20
Nonparametric Curve Estimation

In this Chapter we discuss nonparametric estimation of probability density functions and regression functions which we refer to as **curve estimation** or **smoothing**.

In Chapter 7 we saw that it is possible to consistently estimate a cumulative distribution function F without making any assumptions about F. If we want to estimate a probability density function $f(x)$ or a regression function $r(x) = \mathbb{E}(Y|X = x)$ the situation is different. We cannot estimate these functions consistently without making some smoothness assumptions. Correspondingly, we need to perform some sort of smoothing operation on the data.

An example of a density estimator is a **histogram**, which we discuss in detail in Section 20.2. To form a histogram estimator of a density f, we divide the real line to disjoint sets called **bins**. The histogram estimator is a piecewise constant function where the height of the function is proportional to number of observations in each bin; see Figure 20.3. The number of bins is an example of a **smoothing parameter**. If we smooth too much (large bins) we get a highly biased estimator while if we smooth too little (small bins) we get a highly variable estimator. Much of curve estimation is concerned with trying to optimally balance variance and bias.

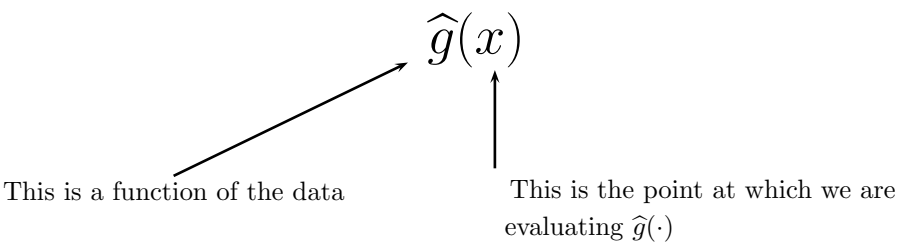

FIGURE 20.1. A curve estimate $\widehat{g}$ is random because it is a function of the data. The point x at which we evaluate $\widehat{g}$ is not a random variable.

20.1 The Bias-Variance Tradeoff

Let g denote an unknown function such as a density function or a regression function. Let $\widehat{g}_n$ denote an estimator of g. Bear in mind that $\widehat{g}_n(x)$ is a random function evaluated at a point x. The estimator is random because it depends on the data. See Figure 20.1.

As a loss function, we will use the **integrated squared error (ISE)**: [1]

$$L(g, \widehat{g}_n) = \int (g(u) - \widehat{g}_n(u))^2 \, du. \tag{20.1}$$

The **risk** or **mean integrated squared error (MISE)** with respect to squared error loss is

$$R(f, \widehat{f}) = \mathbb{E}\Big(L(g, \widehat{g})\Big). \tag{20.2}$$

20.1 Lemma. *The risk can be written as*

$$R(g, \widehat{g}_n) = \int b^2(x) \, dx + \int v(x) \, dx \tag{20.3}$$

where

$$b(x) = \mathbb{E}(\widehat{g}_n(x)) - g(x) \tag{20.4}$$

is the bias of $\widehat{g}_n(x)$ at a fixed x and

$$v(x) = \mathbb{V}(\widehat{g}_n(x)) = \mathbb{E}\bigg(\big(\widehat{g}_n(x) - \mathbb{E}\big(\widehat{g}_n(x)\big)^2\big)\bigg) \tag{20.5}$$

is the variance of $\widehat{g}_n(x)$ at a fixed x.

[1] We could use other loss functions. The results are similar but the analysis is much more complicated.

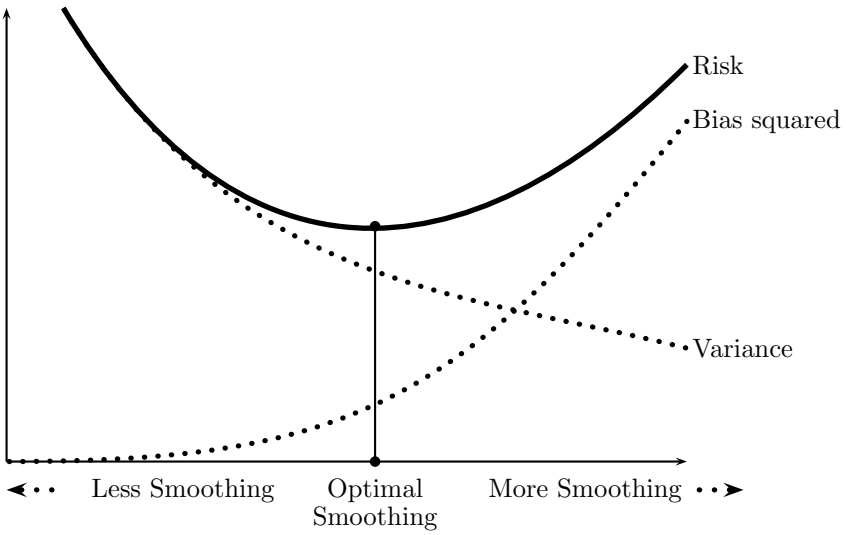

FIGURE 20.2. The Bias-Variance trade-off. The bias increases and the variance decreases with the amount of smoothing. The optimal amount of smoothing, indicated by the vertical line, minimizes the risk = bias^2 + variance.

In summary,

$$\text{RISK} = \text{BIAS}^2 + \text{VARIANCE}. \tag{20.6}$$

When the data are oversmoothed, the bias term is large and the variance is small. When the data are undersmoothed the opposite is true; see Figure 20.2. This is called the **bias-variance tradeoff**. Minimizing risk corresponds to balancing bias and variance.

20.2 Histograms

Let $X_1, \ldots, X_n$ be IID on $[0,1]$ with density f. The restriction to $[0,1]$ is not crucial; we can always rescale the data to be on this interval. Let m be an

integer and define **bins**

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \ \ldots, \ B_m = \left[\frac{m-1}{m}, 1\right]. \tag{20.7}$$

Define the **binwidth** $h = 1/m$, let ν_j be the number of observations in B_j, let $\widehat{p}_j = \nu_j/n$ and let $p_j = \int_{B_j} f(u)du$.

The **histogram estimator** is defined by

$$\widehat{f}_n(x) = \begin{cases} \widehat{p}_1/h & x \in B_1 \\ \widehat{p}_2/h & x \in B_2 \\ \vdots & \vdots \\ \widehat{p}_m/h & x \in B_m \end{cases}$$

which we can write more succinctly as

$$\widehat{f}_n(x) = \sum_{j=1}^{n} \frac{\widehat{p}_j}{h} I(x \in B_j). \tag{20.8}$$

To understand the motivation for this estimator, let $p_j = \int_{B_j} f(u)du$ and note that, for $x \in B_j$ and h small,

$$\mathbb{E}(\widehat{f}_n(x)) = \frac{\mathbb{E}(\widehat{p}_j)}{h} = \frac{p_j}{h} = \frac{\int_{B_j} f(u)du}{h} \approx \frac{f(x)h}{f(x)} = f(x).$$

20.2 Example. Figure 20.3 shows three different histograms based on $n = 1,266$ data points from an astronomical sky survey. Each data point represents the distance from us to a galaxy. The galaxies lie on a "pencilbeam" pointing directly from the Earth out into space. Because of the finite speed of light, looking at galaxies farther and farther away corresponds to looking back in time. Choosing the right number of bins involves finding a good tradeoff between bias and variance. We shall see later that the top left histogram has too few bins resulting in oversmoothing and too much bias. The bottom left histogram has too many bins resulting in undersmoothing and too few bins. The top right histogram is just right. The histogram reveals the presence of clusters of galaxies. Seeing how the size and number of galaxy clusters varies with time, helps cosmologists understand the evolution of the universe. ■

The mean and variance of $\widehat{f}_n(x)$ are given in the following Theorem.

20.3 Theorem. *Consider fixed x and fixed m, and let B_j be the bin containing x. Then,*

$$\mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h} \quad \text{and} \quad \mathbb{V}(\widehat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}. \tag{20.9}$$

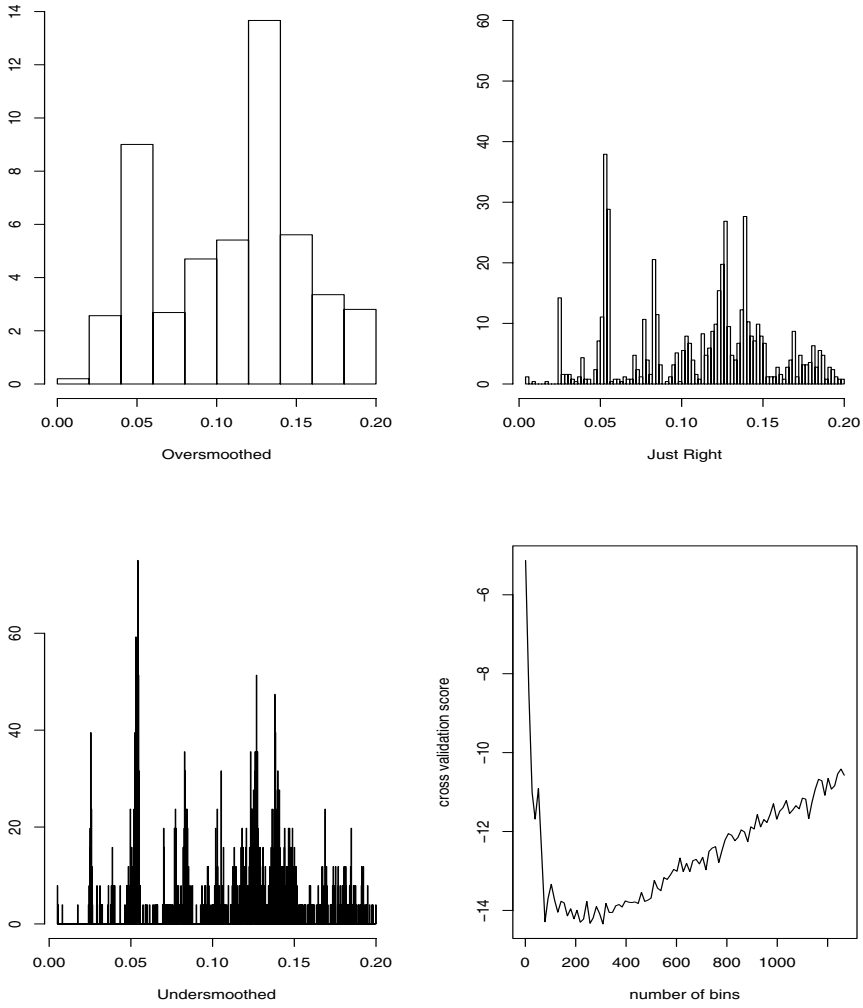

FIGURE 20.3. Three versions of a histogram for the astronomy data. The top left histogram has too few bins. The bottom left histogram has too many bins. The top right histogram is just right. The lower, right plot shows the estimated risk versus the number of bins.

Let's take a closer look at the bias-variance tradeoff using equation (20.9). Consider some $x \in B_j$. For any other $u \in B_j$,

$$f(u) \approx f(x) + (u-x)f'(x)$$

and so

$$\begin{aligned} p_j = \int_{B_j} f(u)du &\approx \int_{B_j} \Big(f(x) + (u-x)f'(x) \Big) du \\ &= f(x)h + hf'(x) \left(h\left(j - \frac{1}{2}\right) - x \right). \end{aligned}$$

Therefore, the bias $b(x)$ is

$$\begin{aligned} b(x) &= \mathbb{E}(\widehat{f}_n(x)) - f(x) = \frac{p_j}{h} - f(x) \\ &\approx \frac{f(x)h + hf'(x)\left(h\left(j-\frac{1}{2}\right) - x\right)}{h} - f(x) \\ &= f'(x)\left(h\left(j - \frac{1}{2}\right) - x \right). \end{aligned}$$

If $\widetilde{x}_j$ is the center of the bin, then

$$\begin{aligned} \int_{B_j} b^2(x)\, dx &\approx \int_{B_j} (f'(x))^2 \left(h\left(j - \frac{1}{2}\right) - x \right)^2 dx \\ &\approx (f'(\widetilde{x}_j))^2 \int_{B_j} \left(h\left(j - \frac{1}{2}\right) - x \right)^2 dx \\ &= (f'(\widetilde{x}_j))^2 \frac{h^3}{12}. \end{aligned}$$

Therefore,

$$\begin{aligned} \int_0^1 b^2(x)dx &= \sum_{j=1}^m \int_{B_j} b^2(x)dx \approx \sum_{j=1}^m (f'(\widetilde{x}_j))^2 \frac{h^3}{12} \\ &= \frac{h^2}{12} \sum_{j=1}^m h\,(f'(\widetilde{x}_j))^2 \approx \frac{h^2}{12} \int_0^1 (f'(x))^2 dx. \end{aligned}$$

Note that this increases as a function of h. Now consider the variance. For h small, $1 - p_j \approx 1$, so

$$\begin{aligned} v(x) &\approx \frac{p_j}{nh^2} \\ &= \frac{f(x)h + hf'(x)\left(h\left(j-\frac{1}{2}\right) - x\right)}{nh^2} \\ &\approx \frac{f(x)}{nh} \end{aligned}$$

where we have kept only the dominant term. So,

$$\int_0^1 v(x)dx \approx \frac{1}{nh}.$$

Note that this decreases with h. Putting all this together, we get:

20.4 Theorem. *Suppose that* $\int (f'(u))^2 du < \infty$. *Then*

$$R(\widehat{f}_n, f) \approx \frac{h^2}{12}\int (f'(u))^2 du + \frac{1}{nh}. \tag{20.10}$$

The value h^* *that minimizes (20.10) is*

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int (f'(u))^2 du}\right)^{1/3}. \tag{20.11}$$

With this choice of binwidth,

$$R(\widehat{f}_n, f) \approx \frac{C}{n^{2/3}} \tag{20.12}$$

where $C = (3/4)^{2/3}\left(\int (f'(u))^2 du\right)^{1/3}$.

Theorem 20.4 is quite revealing. We see that with an optimally chosen binwidth, the MISE decreases to 0 at rate $n^{-2/3}$. By comparison, most parametric estimators converge at rate n^{-1}. The slower rate of convergence is the price we pay for being nonparametric. The formula for the optimal binwidth h^* is of theoretical interest but it is not useful in practice since it depends on the unknown function f.

A practical way to choose the binwidth is to estimate the risk function and minimize over h. Recall that the loss function, which we now write as a function of h, is

$$\begin{aligned} L(h) &= \int (\widehat{f}_n(x) - f(x))^2\, dx \\ &= \int \widehat{f}_n^2(x)\, dx - 2\int \widehat{f}_n(x) f(x) dx + \int f^2(x)\, dx. \end{aligned}$$

The last term does not depend on the binwidth h so minimizing the risk is equivalent to minimizing the expected value of

$$J(h) = \int \widehat{f}_n^2(x)\, dx - 2\int \widehat{f}_n(x) f(x) dx.$$

We shall refer to $\mathbb{E}(J(h))$ as the risk, although it differs from the true risk by the constant term $\int f^2(x)\,dx$.

20.5 Definition. *The* **cross-validation estimator of risk** *is*

$$\widehat{J}(h) = \int \left(\widehat{f}_n(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i) \tag{20.13}$$

where $\widehat{f}_{(-i)}$ is the histogram estimator obtained after removing the i^{th} observation. We refer to $\widehat{J}(h)$ as the cross-validation score or estimated risk.

20.6 Theorem. *The cross-validation estimator is nearly unbiased:*

$$\mathbb{E}(\widehat{J}(x)) \approx \mathbb{E}(J(x)).$$

In principle, we need to recompute the histogram n times to compute $\widehat{J}(h)$. Moreover, this has to be done for all values of h. Fortunately, there is a shortcut formula.

20.7 Theorem. *The following identity holds:*

$$\widehat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)}\sum_{j=1}^{m} \widehat{p}_j^2. \tag{20.14}$$

20.8 Example. We used cross-validation in the astronomy example. The cross-validation function is quite flat near its minimum. Any m in the range of 73 to 310 is an approximate minimizer but the resulting histogram does not change much over this range. The histogram in the top right plot in Figure 20.3 was constructed using $m = 73$ bins. The bottom right plot shows the estimated risk, or more precisely, $\widehat{A}$, plotted versus the number of bins. ■

Next we want a confidence set for f. Suppose $\widehat{f}_n$ is a histogram with m bins and binwidth $h = 1/m$. We cannot realistically make confidence statements about the fine details of the true density f. Instead, we shall make confidence statements about f at the resolution of the histogram. To this end, define

$$\overline{f}_n(x) = \mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h} \quad \text{for } x \in B_j \tag{20.15}$$

where $p_j = \int_{B_j} f(u)du$. Think of $\overline{f}(x)$ as a "histogramized" version of f.

20.9 Definition. *A pair of functions* $(\ell_n(x), u_n(x))$ *is a* $1-\alpha$ **confidence band (or confidence envelope)** *if*

$$\mathbb{P}\Big(\ell(x) \leq \overline{f}_n(x) \leq u(x) \ \text{ for all } x\Big) \geq 1-\alpha. \tag{20.16}$$

20.10 Theorem. *Let* $m = m(n)$ *be the number of bins in the histogram* $\widehat{f}_n$. *Assume that* $m(n) \to \infty$ *and* $m(n)\log n/n \to 0$ *as* $n \to \infty$. *Define*

$$\begin{aligned} \ell_n(x) &= \left(\max\left\{\sqrt{\widehat{f}_n(x)} - c, 0\right\}\right)^2 \\ u_n(x) &= \left(\sqrt{\widehat{f}_n(x)} + c\right)^2 \end{aligned} \tag{20.17}$$

where

$$c = \frac{z_{\alpha/(2m)}}{2}\sqrt{\frac{m}{n}}. \tag{20.18}$$

Then, $(\ell_n(x), u_n(x))$ *is an approximate* $1-\alpha$ *confidence band.*

Proof. Here is an outline of the proof. From the central limit theorem, $\widehat{p}_j \approx N(p_j, p_j(1-p_j)/n)$. By the delta method, $\sqrt{\widehat{p}_j} \approx N(\sqrt{p_j}, 1/(4n))$. Moreover, it can be shown that the $\sqrt{\widehat{p}_j}$'s are approximately independent. Therefore,

$$2\sqrt{n}\left(\sqrt{\widehat{p}_j} - \sqrt{p}_j\right) \approx Z_j \tag{20.19}$$

where $Z_1, \ldots, Z_m \sim N(0,1)$. Let

$$A = \left\{\ell_n(x) \leq \overline{f}_n(x) \leq u_n(x) \ \text{ for all } x\right\} = \left\{\max_x \left|\sqrt{\widehat{f}_n(x)} - \sqrt{\overline{f}(x)}\right| \leq c\right\}.$$

Then,

$$\begin{aligned} \mathbb{P}(A^c) &= \mathbb{P}\left(\max_x \left|\sqrt{\widehat{f}_n(x)} - \sqrt{\overline{f}(x)}\right| > c\right) = \mathbb{P}\left(\max_j \left|\sqrt{\frac{\widehat{p}_j}{h}} - \sqrt{\frac{p_j}{h}}\right| > c\right) \\ &= \mathbb{P}\left(\max_j 2\sqrt{n}\left|\sqrt{\widehat{p}_j} - \sqrt{p_j}\right| > z_{\alpha/(2m)}\right) \\ &\approx \mathbb{P}\left(\max_j |Z_j| > z_{\alpha/(2m)}\right) \leq \sum_{j=1}^m \mathbb{P}\left(|Z_j| > z_{\alpha/(2m)}\right) \\ &= \sum_{j=1}^m \frac{\alpha}{m} = \alpha. \quad \blacksquare \end{aligned}$$

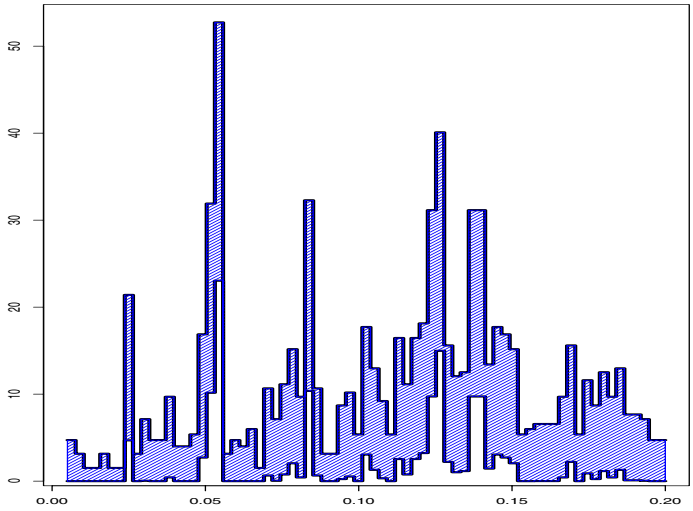

FIGURE 20.4. 95 percent confidence envelope for astronomy data using $m = 73$ bins.

20.11 Example. Figure 20.4 shows a 95 percent confidence envelope for the astronomy data. We see that even with over 1,000 data points, there is still substantial uncertainty. ■

20.3 Kernel Density Estimation

Histograms are discontinuous. **Kernel density estimators** are smoother and they converge faster to the true density than histograms.

Let $X_1, \ldots, X_n$ denote the observed data, a sample from f. In this chapter, a **kernel** is defined to be any smooth function K such that $K(x) \geq 0$, $\int K(x)\, dx = 1$, $\int xK(x)dx = 0$ and $\sigma_K^2 \equiv \int x^2 K(x)dx > 0$. Two examples of kernels are the **Epanechnikov kernel**

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2/5)/\sqrt{5} & |x| < \sqrt{5} \\ 0 & otherwise \end{cases} \tag{20.20}$$

and the Gaussian (Normal) kernel $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$.

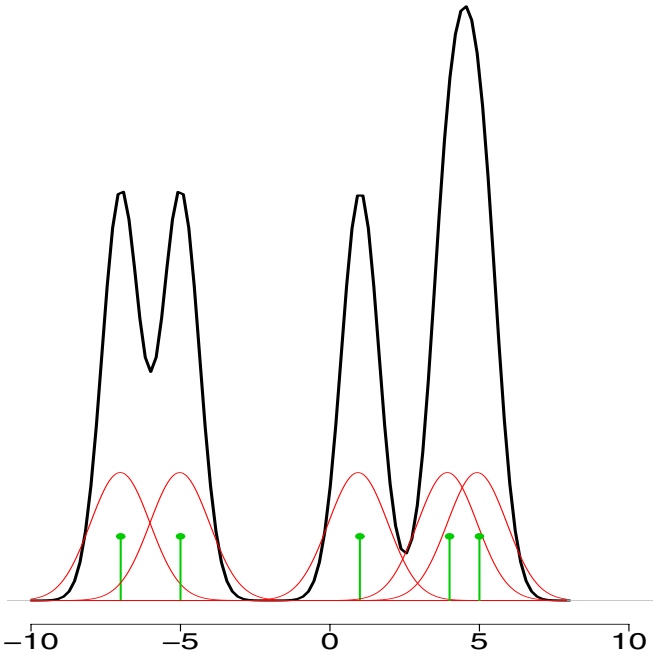

FIGURE 20.5. A kernel density estimator $\widehat{f}$. At each point x, $\widehat{f}(x)$ is the average of the kernels centered over the data points X_i. The data points are indicated by short vertical bars.

20.12 Definition. *Given a kernel K and a positive number h, called the* **bandwidth**, *the* **kernel density estimator** *is defined to be*

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right). \tag{20.21}$$

An example of a kernel density estimator is show in Figure 20.5. The kernel estimator effectively puts a smoothed-out lump of mass of size $1/n$ over each data point X_i. The bandwidth h controls the amount of smoothing. When h is close to 0, $\widehat{f}_n$ consists of a set of spikes, one at each data point. The height of the spikes tends to infinity as $h \to 0$. When $h \to \infty$, $\widehat{f}_n$ tends to a uniform density.

20.13 Example. Figure 20.6 shows kernel density estimators for the astronomy data using three different bandwidths. In each case we used a Gaussian kernel. The properly smoothed kernel density estimator in the top right panel shows similar structure as the histogram. However, it is easier to see the clusters with the kernel estimator. ■

To construct a kernel density estimator, we need to choose a kernel K and a bandwidth h. It can be shown theoretically and empirically that the choice of K is not crucial. [2] However, the choice of bandwidth h is very important. As with the histogram, we can make a theoretical statement about how the risk of the estimator depends on the bandwidth.

20.14 Theorem. *Under weak assumptions on f and K,*

$$R(f, \widehat{f}_n) \approx \frac{1}{4}\sigma_K^4 h^4 \int (f''(x))^2 + \frac{\int K^2(x)dx}{nh} \tag{20.22}$$

where $\sigma_K^2 = \int x^2 K(x)dx$. The optimal bandwidth is

$$h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}} \tag{20.23}$$

where $c_1 = \int x^2 K(x)dx$, $c_2 = \int K(x)^2 dx$ and $c_3 = \int (f''(x))^2 dx$. With this choice of bandwidth,

$$R(f, \widehat{f}_n) \approx \frac{c_4}{n^{4/5}}$$

for some constant $c_4 > 0$.

PROOF. Write $K_h(x, X) = h^{-1} K\left((x - X)/h\right)$ and $\widehat{f}_n(x) = n^{-1} \sum_i K_h(x, X_i)$. Thus, $\mathbb{E}[\widehat{f}_n(x)] = \mathbb{E}[K_h(x, X)]$ and $\mathbb{V}[\widehat{f}_n(x)] = n^{-1}\mathbb{V}[K_h(x, X)]$. Now,

$$\begin{aligned}
\mathbb{E}[K_h(x, X)] &= \int \frac{1}{h} K\left(\frac{x - t}{h}\right) f(t)\, dt \\
&= \int K(u) f(x - hu)\, du \\
&= \int K(u) \left[f(x) - hf'(x) + \frac{1}{2} f''(x) + \cdots\right] du \\
&= f(x) + \frac{1}{2} h^2 f''(x) \int u^2 K(u)\, du \cdots
\end{aligned}$$

since $\int K(x)\, dx = 1$ and $\int x\, K(x)\, dx = 0$. The bias is

$$\mathbb{E}[K_h(x, X)] - f(x) \approx \frac{1}{2}\sigma_k^2 h^2 f''(x).$$

[2] It can be shown that the Epanechnikov kernel is optimal in the sense of giving smallest asymptotic mean squared error, but it is really the choice of bandwidth which is crucial.

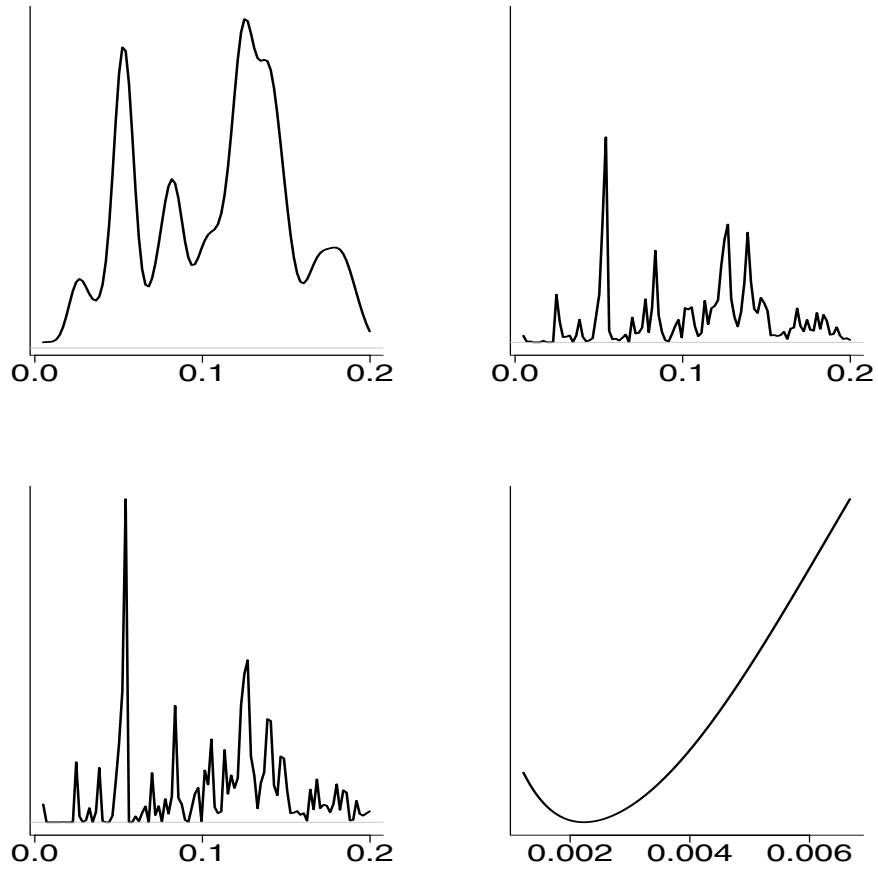

FIGURE 20.6. Kernel density estimators and estimated risk for the astronomy data. Top left: oversmoothed. Top right: just right (bandwidth chosen by cross-validation). Bottom left: undersmoothed. Bottom right: cross-validation curve as a function of bandwidth h. The bandwidth was chosen to be the value of h where the curve is a minimum.

By a similar calculation,

$$\mathbb{V}[\widehat{f}_n(x)] \approx \frac{f(x)\int K^2(x)\,dx}{n\,h_n}.$$

The result follows from integrating the squared bias plus the variance. ■

We see that kernel estimators converge at rate $n^{-4/5}$ while histograms converge at the slower rate $n^{-2/3}$. It can be shown that, under weak assumptions, there does not exist a nonparametric estimator that converges faster than $n^{-4/5}$.

The expression for h^* depends on the unknown density f which makes the result of little practical use. As with the histograms, we shall use cross-validation to find a bandwidth. Thus, we estimate the risk (up to a constant) by

$$\widehat{J}(h) = \int \widehat{f}^2(x)dz - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{-i}(X_i) \tag{20.24}$$

where $\widehat{f}_{-i}$ is the kernel density estimator after omitting the i^{th} observation.

20.15 Theorem. *For any $h > 0$,*

$$\mathbb{E}\left[\widehat{J}(h)\right] = \mathbb{E}\left[J(h)\right].$$

Also,

$$\widehat{J}(h) \approx \frac{1}{hn^2}\sum_i\sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0) \tag{20.25}$$

where $K^*(x) = K^{(2)}(x) - 2K(x)$ and $K^{(2)}(z) = \int K(z-y)K(y)dy$. In particular, if K is a N(0,1) Gaussian kernel then $K^{(2)}(z)$ is the $N(0,2)$ density.

We then choose the bandwidth h_n that minimizes $\widehat{J}(h)$.[3] A justification for this method is given by the following remarkable theorem due to Stone.

20.16 Theorem (Stone's Theorem)**.** *Suppose that f is bounded. Let $\widehat{f}_h$ denote the kernel estimator with bandwidth h and let h_n denote the bandwidth chosen by cross-validation. Then,*

$$\frac{\int\left(f(x) - \widehat{f}_{h_n}(x)\right)^2 dx}{\inf_h \int\left(f(x) - \widehat{f}_h(x)\right)^2 dx} \xrightarrow{\text{P}} 1. \tag{20.26}$$

[3]For large data sets, $\widehat{f}$ and (20.25) can be computed quickly using the fast Fourier transform.

20.17 Example. The top right panel of Figure 20.6 is based on cross-validation. These data are rounded which problems for cross-validation. Specifically, it causes the minimizer to be $h = 0$. To overcome this problem, we added a small amount of random Normal noise to the data. The result is that $\widehat{J}(h)$ is very smooth with a well defined minimum. ■

20.18 Remark. Do not assume that, if the estimator $\widehat{f}$ is wiggly, then cross-validation has let you down. The eye is not a good judge of risk.

To construct confidence bands, we use something similar to histograms. Again, the confidence band is for the smoothed version,

$$\overline{f}_n = \mathbb{E}(\widehat{f}_n(x)) = \int \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)\, du,$$

of the true density f. [4] Assume the density is on an interval (a, b). The band is

$$\ell_n(x) = \widehat{f}_n(x) - q\ \mathsf{se}(x), \quad u_n(x) = \widehat{f}_n(x) + q\ \mathsf{se}(x) \tag{20.27}$$

where

$$\begin{aligned}
\mathsf{se}(x) &= \frac{s(x)}{\sqrt{n}}, \\
s^2(x) &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(x) - \overline{Y}_n(x))^2, \\
Y_i(x) &= \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \\
q &= \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right), \\
m &= \frac{b-a}{\omega}
\end{aligned}$$

where ω is the width of the kernel. In case the kernel does not have finite width then we take ω to be the effective width, that is, the range over which the kernel is non-negligible. In particular, we take $\omega = 3h$ for the Normal kernel.

20.19 Example. Figure 20.7 shows approximate 95 percent confidence bands for the astronomy data. ■

[4] This is a modified version of the band described in Chaudhuri and Marron (1999).

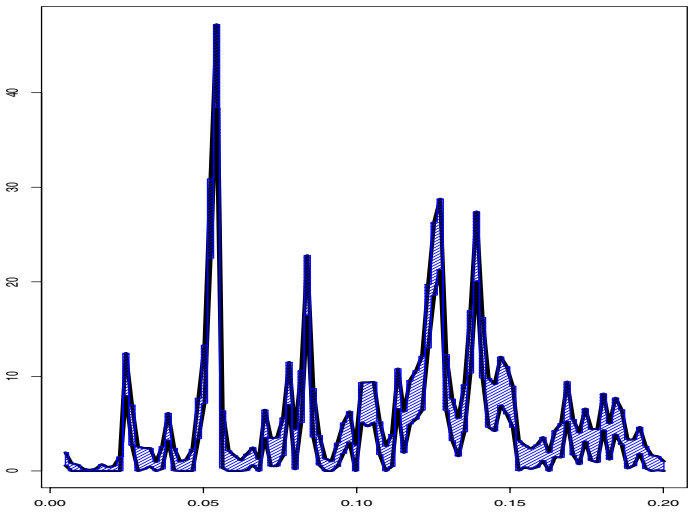

FIGURE 20.7. 95 percent confidence bands for kernel density estimate for the astronomy data.

Suppose now that the data $X_i = (X_{i1}, \ldots, X_{id})$ are d-dimensional. The kernel estimator can easily be generalized to d dimensions. Let $h = (h_1, \ldots, h_d)$ be a vector of bandwidths and define

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \tag{20.28}$$

where

$$K_h(x - X_i) = \frac{1}{nh_1 \cdots h_d}\left\{\prod_{j=1}^{d} K\left(\frac{x_i - X_{ij}}{h_j}\right)\right\} \tag{20.29}$$

where $h_1, \ldots, h_d$ are bandwidths. For simplicity, we might take $h_j = s_j h$ where s_j is the standard deviation of the j^{th} variable. There is now only a single bandwidth h to choose. Using calculations like those in the one-dimensional case, the risk is given by

$$\begin{aligned} R(f, \widehat{f}_n) &\approx \frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int f_{jj}^2(x)dx + \sum_{j \neq k} h_j^2 h_k^2 \int f_{jj} f_{kk} dx\right] \\ &\quad + \frac{\left(\int K^2(x)dx\right)^d}{nh_1 \cdots h_d} \end{aligned}$$

where f_{jj} is the second partial derivative of f. The optimal bandwidth satisfies $h_i \approx c_1 n^{-1/(4+d)}$, leading to a risk of order $n^{-4/(4+d)}$. From this fact, we see

that the risk increases quickly with dimension, a problem usually called the **curse of dimensionality**. To get a sense of how serious this problem is, consider the following table from Silverman (1986) which shows the sample size required to ensure a relative mean squared error less than 0.1 at 0 when the density is multivariate normal and the optimal bandwidth is selected:

Dimension	Sample Size
1	4
2	19
3	67
4	223
5	768
6	2790
7	10,700
8	43,700
9	187,000
10	842,000

This is bad news indeed. It says that having 842,000 observations in a ten-dimensional problem is really like having 4 observations in a one-dimensional problem.

20.4 Nonparametric Regression

Consider pairs of points $(x_1, Y_1), \ldots, (x_n, Y_n)$ related by

$$Y_i = r(x_i) + \epsilon_i \tag{20.30}$$

where $\mathbb{E}(\epsilon_i) = 0$. We have written the x_i's in lower case since we will treat them as fixed. We can do this since, in regression, it is only the mean of Y conditional on x that we are interested in. We want to estimate the regression function $r(x) = \mathbb{E}(Y|X = x)$.

There are many nonparametric regression estimators. Most involve estimating $r(x)$ by taking some sort of weighted average of the Y_i's, giving higher weight to those points near x. A popular version is the Nadaraya-Watson kernel estimator.

20.20 Definition. *The* **Nadaraya-Watson kernel estimator** *is defined by*

$$\widehat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i \tag{20.31}$$

where K is a kernel and the weights $w_i(x)$ are given by

$$w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)}. \tag{20.32}$$

The form of this estimator comes from first estimating the joint density $f(x,y)$ using kernel density estimation and then inserting the estimate into the formula,

$$r(x) = \mathbb{E}(Y|X=x) = \int y f(y|x)dy = \frac{\int y f(x,y)dy}{\int f(x,y)dy}.$$

20.21 Theorem. *Suppose that $\mathbb{V}(\epsilon_i) = \sigma^2$. The risk of the Nadaraya-Watson kernel estimator is*

$$\begin{aligned} R(\widehat{r}_n, r) &\approx \frac{h^4}{4}\left(\int x^2 K^2(x)dx\right)^4 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx \\ &\quad + \int \frac{\sigma^2 \int K^2(x)dx}{nhf(x)}dx. \end{aligned} \tag{20.33}$$

The optimal bandwidth decreases at rate $n^{-1/5}$ and with this choice the risk decreases at rate $n^{-4/5}$.

In practice, to choose the bandwidth h we minimize the cross validation score

$$\widehat{J}(h) = \sum_{i=1}^n (Y_i - \widehat{r}_{-i}(x_i))^2 \tag{20.34}$$

where $\widehat{r}_{-i}$ is the estimator we get by omitting the i^{th} variable. Fortunately, there is a shortcut formula for computing $\widehat{J}$.

20.22 Theorem. *$\widehat{J}$ can be written as*

$$\widehat{J}(h) = \sum_{i=1}^n (Y_i - \widehat{r}(x_i))^2 \frac{1}{\left(1 - \frac{K(0)}{\sum_{j=1}^n K\left(\frac{x_i - x_j}{h}\right)}\right)^2}. \tag{20.35}$$

20.23 Example. Figures 20.8 shows cosmic microwave background (CMB) data from BOOMERaNG (Netterfield et al. (2002)), Maxima (Lee et al. (2001)), and DASI (Halverson et al. (2002))). The data consist of n pairs (x_1, Y_1), ..., (x_n, Y_n) where x_i is called the multipole moment and Y_i is the

estimated power spectrum of the temperature fluctuations. What you are seeing are sound waves in the cosmic microwave background radiation which is the heat, left over from the big bang. If $r(x)$ denotes the true power spectrum, then

$$Y_i = r(x_i) + \epsilon_i$$

where ϵ_i is a random error with mean 0. The location and size of peaks in $r(x)$ provides valuable clues about the behavior of the early universe. Figure 20.8 shows the fit based on cross-validation as well as an undersmoothed and oversmoothed fit. The cross-validation fit shows the presence of three well-defined peaks, as predicted by the physics of the big bang. ■

The procedure for finding confidence bands is similar to that for density estimation. However, we first need to estimate σ^2. Suppose that the x_i's are ordered. Assuming $r(x)$ is smooth, we have $r(x_{i+1}) - r(x_i) \approx 0$ and hence

$$Y_{i+1} - Y_i = \Big[r(x_{i+1}) + \epsilon_{i+1}\Big] - \Big[r(x_i) + \epsilon_i\Big] \approx \epsilon_{i+1} - \epsilon_i$$

and hence

$$\mathbb{V}(Y_{i+1} - Y_i) \approx \mathbb{V}(\epsilon_{i+1} - \epsilon_i) = \mathbb{V}(\epsilon_{i+1}) + \mathbb{V}(\epsilon_i) = 2\sigma^2.$$

We can thus use the average of the $n-1$ differences $Y_{i+1} - Y_i$ to estimate σ^2. Hence, define

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2. \tag{20.36}$$

As with density estimate, the confidence band is for the smoothed version $\overline{r}_n(x) = \mathbb{E}(\widehat{r}_n(x))$ of the true regression function r.

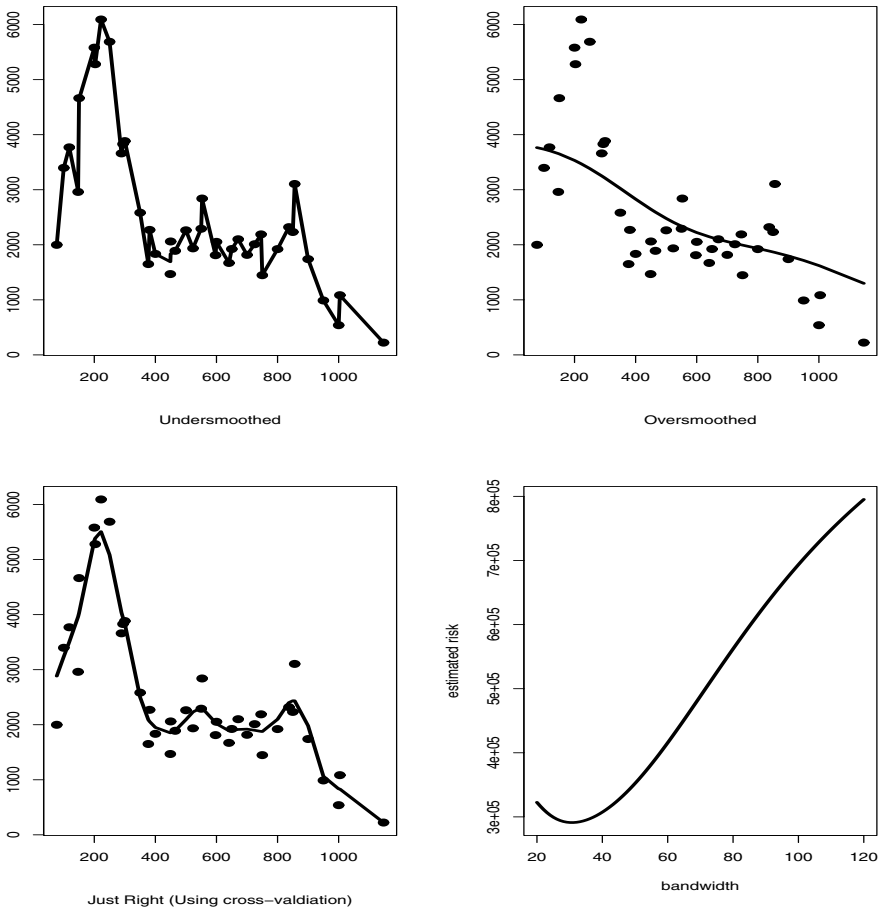

FIGURE 20.8. Regression analysis of the CMB data. The first fit is undersmoothed, the second is oversmoothed, and the third is based on cross-validation. The last panel shows the estimated risk versus the bandwidth of the smoother. The data are from BOOMERaNG, Maxima, and DASI.

Confidence Bands for Kernel Regression

An approximate $1-\alpha$ confidence band for $\overline{r}_n(x)$ is

$$\ell_n(x) = \widehat{r}_n(x) - q\ \widehat{\mathsf{se}}(x), \quad u_n(x) = \widehat{r}_n(x) + q\ \widehat{\mathsf{se}}(x) \tag{20.37}$$

where

$$\begin{aligned} \widehat{\mathsf{se}}(x) &= \widehat{\sigma}\sqrt{\sum_{i=1}^{n} w_i^2(x)}, \\ q &= \Phi^{-1}\left(\frac{1+(1-\alpha)^{1/m}}{2}\right), \\ m &= \frac{b-a}{\omega}, \end{aligned}$$

$\widehat{\sigma}$ is defined in (20.36) and ω is the width of the kernel. In case the kernel does not have finite width then we take ω to be the effective width, that is, the range over which the kernel is non-negligible. In particular, we take $\omega = 3h$ for the Normal kernel.

20.24 Example. Figure 20.9 shows a 95 percent confidence envelope for the CMB data. We see that we are highly confident of the existence and position of the first peak. We are more uncertain about the second and third peak. At the time of this writing, more accurate data are becoming available that apparently provide sharper estimates of the second and third peak. ■

The extension to multiple regressors $X = (X_1, \ldots, X_p)$ is straightforward. As with kernel density estimation we just replace the kernel with a multivariate kernel. However, the same caveats about the curse of dimensionality apply. In some cases, we might consider putting some restrictions on the regression function which will then reduce the curse of dimensionality. For example, **additive regression** is based on the model

$$Y = \sum_{j=1}^{p} r_j(X_j) + \epsilon. \tag{20.38}$$

Now we only need to fit p one-dimensional functions. The model can be enriched by adding various interactions, for example,

$$Y = \sum_{j=1}^{p} r_j(X_j) + \sum_{j<k} r_{jk}(X_j X_k) + \epsilon. \tag{20.39}$$

Additive models are usually fit by an algorithm called **backfitting**.

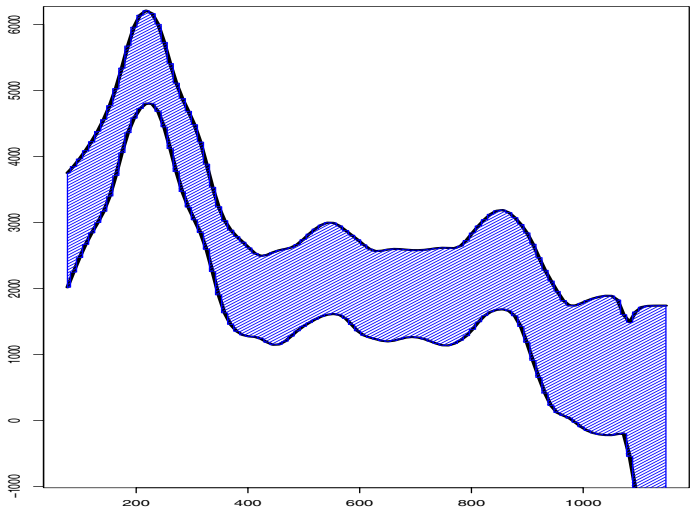

FIGURE 20.9. 95 percent confidence envelope for the CMB data.

Backfitting

1. Initialize $r_1(x_1), \ldots, r_p(x_p)$.
2. For $j = 1, \ldots, p$:
   (a) Let $\epsilon_i = Y_i - \sum_{s \neq j} r_s(x_i)$.
   (b) Let r_j be the function estimate obtained by regressing the ϵ_i's on the j^{th} covariate.
3. If converged STOP. Else, go back to step 2.

Additive models have the advantage that they avoid the curse of dimensionality and they can be fit quickly, but they have one disadvantage: the model is not fully nonparametric. In other words, the true regression function $r(x)$ may not be of the form (20.38).

20.5 Appendix

Confidence Sets and Bias. The confidence bands we computed are not for the density function or regression function but rather for the smoothed

function. For example, the confidence band for a kernel density estimate with bandwidth h is a band for the function one gets by smoothing the true function with a kernel with the same bandwidth. Getting a confidence set for the true function is complicated for reasons we now explain.

Let $\widehat{f}_n(x)$ denote an estimate of the function $f(x)$. Denote the mean and standard deviation of $\widehat{f}_n(x)$ by $\overline{f}_n(x)$ and $s_n(x)$. Then,

$$\frac{\widehat{f}_n(x) - f(x)}{s_n(x)} = \frac{\widehat{f}_n(x) - \overline{f}_n(x)}{s_n(x)} + \frac{\overline{f}_n(x) - f(x)}{s_n(x)}.$$

Typically, the first term converges to a standard Normal from which one derives confidence bands. The second term is the bias divided by the standard deviation. In parametric inference, the bias is usually smaller than the standard deviation of the estimator so this term goes to 0 as the sample size increases. In nonparametric inference, optimal smoothing leads us to balance the bias and the standard deviation. Thus the second term does not vanish even with large sample sizes. This means that the confidence interval will not be centered around the true function f.

20.6 Bibliographic Remarks

Two very good books on density estimation are Scott (1992) and Silverman (1986). The literature on nonparametric regression is very large. Two good starting points are Hardle (1990) and Loader (1999). The latter emphasizes a class of techniques called local likelihood methods.

20.7 Exercises

1. Let $X_1, \ldots, X_n \sim f$ and let $\widehat{f}_n$ be the kernel density estimator using the boxcar kernel:

$$K(x) = \begin{cases} 1 & -\frac{1}{2} < x < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Show that

$$\mathbb{E}(\widehat{f}(x)) = \frac{1}{h}\int_{x-(h/2)}^{x+(h/2)} f(y)dy$$

   and

$$\mathbb{V}(\widehat{f}(x)) = \frac{1}{nh^2}\left[\int_{x-(h/2)}^{x+(h/2)} f(y)dy - \left(\int_{x-(h/2)}^{x+(h/2)} f(y)dy\right)^2\right].$$

(b) Show that if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, then $\widehat{f}_n(x) \xrightarrow{\text{P}} f(x)$.

2. Get the data on fragments of glass collected in forensic work from the book website. Estimate the density of the first variable (refractive index) using a histogram and use a kernel density estimator. Use cross-validation to choose the amount of smoothing. Experiment with different binwidths and bandwidths. Comment on the similarities and differences. Construct 95 percent confidence bands for your estimators.

3. Consider the data from question 2. Let Y be refractive index and let x be aluminum content (the fourth variable). Perform a nonparametric regression to fit the model $Y = f(x) + \epsilon$. Use cross-validation to estimate the bandwidth. Construct 95 percent confidence bands for your estimate.

4. Prove Lemma 20.1.

5. Prove Theorem 20.3.

6. Prove Theorem 20.7.

7. Prove Theorem 20.15.

8. Consider regression data $(x_1, Y_1), \ldots, (x_n, Y_n)$. Suppose that $0 \leq x_i \leq 1$ for all i. Define bins B_j as in equation (20.7). For $x \in B_j$ define

$$\widehat{r}_n(x) = \overline{Y}_j$$

where $\overline{Y}_j$ is the mean of all the Y_i's corresponding to those x_i's in B_j. Find the approximate risk of this estimator. From this expression for the risk, find the optimal bandwidth. At what rate does the risk go to zero?

9. Show that with suitable smoothness assumptions on $r(x)$, $\widehat{\sigma}^2$ in equation (20.36) is a consistent estimator of σ^2.

10. Prove Theorem 20.22.