a third estimator for h_n, which we label h_{STE} can be proposed as the value of h that satisfies

$$h_{STE} - n^{-\frac{1}{5}}\{\frac{\hat{\sigma}_1^2 \int K^2(\psi)d\psi \int dx}{\mu_2^2(n^{-1}\sum_{i=1}^n\{\hat{m}^{(2)}(X_i; g_{AMSE}(h_{STE}))\}^2)}\}^{\frac{1}{5}} = 0$$

where,

$$g_{AMSE}(h) = C_2(K)C_1(K)^{-\frac{5}{7}}\left(\frac{n^{-1}\sum_{i=1}^n(\hat{\beta}_2 + \hat{\beta}_3 X_i + \hat{\beta}_4\frac{X_i^2}{2})^2}{|\sum_{i=1}^n(\hat{\beta}_2 + \hat{\beta}_3 X_i + \hat{\beta}_4\frac{X_i^2}{2})n^{-1}\hat{\beta}_4|}\right)^{\frac{1}{7}} h^{\frac{5}{7}}.$$

The motivation here can clearly be traced back to Sheather and Jones (1991) and Park and Marron (1990) in the context of density estimation. h_{ROT}, h_{DPI} and h_{STE} provide state of the art plug in estimators for bandwith in the context of univariate nonparametric regressions. Regarding the properties of these plug in estimators, it can be shown that,

$$\frac{h_{DPI}}{h_{MISE}} - 1 = O_p(n^{-\frac{2}{7}}) \ , \ \frac{h_{STE}}{h_{MISE}} - 1 = O_p(n^{-\frac{2}{7}})$$

where h_{MISE} is the value of h that minimizes $MISE$. This is far superior to the cross-validation choice of h which produces $\frac{h_{CV}}{h_{MISE}} - 1 = O_p(n^{-\frac{1}{10}})$ as shown in Hardle et al. (1988).

Appendix

Lemma 7. *Let $f_n(x) : \mathbb{X} \to \mathbb{R}$ for $n = 1, 2, \ldots$ be measurable. Then $\sup\{f_1, \ldots, f_n\}$, $\inf\{f_1, \ldots, f_n\}$, $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$ and $\liminf_n f_n$ are all measurable functions.*

Proof. Let $h(x) = \sup\{f_1(x), \ldots, f_n(x)\}$. Then, $S_h = \{x \in \mathbb{X} : h(x) > a\} = \cup_{i=1}^n \{x : f_i(x) > a\}$. Consequently, since f_i is measurable, $S_h \in \mathcal{F}$. Similarly if $g(x) = \sup f_n(x)$, $S_g = \{x \in \mathbb{X} : g(x) > a\} = \cup_{i=1}^\infty \{x : f(x) > a\} \in \mathcal{F}$. The same argument can be made for inf. Since $\limsup_{n\to\infty} = \inf_n \sup_{k\geq n} f_k$, lim sup is measurable. The same for $\liminf_{n\to\infty} f_n$. □

Lemma 8. *Let $f : (\Omega, \mathcal{F}) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ be a non-negative measurable function. Then, there exists a sequence $\varphi_n(\omega) : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ of simple non-negative measurable functions such that*

a) $\varphi_n(\omega) \leq \varphi_{n+1}(\omega)$, for all $\omega \in \Omega$, $n = 1, 2, \ldots$

b) $\lim_{n\to\infty} \varphi_n(\omega) = f(\omega)$, for all $\omega \in \Omega$.

Proof. For each $n = 1, 2, \ldots$ define the sets

$$E_{k,n} = \begin{cases} \left\{\omega \in \Omega : \frac{k}{2^n} \leq f(\omega) < \frac{k}{2^n} + \frac{1}{2^n}\right\} & \text{for } k = 0, 1, \ldots, n2^n - 1 \\ \{\omega \in \Omega : f(w) \geq n\} & \text{for } k = n2^n. \end{cases}$$

For each n, the sets $\{E_{k,n} : k = 0, 1, \ldots, n2^n\}$ are disjoint, belong to $\mathcal{F}$ and $\cup_{k=0}^{n2^n} E_{k,n} = \Omega$. Put $\varphi_n(\omega) = \frac{k}{2^n}$ on $E_{k,n}$. Then, by construction $0 \leq \varphi_n(\omega) \leq \varphi_{n+1}(\omega)$, $\forall\omega$ and $n = 1, 2, \ldots$. For any $\omega \in \Omega$, $\varphi_1(\omega), \varphi_2(\omega), \ldots$ is a nondecreasing sequence of real numbers which is bounded above by $f(\omega)$. Hence, $\lim_{n\to\infty} \varphi_n(\omega)$ exists as $f(\omega)$. Since n is a natural number, the function $\varphi_n(\omega)$ takes at most $n2^n$ values. □

Let $x = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T \in \mathbb{R}^n$, $\alpha = \begin{pmatrix} \alpha_1 & \cdots & \alpha_n \end{pmatrix}^T \in \mathbb{R}^n$ with $\alpha_i \in \{0, 1, \cdots\}$ for all $i = 1, \cdots, n$. We define $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, $|\alpha| = \sum_{i=1}^n \alpha_i$ and $\alpha! = \alpha_1! \cdots \alpha_n!$.

Lemma 9. *(Multinomial Theorem) For any $x = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T \in \mathbb{R}^n$ and for any $k \in \mathbb{N}$ we have*

$$(x_1 + \cdots + x_n)^k = \sum_{|\alpha|=k} \frac{k!}{\alpha!} x^\alpha \tag{34}$$

Proof. The proof is by induction. First, if $n = 2$ we have from the Binomial Theorem

$$(x_1 + x_2)^k = \sum_{j=0}^k \frac{k!}{(k-j)!j!} x_1^j x_2^{k-j} = \sum_{\alpha_1+\alpha_2=k} \frac{k!}{(\alpha_1!\alpha_2!)} x_1^{\alpha_1} x_2^{\alpha_2} = \sum_{|\alpha|=k} \frac{k!}{(\alpha!)} x^\alpha.$$

Now, suppose that equation (34) is true for all $n < N$. We now prove it holds for N.

$$\begin{aligned}(x_1 + \cdots + x_N)^k &= ((x_1 + \cdots + x_{N-1}) + x_N) = (\tilde{x} + x_N)^k \\ &= \sum_{\alpha_1+\alpha_2=k} \frac{k!}{(\alpha_1!\alpha_2!)} \tilde{x}^{\alpha_1} x_N^{\alpha_2} = \sum_{\alpha_1+\alpha_2=k} \frac{k!}{(\alpha_1!\alpha_2!)} \sum_{|\beta|=\alpha_1} \frac{\alpha_1!}{\beta!} \tilde{x}^\beta x_N^{\alpha_2}\end{aligned}$$

where $\beta = \begin{pmatrix} \beta_1 & \cdots & \beta_{N-1} \end{pmatrix}^T$. Let $\alpha = \begin{pmatrix} \beta_1 & \cdots & \beta_{N-1} & \alpha_2 \end{pmatrix}$, then $\beta!\alpha_2! = \alpha!$ and $\tilde{x}^\beta x_N^{\alpha_2} = x^\alpha$. Thus,

$$(x_1 + \cdots + x_N)^k = \sum_{|\alpha|=k} \frac{k!}{\alpha!} x^\alpha.$$

□

Let $\frac{\partial}{\partial_j} f$ be the partial derivative operator with respect to the j^{th} argument of a function $f : \mathbb{R}^n \to \mathbb{R}$ and $\partial^\alpha f(x) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} f(x)$.

Lemma 10. *(Multivariate Taylor's Theorem) Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is $k+1$ times continuously differentiable in an open and convex subset S of $\mathbb{R}^n$. If $a, a+h \in S$, then*

$$f(a+h) = \sum_{|\alpha| \le k} \frac{1}{\alpha!} h^\alpha \partial^\alpha f(a) + R_{a,k}(h)$$

where the remainder $R_{a,k}(h)$ is given by

$$R_{a,k}(h) = \sum_{|\alpha|=k+1} \frac{1}{\alpha!} h^\alpha \partial^\alpha f(a + ch) \text{ for some } c \in (0,1)$$

or

$$R_{a,k}(h) = (k+1) \sum_{|\alpha|=k+1} \frac{1}{\alpha!} h^\alpha \int_0^1 (1-t)^k \partial^\alpha f(a+th) dt$$

Proof. Let S be a convex open subset of $\mathbb{R}^n$ containing x and a and $x^* = a + th$ for $h = x - a$ and $t \in (0,1)$. Let $g(t) = f(a+th)$ and note that

$$\begin{aligned}
\frac{d}{dt} g(t) &= h \cdot \begin{pmatrix} \frac{\partial}{\partial_1} f(a+th) & \cdots & \frac{\partial}{\partial_n} f(a+th) \end{pmatrix} = \sum_{j=1}^n \frac{\partial}{\partial_j} f(a+th) h_j = h \cdot \nabla f(a+th) \\
\frac{d^2}{dt^2} g(t) &= \sum_{j=1}^n h_j \frac{d}{dt} \frac{\partial}{\partial_j} f(a+th) \\
&= \sum_{j=1}^n h_j \sum_{i=1}^n \frac{\partial^2}{\partial_{ji}} f(a+th) h_i \\
&= \sum_{j=1}^n h_j h \cdot \begin{pmatrix} \frac{\partial^2}{\partial_{j1}} f(a+th) & \cdots & \frac{\partial^2}{\partial_{jn}} f(a+th) \end{pmatrix} \\
&= \sum_{j=1}^n h_j (h \cdot \nabla \frac{\partial}{\partial_j} f(a+th)) \\
&= h \cdot \nabla (h \cdot \nabla f(a+th)) = (h\nabla)^2 f(a+th).
\end{aligned}$$

Repeating the argument we have $\frac{d^j}{dt^j} g(t) = (h\nabla)^j f(a+th)$. Now, by the univariate version of Taylor's Theorem we have

$$g(1) = g(0) + \frac{d}{dt} g(0) + \frac{1}{2!} \frac{d^2}{dt^2} g(0) + \cdots + \frac{1}{k!} \frac{d^k}{dt^k} g(0) + R_k$$

where R_k denotes a remainder. Using $g(t) = f(a+th)$ and $\frac{d^j}{dt^j}g(t) = (h\nabla)^j f(a+th)$ we have

$$f(a+h) = \sum_{j=0}^{k} \frac{1}{j!}(h\nabla)^j f(a) + R_{a,k}(h)$$

By Lemma 9 we write

$$\begin{aligned} f(a+h) &= \sum_{j=0}^{k} \frac{1}{j!} \sum_{|\alpha|=j} \frac{j!}{\alpha!} h^\alpha \partial^\alpha f(a) + R_{a,k}(h) \\ &= \sum_{|\alpha|\leq k} \frac{1}{\alpha!} h^\alpha \partial^\alpha f(a) + R_{a,k}(h). \end{aligned}$$

The structure of the remainder follows from the univariate version of Taylor's Theorem. □

Lemma 11. *(Markov's Inequality) Let $E(|X(\omega)|) < \infty$. Then, for $a > 0$*

$$P\left(|X(\omega)| \geq a\right) \leq \frac{1}{a} E\left(|X(\omega)|\right).$$

Proof. Note that, $aI_{\{|X|\geq a\}} \leq |X(\omega)|$ and $a\int I_{\{|X(\omega)|\geq a\}} dP \leq \int |X(\omega)| dP$. Therefore,

$$P(|X(\omega)| \geq a) \leq \frac{1}{a} E\left(|X(\omega)|\right).$$

□

References

Hall, P., Marron, J. S., 1990. On variance estimation in nonparametric regression. Biometrika 77, 415–419.

Hardle, W., Hall, P., Marron, J. S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? Journal of the American Satistical Association 83, 86–95.

Park, B., Marron, J. S., 1990. Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association 85, 66–72.

Ruppert, D., Sheather, S., Wand, M. P., 1995. An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association 90, 1257–1270.

Sheather, S., Jones, M. C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society B 53, 683–690.