6
Density Estimation

Let F be a distribution with probability density $f = F'$ and let

$$X_1, \ldots, X_n \sim F$$

be an IID sample from F. The goal of **nonparametric density estimation** is to estimate f with as few assumptions about f as possible. We denote the estimator by $\widehat{f}_n$. As with nonparametric regression, the estimator will depend on a smoothing parameter h and choosing h carefully is important.

6.1 Example (Bart Simpson). The top left plot in Figure 6.1 shows the density

$$f(x) = \frac{1}{2}\phi(x; 0, 1) + \frac{1}{10}\sum_{j=0}^{4} \phi(x; (j/2) - 1, 1/10) \tag{6.2}$$

where $\phi(x; \mu, \sigma)$ denotes a Normal density with mean μ and standard deviation σ. Marron and Wand (1992) call this density "the claw" although we will call it the Bart Simpson density. Based on 1000 draws from f, I computed a kernel density estimator, described later in the chapter. The top right plot is based on a small bandwidth h which leads to undersmoothing. The bottom right plot is based on a large bandwidth h which leads to oversmoothing. The bottom left plot is based on a bandwidth h which was chosen to minimize estimated risk. This leads to a much more reasonable density estimate. ■

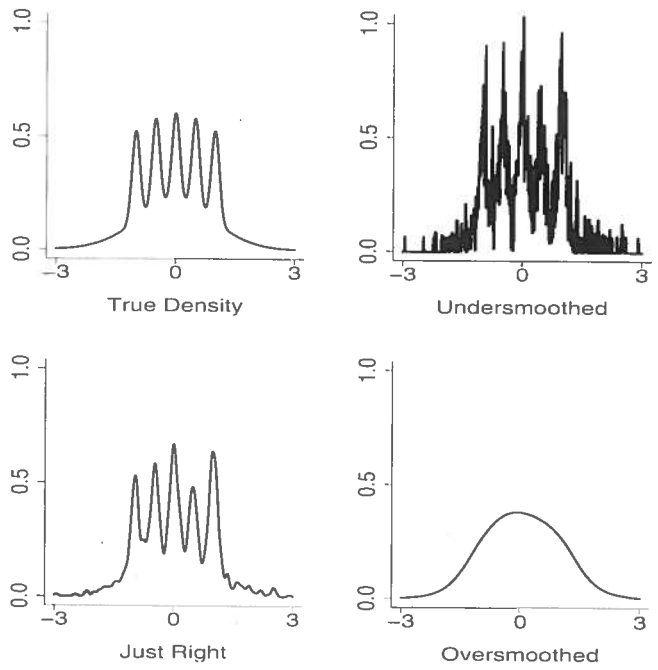

FIGURE 6.1. The Bart Simpson density from Example 6.1. Top left: true density. The other plots are kernel estimators based on $n = 1000$ draws. Bottom left: bandwidth $h = 0.05$ chosen by leave-one-out cross-validation. Top right: bandwidth $h/10$. Bottom right: bandwidth $10h$.

6.1 Cross-Validation

We will evaluate the quality of an estimator $\widehat{f}_n$ with the risk, or integrated mean squared error, $R = \mathbb{E}(L)$ where

$$L = \int (\widehat{f}_n(x) - f(x))^2 dx$$

is the integrated squared error loss function. The estimators will depend on some smoothing parameter h and we will choose h to minimize an estimate of the risk. The usual method for estimating risk is **leave-one-out cross-validation**. The details are different for density estimation than for regression. In the regression case, the cross-validation score was defined as $\sum_{i=1}^{n}(Y_i - \widehat{r}_{(-i)}(x_i))^2$ but in density estimation, there is no response variable Y. Instead, we proceed as follows.

The loss function, which we now write as a function of h, (since $\widehat{f}_n$ will depend on some smoothing parameter 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 h so minimizing the loss is equivalent to minimizing the expected value of

$$J(h) = \int \widehat{f}_n^2(x) \, dx - 2 \int \widehat{f}_n(x) f(x) dx. \tag{6.3}$$

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

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

$$\widehat{J}(h) = \int \left(\widehat{f}_n(x) \right)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i) \tag{6.5}$$

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

6.2 Histograms

Perhaps the simplest nonparametric density estimator is the histogram. Suppose f has its support on some interval which, without loss of generality, we take to be $[0, 1]$. Let m be an integer and define **bins**

$$B_1 = \left[0, \frac{1}{m}\right), \quad B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \quad \ldots, \quad B_m = \left[\frac{m-1}{m}, 1\right]. \tag{6.6}$$

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

The **histogram estimator** is defined by

$$\widehat{f}_n(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} I(x \in B_j). \tag{6.7}$$

To understand the motivation for this estimator, note that, for $x \in B_j$ and small,

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

6.8 Example. Figure 6.2 shows three different histograms based on $n = 1,2$ data points from an astronomical sky survey. These are the data from Examp 4.3. Each data point represents a "redshift," roughly speaking, the distan from us to a galaxy. Choosing the right number of bins involves finding good tradeoff between bias and variance. We shall see later that the top le histogram has too many bins resulting in oversmoothing and too much bia The bottom left histogram has too few bins resulting in undersmoothing. Th top right histogram is based on 308 bins (chosen by cross-validation). Th histogram reveals the presence of clusters of galaxies. ■

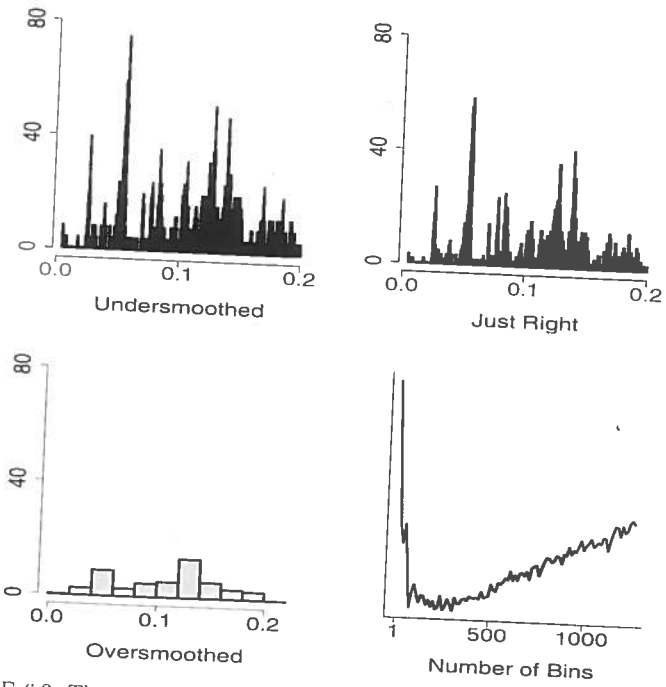

FIGURE 6.2. Three versions of a histogram for the astronomy data. The top left histogram has too many bins. The bottom left histogram has too few bins. The top right histogram uses 308 bins (chosen by cross-validation). The lower right plot shows the estimated risk versus the number of bins.

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

$$\mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h} \quad \text{and} \quad \mathbb{V}(\widehat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}. \tag{6.10}$$

6.11 Theorem. *Suppose that f' is absolutely continuous and that $\int (f'(u))^2 du < \infty$. Then*

$$R(\widehat{f}_n, f) = \frac{h^2}{12}\int (f'(u))^2 du + \frac{1}{nh} + o(h^2) + o\left(\frac{1}{n}\right). \tag{6.12}$$

The value h^* that minimizes (6.12) is

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int (f'(u))^2 du}\right)^{1/3}. \tag{6.13}$$

With this choice of binwidth,

$$R(\widehat{f}_n, f) \sim \frac{C}{n^{2/3}} \tag{6.14}$$

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

The proof of Theorem 6.11 is in the appendix. We see that with an optimally chosen binwidth, the risk decreases to 0 at rate $n^{-2/3}$. We will see shortly that kernel estimators converge at the faster rate $n^{-4/5}$ and that, in a certain sense, no faster rate is possible; see Theorem 6.31. 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. In practice, we use cross-validation as described in Section 6.1. There is a simple formula for computing the cross-validation score $\widehat{J}(h)$.

6.15 Theorem. *The following identity holds:*

$$\widehat{J}(h) = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)}\sum_{j=1}^{m} \widehat{p}_j^2. \tag{6.16}$$

6.17 Example. We used cross-validation in the astronomy example. We find that $m = 308$ is an approximate minimizer. The histogram in the top right plot in Figure 6.2 was constructed using $m = 308$ bins. The bottom right plot shows the estimated risk, or more precisely, $\widehat{J}$, 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 $h = 1/m$. For reasons explained in Section 5.7, it is difficult to construct a confidence set for f. Instead, we shall make confidence statements about f at the resolution of the histogram. Thus, define

$$\overline{f}_n(x) = \mathbb{E}(\widehat{f}_n(x)) = \sum_{j=1}^{m} \frac{p_j}{h} I(x \in B_j) \tag{6.18}$$

where $p_j = \int_{B_j} f(u)du$. Think of $\overline{f}_n(x)$ as a "histogramized" version of f. Recall that a pair of functions (ℓ, u) is a $1-\alpha$ confidence band for $\overline{f}_n$ if

$$\mathbb{P}\big(\ell(x) \leq \overline{f}_n(x) \leq u(x) \quad \text{for all } x\big) \geq 1-\alpha. \tag{6.19}$$

We could use the type of reasoning as in (5.100) but, instead, we take a simpler route.

6.20 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*

$$\begin{aligned} \ell_n(x) &= \left(\max\left\{\sqrt{\widehat{f}_n(x)} - c, 0\right\}\right)^2 \\ u_n(x) &= \left(\sqrt{\widehat{f}_n(x)} + c\right)^2 \end{aligned} \tag{6.21}$$

where

$$c = \frac{z_{\alpha/(2m)}}{2}\sqrt{\frac{m}{n}}. \tag{6.22}$$

Then, $(\ell_n(x), u_n(x))$ is an approximate $1-\alpha$ confidence band for $\overline{f}_n$.

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

$$2\sqrt{n}\left(\sqrt{\widehat{p}_j} - \sqrt{p_j}\right) \approx Z_j \tag{6.23}$$

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,

$$\mathbb{P}(A^c) = \mathbb{P}\left(\max_x \left|\sqrt{\widehat{f}_n(x)} - \sqrt{\overline{f}(x)}\right| > c\right)$$

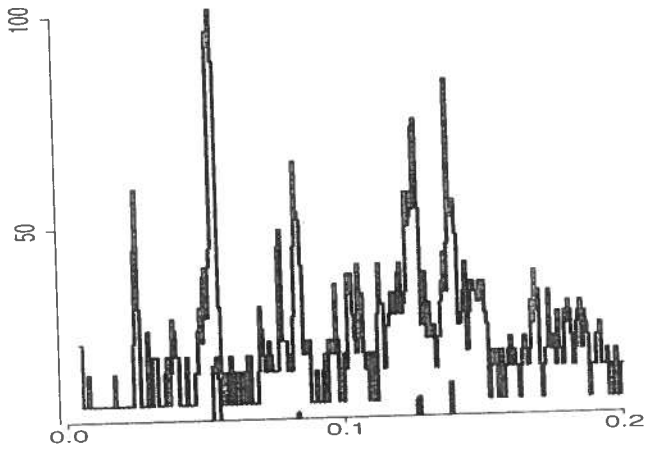

FIGURE 6.3. Ninety-five percent confidence envelope for astronomy data using $m = 308$ bins.

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

6.24 Example. Figure 6.3 shows a 95 percent confidence envelope for the astronomy data. We see that even with over 1000 data points, there is still substantial uncertainty about f as reflected by the wide bands. ■

6.3 Kernel Density Estimation

Histograms are not smooth. In this section we discuss kernel density estimators which are smoother and which converge to the true density faster. Recall that the word **kernel** refers to any smooth function K satisfying the conditions given in (4.22). See Section 4.2 for examples of kernels.

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

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right). \tag{6.26}$$

This amounts to placing a smoothed out lump of mass of size $1/n$ over each data point X_i; see Figure 6.4.

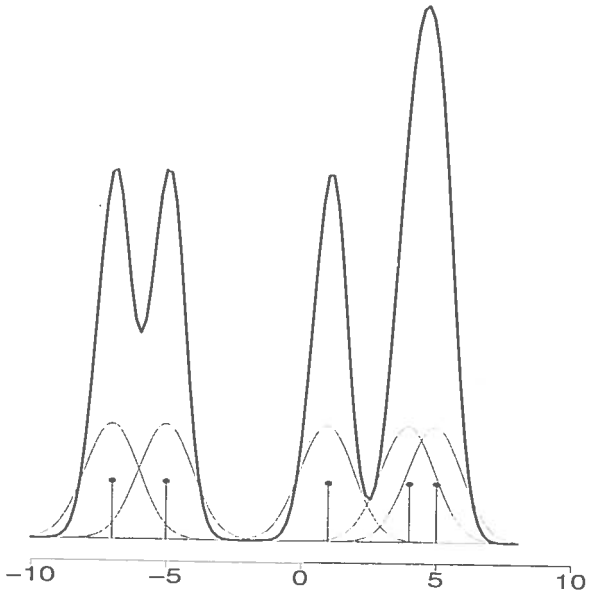

FIGURE 6.4. A kernel density estimator $\widehat{f}_n$. At each point x, $\widehat{f}_n(x)$ is the average of the kernels centered over the data points X_i. The data points are indicated by short vertical bars. The kernels are not drawn to scale.

As with kernel regression, the choice of kernel K is not crucial, but the choice of bandwidth h is important. Figure 6.5 shows density estimates with several different bandwidths. (This is the same as Figure 4.3.) Look also at Figure 6.1. We see how sensitive the estimate $\widehat{f}_n$ is to the choice of h. Small bandwidths give very rough estimates while larger bandwidths give smoother estimates. In general we will let the bandwidth depend on the sample size so we write h_n. Here are some properties of $\widehat{f}_n$.

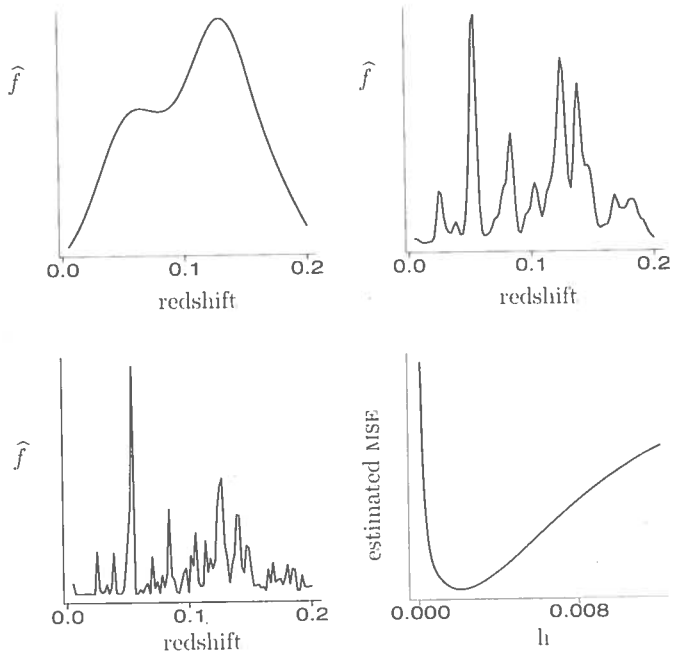

FIGURE 6.5. Kernel density estimators and estimated risk for the astronomy data. Top left: oversmoothed. Top right: just right (bandwidth chosen by cross-validation). Bottom left: undersmoothed. Bottom right: cross-validation curve as a function of bandwidth h. The bandwidth was chosen to be the value of h where the curve is a minimum.

6.27 Theorem. *Assume that f is continuous at x and that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Then $\widehat{f}_n(x) \xrightarrow{\text{P}} f(x)$.*

6.28 Theorem. *Let $R_x = \mathbb{E}(f(x) - \widehat{f}(x))^2$ be the risk at a point x and let $R = \int R_x\, dx$ denote the integrated risk. Assume that f'' is absolutely continuous and that $\int (f'''(x))^2 dx < \infty$. Also, assume that K satisfies (4.22). Then,*

$$R_x = \frac{1}{4}\sigma_K^4 h_n^4 (f''(x))^2 + \frac{f(x)\int K^2(x)dx}{nh_n} + O\left(\frac{1}{n}\right) + O(h_n^6)$$

and

$$R = \frac{1}{4}\sigma_K^4 h_n^4 \int (f''(x))^2 dx + \frac{\int K^2(x)dx}{nh} + O\left(\frac{1}{n}\right) + O(h_n^6) \tag{6.29}$$

where $\sigma_K^2 = \int x^2 K(x)dx$.

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,

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

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

$$\mathbb{E}[K_{h_n}(x,X)] - f(x) = \frac{1}{2}\sigma_K^2 h_n^2 f''(x) + O(h_n^4).$$

By a similar calculation,

$$\mathbb{V}[\widehat{f}_n(x)] = \frac{f(x)\int K^2(x)\,dx}{n\,h_n} + O\left(\frac{1}{n}\right).$$

The first result then follows since the risk is the squared bias plus variance. The second result follows from integrating the first. ∎

If we differentiate (6.29) with respect to h and set it equal to 0, we see that the asymptotically optimal bandwidth is

$$h_* = \left(\frac{c_2}{c_1^2 A(f)n}\right)^{1/5} \tag{6.30}$$

where $c_1 = \int x^2K(x)dx$, $c_2 = \int K(x)^2dx$ and $A(f) = \int (f''(x))^2dx$. This is informative because it tells us that the best bandwidth decreases at rate $n^{-1/5}$. Plugging h_* into (6.29), we see that if the optimal bandwidth is used then $R = O(n^{-4/5})$. As we saw, histograms converge at rate $O(n^{-2/3})$ showing that kernel estimators are superior in rate to histograms. According to the next theorem, there does not exist an estimator that converges faster than $O(n^{-4/5})$. For a proof, see, for example, Chapter 24 of van der Vaart (1998).

6.31 Theorem. *Let $\mathcal{F}$ be the set of all probability density functions and let $f^{(m)}$ denote the m^{th} derivative of f. Define*

$$\mathcal{F}_m(c) = \left\{f \in \mathcal{F} : \int |f^{(m)}(x)|^2dx \le c^2\right\}.$$

For any estimator $\widehat{f}_n$,

$$\sup_{f \in \mathcal{F}_m(c)} \mathbb{E}_f \int (\widehat{f}_n(x) - f(x))^2 dx \geq b \left(\frac{1}{n}\right)^{2m/(2m+1)} \tag{6.32}$$

where $b > 0$ *is a universal constant that depends only on* m *and* c.

In particular, taking $m = 2$ in the previous theorem we see that $n^{-4/5}$ is the fastest possible rate.

In practice, the bandwidth can be chosen by cross-validation but first we describe another method which is sometimes used when f is thought to be very smooth. Specifically, we compute h_* from (6.30) under the idealized assumption that f is Normal. This yields $h_* = 1.06\sigma n^{-1/5}$. Usually, σ is estimated by $\min\{s, Q/1.34\}$ where s is the sample standard deviation and Q is the interquartile range.[1] This choice of h_* works well if the true density is very smooth and is called the **Normal reference rule**.

The Normal Reference Rule

For smooth densities and a Normal kernel, use the bandwidth

$$h_n = \frac{1.06\,\widehat{\sigma}}{n^{1/5}}$$

where

$$\widehat{\sigma} = \min\left\{s, \frac{Q}{1.34}\right\}.$$

Since we don't want to necessarily assume that f is very smooth, it is usually better to estimate h using cross-validation. Recall from Section 6.1 that the cross-validation score is

$$\widehat{J}(h) = \int \widehat{f}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i) \tag{6.33}$$

where $\widehat{f}_{-i}$ denotes the kernel estimator obtained by omitting X_i. The next theorem gives a simpler expression for $\widehat{J}$.

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

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

[1] Recall that the interquartile range is the 75th percentile minus the 25th percentile. The reason for dividing by 1.34 is that $Q/1.34$ is a consistent estimate of σ if the data are from a $N(\mu, \sigma^2)$.

Also,

$$\widehat{J}(h) = \frac{1}{hn^2}\sum_i\sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0) + O\left(\frac{1}{n^2}\right) \tag{6.35}$$

where $K^*(x) = K^{(2)}(x) - 2K(x)$ *and* $K^{(2)}(z) = \int K(z-y)K(y)dy$.

6.36 Remark. When K is a N(0,1) Gaussian kernel then $K^{(2)}(z)$ is the $N(0,2)$ density. Also, we should point out that the estimator $\widehat{f}_n$ and the cross-validation score (6.35) can be computed quickly using the fast Fourier transform: see pages 61–66 of Silverman (1986).

A justification for cross-validation is given by the following remarkable theorem due to Stone (1984).

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

$$\frac{\int\left(f(x) - \widehat{f}_{\widehat{h}}(x)\right)^2 dx}{\inf_h \int\left(f(x) - \widehat{f}_h(x)\right)^2 dx} \xrightarrow{\text{a.s.}} 1. \tag{6.38}$$

The bandwidth for the density estimator in the upper right panel of Figure 6.5 is based on cross-validation. In this case it worked well but of course there are lots of examples where there are problems. 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.

Another approach to bandwidth selection called **plug-in bandwidths.** The idea is as follows. The (asymptotically) optimal bandwidth is given in equation (6.30). The only unknown quantity in that formula is $A(f) = \int(f''(x))^2 dx$. If we have an estimate $\widehat{f}''$ of f'', then we can plug this estimate into the formula for the optimal bandwidth h_*. There is a rich and interesting literature on this and similar approaches. The problem with this approach is that estimating f'' is harder than estimating f. Indeed, we need to make stronger assumptions about f to estimate f''. But if we make these stronger assumptions then the (usual) kernel estimator for f is not appropriate. Loader (1999b) has investigated this issue in detail and provides evidence that the plug-in bandwidth approach might not be reliable. There are also methods that apply corrections to plug-in rules; see Hjort (1999).

A generalization of the kernel method is to use **adaptive kernels** where one uses a different bandwidth $h(x)$ for each point x. One can also use a