Chapter 4

Nonparametric Regression

4.1 Introduction

Consider (X,Y) a bivariate random variable. Y will be called the response variable and X the covariate. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be an independent identically distributed sample from (X,Y). Consider the model

$$Y_i = m(X_i) + e_i, \qquad i = 1,\ldots,n,$$

with design variable X_i and error term e_i. Examples of parametric regression models are:

- linear regression model
$$m(X_i) = \beta_0 + \beta_1 X_i$$
where β_0 and β_1 are unknown parameters
- quadratic regression model
$$m(X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$$
where β_0, β_1 and β_2 are unknown parameters
- nonlinear regression model
$$m(X_i) = \frac{\beta_0}{1 + \beta_1 \exp(-\beta_2 X_i)}$$
where β_0, β_1 and β_2 are unknown parameters.

In a nonparametric regression model no assumptions are made on the form of the function m.

Example 4.1 (Motorcycle data) *The data concern 133 observations of a variable X which is the time (in milliseconds) after a simulated impact with motorcycles, and a variable Y which is the head acceleration (in g) of a PMTO (post mortem human test object). Figure 4.1a shows the local cubic kernel estimate on the Motorcycle data set.*

Example 4.2 (Old Faithful geyser data) *A version of the eruptions data from the Old Faithful geyser in Yellowstone National Park, Wyoming. This version comes from Azzalini and Bowman (1990) and is of continuous measurement from August 1 to August 15, 1985. Figure 4.1b shows a local linear fit to the data.*

In nonparametric regression one makes the distinction between a fixed design model and a random design model.

- In the fixed design case the response variables are assumed to satisfy
$$Y_i = m(x_i) + e_i, \qquad i = 1,\ldots,n,$$
where the $x_1,\ldots,x_n$ are **nonrandom** numbers and $e_1,\ldots,e_n$ are independent random variables with
$$\mathbf{E}[e_i] = 0 \qquad \text{and} \qquad \mathbf{Var}[e_i] = \sigma^2(x_i). \tag{4.1}$$

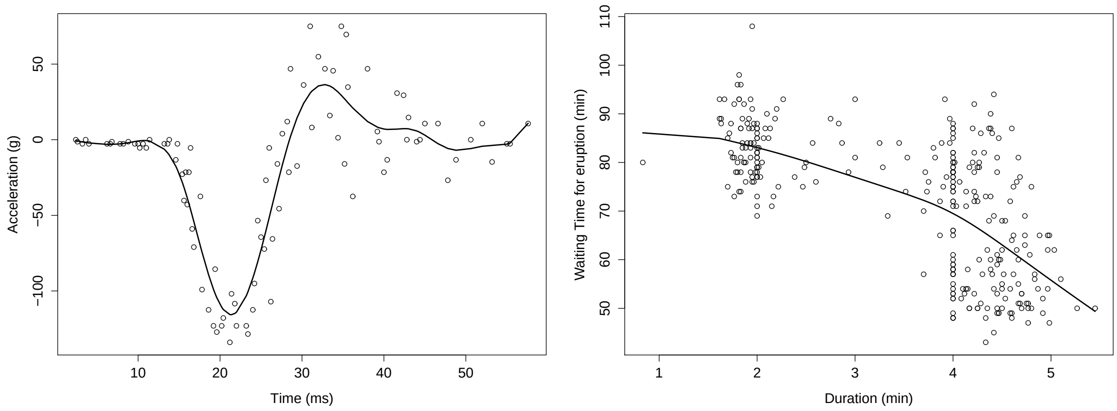


Figure 4.1: (a) Scatterplot of the Motorcycle data set with local cubic kernel estimate. (b) Scatterplot of the Old Faithful geyser data set with local linear kernel estimate.

We call m the mean regression function, or simply the regression function, since from model (4.1)

$$\mathbf{E}[Y_i] = m(x_i)$$

while

$$\mathbf{Var}[Y_i] = \sigma^2(x_i)$$

is called the variance function. The following distinction can now be made:

- $\sigma^2(x_i) = \sigma^2$: homoscedasticity
- heteroscedasticity: $\sigma^2(x_i)$

Hence in the fixed design context the nonrandom numbers are chosen by the experimentor. Some special cases of fixed designs are regular design and equally spaced design.

- The random design regression model is given by

$$Y_i = m(X_i) + \sigma(X_i)e_i, \qquad i = 1, \ldots ,n,$$

where, conditional on $X_1, \ldots ,X_n$, the e_i are independent random variables with

$$\mathbf{E}[e_i|X = x] = 0 \qquad \text{and} \qquad \mathbf{Var}[e_i|X = x] = 1.$$

In this random design context we have that

$$\begin{aligned} \mathbf{E}[Y|X = x] &= \mathbf{E}[m(X) + \sigma(X)e|X = x] \\ &= m(x) + \sigma(x)\,\mathbf{E}[e|X = x] \\ &= m(x), \end{aligned}$$

and

$$\begin{aligned} \mathbf{Var}[Y|X = x] &= \mathbf{E}[Y^2|X = x] - m^2(X) \\ &= \mathbf{E}[m^2(X) + \sigma^2(X)e^2 + 2m(X)\sigma(X)e|X = x] - m^2(X) \\ &= \sigma^2(x). \end{aligned}$$

Hence $m(x)$ is the conditional mean of Y given $X = x$ and $\sigma^2(x)$ is the conditional variance of Y given $X = x$.

4.2 Nadaraya-Watson regression estimator

We have that

$$m(x) = \mathbf{E}[Y|X=x] = \int y f_{Y|X}(x,y)\,dy = \frac{1}{f_X(x)} \int y f_{XY}(x,y)\,dy. \tag{4.2}$$

The unknown quantity $f_{XY}(x,y)$ can be estimated by a bivariate kernel density estimator with product kernel

$$\widehat{f}_{XY}(x,y) = \frac{1}{nhh^\star}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)K\left(\frac{y-Y_i}{h^\star}\right),$$

with bandwidth h and $h^\star$ in the X-direction and Y-direction respectively. We can estimate $\int y f_{XY}(x,y)\,dy$ by

$$\begin{aligned}\int y\widehat{f}_{XY}(x,y)\,dy &= \frac{1}{nhh^\star}\sum_{i=1}^n \int yK\left(\frac{x-X_i}{h}\right)K\left(\frac{y-Y_i}{h^\star}\right)dy \\ &= \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{1}{h^\star}\int yK\left(\frac{y-Y_i}{h^\star}\right)dy \\ &= \frac{h^\star}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{1}{h^\star}\int (Y_i+uh^\star)K(u)\,du \\ &= \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)Y_i,\end{aligned}$$

if $\int K(u)\,du = 1$ and $\int uK(u)\,du = 0$. By replacing $f_X(x)$ by its kernel density estimator in (4.2) we obtain the Nadaraya-Watson kernel regression estimator (independently introduced by Nadaraya (1964) and Watson (1964))

$$\widehat{m}(x) = \sum_{i=1}^n \frac{K\left(\frac{x-X_i}{h}\right)Y_i}{\sum_{j=1}^n K\left(\frac{x-X_i}{h}\right)}. \tag{4.3}$$

Remark 4.1 *Note that this estimator can be written in the following form*

$$\widehat{m}(x) = \sum_{i=1}^n\left(\frac{w_i}{\sum_{j=1}^n w_j}\right)Y_i \quad \textit{with} \quad w_i = K\left(\frac{x-X_i}{h}\right).$$

This is a linear combination of the Y_i's. In general, an estimator of the form

$$\sum_{i=1}^n W_i(x;X_1,\ldots,X_n)Y_i$$

is called a linear smoother.

Next, we will show that the Nadaraya-Watson estimator (4.3) is a consistent estimator for $m(x)$. Let

$$\widehat{m}(x) = \sum_{i=1}^n \frac{\frac{1}{nh}K\left(\frac{x-X_i}{h}\right)Y_i}{\frac{1}{nh}\sum_{j=1}^n K\left(\frac{x-X_i}{h}\right)} \equiv \frac{\widehat{r}(x)}{\widehat{f}_X(x)}$$

and $m(x) \equiv \dfrac{r(x)}{f_X(x)}$. From Chapter 2 we already know that if $f_X(\cdot)$ is continuous at x, $h \to 0$, $nh \to \infty$ as $n \to \infty$ and some conditions on the kernel function (see Theorem 2.3) then

$$\widehat{f}_X(x) \xrightarrow{p} f_X(x)$$

and (if f_X has at least two derivatives)

$$\begin{aligned}\mathbf{E}[\widehat{f}_X(x)] &= f_X(x) + \frac{h^2}{2}\mu_2 f_X''(x) + o(h^2)\\ \mathbf{Var}[\widehat{f}_X(x)] &= \frac{R(K)}{nh} f_X(x) + o\Big(\frac{1}{nh}\Big).\end{aligned}$$

For the bias of $\widehat{r}(x)$ we have

$$\begin{aligned}\mathbf{E}[\widehat{r}(x)] &= \frac{1}{h}\mathbf{E}\left[K\Big(\frac{x-X}{h}\Big)Y\right]\\ &= \frac{1}{h}\mathbf{E}\left\{\mathbf{E}\Big[K\Big(\frac{x-X}{h}\Big)Y|X\Big]\right\}\\ &= \frac{1}{h}\mathbf{E}\left\{K\Big(\frac{x-X}{h}\Big)\mathbf{E}[Y|X]\right\}\\ &= \frac{1}{h}\mathbf{E}\left\{K\Big(\frac{x-X}{h}\Big)m(X)\right\}\\ &= \frac{1}{h}\int K\Big(\frac{x-y}{h}\Big)m(y)f_X(y)\,dy\\ &= \frac{1}{h}\int K\Big(\frac{x-y}{h}\Big)r(y)\,dy\\ &= \frac{1}{h}\int K\Big(\frac{u}{h}\Big)r(x-u)\,du.\end{aligned}$$

Applying Bochner's lemma (Lemma 2.1) yields

$$\lim_{n\to\infty}\mathbf{E}[\widehat{r}(x)] = r(x)$$

provided that $m(\cdot)$ and $f(\cdot)$ are continuous at the point x, $h \to 0$ as $n \to \infty$ and under the necessary assumptions on K. Also, using a Taylor expansion yields (if $m(\cdot)$ and $f(\cdot)$ have at least two derivatives)

$$\mathbf{E}[\widehat{r}(x)] = r(x) + \frac{h^2}{2}\mu_2 r''(x) + o(h^2). \tag{4.4}$$

For the variance of $\widehat{r}(x)$ we have

$$\begin{aligned}\mathbf{Var}[\widehat{r}(x)] &= \mathbf{Var}\left[\frac{1}{nh}\sum_{i=1}^{n}K\Big(\frac{x-X_i}{h}\Big)Y_i\right]\\ &= \frac{1}{n}\mathbf{Var}\left[\frac{1}{h}K\Big(\frac{x-X}{h}\Big)Y\right]\\ &= \frac{1}{n}\left\{\mathbf{E}\Big[\frac{1}{h^2}K^2\Big(\frac{x-X}{h}\Big)Y^2\Big] - \mathbf{E}^2\Big[\frac{1}{h}K\Big(\frac{x-X}{h}\Big)Y\Big]\right\}.\end{aligned}$$

For the first term we have

$$\begin{aligned}\mathbf{E}\Big[\frac{1}{h^2}K^2\Big(\frac{x-X}{h}\Big)Y^2\Big] &= \mathbf{E}\left\{\mathbf{E}\Big[\frac{1}{h^2}K^2\Big(\frac{x-X}{h}\Big)Y^2|X\Big]\right\}\\ &= \mathbf{E}\left\{\frac{1}{h^2}K^2\Big(\frac{x-X}{h}\Big)\mathbf{E}[Y^2|X]\right\}\\ &= \mathbf{E}\left\{\frac{1}{h^2}K^2\Big(\frac{x-X}{h}\Big)\big(\mathbf{Var}[Y|X]+m^2(X)\big)\right\}\\ &= \frac{1}{h^2}\mathbf{E}\left\{K^2\Big(\frac{x-X}{h}\Big)\sigma^2(X)\right\} + \frac{1}{h^2}\mathbf{E}\left\{K^2\Big(\frac{x-X}{h}\Big)m^2(X)\right\}\end{aligned}$$

Putting everything together we have

$$nh\,\mathbf{Var}[\widehat{r}(x)] = \frac{1}{h}\mathbf{E}\left\{K^2\left(\frac{x-X}{h}\right)\sigma^2(X)\right\} + \frac{1}{h}\mathbf{E}\left\{K^2\left(\frac{x-X}{h}\right)m^2(X)\right\} - h\big(\mathbf{E}[\widehat{r}(x)]\big)^2.$$

Taking the limit for $n \to \infty$ gives (and by applying Bochner's lemma (Lemma 2.1))

$$\lim_{n\to\infty} nh\,\mathbf{Var}[\widehat{r}(x)] = \sigma^2(x)f_X(x)R(K) + m^2(x)f_X(x)R(K) + 0$$

for $\sigma(\cdot), m(\cdot)$ and $f_X(\cdot)$ continuous at the point x. For $h \to 0$ and $nh \to \infty$ as $n \to \infty$, the MSE of $\widehat{r}(x)$ is

$$\begin{aligned}\mathbf{E}[\widehat{r}(x) - r(x)]^2 &= \frac{h^4}{4}\mu_2^2\{r''(x)\}^2 + \frac{1}{nh}\{\sigma^2(x) + m^2(x)\}f_X(x)R(K) + o\left(h^4 + \frac{1}{nh}\right) \\ &\to 0, \text{ as } n \to \infty.\end{aligned}$$

Hence, by Chebyshev's inequality we have that for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}[|\widehat{r}(x) - r(x)| \geq \epsilon] = 0 \quad \text{or} \quad \widehat{r}(x) \xrightarrow{\text{P}} r(x)$$

and from Chapter 2

$$\lim_{n\to\infty} \mathbf{P}[|\widehat{f}_X(x) - f_X(x)| \geq \epsilon] = 0 \quad \text{or} \quad \widehat{f}_X(x) \xrightarrow{\text{P}} f_X(x).$$

Then by Slutsky's theorem we have

$$\widehat{m}(x) = \frac{\widehat{r}(x)}{\widehat{f}_X(x)} \xrightarrow{\text{P}} \frac{r(x)}{f_X(x)} = m(x),$$

provided that $f_X(x) > 0$. Next, we can show the bias and variance expressions for the Nadaraya-Watson estimator. First, consider the following expansion

$$\begin{aligned}\widehat{m}(x) - m(x) &= \left(\frac{\widehat{r}(x)}{\widehat{f}_X(x)} - m(x)\right)\left\{\frac{\widehat{f}_X(x)}{f_X(x)} + \left(1 - \frac{\widehat{f}_X(x)}{f_X(x)}\right)\right\} \\ &= \frac{\widehat{r}(x) - m(x)\widehat{f}_X(x)}{f_X(x)} + \frac{1}{f_X(x)}(\widehat{m}(x) - m(x))(f_X(x) - \widehat{f}_X(x)). \qquad (4.5)\end{aligned}$$

For the bias part we have

$$\begin{aligned}\mathbf{E}[\widehat{m}(x) - m(x)] &= \mathbf{E}\left[\frac{\widehat{r}(x) - m(x)\widehat{f}_X(x)}{f_X(x)}\right] + \frac{1}{f_X(x)}\mathbf{E}[(\widehat{m}(x) - m(x))(f_X(x) - \widehat{f}_X(x))] \\ &= B_1 + B_2.\end{aligned}$$

For the first term and using (4.4),

$$\begin{aligned}B_1 &= \frac{1}{f_X(x)}\left\{r(x) + \frac{h^2}{2}\mu_2 r''(x) - m(x)\big[f(x) + \frac{h^2}{2}f''(x)\mu_2\big]\right\} + o(h^2) \\ &= m(x) + \frac{h^2}{2}\mu_2\frac{r''(x)}{f_X(x)} - m(x) - m(x)\frac{h^2}{2}\mu_2\frac{f_X''(x)}{f_X(x)} + o(h^2) \\ &= \frac{h^2}{2}\mu_2\frac{m''(x)f_X(x) + 2m'(x)f_X'(x) + m(x)f_X''(x) - m(x)f_X''(x)}{f_X(x)} + o(h^2) \\ &= \frac{h^2}{2}\mu_2\left(2m'(x)\frac{f_X'(x)}{f_X(x)} + m''(x)\right) + o(h^2) \qquad (4.6)\end{aligned}$$

and the second term gives

$$\begin{aligned}B_2 &\leq \frac{1}{f_X(x)}\big|\mathbf{E}[(\widehat{m}(x) - m(x))(f_X(x) - \widehat{f}_X(x))]\big| \\ &\overset{\text{Cauchy-Schwartz}}{\leq} \frac{1}{f_X(x)}\sqrt{\mathbf{E}[\widehat{m}(x) - m(x)]^2}\sqrt{\mathbf{E}[f_X(x) - \widehat{f}_X(x)]^2}.\end{aligned}$$

The asymptotic order of MSE($\widehat{m}(x)$) will be the same as MSE($\widehat{r}(x)$) i.e., for a bandwidth $h = O(n^{-1/5})$ we have MSE($\widehat{m}(x)$) $= O(n^{-4/5})$. Also, for a bandwidth $h = O(n^{-1/5})$ we know from Chapter 2 that MSE($\widehat{f}_X(x)$) $= O(h^4 + 1/(nh)) = O(n^{-4/5})$. Then, for the term B_2 we have

$$B_2 = O(\sqrt{n^{-4/5}})O(\sqrt{n^{-4/5}}) = O(n^{-4/5}). \tag{4.7}$$

Combining (4.6) and (4.7) gives the bias of the Nadaraya-Watson estimator

$$\text{bias}[\widehat{m}(x)] = \frac{h^2}{2}\mu_2\left(2m'(x)\frac{f'_X(x)}{f_X(x)} + m''(x)\right) + o(h^2) + O(n^{-4/5}) = \frac{h^2}{2}\mu_2\left(2m'(x)\frac{f'_X(x)}{f_X(x)} + m''(x)\right) + o(h^2).$$

Next, by using (4.5)

$$\begin{aligned}
\mathbf{Var}[\widehat{m}(x)] &= \mathbf{Var}\left[\frac{\widehat{r}(x) - m(x)\widehat{f}_X(x)}{f_X(x)} + \frac{1}{f_X(x)}\{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}\right] \\
&= \frac{1}{f_X^2(x)}\,\mathbf{Var}\big[\widehat{r}(x) - m(x)\widehat{f}_X(x)\big] + \frac{1}{f_X^2(x)}\,\mathbf{Var}\big[\{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}\big] \\
&+ \frac{2}{f_X^2(x)}\,\mathbf{Cov}[\widehat{r}(x) - m(x)\widehat{f}_X(x),\{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}] \\
&= V_1 + V_2 + V_3.
\end{aligned} \tag{4.8}$$

For each of the three terms we have and using independence

$$\begin{aligned}
V_1 &= \frac{1}{f_X^2(x)}\,\mathbf{Var}\big[\widehat{r}(x) - m(x)\widehat{f}_X(x)\big] \\
&= \frac{1}{f_X^2(x)}\,\mathbf{Var}\left[\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)Y_i - \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)m(x)\right] \\
&= \frac{1}{nh^2 f_X^2(x)}\,\mathbf{Var}\left[K\left(\frac{x - X}{h}\right)\{Y - m(x)\}\right] \\
&= \frac{1}{nh^2 f_X^2(x)}\,\mathbf{E}\left[K^2\left(\frac{x - X}{h}\right)\{Y - m(x)\}^2\right] - \frac{1}{nh^2 f_X^2(x)}\,\mathbf{E}^2\left[K\left(\frac{x - X}{h}\right)\{Y - m(x)\}\right] \\
&= V_{11} + V_{12}.
\end{aligned}$$

Analyzing each term separate, using Taylor series and $h \to 0$ yields

$$\begin{aligned}
V_{11} &= \frac{1}{nh^2 f_X^2(x)}\,\mathbf{E}\left[K^2\left(\frac{x - X}{h}\right)\mathbf{E}[\{Y^2 - 2Ym(x) + m^2(x)\}|X]\right] \\
&= \frac{1}{nh^2 f_X^2(x)}\,\mathbf{E}\left[K^2\left(\frac{x - X}{h}\right)\{\sigma^2(X) + m^2(X) - 2m(X)m(x) + m^2(x)\}\right] \\
&= \frac{1}{nhf_X^2(x)}\int K^2(u)\sigma^2(x - uh)f_X(x - uh)\,du + \frac{1}{nhf_X^2(x)}\int K^2(u)m^2(x - uh)f_X(x - uh)\,du \\
&\quad - \frac{2m(x)}{nhf_X^2(x)}\int K^2(u)m(x - uh)f_X(x - uh)\,du + \frac{m^2(x)}{nhf_X^2(x)}\int K^2(u)f_X(x - uh)\,du \\
&= \frac{1}{nhf_X^2(x)}\int K^2(u)\{\sigma^2(x) + o(1)\}\{f_X(x) + o(1)\}\,du + \frac{1}{nhf_X^2(x)}\int K^2(u)\{m^2(x) + o(1)\}\{f_X(x) + o(1)\}\,du \\
&\quad - \frac{2m(x)}{nhf_X^2(x)}\int K^2(u)\{m(x) + o(1)\}f_X(x - uh)\,du + \frac{m^2(x)}{nhf_X^2(x)}\int K^2(u)\{f_X(x) + o(1)\}\,du \\
&= \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + \frac{m^2(x)}{nhf_X(x)}\int K^2(u)\,du - \frac{2m^2(x)}{nhf_X(x)}\int K^2(u)\,du + \frac{m^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\left(\frac{1}{nh}\right) \\
&= \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\left(\frac{1}{nh}\right)
\end{aligned}$$

and

$$\begin{aligned} V_{12} &= \frac{1}{nh^2 f_X^2(x)} \mathbf{E}^2\left[K\left(\frac{x-X}{h}\right)\mathbf{E}[\{Y-m(x)\}|X]\right] \\ &= \frac{1}{nf_X^2(x)}\left[\int K(u)\{m(x-uh)-m(x)\}f_X(x-uh)\,du\right]^2 \\ &= \frac{1}{nf_X^2(x)}\left[\int K(u)\{m(x)+o(1)-m(x)\}\{f_X(x)+o(1)\}\,du\right]^2 = o\left(\frac{1}{n}\right). \end{aligned}$$

Hence, for the first term V_1 we have

$$V_1 = \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\left(\frac{1}{nh}\right) + o\left(\frac{1}{n}\right) = \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\left(\frac{1}{nh}\right).$$

To obtain the order of the next two terms in (4.8) we need the following two technical lemmas.

Lemma 4.1 *For any two random variables X and Y with finite variances*

$$\mathbf{Var}[X \pm Y] \leq 2\,\mathbf{Var}\,X + 2\,\mathbf{Var}\,Y.$$

Proof. From the variance of the sum of two random variables we have

$$0 \leq \mathbf{Var}[X \pm Y] = \mathbf{Var}\,X + \mathbf{Var}\,Y \pm 2\,\mathbf{Cov}[X,Y]$$

and

$$|2\,\mathbf{Cov}[X,Y]| \leq \mathbf{Var}\,X + \mathbf{Var}\,Y.$$

Substituting the latter equation into the first gives the result. ■

Lemma 4.2 *Let X and Y be any two random variables with $\mathbf{E}[X] < \infty$, $\mathbf{Var}\,X < \infty$, $\mathbf{Var}\,Y < \infty$. Further, assume there exists a $B \geq 0$ such that $\mathbf{P}[|Y| \leq B] = 1$, then*

$$\mathbf{Var}[XY] \leq 2\|Y\|_\infty^2\,\mathbf{Var}\,X + 2(\mathbf{E}[X])^2\,\mathbf{Var}\,Y$$

where $\|Y\|_\infty = \inf\{B \geq 0 : \mathbf{P}[|Y| \leq B] = 1\}$.

Proof. Using Lemma 4.1 with $X = (X - \mathbf{E}[X])Y$ and $Y = \mathbf{E}[X]Y$ yields

$$\begin{aligned} \mathbf{Var}[XY] &\leq 2\,\mathbf{Var}[(X-\mathbf{E}[X])Y] + 2\,\mathbf{Var}[\mathbf{E}[X]Y] \\ &= 2\,\mathbf{Var}[(X-\mathbf{E}[X])Y] + 2(\mathbf{E}[X])^2\,\mathbf{Var}[Y]. \end{aligned}$$

The first term is (with the 2 omitted)

$$\begin{aligned} \mathbf{Var}[(X-\mathbf{E}[X])Y] &= \mathbf{E}\{[(X-\mathbf{E}[X])Y]^2\} - \{\mathbf{E}[(X-\mathbf{E}[X])Y]\}^2 \\ &\leq \mathbf{E}\{[(X-\mathbf{E}[X])Y]^2\} \\ &\leq \mathbf{E}\{|X-\mathbf{E}[X]|^2|Y|^2\} \\ &\leq \|Y\|_\infty^2\,\mathbf{Var}\,X. \end{aligned}$$

■

Next, our goal is to show that the order of the two last terms in (4.8) are of lower order than the first term. Using Lemma 4.2, the order second term V_2 is

$$\begin{aligned} V_2 &= \frac{1}{f_X^2(x)}\mathbf{Var}\left[(\widehat{m}(x)-m(x))(f_X(x)-\widehat{f}_X(x))\right] \\ &\leq \frac{\left[2(\inf\{B \geq 0 : \mathbf{P}[|f_X(x)-\widehat{f}_X(x)| \leq B] = 1\})^2\,\mathbf{Var}[\widehat{m}(x)] + 2\{\mathbf{E}[\widehat{m}(x)-m(x)]\}^2\,\mathbf{Var}[\widehat{f}_X(x)]\right]}{f_X^2(x)}. \end{aligned}$$

By Theorem 2.6 we know that there exist a B such that $\inf\{B \geq 0 : \mathbf{P}[|f_X(x) - \widehat{f}_X(x)| \leq B] = 1\})$. In fact, Theorem 2.6 states that B goes to zero almost surely and hence the first term in the inequality is $o(1)$. Consequently

$$V_2 = o(1)O\left(\frac{1}{nh}\right) + O(h^4)O\left(\frac{1}{nh}\right) = o\left(\frac{1}{nh}\right).$$

Finally, for the third term (using the covariance inequality)

$$\begin{aligned} V_3 &= \frac{2}{f_X^2(x)}\mathbf{Cov}[\widehat{r}(x) - m(x)\widehat{f}_X(x), \{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}] \\ &= O\left(\sqrt{\mathbf{Var}[\widehat{r}(x) - m(x)\widehat{f}_X(x)]}\sqrt{\mathbf{Var}[\{\widehat{m}(x) - m(x)\}\{f_X(x) - \widehat{f}_X(x)\}]}\right) \\ &= O\left(\frac{1}{\sqrt{nh}}\right)o\left(\frac{1}{\sqrt{nh}}\right) = o\left(\frac{1}{nh}\right). \end{aligned}$$

Combing the three terms together gives the variance of the Nadaraya-Watson estimator

$$\mathbf{Var}[\widehat{m}(x)] = \frac{\sigma^2(x)}{nhf_X(x)}\int K^2(u)\,du + o\left(\frac{1}{nh}\right).$$

4.3 Local polynomial regression

There exist a vast number of methods to construct nonparametric regression estimates e.g., splines, wavelets, support vector machines, local polynomial regression, Nadaraya-Watson regression, Gasser-Müller estimator, Priestly-Chao estimator, orthogonal series estimator, k nearest neighbors, etc. A good overview can be found in Wasserman (2006). A thorough theoretical study of nonparametric regression is given in Györfi et al. (2002). In what follows we will focus on local polynomial regression which is a very popular method in statistics. Most of this material is based on Fan and Gijbels (1996).

4.3.1 Local polynomial regression framework

Consider the bivariate data $(X_1,Y_1),\ldots,(X_n,Y_n)$, which form an independent and identically distributed (i.i.d.) sample from a population (X,Y). Our interest is to estimate the regression function $m(x_0) = \mathbf{E}[Y|X = x_0]$ and its derivatives $m'(x_0), m''(x_0),\ldots,m^{(p)}(x_0)$. To understand the estimation methodology, we can regard the data as being generated from the model

$$Y = m(X) + \sigma(X)e, \tag{4.9}$$

where $\mathbf{E}[e] = 0, \mathbf{Var}[e] = 1$ and X and e are independent. We always denote the conditional variance of Y given $X = x_0$ by $\sigma^2(x_0)$ and the marginal density of X i.e., the *design density*, by f_X.

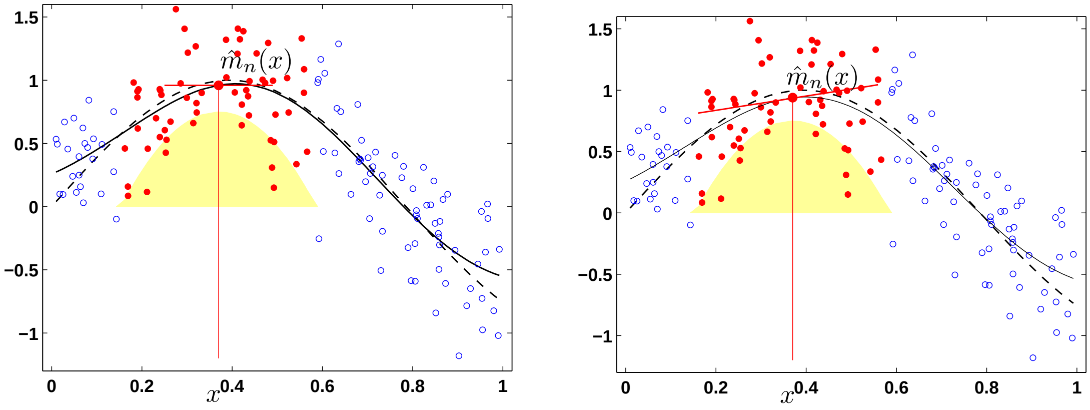
Suppose that the $(p+1)$th derivative of the regression function m at the point x_0 exists. We then approximate the unknown regression function m locally by a polynomial of order p. Figure 4.2 illustrates the idea of local constant $(p = 0)$ and linear kernel regression $(p = 1)$.

A Taylor expansion gives, for x in the neighborhood of x_0,

$$\begin{aligned} m(x) &= m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p + o\big(|x - x_0|^p\big) \\ &= \sum_{j=0}^{p}\frac{m^{(j)}(x_0)}{j!}(x - x_0)^j + o\big(|x - x_0|^p\big) \\ &=: \sum_{j=0}^{p}\beta_j(x - x_0)^j + o\big(|x - x_0|^p\big). \end{aligned} \tag{4.10}$$

This polynomial is fitted locally by the following weighted least squares regression problem:

$$\min_{\beta\in\mathbb{R}^{p+1}}\sum_{i=1}^{n}\Big\{Y_i - \sum_{j=0}^{p}\beta_j\,(X_i - x_0)^j\Big\}^2 K_h(X_i - x_0), \tag{4.11}$$



(a) **(b)**

Figure 4.2: 100 pairs (X_i,Y_i) are generated at random from $Y = \sin(4X)$ (dashed line) with Gaussian errors $e \sim \mathcal{N}(0,1/3)$ and $X \sim \mathcal{U}[0,1]$. The dot around 0.38 (vertical line) is the fitted constant $\hat{m}_n(x)$, and the full circles indicate those observations contributing to the fit at x. The solid region indicates the weights assigned to observations according the Epanechnikov kernel. (a) The full NW estimate is shown by the full line. (b) The full local linear estimate is shown by the full line.

where β_j are the solutions to the weighted least squares problem, $\beta = (\beta_0,\beta_1,\ldots,\beta_p)^T$, $h > 0$ is the bandwidth controlling the size of the local neighborhood and $K_h(\cdot) = K(\cdot/h)/h$ with K a kernel function assigning weights to each point. From the Taylor expansion (4.10) it is clear that $\widehat{m}^{(\nu)}(x_0) = \nu!\widehat{\beta}_\nu$ is an estimator for the νth order derivative $m^{(\nu)}(x_0)$, $\nu = 0,1,\ldots,p$.

It is often more convenient to work with matrix notation. Denote by $\mathbf{X}$ the design matrix of problem (4.11):

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix},$$

and put

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{and} \qquad \widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \vdots \\ \widehat{\beta}_p \end{pmatrix}.$$

Further, $\mathbf{W}$ is the $n \times n$ diagonal matrix of weights

$$\mathbf{W} = \text{diag}\{K_h(X_i - x_0)\}.$$

The weighted least squares problem (4.11) can be written as

$$\min_\beta (\mathbf{Y} - \mathbf{X}\,\beta)^T\, \mathbf{W} (\mathbf{Y} - \mathbf{X}\,\beta),$$

with $\beta = (\beta_0,\ldots,\beta_p)^T$. The solution vector is provided by weighted least squares theory and is given by

$$\widehat{\beta} = (\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathbf{Y}\,. \tag{4.12}$$

There are several important issues which have to be discussed. First of all there is the choice of the bandwidth h, which plays a very crucial role (see also the binwidth for histograms). A too large bandwidth under-parametrizes the regression function, causing a large modelling bias, while a too small bandwidth over-parametrizes the unknown function and results in noisy estimates. In what follows, we will show how to obtain ideal theoretical bandwidth choices. As we will see, this theoretical choice is not directly usable since it depends on unknown quantities. Finding a practical procedure for selecting the bandwidth parameter is one of the most important tasks.

Another issue in local polynomial fitting is the choice of the order of the local polynomial. Since the modelling bias is primarily controlled by the bandwidth, this issue is less crucial however. For a given bandwidth h, a large value of p would expectedly reduce the modelling bias, but would cause a large variance and considerable computational cost.

How good are the local polynomial estimators compared to other estimators? An answer to this question is provided by studying the efficiency of the local polynomial fit. It is beyond the scope of the course to prove the efficiency of local polynomial estimators, but it can be shown that local polynomial fitting is nearly optimal in an asymptotic minimax sense (Fan and Gijbels, 1996, Chapter 3).

From a computational point of view local polynomial estimators are attractive, due to their simplicity. It might be desirable however to speed up the computations especially when computing intensive procedures e.g., bandwidth selection, are to be implemented (Fan and Marron, 1994).

4.4 Advantages of local polynomial fitting

Local polynomial fitting is an attractive method both from theoretical and practical point of view. Other commonly used kernel estimators, such as the Nadaraya-Watson (NW) estimator and the Gasser-Müller (GM) estimator suffer from some drawbacks. In summary, the NW estimator leads to an undesirable form of the bias, while the GM estimator has to pay a price in variance when dealing with a random design model. Local polynomial fitting also has other advantages. The method adopts to various types of designs such as random and fixed designs, highly clustered and nearly uniform designs. Furthermore, there is an absence of boundary effects: the bias at the boundary stays automatically of the same order as in the interior, without the use of specific boundary kernels! The local polynomial approximation method is appealing on general scientific grounds: the least squares principle to be applied opens the way to a wealth of statistical knowledge and thus easy generalizations.

4.4.1 Bias and variance of local polynomial fitting

When dealing with the bandwidth selection problem, a key issue is to have a good insight into bias and variance of the estimators, since a trade-off between these two quantities forms the core of many bandwidth selection criteria. The conditional bias and variance of the estimator $\widehat{\beta}$ are derived immediately from its definition in (4.12)

$$\begin{aligned}
\mathbf{E}[\widehat{\beta}\,|\mathbb{X}] &= (\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathbf{m} \\
&= (\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}(\mathbf{m}-\mathbf{X}\,\beta+\mathbf{X}\,\beta) \\
&= \beta+(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}(\mathbf{m}-\mathbf{X}\,\beta) \\
&= \beta+(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathbf{r} \qquad (4.13)
\end{aligned}$$

$$\begin{aligned}
\mathbf{Var}[\widehat{\beta}\,|\mathbb{X}] &= (\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathrm{diag}\{\sigma^2(X_i)\}\,\mathbf{W}^T\,\mathbf{X}(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}(\mathbf{X}^T\,\Sigma\,\mathbf{X})(\mathbf{X}^T\,\mathbf{W}\,\mathbf{X})^{-1}, \qquad (4.14)
\end{aligned}$$

where $\mathbb{X}=(X_1,\ldots,X_n)$, $\mathbf{m}=(m(X_1),\ldots,m(X_n))^T$, $\beta=(m(x_0),\ldots,m^{(p)}(x_0)/p!)^T$, $\mathbf{r}=\mathbf{m}-\mathbf{X}\,\beta$, the vector of residuals of the local polynomial approximation and $\Sigma=\mathrm{diag}\{K_h^2(X_i-x_0)\sigma^2(X_i)\}$.

These exact bias and variance expressions are not directly usable, since they depend on unknown quantities: the residual $\mathbf{r}$and the diagonal matrix Σ. Hence, there is a need for approximating the bias and variance. We first of all show how to derive the asymptotic expression for the conditional variance given in (4.21). Denote by $S_n=\mathbf{X}^T\,\mathbf{W}\,\mathbf{X}$ and $S_n^\star=\mathbf{X}^T\,\Sigma\,\mathbf{X}$ the $(p+1)\times(p+1)$ matrix $(S_{n,j+l}^\star)_{0\le j,l\le p}$ with $S_{n,j}^\star=\sum_{i=1}^n(X_i-x_0)^jK_h^2(X_i-x_0)\sigma^2(X_i)$. Then, the conditional variance in (4.21) can be re-expressed as

$$S_n^{-1}S_n^\star S_n^{-1}, \qquad (4.15)$$

and the task is now to find the approximations for the two matrices S_n and $S_n^\star$. Since $S_{n,j}=\sum_{i=1}^n K_h(X_i-x_0)(X_i-x_0)^j$, we have that

$$\begin{aligned}
S_{n,j} &= \mathbf{E}[S_{n,j}]+\frac{S_{n,j}-\mathbf{E}[S_{n,j}]}{\sqrt{\mathbf{Var}[S_{n,j}]}}\sqrt{\mathbf{Var}[S_{n,j}]} \\
&= \mathbf{E}[S_{n,j}]+O_p\big(\sqrt{\mathbf{Var}[S_{n,j}]}\big). \qquad (4.16)
\end{aligned}$$

Because the data is i.i.d., we have

$$\begin{aligned}
\mathbf{E}[S_{n,j}] &= n\,\mathbf{E}[K_h(X-x_0)(X-x_0)^j] \\
&= \frac{n}{h}\int K\left(\frac{x-x_0}{h}\right)(x-x_0)^j f_X(x)\,dx \\
&= nh^j \int u^j K(u) f_X(x_0+uh)\,du \\
&= nh^j \int u^j K(u)(f_X(x_0)+o(1))\,du \\
&= nh^j \left[f_X(x_0)\int u^j K(u)\,du + o(1)\right] \\
&= nh^j f_X(x_0)\mu_j\left[1+o(1)\right],
\end{aligned}$$

with $\mu_j = \int u^j K(u)\,du$. Substituting the above expression into (4.16) and using the Cauchy-Schwartz inequality yields

$$S_{n,j} = nh^j f_X(x_0)\mu_j\left[1+o(1)\right] + O_p\big(\sqrt{n\,\mathbf{E}[(X-x_0)^{2j}K_h^2(X-x_0)]}\big). \tag{4.17}$$

Next we need to find the order of the last term. Similarly, we have that

$$\begin{aligned}
n\,\mathbf{E}[(X-x_0)^{2j}K_h^2(X-x_0)] &= \frac{n}{h^2}\int K^2\left(\frac{x-x_0}{h}\right)(x-x_0)^{2j} f_X(x)\,dx \\
&= nh^{2j-1}\int K^2(u)u^{2j} f_X(x_0+uh)\,du.
\end{aligned}$$

It immediately follows that

$$\begin{aligned}
S_{n,j} &= nh^j f_X(x_0)\mu_j\left[1+o(1)\right] + O_p\big(\sqrt{nh^{2j-1}}\big) \\
&= nh^j f_X(x_0)\mu_j\left[1+o(1)+O_p\big(1/\sqrt{nh}\big)\right] \\
&= nh^j f_X(x_0)\mu_j\left[1+o_p(1)\right],
\end{aligned} \tag{4.18}$$

provided that $h \to 0$ and $nh \to \infty$. Since

$$S_n = \mathbf{X}^T\,\mathbf{W}\,\mathbf{X} = \begin{pmatrix} S_{n,0} & S_{n,1} & \cdots & S_{n,p} \\ S_{n,1} & S_{n,2} & \cdots & S_{n,p+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,p} & S_{n,p+1} & \cdots & S_{n,2p} \end{pmatrix}$$

it follows that

$$\begin{aligned}
S_n &= \begin{pmatrix} nh^0 f_X(x_0)\mu_0\left[1+o_p(1)\right] & nhf_X(x_0)\mu_1\left[1+o_p(1)\right] & \cdots & nh^p f_X(x_0)\mu_p\left[1+o_p(1)\right] \\ nh^1 f_X(x_0)\mu_1\left[1+o_p(1)\right] & nh^2 f_X(x_0)\mu_2\left[1+o_p(1)\right] & \cdots & nh^{p+1} f_X(x_0)\mu_{p+1}\left[1+o_p(1)\right] \\ \vdots & \vdots & \ddots & \vdots \\ nh^p f_X(x_0)\mu_p\left[1+o_p(1)\right] & nh^{p+1} f_X(x_0)\mu_{p+1}\left[1+o_p(1)\right] & \cdots & nh^{2p} f_X(x_0)\mu_{2p}\left[1+o_p(1)\right] \end{pmatrix} \\
&= nf_X(x_0)\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & h & 0 & \cdots & 0 \\ 0 & 0 & h^2 & & \vdots \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & h^p \end{pmatrix}\begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_p \\ \mu_1 & \mu_2 & \cdots & \mu_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_p & \mu_{p+1} & \cdots & \mu_{2p} \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & h & 0 & \cdots & 0 \\ 0 & 0 & h^2 & & \vdots \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & h^p \end{pmatrix}\left[1+o_p(1)\right] \\
&= nf_X(x_0)HSH\left[1+o_p(1)\right],
\end{aligned} \tag{4.19}$$

where $H = \text{diag}\{1, h, \ldots, h^p\}$. Using similar arguments, we have that

$$\begin{aligned} S^\star_{n,j} &= nh^{j-1} f_X(x_0)\sigma^2(x_0) \int u^j K^2(u)\, du \left[1 + o_p(1)\right] \\ &= nh^{j-1} f_X(x_0)\sigma^2(x_0)\nu_j \left[1 + o_p(1)\right] \end{aligned}$$

and hence

$$S^\star_n = nh^{-1} f_X(x_0)\sigma^2(x_0) H S^\star H \left[1 + o_p(1)\right] \tag{4.20}$$

with

$$S^\star = \begin{pmatrix} \nu_0 & \nu_1 & \cdots & \nu_p \\ \nu_1 & \nu_2 & \cdots & \nu_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_p & \nu_{p+1} & \cdots & \nu_{2p} \end{pmatrix}.$$

Now, starting from (4.15) and using (4.19) and (4.20) we find that

$$\mathbf{Var}[\widehat{\beta}\,|\mathbb{X}] = \frac{\sigma^2(x_0)}{f_X(x_0) nh} H^{-1} S^{-1} S^\star S^{-1} H^{-1} \left[1 + o_p(1)\right],$$

and since $\widehat{m}_\nu(x_0) = \nu! \varepsilon^T_{\nu+1}\,\widehat{\beta}$, with $\varepsilon_{\nu+1} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ the unit vector with 1 on the $(\nu+1)$th place, the following theorem follows readily.

Theorem 4.1 (Variance of the local polynomial regression estimator) *Assume that $f_X(x_0) > 0$, $f_X(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighborhood of x_0. Further assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional variance of $\widehat{m}_\nu(x_0)$ is given by*

$$\mathbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \varepsilon^T_{\nu+1} S^{-1} S^\star S^{-1} \varepsilon_{\nu+1} \frac{\nu!^2 \sigma^2(x_0)}{f_X(x_0) nh^{1+2\nu}} + o_p\left(\frac{1}{nh^{1+2\nu}}\right). \tag{4.21}$$

Second, we derive the asymptotic expression for the bias. Here, we have to distinguish between the case that $p-\nu$ is odd and $p-\nu$ is even. Let's consider the case $p-\nu$ first. By using a Taylor expansion the conditional bias, see (4.24), $S_n^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathbf{r}$ of $\widehat{\beta}$ can be written as

$$\begin{aligned} \text{bias}[\widehat{\beta}\,|\mathbb{X}] &= S_n^{-1}\,\mathbf{X}^T\,\mathbf{W} \left\{ \beta_{p+1} \begin{pmatrix} (X_1 - x_0)^{p+1} \\ \vdots \\ (X_n - x_0)^{p+1} \end{pmatrix} + o_p\big(\begin{pmatrix} (X_1 - x_0)^{p+1} \\ \vdots \\ (X_n - x_0)^{p+1} \end{pmatrix} \big) \right\} \\ &= S_n^{-1} \left\{ \beta_{p+1} \begin{pmatrix} S_{n,p+1} \\ \vdots \\ S_{n,2p+1} \end{pmatrix} + o_p\big(\begin{pmatrix} nh^{p+1} \\ \vdots \\ nh^{2p+1} \end{pmatrix} \big) \right\} = S_n^{-1} \left\{ \beta_{p+1} c_n + o_p\big(\begin{pmatrix} nh^{p+1} \\ \vdots \\ nh^{2p+1} \end{pmatrix} \big) \right\}, \end{aligned} \tag{4.22}$$

with $c_n = (S_{n,p+1}, \ldots, S_{n,2p+1})^T$ and $\beta_{p+1} = m^{(p+1)}(x_0)/(p+1)!$. Applying (4.18) and (4.19), we obtain from (4.22)

$$\begin{aligned} \text{bias}[\widehat{\beta}\,|\mathbb{X}] &= \frac{1}{n f_X(x_0)} H^{-1} S^{-1} H^{-1} \left\{ \beta_{p+1} \begin{pmatrix} nh^{p+1} f_X(x_0)\mu_{p+1}[1 + o_p(1)] \\ \vdots \\ nh^{2p+1} f_X(x_0)\mu_{2p+1}[1 + o_p(1)] \end{pmatrix} + o_p\big(\begin{pmatrix} nh^{p+1} \\ \vdots \\ nh^{2p+1} \end{pmatrix} \big) \right\} \left[1 + o_p(1)\right] \\ &= H^{-1} S^{-1} \beta_{p+1} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1/h & \cdots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 1/h^p \end{pmatrix} \left\{ \begin{pmatrix} h^{p+1} & 0 & \cdots & 0 \\ 0 & h^{p+2} & \cdots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & h^{2p+1} \end{pmatrix} \begin{pmatrix} \mu_{p+1} \\ \vdots \\ \mu_{2p+1} \end{pmatrix} [1 + o_p(1)] \right\} \left[1 + o_p(1)\right] \\ &= H^{-1} S^{-1} c_p \beta_{p+1} h^{p+1} [1 + o_p(1)]. \end{aligned} \tag{4.23}$$

The above derivation of course holds for any value of $p - \nu$, but the problem is that for $p - \nu$ even the $(\nu + 1)$th element of the vector $S^{-1}c_p$ is zero. This can be easily seen by writing out the structure of the matrix S and the vector c_p, and recalling that odd order moments of a symmetric kernel are zero. Hence the main term of (4.23) is zero, and one clearly has to proceed to higher order expansions. This essentially means that in all derivations we derived to obtain (4.23) some extra terms have to be taken along. In what follows, we derive the case $p - \nu$ odd. Using (4.23), the conditional bias of the local polynomial estimator is then given by

$$\begin{aligned}
\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}] &= \text{bias}[\nu!\varepsilon_{\nu+1}^T\,\widehat{\beta}\,|\mathbb{X}] = \nu!\varepsilon_{\nu+1}^T H^{-1}S^{-1}c_p\beta_{p+1}h^{p+1}[1 + o_p(1)] \\
&= \frac{\nu!}{(p+1)!}m^{(p+1)}(x_0)\varepsilon_{\nu+1}^T \begin{pmatrix} h^{p+1} & 0 & \cdots & 0 \\ 0 & h^p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & h \end{pmatrix} S^{-1}c_p[1 + o_p(1)] \\
&= \varepsilon_{\nu+1}^T S^{-1}c_p\frac{\nu!}{(p+1)!}m^{(p+1)}(x_0)h^{p+1-\nu} + o_p\big(h^{p+1-\nu}\big).
\end{aligned}$$

We can now finalize the following theorem.

Theorem 4.2 (bias of the local polynomial regression estimator) *Assume that $f_X(x_0) > 0$, $f_X(\cdot)$ and $m^{(p+1)}(\cdot)$ are continuous in a neighborhood of x_0. Further assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional bias of $\widehat{m}_\nu(x_0)$ for $p - \nu$ odd is given by*

$$\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \varepsilon_{\nu+1}^T S^{-1}c_p\frac{\nu!}{(p+1)!}m^{(p+1)}(x_0)h^{p+1-\nu} + o_p\big(h^{p+1-\nu}\big). \tag{4.24}$$

Further, for $p - \nu$ even the asymptotic conditional bias of $\widehat{m}_\nu(x_0)$ is given by

$$\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \varepsilon_{\nu+1}^T S^{-1}\tilde{c}_p\frac{\nu!}{(p+2)!}\left\{m^{(p+2)}(x_0) + (p+2)m^{(p+1)}(x_0)\frac{f_X'(x_0)}{f_X(x_0)}\right\}h^{p+2-\nu} + o_p\big(h^{p+2-\nu}\big),$$

provided that $f_X'(\cdot)$ and $m^{(p+2)}(\cdot)$ are continuous in a neighborhood of x_0 and $nh^3 \to \infty$ and $\tilde{c}_p = (\mu_{p+2}, \ldots, \mu_{2p+2})^T$.

A deeper result than the previous theorem, specifying higher order terms in the asymptotic bias and variance expressions, can be found in Fan et al. (1996). From the previous theorem it is already clear there is a theoretical difference between the cases $p - \nu$ odd and $p - \nu$ even. For $p - \nu$ even, the leading term $O_p\big(h^{p+1}\big)$ in the bias expression is zero due to symmetry of the kernel K and hence the second order term is represented in the theorem. For $p - \nu$ odd, the asymptotic bias has a simpler structure and does not involve $f_X'(x_0)$, a factor appearing in the asymptotic bias when $p - \nu$ is even. This theorem is in fact a generalization of what has already been observed for the special case of the local constant fit ($p = 0$) used for estimating the regression function ($\nu = 0$). The estimator resulting from such a fit, the Nadaraya-Watson (NW) estimator (Nadaraya, 1964; Watson, 1964) has an additional term in the asymptotic bias expression. The NW estimator is defined as

$$\widehat{m}_0(x_0) = \sum_{i=1}^n \frac{K\left(\frac{X_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right)}Y_i. \tag{4.25}$$

It can be shown that polynomial fits with $p - \nu$ odd outperform those with $p - \nu$ even.

4.4.2 Equivalent kernels

Next, we will show how the local polynomial approximation method assigns weights to each point. Note that (4.25) can rewritten as

$$\widehat{m}_0(x_0) = \sum_{i=1}^n W_0^n\left(\frac{X_i - x_0}{h}\right)Y_i \qquad \text{with} \quad W_0^n\left(\frac{X_i - x_0}{h}\right) = \frac{K\left(\frac{X_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right)}.$$

Is it possible to write a similar expression for the more general local polynomial estimator? If possible, this will provide further insight into the method and serves as a technical tool for understanding and deriving its asymptotic properties. The answer is YES!

Recall the notation

$$S_{n,j} = \sum_{i=1}^{n} K_h(X_i - x_0)(X_i - x_0)^j$$

and let $S_n = \mathbf{X}^T\,\mathbf{W}\,\mathbf{X}$ denote the $(p+1)\times(p+1)$ matrix $(S_{n,j+l})_{0\leq i,j\leq p}$. Then, the estimator $\widehat{\beta}_\nu$ can be written as

$$\begin{aligned}\widehat{\beta}_\nu = \varepsilon_{\nu+1}^T\,\widehat{\beta} &= \varepsilon_{\nu+1}^T S_n^{-1}\,\mathbf{X}^T\,\mathbf{W}\,\mathbf{Y}\\ &= \sum_{i=1}^{n} W_\nu^n\left(\frac{X_i - x_0}{h}\right)Y_i,\end{aligned}$$

with $W_\nu^n(t) = \varepsilon_{\nu+1}^T S_n^{-1}(1,th,\ldots,(th)^p)^T K(t)/h$. The above expression reveals that the estimator $\widehat{\beta}_\nu$ is very much like a conventional kernel estimator except that the "kernel" W_ν^n depends on the design points AND locations. This explains why the local polynomial fit can adapt automatically to various designs and to boundary estimation.

Substituting (4.19) into the definition of W_ν^n yields

$$W_\nu^n(t) = \frac{1}{nh^{\nu+1}f_X(x_0)}\varepsilon_{\nu+1}^T S^{-1}(1,t,\ldots,t^p)^T K(t)[1+o_p(1)]$$

and therefore

$$\widehat{\beta}_\nu = \frac{1}{nh^{\nu+1}f_X(x_0)}\sum_{i=1}^{n} K_\nu^\star\left(\frac{X_i - x_0}{h}\right)Y_i[1+o_p(1)] \tag{4.26}$$

with

$$K_{\nu,p}^\star(t) = \varepsilon_{\nu+1}^T S^{-1}(1,t,\ldots,t^p)^T K(t) = \left(\sum_{l=0}^{p} S^{\nu l}t^l\right)K(t), \tag{4.27}$$

with $S^{-1} = \left(S^{jl}\right)_{0\leq j,l\leq p}$. This kernel satisfies the following moment conditions:

$$\int u^q K_{\nu,p}^\star(u)\,du = \delta_{\nu,q} \quad 0\leq \nu,q\leq p. \tag{4.28}$$

Table 4.1 gives the forms of some equivalent kernel functions. To emphasize the dependence of p, we use $K_{\nu,p}^\star$ to denote the equivalent kernel given by (4.27). As an illustration we plot in Figure 4.3 the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)_+$ as well as the equivalent kernel $K_{\nu,p}^\star$ for some values of ν and p.

ν	p	Equivalent kernel function $K_{\nu,p}^\star(t)$
0	1	$K(t)$
0	3	$(\mu_4-\mu_2^2)^{-1}(\mu_4-\mu_2 t^2)K(t)$
1	2	$\mu_2^{-1}tK(t)$
2	3	$(\mu_4-\mu_2^2)^{-1}(t^2-\mu_2)K(t)$

Table 4.1: Equivalent kernel functions $K_{\nu,p}^\star$. Taken from Fan and Gijbels (1996, p. 66)

The conditional bias and variance of the estimator $\widehat{m}_\nu(x_0)$, given in (4.21) and (4.24) respectively, can equally well be re-expressed in terms of the equivalent kernel $K_{\nu,p}^\star$, leading to the asymptotic expression (for $p-\nu$ odd)

$$\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \left(\int t^{p+1}K_{\nu,p}^\star(t)\,dt\right)\frac{\nu!}{(p+1)!}m^{(p+1)}(x_0)h^{p+1-\nu} + o_p\left(h^{p+1-\nu}\right), \tag{4.29}$$

and its variance equals

$$\mathbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}] = \left(\int K_{\nu,p}^{\star 2}(t)\,dt\right)\frac{\nu!^2\sigma^2(x_0)}{f_X(x_0)nh^{1+2\nu}} + o_p\left(\frac{1}{nh^{1+2\nu}}\right). \tag{4.30}$$

These expressions can be obtained from (4.26) and (4.28).

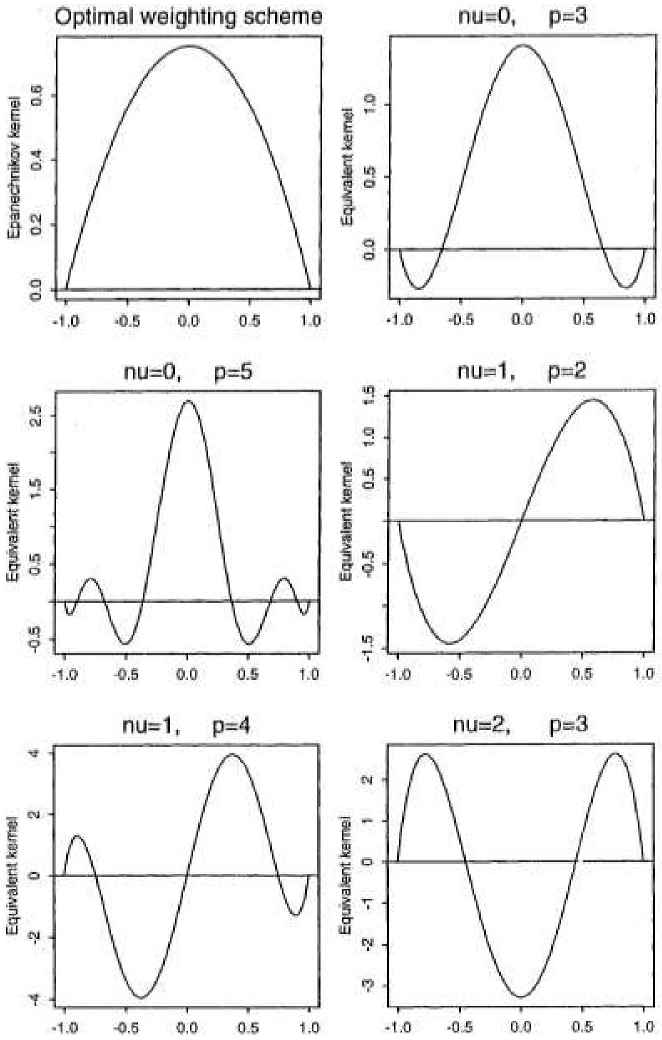


Figure 4.3: Epanechnikov kernel and its equivalent kernel for some values of p and ν. Taken from Fan and Gijbels (1996).

4.4.3 Ideal choice of bandwidth

The choice of the bandwidth parameter is rather crucial and hence should be done with a lot of care. As was the case for density estimation, we can make a distinction between a global or local varying bandwidth. A theoretical optimal local bandwidth for estimating $m^{(\nu)}(x_0)$ is obtained by minimizing the conditional mean squared error given by

$$\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}]^2 + \textbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}]. \tag{4.31}$$

The ideal choice of bandwidth can be approximated by the asymptotically optimal local bandwidth, i.e. the bandwidth which minimizes the asymptotic MSE. It is easy to show that minimizing (4.31) using (4.29) and (4.30), leads to (for $p-\nu$ odd)

$$\begin{aligned} h_{opt}(x_0) &= \left[\frac{(p+1)!^2(2\nu+1)\int K^{\star 2}_{\nu,p}(t)\,dt}{2(p+1-\nu)\left(\int t^{p+1}K^{\star}_{\nu,p}(t)\,dt\right)^2}\right]^{1/(2p+3)} \left[\frac{\sigma^2(x_0)}{\{m^{(p+1)}(x_0)\}^2 f_X(x_0)}\right]^{1/(2p+3)} n^{-1/(2p+3)} \\ &= C_{\nu,p}(K)\left[\frac{\sigma^2(x_0)}{\{m^{(p+1)}(x_0)\}^2 f_X(x_0)}\right]^{1/(2p+3)} n^{-1/(2p+3)}. \end{aligned}$$

The constant $C_{\nu,p}(K)$ is easy to calculate and Table 4.2 lists some of them for different ν and p. If we want a global

ν	p	Gaussian	Uniform	Epanechnikov
0	1	0.776	1.351	1.719
0	3	1.161	2.813	3.243
1	2	0.884	1.963	2.275
2	3	1.006	2.604	2.893

Table 4.2: Constant $C_{\nu,p}(K)$ for different kernel functions

measure of error, we could opt for a weighted MISE given by

$$\int (\text{bias}[\widehat{m}_\nu(x_0)|\mathbb{X}]^2 + \textbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}])w(x)\,dx,$$

with $w \geq 0$ some weight function, leads to a theoretical optimal constant bandwidth. Usually w is taken to be the design density f_X. It can be shown that an asymptotically optimal constant bandwidth is given by (with $w = f_X$)

$$h_{opt} = C_{\nu,p}(K)\left[\frac{\int \sigma^2(x)\,dx}{\int \{m^{(p+1)}(x)\}^2 f_X(x)\,dx}\right]^{1/(2p+3)} n^{-1/(2p+3)}. \tag{4.32}$$

In the latter it is assumed that the integrals are finite and the that the denominator does not vanish.

In practice (4.32) is not usable since it depends on several unknown quantities. In what follows we present a simple way of estimating these quantities. The most simple way to do this is by fitting a polynomial of order $p + 3$ **globally** to $m(x)$ (via ordinary least squares), leading to the parametric fit

$$\widetilde{m}(x) = \widetilde{\alpha}_0 + \cdots + \widetilde{\alpha}_{p+3}x^{p+3}.$$

The choice of a global fit results results in a derivative function $\widetilde{m}^{(p+1)}(x)$ which is of quadratic form, allowing for certain flexibility in estimating the curvature. Assuming a constant error variance $\sigma^2(x) = \sigma^2$, then we can estimate this e.g., in a model-free way (Hall et al., 1990)

$$\widetilde{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n-2}(0.809Y_{[i]} - 0.5Y_{[i+1]} - 0.309Y_{[i+2]})^2,$$

where $Y_{[j]}$ denotes the jth order observation corresponding to the ordered $X_{[j]}$. Other model-free error variance estimators, not necessarily restricted to the one dimensional case, can be found in Devroye et al. (2013) and De Brabanter et al. (2014). Of course, one could also use a model based estimator. Further assume that $x \in [a,b]$, then (4.32) can be estimated by

$$\widetilde{h}_{opt} = C_{\nu,p}(K)\left[\frac{\widetilde{\sigma}^2 \int_a^b dx}{\int \{\widetilde{m}^{(p+1)}(x)\}^2 f_X(x)\,dx}\right]^{1/(2p+3)} n^{-1/(2p+3)}.$$

Using the strong law of large numbers, the final rule of thumb bandwidth selector $\widetilde{h}_{\text{ROT}}$ (for $p - \nu$ odd) is given by

$$\boxed{\widetilde{h}_{\text{ROT}} = C_{\nu,p}(K)\left[\frac{\widetilde{\sigma}^2(b-a)}{\sum_{i=1}^n \widetilde{m}^{(p+1)}(X_i)^2}\right]^{1/(2p+3)}} \tag{4.33}$$

Although (4.33) is derived under certain conditions, it can be applied in situations where these conditions are not strictly fulfilled.

4.4.4 Design adaptation property

The bias and variance expressions in (4.21) and (4.24) are obtained under the random design model, but remain valid for fixed designs. Hence, local polynomial estimators adapt to both random and fixed designs. This is in contrast with the Gasser-Müller estimator which cannot adapt to random designs: the unconditional variance is higher by a factor 1.5 for random designs. More explanation about this statement can be found in Mack and Müller (1989).

Recall that for $p - \nu$ even, additional terms arise in the asymptotic conditional bias. For example, when estimating the regression function $m(x_0)(\nu = 0)$, an extra term $m'(x_0)f'(x_0)/f(x_0)$ appears in the asymptotic bias of the Nadaraya-Watson estimator (4.25). The bias of this estimator depends on the intrinsic part $m''(x_0)$ interplaying with the artifact $m'(x_0)f'(x_0)/f(x_0)$. Keeping $m''(x_0)$ fixed, we first remark that in the highly clustered (asymmetric) design where $|f'(x_0)/f(x_0)|$ is large, the bias of the Nadaraya-Watson estimator can be large. Thus this estimator cannot adapt to highly clustered designs. Similar artifacts hold true for polynomial fits of an even order $p - \nu$. Local polynomial fitting with $p - \nu$ odd however rules out such artifacts and results in design-adaptive estimators.

4.4.5 Automatic boundary carpentry

In applications design points always have a bounded support. For estimating $m^{(\nu)}(x_0)$, with x_0 a point close to the boundary, the local neighborhood $x_0 \pm h$ can lie outside the design region. Hence, certain symmetric moment conditions, valid for all interior points, are no longer valid for x_0 in a boundary region, causing a large boundary bias for most of the smoothing techniques. If a bandwidth is chosen to be 25% of the data range, then for about 50% of the data range the local neighborhood will lie partly outside the design region. Hence the boundary region is about 50% of the whole data range. In higher dimensions these figures are even more striking, reflecting even more severe problems. Since many of the smoothing techniques show the aforementioned bias problem at the boundary, considerable efforts have been devoted to methods for correcting this boundary bias. Two popular approaches are boundary kernel methods and reflection methods. But none of these methods are as simple and as efficient as the automatic boundary correction when using local polynomial fitting. Without loss of generality we assume that the design density has a bounded support [0,1]. A left boundary point is thought of as being of the form $x := ch$, with $c > 0$, whereas a right boundary point is of the form $x = 1 - ch$.

The behavior of the estimator $\widehat{m}^{(\nu)}(x_0)$ for points x_0 at the interior of the support has been studied in the previous sections. In this section we address the question of how local polynomial estimators behave at boundary points. For most regression smoothers the rate of convergence at boundary points is slower than that at points in the interior. In the literature one refers to this problem as boundary effects or edge effects. These effects are visually very disturbing in practice, and in addition they can play a dominant role in theoretical analysis. Hence, in the case of boundary effects there is a strong request for boundary modifications, in order to overcome the problem.

The aforementioned automatic boundary carpentry can be easily seen from the representation of the local polynomial estimator in terms of an equivalent kernel (4.27). Consider a left boundary point $x = ch$. Similar as before, the finite moments are

$$S_{n,j} = nh^j f(0+)\mu_{j,c}\{1 + o_p(1)\}$$

where $\mu_{j,c} = \int_{-c}^{+\infty} u^j K(u)\,du$. This leads to the following equivalent kernel at the boundary

$$K^{\star}_{\nu,c,p}(t) = \varepsilon^T_{\nu+1} S_c^{-1}(1,t,\ldots,t^p)^T K(t) \quad \text{with} \quad S_c = (\mu_{j+l},c)_{0\le j,l\le p}.$$

This equivalent kernel differs from $K^{\star}_{\nu,p}$ only in the matrix S. This reflects the automatic adaptation to the boundary. Figure 4.4 shows the Epanechnikov kernel and some of its equivalent kernels for some boundary points and for various values of p and ν.

Theorem 4.3 (Fan and Gijbels (1996)) *Assume that $f(0+) > 0$ and that $f(\cdot), m^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are right continuous at the point 0. Then, the conditional MSE of the estimator $m^{(\nu)}(x)$ at the left boundary point $x = ch$ is given by*

$$\left[\left\{\int_{-c}^{+\infty} t^{p+1}K^{\star}_{\nu,c,p}(t)\,dt\right\}^2 \left\{\nu!\frac{m^{(p+1)}(0+)}{(p+1)!}\right\}^2 h^{2(p+1-\nu)} + \int_{-c}^{+\infty} K^{\star 2}_{\nu,c,p}(t)\,dt\,\frac{\nu!^2\sigma^2(0+)}{f(0+)nh^{1+2\nu}}\right]\{1 + o_p(1)\}.$$

4.4.6 Which order of polynomial fit?

Fitting polynomials of higher order leads to a possible reduction of the bias, but on the other hand also to an increase of the variability, caused by introducing more local parameters. Intuitively it is clear that in a flat non-sloped region a local constant or linear fit is recommendable, whereas at peaks and valleys local quadratic and cubic fits are preferable. Thus, for a very spatially inhomogeneous curve, the order of the polynomial approximation should be adjusted to the curvature of the unknown regression function. We would like to mention, however, that for many applications

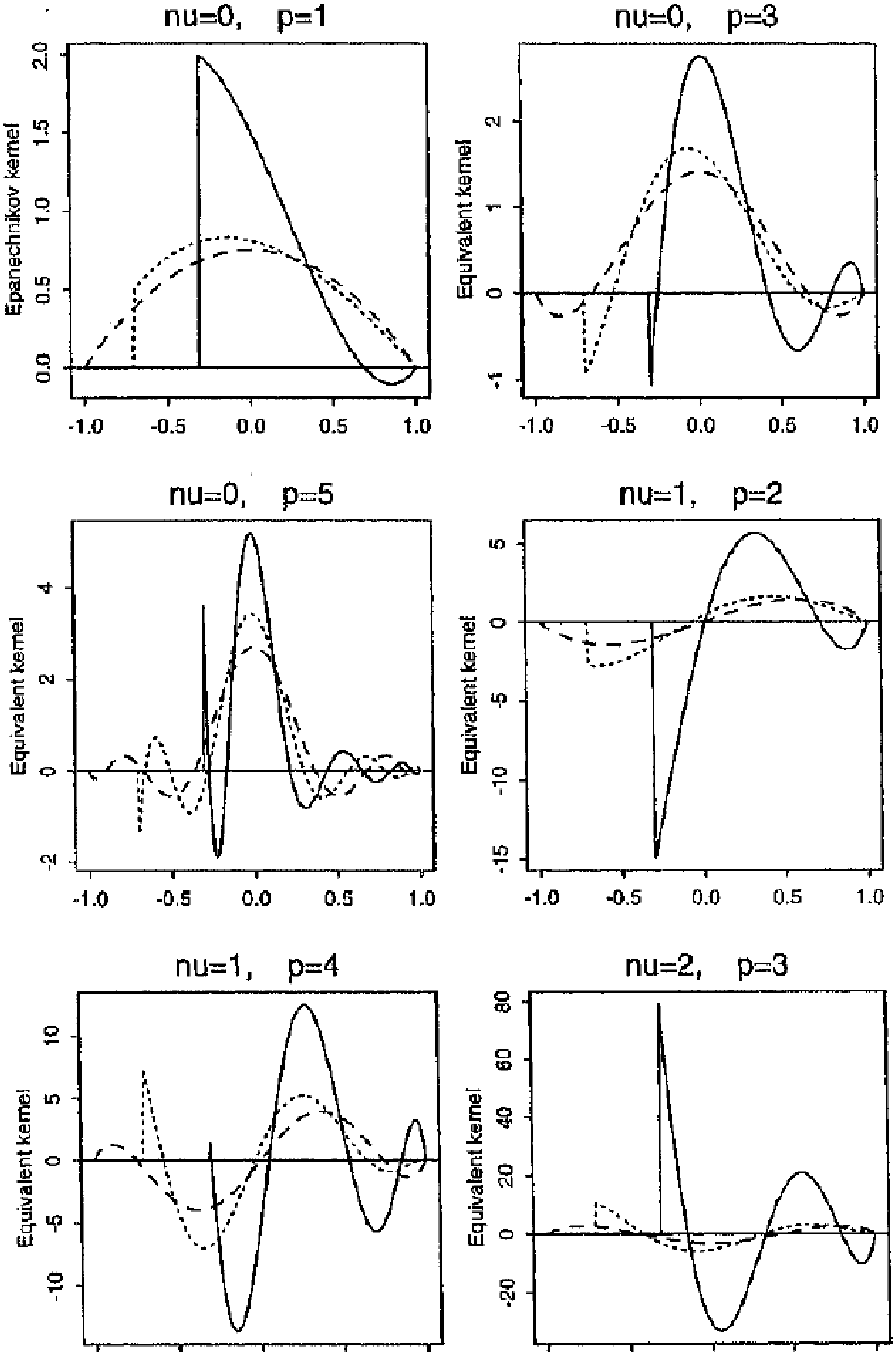


Figure 4.4: The Epanechnikov kernel and its equivalent kernels at the boundary points $c = 0.3$ (solid line) and $c = 0.7$ (dotted line) and interior points $c \geq 1$ (dashed line) for various values of p and ν. Taken from Fan and Gijbels (1996).

the choice $p = \nu + 1$ suffices. A variable order selection carries a possible price including the stochastic element introduced by the selection procedure and computational costs. Such an order selection procedure is mainly proposed for recovering spatially inhomogeneous curves.

Increase in variability

Suppose we fit a local polynomial of order p in order to estimate the derivative $m^{(\nu)}(x_0)$. The bias of such a fit will be of order $h^{p+1-\nu}$ (for $p-\nu$ odd) or of order $h^{p+1-\nu}$ (for $p-\nu$ even) as can be seen from Theorem 4.2. So, higher order polynomial approximations result in a smaller order of the bias. But let's see what happens to the variance if we increase the order of the approximation. The asymptotic variance of the estimator $\widehat{m}_\nu(x_0)$ is given by (see Theorem 4.1)

$$\begin{aligned}\mathbf{Var}[\widehat{m}_\nu(x_0)|\mathbb{X}] &= \varepsilon_{\nu+1}^T S^{-1} S^\star S^{-1} \varepsilon_{\nu+1} \frac{\nu!^2 \sigma^2(x_0)}{f_X(x_0) n h^{1+2\nu}} \{1 + o_p(1)\} \\ &= \left(\int K_{\nu,p}^{\star 2}(t)\, dt \right) \frac{\nu!^2 \sigma^2(x_0)}{f_X(x_0) n h^{1+2\nu}} \{1 + o_p(1)\}\end{aligned}$$

which is of order $n^{-1}h^{-(1+2\nu)}$ and hence not affected by the order of the polynomial fit. But let's take a look at the constant terms. To simplify, take $\nu = 0$. This is by no means a restriction; conclusions drawn for the case of estimating the regression function carry over to the estimation of its derivative functions. The asymptotic variance of the estimator for the regression function is of the form

$$V_p \frac{\sigma^2(x_0)}{f_X(x_0) n h},$$

where V_p is the (1,1)th element of the matrix $S^{-1}S^\star S^{-1}$. Table 4.3 shows how much the variance increases with the order of the approximation for several kernel functions, relative to the variance of the Nadaraya-Watson estimator (local constant fit, $p = 0$). Table 4.3 summarizes the values for V_p/V_0 for various commonly used kernel functions. Note that there is no loss in terms of asymptotic variance by doing a local linear instead of a local constant fit. This remark applies to the comparison of any even order approximation with its consecutive odd order approximation. However, the asymptotic variance increases when moving from an odd order approximation to its consecutive even order approximation. For example in the case of the Epanechnikov kernel, the variance increases by a factor of 2.0833 when a local quadratic instead of a local linear fit is used. The increase in variability is the most pronounced for the uniform kernel. A graphical representation of Table 4.3 is given in Figure 4.5.

p	Gaussian	Uniform	Epanechnikov	Biweight	Triweight
1	1	1	1	1	1
2	1.6876	2.25	2.0833	1.9703	1.9059
3	1.6876	2.25	2.0833	1.9703	1.9059
4	2.2152	3.5156	3.1550	2.8997	2.7499
5	2.2152	3.5156	3.1550	2.8997	2.7499
6	2.6762	4.7852	4.2222	3.8133	3.5689
7	2.6762	4.7852	4.2222	3.8133	3.5689
8	3.1224	6.0562	5.2872	4.7193	4.3753
9	3.1224	6.0562	5.2872	4.7193	4.3753
10	3.5704	7.3281	6.3509	5.6210	5.1744

Table 4.3: Increase of the variability with the order of the polyno1nial approximation p. Results taken from Fan and Gijbels (1995).

It's an odd world

It now becomes clear that odd order fits are preferable. A fit of odd order $2p+1$ introduces an extra parameter in comparison with a fit of even order $2p$, but there is no increase of variability caused by this. With this extra parameter

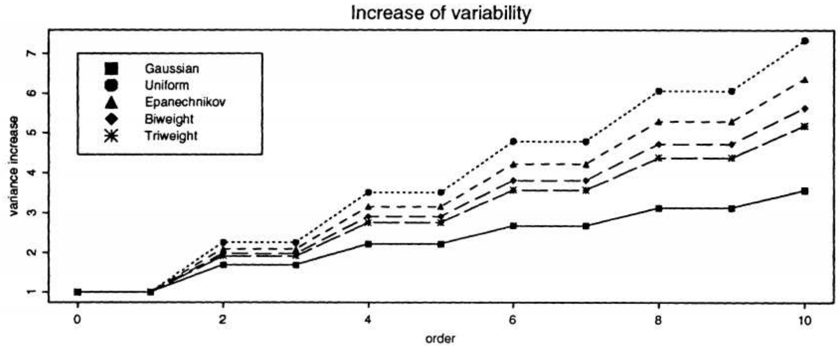


Figure 4.5: Increase of the variability with the order of the polynomial approximation p. Taken from Fan and Gijbels (1995).

an opportunity is created for a significant bias reduction especially in the boundary regions and in highly clustered design regions. Moreover, even order fits suffer from low efficiency, as was established by Fan (1993) for the local constant fit. In addition serious boundary effects appear when using even order fits; this contrasts with odd order fits which have the nice boundary adaptive property. The above asymptotic considerations demonstrate that it is "an odd world": odd order polynomial fits are preferable to even order polynomial fits. An odd world indeed, and that is why we say that $p - \nu$ odd is natural.

4.5 Data driven bandwidth choices: Cross-validation

4.5.1 Leave-one-out cross-validation (LOO-CV)

Consider the integrated squared error (ISE) as a measure of accuracy for the estimator $\widehat{m}_h(x)$, let h denote the bandwidth of the kernel, and given a data set $\mathcal{D}_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$. Note that h should be strictly positive. The main idea is to construct an estimate $\widehat{m}_h$ such that the ISE is small. Let F_X denote the distribution over the input space, then

$$\int |\widehat{m}_h(x) - m(x)|^2 \, dF_X(x) = \int m^2(x)\, dF_X(x) + \int \widehat{m}_h^2(x)\, dF_X(x) - 2\int \widehat{m}_h(x) m(x)\, dF_X(x). \tag{4.34}$$

Since the first term in (4.34) is independent of h, minimizing (4.34) is equivalent to minimizing

$$\int \widehat{m}_h^2(x)\, dF_X(x) - 2\int \widehat{m}_h(x) m(x)\, dF_X(x). \tag{4.35}$$

In practice this would be impossible to compute since this quantity depends on the unknown real-valued (true) function m and the density f. The first term of (4.35) can be entirely computed from the data $\mathcal{D}_n$ and the second term can be written as

$$\int \widehat{m}_h(x) m(x)\, dF_X(x) = \mathbf{E}[\widehat{m}_h(X) m(X) | \mathcal{D}_n]. \tag{4.36}$$

If one estimates (4.36) by its empirical version $n^{-1}\sum_{i=1}^n Y_i \widehat{m}_h(X_i)$ the selection will be a biased estimator of the ISE. The bias is due to the fact that the observation Y_i is used in $\widehat{m}_h(X_i)$ to predict itself. However, there exist several methods to find an unbiased estimate of the ISE e.g. plug-in methods, leave-one-out (LOO) technique and a modification so that bias cancels out asymptotically. Here we will use the LOO technique in which one observation is left out. Therefore, a better estimator for (4.36) instead of its straight empirical version is

$$\frac{1}{n}\sum_{i=1}^n Y_i \widehat{m}_h^{(-i)}(X_i), \tag{4.37}$$

where $\widehat{m}_h^{(-i)}(X_i)$ denotes the LOO estimator with point i left out from the training. Similarly, the first term of (4.35) can be written as

$$\frac{1}{n}\sum_{i=1}^n \left|\widehat{m}_h^{(-i)}(X_i)\right|^2. \tag{4.38}$$

From (4.37) and (4.38), the LOO-CV function is given by

$$\text{LOO-CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{m}_h^{(-i)}(X_i)\right|^2.$$

The LOO cross-validated selection of h is

$$\boxed{\widehat{h}_{\text{LOO-CV}} = \arg\min_h \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{m}_h^{(-i)}(X_i)\right|^2}$$

It is interesting to know that this LOO-CV criterion actually estimates the following quantity (under independent errors)

$$\frac{1}{n}\sum_{i=1}^{n}|m(X_i) - \widehat{m}_h(X_i)|^2 + \frac{1}{n}\sum_{i=1}^{n}\sigma^2(X_i).$$

4.5.2 v-fold Cross-Validation

In general there is no reason that training sets should be of size $n-1$ as in the LOO-CV case. There is the possibility that small perturbations, when single observations are left out, make LOO-CV(h) too variable, if the fitted values $\widehat{m}_h(x)$ do not depend smoothly on the empirical distribution $\hat{F}_n$ or if the loss function $L(Y, \widehat{m}_h(X))$ is not continuous. These potential problems can be avoided, to a large extent, by leaving out groups of observations, rather than single observations. Also, it offers a computational advantage since we do not have to compute n estimates but only v in v-fold CV. The latter plays an important role for large data sets. Note that v-fold CV with $v = n$ is LOO-CV.

The use of groups have the desired effect of reducing variance, but at the cost of increasing bias. According to Beran (1984) and Burman (1989) the bias of v-fold CV yields

$$a_0\left[(v-1)^{-1}n^{-1}\right].$$

For LOO-CV the bias is of order $O(n^{-2})$, but when v is small the bias term is not necessarily small. Therefore, the use of 2-fold CV is never recommended. The term a_0, depending on the loss function L used in the CV procedure and the empirical distribution $\hat{F}_n$, is of the order of the number of effective parameters being estimated. As a result, if the number of effective parameters is not small, the v-fold CV is a poor estimate of the prediction error. However, there are adjustments possible to reduce the bias in v-fold CV, see e.g. Burman (1989, 1990); Tibshirani and Tibshirani (2009); Arlot and Celisse (2010). These adjustments to the v-fold CV procedure reduce the bias to

$$a_1\left[(v-1)^{-1}n^{-2}\right],$$

for some constant a_1 depending on the loss function L used in the CV procedure and the empirical distribution $\hat{F}_n$.

Precise understanding of how **Var**[v-fold CV] depends on the splitting scheme is rather complex since the number of splits (folds) v is linked with the number of points used as validation. Furthermore, the variance of CV strongly depends on the framework and on the stability of the algorithm. Therefore, radically different results have been obtained in different frameworks, in particular on the value of v for which the v-fold CV estimator has a minimal variance, see e.g. Burman (1989) and Hastie et al. (2009, Chapter 7).

What is a suitable value for v? Davison and Hinkley (2003) have suggested the following rule of thumb. Take $v = \min(\sqrt{n}, 10)$, because taking $v > 10$ maybe computationally too expensive while taking groups of size at least $\sqrt{n}$ should perturb the data sufficiently to give a small variance of the estimate.

4.6 Local polynomial regression in R

4.6.1 Toy example

Create a toy example data set (200 points and normal random noise with $\sigma = 0.2$) and load the appropriate library:

```
> set.seed(729)
> x <- seq(0, 1, length.out = 200)
> y <- (sin(2*pi*(x-0.5)))^2 + rnorm(200, 0, 0.2)
> d <- data.frame(x,y)
> library(locpol)
```

Suppose we want to fit a local linear regression ($p = 1$) with Gaussian kernel to the data. We choose the bandwidth with the rule of thumb selector and via cross-validation:

```
> p <- 1
# Rule of Thumb
> hrot <- thumbBw(d$x, d$y, deg = p, kernel = gaussK)

0.03095715

# Cross-validation
> hcv <- regCVBwSelC(d$x,d$y,deg=p,kernel=gaussK,interval=c(seq(0.001,0.5,length.out=1000)))

0.03165582
```

In this case both bandwidths are not that different. This is of course not always the case. In practice it is recommended to use cross-validation. The problem with cross-validation is that it also can have multiple local minima. Therefore, it is recommended to specify a range or interval to look for a (local) minimum. Both methods assume that the data is i.i.d! Finally, we can fit the local linear regression estimate (for both bandwidths) and evaluate in each point of the data set by

```
> r_rot <- locpol(y~x, d, deg = p, bw = hrot, kernel = gaussK, xeval = d$x)
> r_cv <- locpol(y~x, d, deg = p, bw = hcv, kernel = gaussK, xeval = d$x)
```

Figure 4.6 shows the local linear estimate based on bandwidth $\widehat{h}_{\text{ROT}}$ (full line) and the true regression function (dashed line).

```
> plot(x, y)
> lines(x, fitted(r_rot), lwd = 2, col = "red")
> lines(x, (sin(2*pi*(x-0.5)))^2, lwd = 2, col = "blue", lty = 2)
```

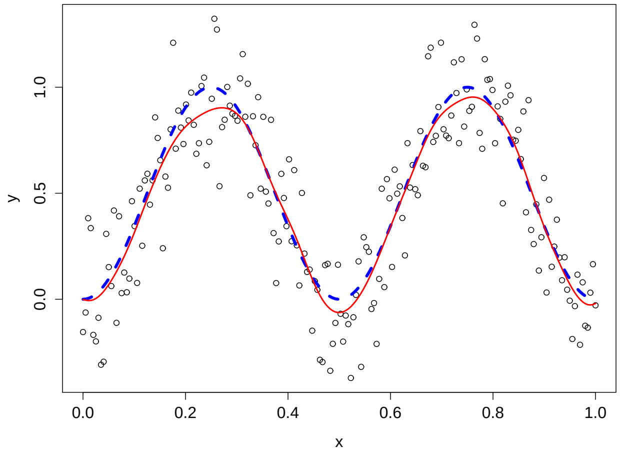


Figure 4.6: Local linear estimate based on bandwidth $\widehat{h}_{\text{ROT}}$ (full line) and the true regression function $\sin^2[2\pi(x-0.5)]$ (dashed line).

4.6.2 LIDAR data example

The lidar data frame has 221 observations from a light detection and ranging (LIDAR) experiment and can be found in the R package *SemiPar*. Loading the data set and finding the bandwidth for a local cubic fit ($p = 3$) based on a Gaussian kernel is done as follows:

```
> library(SemiPar)
> data(lidar)
> d <- data.frame(x = lidar$range, y = lidar$logratio)
> p <- 3
> hrot <- thumbBw(lidar$range, lidar$logratio, deg = p, kernel = gaussK)

10.89228

> hcv <- regCVBwSelC(lidar$range, lidar$logratio, deg = p, kernel = gaussK,...
                   interval=c(seq(22, 26, length.out = 10000)))

24.97158

> r_rot <- locpol(y˜x, d, deg = p, bw = hrot, kernel = gaussK, xeval = d$x)
> r_cv <- locpol(y˜x, d, deg = p, bw = hcv, kernel = gaussK, xeval = d$x)
> plot(d$x, d$y, xlab = "Range", ylab = "Logratio")
> lines(d$x, fitted(r_rot), lwd=4, col = "red", lty = 2)
> lines(d$x, fitted(r_cv), lwd=2, col = "blue")
```

The result is given in Figure 4.7. It is clear that the estimate based on $\widehat{h}_{\text{ROT}} = 10.89228$ is slightly more wiggly than the one based on $\widehat{h}_{\text{LOO-CV}} = 24.97158$. Do you know why? *Hint:* Check the assumptions on which the ROT is based and compare with LOO-CV!

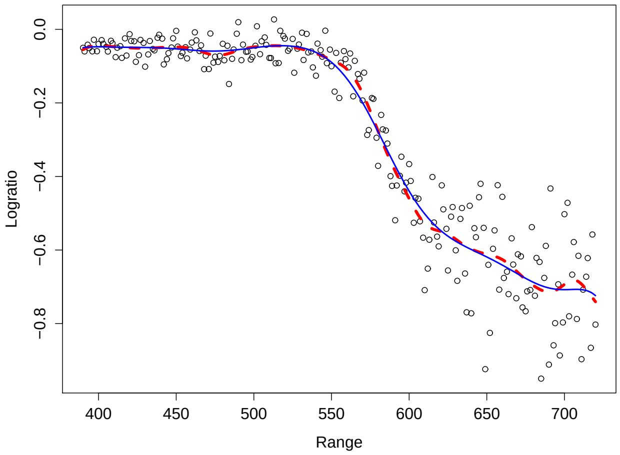


Figure 4.7: Local cubic estimate based on bandwidth $\widehat{h}_{\text{ROT}}$ (dashed line) and $\widehat{h}_{\text{LOO-CV}}$ (full line).