10 Nonparametric Regression

The notes in this section are taken from a lecture course on this topic that I gave previously. We will not cover all of this material in one lecture, but the detailed notes are provided so that you can learn about more about the topics that interest you.

10.1 Introduction

10.1.1 Difference between parametric and nonparametric regression

The basis for regression is a set of observations of pairs of variables (X_i, Y_i), $i = 1, \ldots, n$. We are interested in finding a connection between X and Y. We assume that Y is random, but X can be either random or fixed; we focus mostly on the case that the X_i's are fixed. In parametric regression we assume a particular type of dependence of Y on X (e.g. linear regression: $\mathbb{E}Y = AX$, log-linear regression $\log(\mathbb{E}Y) = AX$, etc). In other words, we assume a priori that the unknown regression function f belongs to a parametric family $\{g(x, \theta) : \theta \in \Theta\}$, where $g(\cdot, \cdot)$ is a given function, and $\Theta \subset \mathbb{R}^k$. Estimation of f is the equivalent to estimation of the parameter vector θ.

In nonparametric regression, by contrast, we do not want to make any assumption about how $\mathbb{E}Y$ depends on X, but want to fit an arbitrary functional dependence. We assume that we observe a function with error:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

Often the errors are assumed to be normally distributed, $\varepsilon_i \sim N(0, \sigma^2)$, independently. The aim is to estimate the unknown function f.

In nonparametric estimation it is usually assumed that f belongs to some large class $\mathcal{F}$ of functions. For example, $\mathcal{F}$ can be the set of all the continuous functions or the set of all smooth (differentiable) functions. For proving certain properties of estimators, we will consider sets of functions with k derivatives, which are called Hölder spaces of functions.

We will described several different approaches to nonparametric regression — kernel smoothing, spline smoothing, general additive models and wavelet estimation.

10.1.2 Nonparametric regression model

Throughout this chapter we will assume the following model of nonparametric regression:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

with independent errors $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and a function $f : [0, 1] \to \mathbb{R}$.

Now suppose that we observe data (x_i, y_i), $i = 1, \ldots, n$, which is a realisation of iid random variables (X_i, Y_i). The aim is to estimate the unknown function $f(x) = \mathbb{E}(Y_i | X_i = x)$, namely to construct an estimator $\widehat{f}_n(x)$ for all $x \in [0, 1]$ which is consistent and efficient, and to be able to test hypotheses about $f(x_0)$ for a fixed x_0 and about $f(x)$ for all x simultaneously.

The maximum likelihood estimator (MLE) of $f(x)$ gives estimates of f only at points x_i where we observe the data: $\widehat{f}(x_i) = y_i$. Since $\mathbb{E}[\varepsilon_i] = 0$, this estimator is unbiased at x_i, as $\mathbb{E}\widehat{f}(x_i) = \mathbb{E}Y_i = f(x_i)$. However, the MLE (and the model) does not give any information about $f(x)$ for $x \neq x_i$. The model is not fully identifiable hence some additional assumptions about f are needed. A key assumption we will make about f that it is smooth.

10.1.3 Estimators

There are two major approaches to nonparametric estimation.

1. **Smoothing**: fitting a flexible smooth curve to data. We will consider two methods: kernel smoothing and spline smoothing. The main question in this context is how smooth should this curve be, and do we have to decide that in advance, or can we let the data to decide?

2. **Orthogonal projection estimation**: represent the regression function f as a series in an orthogonal basis, and estimate the coefficients from the data. We will consider wavelet bases. Wavelets can be spiky, so they are well suited for modelling not very smooth functions, e.g., with jumps or sharp spikes. The main question is how to estimate the coefficients, so that the function estimate is neither too smooth nor too spiky.

10.1.4 Consistency

The key requirement for any estimator is consistency, that is, the more data we have, the closer the estimator is to the function of interest. We encountered consistency in the context of estimators of parameters, and there is a corresponding definition for functions.

Definition 10.1. *$\widehat{f}_n$ is a (weakly) consistent estimator of f in distance d based on n observations iff*

$$\forall \epsilon > 0, \quad \mathbb{P}(d(\widehat{f}_n, f) > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

In the rest of this chapter, when we refer to consistency we will mean weak consistency. We consider two distances on function spaces $d(\widehat{f}_n, f)$.

1) Pointwise at x_0 (local): $d(\widehat{f}_n, f) = |\widehat{f}_n(x_0) - f(x_0)|$, for some $x_0 \in [0,1]$.

2) Integrated (global) : $d(\widehat{f}_n, f) = ||\widehat{f}_n - f||_2 = \sqrt{\int_0^1 (\widehat{f}_n(x) - f(x))^2 dx}$.

   Here $||\cdot||_2$ is defined by

   $$||g||_2^2 \stackrel{def}{=} \int_0^1 [g(x)]^2 dx.$$

   It is a norm in Hilbert space $L^2[0,1] = \{g : [0,1] \to \mathbb{R} \text{ such that } ||g||_2 < \infty\}$.

Markov's inequality is a tool to verify consistency:

$$\mathbb{P}(d(\widehat{f}_n, f) > \epsilon) \leq \epsilon^{-2} \mathbb{E}[d(\widehat{f}_n, f)^2].$$

For these distances, $\mathbb{E}[d(\widehat{f}_n, f)]^2$ has particular names.

1) Mean squared error (MSE):

$$\text{MSE}(\widehat{f}_n(x_0)) = \mathbb{E}[|\widehat{f}_n(x_0) - f(x_0)|^2] = v(x_0) + [b(x_0)]^2$$

2) Mean integrated squared error (MISE):

$$\text{MISE}(\widehat{f}_n) = \mathbb{E}[||\widehat{f}_n - f||^2] = \mathbb{E}\left[\int_0^1 |\widehat{f}_n(x) - f(x)|^2 dx\right] = \int_0^1 v(x)dx + \int_0^1 [b(x)]^2 dx,$$

   where $b(x) = \text{bias}(\widehat{f}(x)) = \mathbb{E}\left[\widehat{f}(x)\right] - f(x)$ and $v(x) = \text{Var}(\widehat{f}(x))$ are the bias and the variance of $\widehat{f}(x)$.

Therefore, M(I)SE($\widehat{f_n}$) $\to 0$ as $n \to \infty$ implies consistency in the corresponding distance. We will also study the rate of convergence of the estimators, that is, how fast MISE and MSE decrease to 0 as a function of sample size n.

10.1.5 Notation

The indicator function of a set A is

$$\mathbf{1}_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A. \end{cases}$$

Informally, we will also write $\mathbf{1}(|x| \leq 1)$ for $\mathbf{1}_{|x|\leq 1}(x)$.

Denote the support of a function g, the set of arguments where g is nonzero, by

$$\text{supp}(g) = \{x : g(x) \neq 0\}.$$

10.2 Kernel estimators

10.2.1 Designs

Definition 10.2. *A set $(X_1, \ldots, X_n)$ is called a design*

Definition 10.3. *A design $(X_1, \ldots, X_n)$ is called fixed if the values $x_1, \ldots, x_n$ are non random*

Example 10.1. *An equispaced (regular) design $x_1 < x_2 < \ldots < x_n$ is a fixed design such that $x_i - x_{i-1} = 1/n$, e.g. $x_i = i/n$;$x_i = \frac{i-1}{n}$; $x_i = \frac{1}{2n} + \frac{i-1}{n}$.*

Definition 10.4. *A design $(X_1, \ldots, X_n)$ is called random iff $X_1, \ldots, X_n$ are iid random variables, $X_i \sim p(x)$.*

Example 10.2. *$x_i \sim U[0, 1]$ with $p(x) = 1$ for $x \in [0, 1]$.*

10.2.2 Nadaraya-Watson estimator

Definition 10.5. *A function $K(x)$ is called a kernel iff $\int_{-\infty}^{\infty} K(x)d(x) = 1$.*

If $K(x) \geq 0$, $K(x)$ is a probability density.

Definition 10.6. *If $K(x) = K(-x)$, then $K(x)$ is a symmetric kernel.*

Definition 10.7. *A kernel K has order m iff $\int_{-\infty}^{\infty} x^\ell K(x)dx = 0$ for all $\ell = 1, 2, \ldots, m-1$ and $\int_{-\infty}^{\infty} x^m K(x)dx \neq 0$.*

If K is symmetric, then K has order ≥ 2.

Example 10.3. *All these kernels are symmetric of order 2, except the last one.*

a) Uniform (box, rectangular) kernel $K(x) = \frac{1}{2}\mathbf{1}(|x| \leq 1)$.

b) Triangular kernel $K(x) = (1 - |x|)\mathbf{1}(|x| \leq 1)$.

c) Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

d) Cosine kernel $K(x) = \frac{\pi}{4}\cos(\pi x/2)\mathbf{1}(|x| \le 1)$.

e) Sinc kernel $K(x) = \frac{\sin(\pi x)}{\pi x}$. *This kernel has infinite order, since* $\int_{-\infty}^{+\infty} \sin(\pi x)x^{m-1}dx = 0$ *for all integer* $m \ge 1$.

Remark 10.1. *If* $K(x)$ *is a kernel, then* $K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$ *is also a kernel.* h *is called the* **bandwidth**.

Example 10.4. *If* $K(x) = \frac{1}{2}\mathbf{1}(|x| \le 1)$ *is a kernel then* $K(x) = \frac{1}{4}\mathbf{1}(|x| \le 2)$ *is a kernel.*

Definition 10.8. ***The Nadaraya-Watson Estimator***

$$\widehat{f}_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x)}{\sum_{j=1}^n K_h(X_j - x)}, \text{ when } \sum_{i=1}^n K_h(X_i - x) \neq 0,$$

otherwise $\widehat{f}_n^{NW}(x) = 0$.

Motivation for the Nadaraya-Watson estimator.

Recall that $f(x)$ can be written as

$$f(x) = \mathbb{E}(Y_i \mid X_i = x) = \int yp(y \mid x)dy = \int \frac{yp(x,y)}{p(x)}dy.$$

Consider the following kernel density estimators:

$$\widehat{p}_n(x) = \frac{1}{n}\sum_{i=1}^n K_h(x_i - x), \quad \widehat{p}_n(x,y) = \frac{1}{n}\sum_{i=1}^n K_h(x_i - x)K_h(y_i - y). \tag{124}$$

Plugging $\widehat{p}_n(x)$ and $\widehat{p}_n(x,y)$ into $\mathbb{E}(Y_i|X_i = x)$, we have

$$\widehat{f}_h(x) = \int_{-\infty}^{\infty} \frac{y\widehat{p}_n(x,y)}{\widehat{p}_n(x)}dy.$$

Now we simplify the numerator, assuming that the kernel is symmetric

$$\int_{-\infty}^{\infty} y\widehat{p}_n(x,y)dy = \frac{1}{n}\int_{-\infty}^{\infty} y\sum_{i=1}^n K_h(x_i - x)K_h(y_i - y) = \frac{1}{n}\sum_{i=1}^n K_h(x_i - x)\int_{-\infty}^{\infty} yK_h(y - y_i)dy,$$

and the last integral is

$$\begin{aligned}\frac{1}{h}\int_{-\infty}^{\infty} yK\left(\frac{y - y_i}{h}\right)dy &= [z = (y - y_i)/h] = \int_{-\infty}^{\infty}(hz + y_i)K(z)dz \\ &= y_i\int_{-\infty}^{\infty} K(z)dz + h\int_{-\infty}^{\infty} zK(z)dz = y_i\end{aligned}$$

assuming that the order of the kernel K is at least 2.

Therefore, an estimator of f can be written as

$$\widehat{f}_h^{NW}(x) = \frac{n^{-1}\sum_{i=1}^n K_h(x_i - x)y_i}{n^{-1}\sum_{i=1}^n K_h(x_i - x)}\mathbf{1}\left(\sum_{i=1}^n K_h(x_i - x) \neq 0\right)$$

which coincides with the **Nadaraya-Watson estimator**. Thus, we proved the following proposition.

Proposition 10.1. *If $K(x)$ is a symmetric kernel of order ≥ 2, under random design,*

$$\widehat{f}_h^{NW}(x) = \int_{-\infty}^{\infty} \frac{y\widehat{p}_n(x,y)}{\widehat{p}_n(x)} dy \, \mathbf{1}(\widehat{p}_n(x) \neq 0),$$

where $\widehat{p}_n(x)$ and $\widehat{p}_n(x,y)$ are kernel density estimators defined by (124).

If we know $p(x)$, then we can write $\widehat{f}(x) = \frac{1}{np(x)} \sum_{i=1}^n y_i K_h(x_i - x)$

If $X_i \sim U[0,1]$ then $p(x) = 1$ and $\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^n y_i K_h(x_i - x)$. This estimator also works for a regular fixed design.

Example 10.5. *Consider the box kernel $K(z) = 0.5\mathbf{1}(z \in [-1,1])$. Then, for x and h such that $|x_i - x| \leq h$ for at least one i, the Nadaraya-Watson estimator can be written as*

$$\widehat{f^{NW}}(x) = \frac{\sum_{i=1}^n h^{-1} Y_i K(\frac{x_i - x}{h})}{h^{-1} \sum_{i=1}^n \frac{1}{n} K(\frac{x_i - x}{h})} = \frac{\sum_{i=1}^n Y_i \frac{1}{2h} \mathbf{1}(|\frac{x_i - x}{h}| \leq 1)}{\sum_{i=1}^n \frac{1}{2h} \mathbf{1}(|\frac{x_i - x}{h}| \leq 1)} = \frac{\sum_{i:\,|x_i - x| \leq h} Y_i}{\sum_{i:\,|x_i - x| \leq h} 1}.$$

The Nadaraya-Watson estimator is an example of a linear estimator.

Definition 10.9. *Estimator $\widehat{f}(x)$ is called linear if it can be written as a linear function of y, i.e. $\widehat{f}(x) = \sum_{i=1}^n W_i(x) Y_i = W^T(x) Y$ where $Y = (y_1, \ldots, y_n)^T$, $W(x) = (w_1(x), \ldots, w_n(x))^T$ and $W(x)$ does not depend on y, only on $(x_1, \ldots, x_n)$.*

If an estimator is linear, then it is easy to find its distribution, and hence to construct a confidence interval and a confidence band (see Section 10.2.8).

Now we study the bias and the variance of the Nadaraya-Watson estimator in two frameworks, asymptotic as the sample size n grows to infinity, and for a fixed sample size.

10.2.3 Asymptotic properties of the Nadaraya-Watson estimator

As we saw in Section 10.1.4, to study consistency of an estimator, it is sufficient to study the asymptotic behaviour of its bias and variance. Thus, to study consistency of the NW estimator, we investigate asymptotic expressions for its bias and variance under the following assumptions.

Assumptions

1. Asymptotic: $n \to \infty, h \to 0$, $nh \to \infty$,
2. Design $x_1, \ldots, x_n$ is regular deterministic,
3. $x \in (0,1)$,
4. $\exists\, f''$,
5. Kernel:

$$\int_{-\infty}^{+\infty} xK(x)dx = 0, \quad 0 < \mu_2(K) \stackrel{def}{=} \int_{-\infty}^{+\infty} x^2 K(x)dx < \infty,$$

$$||K||_2^2 = \int_{-\infty}^{+\infty} [K(x)]^2 dx < \infty.$$

In particular, we assume that the unknown function f has a bounded second derivative and the kernel is of order 2.

A **key tool** to deriving the asymptotic expressions for the bias and the variance is approximation of a sum by an integral. Since the design (x_i) is regular deterministic, i.e. $x_i - x_{i-1} = 1/n$, for any function $g(x)$,

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \approx \int_0^1 g(z)dz.$$

In particular, the denominator of the NW estimator is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) \approx \int_0^1 K_h(z-x)dz &= \int_0^1 K\left(\frac{z-x}{h}\right) d\left(\frac{z}{h}\right) = \int_{\frac{0-x}{h}\to}^{\frac{1-x}{h}} K(v)dv \\ &\approx \int_{-\infty}^{+\infty} K(v)dv = 1\end{aligned}$$

since $n \to \infty$, $-x/h \to -\infty$ and $(1-x)/h \to +\infty$ as $h \to 0$. Here it is important that $x \neq 0$ and $x \neq 1$, that is, it is not at the boundary.

Asymptotic bias of the NW estimator: $b(x) \approx \frac{\mu_2(K)h^2}{2} f''(x)$.

$$\begin{aligned}
b(x) = \mathbb{E}\widehat{f}(x) - f(x) &= \sum_{i=1}^{n} w_i(x)[f(X_i) - f(x)] \quad \text{[Taylor Expansion]} \\
&\approx \sum_{i=1}^{n} w_i(x)\left[f(x) + f'(x)(X_i - x) + f''(x)\frac{(X_i-x)^2}{2} - f(x)\right] \\
&= \sum_{i=1}^{n} \frac{K_h(X_i - x)}{\sum_{j=1}^{n} K_h(X_j - x)}\left[f'(x)(X_i - x) + f''(x)\frac{(X_i-x)^2}{2}\right] \\
&\approx \frac{1}{n}\left[f'(x)\sum_{i=1}^{n}(X_i - x)K_h(X_i - x) + f''(x)\sum_{i=1}^{n} K_h(X_i - x)\frac{(X_i-x)^2}{2}\right] \\
&\approx f'(x)\int_0^1 (z-x)K_h(z-x)dz + f''(x)\int_0^1 K_h(z-x)\frac{(z-x)^2}{2}dz \\
&\approx f'(x)h\int_{-x/h}^{(1-x)/h} K(v)vdv + f''(x)\frac{h^2}{2}\int_{-x/h}^{(1-x)/h} K(v)v^2dv \\
&\approx f'(x)h\int_{-\infty}^{\infty} K(v)vdv + f''(x)\frac{h^2}{2}\int_{-\infty}^{\infty} K(v)v^2dv \\
&= \frac{\mu_2(K)h^2}{2} f''(x).
\end{aligned}$$

Asymptotic variance of the NW estimator: $v(x) \approx \frac{\sigma^2}{nh}||K||_2^2$:

$$\begin{aligned}
v(x) = \sigma^2 \sum_{i=1}^{n} [w_i(x)]^2 &= \sigma^2 \sum_{i=1}^{n} \frac{[K_h(X_i - x)]^2}{\left[\sum_{j=1}^{n} K_h(X_j - x)\right]^2} \\
&\approx_{\left\{\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) \approx 1\right\}} \frac{\sigma^2}{n^2} \sum_{i=1}^{n} [K_h(X_i - x)]^2 \\
\left\{\frac{1}{n}\sum_{i=1}^{n} \to \int_0^1\right\} &\approx \frac{\sigma^2}{n} \int_0^1 [K_h(z - x)]^2 dz = \frac{\sigma^2}{nh} \int_0^1 \left[K\left(\frac{z-x}{h}\right)\right]^2 d\left(\frac{z-x}{h}\right) \\
\left\{v = \frac{z-x}{h}\right\} &= \frac{\sigma^2}{nh} \int_{-x/h}^{(1-x)/h} [K(v)]^2 dv \approx \frac{\sigma^2}{nh} \int_{-\infty}^{\infty} [K(v)]^2 dv \\
&= \frac{\sigma^2}{nh} ||K||_2^2.
\end{aligned}$$

Therefore, **the asymptotic MISE (AMISE)** is:

$$\begin{aligned}
\text{AMISE} &= \int_0^1 \left[|b(x)|^2 + v(x)\right] dx \approx \int_0^1 \left[\frac{\mu_2(K) h^2}{2} f''(x)\right]^2 dx + \int_0^1 \frac{\sigma^2}{nh} ||K||_2^2 dx \\
&= \frac{||f''||_2^2}{4} h^4 [\mu_2(K)]^2 + \frac{\sigma^2}{n} \frac{||K||_2^2}{h}.
\end{aligned}$$

We are in general interested in having the "best" estimator of the function. This can be interpreted as finding h and K that minimise this error. We start with optimising over the kernel, introducing canonical kernels.

10.2.4 Canonical Kernel

Given a kernel $K(x)$ of order 2, consider a scale family of kernels:

$$\left\{K_\delta(x) = \frac{1}{\delta} K\left(\frac{x}{\delta}\right), \delta > 0\right\}$$

Definition 10.10. *The* **canonical bandwidth**, δ_0, *is defined by*

$$\delta_0 = \left(\frac{||K||_2^2}{[\mu_2(K)]^2}\right)^{\frac{1}{5}},$$

where $\mu_2(K) = \int_{-\infty}^{+\infty} x^2 K(x) dx$ *and* $||K||_2 = \sqrt{\int_{-\infty}^{+\infty} [K(x)]^2 dx}$.

Then, given a scale family of kernels $\left\{K_\delta(x) = \frac{1}{\delta} K\left(\frac{x}{\delta}\right), \delta > 0\right\}$, the **canonical kernel**, K_{δ_0}, is

$$K_{\delta_0}(x) = \frac{1}{\delta_0} K\left(\frac{x}{\delta_0}\right).$$

Choosing the canonical kernel in the scale family allows comparison across families of kernels. For example, we shall see that if we choose a canonical kernel, the optimal bandwidth does not depend on the kernel.

Lemma 10.1. *For a scale family* $\{K_\delta, \delta > 0\}$*, the canonical bandwidth* δ_0 *satisfies*

$$||K_{\delta_0}||_2^2 = [\mu_2(K_{\delta_0})]^2.$$

Proof. We show that if $||K_h||_2^2 = [\mu_2(K_h)]^2$ if and only if $h = \delta_0$. Consider separately the right and left hand sides.

$$||K_h||_2^2 = \int_{-\infty}^{\infty} [K_h(x)]^2 \, dx = \frac{1}{h} \int_{-\infty}^{\infty} \left[K\left(\frac{x}{h}\right)\right]^2 d\left(\frac{x}{h}\right) = \frac{1}{h}||K||_2^2$$

$$\mu_2(K_h) = \int_{-\infty}^{+\infty} x^2 K_h(x) dx = h^2 \int_{-\infty}^{+\infty} \left(\frac{x}{h}\right)^2 K\left(\frac{x}{h}\right) d\frac{x}{h} = h^2 \mu_2(K)$$

Therefore, $||K_h||_2^2 = \mu_2(K_h)^2 \Leftrightarrow \frac{1}{h}||K||_2^2 = \left[h^2\mu_2(K)\right]^2$ which implies that

$$h = \left(\frac{||K||_2^2}{[\mu_2(K)]^2}\right)^{\frac{1}{5}} = \delta_0.$$

□

10.2.5 Optimal kernel and optimal bandwidth

We are looking for the kernel and the bandwidth that minimise the asymptotic MISE. The AMISE is given by

$$\text{AMISE} \approx \frac{||f''(x)||_2^2}{4}\left[h^2\mu_2(K)\right]^2 + \frac{\sigma^2}{n}\frac{||K||_2^2}{h}.$$

For a canonical kernel, the AMISE factorises into a term that depends on bandwidth and a term that depends on the kernel:

$$\text{AMISE} \approx ||K||_2^2 \left[h^4 \frac{||f''(x)||_2^2}{4} + h^{-1}\frac{\sigma^2}{n}\right].$$

For any kernel, we can also define the **optimal bandwidth**, h_{opt}, by minimising the AMISE over h. First, we take a derivative of the AMISE with respect to h:

$$\frac{\partial}{\partial h}\text{AMISE} = \left[4h^3 C_1 - h^{-2}\frac{C_2}{n}\right] = 0$$

where $C_1 = ||f''(x)||_2^2\mu_2(K)^2/4$, and $C_2 = \sigma^2||K||_2^2$, which is solved by

$$h_{\text{opt}} = \left(\frac{C_2}{4nC_1}\right)^{\frac{1}{5}} = \left(\frac{\sigma^2||K||_2^2}{n||f''(x)||_2^2\mu_2(K)^2}\right)^{\frac{1}{5}}$$

which corresponds to the minimum of AMISE. For a canonical kernel we note that $||K||_2^2 = \mu_2(K)^2$ and so the optimal bandwidth does not depend on the kernel but it does depend on the unknown function.

Using the optimal bandwidth, the AMISE becomes

$$\text{AMISE} = \frac{5\sigma^{\frac{8}{5}}||f''(x)||_2^{\frac{2}{5}}}{4n^{\frac{4}{5}}}\left(\sqrt{\mu_2(K)}||K||_2^2\right)^{\frac{4}{5}}.$$

Optimal kernel: choose the kernel K to minimize the AMISE. From the preceding expression, this corresponds to minimising the quantity $\sqrt{\mu_2(K)}||K||_2^2$. We note that this is independent of bandwidth, in the sense that $\sqrt{\mu_2(K)}||K||_2^2 = \sqrt{\mu_2(K_\delta)}||K\delta||_2^2$ for all δ. However, rescaling by δ in this way will change the corresponding optimal bandwidth, so that the rescaled kernel with its optimal bandwidth is unchanged. We can use this freedom to set $\mu_2(K) = 1$ (which requires rescaling by $\delta = 1/\sqrt{\mu_2(K)}$). For this choice, minimising the bandwidth-optimised AMISE is equivalent to minimising $||K||_2^2$ under the constraints:

$$\int K(x)dx = 1, \quad \int xK(x)dx = 0, \quad \int x^2K(x)dx = 1.$$

The canonical kernel that minimises $||K||_2$ under these constraints is

$$K^{\text{opt}}(x) = \frac{3}{4}\frac{1}{\sqrt{5}}\left(1 - \frac{x^2}{5}\right)\mathbf{1}(|x| \leq \sqrt{5}).$$

This kernel is called the Epanechnikov kernel. For the Epanechnikov kernel, $||K||_2^2 = 3/5\sqrt{5}$ and $\mu_2(K) = 1$ by construction, so the optimal bandwidth is

$$h_{\text{opt}} = \left(\frac{3\sigma^2}{5\sqrt{5}n||f''(x)||_2^2}\right)^{\frac{1}{5}}.$$

Therefore, the **optimal kernel with the optimal bandwidth**, $K_{h_{\text{opt}}}$, is given by

$$K_{h_{\text{opt}}}(x) = \frac{1}{h_{\text{opt}}}K\left(\frac{x}{h_{\text{opt}}}\right) = \frac{3}{4}\frac{1}{\sqrt{5}h_{\text{opt}}}\left(1 - \frac{x^2}{5h_{\text{opt}}^2}\right)\mathbf{1}(|x| \leq \sqrt{5}h_{\text{opt}}),$$

and the Nadaraya-Watson estimator constructed with this kernel has the smallest AMISE.

The efficiency of a kernel family $\{K_\delta, \delta > 0\}$ for a given kernel K is defined as

$$\frac{\sqrt{\mu_2(K)}\,||K||_2^2}{\sqrt{\mu_2(K^{\text{opt}})}||K^{\text{opt}}||_2^2} = \frac{\sqrt{\mu_2(K_{\delta_0})}||K_{\delta_0}||_2^2}{\sqrt{\mu_2(K^{\text{opt}}_{\delta_0^{\text{opt}}})}||K^{\text{opt}}_{\delta_0^{\text{opt}}}||_2^2} = \left(\frac{\mu_2(K_{\delta_0})}{\mu_2(K^{\text{opt}}_{\delta_0^{\text{opt}}})}\right)^{\frac{5}{2}} = \left(\frac{||K_{\delta_0}||_2^2}{||K^{\text{opt}}_{\delta_0^{\text{opt}}}||_2^2}\right)^{\frac{5}{4}}$$

where δ_0 is the canonical bandwidth for this kernel family, K^{opt} is the Epanechnikov kernel and δ_0^{opt} is its canonical bandwidth. The efficiency to the fourth fifths power gives the ratio of the AMISE for this family of kernels relative to the optimal kernel family. For many kernel families, the efficiencies are close to 1, for instance, it is 0.951 for the Gaussian kernel family, 0.930 for the box kernel family and 0.986 for the triangular kernel family.

Note that since the optimal bandwidth depends on the unknown function, this expression gives a theoretical bound but it is not applicable in practice. One way to avoid dependency on the unknown function is to take $h_{\text{opt}} = Cn^{-1/5}$ which gives the same order of MISE in n but not the optimal constant. Another way to find the best h that is used in practice is to use another approximation of MISE which results in the approach called cross-validation.

10.2.6 Non-asymptotic properties of the Nadaraya-Watson estimator

Nonasymptotic properties of the Nadaraya-Watson estimator can be found in the form of upper bounds on the absolute value of the bias and the variance, and hence on the MSE

and MISE. We shall see that the upper bounds are the same functions of the sample size n. The constants in the upper bounds inform us how the errors depend on other features of the model, such as the kernel, the smoothness of the function, design, etc.

Before we state the upper bounds, we will define a class of smooth functions, the Hölder Class $\boldsymbol{H^{\beta}(M)}$. When the parameter β is an integer, the class $\boldsymbol{H^{\beta}(M)}$ contains functions with β derivatives whose absolute values are bounded by M. However, the class is defined for arbitrary values $\beta > 0$.

Definition 10.11. *The Hölder Class* $\boldsymbol{H^{\beta}(M)}$ *of functions on* $[0,1]$ *with* $\beta > 0,\ M > 0$ *is defined as the set of functions* f *that satisfy the following conditions with* $k = \lfloor \beta \rfloor$*:*

1. $|f^{(k)}(x)| \leq M$ *for all* $x \in [0,1]$,
2. $|f^{(k)}(x) - f^{(k)}(y)| \leqslant M|x-y|^{\beta-k},\ \forall x, y \in [0,1]$,
   where $f^{(k)}$ *is the* k*th derivative of* f.

If $\beta \in (0,1)$, $k = 0$ *and* $f^{(0)}(x) = f(x)$.

Example: if $\beta = 1$, the Hölder class $\boldsymbol{H^{1}(M)}$ contains functions such that $|f'(x)| \leq M$ for all $x \in [0,1]$.

Example: the function $f(x) = \sqrt{|x - 0.5|}$, $x \in [0,1]$, does not have a derivative for all $x \in [0,1]$ but it belongs to the Hölder class $\boldsymbol{H^{\beta}(M)}$ with $\beta = 1/2$ and $M = 1$ due to the inequality

$$|\sqrt{|z|} - \sqrt{|y|}| \leqslant \sqrt{|z-y|} \quad \forall z, y \in [0,1].$$

Now we derive upper bounds on the absolute value of the bias and the variance of the Nadaraya-Watson estimator of a function f that belongs to a Hölder class $\boldsymbol{H^{\beta}(M)}$ with $\beta \in (0,1)$.

Proposition 10.2. *Suppose that* $f \in \mathbb{H}^{\beta}(M)$ *on* $[0,1]$, *with* $\beta \in (0,1]$ *and* $M > 0$. *Let* $\widehat{f}_n^{NW}$ *be the Nadaraya – Watson estimator of* f.

Assume also that:

a) the design $(X_1, \ldots, X_n)$ *is regular deterministic;*

b) $var(\varepsilon_i) = \sigma^2$*;*

c) $\exists\, \lambda_0 > 0$ *such that* $\forall\, x \in [0,1]$,

$$\frac{1}{n} \sum_{i=1}^{n} K_h\left(X_i - x\right) \geq \lambda_0;$$

d) $supp(K) \subseteq [-1,1]$ *(i.e.* $K(x) = 0$ *for* $x \notin [-1,1]$*),*
and $\exists\, K_{\max} \in (0,\infty)$ *such that* $0 \leq K(u) \leq K_{\max}$, $\forall u \in \mathbb{R}$.

Then, for all $x_0 \in [0,1]$ *and* $h \geq 1/(2n)$,

$$|b(x_0)| \leq Mh^{\beta}, \quad v(x_0) \leq \frac{\sigma^2 K_{\max}}{nh\lambda_0}.$$

Proof. 1. The bias of the NW estimator when $f \in H^B(M)$ with $\beta \in (0,1)$ is:

$$\text{bias}(\widehat{f^{NW}}(x)) = \mathbb{E}(\widehat{f^{NW}}(x)) - f(x) = \sum_{i=1}^{n} W_i^{NW}(x)\,[f(x_i) - f(x)]\,.$$

Note that $\forall\, x$, $\sum_{i=1}^n W_i^{NW}(x) = 1$, since

$$\sum W_i(x) = \frac{\sum_{i=1}^n K_h(x_i - x)}{\sum_{i=1}^n K_h(x_i - x)} \mathbf{1}\left(\sum K_h(x_i - x) \neq 0\right) = 1.$$

Therefore, the bias is given by

$$\text{bias}(\widehat{f}^{NW}(x)) = \sum_{i=1}^{n} W_i^{NW}(x)[f(x_i) - f(x)].$$

Since the support of K is $[-1,1]$, the support of $K_h(x) = \frac{1}{h}K(x/h)$ is $[-h,h]$, therefore the sum is only over those i where $|x_i - x| \leq h$, that is,

$$\begin{aligned}
|\text{bias}(\widehat{f}^{NW}(x))| &= \frac{\left|\sum_i K(\frac{x_i - x}{h})(f(x_i) - f(x))\right|}{\sum_i K(\frac{x_i - x}{h})} = \frac{\left|\sum_{i:\,|x_i - x| \leq h} K(\frac{x_i - x}{h})[f(x_i) - f(x)]\right|}{\sum_i K(\frac{x_i - x}{h})} \\
&\leq \frac{\sum_{i\,:|x_i - x| \leq h} K(\frac{x_i - x}{h})\,|f(x_i) - f(x)|}{\sum_i K(\frac{x_i - x}{h})} \leq \frac{\sum_{i\,:|x_i - x| \leq h} K(\frac{x_i - x}{h})\,M|x_i - x|^\beta}{\sum_i K(\frac{x_i - x}{h})} \\
&\leq Mh^\beta,
\end{aligned}$$

using $K(z) \geq 0$ for all z. In particular, the bias is small when h is small, that is, $\text{bias}(\widehat{f}^{NW}(x)) \to 0$ if $h \to 0$. The extension of the proof to $\beta = 1$ is left as an exercise.

2. The variance of the NW estimator can be written as

$$v(x) = \text{Var}(\widehat{f}_n^{NW}(x)) = \text{Var}\left(\sum_{i=1}^n w_i(x)(Y_i)\right) = \sum_{i=1}^n [w_i(x)]^2 \text{Var}(Y_i)$$

since the Y_i's are independent. From assumptions (a) & (b), we know that $\text{Var}(Y_i) = \sigma^2$, since the x_i's are fixed. Therefore,

$$\begin{aligned}
v(x) &= \sigma^2 \sum_{i=1}^n \frac{[K_h(X_i - x)]^2}{\left[\sum_{j=1}^n K_h(X_j - x)\right]^2} && \text{assumption d) } K(z) \geq 0 \text{ for all } z \\
&\leqslant \sigma^2 \frac{\frac{K_{\max}}{h}\sum_{i=1}^n K_h(X_i - x)}{\left[\sum_{j=1}^n K_h(X_j - x)\right]^2} && \text{assumption d), } \forall u, K(u) \leqslant K_{\max} \text{ implies } K_h(X_i - x) = \tfrac{1}{h}K\left(\tfrac{X_i - x}{h}\right) \leqslant \tfrac{K_{\max}}{h} \\
&\leqslant \sigma^2 \frac{\frac{K_{\max}}{h}\sum_{i=1}^n K_h(X_i - x)}{n\lambda_0 \sum_{j=1}^n K_h(X_j - x)} && \text{assumption c) } \exists \lambda_0 > 0 \text{ such that } \forall x \in [0,1], \\
&= \frac{\sigma^2 K_{\max}}{nh\lambda_0} && \sum_{i=1}^n K_h(X_i - x) \geqslant n\lambda_0.
\end{aligned}$$

□

Now we consider the bounds on the MSE of the NW estimator. Under the conditions of Proposition 10.2,

$$\mathrm{MSE}(\widehat{f}_n^{NW}(x_0)) = [\mathrm{bias}(\widehat{f}_n^{NW}(x_0))]^2 + \mathrm{Var}(\widehat{f}_n^{NW}(x_0)) \le M^2 h^{2\beta} + \frac{\sigma^2 K_{\max}}{nh\lambda_0}.$$

The upper bound on MSE is the smallest if

$$h = h_n = \left(\frac{\sigma^2 K_{\max}}{2\beta M^2 \lambda_0 n}\right)^{1/(2\beta+1)},$$

and the corresponding MSE bound is

$$\begin{aligned}\mathrm{MSE}(\widehat{f}_{n,h_n}^{NW}(x_0)) &\le M^2 \left(\frac{\sigma^2 K_{\max}}{2\beta M^2 \lambda_0 n}\right)^{2\beta/(2\beta+1)} + \frac{\sigma^2 K_{\max}}{n\lambda_0}\left(\frac{2\beta M^2 \lambda_0 n}{\sigma^2 K_{\max}}\right)^{1/(2\beta+1)} \\ &\le (1+2\beta) M^{2/(2\beta+1)} \left(\frac{\sigma^2 K_{\max}}{2\beta\lambda_0 n}\right)^{2\beta/(2\beta+1)} \to 0 \quad \text{as} \quad n \to \infty.\end{aligned}$$

Hence, the Nadaraya – Watson estimator with $h = h_n = \left(\frac{\sigma^2 K_{\max}}{2\beta M^2 \lambda_0 n}\right)^{1/(2\beta+1)}$ and kernel K satisfying conditions of Proposition 10.2, is consistent for estimating functions from Hölder class $\mathbb{H}^\beta(M)$ for $\beta \in (0, 1]$.

Example 10.6. *(continued) Derive upper bounds on the absolute value of the bias and the variance of the NW estimator with the box kernel $K(z) = \frac{1}{2}\mathbf{1}(z \in [-1, 1])$ under the nonparametric regression model with $\sigma^2 = 1$ and $x_i = i/n$. Let $f \in H^\beta(M)$, $M = 5$,$\beta = 1/2$.*

Now we verify the assumptions of Proposition 10.2. Assumptions a), b) are satisfied. Assumption c) is $\frac{1}{2n}\sum_{i:|x_i - x| \le h} \frac{1}{h} \ge \lambda_0$, $h \ge 1/2n$.

Let's count the number of integers i between 1 and n such that $|i/n - x| \le h$. Since

$$|i/n - x| \le h \Leftrightarrow (nx - nh) \le i \le (nx + nh),$$

we need to count the number of integers in the interval $[nx - nh, nx + nh]$.

In general, in an interval $[a, a + b]$ for some $b > 0$, the number of of integers is $\lfloor b \rfloor$ if a is not integer, and it is $\lfloor b \rfloor$ +1 if a is integer. Here $\lfloor b \rfloor$ is the lower integer part of b, that is, the largest integer that is less than or equal to b, e.g. $\lfloor 5 \rfloor = 5$, $\lfloor 7.3 \rfloor = 7$ and $\lfloor 2.8 \rfloor = 2$.

Therefore, the smallest number of integers in the interval $[nx - nh, nx + nh]$ is $\lfloor 2nh \rfloor$ which is greater than $2nh - 1$ since $\lfloor 2nh \rfloor \le 2nh < \lfloor 2nh \rfloor + 1$ by the definition of the lower integer part. Hence, we need $h > 1/(2n)$, and then we can take $\lambda_0 = 1 - 1/(2nh) > 0$ since

$$\frac{1}{2n} \sum_{i:|x_i - x| \le h} \frac{1}{h} \ge \frac{2nh - 1}{2nh} = 1 - 1/(2nh) = \lambda_0$$

Assumption d) is satisfied with $K_{\max} = 1/2$.

Therefore, by Proposition 10.2, for $n = 12$ and $h > 1/24$,

$$|b(x)| \le Mh^\beta = 5\sqrt{h}, \quad v(x) \le \frac{1}{2nh(1 - 1/(2nh))} = \frac{1}{2nh - 1} = \frac{1}{24h - 1}.$$

The corresponding MSE (and MISE) for $\hat{f}^{NW}(x)$ is bounded by

$$MSE(\hat{f}^{NW}(x)) = b^2(x) + v(x) \le 25h + \frac{1}{24h - 1}.$$

The derivative of the upper bound with respect to h is

$$25 - \frac{24}{(24h-1)^2}$$

which is zero for $h > 1/24$ at

$$h_{opt} = \frac{1}{24}\left(1 + \sqrt{\frac{24}{25}}\right) = 0.0825.$$

This corresponds to the minimum of the MSE since the second derivative with respect to h of the upper bound is $\frac{2\cdot 24^2}{(24h-1)^3}$ which is positive.

Therefore, the optimal bandwidth is 0.0825.

10.2.7 Rates of convergence

We would like to find the estimator of f which is not only consistent, but also achieves the best possible rate of convergence over some class of functions $\mathcal{F}$, such as the Hölder class $H^\beta(M)$. Now we determine the rate of convergence of the NW estimator, in both local and global distances, and address the question whether it is possible to achieve a faster rate of convergence.

Definition 10.12. *ϕ_n is the* **convergence rate of an estimator $\widehat{f}_n$ at point** *x_0 (local rate of convergence) over a class of functions $\mathcal{F}$, if*

$$0 < c \leqslant \sup_{f\in\mathcal{F}} \mathbb{E}\left[\frac{|\widehat{f}_n(x_0) - f(x_0)|}{\phi_n}\right]^2 \leqslant C < \infty,$$

where constants c and C do not dependent on n, and the rate ϕ_n is only related to n and the function class $\mathcal{F}$.

Similarly, the global rate of convergence of estimator $\widehat{f}_n$ over a class of functions $\mathcal{F}$ is ϕ_n if

$$0 < c \leqslant \sup_{f\in\mathcal{F}} \mathbb{E}\left[\frac{||\widehat{f}_n - f||_2}{\phi_n}\right]^2 \leqslant C < \infty,$$

where the constants c and C do not depend on n, and the rate ϕ_n is only related to n and the function class $\mathcal{F}$.

Recall that $||\widehat{f}_n(x) - f(x)||_2 = \sqrt{\int_0^1 [\widehat{f}_n(x) - f(x)]^2 dx}$.

Definition 10.13. *For a class of functions $\mathcal{F}$, $\phi_n^\star$ is the* **local minimax convergence rate**, *if*

$$0 < c \leqslant \inf_{\widehat{f}_n} \sup_{x_0\in(0,1)} \sup_{f\in\mathcal{F}} \mathbb{E}\left[\frac{|\widehat{f}_n(x_0) - f(x_0)|}{\phi_n^\star}\right]^2 = \inf_{\widehat{f}_n} \sup_{x_0\in(0,1)} \sup_{f\in\mathcal{F}} \frac{MSE(\widehat{f}_n(x_0))}{(\phi_n^\star)^2} \leqslant C < \infty,$$

where the constants c and C do not depend on n, and the rate $\phi_n^\star$ is only related to n and the function class $\mathcal{F}$.

Similarly, for a class of functions $\mathcal{F}$, $\phi_n^\star$ is the **global minimax convergence rate***, if*

$$0 < c \leqslant \inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E}\left[\frac{||\widehat{f}_n - f||_2}{\phi_n^\star}\right]^2 = \inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} \frac{MISE(\widehat{f}_n)}{(\phi_n^\star)^2} \leqslant C < \infty,$$

where constants c and C do not depend on n, and the rate $\phi_n^\star$ is only related to n and the function class $\mathcal{F}$.

Definition 10.14. *An estimator $\widehat{f}_n$ is said to achieve a minimax rate of convergence (local or global), if the rate of convergence of this estimator is the corresponding (local or global) minimax rate of convergence.*

Now we investigate whether the local rate of convergence for the Nadaraya-Watson estimator is minimax.

Theorem 10.1. *Let assumptions of Proposition 10.2 hold for all $x \in [0,1]$. Then, the NW estimator $\widehat{f}^{NW}(x)$ with $h = \alpha n^{-1/(2\beta+1)}$ for same $\alpha > 0$ satisfies*

$$\lim_{n\to\infty} \sup_{x_0\in[0,1]} \sup_{f\in H^\beta(M)} \mathbb{E}\left[\left((\widehat{f}_n^{\,NW}(x_0) - f(x_0))n^{\beta/(2\beta+1)}\right)^2\right] \leq C < \infty,$$

where constant C depends only on $\beta, M, \sigma^2, \lambda_0, K_{\max}, \alpha$.

Proof. By Proposition 10.2, $\forall f \in H^\beta(M), \forall x \in [0,1]$,

$$\mathbb{E}\left[\left(\widehat{f}_n^{\,NW}(x) - f(x)\right)^2\right] \leq Cn^{\frac{-2\beta}{2\beta+1}}$$

with $C < \infty$ dependent on $K_{\max}, \lambda_0, \beta, M, \alpha, \sigma^2$ which can be written as

$$\mathbb{E}\left[\left((\widehat{f}_n^{\,NW}(x) - f(x))n^{\beta/2\beta+1}\right)^2\right] \leq C.$$

Taking supremum over $f \in H^\beta(M)$, $x \in [0,1]$ and n, as $n \to \infty$, we have the statement. □

Therefore, the pointwise rate of convergence of the Nadaraya-Watson estimator is $n^{-\beta/(2\beta+1)}$. In fact, it can be shown (Tsybakov, 2009, chapter 2) that this is the local minimax rate of convergence, so the Nadaraya-Watson estimator achieves this minimax rate and so it is in this sense the "best" estimator, but there do exist other estimators that achieve this rate of convergence. It is straightforward to show that the NW estimator also achieves the global minimax rate of convergence.

The upper bounds being used here apply for the Hölder space with $\beta \in (0,1]$. For the Nadaraya-Watson estimator to achieve the minimax convergence rate for $\beta > 1$, one needs to use kernels of higher order. **Local polynomial estimators**, which will be discussed in Section 10.2.12 are locally and globally minimax for $\beta > 1$.

10.2.8 Inference using a linear estimator

In this subsection we consider the nonparametric regression model

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

with independent errors $\varepsilon_i \sim N(0, \sigma^2)$ and a deterministic design $(x_1, \ldots, x_n)$. These assumptions imply that $\mathbb{E}(Y_i) = f(X_i)$ and $\mathrm{Var}(Y_i) = \sigma^2$.

10.2.9 Confidence intervals for $f(x_0)$ based on a linear estimator

Denote $b(x) = \text{bias}(\widehat{f}(x)) = \mathbb{E}\left[\widehat{f}(x) - f(x)\right]$ and $v(x) = \text{Var}(\widehat{f}(x))$. Then, for a **linear estimator** $\widehat{f}(x) = \sum_{i=1}^{n} Y_i w_i(x)$,

$$\begin{aligned} \mathbb{E}\left(\widehat{f}(x)\right) &= \sum_{i=1}^{n} f(x_i) w_i(x) = b(x) + f(x) \\ \text{Var}\left(\widehat{f}(x)\right) &= \sigma^2 \sum_{i=1}^{n} [w_i(x)]^2 = v(x), \end{aligned}$$

therefore $\widehat{f}(x) \sim N\left(b(x) + f(x), v(x)\right)$.

The variance depends on the weights $w_i(x)$ and σ which are known, so it can be calculated exactly. If we knew the bias, which depends on the unknown function, we could construct $(1-\alpha)100\%$ confidence interval using the fact that the following inequality

$$-z_{\frac{\alpha}{2}} \leqslant \frac{\widehat{f}(x) - [b(x) + f(x)]}{\sqrt{v(x)}} \leqslant z_{\frac{\alpha}{2}}$$

holds with probability $1-\alpha$, that is,

$$f(x) \in [\widehat{f}(x) - b(x) - z_{\frac{\alpha}{2}}\sqrt{v(x)}, \widehat{f}(x) - b(x) + z_{\frac{\alpha}{2}}\sqrt{v(x)}].$$

Here $z_\alpha = \Phi^{-1}(1-\alpha)$ where $\Phi(x)$ is the cumulative distribution function of $N(0,1)$.

However, the bias is unknown, so it is not possible to construct the exact confidence interval. There are two approaches to addressing this issue. The first one is to construct an asymptotic confidence interval where the estimator is constructed in such a way that asymptotically the bias is much smaller than the variance, and therefore may be treated as 0. For the NW estimator, this means choosing a smaller bandwidth. The second one is to use an upper bound on the bias to construct a conservative confidence interval.

- $(1-\alpha)100\%$ Conservative Confidence Interval for $f(x)$.

  If $|b(x)| \leqslant b_0(x)$ & $v(x) \leqslant v_0(x)$, then

  $$f(x) \in \widehat{f}(x) \pm \left(b_0(x) + z_{\frac{\alpha}{2}}\sqrt{v_0(x)}\right).$$

- $(1-\alpha)100\%$ Asymptotic Confidence Interval for $f(x)$.

  Choose the estimator $\widehat{f}(x)$ so that $b(x)^2 \ll v(x)$, thus we can assume $b(x) \approx 0$:

  $$f(x) \in \widehat{f}(x) \pm z_{\frac{\alpha}{2}}\sqrt{v(x)}.$$

  The asymptotic expression for the variance is often used in this case.

10.2.10 Confidence intervals using the Nadaraya-Watson estimator

For a Nadaraya-Watson estimator $f \in H^\beta(M)$ on $x \in [0,1]$, under the conditions of Proposition 10.2,

$$v(x) \leqslant \frac{\sigma^2 K_{\max}}{nh\lambda_0}, \quad |b(x)| \leqslant Mh^\beta.$$

Therefore, a $(1-\alpha)100\%$ **Conservative Confidence Interval** for $f(x)$ is

$$\widehat{f}^{NW}(x) \pm \left(Mh^\beta + z_{\alpha/2}\sigma\sqrt{K_{\max}/(nh\lambda_0)}\right)$$
$$= \left[\widehat{f}^{NW}(x) - Mh^\beta - z_{\alpha/2}\sigma\sqrt{K_{\max}/(nh\lambda_0)}, \widehat{f}^{NW}(x) + Mh^\beta + z_{\alpha/2}\sigma\sqrt{K_{\max}/(nh\lambda_0)}\right].$$

Alternatively, taking the limit $n \to \infty$ and $h \to 0$,

$$v(x) \approx \frac{\sigma^2}{nh}||K||_2^2, \quad b(x) \approx \frac{\mu_2(K)h^2}{2}f''(x) \approx 0.$$

Therefore, a $(1-\alpha)100\%$ **Asymptotic Confidence Interval** for $f(x)$ is

$$\widehat{f}^{NW}(x) \pm z_{\alpha/2}\sigma\sqrt{||K||_2^2/(nh))}$$
$$= \left[\widehat{f}^{NW}(x) - z_{\alpha/2}\sigma\sqrt{||K||_2^2/(nh)}, \widehat{f}^{NW}(x) + z_{\alpha/2}\sigma\sqrt{||K||_2^2/(nh)}\right].$$

10.2.11 Asymptotic Confidence Band for f

Assume that the bias of $\widehat{f}(x)$ is much smaller than its standard deviation and is close to 0, i.e. $|b(x)| \ll \sqrt{v(x)}$ and $b(x) \approx 0$. Then, an asymptotic $(1-\alpha)100\%$ confidence band based on the NW estimator is given by

$$\left\{f : |f(x) - \widehat{f}(x)| \leqslant c_\alpha\sqrt{v(x)},\ \forall\, x \in [a,b]\right\}$$

with

$$c_\alpha \approx \sqrt{2\log\left(\frac{a_0}{\alpha h}\right)},\ \text{where } a_0 = \frac{|b-a|}{\pi}\frac{||K'||_2}{||K||_2},$$

(see Wasserman, section 5.7). For the NW estimator, we can use $v(x) \approx \frac{\sigma^2}{nh}||K||_2^2$.

Confidence bands can be used to test hypotheses about f, e.g.

$$H_0 : f(x) = \text{constant}\ \forall x \in [0,1].$$

10.2.12 Local polynomial estimators.

Motivation and definition The Nadaraya-Watson estimator can be viewed as a local constant least squares approximation of the unknown function. If the kernel K takes only nonnegative values, then for each $x \in [0,1]$, $\widehat{f}_n^{NW}(x)$ satisfies

$$\widehat{f}_n^{NW}(x) = \arg\min_{\theta_x \in \mathbb{R}}\left\{\sum_{i=1}^n (Y_i - \theta_x)^2 K\left(\frac{X_i - x}{h}\right)\right\}$$
$$= \arg\min_{\theta_x \in \mathbb{R}}\left\{\sum_{i=1}^n (\theta_x^2 - 2\theta_x Y_i + Y_i^2) K\left(\frac{X_i - x}{h}\right)\right\}$$
$$= \arg\min_{\theta_x \in \mathbb{R}}\left\{\theta_x^2 \cdot \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) - \theta_x \cdot 2\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right) + C_{X_i,Y_i}(x)\right\}$$

Therefore, if $\sum_{j=1}^{n} K_h(X_j - x) \neq 0$, the value of θ_x that minimises this weighed sum of squares coincides with the Nadaraya-Watson estimator:

$$f_n^{NW}(x) = \frac{\sum_{i=1}^{n} Y_i K_h(X_i - x)}{\sum_{j=1}^{n} K_h(X_j - x)}.$$

This estimator can be generalised further by considering a local polynomial rather than a local constant approximation. For a function $f(x)$, if $\exists f^{(k)}(x)$, then for x_i sufficiently close to x,

$$\begin{aligned} f(x_i) \approx & f(x) + f'(x)(x_i - x) + \cdots + \frac{f^{(k)}(x)}{k!}(x_i - x)^k = \sum_{j=0}^{k} \frac{f^{(j)}(x)}{j!}(x_i - x)^j \\ = & \sum_{j=0}^{k} \left[f^{(j)}(x)h^j\right] \left[\frac{1}{j!}\left(\frac{x_i - x}{h}\right)^j\right] = U_{x,i}^T \theta_x \end{aligned}$$

where

$$\begin{aligned} \theta_x &= \left(f(x), f'(x)h, f''(x)h^2, \ \ldots \ , f^{(k)}(x)h^k\right)^T \\ U_{x,i} &= \left(1, \frac{x_i - x}{h}, \frac{1}{2!}\left(\frac{x_i - x}{h}\right)^2, \ldots, \frac{1}{k!}\left(\frac{x_i - x}{h}\right)^k\right)^T \end{aligned}$$

Definition 10.15. *A local polynomial estimator of $f(x)$ of order k , denoted $LP(k)$ estimator, is defined by*

$$\widehat{f}_n^{LP}(x) = \widehat{\theta}_0(x)$$

where for each x $\widehat{\theta}(x) = \left(\widehat{\theta}_0(x), \widehat{\theta}_1(x), \ldots, \widehat{\theta}_k(x)\right)^T$ is the solution of

$$\widehat{\theta}(x) = \arg \min_{\theta_x \in \mathbb{R}^{k+1}} \left\{ \sum_{i=1}^{n} (Y_i - U_{x,i}^T \theta_x)^2 K\left(\frac{X_i - x}{h}\right) \right\}.$$

For each $m = 1, \ldots, k$, $\widehat{\theta}_m(x)/h^m$ is an estimator of $f^{(m)}(x)$.

Therefore, the local polynomial estimator provides simultaneous estimators not only for $f(x)$ but also for all existing derivatives of f.

This estimator can be written explicitly. Noticing that the expression to be minimised is quadratic in the vector θ_x, we can open the brackets to obtain

$$\begin{aligned} \widehat{\theta}_x &= \arg\min_{\theta_x} \left\{ \sum_{i=1}^{n} (Y_i - U_{x,i}^T \theta_x)^2 K\left(\frac{X_i - x}{h}\right) \right\} \\ &= \arg\min_{\theta_x} \left\{ \sum_{i=1}^{n} (\theta_x^T U_{x,i} U_{x,i}^T \theta_x - 2U_{x,i}^T \theta_x Y_i + Y_i^2) K\left(\frac{X_i - x}{h}\right) \right\} \\ &= \arg\min_{\theta_x} \left\{ \theta_x^T \cdot \sum_{i=1}^{n} U_{x,i} U_{x,i}^T K\left(\frac{X_i - x}{h}\right) \cdot \theta_x - \theta_x^T \cdot 2 \sum_{i=1}^{n} Y_i U_{x,i} K\left(\frac{X_i - x}{h}\right) + C_{X_i,Y_i}(x) \right\} \end{aligned}$$

which is equivalent to

$$\widehat{\theta}_x = \arg\min_{\theta_x} \left\{ \theta_x^T \cdot B(x) \cdot \theta_x - 2\theta_x^T \cdot a(x) \right\}$$

where the matrix $B(x)$ and vector $a(x)$ are defined by

$$B(x) = \frac{1}{nh}\sum_{i=1}^{n} U_{x,i}U_{x,i}^T K\left(\frac{X_i - x}{h}\right) \qquad a(x) = \frac{1}{nh}\sum_{i=1}^{n} Y_i U_{x,i} K\left(\frac{X_i - x}{h}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} U_{x,i}U_{x,i}^T K_h(X_i - x) \qquad = \frac{1}{n}\sum_{i=1}^{n} Y_i U_{x,i} K_h\left(X_i - x\right)$$

Hence, if $B(x)$ is invertible,

$$\widehat{\theta}_x = B^{-1}(x)a(x).$$

Therefore, the Local Polynomial estimator can be written as

$$\widehat{f}_n^{LP}(x) = \widehat{\theta}_0(x) = e_1^T B^{-1}(x)a(x)$$

where the matrix $B(x)$ and vector $a(x)$ are defined above and $e_1^T = (1, 0, 0, \cdots, 0)$.

Note that the local polynomial estimator $\widehat{f}_n^{LP}(x)$ is **linear**:

$$f_n^{LP}(x) = e_1^T B^{-1}(x)a(x) = e_1^T B^{-1}(x) \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i U_{x,i} K_h\left(X_i - x\right)$$
$$= \sum_{i=1}^{n} Y_i \cdot \frac{1}{n} K_h\left(X_i - x\right) \sum_{j=0}^{k} [B^{-1}(x)]_{0,j} \frac{1}{j!}\left(\frac{x_i - x}{h}\right)^j$$
$$= \sum_{i=1}^{n} Y_i w_i(x)$$

with weights

$$w_i(x) = \frac{1}{n} K_h\left(X_i - x\right) \sum_{j=0}^{k} [B^{-1}(x)]_{0,j} \frac{1}{j!}\left(\frac{x_i - x}{h}\right)^j$$

that are independent of $Y_1, \ldots, Y_n$.

Bias, variance, consistency and the rate of convergence for local polynomial estimator

Proposition 10.3. *Suppose that $f \in H^\beta(M)$ on $[0,1]$, with $\beta > 0$ and $M > 0$, and*

a) the design $(X_1, \ldots, X_n)$ is regular deterministic;

b) $\mathbb{E}(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$;

c) $\exists\, \lambda_0 > 0$ such that $\forall\, x \in [0,1]$, the smallest eigenvalue $\lambda_{\min}(B(x))$ of $B(x)$ satisfies

$$\lambda_{\min}(B(x)) \geqslant \lambda_0 \quad , \text{ where } B(x) = \frac{1}{n}\sum_{i=1}^{n} U_{x,i}U_{x,i}^T K_h(X_i - x);$$

d) $supp(K) \subseteq [-1,1]$ and $\exists\, K_{\max} \in (0,\infty)$ such that $\forall u, |K(u)| \leqslant K_{\max}$.

Let $\widehat{f}_n^{LP}$ be the Local Polynomial estimator of f which satisfies the above assumptions with $k = \lfloor \beta \rfloor$. Then, for all $x \in [0,1]$ and $h \geqslant \frac{1}{2n}$,

$$|b(x)| \leqslant \frac{C_K}{k!} M h^{\beta}, \quad v(x) \leqslant \frac{\sigma^2 C_K^2}{nh} \quad \text{with } C_K = \frac{2K_{\max}}{\lambda_0}.$$

Note that if $\beta \in (0,1)$, the LP estimator becomes the NW estimator, and this proposition coincides with Proposition 10.2.

Now we study consistency and the rates of convergence of $\widehat{f}_n^{LP}(x)$. Under the assumptions of Proposition 10.3, MSE of $\widehat{f}_n^{LP}(x)$ is bounded by

$$\text{MSE}\left[\widehat{f}_n^{LP}(x)\right] = [b(x)]^2 + v(x) \leqslant \left[\frac{C_K}{k!} M\right]^2 h^{2\beta} + \frac{\sigma^2 C_K^2}{n} h^{-1}$$

which is minimised at

$$h = h_n = \left(\frac{\frac{\sigma^2 C_K^2}{n}}{2\beta \left(\frac{C_K M}{k!}\right)^2}\right)^{\frac{1}{2\beta+1}} = \left(\frac{\sigma^2 (k!)^2}{2\beta M^2 n}\right)^{\frac{1}{2\beta+1}},$$

with the value of the minimum being

$$\text{MSE}\left[\widehat{f}_{n,h_{opt}}^{LP}(x)\right] \leqslant \left\{\left[\frac{C_K}{k!} M\right]^2 h_{opt}^{2\beta} + \frac{\sigma^2 C_K^2}{n} h_{opt}^{-1}\right\} = C_{LP} \cdot n^{-\frac{2\beta}{2\beta+1}} \to 0 \text{ as } n \to \infty,$$

where C_{LP} is a constant depending only on M, k, σ^2 and C_K (i.e. $K_{\max}, \lambda_0$).

Now we study the local and global minimax rates of convergence of the LP(k) estimator with $h_n = \alpha n^{-\frac{1}{2\beta+1}}$ over $H^{\beta}(M)$ with $k = \lfloor \beta \rfloor$. In this case, under the conditions of Proposition 10.3,

$$\text{MSE}\left[\widehat{f}_{n,h_n}^{LP}(x)\right] \leqslant C_K^2 \left[\frac{\alpha^2 M^2}{[k!]^2} + \alpha^{-1}\sigma^2\right] n^{-\frac{2\beta}{2\beta+1}},$$

which also implies that

$$\text{MISE}\ (\widehat{f}^{LP}(x)) = \int_0^1 \text{MSE}(\widehat{f}^{LP}(x))dx \leqslant C n^{-\frac{2\beta}{2\beta+1}}$$

with the same constant as in the upper bound on the MSE. Therefore, both local and global rates of convergence of LP(k) are $n^{-\frac{\beta}{1+2\beta}}$. Therefore, the local polynomial estimator achieves both local and global minimax rates of convergence. Hence, we proved the following theorem.

Theorem 10.2. *Under the assumptions of Proposition 10.3, the Local Polynomial estimator with the bandwidth $h = h_n = \alpha n^{-\frac{1}{2\beta+1}}$, $\alpha > 0$, satisfies*

$$\begin{aligned} \limsup_{n\to\infty} \sup_{f\in\mathbb{H}^{\beta}(M)} \sup_{x_0\in[0,1]} \mathbb{E}\left[n^{\frac{\beta}{2\beta+1}} |f(x_0) - \widehat{f}_n(x_0)|\right]^2 &\leq C < \infty, \\ \limsup_{n\to\infty} \sup_{f\in\mathbb{H}^{\beta}(M)} \mathbb{E}\left[n^{\frac{\beta}{2\beta+1}} ||f - \widehat{f}_n||_2\right]^2 &\leq C < \infty, \end{aligned}$$

where C is a constant depending only on β, M, a_0, λ_0, $\sigma^2_{\max}$, $K_{\max}$ and α.

10.3 Smoothing Splines

10.3.1 Definition

Definition 10.16. *A smoothing spline is the penalised least squares estimator of f:*

$$\widehat{f}_n^{\text{pen}}(x) = \arg\min_{f\in\mathcal{C}^2}\left[\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda\,\text{pen}(f)\right] \tag{125}$$

with penalty function $\text{pen}(f) = \int[f''(x)]^2dx = ||f''||_2^2$*;* $\lambda > 0$ *is called the regularisation parameter.*

The solution to this minimisation problem has a simple form that is called a **natural cubic spline**.

Definition 10.17. *Let $a \le t_1 < .. < t_N \le b$ be a set of ordered points - called knots. A cubic spline is a continuous function g such that*

- *$g(x)$ is cubic on $[t_j, t_{j+1}]$, for each $j = 1, .., N-1$:*

$$g(x) = b_{j0} + b_{j1}x + b_{j2}x^2 + b_{j3}x^3, \quad x \in [t_j, t_{j+1}],$$

- *both g' and g'' are continuous at t_i, $i = 1, .., N$.*

A spline that is linear beyond the boundary knots is called a natural spline.

- *$g(x)$ is linear on $[a, t_1]$ and $[t_N, b]$*

$$\begin{aligned} g(x) &= b_{00} + \ b_{01}x \ , \quad x \in [a \ , t_1] \\ g(x) &= b_{N0} + b_{N1}x, \quad x \in [t_N, b] \end{aligned}$$

Theorem 10.3. *(without proof) Solution $\widehat{f}_n^{\text{pen}}$ of the above problem is a* **natural cubic spline** *with knots at the data points.*

Theorem 10.4. *Let knots $a \leqslant t_1 < \cdots < t_N \leqslant b$. For $j = 3, \ldots, N$, define*

$$\begin{aligned} h_1(x) &= 1, h_2(x) = x, \\ h_j(x) &= (x - t_{j-2})_+^3 - \frac{(t_N - t_{j-2})}{(t_N - t_{N-1})}(x - t_{N-1})_+^3 \\ &\qquad + \frac{(t_{N-1} - t_{j-2})}{(t_N - t_{N-1})}(x - t_N)_+^3, \quad \forall\ 3 \leqslant j \leqslant N, \end{aligned}$$

where $(x-y)_+^3 = \max\left\{(x-y)^3, 0\right\}$

The set of functions $(h_j)_{j=1}^N$ forms a basis for the set of natural cubic splines at these knots.

Thus, any natural cubic spline $g(x)$ can be written as

$$g(x) = \sum_{j=1}^{N}\beta_j h_j(x).$$

By Theorem 10.3, the solution of the minimisation problem that defines the smoothing spline is a natural cubic spline, and by Theorem 10.4, it can be written as the linear combination of the basis functions $h_j(x)$, $j = 1, 2, \ldots, N$. Hence, minimising over functions f

$$\begin{aligned}\widehat{f}_{n,\lambda}^{SS} &= \arg\min_{f\in C^2}\left\{\sum_{i=1}^{N}\left(Y_i - f(x_i)\right)^2 + \lambda\int\left[f''(x)\right]^2 dx\right\}\\ &= \arg\min_{f\in C^2}\left\{\sum_{i=1}^{N}\left(f(x_i)^2 - 2f(x_i)Y_i + Y_i^2\right) + \lambda\int\left[f''(x)\right]^2 dx\right\}\end{aligned}$$

is equivalent to minimising the following expression over the $(n+2)$-dimensional vector β:

$$\begin{aligned}\widehat{\beta} &= \arg\min_{\beta\in\mathbb{R}^N}\left\{\sum_{i=1}^{N}\left[\sum_{j=1}^{N}\beta_j h_j(x_i)\right]^2 - 2\sum_{i=1}^{N}\left[\sum_{j=1}^{N}\beta_j h_j(x_i)\right]Y_i + \lambda\int\left[\sum_{j=1}^{N}\beta_j h_j''(x)\right]^2 dx\right\}\\ &= \arg\min_{\beta\in\mathbb{R}^N}\left\{\beta^T H^T H\beta - 2\beta^T H^T Y + \lambda\beta^T\Omega\beta\right\},\end{aligned}$$

where $N\times N$ matrix H has entries $H_{ij} = h_j(x_i)$, $i = 1, \ldots, N$, $j = 1, \ldots, N$, and $N\times N$ matrix Ω has elements $\Omega_{j\ell} = \int h_j''(x)h_\ell''(x)dx$, $j, \ell = 1, \ldots, N$.

Hence, if $\left(H^T H + \lambda\Omega\right)$ is invertible,

$$\widehat{\beta} = \left[\left(H^T H + \lambda\Omega\right)^{-1} H^T Y\right].$$

Therefore, we have proved the following theorem.

Theorem 10.5. *A smoothing spline can be written as*

$$\widehat{f}_{n,\lambda}^{SS} = \sum_{j=1}^{N}\widehat{\beta}_j h_j(x)$$

where $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_N)^T$ *is given by*

$$\widehat{\beta} = (H^T H + \lambda\Omega)^{-1} H^T Y$$

where $Y = (Y_1, \ldots, Y_n)^T$*, and matrices* $H = (H_{ij})$ *and* $\Omega = (\Omega_{jl})$ *have entries*

$$H_{ij} = h_j(x_i), \quad \Omega_{jl} = \int_a^b h_j''(x)h_l''(x)dx, \quad i \in 1, \ldots, n, \quad j, l \in 1, \ldots, N$$

The smoothing spline is a linear estimator since it can be written as

$$\widehat{f}_{N,\lambda}^{SS} = \sum_{i=1}^{N} w_i(x)Y_i$$

with weights

$$w_i(x) = \sum_{j=1}^{N} h_j(x)\left[\left(H^T H + \lambda\Omega\right)^{-1} H^T\right]_{ji}$$

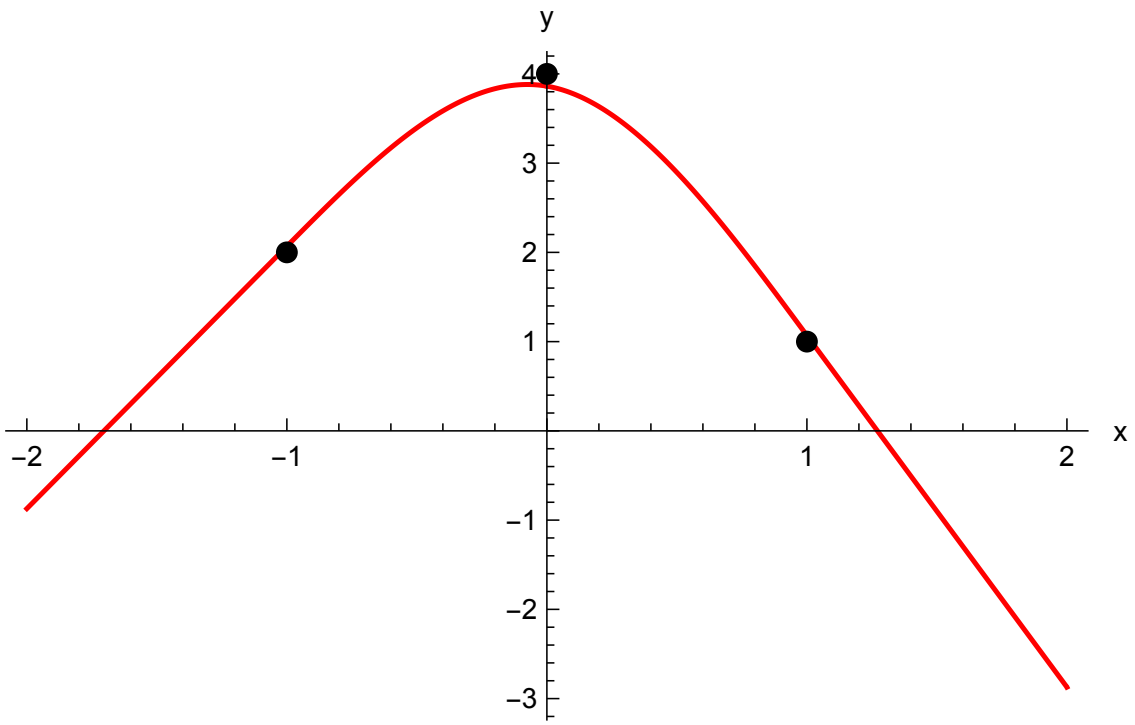


Figure 42: Smoothing spline for example 10.7.

Example 10.7. *Construct a smoothing spline on* $[-2,2]$ *given data* $(-1,2)$, $(0,4)$, $(1,1)$. *Take* $\lambda = 0.01$, *and construct the smoothing spline using*

$$\widehat{f}_n^{SS}(x) = \sum_{i=1}^{N}\sum_{j=1}^{N}[(H^TH + \lambda\Omega)^{-1}H^T]_{ji}h_j(x)Y_i.$$

The matrices necessary for the calculation are $H = (H_{ij})$, $H_{ij} = h_j(x_i)$*:*

$$H = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 6 \end{pmatrix}, \quad H^TH = \begin{pmatrix} 3 & 0 & 7 \\ 0 & 2 & 6 \\ 7 & 6 & 37 \end{pmatrix}$$

and $\Omega = (\Omega_{j\ell})$, $\Omega_{j\ell} = \int h_j''(x)h_\ell''(x)dx$*:*

$$\Omega = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 24 \end{pmatrix}$$

We find the coefficients of the natural spline are $\hat{\beta}^T = (5.00917, 2.94037, -1.14679)$. *The data and smoothing spline are shown in Figure 42.*

10.3.2 Choice of Regularisation Parameter λ

In papplications, λ is usually chosen using cross-validation

$$\widehat{\lambda} = \arg\min_{\lambda>0}\left\{\sum_{i=1}^{n}\left(Y_i - \widehat{f}_{\lambda,-i}(x_i)\right)^2\right\}$$

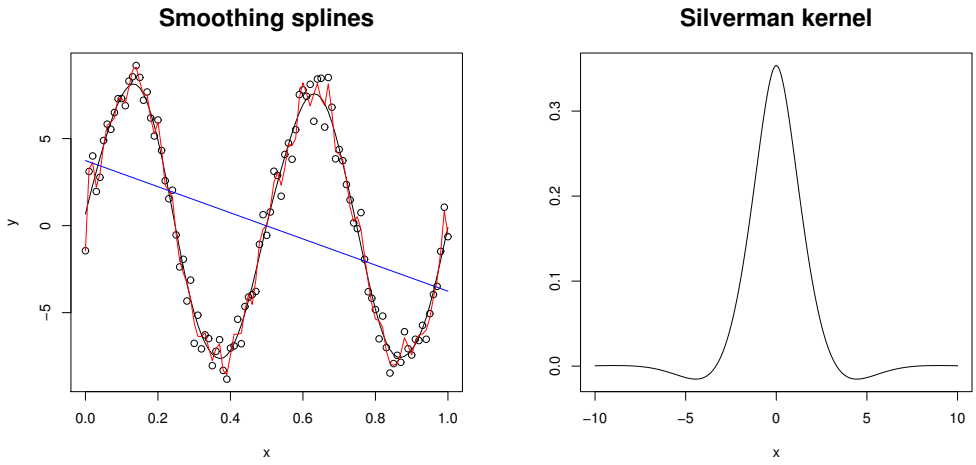


Figure 43: Left: smoothing spline estimator Right: Silverman kernel

where $\widehat{f}_{\lambda,-i}$ is a smoothing spline based on all data points except the i'th. The expression to be minimised is an unbiased estimator of MISE.

Smoothing spline estimators with different regularisation parameters λ are plotted in Figure 43 (Left). The black line corresponds to λ is chosen by cross-validation, the red line - to $\lambda = 0.05$, and the blue line - to $\lambda = 2$. For small $\lambda = 0.05$, where the leading contribution comes from the likelihood, the fitted curve is close to the data points but is not particularly smooth. For larger $\lambda = 2$, the penalisation term dominates the likelihood term, and the linear curve is such that the penalty term is zero (since the second derivative of a linear function is 0). λ chosen by cross-validation provides the estimator with the trade-off between fit to the observed data and smoothness.

10.3.3 Smoothing Spline as a Kernel Estimator

For large N, the smoothing spline is asymptotically equivalent to a kernel estimator:

$$\widehat{f}^{SS}(x) \approx \widehat{f}^{NW}(x),$$

where $\widehat{f}^{NW}(x)$ is the Nadaraya-Watson estimator with the Silverman kernel:

$$K(z) = \frac{1}{2} e^{-|z|/\sqrt{2}} \sin(|z|/\sqrt{2} + \pi/4),$$

plotted in Figure 43 (right), and the bandwidth h can be expressed in terms of λ as $h = \lambda^{1/4}$. Note that this kernel can take negative values. In particular, the smoothing spline has the same optimality properties as a kernel estimator, such as consistency and the optimal rates of convergence.

10.4 Generalized Additive Models

So far we have only talked about regression models with one covariate. However, a more common regression problem would have multiple covariates and take the form

$$Y_i = f(x_{1i}, x_{2i}, \ldots, x_{mi}) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $x_1, \ldots, x_m$ are a set of covariates. Fitting of multivariate regression models is more challenging, not least because large amounts of data are in general required to ensure convergence. The optimal rate of convergence for $f \in H^2(M)$ (i.e., functions with an integrable second derivative) is $n^{-4/5}$ with one covariate, but this degrades to $n^{-4/(4+m)}$ when there are m covariates. If n is the sample size required to achieve a certain accuracy with one covariate, then the sample size required to achieve the same accuracy with m covariates is $n^{(4+m)/5}$ and therefore grows exponentially with m. Nonetheless, generalisations of most univariate nonparametric methods exist and we will describe some of these here.

10.4.1 Multivariate local polynomial regression

Kernel regression can be carried out with multiple covariates, but requires generalisation of the kernel function so that it is a function of m variables. The one-dimensional bandwidth h is replaced by a bandwidth matrix H, allowing a family of kernels to be defined via

$$K_H(\mathbf{x}) = \frac{1}{\sqrt{\det(H)}} K\left(H^{-1/2}\mathbf{x}\right).$$

A common approach is to rescale the covariates so that they have the same mean and variance (at least approximately) and then use an isotropic kernel $h^{-m}K(||\mathbf{x}||_2/h)$ where $K(\cdot)$ is a one-dimensional kernel.

Given a choice of kernel, the local polynomial estimator of order k is found in the same way as before. Firstly we note that an arbitrary function of m variables can be expanded as

$$\begin{aligned} f(x_1, \ldots, x_m) = f(\mathbf{z}) &+ \frac{\partial f}{\partial x_1}(x)(x_1 - z_1) + \frac{\partial f}{\partial x_2}(x)(x_2 - z_2) + \cdots + \frac{\partial f}{\partial x_m}(x)(x_m - z_m) \\ &+ \frac{1}{2!}\left(\frac{\partial^2 f}{\partial x_1^2}(x)(x_1 - z_1)^2 + 2\frac{\partial^2 f}{\partial x_1 \partial x_2}(x)(x_1 - z_1)(x_2 - z_2) + \right. \\ &\qquad \left. \cdots + \frac{\partial^2 f}{\partial x_m^2}(x)(x_m - z_m)^2\right) + \cdots \\ &+ \frac{1}{k!}\left(\frac{\partial^k f}{\partial x_1^k}(x)(x_1 - z_1)^k + \cdots + \frac{\partial^k f}{\partial x_m^k}(x)(x_m - z_m)\right). \end{aligned}$$

There are a total of $M_k = {}_{m+k}C_m = (m+k)!/(m!k!)$ distinct partial derivative terms in this expansion. We can define analogues of the parameter vector θ and the design vector $U_{x,i}$ with this many components

$$\begin{aligned} \theta_{\mathbf{x}} &= (\theta^0, \theta_1^1, \theta_2^1, \ldots, \theta_m^1, \theta_{11}^2, \theta_{12}^2, \ldots \theta_{mm}^2, \ldots, \theta_{mm\cdots m}^k) \\ U_{\mathbf{x},i} &= \left(1, \frac{x_{1i} - x_1}{h}, \frac{x_{2i} - x_2}{h}, \ldots, \frac{x_{mi} - x_m}{h}, \frac{1}{2!}\left(\frac{x_{1i} - x_1}{h}\right)^2, \left(\frac{x_{1i} - x_1}{h}\right)\left(\frac{x_{2i} - x_2}{h}\right), \right. \\ &\qquad \left. \ldots, \frac{1}{2!}\left(\frac{x_{mi} - x_m}{h}\right)^2, \ldots, \frac{1}{k!}\left(\frac{x_{mi} - x_m}{h}\right)^k\right). \end{aligned}$$

In the above, $h^m = \sqrt{\det(H)}$, $\theta^d_{j_1 \dots j_d}$ corresponds to $h^d \partial^d f / \partial x_{j_1} \cdots \partial x_{j_d}$ and the estimator of this quantity provides an estimate of this particular derivative of the function. Note that we must be careful to ensure the ordering of derivatives in θ and $U_{\mathbf{x},i}$ is consistent.

Using this notation the solution for the local polynomial least squares estimator

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^{M_d}} \left\{ \sum_{i=1}^{n} \left(Y_i - U_{\mathbf{x},i}^T \theta_{\mathbf{x}} \right)^2 K_H(\mathbf{x}_i - \mathbf{x}) \right\}$$

takes exactly the same form as before, namely $\hat{\theta}_{\mathbf{x}} = B^{-1}(\mathbf{x}) a(\mathbf{x})$ where

$$B(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} U_{\mathbf{x},i} U_{\mathbf{x},i}^T K_H(\mathbf{x}_i - \mathbf{x}), \qquad a(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} Y_i U_{\mathbf{x},i} K_H(\mathbf{x}_i - \mathbf{x}).$$

10.4.2 Multivariate splines

In a similar way, the notion of a spline can be generalized to more than one dimension. Once again, we aim to minimize the sum of squares, but penalise functions that are not sufficiently smooth. This is formulated in general as

$$\hat{f}^{SS}_{n,\lambda} = \arg\min_f \left\{ \sum_{i=1}^{n} (Y_i - f(x_{1i}, \dots, x_{mi}))^2 + \lambda J_n(f) \right\}$$

where

$$J_n(f) = \int \int \cdots \int \left[\left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_3} \right)^2 + \right.$$
$$\left. \cdots + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_2 \partial x_3} \right)^2 + \cdots \left(\frac{\partial^2 f}{\partial x_m^2} \right)^2 \right] \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_m.$$

The solution to the minimization problem is a **thin plate spline**.

Definition 10.18. *A* **thin plate spline** *through a set of knots* $\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_n$ *in* m*-dimensions, with weights* $w_1, \dots w_n$*, is a function of the form*

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i G(||\mathbf{x} - \mathbf{x}_i||_2) + b_0 + \sum_{j=1}^{m} b_j x_j$$

$$\text{where } G(r) \propto \begin{cases} r^{4-m} \ln r, & m = 2 \text{ or } m = 4 \\ r^{4-m}, & \text{otherwise} \end{cases}, \qquad \text{and } ||\mathbf{x}||_2^2 = \sum_{j=1}^{m} x_j^2.$$

In higher dimensions, $m > 4$, this solution diverges at the knots and so it is not a useful smoothing method. In that case the $m = 2$ basis function, $G(r) = r^2 \ln r$, is often used, or the simple solution $G(r) = r^2$. If these alternative solutions are used the resulting solution is in general not the minimizer for the above problem.

Thin plate splines are difficult to fit and so are not used widely in dimensions higher than 2. It is more common to take an approach that reduces the multi-dimensional fit to a set of one-dimensional fits by using an **additive model**.

10.4.3 Additive models

While the preceding methods provide ways to fit general multivariate nonparametric models, they are often hard to visualize and interpret. This motivates assuming a somewhat simpler form for the unknown function, called an **additive model**.

Definition 10.19. *An additive model is a model of the form*

$$Y_i = \alpha + \sum_{j=1}^{m} f_j(x_j) + \epsilon_i, \qquad i = 1, \ldots, n$$

where $f_1, \ldots, f_m$ are smooth functions.

The model above is not identifiable since a constant can be subtracted from any one of the functions and added to α or any of the other functions to leave the model unchanged. The usual approach to making the model identifiable is to set $\hat{\alpha} = \bar{Y} = \sum_{i=1}^{n} Y_i/n$ and forcing $\sum_{i=1}^{n} \hat{f}_j(x_{ji}) = 0$. The resulting functions can be regarded as representing deviations from the mean $\bar{Y}$.

An additive model can be fitted using any of the techniques for one-dimensional problems that have been described in this course using a procedure known as **backfitting**.

Definition 10.20. *The backfitting algorithm obtains estimates of $\hat{f}_j(x_j)$ in the additive model as follows. Fix the estimator $\hat{\alpha} = \bar{Y}$ and choose initial guesses for $\hat{f}_1, \ldots, \hat{f}_m$. Then*

1. *For $j = 1, \ldots, m$:*
   (a) *Compute $\tilde{Y}_i = Y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ki}), i = 1, \ldots, n$.*
   (b) *Apply a one-dimensional nonparametric fitting procedure (smoother) to $\tilde{Y}_i$ as a function of x_j. Set $\hat{f}_j$ equal to the output of this procedure.*
   (c) *Renormalise by setting $\hat{f}_j(x)$ equal to $\hat{f}_j(x) - \sum_{i=1}^{n} \hat{f}_j(x_{ji})/n$.*
2. *Repeat step 1 until the estimators converge.*

10.4.4 Projection pursuit

Projection pursuit regression attempts to approximate the unknown function $f(x_1, \ldots, x_m)$ by one of the form

$$\mu + \sum_{j=1}^{M} r_j(z_j) \qquad \text{where } z_i = \alpha_i^T \mathbf{x}$$

and each α_i is a unit vector. Projection pursuit attempts to find a transformation of the coordinates that makes an additive model fit as well as possible. In practice, projection pursuit is fitted iteratively, using some one-dimensional nonparametric method. We use $S(w; \mathbf{Y}, \mathbf{x})$ to denote the value of the output of this nonparametric method at a point w, where $\mathbf{x}$ is the vector of (one-dimensional) covariates at the observed points and $\mathbf{Y}$ is the vector of measured values. First set $\hat{\mu} = \bar{Y}$ as before and then initialise the residuals $\hat{\epsilon}_i = Y_i - \bar{Y}$. We use $\hat{\epsilon}$ to denote the vector of current residuals, i.e., $(\hat{\epsilon})_i = \hat{\epsilon}_i$. We also scale the covariates so that their variances are equal and then define an $m \times n$ matrix X such that X_{ij} is the value of the i'th covariate for the j'th data point. Then proceed as follows:

1. Set $j = 0$.
2. Find the unit vector α that minimizes
$$I(\alpha) = 1 - \frac{\sum_{i=1}^{n}(\hat{\epsilon}_i - S(\alpha^T \mathbf{x}_i; \hat{\epsilon}, X^T\alpha))^2}{\sum_{i=1}^{n}\hat{\epsilon}_i^2}$$
and then set $z_{ji} = \alpha^T \mathbf{x}_i$ and $\hat{f}_j(z_{ji}) = S(\alpha^T \mathbf{x}_i; \hat{\epsilon}, X^T\alpha)$.
3. Set $j = j + 1$ and update the residuals
$$\hat{\epsilon}_i \leftarrow \hat{\epsilon}_i - \hat{f}_j(z_{ji}).$$
4. If $j = M$ stop, else return to step 2.

10.4.5 Generalized additive models

Definition 10.21. *An generalized additive model is a model in which observed random variables Y_i are assumed to be drawn from a specified distribution in the exponential family, with a specified link function, $g(\cdot)$, and a model for the expectation value of the form*
$$\eta(\mathbf{x}) = g(\mathbb{E}(Y)) = \alpha + \sum_{j=1}^{m} f_j(x_j)$$
where $f_1, \ldots, f_m$ are smooth functions.

Fitting a generalized additive model can be done iteratively, using a method for fitting a general additive model, in the same way that generalized linear models can be found by fitting general linear models using iterative weighted least squares (Fisher's method of scoring).

The general procedure is as follows:

1. Start with observed data $\{(\mathbf{x}_i, y_i) : i = 1, \ldots n\}$ and initial guesses for $\hat{\alpha}$ and $\hat{f}_1, \ldots, \hat{f}_m$.
2. Then repeat the folliwng steps until the estimates for $\hat{f}_1, \ldots, \hat{f}_m$ converge:
   (a) Compute fitted values
   $$\hat{\eta}(\mathbf{x}_i) = \hat{\alpha} + \sum_{j=1}^{m} \hat{f}_j(x_{mi})$$
   $$\text{and } \hat{r}(\mathbf{x}_i) = g^{-1}(\hat{\eta}(\mathbf{x}_i)).$$
   (b) Computed transformed responses
   $$z_i = \hat{\eta}(\mathbf{x}_i) + (y_i - \hat{r}(\mathbf{x}_i))g'(\hat{r}(\mathbf{x}_i)),$$
   where $g'(\cdot)$ denotes the derivative of the link function.
   (c) Compute weights
   $$w_i = \left[(g'(\hat{r}(\mathbf{x}_i))^2\sigma^2\right]^{-1}.$$

(d) Compute the weighted general additive model for z_i as a function of $\mathbf{x}_i$ with weights w_i.

Note that the above procedure relies on being able to fit a weighted nonparametric model, but all of the methods described above have assumed equal variance. However, it is straightforward to generalise the previous methods to the weighted context. For example, the extension of the Nadaraya-Watson estimator to the weighted case is

$$\hat{f}_n^{wNW}(x) = \frac{\sum_{i=1}^n w_i Y_i K_h(X_i - x)}{\sum_{j=1}^n w_j K_h(X_j - x)}.$$

Example 10.8. *Construct a general additive model, using smoothing splines, on the interval* $[-2,2] \times [-2,2]$ *given data* $(-1,-1,1)$, $(-1,0,3)$, $(-1,1,0)$, $(0,-1,2)$, $(0,0,4)$, $(0,1,1)$, $(1,-1,6)$, $(1,0,3)$, $(1,1,2)$. *Use* $\lambda = 0.01$ *in both dimensions.*

We note that in this case we have data on a regular grid. The backfitting procedure fits a function in one dimension at a time, and so we will need to fit a smoothing spline with multiple observations at a given point. For equal numbers of observations at each point, n_s*, this is a trivial extension of the procedure described above. The spline takes the same form, but we replace* Y_i *by the average of the* $Y_i's$ *at each value of* x*, and we change the smoothing parameter to* λ/n_s*.*

First we estimate $\hat{\alpha} = \bar{Y} = 22/9$ *and subtract this from each point. We then fit a smoothing spline to the data* $(-1,-10/9)$, $(0,-1/9)$, $(1,11/9)$ *using* $\lambda = 0.01/3$*. The* H *and* Ω *matrices are the same as in Example* 3.1

$$H = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 6 \end{pmatrix}, \qquad \Omega = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 24 \end{pmatrix}.$$

and we derive $\hat{\beta}_1 = [(H^T H + \lambda\Omega)^{-1} H^T Y]$ *as before*

$$\hat{\beta}_1^T = (-0.188781, 0.923948, 0.0809061).$$

This gives fitted values at $x = -1, 0, 1$ *of*

$$\hat{f}_1(-1) = -1.11273, \quad \hat{f}_1(0) = -0.107875, \quad \hat{f}_1(1) = 1.2206.$$

We need to correct the fit by subtracting $\sum_{i=1}^3 \hat{f}_1(x_{1i})/3$*, but this number is very close to zero so the values do not change.*

We now need to fit for the second dimension, x_2*. The first stage, in general, is to subtract* $\hat{f}_1(x_{1i})$ *from* Y_i *for each* i*. In this case we have multiple observations at each value of* x_2 *and so we then need to average the* Y_i*'s for each* x_2*. Since the grid is regular, we effectively subtract* $\sum_{i=1}^3 \hat{f}_1(x_{1i})/3$ *from each value, but this has been fixed to equal* 0 *and so does not change the averaged values. This happens generically when the data is on a regular grid and means the backfitting algorithm converges in one iteration.*

The data to fit in x_2 *is* $(-1,5/9)$, $(0,8/9)$, $(1,-13/9)$ *with* $\lambda = 0.01/3$ *again. The* H *and* Ω *matrices are unchanged so we obtain*

$$\hat{\beta}_2^T = (1.51025, 0.941748, -0.647249).$$

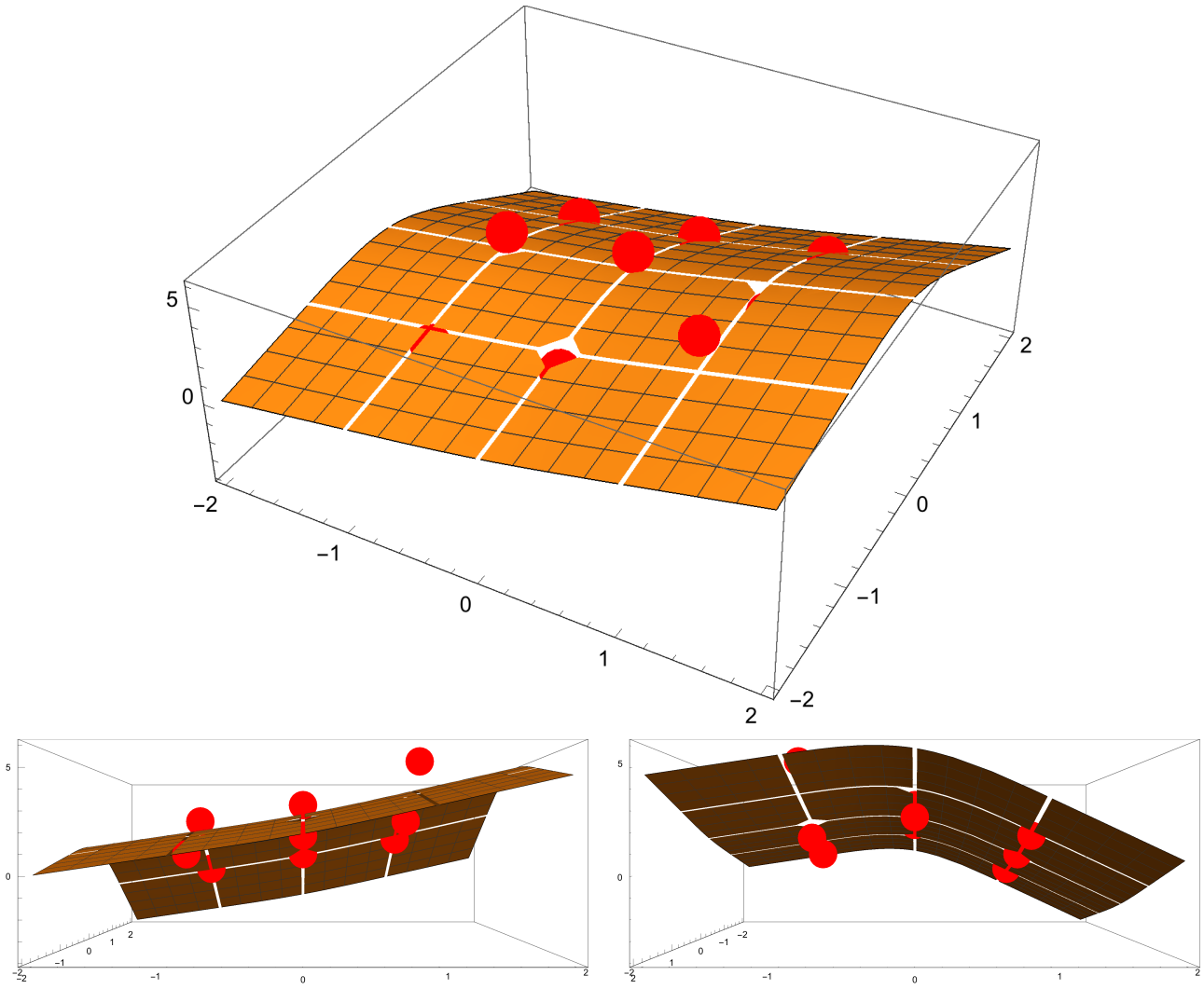


Figure 44: Data (red points) and general additive model fit (shaded surface) for example 10.8. The top plot shows the full surface, while the bottom two plots show the surface from the x_1 and x_2 sides respectively.

The algorithm has now converged and we obtain our general additive model estimate of $f(x_1, x_2)$ *as*

$$\hat{f}(x_1, x_2) = \frac{22}{9} + \sum_{i=1}^{3} \beta_{1i} h_i(x_1) + \sum_{i=1}^{3} \beta_{2i} h_i(x_2)$$

$$\textit{where } h_1(x) = 1, \qquad h_2(x) = x, \qquad h_3(x) = (x+1)_+^3 - 2(x)_+^3 + (x-1)_+^3.$$

The raw data and the GAM estimate are shown in Figure 44.

10.5 Wavelet Estimators

We return again to the nonparametric regression model

$$Y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \qquad \mathbb{E}(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2,\ \text{independently.}$$

In this subsection we will assume that the design is regular deterministic, that is $x_i - x_{i-1} = 1/n$ for all i. In particular, we consider $x_i = \frac{i}{n}$.

10.5.1 Orthonormal basis and projection estimator

We will denote the set of square-integrable functions by $L^2 = \left\{ f : ||f||_2 = \sqrt{\int f^2(x)dx} < \infty \right\}$.

Definition 10.22. *A set of functions $\{\varphi_k(x)\}_{k=0}^{\infty}$ is called an orthonormal basis of $L^2[0,1]$, if*

- $\forall\, g \in L^2, \exists\, (a_k)_{k=0}^{\infty}$ *such that* $g(x) = \sum_{k=0}^{\infty} a_k \varphi_k(x)$ *(the set spans $L^2[0,1]$),*
- $\forall\, x, \sum_{k=0}^{\infty} a_k \varphi_k(x) = 0 \Rightarrow$ *all* $a_k = 0$ *(linear independence),*
- $j \neq k, \int \varphi_k(x)\varphi_j(x) = 0$ *(orthogonality),*
- $\forall\, k, ||\varphi_k||_2 = 1$ *(normalisation).*

Therefore, any function $f \in L^2[0,1]$ can be written as

$$f(x) = \sum_{k=0}^{\infty} \theta_k \varphi_k(x).$$

Due to orthonormality of the basis, the coefficients θ_k have a simple expression: $\theta_k = \int_0^1 f(x)\varphi_k(x)dx$, since

$$\int_0^1 f(x)\varphi_k(x)dx = \int_0^1 \left[\sum_{j=0}^{\infty} \theta_j \varphi_j(x)\right] \varphi_k(x)dx = \sum_{j=0}^{\infty} \theta_j \left[\int_0^1 \varphi_j(x)\varphi_k(x)dx\right] = \theta_k$$

Examples of orthonormal bases:

1. Fourier basis: $\varphi_{2k}(x) = 1$, $\varphi_{2k}(x) = \cos(2\pi kx)$, $\varphi_{2k+1}(x) = \sin(2\pi kx)$, $k = 1, 2, \ldots$, $x \in [0,1]$ (Tsybakov, 2009).
2. A wavelet basis (Vidakovic, 1999)
3. An orthogonal polynomial basis, such as Chebyshev, Lagrange, Laguerre polynomials (more commonly used in the context of density estimation)

Projection estimator

Assume that $f \in L^2[0,1]$, and $\{\varphi_k(x)\}_{k=0}^{\infty}$ is an orthonormal basis of $L^2[0,1]$. Then, we can write

$$f(x) = \sum_{k=0}^{\infty} \theta_k \varphi_k(x)$$

for some real coefficients $\theta_0, \theta_1, \ldots$. A projection estimation of f is based on a simple idea: approximate f by its projection $\sum_{k=0}^{N} \theta_k \varphi_k(x)$ on the linear span of the first $N+1$ functions of the basis, and replace θ_k by their estimators. Thus, a projection estimator is constructed in three steps.

(1) for large N, approximate $f(x) \approx \sum_{k=0}^{N} \theta_k \varphi_k(x)$

(2) construct an estimator $\widehat{\theta}_k$ of θ_k from data $(y_1, \ldots, y_n)$, $k = 0, 1, \ldots, N$

(3) plug in the estimator $\widehat{\theta}_k$ in the approximation: $\widehat{f_N}(x) = \sum_{k=0}^{N} \widehat{\theta}_k \varphi_k(x)$

From the expression for θ_k in terms of f and φ_k, if we know only values of $f(x)$ at points $x_i = i/n$, $i = 1, \ldots, n$, then for large n the integral can be approximated by a sum:

$$\theta_k \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i) \varphi_k(x_i).$$

Since we observe values of $f(x_i)$ with error, we plug in these observation in the above expression to obtain the following estimator for θ_k:

$$\widehat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} Y_i \varphi_k(x_i).$$

Inserting this expression into the estimator of the function, we obtain a **projection estimator**:

$$\widehat{f}_N(x) = \sum_{k=0}^{N} \left[\frac{1}{n} \sum_{i=1}^{n} f(x_i) \varphi_k(x_i) \right] \varphi_k(x) = \sum_{i=1}^{n} Y_i \left[\sum_{k=0}^{N} \frac{1}{n} \varphi_k(x_i) \varphi_k(x) \right]$$

which is a linear estimator with weights $w_i(x) = \sum_{k=0}^{N} \frac{1}{n} \varphi_k(x_i) \varphi_k(x)$ which do not depend on Y_i. The choice of N corresponds to choosing the smoothness of the function $\widehat{f}_N$.

10.5.2 Wavelet basis

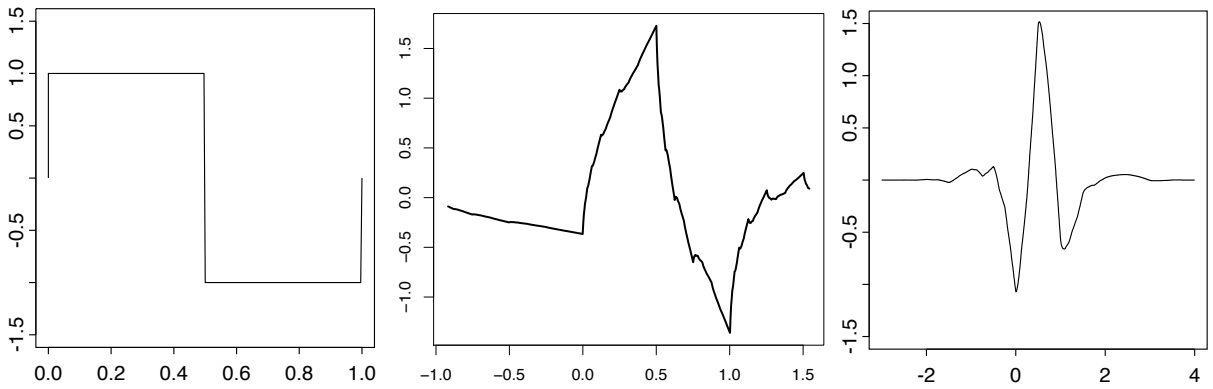
A wavelet basis is constructed using two functions, a scaling function $\phi(x)$ and a wavelet function $\psi(x)$ that are also called the father and mother wavelet respectively. They satisfy the following properties:

$$\int \phi(x) dx = 1, \quad \int \psi(x) dx = 0.$$

Definition 10.23. *Given a wavelet function ψ and a scaling function ϕ, a wavelet basis on $[0, 1]$ is*

$$\{\phi, \psi_{jk}, j = 0, 1, \ldots, k = 0, \ldots, 2^j - 1\},$$

where $\phi_{jk}(x) = 2^{j/2} \phi(2^j x - k)$, $\psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$.



(a) Haar mother wavelet (b) Daubechies mother wavelet, $s = 2$ (c) Daubechies wavelet, $s = 4$

Figure 45: Haar and Daubechies wavelet functions

Under certain additional conditions on the scaling function $\phi(x)$ and the wavelet function $\psi(x)$, this basis is *orthonormal*. Then, any $f \in L^2[0,1]$ can be decomposed in a **wavelet basis**:

$$f(x) = \theta_0\phi(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\theta_{jk}\psi_{jk}(x),$$

and $\theta = \{\theta_0, \theta_{jk}\}$ is a set of **wavelet coefficients**:

$$\theta_0 = \int_0^1 \phi(x)f(x)dx, \quad \theta_{jk} = \int_0^1 \psi_{jk}(x)f(x)dx.$$

Wavelets (ϕ, ψ) are said to have regularity s if they have s derivatives and ψ has s vanishing moments ($\int x^k\psi(x)dx = 0$ for integer $k \le s$).

Examples of wavelet functions are plotted in Figure 45, and the structure of the wavelet basis is illustrated in Figure 46.

Example 10.9. *The Haar wavelet basis is determined by the scaling function $\phi(x) = \mathbf{1}_{(0,1]}(x)$ and the wavelet function $\psi(x) = \mathbf{1}_{(0,1/2]}(x) - \mathbf{1}_{(1/2,1]}(x)$ which satisfy*

$$\int \phi(x)dx = 1, \quad \int \psi(x)dx = 0, \quad \int \psi_{jk}(x)dx = 0.$$

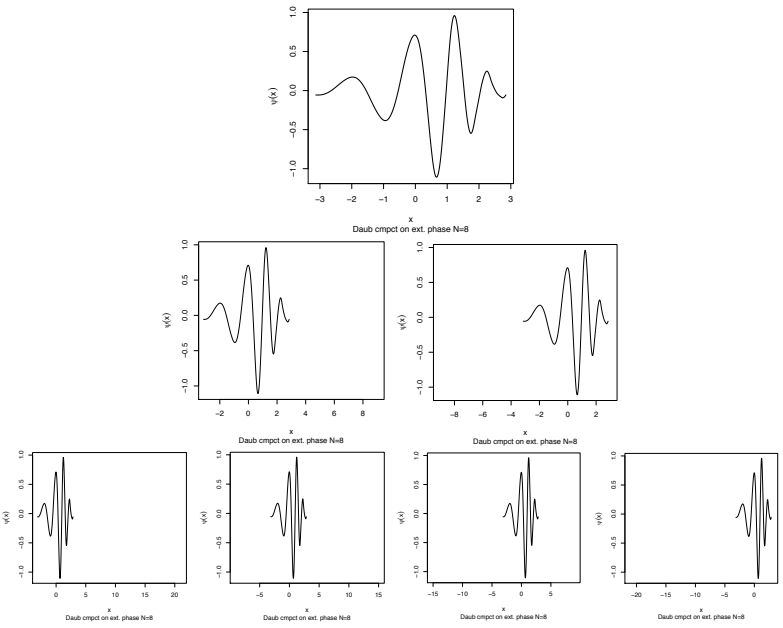
Check that the basis $\{\phi, \psi_{jk}, j = 0, 1, \ldots, k = 0, \ldots, 2^j - 1\}$ defined by these functions is orthonormal, that is, that the functions are normalised

$$||\phi||_2^2 = \int \phi^2(x)dx = 1, \qquad ||\psi||_2^2 = \int \psi^2(x)dx = 1, \qquad ||\psi_{jk}||_2^2 = \int \psi_{jk}^2(x)dx = 1,$$

and are orthogonal:

$$\int \phi(x)\psi_{jk}(x)dx = 0, \qquad \int \psi_{jk}(x)\psi_{\ell m}(x) = 0 \text{ for } (j,k) \neq (\ell, m).$$

Local polynomial and kernel estimators provide localisation in time only. A Fourier basis provides localisation in frequency only. The advantage of a wavelet basis is that it provides localisation in both time and frequency, at the expense of having two indices. The wavelet transform provides a sparse representation of most functions (it is the basis of JPEG2000).



...

Figure 46: Daubechies wavelet transform, $s = 8$

10.5.3 Wavelet estimators

A **wavelet estimator** can be constructed following the same structure as a projection estimator:

1) derive an estimate $\widehat{\theta}_{jk}$ from noisy discrete wavelet coefficients

2) substitute into the series expansion to obtain the estimate of f, to obtain a wavelet estimator $\widehat{f}$:

$$\widehat{f}(x) = \widehat{\theta}_0\phi(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\widehat{\theta}_{jk}\psi_{jk}(x).$$

For example, a **wavelet projection estimator** can be constructed as

$$\widehat{f}_{J_0}(x) = \widehat{\theta}_0\phi(x) + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}\widehat{\theta}_{jk}\psi_{jk}(x),$$

with

$$\widehat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i\phi(x_i), \quad \widehat{\theta}_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_{jk}(x_i),\ j < J_0.$$

From this definition it follows that $\widehat{\theta}_{jk} = 0$ for $j \geq J_0$. It is a linear estimator.

The number of nonzero coefficients of $\widehat{f}_{J_0}(x)$ is

$$1 + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1} 1 = 1 + \sum_{j=0}^{J_0-1} 2^j = 1 + \frac{2^{J_0}-1}{2-1} = 2^{J_0}.$$

Example 10.10. *For the Haar wavelet projection estimator, the variance is*

$$Var(\widehat{f_{J_0}}(x)) = \frac{\sigma^2}{n}\left[(\phi(x))^2 + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}(\psi_{jk}(x))^2\right] = \frac{\sigma^2}{n}\left[1 + \sum_{j=0}^{J_0-1} 2^j\right] = \frac{2^{J_0}}{n}\sigma^2,$$

since $(\phi(x))^2 = 1$ *for all* $x \in [0,1]$, *and* $(\psi_{jk}(x))^2 = 2^j$ *for* (j,k) *such that* $x \in supp(\psi_{jk})$, *i.e. if* $\frac{k}{2^j} \leq x < \frac{k+1}{2^j}$ *(just one* $k = \lfloor x2^j \rfloor$ *for each* j *satisfies this condition).*

We will also consider wavelet thresholding estimators which are examples of nonlinear estimators (see Section 10.5.10).

10.5.4 Multiresolution analysis (MRA)

In this section there is a brief explanation of why wavelet functions, together with the scaling function, form a basis.

Definition 10.24. *A multiresolution analysis (MRA) is a sequence of closed subspaces* V_n, $n \in \{0,1,2,..\}$ *in* $L^2(\mathbb{R})$ *such that*

1. $V_0 \subset V_1 \subset V_2 \subset \ldots, \quad \text{Clos}(\bigcup_j V_j) = L^2(\mathbb{R})$, *where* $\text{Clos}(A)$ *stands for the closure of a set* A.

2. *Subspaces* V_j *are self-similar:*

$$g(2^j x) \in V_j \quad \Leftrightarrow \quad g(x) \in V_0,$$

3. *There exists a scaling function* $\phi \in V_0$ *such that* $\int_{\mathbb{R}} \phi(x)dx \neq 0$ *whose integer-translates span the space* V_0*:*

$$V_0 = \left\{ g \in L^2(\mathbb{R}) : \quad g(x) = \sum_{k \in \mathbb{Z}} c_k \phi(x-k) \text{ for some } (c_k)_{k \in \mathbb{Z}} \right\},$$

and for which the set of functions $\{\phi(\cdot - k), \quad k \in \mathbb{Z}\}$ *is an orthonormal basis.*

Property 2 of MRA implies that for any $h(x) \in V_j$ $\exists\, g \in V_0$ such that

$$h(x) = g(2^j x) = \sum_{k \in \mathbb{Z}} c_k \phi(2^j x - k),$$

and hence $\{\phi(2^j x - k)\}_{k \in \mathbb{Z}}$ or, equivalently, $\{\phi_{jk}\}_{k \in \mathbb{Z}}$, form an orthonormal basis of V_j. In particular, since $\phi(x) \in V_0$ we have

$$\phi(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} h_k \phi(2x - k). \tag{126}$$

The coefficients in this expansion satisfy

$$\sum_k h_k = \sqrt{2}, \qquad \sum_k h_k h_{k-2l} = \delta_{0l}.$$

We then define another function (the mother wavelet)

$$\psi(x) = \sqrt{2} \sum g_k \phi(2x - k)$$

and require that $\psi(x-m)$ is orthogonal to $\phi(x)$ for all integers m, and that $\{\psi(x-m) : m \in \mathbb{Z}\}$ is an orthonormal set. These conditions impose constraints on the coefficients $\{g_k\}$

$$\sum_k g_k h_{k+2m} = 0 \; \forall m \in \mathbb{Z}, \qquad \sum_k g_k g_{k-2l} = \delta_{0l}$$

which can be satisfied by the choice $g_k = (-1)^{1-k} h_{1-k}$. It is clear that the space of functions spanned by $\{\psi(x-m) : m \in \mathbb{Z}\}$, which we denote W_0, is orthogonal to that spanned by $\{\phi(x-m) : m \in \mathbb{Z}\}$, which is V_0. The direct sum $W_0 \oplus V_0$ can be seen to coincide with V_1 (we will not prove this here, but roughly speaking V_1 is twice the size of V_0 so it makes sense that adding two orthogonal spaces of the same size as V_0 together can generate V_1).

We can continue this procedure to larger j. For each $j \geq 0$, we define the *"difference" space* W_j: $V_{j+1} = V_j \oplus W_j$, for which an orthonormal basis is given by $\{\psi_{jk}(x) : k \in \mathbb{Z}\}$. We see that $L^2(\mathbb{R}) = V_0 \oplus W_1 \oplus W_2 \oplus \ldots \oplus W_j \oplus \ldots$, and the set $\{\phi(x),\, \psi_{jk}(x) : j = 0, 1, 2, ..,\ k \in \mathbb{Z}\}$ forms an orthonormal basis of $L^2(\mathbb{R})$.

10.5.5 Filter characterisation of the wavelet transform

We now prove some of the results used to describe the MRA above.

Proposition 10.4. *1.* $\sum_{k\in\mathbb{Z}} h_k = \sqrt{2}, \quad \sum_{k\in\mathbb{Z}} g_k = 0$

2. $\sum_{k\in\mathbb{Z}} h_k^2 = 1, \quad \sum_{k\in\mathbb{Z}} g_k^2 = 1$

3. For all $\ell \neq 0$, $\quad \sum_{k\in\mathbb{Z}} h_k h_{k-2\ell} = 0, \quad \sum_{k\in\mathbb{Z}} g_k g_{k-2\ell} = 0$

4. For all $\ell \in \mathbb{Z}$, $\quad \sum_{k\in\mathbb{Z}} g_k h_{k-2\ell} = 0.$

Proof of Properties 1 and 2. 1. To prove $\sum_{k\in Z} h_k = \sqrt{2}$, we integrate the scaling equation:

$$\begin{aligned} 1 &= \int \phi(x)dx = \sum_{k\in Z} h_k \sqrt{2} \int \phi(2x-k)dx = [v = 2x-k] = \sum_{k\in Z} h_k 2^{-1/2} \int \phi(v)dv \\ &= \frac{1}{\sqrt{2}} \sum_{k\in Z} h_k \end{aligned}$$

which implies the result.

Similarly, to prove $\sum_{k\in Z} g_k = 0$, we integrate the wavelet equation:

$$\begin{aligned} 0 &= \int \psi(x)dx = \sqrt{2} \sum_{k\in Z} g_k \int \phi(2x-k)dx = [v = 2x-k] = 2^{-1/2} \sum_{k\in Z} g_k \int \phi(v)dv \\ &= 2^{-1/2} \sum_{k\in Z} g_k \end{aligned}$$

which implies that $\sum_{k\in Z} g_k = 0$.

2. To prove $\sum_{k\in Z} h_k^2 = 1$, we integrate the squared scaling equation:

$$\begin{aligned} 1 &= \int \phi(x)^2 dx = 2 \int \left[\sum_{k\in Z} h_k \phi(2x-k)\right]^2 dx = \sum_{k,m} h_k h_m \int \phi(2x-k)\phi(2x-m)d(2x) \\ &= \sum_k h_k^2 \end{aligned}$$

since $\int \phi(2x-k)\phi(2x-m)d(2x) = 1$ if $k = m$ and is 0 otherwise.

$\sum_{k\in Z} g_k^2 = 1$ is proved similarly, by integrating the squared wavelet equation. □

The two filter decompositions (for $\phi(x)$, with coefficients $\{h_k\}$ and $\psi(x)$ with coefficients $\{g_k\}$ satisfying $g_k = (-1)^k h_{1-k}$) have other properties which we will use later to show that a finite dimensional version of wavelet decomposition, a discrete wavelet transform performed via the cascade algorithm, transforms iid Gaussian random variables to iid Gaussian random variables.

Example 10.11. *Determine filters g_k, h_k for the Haar wavelet transform.*

For the Haar wavelets, the scaling equation is

$$\mathbf{1}_{(0,1]}(x) = \mathbf{1}_{(0,1/2]}(x) + \mathbf{1}_{(1/2,1]}(x) = \mathbf{1}_{(0,1]}(2x) + \mathbf{1}_{(0,1]}(2x-1)$$

That is,

$$\phi(x) = \phi(2x) + \phi(2x-1) = \sqrt{2}\sum_{k\in\mathbb{Z}} h_k \phi(2x-k)$$

which implies that the only nonzero values of h_k are $h_0 = h_1 = 1/\sqrt{2}$.

The Haar wavelet function satisfies the following:

$$\psi(x) = \mathbf{1}_{(0,1/2]}(x) - \mathbf{1}_{(1/2,1]}(x) = \mathbf{1}_{(0,1]}(2x) - \mathbf{1}_{(0,1]}(2x-1) = \frac{1}{\sqrt{2}}\left(\phi(2x) - \phi(2x-1)\right)$$

which implies that $g_0 = 1/\sqrt{2}$, $g_1 = -1/\sqrt{2}$ and the remaining g_k are 0.

10.5.6 Discrete wavelet transform (DWT)

In typical realistic settings, we observe only a finite number of noisy values of the function. How can we obtain (noisy) wavelet coefficients based on this partial information?

10.5.7 Motivation

We want to discretise the wavelet transform:

$$\theta_{jk} = \int_0^1 f(x)\psi_{jk}(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} \psi_{jk}(i/n) f(i/n) = \frac{1}{\sqrt{n}}(Wf_n)_{(jk)} = \frac{w_{jk}}{\sqrt{n}} =: \tilde{\theta}_{jk},$$

where W, an $n \times n$ matrix defined by $W_{1i} = \phi(x_i)$, $W_{li} = \psi_{jk}(x_i)$ with $l = 2^j + k + 1$, is (approximately) orthonormal and f_n is a vector $f_n = (f(1/n), \ldots, f(1))$. We assume $n = 2^J$ for some integer J. The subscript (jk) in the above denotes the row, $l = 2^j + k + 1$, corresponding to a particular pair (j, k).

If the function f is bounded, the approximate wavelet coefficients $\tilde{\theta}_{jk}$ are close to the exact coefficients θ_{jk}: $|\tilde{\theta}_{jk} - \theta_{jk}| \le C/n$. For Haar wavelets, $\theta_{jk} = \tilde{\theta}_{jk}$ since the Haar wavelets are constants on each interval $(i/n, (i+1)/n)$ for $n = 2^J$ for some integer J.

Use the linear transform defined by a matrix W as a discrete wavelet transform. There are other ways to derive the approximation, so that $|\tilde{\theta}_{jk} - \theta_{jk}| \le C/n$ and matrix W is orthonormal ($WW^T = I$). In practice, it is done via the **cascade algorithm** which is derived from filter properties of wavelet transform. In this case, $|\tilde{\theta}_{jk} - \theta_{jk}| \le C/n$ and the matrix W satisfies $WW^T = I$ due to the filter properties (Proposition 10.4).

Applying the discretised wavelet transform W to data yields

$$\begin{aligned} d_{jk} &= w_{jk} + \varepsilon_{jk}, \quad 0 \le j \le J-1,\ k = 0, \ldots, 2^j - 1, \\ c_{00} &= u_{00} + \varepsilon_0, \end{aligned}$$

where d_{jk} and c_{00} are discrete wavelet and scaling coefficients of observations (y_i), and ε_{jk} and ε_0 are discrete wavelet coefficients of the noise (ϵ_i). If $\epsilon_i \sim N(0, \sigma^2)$ independent, then $\varepsilon_{jk} \sim N(0, \sigma^2)$ and $\varepsilon_0 \sim N(0, \sigma^2)$ independently due to $WW^T = I$.

10.5.8 Cascade algorithm

The wavelet and scaling equations are the basis for the cascade algorithm that can be used to calculate the wavelet coefficients. The algorithm is very fast, taking $2n$ steps where n is the number of the observations. The algorithm is constructed by using recurrent equations for wavelet and scaling coefficients that are derived from the wavelet and the scaling equations in the following way.

Suppose we observe values of $f(x_i)$, $x_i = i/n$, $i = 1, \ldots, n$. Denote the corresponding "noiseless" discrete scaling coefficients by u_{jk} and discrete wavelet coefficients by w_{jk} (recall that $\theta_{jk} \approx w_{jk}/\sqrt{n}$ and $\theta_0 \approx u_{00}/\sqrt{n}$). Then, the wavelet coefficients satisfy the following (using the wavelet equation):

$$\begin{aligned}
\theta_{jk} &= \int_0^1 f(x)\psi_{jk}(x)dx = \int_0^1 f(x)\psi(2^j x - k)2^{j/2}dx \\
&= \int_0^1 f(x)\left[\sqrt{2}\sum_{m\in Z} g_m \phi\left(2(2^j x - k) - m\right)\right] 2^{j/2}dx \\
&= \int_0^1 f(x)\left[\sum_{m\in Z} g_m \phi\left(2^{j+1}x - 2k - m\right) 2^{(j+1)/2}\right] dx \\
&= \sum_{m\in Z} g_m \int_0^1 f(x)\phi_{j+1,2k+m}(x)dx.
\end{aligned}$$

Here, $\int_0^1 f(x)\phi_{jk}(x)dx$ are scaling coefficients of f that are not used directly for estimation but are useful for computational purposes. For the discrete wavelet and scaling coefficients w_{jk} and u_{jk}, we can write the following recurrence relation:

$$w_{jk} = \sum_{m\in Z} g_m u_{j+1,2k+m}.$$

Using the scaling equation, we can derive a similar connection between the scaling coefficients at consecutive levels j and $j+1$:

$$u_{jk} = \sqrt{n}\int_0^1 f(x)\phi_{jk}(x) = \sum_{m\in Z} h_m u_{j+1,2k+m}.$$

These recurrence equations are used in the cascade algorithm. They also apply to noisy scaling and wavelet coefficients c_{jk} and d_{jk}.

We need to have a starting point. Assuming that $\text{supp}(\phi) = [0, 1]$, like for the Haar scaling function, the scaling coefficients at level J for $k = 0, 1, .., 2^J - 1$ satisfy:

$$\begin{aligned}
&\int_0^1 f(x)2^{J/2}\phi(2^J x - k)dx = 2^{J/2}\int_{k/2^J}^{(k+1)/2^J} f(x)\phi(2^J x - k)dx \\
\approx\ & f((k+1)/n)\int_{k/2^J}^{(k+1)/2^J} 2^{J/2}\phi(2^J x - k)dx = [v = 2^J x - k] = f(x_{k+1})2^{-J/2}\int_0^1 \phi(v)dv \\
\approx\ & \frac{f(x_{k+1})}{\sqrt{n}}.
\end{aligned}$$

Therefore, we can set $u_{J,k} = f(x_{k+1})$, $k = 0, 1, \ldots, 2^J - 1 = n - 1$. For noisy observations (Y_i), we can start with noisy discrete scaling coefficients $c_{J,k} = Y_{k+1}$.

Assumptions for the cascade algorithm.

1. Y_i are (noisy) observations of a function f at points x_i, $i = 1, .., n$
2. points (x_i) form a regular fixed design $(x_i - x_{i-1} = \frac{1}{n})$.
3. $n = 2^J$ for some integer J.

Cascade algorithm

1. Set $c_{Jk} = Y_{k+1}$ for $k = 0, 1, .., 2^J - 1$, set $j = J - 1$;
2. Set
$$c_{jk} = \sum_{m\in\mathbb{Z}} h_m c_{j+1,2k+m}, \quad d_{jk} = \sum_{m\in\mathbb{Z}} g_m c_{j+1,2k+m};$$
3. if $j = 0$ stop; else set $j := j - 1$ and repeat step 2.

Output: discrete wavelet coefficients c_{00}, d_{jk} for $0 \le j \le J - 1, \quad k = 0, \ldots, 2^j - 1$.

Using the expressions for the Haar wavelet filters h_k and g_k, the recurrent step of the cascade algorithm for the Haar wavelet transform is

$$u_{jk} = \frac{1}{\sqrt{2}} \left(u_{j+1,2k} + u_{j+1,2k+1} \right), \quad w_{jk} = \frac{1}{\sqrt{2}} \left(u_{j+1,2k} - u_{j+1,2k+1} \right).$$

To reconstruct the function from the wavelet coefficients, this algorithm can be inverted.

10.5.9 Summary

- The number of data points $n = 2^J$.
- Cascade algorithm: set $c_{J0} = Y_1, \ldots, c_{J,2^J-1} = Y_n$, and compute recursively
$$c_{jk} = \sum_m h_m c_{j+1,2k+m}, \quad d_{jk} = \sum_m g_m c_{j+1,2k+m}.$$
- The output of the the cascade algorithm are discrete wavelet coefficients: c_{00} & d_{jk}, $j < J$ that satisfy
$$d_{jk} \sim N(w_{jk}, \sigma^2), \qquad c_{00} \sim N(u_{00}, \sigma^2), \quad \text{independently.}$$
- To construct an estimator of f, choose estimators $\widehat{w}_{jk}$, $\widehat{u}_{00} (= c_{00})$, and hence construct the corresponding estimators
$$\widehat{\theta}_0 = \frac{\widehat{u}_{00}}{\sqrt{n}}, \qquad \widehat{\theta}_{jk} = \frac{\widehat{w}_{jk}}{\sqrt{n}}.$$
These estimators are then used to obtain an estimator of the function f:
$$\widehat{f}(x) = \widehat{\theta}_0 \phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \widehat{\theta}_{jk} \psi_{jk}(x).$$

For example, a linear projection estimator $\widehat{f_{J_0}}(x)$ for $f(x)$ can be constructed using the output of the cascade algorithm:

$$\widehat{w}_{jk} = d_{jk},\ j \le J_0 - 1; \quad \widehat{w}_{jk} = 0,\ j \ge J_0; \quad \widehat{u}_{00} = c_{00}.$$

For Haar wavelets, the linear projection estimator $\widehat{f_{J_0}}$ coincides with the wavelet estimator based on discrete wavelet coefficients with $\widehat{w}_{jk} = d_{jk}$ for $j \le J_0 - 1$ and $\widehat{w}_{jk} = 0$ for $j > J_0$.

10.5.10 Thresholding Estimators for threshold λ

Hard thresholding estimator

$$\widehat{w}_{jk} = d_{jk} I\left(|d_{jk}| > \lambda\right) = \begin{cases} d_{jk}, & \text{if } |d_{jk}| > \lambda \\ 0, & \text{if } |d_{jk}| < \lambda \end{cases}$$

Soft thresholding estimator

$$\widehat{w}_{jk} = \begin{cases} d_{jk} - \lambda, & d_{jk} > \lambda \\ 0, & -\lambda \le d_{jk} \le \lambda \\ d_{jk} + \lambda, & d_{jk} < -\lambda \end{cases}$$

There is a default choice of threshold λ that is called the *universal threshold*:

$$\lambda = \sigma\sqrt{2\log n}.$$

In practice, the standard deviation σ is estimated as the median absolution deviation (MAD):

$$\widehat{\sigma} = 1.4826\ \mathrm{MAD}(d_{J-1,0}, \ldots, d_{J-1,2^J-1})$$

where $\mathrm{MAD}(x_1, \ldots, x_n) = \mathrm{median}(|x_i - \mathrm{median}(x_i)|)$.

10.5.11 Inference on f using wavelet estimators

10.5.12 Asymptotic confidence intervals for $f(x)$

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i = \frac{i}{n} \quad \varepsilon_i \sim N(0, \sigma^2)$$

To construct an asymptotic confidence interval for $f(x)$, we use the linear estimator

$$\widehat{f_{J_0}(x)} = \widehat{\theta}_0\phi(x) + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^{j_0}-1} \widehat{\theta}_{jk}\psi_{jk}(x),$$

where

$$\widehat{\theta}_0 = \frac{1}{\sqrt{n}}\widehat{u}_{00}, \quad \widehat{u}_{00} = c_{00} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\phi(x_i)$$

$$\widehat{\theta}_{jk} = \frac{1}{\sqrt{n}}\widehat{w}_{jk}, \quad \widehat{w}_{jk} = d_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_{jk}(x_i)$$

Recall that this estimator is linear:

$$\Rightarrow \widehat{f_{J_0}(x)} = \sum_{i=1}^{n} w_i(x) Y_i, \quad w_i(x) = \frac{1}{n}\phi(x_i)\phi(x) + \frac{1}{n}\sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}\psi_{jk}(x_i)\psi_{jk}(x),$$

therefore, given independent observations of $Y_i \sim N(f(x_i), \sigma^2)$ for $i = 1, \ldots, n$,

$$\widehat{f_{J_0}(x)} \sim N\left(f(x), \sigma^2 \sum_{i=1}^{n} w_i^2(x)\right) \quad \text{for large } n.$$

For Haar wavelets, we derived that $\sum_{i=1}^{n} w_i^2(x) = 2^{J_0}/n$.

Therefore, an asymptotic $(1-\alpha)100\%$ confidence interval for $f(x)$ based on the Haar wavelets projection estimator $\widehat{f_{J_0}}(x)$, assuming that J_0 is large enough so that the bias is much smaller than the variance, is

$$\widehat{f_{J_0}}(x) \pm z_{\alpha/2}\frac{2^{J_0/2}\sigma}{\sqrt{n}}.$$

Note that if J_0 is too large, then the confidence interval is large. Therefore, there is a tradeoff between bias and variance that results in "optimal" choice of J_0. This is discussed by considering the MISE of $\widehat{f_{J_0}}(x)$.

10.5.13 Hypothesis testing

Local support of the wavelet basis is useful when it is of interest to test whether a function is a constant on a certain subinterval of $[0, 1]$. We want to test the hypothesis

$$H_0 : f(x) = \textit{constant} \text{ on } (a, b)$$

using Haar wavelets.

Due to the support of ψ_{jk} being $[k/2^j, (k+1)/2^j]$, for $(a, b) = (m2^{-\ell}, (m+1)2^{-\ell})$ for some positive integers m and ℓ this hypothesis is equivalent to the following hypothesis about the Haar wavelet coefficients of function f:

$$H_0 : \theta_{jk} = 0 \text{ for } (j, k) \text{ such that } a < \frac{k+1/2}{2^j} < b$$

that is, the change point of ψ_{jk} is inside (a, b). The equivalent null hypothesis can also be written as

$$H_0 : w_{jk} = 0 \text{ for } (j, k) \text{ such that } a < \frac{k+1/2}{2^j} < b$$

since $(\theta_{jk} = w_{jk}/\sqrt{n})$ for Haar wavelets.

Test this hypothesis using observed discrete wavelet coefficients $d_{jk} \sim N(w_{jk}, \sigma^2)$, $j = 0, \ldots, J-1$, $k = 0, \ldots, 2^j - 1$, independently.

Given only $n = 2^J$ observations, we can test this hypothesis only using the wavelet coefficients with $j < J$:

$$H_0 : w_{jk} = 0 \text{ for } (j, k) \text{ such that } a < \frac{k+1/2}{2^j} < b \,\&\, j < J.$$

Test statistic:

$$T = \sigma^{-2} \sum_{j,k:\, a<\frac{k+1/2}{2^j}<b,\, j<J} d_{jk}^2$$

which has a χ^2_m distribution under the null hypothesis where m is the number of coefficients tested to be zero, that is, $m = \text{Card}\{(j,k) : a < \frac{k+1/2}{2^j} < b,\, 0 \le j < J,\, 0 \le k \le 2^j - 1\}$.

Example 10.12. *Data:* $\mathbf{y} = (-1.0, -0.2, 0.8, 0.6, 0.0, -0.4, -0.3, -0.5)$, $x_i = i/8$, $i = 1, .., 8$, $n = 8$. *The data follows the nonparametric regression model with* $\sigma = 0.2$.

1. Test $H_0 : f(x) = const$ *on* $(1/4, 1/2)$.

Corresponding hypothesis for the wavelet coefficients is $H_0 : w_{jk} = 0$ *for* (j,k) *that satisfy* $1/4 < \frac{k+1/2}{2^j} < 1/2$, $j < J - 1 = 2$ *then* $(2^j/4 - 1/2) < k < 2^j/2 - 1/2$

Since $n = 8 = 2^3$, *we have* $J = 3$ *and hence we consider* $0 \le j \le 2$*:*
$j = 2$*:* $1/2 < k < 3/2$, *i.e.* $k = 1$ *and hence* $(j,k) = (2,1)$ *satisfies the condition*
$j = 1$*:* $0 < k < 1/2$ *no integer in the interval, so none*
$j = 0$*:* $-1/4 < k < 0$ *none.*

Therefore, the equivalent hypothesis is $H_0 : w_{21} = 0$. *Since the corresponding noisy discrete Haar wavelet coefficient* $d_{21} \sim N(w_{21}, \sigma^2)$, *under the null hypothesis* $T = d_{21}^2/\sigma^2 \sim \chi^2_1$, *therefore we reject* H_0 *at a 5% significance level if* $T = d_{21}^2/\sigma^2 > \chi^2_1(5\%) = 3.841$. *Since for this data* $d_{21} = 0.1414$ *and hence* $T = d_{21}^2/\sigma^2 = 0.5 < 3.841$, *there is not sufficient data to reject the null hypothesis at a 5% significance level.*

2. Now test $H_0 : f(x) = const$ *on* $(1/2, 1)$.

The corresponding hypothesis for the wavelet coefficients is $H_0 : w_{jk} = 0$ *for* (j,k) *s.t.* $1/2 < \frac{k+1/2}{2^j} < 1$, *that is, for* (j,k) *such that*
$\Leftrightarrow 2^j/2 - 1/2 < k < 2^j - 1/2$.

$j \le J - 1 = 2$. *Check this condition for each* $0 \le j \le 2$*:*
$j = 2$*:* $3/2 < k < 7/2$, *that is,* $k = 2, 3$
$j = 1$*:* $1/2 < k < 3/2$, *that is,* $k = 1$
$j = 0$*:* $0 < k < 1/2$ *none*

Therefore, the equivalent hypothesis is

$$H_0 : w_{11} = w_{22} = w_{23} = 0.$$

The test statistic is $T = (d_{11}^2 + d_{22}^2 + d_{23}^2)/\sigma^2 \sim \chi^2_3$ *under* H_0. *That is, we reject the null hypothesis at a 5% significance level if* $T > \chi^2_3(5\%) = 7.815$. *For this data,* $T = (0.2^2 + 0.2828427^2 + 0.1414214^2)/0.04 = 3.5 < 7.815$, *therefore there is not sufficient data to reject the null hypothesis at a 5% significance level.*

Remark 10.2. *For an arbitrary interval* (a,b) *(that is, not of the form* $(m2^{-\ell}, (m+1)2^{-\ell})$*), the equivalent null hypothesis in terms of Haar wavelet coefficients is*

$$H_0 : w_{jk} = 0 \text{ for } (j,k) \text{ such that } \{a < \frac{k}{2^j} < b \text{ or } a < \frac{k+1/2}{2^j} < b \text{ or } a < \frac{k+1}{2^j} < b\},$$

for $j = 0, 1, \ldots, J-1$ *and* $k = 0, 1, \ldots, 2^j - 1$. *That is, in the more general case we need to check if any of the three points where the Haar wavelet* ψ_{jk} *jumps between different constant values is inside the interval* (a,b).

For an interval of the type $(m2^{-\ell}, (m+1)2^{-\ell})$ *it is not necessary to check the end point since they are either at the same place with regard to* (a,b) *(that is, inside or outside) as the mid point* $(k + 1/2)2^{-j}$ *or on the boundary of the interval.*

10.5.14 MISE (mean integrated square error) of wavelet estimators

Suppose a function f has the following wavelet decomposition:

$$f(x) = \theta_0\phi(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\theta_{jk}\psi_{jk}(x),$$

and consider a wavelet estimator

$$\hat{f}(x) = \hat{\theta}_0\phi(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\hat{\theta}_{jk}\psi_{jk}(x).$$

Lemma 10.2. *(Parseval identity). For a function f and its wavelet estimator $\hat{f}(x)$,*

$$||f - \hat{f}||_2^2 = (\theta_0 - \hat{\theta}_0)^2 + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}(\hat{\theta}_{jk} - \theta_{jk})^2.$$

This is due to the wavelet basis being orthonormal.

Consider the following estimator of the wavelet coefficients for $j = 0, .., J_0 - 1$ for some J_0:

$$\hat{\theta}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)Y_i,$$

and $\hat{\theta}_{jk} = 0$ for $j \geq J_0$. The estimator of the scaling coefficient is $\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)Y_i$. Sometimes we refer to θ_0 as $\theta_{-1,0}$, and to $\phi(x)$ as $\psi_{-1,0}(x)$.

The corresponding wavelet estimator is

$$\hat{f}_{J_0}(x) = \sum_{j\leq J_0-1}\sum_{k}\hat{\theta}_{jk}\psi_{jk}(x) = \frac{1}{n}\sum_{i=1}^{n}Y_i\sum_{j\leq J_0-1}\sum_{k}\psi_{jk}(x_i)\psi_{jk}(x).$$

This wavelet estimator

$$\hat{f}_{J_0}(x) = \frac{1}{n}\sum_{i=1}^{n}Y_i\sum_{j\leq J_0-1}\sum_{k}\psi_{jk}(x_i)\psi_{jk}(x)$$

is linear since it can be written as

$$\hat{f}_{J_0}(x) = \sum_{i=1}^{n}Y_iW_i(x),$$

with $W_i(x) = \frac{1}{n}\sum_{j\leq J_0-1,k}\psi_{jk}(x_i)\psi_{jk}(x)$, i.e., that is independent of the Y_i's.

By Lemma 10.2,

$$\mathbb{E}||f - \hat{f}||_2^2 = \mathbb{E}(\theta_0 - \hat{\theta}_0)^2 + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2,$$

hence it is sufficient to find MSE of $\hat{\theta}_{jk}$, $\mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2$.

We know that

$$\mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2 = \mathrm{Var}(\hat{\theta}_{jk}) + \left[\mathrm{bias}(\hat{\theta}_{\mathrm{jk}})\right]^2.$$

Therefore, we need to find the variance and the bias of $\hat{\theta}_{jk}$.

Variance

For $j \leq J_0 - 1$,

$$\begin{aligned}
\mathrm{Var}\left(\hat{\theta}_{jk}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)Y_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\psi_{jk}^2(x_i)\mathrm{Var}\left(Y_i\right) = \frac{\sigma^2}{n}\,\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}^2(x_i) \\
&= \frac{\sigma^2}{n}(1+o(1)),
\end{aligned}$$

due to the independence of the Y_i's and $\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}^2(x_i) \approx \int_0^1 \psi_{jk}^2(x)dx = 1$.

Bias

For $j \leq J_0 - 1$, the bias is

$$\mathbb{E}\left(\hat{\theta}_{jk} - \theta_{jk}\right) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)\psi_{jk}(x_i) - \int_0^1 f(x)\psi_{jk}(x)dx.$$

Assume that $f \in H^{\beta}(M_f)$ and is bounded, i.e. $|f(x)| \leq C_f$ for all $x \in [0,1]$. We assume that the wavelet function ψ is such that $|\psi(x) - \psi(y)| \leq M_\psi|x-y|$ for all $x, y \in [0,1]$, and it is bounded: $|\psi(x)| \leq C_\psi$ for all $x \in [0,1]$ (and that the same conditions hold for the scaling function ϕ). We also assume that $\mathrm{supp}(\psi) \subseteq [0,1]$ and $\mathrm{supp}(\phi) \subseteq [0,1]$.

Under these assumptions with $\beta \in (0,1]$, the absolute value of the bias is bounded by

$$\begin{aligned}
\left|\mathbb{E}\left(\hat{\theta}_{jk} - \theta_{jk}\right)\right| &\leq \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left|f(x)\psi_{jk}(x) - f(x_i)\psi_{jk}(x_i)\right|dx \\
&\leq \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left[\left|f(x)\psi_{jk}(x) - f(x)\psi_{jk}(x_i)\right| + \left|f(x)\psi_{jk}(x_i) - f(x_i)\psi_{jk}(x_i)\right|\right]dx \\
&\leq \max_x |f(x)| 2^{j/2}\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left|\psi(2^j x - k) - \psi(2^j x_i - k)\right|dx \\
&\quad + \sum_{i=1}^{n}|\psi_{jk}(x_i)|\int_{x_{i-1}}^{x_i}\left|f(x) - f(x_i)\right|dx.
\end{aligned}$$

Considering the first term on the right hand side, we have

$$\begin{aligned}
\int_{x_{i-1}}^{x_i}\left|\psi(2^j x - k) - \psi(2^j x_i - k)\right|dx &\leq M_\psi\int_{x_{i-1}}^{x_i}|2^j x - k - (2^j x_i - k)|dx \\
&\leq 0.5 M_\psi 2^j n^{-2}.
\end{aligned}$$

The intersection of the interval of integration $[(i-1)/n, i/n]$ and the support of ψ_{jk}

$$\mathrm{supp}(\psi_{jk}) = [k2^{-j}, (k+1)2^{-j}] = [k2^{J-j}/n, (k+1)2^{J-j}/n]$$

is nonempty (and consists of more than a single point) iff $k2^{J-j} < i-1 < (k+1)2^{J-j}$ or $k2^{J-j} < i < (k+1)2^{J-j}$, i.e. $k2^{J-j}+1 \le i <\le (k+1)2^{J-j}$. There are 2^{J-j} of such i. Thus,

$$\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}|\psi_{jk}(x)-\psi_{jk}(x_i)|dx \le 0.5M_\psi 2^j n^{-2}2^{J-j} = 0.5M_\psi n^{-2}2^J = 0.5M_\psi n^{-1},$$

using $n=2^J$ and hence

$$\max_x |f(x)|2^{j/2}\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left|\psi(2^jx-k)-\psi(2^jx_i-k)\right|dx \le 0.5C_fM_\psi 2^{j/2}n^{-1}.$$

For the second term, we have

$$\int_{x_{i-1}}^{x_i}|f(x)-f(x_i)|\,dx \;\;\le\;\; M_f\int_{x_{i-1}}^{x_i}|x-x_i|^\beta \le \frac{M_f}{(\beta+1)n^{\beta+1}},$$

and using the restriction to the support of ψ_{jk}

$$|\psi_{jk}(x_i)| \le 2^{j/2}C_\psi \mathbf{1}(k2^{J-j}+1 < i < (k+1)2^{J-j}),$$
$$\Rightarrow \sum_{i=1}^{n}|\psi_{jk}(x_i)| \le 2^{j/2}C_\psi\sum_{i=1}^{n}\mathbf{1}(k2^{J-j}+1 \le i \le (k+1)2^{J-j}) \le 2^{J-j/2}C_\psi \le C_\psi n2^{-j/2}.$$

Thus,

$$|\mathbb{E}\hat\theta_{jk}-\theta_{jk}| \;\;\le\;\; 0.5C_fM_\psi 2^{j/2}n^{-1} + \frac{M_fC_\psi}{(\beta+1)}2^{-j/2}n^{-\beta}$$

again using $n=2^J$ and $j<J$.

MSE $(\hat\theta_{jk})$ for $j \ge J_0$

For $j \ge J_0$, $\hat\theta_{jk}=0$, and therefore the MSE $(\hat\theta_{jk}) = \mathbb{E}(\hat\theta_{jk}-\theta_{jk})^2 = \theta_{jk}^2$.
For $f \in H^\beta(M_f)$, $|\theta_{jk}| \le M_f 2^{-j(\beta+1/2)}$ for all j,k.

Now we summarise the properties of **bias and variance of $\hat\theta_{jk}$** that we have derived.

Lemma 10.3. *Assume that*

- *$f \in H^\beta(M_f)$, $\beta \in (0,1)$, and $|f(x)| \le C_f$ for all $x \in [0,1]$;*
- *ψ is such that supp$(\psi) \subseteq [0,1]$, $|\psi(x)-\psi(y)| \le M_\psi|x-y|$ for all $x,y \in [0,1]$, and it is bounded: $|\psi(x)| \le C_\psi$ for all $x \in [0,1]$ (and that the same conditions hold for the scaling function ϕ).*

Then, for $\hat\theta_{jk} = \frac{1}{n}\sum_{i=1}^n \psi_{jk}(x_i)Y_i$,

$$\begin{aligned} Var\left(\hat\theta_{jk}\right) &= \frac{\sigma^2}{n}(1+o(1)) \quad \text{as } n\to\infty, \\ |\text{bias}(\hat\theta_{\text{jk}})| &\le c_1 2^{j/2}n^{-1} + c_2 2^{-j/2}n^{-\beta}, \end{aligned}$$

where $c_1 = 0.5C_fM_\psi$ and $c_2 = \frac{M_fC_\psi}{(\beta+1)}$.

MISE of $\hat{f}_{J_0}(x)$

Under the assumptions of Lemma 10.3, the MISE of the linear wavelet estimator is

$$\begin{aligned}
\mathbb{E}||f-\hat{f}_{J_0}||_2^2 &= \mathbb{E}(\theta_0-\hat{\theta}_0)^2+\sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}\mathbb{E}(\hat{\theta}_{jk}-\theta_{jk})^2+\sum_{j=J_0}^{\infty}\sum_{k=0}^{2^j-1}\theta_{jk}^2 \\
&\leq 2^{J_0}\frac{\sigma^2}{n}(1+o(1))+2c_1^2n^{-2}[1+\sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}2^j] \\
&\quad +2c_2^2n^{-2\beta}[1+\sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}2^{-j}]+M_f^2\sum_{j=J_0}^{\infty}\sum_{k=0}^{2^j-1}2^{-j(2\beta+1)} \\
&= 2^{J_0}\frac{\sigma^2}{n}(1+o(1))+2c_1^2n^{-2}(2^{2J_0}+2)/3+2c_2^2n^{-2\beta}(J_0+1)+M_f^2\frac{2^{-2\beta J_0}}{1-2^{-2\beta}} \\
&\leq \sigma^2\frac{N}{n}(1+o(1))+\tilde{c}_1n^{-2}N^2+\tilde{c}_2n^{-2\beta}\log n+\tilde{c}_3N^{-2\beta}+\tilde{c}_4n^{-2}
\end{aligned}$$

where $N=2^{J_0}<2^J=n$ and $\tilde{c}_1=2c_1^2/3$, $\tilde{c}_2=2c_2^2$, $\tilde{c}_3=M_f^2(1-2^{-2\beta})^{-1}$ and $\tilde{c}_4=4c_1^2/3$.

For the estimator to be consistent, we need the MISE to tend to 0 as $n\to\infty$, therefore we need $N/n\to 0$ and $N\to\infty$ as $n\to\infty$. In this case, the second term is much smaller than the first one, and $\log N<\log n$. Therefore, to find the optimal N (and hence the optimal J_0) that minimises the upper bound on the MISE, we can consider just 2 remaining terms:

$$MISE(\hat{f}_{J_0}) \leq \sigma^2\frac{N}{n}(1+o(1))+\tilde{c}_3N^{-2\beta}(1+o(1))$$

This expression is minimised when $N=cn^{1/(2\beta+1)}$, that is, when $2^{J_0}=c2^{J/(2\beta+1)}$ which implies that $J_0=\frac{J}{2\beta+1}(1+o(1))$ as $n\to\infty$ (and hence as $J\to\infty$).

Therefore, the linear wavelet estimator with $J_0=\frac{J}{2\beta+1}$ has MISE bounded by

$$\text{MISE}(\hat{f}_{J_0}) \leq Cn^{-2\beta/(2\beta+1)}$$

that is, it achieves the global minimax rate of convergence, and it has the same rate of convergence as the kernel estimator with the optimal bandwidth.

Note that this estimator is non-adaptive, that is, we need to know β, the smoothness of the unknown function, to estimate f well. The wavelet thresholding estimator with the threshold $(1+d)\sigma\sqrt{2\log n}$ for any $d\in(0,1)$ (that is, slightly larger than the universal threshold) achieves the optimal rate of convergence (up to a factor of $\log n$) **adaptively**, that is, without using the smoothness of f.