6

KERNEL DENSITY ESTIMATORS

It is remarkable that the histogram stood as the only nonparametric density estimator until the 1950s, when substantial and simultaneous progress was made in density estimation and in spectral density estimation. In a little-known technical report, Fix and Hodges (1951) introduced the basic algorithm of nonparametric density estimation. They addressed the problem of statistical discrimination when the parametric form of the sampling density was not known. Fortunately, this paper has been reprinted with commentary by Silverman and Jones (1989). During the following decade, several general algorithms and alternative theoretical modes of analysis were introduced by Rosenblatt (1956), Parzen (1962), and Cencov (1962). There followed a second wave of important and primarily theoretical papers by Watson and Leadbetter (1963), Loftsgaarden and Quesenberry (1965), Schwartz (1967), Epanechnikov (1969), Tarter and Kronmal (1970), and Wahba (1971). The natural multivariate generalization was introduced by Cacoullos (1966). Finally, in the 1970s came the first papers focusing on the practical application of these methods: Scott et al. (1978) and Silverman (1978b). These and later multivariate applications awaited the computing revolution.

The basic kernel estimator may be written compactly as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i), \tag{6.1}$$

Multivariate Density Estimation, First Edition. David W. Scott.

where $K_h(t) = K(t/h)/h$, which is a notation introduced by Rosenblatt (1956). The kernel estimator can be motivated not only as the limiting case of the averaged shifted histogram as in (5.14) but also by other techniques demonstrated in Section 6.1. In fact, virtually all nonparametric algorithms are asymptotically kernel methods, a fact demonstrated empirically by Walter and Blum (1979) and proved rigorously by Terrell and Scott (1992). Woodroofe (1970) called the general class "delta sequences."

6.1 MOTIVATION FOR KERNEL ESTIMATORS

From the vantage point of a statistician or instructor, the averaging of shifted histograms seems the most natural motivation for kernel estimators. However, following other starting points in numerical analysis, time series, and signal processing provides, a deeper understanding of kernel methods. When trying to understand a particular theoretical or practical point concerning a nonparametric estimator, not all approaches are equally powerful. For example, Fourier analysis provides sophisticated tools for theoretical purposes. The bias-variance trade-off can be recast in terms of low-pass and high-pass filters in signal processing. Each is describing the same entity but with different mathematics.

6.1.1 Numerical Analysis and Finite Differences

The kernel estimator originated as a numeric approximation to the derivative of the cumulative distribution function (Rosenblatt, 1956). The empirical probability density function, which was defined in Equation (2.2) as the formal derivative of the empirical cdf $F_n(x)$ is a sum of Dirac delta functions, which is useless as an estimator of a smooth density function. Consider, however, a one-sided finite difference approximation to the derivative of $F_n(\cdot)$:

$$\begin{aligned}\hat{f}(x) &= \frac{F_n(x) - F_n(x-h)}{h}\\ &= \frac{1}{nh}\sum_{i=1}^{n} I_{[x-h,x)}(x_i) = \frac{1}{nh}\sum_{i=1}^{n} I_{(0,1]}\left(\frac{x-x_i}{h}\right),\end{aligned} \tag{6.2}$$

which from Equation (6.1) is clearly a kernel estimator with $K = U(0,1]$. As $\mathrm{E}[F_n(x)] = F(x)$ for all x, then with the Taylor's series

$$F(x-h) = F(x) - hf(x) + \frac{1}{2}h^2 f'(x) - \frac{1}{6}h^3 f''(x) + \cdots,$$

the bias is easily computed as

$$\mathrm{Bias}\{\hat{f}(x)\} = \mathrm{E}[\hat{f}(x)] - f(x) = -\frac{1}{2}hf'(x) + O(h^2).$$

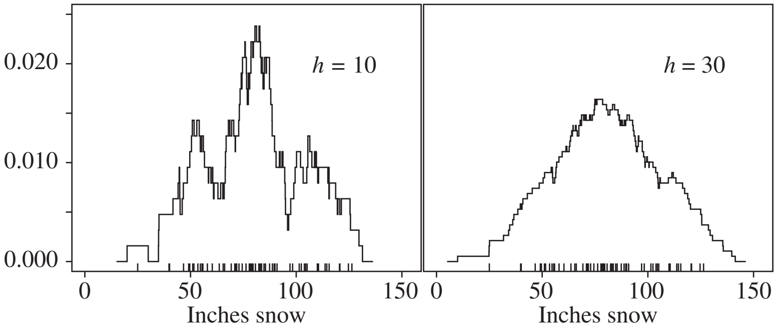

FIGURE 6.1 Central difference estimates of the Buffalo snowfall data.

Thus the integrated squared bias is $h^2R(f')/4$, which is comparable to the order of the integrated squared bias (ISB) of the histogram in Theorem 3.1 rather than the $O(h^4)$ ISB of the frequency polygon (FP). Furthermore, the ISB of (6.2) is three times larger than the ISB of the histogram. (The integrated variances are identical; see Problem 6.1.) Thus, the one-sided kernel estimator (6.2) is inferior to a histogram.

Without comment, Rosenblatt proposed a two-sided or central difference estimator of f:

$$\hat{f}(x) = \frac{F_n\left(x+\frac{h}{2}\right) - F_n\left(x-\frac{h}{2}\right)}{h}. \tag{6.3}$$

The bias of (6.3) turns out to be $h^2 f''(x)/24$ (see Problem 6.2). Thus the squared bias is $O(h^4)$, matching that of the FP. The corresponding kernel is $K = U(-0.5, 0.5)$. Recall that the histogram placed a rectangular block into the bin where each data point fell. The one-sided estimator (6.2) places the left edge of a rectangular block at each data point, whereas the two-sided estimator (6.3) places the center of a rectangular block at each data point (see Tarter and Kronmal (1976)).

Figure 6.1 displays two central difference estimates of the Buffalo snowfall data. Compared with the FP, these estimates are inferior graphically, as the estimate contains $2n$ jumps that are not even equally spaced. However, most criteria such as MISE are not particularly sensitive to such local noisy behavior.

Quasi-Newton optimization codes routinely make use of numerical central difference estimates of derivatives. Some codes use even "higher order" approximations to the first derivative (see Section 6.2.3.1).

6.1.2 Smoothing by Convolution

An electrical engineer facing a noisy function will reach into a grab bag of convolution filters to find one which will smooth away the undesired high-frequency components. The convolution operation "$*$", which is defined as

$$(f * w)(x) = \int_{-\infty}^{\infty} f(t)\, w(x-t)\, dt,$$

replaces the value of a function, $f(x)$, by a local weighted average of the function's values, according to a weight function $w(\cdot)$ that is usually symmetric and concentrated around 0. Statisticians also rely on the operation of averaging to reduce variance. Therefore, the empirical probability density function (2.2), which is too noisy, may be filtered by convolution, with the result that

$$\begin{aligned}\left[\frac{dF_n}{dx}\right] * w &= \int_{-\infty}^{\infty} \left[\frac{1}{n}\sum_{i=1}^{n} \delta(t-x_i)\right] w(x-t)\,dt \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\int_{-\infty}^{\infty} \delta(t-x_i)\,w(x-t)\,dt\right] = \frac{1}{n}\sum_{i=1}^{n} w(x-x_i), \qquad (6.4)\end{aligned}$$

which is the second kernel form given in (6.1) but without the smoothing parameter h appearing explicitly as, for example, $w(t) = K_h(t)$. In general, the shape and extent of the convolution filter weight function w will depend on the sample size. The kernel estimator (6.1) uses a single "shape" for all sample sizes, and the width of the kernel is explicitly controlled through the smoothing parameter h. The literature on filter design often uses different terminology. For example, the width of the filter w may be controlled by the half-power point, where the filter reaches half its value at the origin.

6.1.3 Orthogonal Series Approximations

The heuristic introduction to smoothing by convolution may be formalized by an orthogonal series approximation argument. For simplicity, suppose that the density function f is periodic on the interval [0, 1] so that the ordinary Fourier series basis, $\phi_\nu(t) = \exp\,(2\pi i\nu t)$, is appropriate. Every function, even noisy functions, may be expressed in terms of these basis functions as

$$f(x) = \sum_{\nu=-\infty}^{\infty} f_\nu\,\phi_\nu(x) \quad \text{where} \quad f_\nu = <f, \phi_\nu> = \int_0^1 f(x)\,\phi_\nu^*(x)\,dx. \qquad (6.5)$$

The basis functions are orthonormal, that is, $\int \phi_\nu^*(x)\phi_\mu(x)dx = \delta_{\mu\nu}$, where $\delta_{\mu\nu}$ is the Kronecker delta function and ϕ^* denotes complex conjugate. As f is a density function, the coefficient f_ν in Equation (6.5) may be expressed in statistical terms as

$$f_\nu = \mathrm{E}\,[\phi_\nu^*(X)]\,; \quad \text{hence} \quad \hat{f}_\nu = \frac{1}{n}\sum_{\ell=1}^{n} \phi_\nu^*(x_\ell) \qquad (6.6)$$

is an unbiased and consistent estimator of the Fourier coefficient f_ν. (Note that the sum is over ℓ to avoid confusion with $i = \sqrt{-1}$.) As an extreme example, consider the Fourier coefficients of the empirical probability density function (2.2):

$$f_\nu = \int_0^1 \left[\frac{1}{n}\sum_{\ell=1}^{n} \delta(x-x_\ell)\right] \phi_\nu^*(x)\,dx = \frac{1}{n}\sum_{\ell=1}^{n} \phi_\nu^*(x_\ell) = \hat{f}_\nu,$$

where $\hat{f}_\nu$ is defined in (6.6). Since $\hat{f}_\nu$ and f_ν for the empirical pdf are *identical*, the following is formally true for any sample $\{x_\ell\}$:

$$\sum_{\nu=-\infty}^{\infty} \hat{f}_\nu \, \phi_\nu(x) = \frac{1}{n}\sum_{\ell=1}^{n} \delta(x - x_\ell), \tag{6.7}$$

which is the empirical probability density function.

Cencov (1962), Kronmal and Tarter (1968), and Watson (1969) suggested smoothing the empirical density function by including only a few selected terms from Equation (6.7). Excluding terms of the form $|\nu| > k$ corresponds to what the engineers call "boxcar" filter weights

$$w_\nu(k) = \begin{cases} 1 & |\nu| \le k \\ 0 & \text{otherwise.} \end{cases} \tag{6.8}$$

As the Fourier transform of the boxcar function is the *sinc* function, $\sin(x)/(\pi x)$, the estimate will be rough and will experience "leakage"; that is, sample points relatively distant from a point x will influence $\hat{f}(x)$. Wahba (1977) suggests applying a smooth tapering window to this series, which provides more fine tuning of the resulting estimate. She introduces two parameters, λ and p, that control the shape and extent of the tapering window:

$$w_\nu(\lambda, p) = \frac{1}{1 + \lambda(2\pi\nu)^{2p}} \qquad \text{for } |\nu| \le n/2. \tag{6.9}$$

Both forms of the weighted Fourier estimate may be written explicitly as

$$\hat{f}(x) = \sum_\nu w_\nu \left[\frac{1}{n} \sum_{\ell=1}^{n} \phi_\nu^*(x_\ell) \right] \phi_\nu(x) = \frac{1}{n}\sum_{\ell=1}^{n} \left[\sum_\nu w_\nu \, \phi_\nu^*(x_\ell) \, \phi_\nu(x) \right], \tag{6.10}$$

where the order of summations has been exchanged. With the Fourier basis, the orthogonal series estimator (6.10) equals

$$\hat{f}(x) = \frac{1}{n}\sum_{\ell=1}^{n} \left[\sum_\nu w_\nu \, e^{2\pi i \nu (x - x_\ell)} \right].$$

This estimator is now in the convolution form (6.4) of a fixed kernel estimator, with the filter (or kernel) defined by the quantity in brackets. Some examples of these kernel functions are shown in Figure 6.2. Wahba's equivalent kernels are smoother and experience less leakage. The parameters λ and p can be shown to have interpretations corresponding to the smoothing parameter and to the order of the finite difference approximation, respectively.

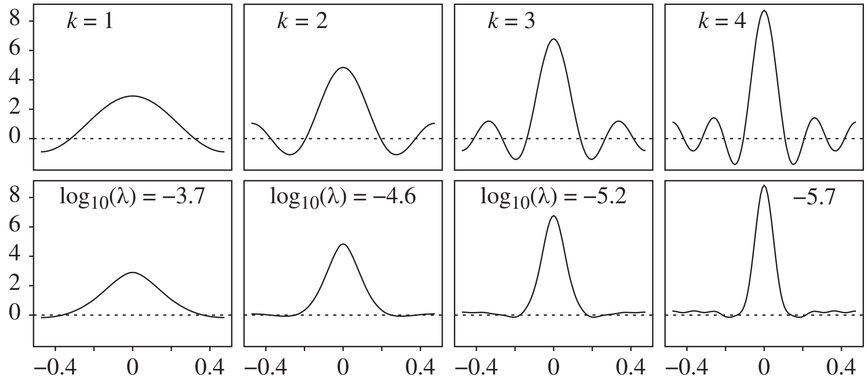

FIGURE 6.2 Examples of equivalent kernels for orthogonal series estimators. The four Wahba kernels (bottom row) have been selected to match the peak height of the corresponding Kronmal–Tarter–Watson kernels (top row). The Kronmal–Tarter–Watson kernels are independent of sample size; the Wahba examples are for $n = 16$.

6.2 THEORETICAL PROPERTIES: UNIVARIATE CASE

6.2.1 MISE Analysis

The statistical analysis of kernel estimators is much simpler than for histograms, as the kernel estimator (6.1) is the *arithmetic mean* of n independent and identically distributed random variables,

$$K_h(x,X_i) \equiv \frac{1}{h}K\left(\frac{x-X_i}{h}\right).$$

Therefore,

$$\mathrm{E}\{\hat{f}(x)\} = \mathrm{E}K_h(x,X) \quad \text{and} \quad \mathrm{Var}\{\hat{f}(x)\} = \frac{1}{n}\mathrm{Var}K_h(x,X). \tag{6.11}$$

The expectation equals

$$\mathrm{E}K_h(x,X) = \int \frac{1}{h}K\left(\frac{x-t}{h}\right)f(t)\,dt = \int K(w)f(x-hw)\,dw \tag{6.12}$$

$$= f(x)\int K(w) - hf'(x)\int wK(w) + \frac{1}{2}h^2 f''(x)\int w^2K(w) + \cdots, \tag{6.13}$$

and the variance is given by

$$\mathrm{Var}K_h(x,X) = \mathrm{E}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right]^2 - \left[\mathrm{E}\frac{1}{h}K\left(\frac{x-X}{h}\right)\right]^2. \tag{6.14}$$

The second term in (6.14) was computed in (6.13) and is approximately equal to $[f(x)\int K(w)+\cdots]^2$, while the first term may be approximated by

$$\int \frac{1}{h^2}K\left(\frac{x-t}{h}\right)^2 f(t)\,dt = \int \frac{1}{h}K(w)^2 f(x-hw)\,dw \approx \frac{f(x)R(K)}{h}. \tag{6.15}$$

From Equation (6.13), if the kernel K satisfies

$$\int K(w)=1, \quad \int wK(w)=0, \quad \text{and} \quad \int w^2K(w)\equiv\sigma_K^2 > 0,$$

then the expectation of $\hat{f}(x)$ will equal $f(x)$ to order $O(h^2)$. In fact,

$$\text{Bias}(x)=\frac{1}{2}\sigma_K^2h^2f''(x)+O(h^4) \quad \Rightarrow \quad \text{ISB}=\frac{1}{4}\sigma_K^4h^4R(f'')+O(h^6). \tag{6.16}$$

Similarly, from (6.14), (6.15), and (6.13),

$$\begin{aligned}\text{Var}(x) &= \frac{f(x)R(K)}{nh}-\frac{f(x)^2}{n}+O\left(\frac{h}{n}\right) \Rightarrow \\ \text{IV} &= \frac{R(K)}{nh}-\frac{R(f)}{n}+\cdots.\end{aligned} \tag{6.17}$$

These results are summarized in the following theorem.

Theorem 6.1: *For a nonnegative univariate kernel density estimator*,

$$\begin{aligned}\text{AMISE} &= \frac{R(K)}{nh}+\frac{1}{4}\sigma_K^4h^4R(f'') \\ h^* &= \left[\frac{R(K)}{\sigma_K^4R(f'')}\right]^{1/5} n^{-1/5} \\ \text{AMISE}^* &= \frac{5}{4}[\sigma_K R(K)]^{4/5}R(f'')^{1/5}n^{-4/5}.\end{aligned} \tag{6.18}$$

The conditions under which the theorem holds have been explored by many authors, including Parzen (1962). A simple set of conditions is that the kernel K be a continuous probability density function with finite support, $K\in L_2, \mu_K=0, 0<\sigma_K^2<\infty$, and that f'' be absolutely continuous and $f'''\in L_2$ (Scott, 1985b).

It is easy to check that the ratio of asymptotic integrated variance (AIV) to asymptotic integrated squared bias (AISB) in the asymptotic mean integrated squared error (AMISE)* is 4:1. That is, the ISB comprises only 20% of the AMISE. The similarity to the FP results in Theorem 4.1 is clear. If K is an isosceles triangle, then the results in Theorem 6.1 match those for the naive average shifted histogram (ASH) with $m = \infty$ in Equation (5.8) (see Problem 6.3). Since $R(\phi''(x|0,\sigma^2)) = 3/(8\sqrt{\pi}\sigma^5)$, the normal reference rule bandwidth with a normal kernel is

$$\boxed{\text{normal reference rule:} \qquad h = (4/3)^{1/5}\sigma n^{-1/5} \approx 1.06\,\hat{\sigma} n^{-1/5}.} \tag{6.19}$$

6.2.2 Estimation of Derivatives

Occasionally, there arises a need to estimate the derivatives of the density function; for example, when looking for modes and bumps. Derivatives of an ordinary kernel estimate behave consistently if the kernel is sufficiently differentiable and if wider bandwidths are selected. Larger smoothing parameters are required as the derivative of the estimated function is noisier than the estimated function itself. Take as an estimator of the rth derivative of f the rth derivative of the kernel estimate:

$$\hat{f}^{(r)}(x) = \frac{d^r}{dx^r}\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) = \frac{1}{nh^{r+1}}\sum_{i=1}^{n} K^{(r)}\left(\frac{x-x_i}{h}\right). \tag{6.20}$$

A calculation similar to that leading to Equation (6.15) shows that

$$\operatorname{Var}\{\hat{f}^{(r)}(x)\} \approx \frac{n}{(nh^{r+1})^2}\mathrm{E}\left[K^{(r)}\left(\frac{x-X}{h}\right)^2\right] \approx \frac{f(x)R(K^{(r)})}{nh^{2r+1}};$$

hence, the asymptotic integrated variance of $\hat{f}^{(r)}$ is $R(K^{(r)})/(nh^{2r+1})$. After an expansion similar to (6.13) to find the bias, the expectation of the first derivative estimator is

$$\mathrm{E}\hat{f}'(x) = \frac{1}{h}\left[f_x\int K' - hf_x'\int wK' + \frac{h^2}{2}f_x''\int w^2K' - \frac{h^3}{6}f_x'''\int w^3K' + \cdots\right],$$

where $f_x^{(r)} \equiv f^{(r)}(x)$. Assuming K is symmetric, $\int w^rK' = 0$ for even r. Integrating by parts, $\int wK' = -1$ and $\int w^3K' = -3\sigma_K^2$. Hence, the pointwise bias is of order h^2 and involves the *third* derivative of f. A general theorem is easily given (see Problem 6.6).

Theorem 6.2: *Based on a nonnegative univariate kernel density estimator* $\hat{f}$,

$$\mathrm{AMISE}(\hat{f}^{(r)}) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{1}{4}h^4\sigma_K^4 R(f^{(r+2)}), \tag{6.21}$$

$$h_r^* = \left[\frac{(2r+1)R(K^{(r)})}{\sigma_K^4 R(f^{(r+2)})}\right]^{1/(2r+5)} n^{-1/(2r+5)}$$

$$\mathrm{AMISE}^*(\hat{f}^{(r)}) = \frac{2r+5}{4}R(K^{(r)})^{\frac{4}{2r+5}}\left[\sigma_K^4 R(f^{(r+2)})/(2r+1)\right]^{\frac{2r+1}{2r+5}} n^{\frac{-4}{2r+5}}.$$

While the order of the bias term remains $O(h^4)$, each additional derivative order introduces two extra powers of h in the variance. The optimal smoothing parameters h^* for the first and second derivatives are $O(n^{-1/7})$ and $O(n^{-1/9})$, respectively, while the AMISE* is $O(n^{-4/7})$ and $O(n^{-4/9})$. If the optimal density rate $h^* = O(n^{-1/5})$ is used in the estimate of the second derivative, the asymptotic IV in the AMISE does not vanish, since $nh^5 = O(1)$. The estimation of an additional derivative is more difficult than estimating an additional dimension. For example, the optimal AMISE rate for the second derivative is $O(n^{-4/9})$, which is the same (slower) rate as for the optimal AMISE of a 5-D multivariate frequency polygon density estimator.

6.2.3 Choice of Kernel

Much of the first decade of theoretical work focused on various aspects of estimation properties relating to the characteristics of a kernel. Within a particular class of kernel (e.g., the order of its first nonzero moment), the quality of a density estimate is now widely recognized to be primarily determined by the choice of smoothing parameter, and only in a minor way by the choice of kernel, as will become evident in Table 6.2. Thus the topic could be de-emphasized. However, there has been a recent spurt of useful research on kernel design in special situations. While many potential hazards face the user of density estimation (e.g., underestimating the smoothness of the unknown density), the specification of desired properties for the kernel is entirely at the disposal of the worker, who should have a good understanding of the following results.

6.2.3.1 Higher Order Kernels Bartlett (1963) considered the possibility of carefully choosing the kernel to further reduce the contribution of the bias to the MISE. If the requirement that the kernel estimate should itself be a true density is relaxed, then it is possible to achieve significant improvement in the MISE. Suppose a kernel of order p is chosen so that

$$\int K = 1; \quad \int w^i K = 0, \quad i = 1, \cdots, p-1; \quad \text{and} \quad \int w^p K \neq 0, \tag{6.22}$$

then continuing the expansion in Equation (6.13), the pointwise kernel bias becomes [letting $\mu_i \equiv \int w^i K(w)dw$]

$$\text{Bias}\{\hat{f}(x)\} = \frac{1}{p!} h^p \mu_p f^{(p)}(x) + \cdots .$$

Since the formulas for the pointwise and integrated variances are unchanged, the following theorem may be proved.

Theorem 6.3: *Assuming that f is sufficiently differentiable and that the kernel K is of order p,*

$$\begin{aligned} \text{AMISE}(h) &= \frac{R(K)}{nh} + \frac{1}{(p!)^2}\mu_p^2 R(f^{(p)})\, h^{2p} \\ h^* &= \left[\frac{(p!)^2 R(K)}{2p\mu_p^2 R(f^{(p)})}\right]^{1/(2p+1)} n^{-1/(2p+1)} \\ \text{AMISE}^* &= \tfrac{2p+1}{2p}\left[2p\,\mu_p^2 R(K)^{2p} R(f^{(p)})/(p!)^2\right]^{1/(2p+1)} n^{-2p/(2p+1)}. \end{aligned} \tag{6.23}$$

Asymptotically, this result indicates it is possible to approach the usual parametric rate of $O(n^{-1})$ for the AMISE. However, the width of the optimal bandwidths increases as the order of the kernel p increases, suggesting that much of the benefit may be quite asymptotic for large p.

Table 6.1 shows some higher order kernels, together with the optimal AMISE for the case of standard normal data. These kernels are shown in Figure 6.3. Selecting representative higher order kernels is difficult, for reasons given in Section 6.2.3.2. Only *even* values of p are considered, because all *odd* "moments" of symmetric kernels vanish. Each increase of 2 in p adds another (even) moment constraint in Equation (6.22). If the kernels are polynomials, then the degree of the polynomial must also increase so that there are sufficient degrees of freedom to satisfy the constraints. The kernels in Table 6.1 begin with the so-called Epanechnikov kernel and are the unique continuous polynomial kernels of degree p that satisfy the constraints *and* have their support on the interval $[-1,1]$.

TABLE 6.1 Some Simple Polynomial Higher Order Kernels

p	K_p on $(-1,1)$	$N(0,1)$ AMISE*
2	$\frac{3}{4}(1-t^2)$	$0.320n^{-4/5}$
4	$\frac{15}{32}(1-t^2)(3-7t^2)$	$0.482n^{-8/9}$
6	$\frac{105}{256}(1-t^2)(5-30t^2+33t^4)$	$0.581n^{-12/13}$
8	$\frac{315}{4{,}096}(1-t^2)(35-385t^2+1001t^4-715t^6)$	$0.681n^{-16/17}$

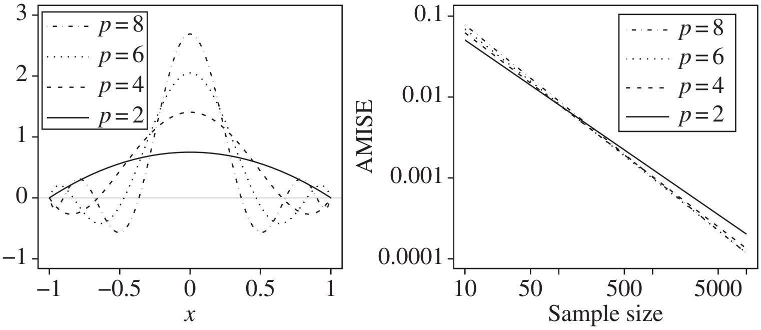

FIGURE 6.3 Examples of higher-order kernels that are low-order polynomials. The right panel shows the corresponding $N(0,1)$ AMISE* curves on a log–log scale.

The plots of the AMISE for normal data in Figure 6.3 suggest that the higher-order kernels require several thousand data points before a substantial gain may be realized. For rougher data, the gains are even more asymptotic. The improvement made possible by going to a higher order kernel is not simply a constant multiplicative factor but rather an exponential change in the order of convergence of n. Of course, for small samples, the difference between MISE and AMISE may be substantial, particularly for higher order kernels. The exact MISE may be obtained by numerical integration of the bias and variance equations. For normal data, the exact MISE was obtained for several sample sizes with the kernels in Table 6.1 plus the histogram. The results are depicted in Figure 6.4. The individual MISE curves are plotted against h/h^*, so that the minimum is centered on 1. These figures suggest that in most practical situations, kernels of order 2 and 4 are sufficient. The largest gain in MISE is obtained when going from the histogram to the order-2 kernel, with diminishing returns beyond that. Higher order kernels also seem sensitive to oversmoothing. The order-8 kernels are inferior to the histogram if $h > 2h^*$.

These higher order kernels have negative components. They will be referred to as “negative kernels,” although the more accurate phrase is “not nonnegative.” The introduction of negative kernels does provide improvement in the MISE but at the cost of having to provide special explanations. This negativity is particularly a nuisance in multiple dimensions where the regions of negative estimate can be scattered all over the domain. Statisticians may be comfortable ignoring such features, but care should be taken in their actual use. In practice, negative portions of the estimate could be clipped at 0. Clipping introduces discontinuities in the derivative of the estimate, and the modified density estimate now integrates to slightly more than 1.

In Figure 6.5, ASH estimates from Equation (5.5) using the order 2 and 4 kernels in Table 6.1 for weights in Equation (5.6) are applied to the steel surface data (Bowyer, 1980; Silverman, 1986). The data are measurements from an arbitrary origin of the actual height of a machined flat surface at a grid of 15,000 points. The bandwidths were selected so that the values at the mode matched. This example clearly indicates the reason many statisticians are willing to use negative kernels with large datasets.

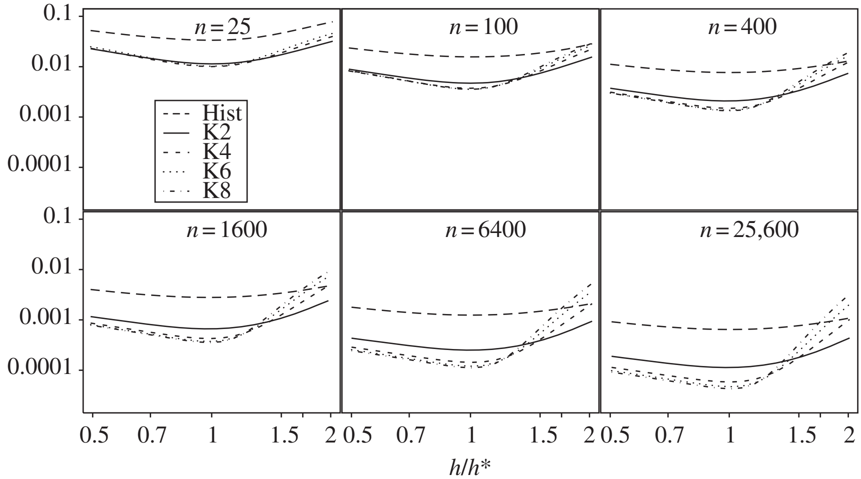

FIGURE 6.4 Exact MISE using higher order kernels with normal data for several sample sizes. The histogram MISE is included for reference.

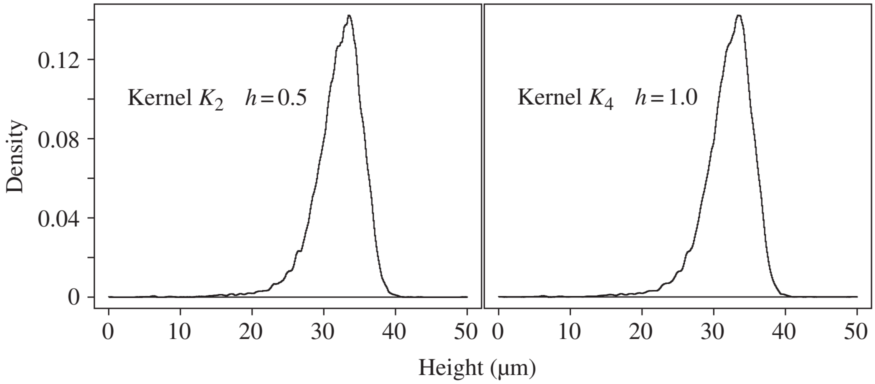

FIGURE 6.5 Positive and negative kernel estimates of the steel surface data. Kernels used were K_2 and K_4 from Table 6.1.

The negativity problem is quite marginal (minimum value −0.00008) and there is some improvement in the appearance of the estimate at the peak (simultaneously lower bias and variance).

A second problem introduced by the use of negative kernels involves the subjective choice of bandwidth. With negative kernels, the estimates can appear rough for moderately oversmoothed values of h, and not just for small, undersmoothed values, as with positive kernels. This dual roughness can be a problem for the novice, especially given the promise of higher order "better" estimates. This phenomenon is easy to demonstrate (see Figure 6.6 with the snowfall data).

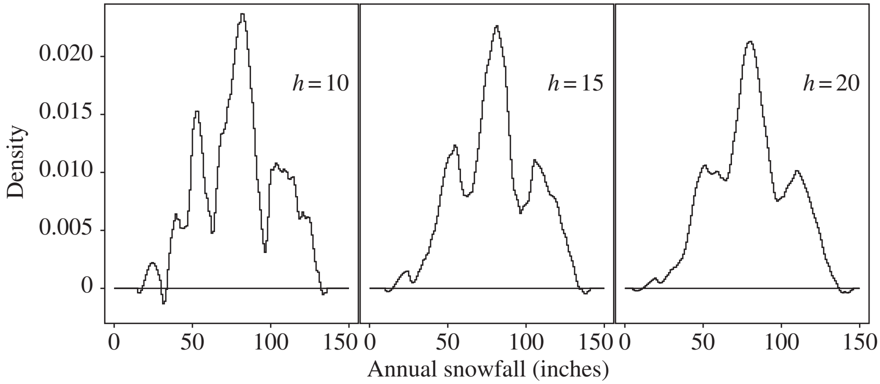

FIGURE 6.6 Kernel K_4 applied to the Buffalo snowfall data with three smoothing parameters. The ASH estimate is depicted in its histogram form.

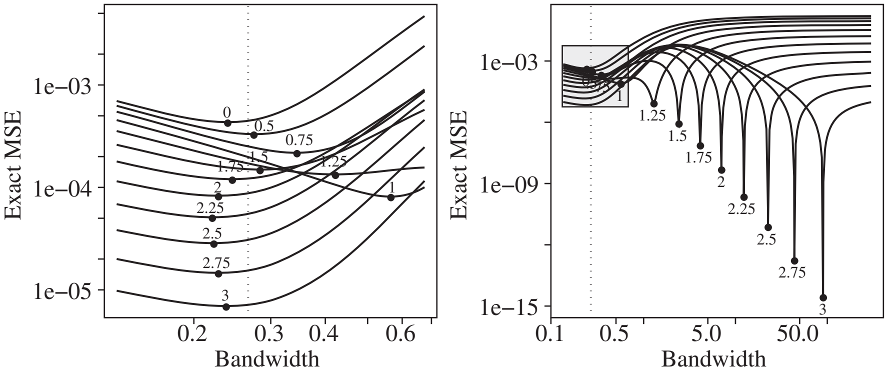

FIGURE 6.7 The exact MSE as a function of the bandwidth h for $f \sim N(0,1)$ and $n = 1000$ for selected values of x between 0 and 3. The globally optimal bandwidth $h = 0.266$ is indicated by the vertical dotted line. The bandwidth range is expanded in the right frame.

It is interesting to speculate about the relative merits of using a higher order kernel versus using a lower order kernel with an adaptive procedure. (An adaptive kernel behaves much like an adaptive frequency polygon, as in Theorem 4.2.) There is reason to believe that there is a role for both. Asymptotically, higher order kernels outperform adaptive procedures. The sensitivity to errors in the optimal bandwidth grows with the kernel order, suggesting that adaptivity is more important with higher order kernels.

Consider, for example, the exact MSE(h,x) for the $N(0,1)$ density with a normal kernel. A closed form expression for MSE(h,x) was given by Fryer (1976). A plot of MSE versus h for several values of x between 0 and 3 is shown in Figure 6.7. The global best bandwidth for $n = 1000$ is $h^* = 0.266$. In the left frame, the best bandwidth as x varies is close to this value, except when $x \approx 1$. Recall the bias is a function of $f''(x)$ and the second derivative for a normal density vanishes at $x = \pm 1$; hence, the optimal local bandwidth will be wider there for reduced variance. Plotting over a wider range of bandwidths reveals a surprise in the right frame of Figure 6.7. For those points x in the tail, there are some very large bandwidths that give surprisingly small MSEs. The reason is that by averaging over a very large neighborhood, the bias may be eliminated by artificially matching the true density. The potential utility of these large bandwidths is discussed below in Section 6.6.4.1. There is also the "local" best adaptive bandwidth around $h = 0.266$ (which is a reasonable target). Schucany (1989) has reported some success in this task. But on the whole, good adaptive estimation remains a difficult task. The MSE function is remarkably complex given the simplicity of the $N(0,1)$ density.

Let us return to a discussion of the higher order kernel approach.

While taking the limit of kernels as $p \to \infty$ may not seem wise, Davis (1975) investigated the properties of a particular "$p = \infty$" kernel, the "sinc" function sin $(x)/(\pi x)$. She showed that the MISE$^* = O(n/\log n)$. Marron and Wand (1992) have examined the MISE of a variety of more complex densities and kernels in the normal mixture family and have computed sample sizes required to justify the use of a higher order kernel.

Finally, higher order kernels can be used toward estimating the derivative of a density. However, given the nonmonotone appearance of such kernels, the derivative estimates are likely to exhibit kernel artifacts and should be reserved for data-rich situations.

Terrell and Scott (1980), using a generalized jackknife technique similar to that of Schucany and Sommers (1977), proposed an alternative method of reducing the bias. The jackknife method reduces bias by playing two estimators against each other. In density estimation, the procedure involves constructing the ratio of two positive kernel estimators with different bandwidths; for example,

$$\hat{f}(x) = \hat{f}_h(x)^{4/3} \div \hat{f}_{2h}(x)^{1/3}. \tag{6.24}$$

The result follows from jackknifing the $O(h^2)$ bias term in a Taylor's series expansion of the log bias. The expectation E_h of the usual kernel estimate is $E_h \equiv \mathrm{E}[\hat{f}_h(x)] = f(x)[1 + c_2h^2/f(x) + c_4h^4/f(x) + \cdots]$, where $c_2 = h^2\sigma_K^2 f''(x)/2$, etc. Then by Taylor's expansion,

$$\begin{aligned}\log E_h &= \log f(x) + c_2h^2/f(x) + \left[c_4 f(x) - c_2^2/2\right]h^4/f(x)^2 + \cdots \\ \log E_{2h} &= \log f(x) + 4c_2h^2/f(x) + 16\left[c_4 f(x) - c_2^2/2\right]h^4/f(x)^2 + \cdots .\end{aligned}$$

Then $\frac{4}{3}\log E_h - \frac{1}{3}E_{2h} = \log f(x) + O(h^4)$, which, after taking exponentials, suggests the form given in Equation (6.24) (See Problem 6.14 for details). The authors show that $h^* = 1.42\sigma n^{-1/9}$ in the case $f = N(0,1)$ and $K = N(0,\sigma^2)$. The resulting estimate is nonnegative and continuous, but its integral is usually slightly greater than 1. Generally, exceeding the rate $n^{-4/5}$ requires violating 1 of the 2 posits of a density function: nonnegativity and total probability mass of 1.

6.2.3.2 Optimal Kernels The kernel density estimate inherits all the properties of its kernel. Hence, it is important to note that the naive Rosenblatt kernel is discontinuous, the ASH triangle kernel has a discontinuous derivative, and the Cauchy kernel has no moments. A conservative recommendation is to choose a smooth, clearly unimodal kernel that is symmetric about the origin. However, strange kernel shapes are seldom visible in the final estimate, except perhaps in the tails, because of all the averaging.

The question of finding an optimal kernel for nonnegative estimates was considered by Epanechnikov (1969); the same variational problem was considered by Bartlett (1963), and in another context by Hodges and Lehmann (1956). From Equation (6.18), the kernel's contribution to the optimal AMISE is the following dimensionless factor:

$$\text{AMISE}^* \propto [\sigma_K R(K)]^{4/5}. \tag{6.25}$$

The problem of finding the smoothest density for the oversmoothed bandwidth problem is similar to the problem of minimizing (6.25), which may be written as

$$\min_K R(K) \quad \text{s/t} \quad \sigma_K^2 = \sigma^2.$$

The solution is a scaled version of the so-called Epanechnikov's kernel:

$$K_2^*(t) = \frac{3}{4}(1-t^2)I_{[-1,1]}(t).$$

It is interesting that the optimal kernel has finite support. The optimal kernel is not differentiable at $t = \pm\, 1$.

The variance of K_2^* is 1/5 and $R(K_2^*) = 3/5$ in (6.25). Since the AMISE is also proportional to $n^{-4/5}$, other kernels require (see Problem 6.8)

$$\frac{\sigma_K R(K)}{\sigma_{K_2^*} R(K_2^*)} = \frac{\sigma_K R(K)}{3/(5\sqrt{5})} \tag{6.26}$$

times as much data to achieve the same AMISE as the optimal Epanechnikov kernel. Table 6.2 lists many commonly used kernels and computes their asymptotic relative efficiency. The optimal kernel shows only modest improvement. Therefore, the kernel can be chosen for other reasons (ease of computation, differentiability, etc.) without undue concern for loss of efficiency. It is somewhat surprising that the popular normal kernel is so wasteful. Given the computational overhead computing exponentials, it is difficult to recommend the actual use of the normal kernel except as a point of reference.

In the last part of the table, a few absurd kernels are listed to illustrate that a very large inefficiency is still less than 2. From Theorem 4.1, the frequency polygon estimator belongs in the same grouping as the positive kernel estimators. The entries in the table by the FP were obtained by matching the AMISE expressions

TABLE 6.2 Some Common and Some Unusual Kernels and Their Relative Efficiencies. All kernels are supported on $[-1,1]$ unless noted otherwise.

Kernel	Equation	$R(K)$	σ_K^2	$\sigma_K R(K)$	Efficiencies
Uniform	$U(-1,1)$	1/2	1/3	0.2887	1.0758
Triangle	$(1-\|t\|)_+$	2/3	1/6	0.2722	1.0143
Epanechnikov	$\frac{3}{4}(1-t^2)_+$	3/5	1/5	0.2683	1
Biweight	$\frac{15}{16}(1-t^2)_+^2$	5/7	1/7	0.2700	1.0061
Triweight	$\frac{35}{32}(1-t^2)_+^3$	$\frac{350}{429}$	1/9	0.2720	1.0135
Normal	$N(0,1)$	$1/2\sqrt{\pi}$	1	0.2821	1.0513
Cosine arch	$\frac{\pi}{4}\cos\frac{\pi}{2}t$	$\pi^2/16$	$1-\frac{8}{\pi^2}$	0.2685	1.0005
Indifferent FP	See Problem 6.17	11/20	1/4	0.2750	1.0249
Dble. exp.	$\frac{1}{2}e^{-\|t\|}, \|t\| \leq \infty$	1/4	2	0.3536	1.3176
Skewed	$2860(t+\frac{2}{7})_+^3(\frac{5}{7}-t)_+^9$	$\frac{7436}{3059}$	2/147	0.2835	1.0567
Dble. Epan.	$3\|t\|(1-\|t\|)_+$	3/5	3/10	0.3286	1.2247
Shifted exp.	$e^{-(t+1)}, t > -1$	1/2	1	0.5743	1.8634
FP	See Theorem 4.1	2/3	$7/12\sqrt{5}$	0.3405	1.2690

in Theorems 4.1 and 6.1. The conclusion is that the FP is indeed in the same class, but inefficient. Finally, note that the limiting kernel of the averaged shifted histogram (isosceles triangle) is superior to the limiting kernel of the averaged shifted frequency polygon (indifferent FP), although the FP itself is superior to the histogram.

The symmetric Beta density functions, when transformed to the interval $(-1,1)$ so that the mean is 0, are a useful choice for a class of kernels:

$$K_k(t) = \frac{(2k+1)!!}{2^{k+1}k!}(1-x^2)_+^k, \tag{6.27}$$

where the double factorial notation means $(2k+1)!! = (2k+1)(2k-1)\cdots 5\cdot 3\cdot 1$. The Epanechnikov and biweight kernels are in this class. So is the normal density as $k\to\infty$ (see Problem 6.18).

The search for optimal high-order kernels is quite different and not so fruitful. Suppose that $K_4^*, K_6^*, \ldots$ are the optimal order-4, order-6, $\ldots$ kernels, respectively. From Theorem 6.2, the kernel's contribution to the AMISE* for order-4 kernels is the following nonnegative and dimensionless quantity:

$$\text{AMISE}_4^* \propto [R(K)^8 \mu_4^2]^{1/9}. \tag{6.28}$$

Consider the following fourth-order kernel, which is a mixture of K_4^* and K_6^*:

$$K_\epsilon(t) = \epsilon\, K_4^*(t) + (1-\epsilon)K_6^*(t), \qquad 0 \leq \epsilon \leq 1.$$

K_ϵ has finite roughness but its fourth moment vanishes as $\epsilon \to 0$. Thus, the fourth moment of K_4^* must be 0 and the criterion in (6.28) equals 0 at the solution. But by definition, then, K_4^* is no longer a fourth-order kernel. As many kernels have zero fourth "moment" but finite roughness, the lower bound of 0 is achieved by many kernels, none of which are in any sense order-4 or interesting in the sense of Epanechnikov. Of course, the AMISE would not in fact be 0, but would involve higher order terms.

Choosing among higher order kernels is quite complex and it is difficult to draw guidelines. In practice, second- and fourth-order methods are probably the most one should consider, as kernels beyond the order-4 provide little further reduction in MISE. For very large samples, where higher order methods do provide a substantial *fractional* reduction, the *absolute* MISE may already be so small that any practical advantage is lost.

The general advice on choosing a kernel based on these observations is to choose a symmetric kernel that is a low-order polynomial. Gasser et al. (1985) have developed a smooth hierarchy of higher order kernels. They show that their kernels are optimal but in a different sense: these kernels have minimum roughness subject to a fixed number of sign changes in the kernel. Such justification does not really warrant the label "optimal" in the sense of Epanechnikov. However, they have provided a valuable formula for low-order polynomial kernels appropriate for various combinations of the kernel order p, and for the rth derivative of the density,

$$K^*_{(k,p,r)} = \sum_{i=0}^{k+2(r-1)} \lambda_i \, t^i \, I_{[-1,1]}(t),$$

where $k \geq p+2$ and (k,p) are both odd or both even, with

$$\lambda_i = \frac{(-1)^{\frac{i+p}{2}}(k+p+2r)!(k-p)(k+2r-i)(k+i)!}{i!(i+p+1)2^{2(k+r)+1}\left(\frac{k-p}{2}\right)!\left(\frac{k+p+2r}{2}\right)!\left(\frac{k+2r-i}{2}\right)!\left(\frac{k+i}{2}\right)!}$$

if $k+i$ is even, and 0 otherwise (see Müller (1988)).

Given the availability of symbolic manipulation programs, it is probably sufficient to solve the set of linear equations governing the particular application simply. A "designer kernel" approach allows the addition of any linear conditions and results in a new kernel of higher polynomial degree. Again, the choice of kernel is not a critical matter.

6.2.3.3 Equivalent Kernels For a variety of reasons, there is no single kernel that can be recommended for all circumstances. One serious candidate is the normal kernel; however, it is relatively inefficient and has infinite support. The optimal Epanechnikov kernel is not continuously differentiable and cannot be used to estimate derivatives. In practice, the ability to switch between different kernels without having to reconsider the calibration problem at every turn is convenient. This task is easy to accomplish, *but only for kernels of the same order*. As Scott (1976) noted, if

TABLE 6.3 Factors for Equivalent Smoothing Among Popular Kernels[a]

From\To	Normal	Uniform	Epan.	Triangle	Biwt.	Triwt.
Normal	1	1.740	2.214	2.432	2.623	2.978
Uniform	0.575	1	1.272	1.398	1.507	1.711
Epanech.	0.452	0.786	1	1.099	1.185	1.345
Triangle	0.411	0.715	0.910	1	1.078	1.225
Biwt.	0.381	0.663	0.844	0.927	1	1.136
Triwt.	0.336	0.584	0.743	0.817	0.881	1

[a] To go from h_1 to h_2, multiply h_1 by the factor in the table in the row labeled K_1 and in the column labeled K_2.

h_1 and h_2 are smoothing parameters to be used with kernels K_1 and K_2, respectively, then Theorem 6.1 implies that asymptotically

$$\frac{h_1^*}{h_2^*} = \left[\frac{R(K_1)/\sigma_{K_1}^4}{R(K_2)/\sigma_{K_2}^4}\right]^{1/5} = \frac{\sigma_{K_2}}{\sigma_{K_1}}\left[\frac{\sigma_{K_1}R(K_1)}{\sigma_{K_2}R(K_2)}\right]^{1/5}. \tag{6.29}$$

Table 6.3 gives a summary of factors for equivalent smoothing bandwidths among popular kernels.

The term in brackets on the right in Equation (6.29) is the ratio of dimensionless quantities for each kernel. Those quantities $\sigma_k R(K)$ are almost equal to each other, as may be seen from Equation (6.26) and in Table 6.2. Thus the term in Equation (6.29) in brackets can be set to 1, so that the task of choosing equivalent smoothing parameters for different kernels can be accomplished by scaling according to the standard deviations:

$$\boxed{\text{Equivalent kernel rescaling:} \qquad h_2^* \approx \frac{\sigma_{K_1}}{\sigma_{K_2}} h_1^*.} \tag{6.30}$$

For example, the “exact” factor going from a normal to triweight bandwidth in Table 6.3 is 2.978, while the approximate rule (6.30) gives the factor 3.

Equivalent bandwidth scaling provides nearly identical estimates not only for optimal smoothing parameters but also for nonoptimal values. This rescaling is often used when computing and plotting a biweight kernel estimate, but using a smoothing parameter derived from a normal kernel cross-validation rule.

If all kernels were presented with equal variances, then no changes in smoothing parameters would be required. However, it would be extremely difficult to remember the formulas of those kernels, as the variances would be incorporated into the kernel forms. On balance, it seems easier to write kernels in a parsimonious form. Marron and Nolan (1988) have proposed scaling all kernels to their “canonical” form having equivalent bandwidths to a normal kernel. This proposal is slightly different than the variance rescaling proposal, since the kernel roughness is also taken into account.

For higher order kernels K_{p1} and K_{p2} of order p, a similar rescaling follows from Theorem 6.3 based on the appropriate higher order "moment" (see Problem 6.20):

$$\frac{h_{p1}^*}{h_{p2}^*} = \left[\frac{\mu_{p2}}{\mu_{p1}}\right]^{\frac{1}{p}} \left\{ \left[\frac{\mu_{p1}}{\mu_{p2}}\right]^{\frac{1}{p}} \frac{R(K_{p1})}{R(K_{p2})} \right\}^{\frac{1}{2p+1}} \approx \left[\frac{\mu_{p2}}{\mu_{p1}}\right]^{\frac{1}{p}}, \tag{6.31}$$

since the quantity in brackets is dimensionless and approximately equal to 1 for most symmetric kernels. When $p = 2$, Equations (6.31) and (6.30) agree. Some of the higher order moments are *negative*, but their ratio is positive in (6.31).

6.2.3.4 Higher Order Kernels and Kernel Design Between kernels of different order, there is no similar notion for choosing bandwidths that give "equivalent smoothing." Furthermore, the lack of a true "optimal kernel" beyond order 2 is troubling. In this section, higher order kernels are reintroduced from two other points of view. The results have some curious implications for bandwidth selection and kernel design, and suggest ways in which further work may be helpful.

The first alternative approach appeals to the numerical analysis argument first introduced by Rosenblatt (1956). Using forward and central difference approximations for the derivative, it was shown in Section 6.1.1 that the equivalent kernels are the order-1 and order-2 kernels $U(0,1)$ and $U(-0.5,0.5)$, respectively. Using the notation

$$\Delta F_n(x,c) \equiv F_n(x+c) - F_n(x-c),$$

the two-point Rosenblatt estimator with kernel $U(-1,1)$ is $\hat{f}_2(x) = \Delta F_n(x,h)/(2h)$. Consider the following 4-, 6-, and 8-point derivative estimators (see Problem 6.21):

$$\begin{aligned}
\hat{f}_4(x) &= \frac{8\Delta F_n(x,h) - \Delta F_n(x,2h)}{12h} \\
\hat{f}_6(x) &= \frac{45\Delta F_n(x,h) - 9\Delta F_n(x,2h) + \Delta F_n(x,3h)}{60h} \\
\hat{f}_8(x) &= \frac{224\Delta F_n(x,h) - 56\Delta F_n(x,2h) + \frac{32}{3}\Delta F_n(x,3h) - \Delta F_n(x,4h)}{280h}.
\end{aligned} \tag{6.32}$$

The biases for these three estimators are $-h^4 f^{(4)}(x)/30$, $-h^6 f^{(6)}(x)/140$, and $-h^8 f^{(8)}(x)/630$, respectively. The equivalent kernels are easily visualized by plotting $\hat{f}_r(x)$ with one data point at 0 with $h = 1$ as in Figure 6.8.

In Figure 6.9, the $p = 2$ naive kernel estimate is plotted for 320 cholesterol levels of patients with heart disease. As $\hat{\sigma} = 43$, the bandwidth chosen was $h = 25$, which is the equivalent bandwidth rule for a $U(-1,1)$ kernel to the normal reference rule (6.19). Specifically, the equivalent bandwidth is computed as $h \approx 3^{1/2}[1.06 \times 43 \times 320^{-1/5}]$,

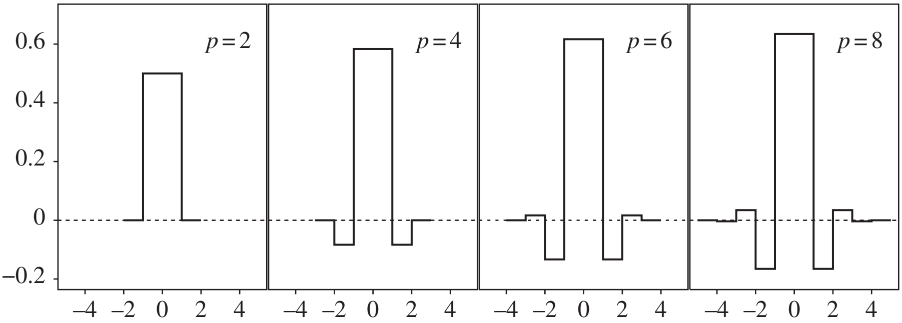

FIGURE 6.8 Higher order boxcar or naive kernels based on finite differences.

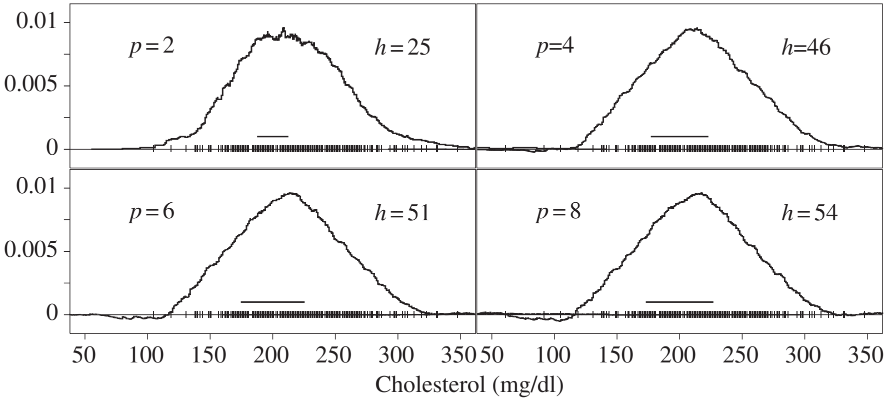

FIGURE 6.9 Higher order naive kernels applied to the cholesterol data for diseased males $(n = 320)$. Bandwidths are indicated by the horizontal lines.

where $3^{1/2}$ is the standard deviation of the $U(-1,1)$ kernel. The bandwidths for the $p =$ fourth-, sixth-, and eighth-order kernels were chosen so that the levels at the modes were equal. The negative side lobes are easily seen. The second- and fourth-order estimates are substantially different. The horizontal line shows the bandwidth h, which is half the width for the central lobe of the kernel. Thus for K_8, the support is eighth times wider than h; the influence of each data point extends 216 units (mg/dl) in both directions. For the order-2, 4-, and 6- kernels, the extent is 25, 92, and 153, respectively. Higher order kernels even with finite support are not local.

An empirical observation is that when a higher order kernel fails to provide further reduction in ISE, then the optimal higher order density estimates are all very similar (except for small local noise due to the added roughness in the tails of the kernel). This similarity occurs when the central lobe of the kernel [over the interval $(-h,h)$] remains of fixed width even as the order of the kernel grows. The sequence of bandwidths is 25, 46, 51, and 54 for the cholesterol data. Thus $p = 4$ seems a plausible choice for these data, possibly with a narrower bandwidth. As p increases, the negative lobes have an unfortunate tendency to grow and spread out.

Alternatively, the bandwidths for the naive higher order kernels could have been chosen to increase the estimate at the mode by an estimate of the bias there; for example, for the order-2 kernel

$$\text{Bias}\{\hat{f}(x)\} = \frac{1}{2}h^2\sigma_K^2 f''(x). \tag{6.33}$$

If the resulting estimate (with the order-4 kernel) is much rougher or if a bandwidth cannot be found that accomplishes the desired increase, then a reasonable conclusion is that the maximum feasible kernel order has been exceeded. The resulting bandwidth for the naive higher order kernel may be transformed by (6.31) to a smoother higher order kernel in Table 6.1.

The second alternative introduction to negative kernels suggests that the problem of bandwidth selection is even more complicated than is apparent. Beginning with a positive kernel estimate and the bias estimate given earlier, the idea is to estimate and remove the bias pointwise by using a kernel estimate of the second derivative. For clarity, a possibly different bandwidth g is used to estimate $f''(x)$:

$$\hat{f}''(x) = \frac{1}{ng^3}\sum_{i=1}^{n} K''\left(\frac{x-x_i}{g}\right). \tag{6.34}$$

Consider the "bias-corrected" kernel estimate, which is obtained by combining Equations (6.33) and (6.34):

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) - \frac{1}{2}h^2\sigma_k^2 \times \frac{1}{ng^3}\sum_{i=1}^{n} K''\left(\frac{x-x_i}{g}\right). \tag{6.35}$$

As $\hat{f}''(x)$ integrates to 0, this modified kernel estimate still integrates to 1. By inspection, (6.35) is itself in the form of a kernel estimator with kernel

$$K_{h,g}(t) = K_h(t) - \frac{h^2\sigma_k^2}{2g^2}K_g''(t). \tag{6.36}$$

A straightforward calculation verifies that this new kernel has a vanishing second moment, so that this constructive procedure in fact results in a order-4 kernel (see Problem 6.22). However, this approach suggests that $g \neq h$, since the bandwidth for the second derivative should be wider than for the density estimate itself. If so, then a well-constructed higher order kernel should slowly "expand" as the sample size increases. Given the relatively slow changes in the bandwidths, there may not be any practical improvement over allowing $g = h$.

6.2.3.5 Boundary Kernels When the unknown density is discontinuous, kernel estimates suffer the same dramatic loss of MISE efficiency as for the frequency polygon. In the case where the location of the discontinuity is known, an elegant fix is available. Without loss of generality, suppose that the discontinuity occurs at zero and that the density vanishes for $x < 0$. Careful examination of the theoretical argument which led to the elimination of the $O(h)$ bias term for a kernel estimate reveals that the critical requirement is that the kernel satisfy $\int tK(t) = 0$. The fundamental requirement is not that the kernel be symmetric; symmetry is only a simple way that the kernel may satisfy the integral equation.

The task, then, is to design finite-support kernels for use with samples in the boundary region $x_i \in (0,h)$. In order that the kernel estimate vanish for $x < 0$, the kernel for a sample point $x_i \in [0,h)$ should cover the interval $[0,x_i+h)$ rather than the interval (x_i-h,x_i+h). As the interval $[0,x_i+h)$ is narrower than $2h$, the roughness of the kernels (and hence the IV) will increase rather dramatically. In a regression setting, Gasser and Müller (1979) and Rice (1984a) have suggested using the wider interval (0, $2h$) for every $x_i \in [0,h)$. This suggestion is equivalent to choosing kernels supported on the interval $(c,c+2)$, for $-1 \le c \le 0$, that satisfy $\int_c^{c+2} tK(t)\,dt = 0$. An attempt to allow the interval width to vary so as to achieve "equivalent smoothing" using the full rule (6.29) seems doomed to failure because a wider interval cannot usually be found with equivalent smoothing. Thus the simple choice of $2h$ is a reasonable compromise.

A designer boundary kernel is described. Assume that the desired boundary kernel is to be a modification of the ordinary biweight kernel, with similar properties at the right-hand endpoint $x = c+2$. This suggests looking at a designer boundary kernel of the form

$$K_c(t) = [c_1 + c_2(t-c)^2]\cdot[t-(c+2)]^2, \qquad -1 \le c \le\ 0,$$

where the constants c_1 and c_2 are determined by the constraints $\int K_c(t)\,dt = 1$ and $\int tK_c(t)\,dt = 0$. The form for $K_c(\cdot)$ ensures that the two constraints are linear in the unknowns c_1, c_2. Solving those equations gives

$$K_c(t) = \frac{3}{4}\left[(c+1) - \frac{5}{4}(1+2c)(t-c)^2\right]\left[t-(c+2)\right]^2 I_{[c,c+2]}(t). \tag{6.37}$$

Figure 6.10 shows some examples of these kernels in two different ways. First, the kernels are shown centered on the sample (taken to be the origin). Second, the kernels are shown as they would be placed on top of the samples, so that they *begin* at zero. Note that if $x_i \in [0,h)$, then the kernel K_c with $c = (0-x_i)/h$ should be used instead of the ordinary biweight kernel.

Clearly, Figure 6.10 indicates that these are *floating boundary kernels*, meaning that the value of the kernel floats at the left boundary. An example of the use of these kernels with a sample of 100 points from the negative exponential density illustrates the effectiveness of the floating boundary kernels (see Figure 6.11). Notice how the unmodified biweight kernel estimate is quite biased in the interval $(0,h)$ and spills into the $x < 0$ region. However, without any checking or indication of a boundary problem, the unmodified kernel estimate appears quite smooth. (This smoothness should serve as a warning to check for errors resulting from the

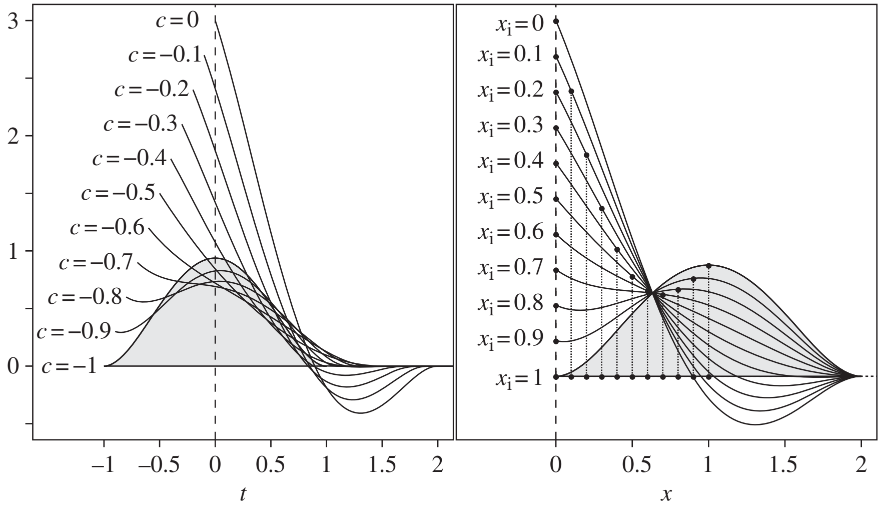

FIGURE 6.10 (Left frame) Examples of the "floating" boundary kernels $K_c(t)$, where $-1 < c < 0$. (Right frame) Assuming the boundary $x \geq 0$, each kernel $K_c(t)$ is drawn centered on the data point, $x_i = -c$, which is indicated by the dashed vertical line.

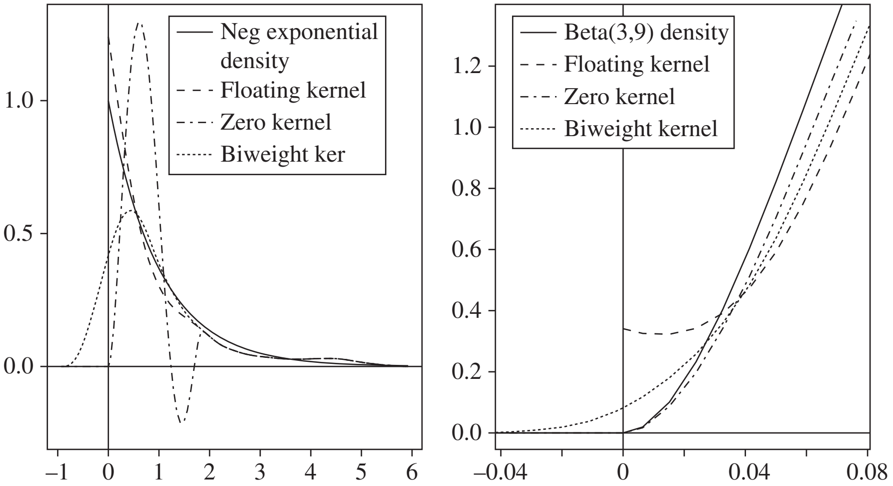

FIGURE 6.11 (Left frame) Example with negative exponential data—with and without boundary modification for $n = 100$ and $h = 0.93$. The "floating" and "zero" boundary kernels are defined in Equations (6.37) and (6.38), respectively. (Right frame) Example with Beta(3,9) density in a neighborhood of 0 for $n = 100$ and $h = 0.11$.

lack of prior knowledge of the existence of a boundary problem.) For the negative exponential density, the asymptotic theory holds if the roughness is computed on the interval $(0,\infty)$: $\int_0^\infty f''(x)^2 dx = 1/2$, so that $h^* = (70/n)^{1/5}$ for the biweight kernel by Theorem 6.1.

The density estimate of the negative exponential data shown as a big-dashed line in Figure 6.11 was constructed using the *zero boundary kernel*,

$$K_c^0(t) = \frac{15}{16}(t-c)^2(2+c-t)^2[(7c^2+14c+8)-7t(c+1)]I_{[c,c+2]}(t). \tag{6.38}$$

This modified biweight kernel was designed with the additional constraint that the boundary kernel and its derivative vanish at the left boundary $x=c$ rather than float as before. This modification was accomplished at the design stage by including the factor $(t-c)^2$, which ensures that the kernel and its derivative vanish at the left-hand endpoint $t=c$. Figure 6.12 displays these kernels in two ways. Clearly, the kernel is inappropriate for negative exponential data, inducing a large, but smooth, oscillation. However, the zero boundary kernel is appropriate for data from the Beta(3,9) density; see Figure 6.11, which shows the application of these kernels to a sample of 100 Beta(3,9) points in the vicinity of the origin. For this density, $R(f'') \approx 24,835$ and $h^* \approx (0.269/n)^{1/5}$. The ordinary estimate spills into the negative region and the "floating" kernel estimate lives up to its name.

While boundary kernels can be very useful, there are potentially serious problems with real data. There are an infinite number of boundary kernels reflecting the spectrum of possible design constraints, and these kernels are not interchangeable. Severe artifacts can be introduced by any one of them in inappropriate situations. Very careful examination is required to avoid being victimized by the particular boundary

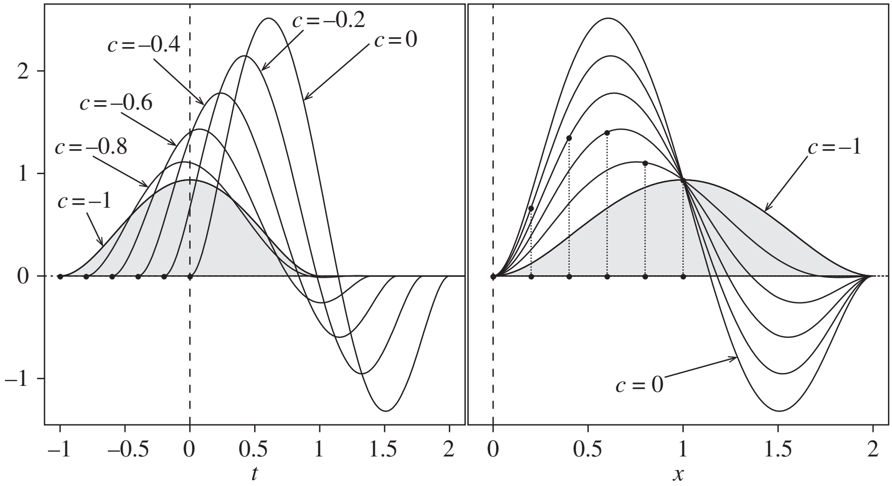

FIGURE 6.12 Examples of "zero" boundary kernels as in Figure 6.10.

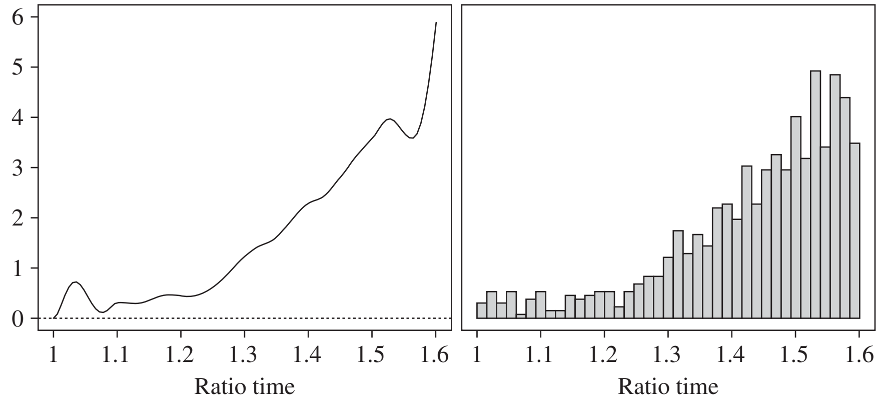

FIGURE 6.13 Density estimate of 857 fastest times in the 1991 Houston Tenneco Marathon. The data are the ratio to the leader's time for the race. Different boundary kernels were used on each extreme. A histogram is shown for comparison.

kernel chosen. Artifacts can unfortunately be introduced by the choice of the support interval for the boundary kernel. Little is known about the best way to avoid this situation, but the Rice–Müller solution seems the best of several possible alternatives that have been attempted. Finally, while the boundary kernels seem to cover a wide region of the density estimate, the effect is generally limited to the interval (0, *h*/2) for appropriately smoothed estimates.

Some data may require a different boundary kernel at each end. For example, the fastest 857 times in the January 20, 1991, Houston Tenneco Marathon were recorded as a ratio to the winning time. Clearly, the features in the density are expected to differ at the two boundaries. An estimate was constructed using K_c^0 on the left and the floating K_c on the right with $h = 0.05$ (see Figure 6.13). The clump among the leaders is real; however, the extra bump on the right appears to be more of an artifact.

Reflection Boundary Technique There is a more conservative technique that can replace the "floating" kernel. If the data are nonnegative and the discontinuity is at $x = 0$, an ordinary kernel estimate is computed but on the augmented data $(-x_n, \ldots, -x_1, x_1, \ldots, x_n)$. The final estimate is obtained by doubling this estimate for $x \geq 0$. The bandwidth should be based on the sample size n and not $2n$. This technique avoids the pitfalls of negative boundary kernels, but is generally of lower order consistency (see Problem 4.4 in the context of the frequency polygon).

6.3 THEORETICAL PROPERTIES: MULTIVARIATE CASE

The theoretical analysis of multivariate kernel estimators is the same as for frequency polygons save for a few details. The initial discussion will be limited to product kernel density estimators. The general kernel analysis will be considered afterward.

6.3.1 Product Kernels

The general form of a product kernel estimator is given by

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j}\right) \right\}. \tag{6.39}$$

The same (univariate) kernel is used in each dimension but with a (possibly) different smoothing parameter for each dimension. The data x_{ij} come from an $n \times d$ matrix. The estimate is defined pointwise, where $\mathbf{x} = (x_1, \ldots, x_d)^T$. Geometrically, the estimate places a probability mass of size $1/n$ centered on each sample point, exactly as in the univariate case. Recall that the limiting form of the naive multivariate ASH is a product triangle kernel estimator. Several bivariate product kernels are displayed in Figure 6.14.

Consider the pointwise bias of the multivariate estimator. Clearly,

$$\begin{aligned}
E\hat{f}(\mathbf{x}) &= E\prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x_j - X_j}{h_j}\right) = \int_{\Re^d} \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x_j - t_j}{h_j}\right) f(\mathbf{t})\, d\mathbf{t} \\
&= \int_{\Re^d} \prod_{j=1}^{d} K(w_j) f(x_1 - h_1 w_1, \ldots, x_n - h_n w_n)\, d\mathbf{w} \\
&\approx \int_{\Re^d} \prod_{j=1}^{d} K(w_j) \left[f(\mathbf{x}) - \sum_{r=1}^{d} h_r w_r f_r(\mathbf{x}) + \sum_{r,s=1}^{d} \frac{h_r h_s}{2} w_r w_s f_{rs}(\mathbf{x}) \right] d\mathbf{w} \\
&= f(\mathbf{x}) + \frac{1}{2}\sigma_K^2 \sum_{j=1}^{d} h_j^2 f_{jj}(\mathbf{x}) + O(h^4).
\end{aligned} \tag{6.40}$$

As before, the $O(h)$ bias terms vanish if the univariate kernels have zero mean. Similarly, the $h_r h_s$ terms vanish (see Problem 6.24). It follows that the integrated squared bias is as given in Theorem 6.4. The pointwise variance is $f(x)R(K)^d / (nh_1 h_2 \cdots h_n)$, from which the integrated variance follows easily.

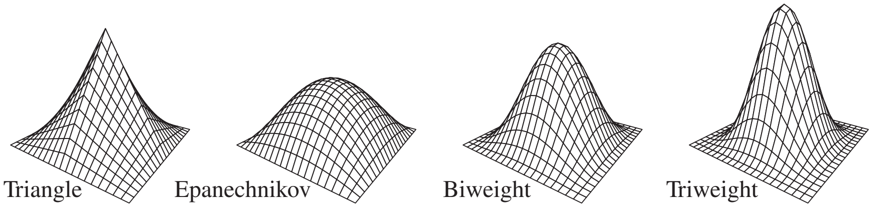

FIGURE 6.14 Product kernel examples for four kernels.

Theorem 6.4: *For a multivariate product kernel estimator, the components of the AMISE are*

$$\text{AISB} = \frac{1}{4}\sigma_K^4 \left[\sum_{i=1}^{d} h_i^4 R(f_{ii}) + \sum_{i \neq j} h_i^2 h_j^2 \int_{\Re^d} f_{ii} f_{jj}\, d\mathbf{x} \right]$$

$$\text{AIV} = \frac{R(K)^d}{n h_1 h_2 \cdots h_d} - \frac{R(f)}{n} + O\left(\frac{h}{n}\right). \qquad (6.41)$$

The order of the optimal smoothing parameters is precisely the same as for the multivariate FP: $h_i^* = O(n^{-1/(4+d)})$ and $\text{AMISE}^* = O(n^{-4/(4+d)})$.

It is a minor inconvenience, but no general closed-form expression for the optimal smoothing parameters exists, save as the solution to d nonlinear equations. Solutions may be found in special cases, for example, when $d \leq 2$ or if $h_i = h$ for all i. For example, with general bivariate normal data and a normal kernel, straightforward integration shows that

$$R(f_{11}) = 3\left[16\pi(1-\rho^2)^{5/2}\sigma_1^5\sigma_2\right]^{-1},$$

$$R(f_{22}) = 3\left[16\pi(1-\rho^2)^{5/2}\sigma_1\sigma_2^5\right]^{-1},$$

$$\int_{\Re^2} f_{11} f_{22}\, dx_1\, dx_2 = (1+2\rho^2)\left[16\pi(1-\rho^2)^{5/2}\sigma_1^3\sigma_2^3\right]^{-1}.$$

From Theorem 6.4, the AMISE is minimized when

$$\begin{aligned} h_i^* &= \sigma_i(1-\rho^2)^{5/12}(1+\rho^2/2)^{-1/6} n^{-1/6} \\ &\approx \sigma_i(1-\rho^2/2-\rho^4/16-\cdots)\, n^{-1/6} \quad i = 1,2, \end{aligned} \qquad (6.42)$$

for which

$$\text{AMISE}^* = \frac{3}{8\pi}(\sigma_1\sigma_2)^{-1}(1-\rho^2)^{-5/6}(1+\rho^2/2)^{1/3} n^{-2/3}.$$

Observe that the AMISE diverges to infinity when the data are perfectly correlated (the *real* curse of dimensionality). In comparison to other bivariate estimates, if $\rho = 0$ and $\sigma_i = 0$, then the bivariate AMISE is equal to 1/400 when $n = 302$. A bivariate FP and histogram require $n = 557$ and $n = 4244$, respectively.

A second example of a special case is the multivariate normal, where all the variables are independent. If a normal kernel is used, then a short calculation with Theorem 6.4 gives the

$$\boxed{\text{normal reference rule:} \qquad h_i^* = \left(\frac{4}{d+2}\right)^{1/(d+4)} \sigma_i \, n^{-1/(d+4)}.} \tag{6.43}$$

As the dimension d varies, the constant in Equation (6.43) ranges over the interval (0.924, 1.059), with a limit equal to 1. The constant is exactly 1 in the bivariate case and smallest when $d = 11$. Hence, an easy-to-remember data-based ruleis

$$\boxed{\text{Scott's rule in } \Re^d: \qquad \hat{h}_i = \hat{\sigma}_i n^{-1/(d+4)}.} \tag{6.44}$$

For other kernels, the equivalent kernel smoothing parameter may be obtained by dividing by the standard deviation of that kernel. Just as in one dimension, these formulas can be used in place of more precise oversmoothing values as independent normal data are very smooth. Any special structure will require narrower bandwidths. For example, the modification based on skewness and kurtosis in $\Re^1$ are identical to the factors for the frequency polygon in Section 4.1.2. If the data are not full-rank, kernel methods perform poorly. Dimension reduction techniques will be considered in Chapter 7.

6.3.2 General Multivariate Kernel MISE

In practice, product kernels are recommended. However, for various theoretical studies, general multivariate kernels will be required. This section presents a brief summary of those studies.

The general multivariate kernel estimator will include not only an arbitrary multivariate density as a kernel but also an arbitrary linear transformation of the data. Let H be a $d \times d$ nonsingular matrix and $K : R^d \to R^1$ be a kernel satisfying conditions given below.

Then the general multivariate kernel estimator is

$$\hat{f}(\mathbf{x}) = \frac{1}{n|H|} \sum_{i=1}^{n} K(H^{-1}(\mathbf{x} - \mathbf{x}_i)). \tag{6.45}$$

It should be apparent from Equation (6.45) that the linear transformation H could be incorporated into the kernel definition. For example, it is equivalent to choose K to be $N(\mathbf{0}_d, \Sigma)$ with $H = I_d$, or to choose K to be $N(\mathbf{0}_d, I_d)$ with $H = \Sigma^{1/2}$ (see Problem 6.25). Thus it is possible to choose a multivariate kernel with a simple covariance structure without loss of generality. It will not, however, be sufficient to consider only product kernels, as that would limit the discussion to multivariate kernels that are independent (and not just uncorrelated) and to kernels that are supported on a rectangular region.

The multivariate kernel will be assumed hereafter to satisfy three moment conditions (note these are matrix equations):

$$\begin{aligned}\int_{\Re^d} K(\mathbf{w})\,d\mathbf{w} &= 1\\ \int_{\Re^d} \mathbf{w}K(\mathbf{w})\,d\mathbf{w} &= \mathbf{0}_d \qquad (6.46)\\ \int_{\Re^d} \mathbf{w}\mathbf{w}^T K(\mathbf{w})\,d\mathbf{w} &= I_d.\end{aligned}$$

If K is indeed a multivariate probability density, then the last two equations summarize many assumptions about the *marginal kernels*, $\{K_i(w_i), i = 1,\ldots,d\}$. The second equation says that the means of the marginal kernels are all zero. The third equation states that the marginal kernels are all pairwise uncorrelated and that each has unit variance. Thus any simple linear transformation is assumed to be captured entirely in the matrix H and not in the kernel.

In matrix notation, it is straightforward to compute the error of the multivariate kernel estimator. For letting $\mathbf{w} = H^{-1}(\mathbf{x}-\mathbf{y})$,

$$\begin{aligned}\mathrm{E}\hat{f}(\mathbf{x}) &= \int_{\Re^d} K(H^{-1}(\mathbf{x}-\mathbf{y}))f(\mathbf{y})\,d\mathbf{y}/|H|\\ &= \int_{\Re^d} K(\mathbf{w})f(\mathbf{x}-H\mathbf{w})\,d\mathbf{w}\\ &= \int_{\Re^d} K(\mathbf{w})\left[f(\mathbf{x}) - \mathbf{w}^T H\nabla f(\mathbf{x}) + \frac{1}{2}\mathbf{w}^T H^T \nabla^2 f(\mathbf{x}) H\mathbf{w}\right] d\mathbf{w} \qquad (6.47)\end{aligned}$$

to second order. Further simplification is possible using the following property of the trace (tr) of a matrix: $\mathrm{tr}\{AB\} = \mathrm{tr}\{BA\}$, assuming that the matrices A and B have dimensions $r \times s$ and $s \times r$, respectively. Now the quadratic form in Equation (6.47) is a 1×1 matrix, which trivially equals its trace. Hence, using the trace identity and exchanging the trace and integral operations yields

$$\mathrm{E}\hat{f}(\mathbf{x}) = f(\mathbf{x}) - 0 + \frac{1}{2}\mathrm{tr}\left\{\int_{\Re^d} \mathbf{w}\mathbf{w}^T K(\mathbf{w})\,d\mathbf{w}\cdot H^T\nabla^2 f(\mathbf{x})H\right\}.$$

As the covariance matrix of K is I_d by assumption (6.46), the integral factor in the trace vanishes. Therefore,

$$\mathrm{Bias}\{\hat{f}(\mathbf{x})\} = \frac{1}{2}\mathrm{tr}\{H^T\nabla^2 f(\mathbf{x})H\} = \frac{1}{2}\mathrm{tr}\{HH^T\nabla^2 f(\mathbf{x})\}.$$

Next, define the scalar $h > 0$ and the $d \times d$ matrix A to satisfy

$$H = hA \quad \text{where } |A| = 1.$$

Choosing A to have determinant equal to 1 means that the elliptical shape of the kernel is entirely controlled by the matrix AA^T and the size of the kernel is entirely controlled by the scalar h. Observe that this parameterization is entirely general and permits different smoothing parameters for each dimension. For example, if

$$H = \begin{pmatrix} h_1 & & 0 \\ & \ddots & \\ 0 & & h_d \end{pmatrix}; \qquad \text{then} \quad H = h \cdot \begin{pmatrix} h_1/h & & 0 \\ & \ddots & \\ 0 & & h_d/h \end{pmatrix},$$

where $h = (h_1 h_2 \cdots h_d)^{1/d}$ is the geometric mean of the d smoothing parameters. Check that $|A| = 1$.

It follows that

$$\text{Bias}\{\hat{f}(\mathbf{x})\} = \frac{1}{2} h^2 \text{tr}\{AA^T \nabla^2 f(\mathbf{x})\}, \tag{6.48}$$

so that

$$\text{AISB} = \frac{1}{4} h^4 \int_{\Re^d} \left[\text{tr}\{AA^T \nabla^2 f(\mathbf{x})\}\right]^2 d\mathbf{x}.$$

As usual, the variance term is dominated by $\text{E}K_H(\mathbf{x} - \mathbf{x}_i)^2$; therefore,

$$\text{Var}\{\hat{f}(\mathbf{x})\} = \frac{f(\mathbf{x})}{n|H|} \int_{\Re^d} K(\mathbf{w})^2 \, d\mathbf{w} \quad \Rightarrow \quad \text{AIV} = \frac{R(K)}{nh^d}. \tag{6.49}$$

Together, these results may be summarized in a theorem.

Theorem 6.5: *For a general multivariate kernel estimator* (6.45) *parameterized by $H = hA$,*

$$\text{AMISE} = \frac{R(K)}{nh^d} + \frac{1}{4} h^4 \int_{\Re^d} \left[\text{tr}\{AA^T \nabla^2 f(\mathbf{x})\}\right]^2 d\mathbf{x}. \tag{6.50}$$

In spite of appearances, this is not using the same bandwidth in each dimension, but rather is applying a general elliptically shaped kernel at an arbitrary rotation.

The integral in Equation (6.50) will be quite complicated to evaluate unless the matrix A has a very simple structure. However, Wand and Jones (1995) provide a clever expression in their Section 4.3 that makes it much easier to evaluate this integral. Define the symmetric matrix $M = 2\nabla^2 f(\mathbf{x}) - \text{Diag}\left[\nabla^2 f(\mathbf{x})\right]$. Put the lower triangular portion of M into a vector $\mathbf{n}$ of length $d(d+1)/2$, a procedure referred to as the half-vectorization operation, $\mathbf{n} = \text{vech}M$. Their final expression is the scalar

$$\left(\text{vech}AA^T\right)^T \left[\int \mathbf{n}\mathbf{n}^T d\mathbf{x}\right] \left(\text{vech}AA^T\right). \tag{6.51}$$

The integral is applied to the elements of the $\frac{d(d+1)}{2} \times \frac{d(d+1)}{2}$ matrix $\mathbf{n}\mathbf{n}^T$. The integrals can be computed or estimated using integration by parts in many cases. Since AA^T is also a symmetric $d \times d$ matrix, $\text{vech}AA^T$ is a vector of length $\frac{d(d+1)}{2}$. See Wand and Jones (1995) and Problem 6.38 for further details.

6.3.3 Boundary Kernels for Irregular Regions

Staniswalis et al. (1993) showed that a boundary kernel for an arbitrarily complicated domain may be constructed by a simple device. Suppose that an estimate at $\mathbf{x}$ is desired. They propose using a kernel with spherical support, with radius h. Only samples $\mathbf{x}_i$ in the sphere of radius h around $\mathbf{x}$ influence the estimate. For each sample $\mathbf{x}_i$ in that region, determine if the diameter on which it falls (the center being the estimation point $\mathbf{x}$) intersects the boundary. If it does, construct a one-dimensional boundary kernel along that diameter. Repeating this construction for all samples in the sphere, the authors prove that the resulting estimate retains the correct order of bias.

6.4 GENERALITY OF THE KERNEL METHOD

6.4.1 Delta Methods

Walter and Blum (1979) catalogued the common feature of the already growing list of different density estimators. Namely, each could be reexpressed as a kernel estimator. Such a demonstration for orthogonal series estimators was given in Section 6.1.3. Reexamining Equation (3.2), even the histogram can be thought of as a kernel estimator. Surprisingly, this result was shown to hold even for estimators that were solutions to optimization problems. For example, consider one of the several maximum penalized likelihood (MPL) criteria suggested by Good and Gaskins (1972):

$$\max_f \left[\sum_{i=1}^n \log f(x_i) - \alpha \int_{-\infty}^{\infty} f'(x)^2\, dx\right] \qquad \text{for some } \alpha > 0. \tag{6.52}$$

Without the *roughness penalty term* in (6.52), the solution would be the empirical pdf. The many MPL estimators were shown to be kernel estimators by de Montricher et al. (1975) and Klonias (1982). The form of the kernel solutions differs in that the weights on the kernels were not all equal to $1/n$. For some other density estimation algorithms, the equivalent kernel has weight $1/n$ but has an adaptive bandwidth. A simple example of this type is the kth nearest-neighbor (k-NN) estimator. The k-NN estimate at x is equivalent to a histogram estimate with a bin centered on x with bin width sufficiently large so that the bin contains k points (in two and three dimensions, the histogram bin shape is a circle and a sphere, respectively). Thus the equivalent kernel in $\Re^d$ is simply a uniform density over the unit ball in $\Re^d$, but with bin widths that adapt to x.

6.4.2 General Kernel Theorem

There is a theoretical basis for the empirical observations of Walter and Blum that many algorithms may be viewed as generalized kernel estimates. Terrell provided a theorem to this effect that contains a constructive algorithm for obtaining the generalized kernel of any density estimator (see Terrell and Scott (1992)). The construction is not limited to nonparametric estimators, a fact that is exploited later.

Theorem 6.6: *Any density estimator that is a continuous and Gâteaux differentiable functional on the empirical distribution function may be written as*

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, x_i, F_n), \tag{6.53}$$

where K is the Gâteaux derivative of $\hat{f}$ under variation of x_i.

The Gâteaux derivative of a functional T at the function ϕ in the direction of the function η is defined to be

$$DT(\phi)[\eta] = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\big[T(\phi + \epsilon\,\eta) - T(\phi)\big]. \tag{6.54}$$

Theorem 6.6, which is proved below, has an analogous multivariate version (Terrell and Scott, 1992). The kernel K simply measures the influence of x_i on $\hat{f}(x)$. As F_n converges to F, then asymptotically, the form of K is independent of the remaining $n-1$ observations. Thus, any continuous density estimator may be written (asymptotically) as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, x_i, n) = \frac{1}{n}\sum_{i=1}^{n} K_n(x, x_i). \tag{6.55}$$

6.4.2.1 Proof of General Kernel Result The empirical cdf in Equation (2.1) can be written in the unusual form

$$F_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} I_{[x_i,\infty)}(\cdot). \tag{6.56}$$

Write the density estimator as an operator $\hat{f}(x) = T_x\{F_n\}$. Define

$$\begin{aligned} K(x,y,F_n) &\equiv \lim_{\epsilon\to 0} \frac{1}{\epsilon}[T_x\{(1-\epsilon)F_n + \epsilon I_{[y,\infty)}\} - (1-\epsilon)T_x\{F_n\}] \qquad (6.57) \\ &= \lim_{\epsilon\to 0} \frac{1}{\epsilon}[T_x\{F_n + \epsilon(I_{[y,\infty)} - F_n)\} - T_x\{F_n\}] + T_x\{F_n\} \\ &= DT_x(F_n)[I_{[y,\infty)} - F_n] + \hat{f}(x), \end{aligned}$$

where $DT(\phi)[\eta]$ is the Gâteaux derivative of T at ϕ in the direction η. Proposition (2.7) of Tapia (1971) shows that the Gâteaux derivative is linear in its second argument, so

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} K(x,x_i,F_n) &= \frac{1}{n}\sum_{i=1}^{n} DT_x(F_n)\left[I_{[x_i,\infty)} - F_n\right] + \hat{f}(x) \\ &= DT_x(F_n)\left[\frac{1}{n}\sum_{i=1}^{n} I_{[x_i,\infty)} - F_n\right] + \hat{f}(x) \\ &= 0 + \hat{f}(x), \end{aligned}$$

where the term in brackets is 0 by Equation (6.56). Note that by linearity, the Gâteaux variation in the direction 0, $DT(\phi)[0]$, is identically 0. This concludes the proof.

6.4.2.2 Characterization of a Nonparametric Estimator An estimator is defined to be nonparametric when it is consistent in the mean square for a large class of density functions. With a little effort, this definition translates into specific requirements for the equivalent kernel, $K_n(x,y)$. From the many previous examples, a nonparametric estimator that is consistent must be *local*; that is, the influence of sample points outside an ϵ–neighborhood of x must vanish as $n \to \infty$. As suggested by the term "delta function sequences" coined by Watson and Leadbetter (1963),

$$\lim_{n\to\infty} K_n(x,x_i,F_n) = \delta(x - x_i).$$

However, K_n must not equal $\delta(x - x_i)$ for finite n, as was the case in Section 6.1.3 with the orthogonal series estimator with all $\hat{f}_\nu$ coefficients included.

Beginning with the bias of the asymptotic kernel form in (6.55)

$$\begin{aligned} \mathrm{E}\hat{f}(x) = \mathrm{E}K_n(x,X) &= \int K_n(x,y)f(y)\,dy \\ &= \int K_n(x,y)\left[f(x) + (y-x)f'(x) + (y-x)^2 f''(\xi_y)/2\right] dy \end{aligned}$$

by the exact version of Taylor's theorem where $\xi_y \in (x,y)$. To be asymptotically unbiased, all three terms in the following must vanish:

$$\begin{aligned} \mathrm{Bias}\{\hat{f}(x)\} = &f(x)\left[\int K_n(x,y)\,dy - 1\right] + f'(x)\int K_n(x,y)\,(y-x)\,dy \\ &+ \frac{1}{2}\int K_n(x,y)\,(y-x)^2 f''(\xi_y)\,dy. \qquad (6.58) \end{aligned}$$

Therefore, the first condition is that

$$\lim_{n\to\infty}\int K_n(x,y)\,dy = 1 \quad \forall x \in \Re^1.$$

Assume that the estimator has been rescaled so that the integral is exactly 1 for all n. Therefore, define the random variable Y to have pdf $K_n(x,\cdot)$. The second condition is that

$$\begin{aligned}\lim_{n\to\infty}\int K_n(x,y)\,y\,dy &= \int K_n(x,y)\,x\,dy = x \quad \forall x\\ \Longrightarrow \quad \lim_{n\to\infty} K_n(x,y) &= \delta(y-x)\end{aligned}$$

as Watson and Leadbetter (1963) and Walter and Blum (1979) suggested. The precise behavior of the bias is determined by the rate at which this happens. For example, suppose that $\int K_n(x,y)\,y\,dy = x$ for all n and that

$$\sigma_{x,n}^2 = \int K_n(x,y)\,(y-x)^2\,dy \neq 0 \tag{6.59}$$

for finite n so that the first two moments of the random variables $Y \sim \left(x,\sigma_{x,n}^2\right)$ and $T \equiv (Y-x)/(\sigma_{x,n}) \sim (0,1)$. Suppose that the density function of T, which is a simple linear transformation of K_n, converges to a nice density:

$$\tilde{L}_n(x,t) = K_n(x, x+\sigma_{x,n}\,t)\,\sigma_{x,n} \to L(x,t) \quad \text{as } n\to\infty.$$

Then the last bias term in (6.58) may be approximated by

$$\begin{aligned}\frac{1}{2}f''(x)\int K_n(x,y)\,(y-x)^2\,dy &= \frac{1}{2}f''(x)\int K_n(x,x+\sigma_{x,n}\,t)\,(\sigma_{x,n}\,t)^2\sigma_{x,n}\,dt\\ &= \frac{1}{2}\sigma_{x,n}^2 f''(x)\int t^2\tilde{L}(x,t)\,dt\\ &\approx \frac{1}{2}\sigma_{x,n}^2 f''(x)\int t^2 L(x,t)\,dt = \frac{1}{2}\sigma_{x,n}^2 f''(x)\end{aligned}$$

since the $\mathrm{Var}(T)=1$, so that the bias is $O\left(\sigma_{x,n}^2\right)$, the familiar rate for second-order kernels. Thus the third condition is that $\sigma_{x,n}\to 0$ as $n\to\infty$.

In order that the variance vanish asymptotically, consider

$$\begin{aligned}\mathrm{Var}\{\hat{f}(x)\} &= \frac{1}{n}\mathrm{Var}\{K_n(x,X)\} \leq \frac{1}{n}\mathrm{E}[K_n(x,X)^2] = \frac{1}{n}\int K_n(x,y)^2 f(y)\,dy\\ &= \frac{1}{n}\int K_n(x,x+\sigma_{x,n}\,t)^2 f(x+\sigma_{x,n}\,t)\,\sigma_{x,n}\,dt\\ &= \frac{1}{n\sigma_{x,n}}\int \tilde{L}_n(x,t)^2\left[f(x)+\cdots\right]dt \approx \frac{f(x)\,R[L(x,\cdot)]}{n\sigma_{x,n}}.\end{aligned}$$

Thus the fourth condition required is that the variance of the equivalent kernel satisfy $n\sigma_{x,n} \to \infty$ as $n \to \infty$.

These findings may be summarized in a theorem, the outline of which is presented in Terrell (1984).

Theorem 6.7: *Suppose $\hat{f}$ is a density estimator with asymptotic equivalent kernel $K_n(x,y)$ and that $\sigma^2_{x,n}$ defined in* (6.59) *is bounded and nonzero. Then $\hat{f}$ is a nonparametric density estimator if, for all $x \in \Re^1$,*

$$\begin{aligned}\lim_{n\to\infty} \int K_n(x,y)\,dy &= 1 \\ \lim_{n\to\infty} \int K_n(x,y) y\,dy &= x \\ \lim_{n\to\infty} \sigma_{x,n} &= 0 \\ \lim_{n\to\infty} n\sigma_{x,n} &= \infty.\end{aligned} \tag{6.60}$$

6.4.2.3 Equivalent Kernels of Parametric Estimators Theorem 6.6 shows how to construct the kernel for any density estimator, parametric or nonparametric. For example, consider the parametric estimation of $f = N(\mu, 1) = \phi(x|\mu, 1)$ by $\hat{\mu} = \bar{x}$. Thus $T_x(F_n) = \phi(x|\bar{x}, 1)$. Examining the argument in the first line in Equation (6.57) and comparing it to the definition of the ecdf in (6.56), it becomes apparent that the empirical pdf $n^{-1}\sum_{i=1}^n \delta(x - x_i)$ is being replaced by

$$\frac{1-\epsilon}{n}\sum_{i=1}^n \delta(x - x_i) + \epsilon\,\delta(x - y),$$

which is the original empirical pdf with a small portion ϵ of the probability mass proportionally removed and placed at $x = y$. The sample mean of this perturbed epdf is $(1-\epsilon)\bar{x} + \epsilon y$. Thus the kernel may be computed directly from Equation (6.57) by

$$\begin{aligned}K(x,y,F_n) &= \lim_{\epsilon\to 0} \frac{1}{\epsilon}\Big[\phi\big(x|(1-\epsilon)\bar{x} + \epsilon\, y, 1\big) - (1-\epsilon)\,\phi\big(x|\bar{x}, 1\big)\Big] \\ &= \frac{1 + (y-\bar{x})(x-\bar{x})}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\bar{x})^2}\end{aligned} \tag{6.61}$$

by a Taylor's series (see Problem 6.28). The asymptotic equivalent kernel is

$$K(x,y) = \lim_{n\to\infty} K(x,y,F_n) = \frac{1 + (y-\mu)(x-\mu)}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\mu)^2}.$$

This kernel is never *local* and so the estimator is *not* nonparametric (if there was any doubt). Note that the *parametric kernel estimator* with kernel (6.61) is quite good, as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1+(x_i-\bar{x})(x-\bar{x})}{\sqrt{2\pi}}\, e^{-(x-\bar{x})^2/2} = \frac{1}{\sqrt{2\pi}}\, e^{-(x-\bar{x})^2/2}.$$

Of course, just as in the parametric setting, this "kernel" estimator will always be $\phi(x|\bar{x},1)$ for all datasets, no matter how non-normal the underlying density is.

6.5 CROSS-VALIDATION

6.5.1 Univariate Data

The goal is to go beyond the theoretical results for optimal bandwidth specification and to achieve practical data-based algorithms. The oversmoothed and normal reference rules provide reasonable initial choices, together with the simple modifications based on sample skewness and kurtosis given in Section 4.1.2. (The bandwidth modification factors for positive kernels and the frequency polygon are identical.) The unbiased and biased cross-validation algorithms for the histogram are easily extended to both kernel and ASH estimators. However, the development of bandwidth selection algorithms for kernel estimators has progressed beyond those for the histogram.

The importance of choosing the optimal bandwidth is easily overstated. From the sensitivity results in Table 3.3, any choice of h within 15–20% of h^* will often suffice. Terrell (1990) has even suggested that an oversmoothed rule be generally used. With real data, it is relatively easy to examine a sequence of estimates based on the sequence of smoothing parameters

$$h = \hat{h}_{\mathrm{OS}}/1.05^k \qquad \text{for } k = 0,1,2,\ldots,$$

starting with the sample oversmoothed bandwidth ($\hat{h}_{\mathrm{OS}}$), and stopping when the estimate displays some instability and very local noise near the peaks. Silverman (1978a) has characterized the expected amount of local noise present in $\hat{f}''(x)$ in the L_∞ norm (see Equation (6.65)). He suggested examining plots of $\hat{f}''(x)$ interactively in addition to $\hat{f}(x)$, a procedure referred to as the *test graph method*.

The biased cross-validation algorithm, which estimates $R(f'')$ from a kernel estimate, attempts to correct the overestimation by $R(\hat{f}''(\cdot))$ that is the direct result of presence of the local noise (see Section 6.5.1.3). However, the power of the interactive approach to bandwidth selection should not be underestimated. The interactive process can be quite painless in systems supporting animation, such as LISP-STAT (Tierney, 1990).

It should be emphasized that from an exploratory point of view, all choices of the bandwidth h lead to useful density estimates. Large bandwidths, on the one hand, provide a picture of the global structure in the unknown density, including general features such as skewness, outliers, clusters, and location. Small bandwidths, on the other hand, reveal local structure which may or may not be present in the true density. Furthermore, the optimality of h is dependent not only on the choice of metric L_p but also on the feature in the density to be emphasized $(F, f, f', f'', \ldots)$.

However, the desire for fully automatic and reliable bandwidth selection procedures has led inevitably to a series of novel algorithms. These objective (not subjective) procedures often have the stated goal of finding a bandwidth that minimizes the actual L_2 error rather than using the bandwidth that minimizes the expected L_2 error (MISE). In an early paper, Wahba (1981) expressed the expectation that her generalized cross-validation algorithm would accomplish this goal. The best L_2 bandwidth, h_{ISE}, remained the target in unbiased cross-validation for Hall and Marron (1987a, b). Scott and Factor (1981) had expressed the view that h_{MISE} was an appropriate target. The MISE-optimal bandwidth depends only on (f_n), whereas the ISE-optimal bandwidth depends on the sample as well, $(f, x, \{x_i\})$. However, it has been shown that the sample correlation between h_{ISE} and $\hat{\sigma}$ approached -0.70 for normal data (Scott and Terrell, 1987; Scott, 1988b). Given such a large negative correlation with the scale of the data, tracking $\hat{h}_{\mathrm{ISE}}$ closely would require guessing whether $\hat{\sigma} > \sigma$, or vice versa, a difficult task.

Of some interest is the fact that while $\hat{h}_{\mathrm{ISE}} \approx h_{\mathrm{MISE}}$,

$$\frac{\sigma_{\hat{h}_{\mathrm{ISE}}}}{h_{\mathrm{MISE}}} = O(n^{-1/10}),$$

so that the ISE-optimal bandwidth is only slowly converging to h_{MISE}, as was shown by Hall and Marron (1987a). Scott and Terrell (1987) showed that unbiased cross-validation (UCV) and biased CV (BCV) bandwidths converged to h_{MISE} at the same slow rate. Some of the more recent extensions have been able to push the relative convergence rate all the way to $O(n^{-1/2})$, which is the best rate possible (Hall and Marron, 1991). These procedures require the introduction of one or two auxiliary smoothing parameters. Given the slow rate of convergence of the unattainable $\hat{h}_{\mathrm{ISE}}$, it is perhaps unclear whether there is a practical advantage to be had in the faster rates. This question is examined further in Section 6.5.1.5.

In an empirical study of the performance of nine cross-validation algorithms and four sampling densities, Jones and Kappenman (1992) report the "broad equivalence of almost all" of these algorithms with respect to the observed ISE. Other simulations (Scott and Factor, 1981; Bowman, 1985; Scott and Terrell, 1987; Park and Marron, 1990) have reported less equivalence among the estimated smoothing parameter values themselves. Jones and Kappenman reported that the fixed AMISE bandwidth h^* outperformed all nine CV algorithms with respect to ISE. Their results reinforce the suggestion that h^* is an appropriate target and that any choice within 15–20% of h^* should be adequate. Most efforts now focus on h_{MISE} as the target bandwidth.

6.5.1.1 Early Efforts in Bandwidth Selection The earliest data-based bandwidth selection ideas came in the context of orthogonal series estimators, which were discussed in Section 6.1.3. Using the Tarter–Kronmal weights (6.8) and the representation in Equation (6.5), the pointwise error is

$$\hat{f}(x) - f(x) = \sum_{v=-m}^{m} \hat{f}_v \phi_\nu(x) - \sum_{v=-\infty}^{\infty} f_v \phi_\nu(x).$$

As the basis functions are orthonormal,

$$\text{ISE} = \sum_{v=-m}^{m} ||\hat{f}_v - f_v||^2 + \sum_{v \notin [-m,m]} ||f_v||^2 .$$

Recall that MISE = E(ISE). Tarter and Kronmal's selection procedure provided unbiased estimates of the increment in MISE in going from a series with $m-1$ terms to one with m terms (noting the equality of the $\pm\nu$ MISE terms):

$$\text{MISE}(m) - \text{MISE}(m-1) = 2\{\text{E}\,||\hat{f}_m - f_m||^2 - ||f_m||^2\}. \tag{6.62}$$

Unbiased estimates of the two terms on the right-hand side may be obtained for the Fourier estimator in Equation (6.6), as $\text{E}\hat{f}_v = f_v$ and $\text{E}\hat{f}_v\hat{f}_v^* = (1-(n-1)|\hat{f}_v|^2)/n$, where $\hat{f}_v^*$ denotes the complex conjugate of $\hat{f}_v$; hence, $\text{Var}(\hat{f}_v) = \text{E}\hat{f}_v\hat{f}_v^* - |\hat{f}_v|^2 = (1-|\hat{f}_v|^2)/n$. The data-based choice for m is achieved when the increment becomes *positive*. Notice that accepting the inclusion of the mth coefficient in the series estimator is the result of the judgment that the additional variance of $\hat{f}_m$ is less than the reduction in bias $||f_m||^2$. Usually, fewer than six terms are chosen, so that only relatively coarse adjustments can be made to the smoothness of the density estimator. Sometimes the algorithm misses higher order terms. But the real significance of this algorithm lies in its claim to be the first unbiased cross-validation algorithm for a density estimator.

Likewise, the credit for the first biased cross-validation algorithm goes to Wahba (1981) with her generalized cross-validation algorithm. She used the same unbiased estimates of the Fourier coefficients as Tarter and Kronmal, but with her smoother choice of weights, she lost the simplicity of examining the incremental changes in MISE. However, those same unbiased estimates of the Fourier coefficients lead to a good estimate of the AMISE. By ignoring all the unknown Fourier coefficients for $|v| > n/2$, a small bias is introduced. Both groups recommend plotting the estimated risk function in order to find the best data-based smoothing parameter rather than resorting to (blind) numerical optimization.

The earliest effort at choosing the kernel smoothing parameter in a fully automatic manner was a modified maximum likelihood algorithm due to Habbema et al. (1974) and Duin (1976). While it has not withstood the test of time, it is significant for having introduced a leave-one-out modification to the usual maximum likelihood (ML) criterion. Choosing the bandwidth h to maximize the usual ML criterion results in the (rough) empirical pdf:

$$0 = \arg\max_h \sum_{i=1}^{n} \log \hat{f}(x_i;h) \quad \Rightarrow \quad \hat{f}(x;h=0) = \frac{1}{n}\sum_{i=1}^{n} \delta(x-x_i),$$

a solution with "infinite" likelihood. The problem arises since, as $h \to 0$, the contribution to the likelihood at $x = x_i$ from the point x_i itself becomes infinite. The authors sought to eliminate that "self-contribution" by modifying the ML criterion:

$$\max_h \sum_{i=1}^{n} \log \hat{f}_{-i}(x_i;h),$$

where $\hat{f}_{-i}(x_i;h)$ is a kernel estimator based on the $n-1$ data points excluding x_i and then evaluated there. In spite of some promising empirical small sample and consistency results (Chow et al., 1983), the algorithm was shown to be overly influenced by outliers and tight clusters (Schuster and Gregory, 1981; Scott and Factor, 1981). With a finite support kernel, for example, the bandwidth cannot be less than $x_{(2)} - x_{(1)}$, which is the distance between the first two order statistics; for many densities the distance between these order statistics does not converge to zero and so the bandwidth does not converge to zero as required.

A simple fixed-point algorithm was suggested by Scott et al. (1977). For kernel estimates, the only unknown in h^* in (6.18) is $R(f'')$. If a normal kernel is used, then a straightforward calculation finds that

$$R(\hat{f}_h'') = \frac{3}{8\sqrt{\pi}n^2h^5}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(1-\Delta_{ij}^2+\frac{1}{12}\Delta_{ij}^4\right)e^{-\frac{1}{4}\Delta_{ij}^2}, \tag{6.63}$$

where $\Delta_{ij} = (x_i - x_j)/h$. Following Equation (6.18), the search for a fixed-point value for h^* is achieved by iterating

$$h_{k+1} = \left[\frac{R(K)}{n\sigma_K^4 R(\hat{f}_{h_k}'')}\right]^{1/5},$$

with h_0 chosen to be the normal reference bandwidth. As the ratio of the optimal bandwidths for estimating f and f'' diverges as $n \to \infty$, it is clear that the algorithm is not consistent. That the algorithm worked well for small samples (Scott and Factor, 1981) is not surprising since the optimal bandwidths are reasonably close to each other for small samples. Note that this algorithm as stated provides no estimate of the MISE. It is a simple matter to use the roughness estimate (6.63) in Equation (6.18), following the lead of Wahba, to obtain

$$\widehat{\text{AMISE}}(h) = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 R(\hat{f}_h''). \tag{6.64}$$

Finding the minimizer of Equation (6.64) not only provides a data-based bandwidth estimate, but also an estimate of the MISE. This idea was resurrected with biased cross-validation using a consistent estimator for $R(f'')$ rather than Equation (6.63). Alternatively, a much wider bandwidth appropriate to f'' rather than f might be used in the iteration. Sheather (1983) proposed such a scheme while trying to estimate the density at the origin, providing an example of the plug-in (PI) estimators discussed by Woodroofe (1970). Sheather's motivation is discussed in Sheather and Jones (1991). In a talk in 1991, Gasser also reported success in this way for choosing a global bandwidth by inflating the bandwidth used in $R(\hat{f}_n'')$ by the factor $n^{-1/10}$.

For a normal kernel, Silverman (1978a) proved that

$$\frac{\sup |\hat{f}'' - \mathrm{E}\hat{f}''|}{\sup |\mathrm{E}\hat{f}''|} \approx 0.4. \tag{6.65}$$

He proposed choosing the bandwidth where it appears that the ratio of the noise to the signal is 0.4. This ratio is different for other kernels. The *test graph* procedure can be used in the bivariate case as well.

6.5.1.2 Oversmoothing The derivation of the oversmoothed rule for kernel estimators will be constructive, unlike the earlier derivations of the histogram and frequency polygon oversmoothed rules. The preferred version uses the variance as a measure of scale. Other scale measures have been considered by Terrell and Scott (1985) and Terrell (1990).

Consider the variational problem

$$\min_f \int_{-\infty}^{\infty} f''(x)^2 dx \qquad \text{s/t} \quad \int f = 1 \text{ and } \int x^2 f = 1. \tag{6.66}$$

Clearly, the solution will be symmetric. The associated Lagrangian is

$$L(f) = \int_{-\infty}^{\infty} f''(x)^2 dx + \lambda_1 \left(\int f - 1 \right) + \lambda_2 \left(\int x^2 f - 1 \right).$$

At a solution, the Gâteaux variation, defined in Equation (6.54), of the Lagrangian in any "direction" η must vanish. For example, the Gâteaux variation of $\Psi(f) = \int f''(x)^2$ is

$$\begin{aligned}\Psi'(f)[\eta] &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[\int \left[(f + \epsilon\, \eta)'' \right]^2 - \int \left[f'' \right]^2 \right] \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[\int \left(f''^2 + 2\epsilon f'' \eta'' + \epsilon^2 \eta''^2 - f''^2 \right) \right] = \int 2 f'' \eta''.\end{aligned}$$

Computing the Gâteaux variation of $L(f)$, we have

$$\begin{aligned} 0 = L'(f)[\eta] &= \int 2f''(x)\eta''(x) + \lambda_1 \int \eta(x) + \lambda_2 \int x^2\eta(x) \\ &= 2f''(x)\eta'(x)\,\Big|_{-c}^{c} - 2f'''(x)\eta(x)\,\Big|_{-c}^{c} + \int \left[2f^{iv}(x) + \lambda_1 + \lambda_2 x^2\right]\eta(x) \end{aligned} \tag{6.67}$$

after integrating by parts twice and where c is the boundary, possibly infinite, for f. Now $\eta(\pm c)$ must vanish so that there is not a discontinuity in the solution; therefore, the second term vanishes. The remaining two terms must vanish for all feasible choices for η; therefore, $f''(\pm c)$ must vanish, leaving only the integral. It follows that the integrand must vanish and that f is a sixth-order polynomial with only even powers (by symmetry). Therefore, the solution must take the form

$$f(x) = a(x-c)^3(x+c)^3$$

so that $f''(\pm c) = 0$. The two constraints in (6.66) impose two linear conditions on the unknowns (a, c), with the result that

$$f^*(x) = \frac{35}{69{,}984}(9-x^2)_+^3 \qquad \text{and} \qquad R[(f^*)''] = \frac{35}{243}.$$

A simple change of variables shows that $R(f'') \geq 35/(243\sigma^5)$; therefore,

$$h^* = \left[\frac{R(K)}{n\sigma_K^4 R(f'')}\right]^{1/5} \leq \left[\frac{243\sigma^5 R(K)}{35 n\sigma_K^4}\right]^{1/5} \Rightarrow \tag{6.68}$$

$$\boxed{\text{Oversmoothing rule:} \qquad h_{\text{OS}} = 3\left[\frac{R(K)}{35\,\sigma_K^4}\right]^{1/5} \sigma n^{-1/5}.} \tag{6.69}$$

For the normal kernel, $h_{\text{OS}} = 1.144\sigma n^{-1/5}$. For the biweight kernel, it is *exactly* $h_{\text{OS}} = 3\sigma n^{-1/5}$. The rule is 1.08 times wider than the normal reference rule.

6.5.1.3 Unbiased and Biased Cross-Validation The presentation in the section will rely heavily on the references for certain details. The overall flavor is the same as in the application to the histogram and frequency polygon.

The remarkable fact is that the UCV justification for the histogram is entirely general. Thus the definition in Equation (3.52) applies in the case of a kernel estimator. For the case of a normal kernel, Rudemo (1982) and Bowman (1984) showed that (replacing $n \pm 1$ with n for simplicity)

$$\boxed{\text{UCV}(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{n^2 h\sqrt{\pi}} \sum_{i<j} \left(e^{-\Delta_{ij}^2/4} - \sqrt{8}\, e^{-\Delta_{ij}^2/2}\right),} \tag{6.70}$$

which is a special case of the general formula (Scott and Terrell, 1987)

$$\mathrm{UCV}(h) = \frac{R(K)}{nh} + \frac{2}{n^2 h} \sum_{i<j} \gamma(\Delta_{ij}),$$

where

$$\gamma(\Delta) = \int K(w)\, K(w+\Delta)\, dw.$$

The BCV algorithm follows from the result (Scott and Terrell, 1987) that

$$\mathrm{E}R(\hat{f}_h'') = R(f'') + \frac{R(K'')}{nh^5} + O(h^2),$$

where $R(K'')/(nh^5)$ is asymptotically a constant, representing the fixed but finite noise that exists in the kernel estimate. Therefore, $R(\hat{f}_h'') - R(K'')/(nh^5)$ is an asymptotically unbiased estimator for the unknown roughness $R(f'')$. Substituting into Equation (6.18), the estimate of the MISE becomes

$$\mathrm{BCV}(h) = \frac{R(K)}{nh} + \frac{\sigma_K^4}{2n^2 h} \sum_{i<j} \phi(\Delta_{ij}), \tag{6.71}$$

where

$$\phi(\Delta) = \int K''(w)\, K''(w+\Delta)\, dw.$$

The similarity of the general UCV and BCV formulas is remarkable, given their quite different origins. In the case of a normal kernel,

$$\boxed{\mathrm{BCV}(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{64n^2 h\sqrt{\pi}} \sum_{i<j} \left(\Delta_{ij}^4 - 12\Delta_{ij}^2 + 12\right) e^{-\Delta_{ij}^2/4}.} \tag{6.72}$$

As before, $\lim_{h\to\infty} \mathrm{BCV}(h) = 0$; therefore, $\hat{h}_{\mathrm{BCV}}$ is taken to be the largest local minimizer of $\mathrm{BCV}(h)$ less than or equal to the oversmoothed bandwidth.

The asymptotic theory of the CV criteria is straightforward once it is recognized that the stochastic part is all contained in the so-called U-statistics, which are double sums of the form

$$U_n = \sum_{i<j} H_n(X_i, X_j).$$

Hall (1984) proved that if the function H_n is symmetric and the random variable $\mathrm{E}[H_n(X,Y) \mid X] = 0$, then together with a certain moment condition,

$$U_n = \mathrm{AN}\left(0, \tfrac{1}{2} n^2\, \mathrm{E}H_n^2\right).$$

The asymptotic normality of UCV and BCV is not obvious because the number of terms in the double sum effectively shrinks, as the bandwidth is a decreasing function of n. Hall used an argument from degenerate Martingale theory to prove this theorem.

Scott and Terrell (1987) then proved that for a fixed bandwidth, h, the UCV and BCV functions were (1) both asymptotically normal; (2) both converged to AMISE(h); and (3) the asymptotic (vertical) variances of UCV and BCV at h are

$$\frac{2R(\gamma)R(f)}{n^2h} \quad \text{and} \quad \frac{\sigma_K^8 R(\phi)R(f)}{8n^2h}, \tag{6.73}$$

respectively. For kernels in the symmetric Beta family of the form (6.27), the (vertical) variance of UCV is at least 80 times greater than for the BCV criterion. This smaller vertical variance suggests that the actual minimizer of the BCV criterion will have smaller variance than $\hat{h}_{\text{UCV}}$. The variances in Equation (6.73) are of order $O(n^{-9/5})$ if $h = O(n^{-1/5})$. However, this rapid rate of decrease can be explained by the fact that the CV functions are themselves going to 0 at the rate $O(n^{-4/5})$. Thus the relevant quantity is the coefficient of variation (*C.V.*)

$$C.V. = \frac{\sqrt{\text{VarUCV}(h)}}{O[\text{UCV}(h)]} = \frac{O(n^{-9/10})}{O(n^{-4/5})} = O(n^{-1/10}),$$

as was claimed earlier. The astute reader will note that the bias has not been counted in this error in the BCV case; however, the bias turns out to be $O(n^{-1})$, and hence the squared bias is $O(n^{-2})$, which is of lower order than the variance.

Using a delta argument outlined below, Scott and Terrell (1987) showed that $\hat{h}_{\text{UCV}}$ and $\hat{h}_{\text{BCV}}$ converged to h_{AMISE} and were asymptotically normal with respective variances given by

$$\frac{2R(f)R[\Delta\gamma'(\Delta)]}{25n^2(h^*)^7\sigma_K^4 R(f'')^2} \quad \text{and} \quad \frac{R(f)R[\Delta\phi'(\Delta)]}{200n^2(h^*)^7 R(f'')^2}. \tag{6.74}$$

Observe that if $h^* = O(n^{-1/5})$, then these variances are $O(n^{-3/5})$, from which the (horizontal) coefficient of variation is found to still be of $O(n^{-1/10})$:

$$C.V. = \frac{\sqrt{\text{Var}\,\hat{h}_{\text{UCV}}}}{O(\hat{h}_{\text{UCV}})} = \frac{O(n^{-3/10})}{O(n^{-1/5})} = O(n^{-1/10}).$$

Again, for densities in the symmetric Beta family, the UCV variance is at least 16 times that of the BCV. These results were confirmed in a simulation study. However, it was noted that BCV performed poorly for several difficult densities without a very large dataset. This finding is not surprising, given that the basis for the BCV formula is AMISE, while the exact MISE is the basis of UCV.

It is instructive to outline the derivation of Equation (6.74). Clearly, the BCV bandwidth satisfies

$$\frac{d}{dh}[\text{BCV}(h)]\bigg|_{h=\hat{h}_{\text{BCV}}} = 0.$$

Noting that $\Delta'_{ij} = -(x_i - x_j)/h^2 = -\Delta_{ij}/h$, the derivative of BCV as defined in (6.71) equals

$$\frac{-R(K)}{nh^2} - \frac{\sigma_K^4}{2n^2h^2}\sum_{i<j}\phi(\Delta_{ij}) + \frac{\sigma_K^4}{2n^2h}\sum_{i<j}\phi'(\Delta_{ij})\frac{-\Delta_{ij}}{h} = 0$$

or

$$\sum_{i<j}[\phi(\Delta_{ij}) + \Delta_{ij}\phi'(\Delta_{ij})]\Big|_{h=\hat{h}_{\mathrm{BCV}}} = -\frac{2nR(K)}{\sigma_K^4}.$$

Define $\psi(\Delta) = \Delta\,\phi'(\Delta)$; then computing approximations to the moments of ϕ and ψ (a nontrivial calculation given in Section 9 of Scott and Terrell (1987)), it follows that

$$\sum_{i<j}\left[\phi(\Delta_{ij}) + \psi(\Delta_{ij})\right] = \mathrm{AN}\left(-2n^2h^5R(f''), n^2hR(f)R(\psi)/2\right).$$

Hence, combining these two results and rearranging,

$$-2n^2R(f'')\hat{h}^5_{\mathrm{BCV}} = \mathrm{AN}\left(-2nR(K)/\sigma_K^4, n^2h^*R(f)R(\psi)/2\right)$$

or

$$\hat{h}^5_{\mathrm{BCV}} = \mathrm{AN}\left(\frac{R(K)}{\sigma_K^4nR(f'')}, \frac{h^*R(f)R(\psi)}{8n^2R(f'')^2}\right). \tag{6.75}$$

Clearly, the asymptotic mean of $\hat{h}^5_{\mathrm{BCV}}$ is $(h^*)^5$. The random variable $\hat{h}_{\mathrm{BCV}}$ is the 1/5 power of that given in (6.75). Applying the delta method, it may be concluded that $\hat{h}_{\mathrm{BCV}}$ is AN with mean h^* and variance which may be computed by the formula

$$\mathrm{Var}\{g(h)\} = \left(\frac{\mathrm{d}g}{\mathrm{d}h}\right)^2_{h=h^*}\mathrm{Var}\{h\}.$$

Now $g(h) = h^5$ and $g'(h) = 5h^4$, so that

$$\mathrm{Var}\{\hat{h}_{\mathrm{BCV}}\} = \mathrm{Var}\{\hat{h}^5_{\mathrm{BCV}}\}/[25(h^*)^8]. \tag{6.76}$$

The variance (6.74) follows immediately combining (6.75) and (6.76).

Despite these apparently favorable findings, BCV does not qualify as a general replacement for UCV. UCV may be noisier but it tends to produce nearly unbiased smoothing parameters. However, there is a need for an auxiliary CV criterion since UCV is susceptible to certain problems. Clearly, BCV has its own set of limitations. But by carefully examining of the trio of smoothing parameters suggested by UCV, BCV, and OS as well as the shapes of the UCV and BCV curves, good bandwidths should be reliably available.

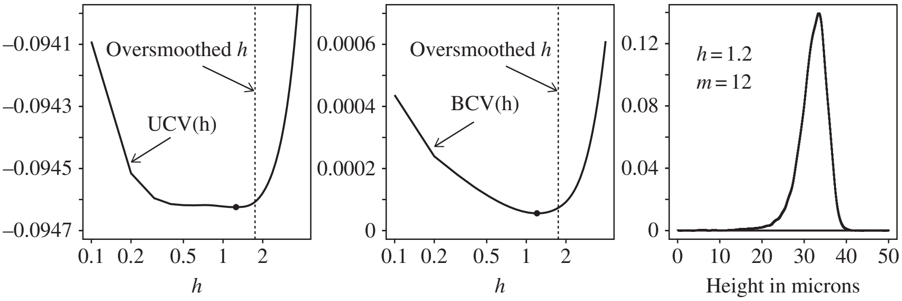

FIGURE 6.15 UCV and BCV estimates of the steel surface data ($n = 15,000$) using the triweight ASH. The UCV bandwidth was tied at 1.2 and 1.3, while the BCV bandwidth was 1.2 (shown): both estimates were virtually identical.

As a practical matter, both UCV and BCV involve $O(n^2)$ computation due to the double sums. If the normal kernel versions are used, the work can easily be prohibitive for $n > 500$. This work can be substantially reduced by using an ASH implementation. The computational details are given in Scott and Terrell (1987) and are not repeated here. Of course, code is available from several sources to perform these and other computations. The ASH implementation is illustrated in Figure 6.15 for the steel surface data. The standard deviation of these data is 3.513, so that the oversmoothed bandwidth for the triweight kernel is $\hat{h}_{\mathrm{OS}} = 1.749$ from Equation (6.69). Observe that the UCV estimate is flat over a relatively wide interval even with such a large dataset. However, the minima of the two criteria are virtually identical, $h_{\mathrm{BCV}} = 1.2$ and $h_{\mathrm{UCV}} = 1.3$ for the triweight kernel. (The triweight estimate with equivalent smoothing parameter $h = 0.67$ is shown in Figure 6.5.) The original data were binned into 500 intervals, so that the use of the ASH implementation of the CV and estimation functions is natural.

6.5.1.4 Bootstrapping Cross-Validation Taylor (1989) investigated a data-based algorithm for choosing h based on bootstrap estimates of the $\mathrm{MSE}\{\hat{f}(x)\}$ and $\mathrm{MISE}\{\hat{f}\}$. The bootstrap resample is not taken from the empirical pdf f_n as in the *ordinary bootstrap*, but rather the bootstrap sample $\{x_1^*, x_2^*, \ldots, x_n^*\}$ is a random sample from the candidate kernel density estimate $\hat{f}(x;h)$ itself. Such a resample is called a *smoothed bootstrap sample* and is discussed further in Chapter 9. Letting E_* represent expectation with respect to the smoothed bootstrap random sample, Taylor examined

$$\begin{aligned}
\mathrm{BMSE}_*(x;h) &= \mathrm{E}_*[\hat{f}_*(x;h) - \hat{f}(x;h)]^2 \\
&= \mathrm{E}_*\left[\frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i^*) - \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)\right]^2 \qquad (6.77) \\
\mathrm{BMISE}_*(h) &= \int_x \mathrm{BMSE}_*(x;h)\,dx.
\end{aligned}$$

The interesting result is that if the resample comes from the empirical density function, then the bootstrap quantities in Equation (6.77) estimate only the variance and not the bias. The bias introduced by the smoothed bootstrap is precisely what is needed to mimic the true unknown bias for that choice of the smoothing parameter h. A bit of extra variance is introduced by the smoothed bootstrap, but it is easily removed.

The algebra involved in computing the BMSE$_*$ is perhaps unfamiliar but straightforward. For example, $\mathrm{E}_*\hat{f}(x) = \mathrm{E}_* K_h(x - x^*)$ and

$$\mathrm{E}_* K_h(x-x^*) = \int K_h(x-y)\hat{f}(y)\,dy = \frac{1}{n}\sum_{i=1}^{n}\int K_h(x-y)K_h(y-x_i)\,dy.$$

The computation of $\mathrm{E}_* K_h(x-x^*)^2$ is a trivial extension. For the particular case of the normal kernel, the convolutions indicated in the bootstrap expectation may be computed in closed form. After a few pages of work using the normal kernel and adjusting the variance, Taylor proposes minimizing

$$\boxed{\mathrm{BMISE}_*(h) = \frac{1+\frac{\sqrt{2}}{n}\sum_{i<j}\left[\sqrt{2}\,e^{-\frac{\Delta_{ij}^2}{4}} - \frac{4}{\sqrt{3}}e^{-\frac{\Delta_{ij}^2}{6}} + e^{-\frac{\Delta_{ij}^2}{8}}\right]}{2nh\sqrt{\pi}}} \tag{6.78}$$

Taylor shows that BMISE$_*(h)$ has the same order variance as UCV(h) and BCV(h), but with a smaller constant asymptotically.

In the Figure 6.16, the three CV functions, with comparable vertical scalings, are plotted for the snowfall data along with the corresponding density estimates using the normal kernel. Many commonly observed empirical results are depicted in these graphs. The BCV and bootstrap curves are similar, although BCV(h) has a sharper minimum. Both the biased and bootstrap CV functions have minima at bandwidths greater than the oversmoothed bandwidth $h_{\mathrm{OS}} = 11.9$. This serves to emphasize how difficult estimating the bias is with small samples. The unbiased CV function better reflects the difficulty in precisely estimating the bias by presenting a curve that is flat over a wide interval near the minimum. There is some visual evidence of three modes in this data. Since these CV functions do not take account of the time series nature of these data and the first-order autocorrelation is -0.6, a smaller bandwidth is probably justifiable (see Chiu (1989) and Altman (1990)).

These computations are repeated for the Old Faithful geyser dataset, which is clearly bimodal, and the results depicted in Figure 6.17. It is interesting to note that both the biased and bootstrap CV functions have two local minima. Fortunately, only one local minima is smaller than the oversmoothed upper bound $h_{\mathrm{OS}} = 0.47$, although the bootstrap curve barely exhibits the second local minimum. The UCV function leads to a narrow bandwidth and a density estimate that seems clearly undersmoothed given the sample size. These observations seem to recur with many "real" datasets

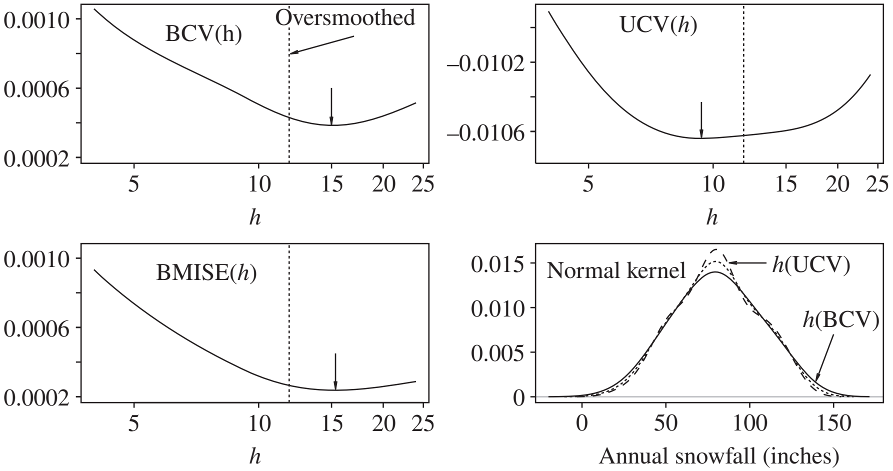

FIGURE 6.16 Normal kernel cross-validation algorithms and density estimates for the snowfall data ($n = 63$). The CV bandwidths are indicated by arrows and the oversmoothed bandwidth by the dashed line. The UCV, BCV, and oversmoothed density estimates are represented by the dashed, solid, and dotted lines, respectively.

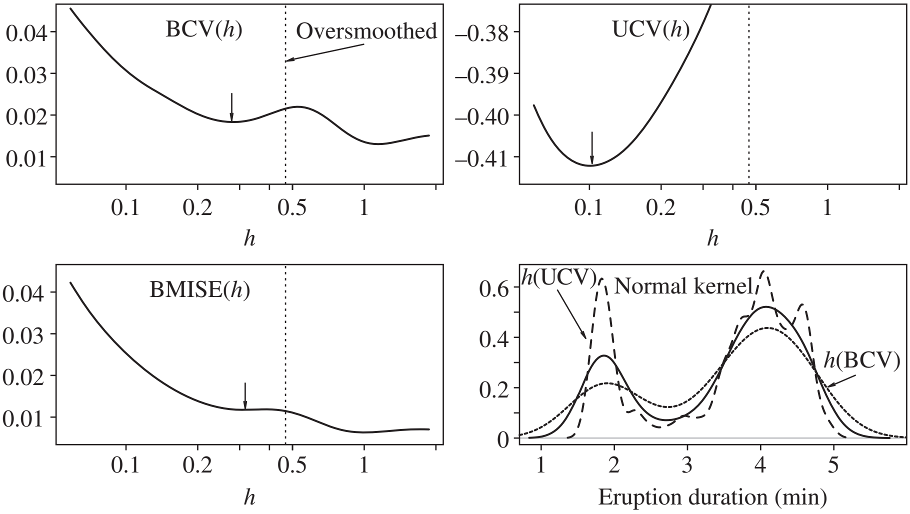

FIGURE 6.17 Normal kernel cross-validation algorithms and density estimates for the geyser dataset ($n = 107$). The CV bandwidths are indicated by arrows and the oversmoothed bandwidth by a solid line. The UCV, BCV, and oversmoothed density estimates are represented by the dashed, solid, and dotted lines, respectively.

with modest sample sizes. Concordance among a subset of these rather differently behaving criteria should be taken seriously.

6.5.1.5 Faster Rates and PI Cross-Validation In a series of papers, several authors have worked to improve the relatively slow $O(n^{-1/10})$ convergence of CV bandwidth algorithms. Along the way to the best $O(n^{-1/2})$ rate, algorithms were proposed with such interesting rates as $O(n^{-4/13})$. All share some features: for example, all have U-statistic formulas similar to UCV and BCV (Jones and Kappenman, 1992). All strive to improve estimates of quantities such as $R(f'')$ and $R(f''')$. The improvements come from expected sources—the use of higher order kernels, for example. Some examples include Chiu (1991) and Sheather and Jones (1991).

For purposes of illustration, the discussion is limited to one particular $O(n^{-1/2})$ PI algorithm due to Hall et al. (1991). They discovered that simply improving the estimation of $R(f'')$ was not sufficient; rather, a more accurate approximation for the AMISE is required (see Problem 6.10):

$$\text{AMISE}(h) = \frac{R(K)}{nh} - \frac{R(f)}{n} + \frac{1}{4}h^4\mu_2^2R(f'') - \frac{1}{24}h^6\mu_2\mu_4R(f'''),$$

which is accurate to $O(n^{-7/5})$. The second term, which is constant, may be ignored. The two unknown roughness functionals are estimated not as in BCV, but rather with two auxiliary smoothing parameters, λ_1 and λ_2, so that

$$\boxed{\widehat{\text{AMISE}}(h) = \frac{R(K)}{nh} + \frac{1}{4}h^4\mu_2^2\hat{R}_{\lambda_1}(f'') - \frac{1}{24}h^6\mu_2\mu_4\hat{R}_{\lambda_2}(f''').} \qquad (6.79)$$

Since there is no simple formula linking the three smoothing parameters in this formula, the authors propose selecting λ_1 and λ_2 based on a robust version of the normal reference rule. Several practical observations may be made. First, since the selection of the two auxiliary bandwidths is a onetime choice, the total computational effort is much less than for the BCV or UCV approaches. Second, this new approximation to the AMISE diverges to $-\infty$ as $h \to \infty$; hence, the PI rule is also looking for a local minimizer less than the oversmoothed bandwidth. However, given the simple form of Equation (6.79), it is relatively easy to provide a closed-form, asymptotic approximation to its (local) minimizer (see Problem 6.32):

$$\boxed{\hat{h}_{\text{PI}} = \left[\frac{\hat{J}_1}{n}\right]^{\frac{1}{5}} + \left[\frac{\hat{J}_1}{n}\right]^{\frac{3}{5}} \cdot \hat{J}_2; \quad \hat{J}_1 = \frac{R(K)}{\mu_2^2\hat{R}_{\lambda_1}(f'')}, \; \hat{J}_2 = \frac{\mu_4\hat{R}_{\lambda_2}(f''')}{\mu_2\hat{R}_{\lambda_1}(f'')}.} \qquad (6.80)$$

A portion of the details of a particular implementation given in the authors' paper is outlined in the next paragraph. The PI bandwidths computed in this fashion indeed have rapidly vanishing noise.

The estimates of $R(f'')$ and $R(f''')$ follow from the identities

$$\begin{aligned}\int f''(x)^2\,dx &= +\int f^{iv}(x)f(x)\,dx\\ &= +\mathrm{E}f^{iv}(X) \leftarrow \frac{1}{n(n-1)\lambda_1^5}\sum_{i=1}^{n}\sum_{j=1}^{n}K^{iv}\left(\frac{x_i-x_j}{\lambda_1}\right)\\ \int f'''(x)^2\,dx &= -\int f^{vi}(x)f(x)\,dx\\ &= -\mathrm{E}f^{vi}(X) \leftarrow \frac{1}{n(n-1)\lambda_2^7}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{vi}\left(\frac{x_i-x_j}{\lambda_2}\right).\end{aligned}$$

These two estimates are suggested by the UCV estimator. The authors chose a particular order-4 polynomial kernel for $K(x)$, which is supported on the interval $(-1,1)$, whose fourth derivative is equal to

$$K^{iv}(x) = \tfrac{135{,}135}{4{,}096}(1-x^2)(46{,}189x^8-80{,}036x^6+42{,}814x^4-7{,}236x^2+189).$$

Of course, for the normal kernel, $\phi^{vi}(x) = (x^6-15x^4+45x^2-15)\,\phi(x)$. Using normal reference rules and the interquartile range (IQR) as the measure of scale, the formulas for the auxiliary smoothing parameters are calculated to be

$$\hat{\lambda}_1 = 4.29 \text{ IQR}\, n^{-1/11} \quad \text{and} \quad \hat{\lambda}_2 = 0.91 \text{ IQR}\, n^{-1/9}.$$

To illustrate the performance of the PI bandwidth, 21 $N(0,1)$ simulations are displayed in Figure 6.18 for several sample sizes. The rapid convergence of the bandwidths to the h^* is apparent. For small samples, there may be no minimizer of the PI AMISE estimate given in Equation (6.79); examine some of the individual risk curves in Figure 6.18. The lack of a local minimum is a feature observed for small samples with the BCV criterion. It would appear that for small samples, the PI formula (6.80) is essentially returning a version of the normal reference rule rather than a true minimizer of the risk function. As the sample size increases, the “strength” of the PI estimate grows as the bowl shape of the AMISE estimate widens.

The PI AMISE function estimates for the Buffalo snowfall data ($n = 63$) and the Old Faithful eruption duration data ($n = 107$) are shown in Figure 6.19. Neither estimate has a local minimum, although the PI formula gives reasonable results. This illustrates the danger inherent in not plotting the risk function. It also suggests that the excellent small sample behavior of PI estimators involves subtle factors. In any case, the greater the number of reasonable bandwidth algorithms available to attack a dataset, the better.

For a very large dataset, such as the steel surface data, the PI, BCV, and UCV bandwidths are often identical. The PI risk curve is shown in Figure 6.19. The value of $\hat{h}_{\mathrm{PI}} = 0.424$ for the Gaussian kernel. The conversion factor to the triweight kernel ($\sigma_K = 1/3$) is 3 or $\hat{h} = 1.27$, which is identical to the UCV and BCV predictions in

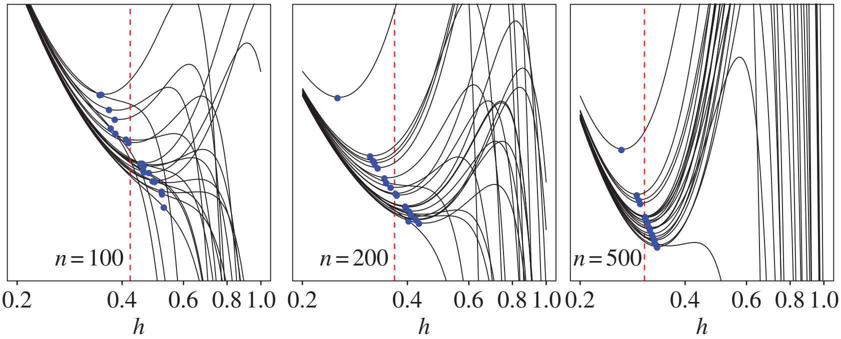

FIGURE 6.18 Twenty-one examples of the AMISE approximation of the plug-in rule with $N(0,1)$ data and a normal kernel. The PI bandwidth for each simulation is shown by the black dot on the risk curve. The vertical dotted line indicates the normal reference rule (with $\sigma = 1$). Note that the horizontal axis is the same for each sample size, but the vertical scale (not labeled) zooms in on the relevant area.

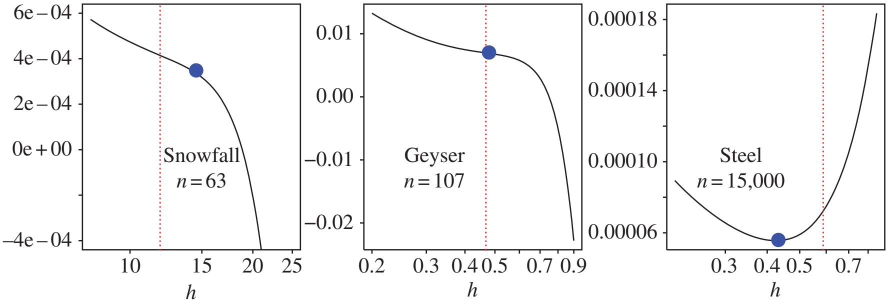

FIGURE 6.19 Plug-in cross-validation curves for the snowfall data ($n = 63$), the geyser dataset ($n = 107$) and the steel surface data ($n = 15{,}000$) for the normal kernel. The plug-in bandwidth obtained by formula (6.80) is indicated by the black dot, and the oversmoothed bandwidth by the dashed line.

Figure 6.15. The PI AMISE curve for the bimodal geyser dataset has no local minima. The PI bandwidth is $\hat{h}_{\mathrm{PI}} = 0.472$ (or $3 \times 0.472 = 1.42$ on the triweight scale). Thus the PI bandwidth matches the oversmoothed bandwidth but misses the local bandwidth found by the BCV and BMISE curves in Figure 6.17. For the smaller snowfall dataset, the PI bandwidth is greater than the oversmoothed bandwidth. Improved PI algorithms are under development, using generalized scale measures appropriate for multimodal data, and rapid changes and new successes can be expected. In any case, the experienced worker can expect to be able to judge the success or failure of any cross-validation bandwidth readily with modern interactive computing. Sheather (2004) indicates that the conservative choice of bandwidths for the auxiliary parameters tends to make the PI bandwidths oversmooth for complex densities.

6.5.1.6 Constrained Oversmoothing Oversmoothing has been presented strictly as a means of bounding the unknown smoothing parameters by choosing a measure of scale, such as the sample range or standard deviation, for the data. It is easy to find cases where the oversmoothed bandwidths are much too wide. If the sampling density is in fact the oversmoothed density, then the bandwidth $\hat{h}_{\mathrm{OS}}$ is not strictly an upper bound, as $\hat{h}_{\mathrm{OS}}$ varies about h^*. In most instances, the two bandwidths will be within a few percentage points of each other.

The variational problems considered in oversmoothing can be generalized to provide much more relevant bandwidths in almost every situation. The basic idea is to add constraints in addition to the one measure of overall scale. The new proposal for constrained oversmoothing (CO) is to require that several of the percentiles in the oversmoothed density match the sample percentiles in the data. Specifically,

$$f_{\mathrm{CO}}^* = \arg\min_f \int f''(x)^2\,dx \quad \text{s/t} \quad \int_{-\infty}^{\alpha_i} f(x)\,dx = F_n(\alpha_i), \quad i \in [1,k],$$

where $k \geq 2$. In other words, the cdf of the oversmoothed distribution should match the empirical cdf in several intermediate locations. This problem has already been solved in two instances when $k = 2$, once with the range constraint and again with the interquartile range constraint. The new suggestion is to choose to match the 10th, 30th, 50th, 70th, and 90th sample percentiles, for example. The resulting constrained oversmoothed density may or may not be close to the true density, but computing the roughness of the CO density provides a significantly improved *practical* estimate for use in the usual asymptotic bandwidth formula. With f_{CO}^* in hand, the constrained oversmoothed smoothing parameter is found as

$$\hat{h}_{\mathrm{CO}} = \left[\frac{R(K)}{n\sigma_K^4 R[(f_{\mathrm{CO}}^*)'']}\right]^{1/5}.$$

In Figure 6.20 for a sample of 1000 $N(0,1)$ points, several possible solutions to the variational problem are displayed along with the computed roughness. The location of the constraints is indicated by the dashed lines. The solution is a quartic spline. The relevant quantity is the fifth root of the ratio of that roughness to the true roughness, $R(\phi'') = 3/(8\sqrt{\pi}) = 0.2115$. The first CO solution is based on matching the 1, 50, and 99% sample percentiles. Now $(0.2115/0.085)^{1/5} = 1.20$, so that the constrained oversmoothed bandwidth is 1.2 times wider than h^*. Thus $h^* < h_{\mathrm{CO}}$ as usual. However, by including more constraints, this inequality will not be strictly observed. For example, two solutions were found with 11 constraints. One used the 1, 10, 20, …, 90, and 99% sample percentiles. Since the roughness of this solution is slightly *greater* than the true roughness, the CO bandwidth will be *smaller* than h^*. In fact, $h_{\mathrm{CO}} = 0.94\ h^*$. A second problem with 11 constraints used a fixed width mesh as shown. The sample percentiles matched in this mesh were 1.0, 3.4, 7.8, 16.6, 30.6, 49.9, 68.9, 82.7, 92.5, 97.3, and 99.0%. For this solution, $h_{\mathrm{CO}} = 0.82\ h^*$. Matching

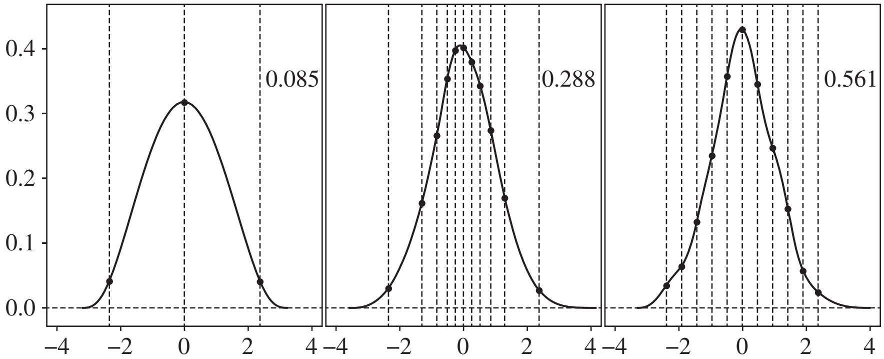

FIGURE 6.20 Constrained oversmoothed density solutions for a $N(0,1)$ sample of 1000 points. The true roughness is 0.212.

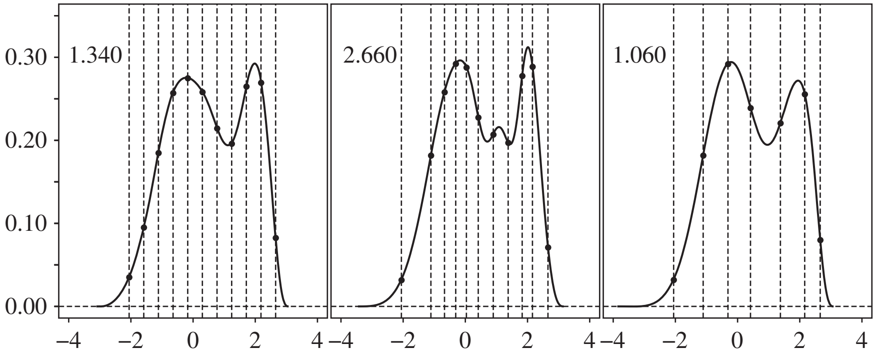

FIGURE 6.21 Constrained oversmoothed density solutions for a sample of 1000 points from the mixture density, $0.75N(0,1)+0.25N(0,1/9)$. The true roughness is 3.225.

5–10 percentiles would seem adequate in practice. For example, the percentile mesh with seven constraints (not shown) has a roughness of 0.336, which corresponds to $h_{\text{CO}} = 0.91\ h^*$.

This procedure was repeated for a sample of 1000 points from a mixture density, $\frac{3}{4}\phi(x|0,1)+\frac{1}{4}\phi(x|2,1/9)$ (see Figure 6.21). The true roughness is 3.225. Solutions for a fixed bin width and 2 percentile meshes are shown. For each solution $h^* < h_{\text{CO}}$ as the roughness of each solution is less than the true roughness. The difference is greatest with the solution in the last frame, for which $h_{\text{CO}} = 1.25\ h^*$. The ordinary oversmoothed rule, which is based on the variance 55/36, leads to a lower bound of 0.0499 for the roughness and $h_{\text{OS}} = 2.30\ h^*$ from Equation (6.68). Thus the constrained oversmoothed procedure gives a conservative bandwidth, but not *so* conservative. The solution in the middle frame indicates how the constrained oversmoothed density solution may contain features not in the true density (the small bump at $x=1$), but still be conservative ($h_{\text{CO}} = 1.04\ h^*$).

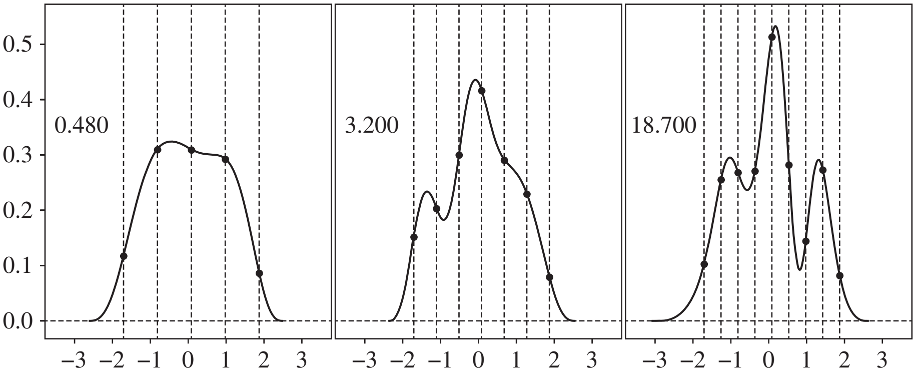

FIGURE 6.22 Constrained oversmoothed density solutions for Buffalo snowfall data.

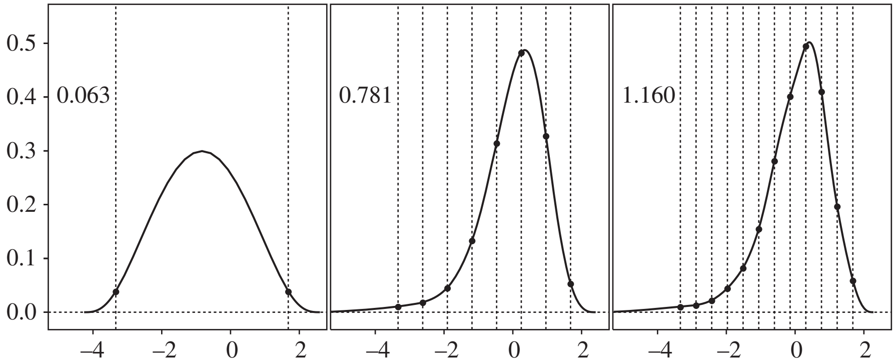

FIGURE 6.23 Constrained oversmoothed density solutions for steel surface data. The bin count data were jittered.

The application to real data is promising. (In all the examples, the original data were centered and standardized.) Several constrained oversmoothed solutions are shown in Figure 6.22 for the Buffalo snowfall data. In this case, all equally spaced meshes were selected. For small samples, this selection seems to work better than percentile meshes.

The final application is to a large dataset. Several constrained oversmoothed solutions are shown in Figure 6.23 for the steel surface data. Again, all equally spaced meshes were selected. Clearly, a roughness of about 0.8 is indicated—this leads to a bandwidth that is 60% of $(0.063/0.8)^{1/5}$, the usual oversmoothing shown in the first frame.

Clearly, the number and location of the constraints serve as surrogate smoothing parameters. However, if not too many are chosen, the solution should still serve as a more useful point of reference (upper bound). It is possible to imagine extending the problem to adaptive meshes.

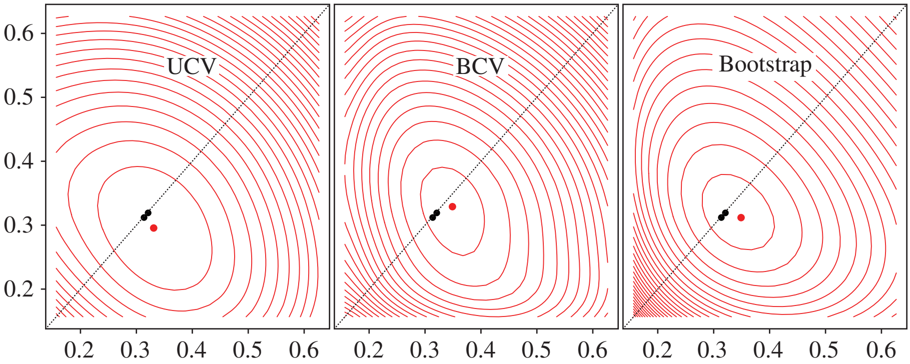

FIGURE 6.24 Estimated MISE(h_x, h_y) using UCV, BCV, and the bootstrap algorithms on 1500 $N(\mathbf{0}_2, I_2)$ points. The two dots on each diagonal are h^* and the oversmoothed bandwidths. The dot locating the minimizer of each criterion is below the diagonal.

6.5.2 Multivariate Data

The efficacy of univariate cross-validation algorithms is still under investigation, although the general options available are reasonably well understood. PI methods can be quite stable, while unbiased cross-validation is very general and easy to implement. Progress on multivariate generalizations has been made, although again the options are still under discussion. These are outlined below.

6.5.2.1 Multivariate Cross-Validation In principle, it is straightforward to extend each of the algorithms in the preceding sections to the multivariate setting. For example, the bootstrap algorithm is easily extended, as are the closed-form expressions for BCV and UCV (Sain et al., 1994). Examples of UCV, BCV, and BMISE$_*(h_1, h_2)$ based on 1500 $N(\mathbf{0}_2, I_2)$ points are shown in Figure 6.24. All are reasonably close in this case.

Figure 6.25 shows the same criteria applied to the standardized log-lipid dataset. However, the BMISE$_*(h_1, h_2)$ estimate for the lipid values ($n = 320$) has its minimum beyond the normal reference rule and the oversmoothed bandwidths. The UCV and BCV bandwidths are similar, although the BCV is (surprisingly) a bit more aggressive. Both show the two and possibly third mode/bump, whereas the Bootstrap estimate is unimodal. In general, it may be expected that the performance of all multivariate CV algorithms is more asymptotic and hence slower to converge in practice.

Retaining a smoothing parameter for each coordinate direction is important. Even if the data have been standardized, it is unlikely that $h_i = h_j$ will be satisfactory, (see Nezames (1980) and Wand and Jones (1991)). Less clear is whether the gain using a full smoothing matrix, which allows for elliptical contours, is sufficient to account for the dramatic increase in the number of parameters that must estimated by cross-validation. In many or most instances, the gains are not realized in practice. Furthermore, a full covariance may degenerate into a singular form, which can have “infinite” likelihood. Different starting values may avoid this problem.

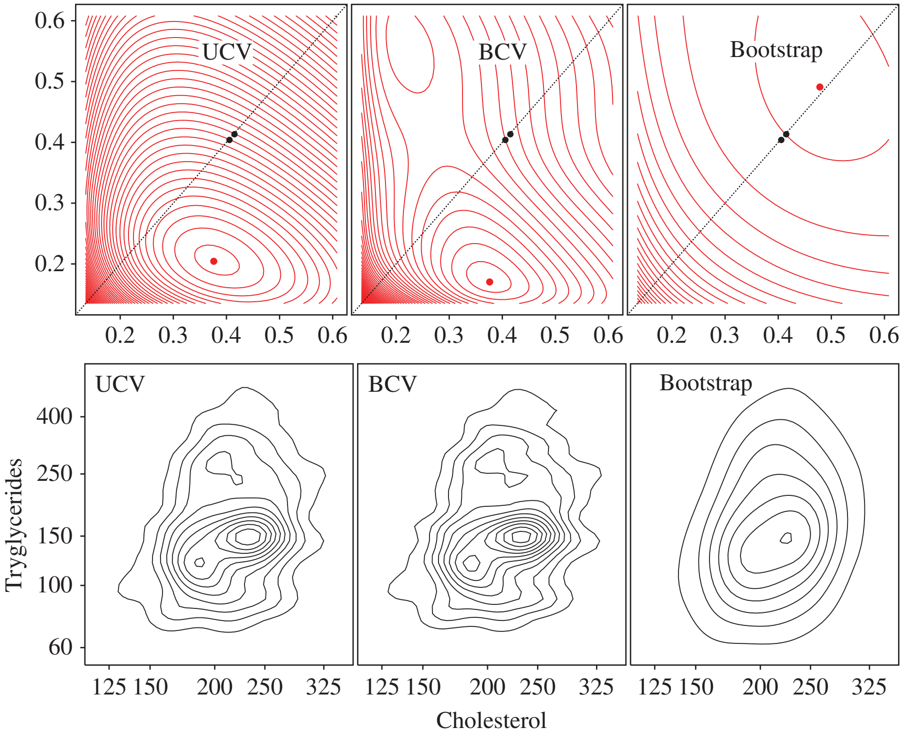

FIGURE 6.25 Same criterion as in Figure 6.24 for the standardized log lipid dataset ($n = 320$), together with the corresponding kernel estimates.

Density estimation is also quite difficult if the data cloud is highly skewed or falls (nearly) onto a lower dimensional manifold. This topic is taken up in detail in Chapter 7. However, marginal transformations using the Tukey ladder, for example, are always recommended. But as in the univariate setting, an absolutely optimal choice of bandwidth is not critical for exploratory purposes.

6.5.2.2 Multivariate Oversmoothing Bandwidths The extension of the oversmoothing bandwidth to $\Re^d$ has been solved by Terrell (1990). The easiest multivariate density is spherically symmetric. Thus the general kernel formulation is required with the constraint that all the marginal bandwidths are equal. By symmetry, finding the form of the multivariate oversmoothed density along the x-axis, for example, will be sufficient. The variational problem turns out to be identical to the one-dimensional problem in Equation (6.67), but in polar coordinates: $2f^{iv} + \lambda_1 + \lambda_2 r^2 = 0$, which implies that $f(r) = a(c^2 - r^2)^3$. The general form of the solution has been given in Theorem 3 of Terrell (1990). The result is that

$$h_{\mathrm{OS}} = \left[\frac{R(K)d}{nC_f}\right]^{1/(d+4)}, \quad \text{where} \quad C_f = \frac{16\Gamma\left(\frac{d+8}{2}\right)d(d+2)}{(d+8)^{(d+6)/2}\pi^{d/2}},$$

for kernels that have an identity covariance matrix. The constants in h_{OS} for $d \geq 1$ are 1.14, 1.08, 1.05, 1.04, 1.03, and so on. The constants decrease to 1.02 when $d = 8$ and slowly increase thereafter, with a limiting value of $\sqrt{e/2} = 1.166$. The constant finally grows to match 1.14 when $d = 348$. For $1 \leq d \leq 10$, the oversmoothed rule starts at 8.0% wider than the normal reference rule given in Equation (6.43) and increases to 10.5% wider when $d = 10$. As $d \to \infty$, the ratio continues grows to $\sqrt{e/2} = 1.166$, for a 16.6% increase. Using the easy-to-remember normal reference rule formula should be sufficient in most cases. Rescaling to other product kernels is easily accomplished by applying the rescaling rule dimension by dimension. The bivariate oversmoothed bandwidth is depicted in Figures 6.24 and 6.25.

6.5.2.3 Asymptotics of Multivariate Cross-Validation Given the very slow rates of convergence of UCV and BCV in one-dimension, and the slower rates of convergence of the AMISE as the dimension increases, it might reasonably be expected that a similar fate would transpire for multivariate cross-validation algorithms. Surprisingly, Sain et al. (1994) showed this was not the case. They found both the UCV and BCV bandwidths converged at the rates

$$n^{-d/(2d+8)} = \left(n^{-\frac{1}{10}}, n^{-\frac{1}{6}}, n^{-\frac{3}{14}}, n^{-\frac{1}{4}}\right) \quad \text{for} \quad 1 \leq d \leq 4. \tag{6.81}$$

Thus the cross-validation problem actually becomes easier as the dimension increases, and apparently the rate becomes $O(n^{-1/2})$ in the limit. Unfortunately, terms which could be ignored when $d < 4$ come into play, and dominate when $d > 4$.

This fact was discovered by Duong (2004) in a truly impressive thesis of wide scope. He re-analyzed the asymptotics of some half-dozen cross-validation algorithms, and introduced a new variation in the multivariate case of the smoothed cross-validation algorithm of Hall et al. (1992). These results were reported in Duong and Hazelton (2005a,b). They show that the formula in Equation (6.81) is correct when $1 \leq d \leq 4$, but that the correct rates for $d > 4$ for UCV and BCV are

$$n^{-2/(d+4)} = \left(n^{-\frac{2}{9}}, n^{-\frac{1}{5}}, n^{-\frac{2}{11}}, n^{-\frac{1}{6}}\right) \quad \text{for} \quad 5 \leq d \leq 8. \tag{6.82}$$

These rates, which are plotted in Figure 6.26, are decreasing and the cross-validation problem does become increasing difficult with dimension, in line with the general curse-of-dimensionality experienced by all nonparametric procedures. Fortunately, in the practical dimensions discussed in this book where $1 \leq d \leq 6$, there is good news. Also shown in Figure 6.26 are the rates for the PI bandwidth of Wand and Jones (1994). The faster rates occur for a diagonal smoothing parameter matrix, and PI-2 refers to the full smoothing matrix case. SCV-2 is described in Duong (2004). The formulae for these three criteria are $\min(8, d+4)/(2d+12)$, $4/(d+12)$, and $2/(d+6)$, respectively.

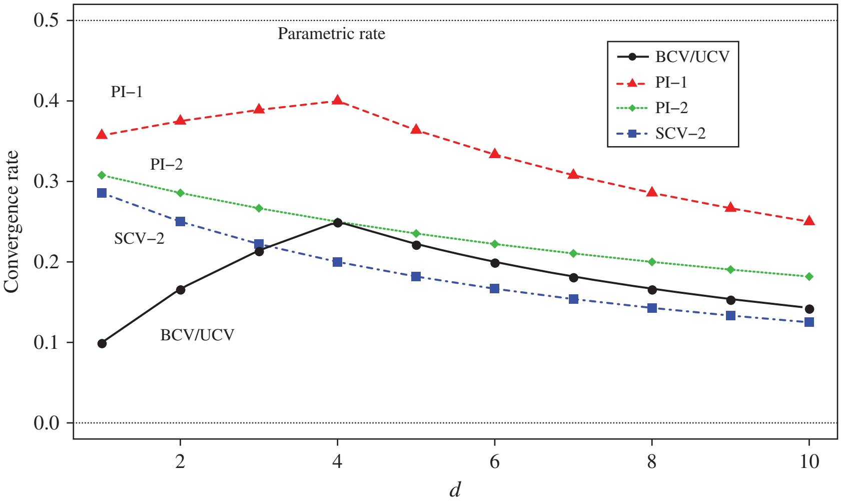

FIGURE 6.26 Magnitude of convergence rate exponents of several cross-validation algorithms. The best rate of $O(n^{-1/2})$ for parametric models would appear as $1/2$ on this graph.

In summary, the array of multivariate cross-validation algorithms available should suggest that all be tried and compared, as each has its failings in unpredictable ways. If all return similar bandwidths, so much the better. UCV remains the most flexible and general algorithm, and it is interesting to see it is quite competitive for $d > 2$. Its useful empirical performance in dimensions 1 and 2 has also been noted.

6.6 ADAPTIVE SMOOTHING

6.6.1 Variable Kernel Introduction

Consider the multivariate fixed kernel estimator

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}-\mathbf{x}_i).$$

The most general adaptive estimator within this simple framework allows the bandwidth h to vary not only with the point of estimation but also with the particular realization from the unknown density f:

$$h \leftarrow h(\mathbf{x}, \mathbf{x}_i, \{\mathbf{x}_j\}) \approx h(\mathbf{x}, \mathbf{x}_i, f).$$

The second form indicates that asymptotically, the portion of the adaptive bandwidth formula dependent upon the whole sample can be represented as a function of the true density. Furthermore, it may be assumed that the optimal adaptive bandwidth function, $h(\mathbf{x}, \mathbf{x}_i)$, is smooth and a slowly varying function. Thus, for finite samples, it will be sufficient to consider two distinct approaches toward adaptive estimation of the density function:

$$\boxed{\hat{f}_1(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{h_\mathbf{x}}(\mathbf{x}-\mathbf{x}_i) \quad \text{where} \quad h_\mathbf{x} \equiv h(\mathbf{x}, \mathbf{x}, f)} \tag{6.83}$$

or

$$\boxed{\hat{f}_2(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{h_i}(\mathbf{x}-\mathbf{x}_i) \quad \text{where} \quad h_i \equiv h(\mathbf{x}_i, \mathbf{x}_i, f).} \tag{6.84}$$

In the first case, a fixed bandwidth is used for all n data points, but that fixed bandwidth changes for each estimation point $\mathbf{x}$. In the second case, a different bandwidth is chosen for each $\mathbf{x}_i$, and then applied to estimate the density globally. Each is justified asymptotically by the local smoothness assumption on $h(\mathbf{x}, \mathbf{x}_i, f)$, since asymptotically only those data points in a small neighborhood of $\mathbf{x}$ contribute to the density value there. Presumably, all the optimal bandwidths in that neighborhood are sufficiently close so that using just one value is adequate. The choice of $\hat{f}_1$ or $\hat{f}_2$ depends on the practical difficulties in specifying the adaptive bandwidth function. For small samples, one may expect some difference in performance between the two estimators. Jones (1990) has given a useful graphical demonstration of the differences between the two adaptive estimators.

Examples of $\hat{f}_1$ include the k-NN estimator of Loftsgaarden and Quesenberry (1965) (see Section 6.4.1) with $h_\mathbf{x}$ equal to the distance to the kth nearest sample point:

$$h_\mathbf{x} = d_k(\mathbf{x}, \{\mathbf{x}_i\}) \approx \left(\frac{k}{nV_d f(\mathbf{x})}\right)^{1/d}, \tag{6.85}$$

where the stochastic distance is replaced by a simple histogram-like formula. The second form was introduced by Breiman et al. (1977), who suggested choosing

$$h_i = h \times d_k\big(\mathbf{x}_i, \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}\big) \approx h \times \left(\frac{k}{nV_d f(\mathbf{x})}\right)^{1/d}. \tag{6.86}$$

The similarity of these two particular proposals is evident. However, the focus on the use of the k-NN distance is simply a matter of convenience.

When estimated optimally point by point, the first form provides the asymptotically best possible estimate, at least from the MISE point of view. However, the estimator is not by construction a density function. For example, the k-NN estimator is easily seen to integrate to ∞ (see Problem 6.34). The second estimator, on the other hand, is by construction a *bona fide* density estimator for nonnegative kernels.

In practice, adaptive estimators in the very general forms given in Equations (6.83) and (6.84) can be very difficult to specify. The "adaptive smoothing function" $h_{\mathbf{x}}$ for $\hat{f}_1$ is ∞-dimensional. The specification for $\hat{f}_2$ is somewhat easier, since the "adaptive smoothing vector" $\{h_i\}$ is only n-dimensional. As usual, the "correct" choices of these quantities rely on further knowledge of unknown derivatives of the density function.

In practice, adaptive estimators are made feasible by significantly reducing the dimension of the adaptive smoothing function. One simple example is to incorporate the distance function $d_k(\cdot,\cdot)$ as in Equations (6.85) and (6.86). Abramson (1982a) proposed a variation on the Breiman et al.'s formula (6.86):

$$h_i = h/\sqrt{f(\mathbf{x}_i)} \quad \text{for } \mathbf{x}_i \in \Re^d .$$

In a companion paper, Abramson (1982b) proves that using a nonadaptive pilot estimate for f is adequate. Observe that these two proposals agree when $d = 2$, where Breiman et al. (1977) discovered empirically that their formula worked well.

In the univariate setting, the simple idea of applying a fixed kernel estimator to transformed data and then backtransforming falls into the second category as well. If u is a smooth monotone function selected from a transformation family such as Box and Cox (1964), then when the fixed kernel estimate of $w = u(x)$ is retransformed back to the original scale, the effect is to implicitly specify the value of h_i, at least asymptotically. The transformation approach has a demonstrated ability to handle skewed data well, and symmetric kurtotic data to a lesser extent (see Wand et al. (1991)). The transformation technique does not work as well with multimodal data.

In each case, the potential instability of the adaptive estimator has been significantly reduced by "stiffening" the smoothing function or vector. This may be seen explicitly by counting the number of smoothing parameters s that must be specified: $s = 1$ for k-NN (k); $s = 2$ for Breiman et al. (h, k); $s = 2$ for Abramson (h, h_{pilot}); and $s = 2$, 3, or 4 for the transformation approach (h plus 1–3 parameters defining the transformation family).

Theoretical and practical aspects of the adaptive problem are investigated further. There is much more known about the former than the latter.

6.6.2 Univariate Adaptive Smoothing

6.6.2.1 Bounds on Improvement Consider a pointwise adaptive estimator with a pth-order kernel:

$$\hat{f}_1(x) = \frac{1}{nh(x)} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h(x)}\right) \quad \text{letting } h(x) \equiv h_x .$$

The pointwise AMSE properties of this estimator can be recognized in Theorem 6.3 for a pth-order kernel:

$$\mathrm{AV}(x) = \frac{R(K)f(x)}{nh(x)} \quad \text{and} \quad \mathrm{ASB}(x) = \frac{\mu_p^2 f^{(p)}(x)^2}{(p!)^2} h(x)^{2p}$$

from which the optimal pointwise AMSE(x) may be obtained:

$$\begin{aligned} h^*(x) &= \left[\frac{(p!)^2 R(K) f(x)}{2p\,\mu_p^2 f^{(p)}(x)^2}\right]^{\frac{1}{2p+1}} n^{-\frac{1}{2p+1}} \\ \mathrm{AMSE}^*(x) &= \frac{2p+1}{2p}\left[\frac{2p\,\mu_p^2 R(K)^{2p} f(x)^{2p} f^{(p)}(x)^2}{(p!)^2}\right]^{\frac{1}{2p+1}} n^{-\frac{2p}{2p+1}}. \end{aligned} \tag{6.87}$$

Recall that the MISE accumulates pointwise errors. Thus accumulating the minimal pointwise errors obtained by using $h^*(x)$ gives the asymptotic lower bound to the adaptive AMISE:

$$\begin{aligned} \mathrm{AAMISE}^* &= \int_{-\infty}^{\infty} \mathrm{AMSE}^*(x)\,dx \\ &= \frac{2p+1}{2p}\left[\frac{2p\,\mu_p^2 R(K)^{2p}}{(p!)^2}\right]^{\frac{1}{2p+1}} \int \left[f(x)^{2p} f^{(p)}(x)^2\right]^{\frac{1}{2p+1}} dx \times n^{-\frac{2p}{2p+1}}. \end{aligned} \tag{6.88}$$

Comparing Equations (6.23) and (6.88), it follows that the bound on the improvement of an adaptive kernel estimator is

$$\frac{\mathrm{AAMISE}^*}{\mathrm{AMISE}^*} = \frac{\int \left[f(x)^{2p} f^{(p)}(x)^2\right]^{\frac{1}{2p+1}} dx}{\left[\int f^{(p)}(x)^2\,dx\right]^{\frac{1}{2p+1}}}. \tag{6.89}$$

An application of Jensen's inequality to the quantity

$$\mathrm{E}[f^{(p)}(X)^2/f(X)]^{1/(2p+1)}$$

shows that the ratio in (6.89) is always ≤ 1 (see Problem 6.35). In Table 6.4, this lower bound ratio is computed numerically for the normal and Cauchy densities. Observe that the adaptivity potential decreases for higher order kernels if the data are normal, but the opposite holds for Cauchy data. The table gives further evidence of the relative ease when estimating the normal density. Rosenblatt derived (6.88) in the positive kernel case $p = 2$.

TABLE 6.4 Ratio of AAMISE* to AMISE* for Two Common Densities as a Function of the Kernel Order

Kernel order	Density	
p	Normal(%)	Cauchy(%)
1	89.3	84.0
2	91.5	76.7
4	94.2	72.0
6	95.6	70.0
8	96.5	68.9

For the case of a positive kernel $p = 2$, the asymptotically optimal adaptive mesh is in fact equally spaced when

$$\frac{f''(x)^2}{f(x)} = c \quad \Rightarrow \quad f(x) = \frac{c}{144}(x-a)^4, \tag{6.90}$$

where a is an arbitrary constant (see Section 3.2.8.3). Thus the null space for kernel estimators occurs in intervals where the density is a pure quartic function. Piecing together pure segments of the form (6.90), while ensuring that f and f' are continuous implies that f is monotone; thus there does not exist an entire null adaptive density in C^1 or C^2 unless boundary kernels are introduced.

6.6.2.2 Nearest-Neighbor Estimators Using the asymptotic value for the adaptive smoothing parameter from Equation (6.85) with a positive kernel ($p = 2$) and $d = 1$,

$$h(x) = \frac{k}{2nf(x)}, \tag{6.91}$$

the adaptive asymptotic integrated squared bias is given by

$$\text{AAISB}(k) = \int_x \text{AISE}(x)dx = \int_x \frac{1}{4}h(x)^4 f''(x)^2 dx = \frac{k^4}{64n^4}\int_x \frac{f''(x)^2}{f(x)^4}dx.$$

Surprisingly, the latter integral is easily seen to diverge for such simple densities as the normal. This divergence does not imply that the bias is ∞ for finite samples, but it does indicate that no choice of k can be expected to provide a satisfactory variance/bias trade-off.

The explanation is quite simple: asymptotically, optimal adaptive smoothing depends not only on the density level as in Equation (6.91) but also on the curvature as in Equation (6.87). Thus the simple rule given by Equation (6.91) does not represent an improvement relative to nonadaptive estimation. This phenomenon of performing worse will be observed again and again with many simple ad hoc adaptive procedures, at least asymptotically. Some do provide significant gain for small samples or certain densities. Other approaches, such as transformation, include the fixed bandwidth setting as a special case, and hence need not perform significantly worse asymptotically.

6.6.2.3 Sample-Point Adaptive Estimators Consider the second form for an adaptive estimator with different smoothing parameters at the sample points:

$$\hat{f}_2(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_i)} K\left(\frac{x-x_i}{h(x_i)}\right) \quad \text{letting } h(x_i) \equiv h_i. \tag{6.92}$$

Terrell and Scott (1992, appendix) proved under certain conditions that

$$\mathrm{AV}\{\hat{f}_2(x)\} = f(x)R(K)/[nh(x)]$$

and

$$\mathrm{ASB}\{\hat{f}_2(x)\} = \{(p!)^{-1}[h(x)^p f(x)]^{(p)}\}^2,$$

where the superscript $^{(p)}$ indicates a pth-order derivative. Abramson (1982a) proposed what is obvious from this expression for the bias, namely, that when $p = 2$, the choice

$$h(x) = h/\sqrt{f(x)} \tag{6.93}$$

implies that the second-order bias term in ASB vanishes! The bias is actually

$$\frac{1}{4!}\left[h(x)^4 f(x)\right]^{(\mathrm{iv})} = \frac{1}{24}h^4\left[\frac{1}{f(x)}\right]^{(\mathrm{iv})},$$

as shown by Silverman (1986). This fourth-order bias is usually reserved for negative $p = 4$th order kernels and apparently contradicts Farrell's (1972) classical result about the best bias rates with positive kernels.

In fact, Terrell and Scott (1992) have provided a simple example that illustrates the actual behavior with normal data ($f = \phi$):

$$\hat{f}_2(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sqrt{\phi(x_i)}}{h} K\left(\frac{(x-x_i)\sqrt{\phi(x_i)}}{h}\right),$$

from which it follows that

$$\mathrm{E}\hat{f}_2(x) = \int \frac{\sqrt{\phi(t)}}{h} K\left(\frac{(x-t)\sqrt{\phi(t)}}{h}\right)\phi(t)\,dt, \tag{6.94}$$

with a similar expression for the variance. The exact adaptive MISE is difficult to obtain, but may be computed numerically for specific choices of h and n. The authors showed that the exact adaptive MISE was half that of the best fixed bandwidth estimator when $n < 200$; however, the fixed bandwidth estimator was superior for $n > 20{,}000$. This finding suggests the procedure does not have $O(n^{-8/9})$ MISE.

In fact, it is easy to demonstrate the source of the difference with Silverman (1986) and Hall and Marron (1988). Consider an asymptotic formula for the MSE of the estimator at $x = 0$, without loss of generality. The subtle point is made most clearly by choosing the boxcar kernel $K = U(-1,1)$ so that Equation (6.94) becomes

$$\mathrm{E}\hat{f}(0) = \frac{1}{2h}\int \phi(t)^{3/2}\left\{I_{[-1,1]}\left(\frac{t\sqrt{\phi(t)}}{h}\right)\right\}dt. \tag{6.95}$$

Usually, the limits of integration would extend from $-h$ to h. However, a closer examination shows that the argument of the kernel is not monotone increasing and, in fact, approaches zero as $|t| \to \infty$. Thus the integral in Equation (6.95) covers three intervals, call them $(-\infty,-b), (-a,a)$, and (b,∞), where a and b are solutions to the equation

$$\frac{t\sqrt{\phi(t)}}{h} = 1. \tag{6.96}$$

Define $c = (2\pi)^{1/4}h$. Then Equation (6.96) takes two forms that give sufficient approximations to the interval endpoints a and b:

$$\begin{aligned} t\,e^{-t^2/4} = c \quad &\Rightarrow \quad a \approx c + c^3/4 + 5c^5/32 \\ \log t - t^2/4 = \log c \quad &\Rightarrow \quad b \approx \left(-4\log c + 4\log\sqrt{-4\log c}\right)^{1/2}. \end{aligned}$$

Now taking a Taylor's series of the integrand in Equation (6.95) and integrating over $(-a,a)$ gives

$$\mathrm{E}\hat{f}(0) = \phi(0) + (2\pi)^{1/2}h^4/40 + O(h^6),$$

which gives the predicted $O(h^4)$ bias. However, the contribution toward the bias from the remaining two intervals totals

$$2\cdot\frac{1}{2h}\int_{-\infty}^{-b}\phi(x)^{3/2}dx = \left(\frac{2}{9\pi}\right)^{1/4}\frac{1}{h}\Phi\left(-b\sqrt{3/2}\right),$$

which, using the approximation for b and the tail approximation $\Phi(x) \approx -\phi(x)/x$ for $x << 0$, equals

$$h^2/\left(24[\log\{(2\pi)^{1/4}h\}]^2\right).$$

Thus the tails exert an undue influence on the estimate in the middle and destroy the apparent gain in bias. With a smoother kernel, the same effect is observed but not so clearly. Abramson recognized this practical problem and suggested putting an upper bound on h_i by "clipping" the pilot estimator in Equation (6.93) away from zero. Other authors have missed that suggestion. However, the asymptotic inefficiency does not negate the good small-sample properties observed by Abramson (1982a), Silverman (1986), and Worton (1989).

This same analysis may be applied to the original proposal of Breiman et al. (1977). The contribution from the tails to the bias turns out to be $O(h/\log h)$. These authors had noted that despite excellent empirical bivariate performance, the univariate performance was poor. This slow bias rate helps to explain that observation.

6.6.2.4 Data Sharpening The bias of a positive kernel estimate, $\hat{f}(x)$, is controlled by the second derivative there (see Equation (6.16)). Thus modes will generally be underestimated and antimodes will always be overestimated. (Even if higher-order negative kernels are employed, this same phenomena will be observed.) Recall that the gradient method minimizes a function by moving in the direction of the *negative* gradient. Samiuddin and El-Sayyad (1990) proposed adjusting the data towards the nearest peak by following the *positive* gradient. Specifically, in the univariate setting where $X_i \sim f(x)$, they propose using a kernel estimate on the adjusted data

$$\begin{aligned} \tilde{x}_i &= x_i + \frac{h^2}{2}\frac{f'(x_i)}{f(x_i)} \qquad \text{for } i = 1,\ldots,n & (6.97) \\ &= x_i - \frac{h^2}{2} \cdot \frac{x_i - \mu}{\sigma^2} \quad \text{for } N(\mu,\sigma^2) \text{ data}, & (6.98) \end{aligned}$$

where we assume the kernel is scaled so that $\sigma_K = 1$. Note that for a normal density, Equation (6.98) may be rewritten as $(\tilde{x}_i - \mu) = (x_i - \mu)\left[1 - h^2/(2\sigma^2)\right]$; therefore, the data are shrunk toward the mean (and mode) by the factor $\frac{1}{2}(2 - h^2/\sigma^2)$. Versions of Equation (6.97) that are data-based and also of a more general form have been considered by Choi and Hall (1999) and Hall and Minnotte (2002).

In order to investigate the effectiveness of this proposal, we focus on the expectation of the kernel estimator using the modified data:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - \tilde{x}_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i - \frac{1}{2}h^2 f'(x_i)/f(x_i)}{h}\right), \tag{6.99}$$

where the true density is used in the adjustment (6.97). Assume that the kernel K is symmetric, $K(-w) = K(w)$, and that $\sigma_K = 1$. Using the change of variables $w = (x - y)/h$ and a Taylor's Series in the kernel, the expectation is given by

$$
\begin{aligned}
\mathrm{E}\hat{f}(x) &= \int \frac{1}{h} K\left(\frac{x-y-\frac{1}{2}h^2 f'(y)/f(y)}{h}\right) f(y)\,dy \\
&= \int K\left(w - \frac{h}{2}\frac{f'(x-hw)}{f(x-hw)}\right) f(x-hw)\,dy \\
&= \int f(x-hw)\left[K(w) - \frac{h}{2}\frac{f'(x-hw)}{f(x-hw)}K'(w) + \frac{h^2}{8}\frac{f'(x-hw)^2}{f(x-hw)^2}K''(w)\right] dw \\
&= \int K(w) f(x-hw) - \frac{h}{2}\int K'(w) f'(x-hw) + \frac{h^2}{8}\int K''(w)\frac{f'(x-hw)^2}{f(x-hw)} \\
&= \left[f(x) + \frac{1}{2}h^2 f''(x) + O(h^4)\right] - \frac{h}{2}\left[h f''(x) + O(h^3)\right] + \frac{h^2}{8}\left[O(h^2)\right] \\
&= f(x) + \frac{1}{2}h^2 f''(x) - \frac{1}{2}h^2 f''(x) + O(h^4) = f(x) + O(h^4).
\end{aligned}
$$

The result is that the bias of the Samiuddin and El-Sayyad proposal is of higher order $O(h^4)$, even with a positive kernel. In the third to last expression, the first integral is the familiar bias expansion. The second integral may be approximated by applying a Taylor's Series to $f'(x-hw)$:

$$
\int K'(w)\left[\sum_{\ell=0}(-hw)^{\ell} f^{(1+\ell)}(x)/\ell!\right] dw = 0 - hf''(x) + 0 + O(h^3),
$$

since $\int K' = 0$, $\int wK' = -1$, $\int w^2 K' = 0$, and $\int w^3 K' = -3\sigma_K^2$. There is little to be gained by deriving the explicit bias expression, which is quite complicated (see Hall and Minnotte (2002)). The leading variance terms turn out to be the same as for the ordinary kernel estimator. While this derivation adjusts the data using $f(x)$, which is unknown, the data-based algorithm explicated by Hall and Minnotte (2002) achieves any higher order desired.

As we have seen so many times before, the search for an algorithm that reduces bias by local adapting or adjustment results in a higher order bias result instead. Hall and Minnotte (2002) and others report promising examples with real data and some simulations. However, the application of the data adjustment (6.97) globally has an obvious deficiency, as can be seen by observing its action on the mixture density in Equation (3.78). From Equation (6.98), for standard normal data, the adjustment is $-h^2 x_i/2$, while for $N(3,1/3^2)$ data, the adjustment is $-9h^2(x_i-3)/2$, which is actually 27 times greater in standard units. Intuitively, the adjustment for the narrower mixture data should be one-third as wide since the standard deviation is 1/3. One possible fix would be to incorporate the magnitude of the second derivative at the local mode. Examining the published case studies, many had the same curvature at the modes and the effect was missed. Nevertheless, this idea holds good promise with a multistage estimation implementation to find the location of the modes and antimodes and local scale. It may also be advisable to use different smoothing parameters in Equations (6.97) and (6.99) and determine both smoothing parameters by unbiased cross-validation. In general, each modal region may require its own smoothing parameter as well.

6.6.3 Multivariate Adaptive Procedures

Multivariate adaptive procedures contain some interesting and unique features. These results are available in more detail in Terrell and Scott (1992).

6.6.3.1 Pointwise Adapting Let $\nabla^2 f(\mathbf{x})$, which is the matrix of second partial derivatives of f at $\mathbf{x}$, be denoted by $S_{\mathbf{x}}$. From Equation (6.48), the pointwise asymptotic bias is

$$\mathrm{AB}(\mathbf{x}) = \frac{1}{2}h^2 \mathrm{tr}\{A^T S_{\mathbf{x}} A\} = \frac{1}{2}h^2 \mathrm{tr}\{AA^T S_{\mathbf{x}}\}. \tag{6.100}$$

In the univariate setting, the bias is controlled entirely by the scale of the kernel, while in the multivariate setting, the shape of the kernel is also available to control the bias. The importance of the shape in minimizing $\mathrm{tr}\{A^T S_{\mathbf{x}} A\}$ depends on the properties of the matrix $S_{\mathbf{x}}$.

Case I $S_{\mathbf{x}}$ is positive or negative definite. As H (and hence A) is full rank by assumption, then $A^T S_{\mathbf{x}} A$ is also positive or negative definite. Because the sum and product of the eigenvalues of a definite matrix equal its trace and determinant, respectively, the matrix with minimum (absolute) trace and determinant equal to 1 has all of its eigenvalues equal. Thus the matrix A should be chosen to satisfy

$$AA^T S_{\mathbf{x}} = |S_{\mathbf{x}}|^{1/d} I_d.$$

Observe that with this choice, the matrix on the right-hand side has the same determinant as $S_{\mathbf{x}}$, and all the eigenvalues of $AA^T S_{\mathbf{x}}$ equal $|S_{\mathbf{x}}|^{1/d}$. Therefore, the best $\mathrm{tr}\{A^T S_{\mathbf{x}} A\} = d|S_{\mathbf{x}}|^{1/d}$. The pointwise asymptotic MSE of $\hat{f}(\mathbf{x})$ follows from Equations (6.49) and (6.100) and may be optimized to yield

$$h^*(\mathbf{x}) = \left[\frac{f(\mathbf{x})R(K)}{nd\,|S_{\mathbf{x}}|^{2/d}}\right]^{1/(d+4)}$$

so that

$$\mathrm{AMSE}^*(\mathbf{x}) = \left(\frac{d+4}{4d}\right)^{\frac{2(d+2)}{d+4}} \left[f(\mathbf{x})R(K)\sqrt{|S_{\mathbf{x}}|}\right]^{4/(d+4)} n^{-4/(d+4)}.$$

Case II The density is *saddle-shaped* at $\mathbf{x}$; that is, the matrix $S_{\mathbf{x}}$ has both positive and negative eigenvalues. The density is curved upward in some directions and downward in others. In this case, it is possible to construct the matrix A so that sum of the eigenvalues of $AA^T S_{\mathbf{x}}$ equals 0 (see Terrell and Scott (1992)). Thus the order h^2 bias terms vanish with higher order terms dominating. Therefore, regions where the density is saddle-shaped asymptotically contribute nothing to the AAMISE compared to regions where the density is definite.

How common are saddle-shaped regions? Consider the multivariate normal density $N(\mathbf{0}_d, I_d)$. Then the gradient and Hessian of $f(\mathbf{x})$ are

$$\nabla f(\mathbf{x}) = -f(\mathbf{x})\mathbf{x} \quad \text{and} \quad \nabla^2 f(\mathbf{x}) = S_{\mathbf{x}} = f(\mathbf{x})(\mathbf{x}\mathbf{x}^T - I_d).$$

There are $d-1$ eigenvalues of the matrix $(\mathbf{x}\mathbf{x}^T - I_d)$ equal to -1 and one equal to $\mathbf{x}^T\mathbf{x} - 1$. Therefore, the multivariate normal density is negative definite when $\|\mathbf{x}\| < 1$ and saddle-shaped *everywhere* outside the unit sphere. Given that the fraction of probability mass inside the unit sphere decreases as the dimension increases, the potential practical significance of this finding seems promising. Finally, observe that exactly on the unit sphere where $\|\mathbf{x}\| = 1$, one of the eigenvalues vanishes. This leads to the final case.

Case III $S_{\mathbf{x}}$ is semidefinite with at least one zero eigenvalue. The density is flat in certain directions. There is no contribution to the bias in those directions and hence the bias contribution is again of lower order than in Case I. The problem at that point can be reduced to a lower dimension by projection if desired.

6.6.3.2 Global Adapting The problem of selecting the best global adaptive fixed kernel estimator was formulated by Deheuvels (1977a, b), who characterized the solution in differential equation form. The global criterion to be optimized is

$$\text{AAMISE}(h, A) = \frac{R(K)}{nh^d} + \frac{1}{4}h^4 \int_{R^d} \text{tr}\{A^T S_{\mathbf{x}} A\}\, d\mathbf{x}.$$

Minimizing over h and A separately gives

$$\text{AAMISE}^* = \left[\min_A \int_{R^d} \text{tr}^2\{A^T S_{\mathbf{x}} A\} d\mathbf{x}\right]^{d/(d+4)} \left[\frac{(d+4)R(K)}{4nd}\right]^{4/(d+4)}.$$

For $N(\mathbf{0}_d, I_d)$ data, the advantage of the adaptive scheme compared to the fixed kernel is great (see Table 6.5 (from Terrell and Scott, 1992)). The minimization was done numerically by Terrell. The advantage comes from the large saddle-shaped portion of the density.

The final global comparison is between the fixed kernel estimator and the k-NN multivariate density estimator. The latter estimator will be treated as an adaptive

TABLE 6.5 Relative Efficiency of Transformed Adaptive Density Estimate Compared to Fixed Kernel for the Multivariate Normal Density

Dimension d	1	2	3	4	5	6
Adapt/fixed efficiency	1	45.5%	30.2%	18.6%	10.6%	5.7%

TABLE 6.6 Relative Efficiency of k-NN Estimator to Fixed Kernel for the Multivariate Normal Density

Dimension d	1	2	3	4	5	15	100
k-NN/fixed efficiency	0	0	0.48	0.87	1.15	1.55	1.49

kernel estimator, with the kernel being a uniform density over the unit sphere and the bandwidth being taken as the asymptotic form

$$h \leftarrow [k/(nf(\mathbf{x})V_d)]^{1/d}.$$

The AMISE of the fixed kernel estimator using this same kernel follows from the general kernel results earlier (see Terrell and Scott, 1992). It is shown that

$$\frac{\text{AMISE}_h^*}{\text{AAMISE}_k^*} = \left(\frac{d-2}{d}\right)^{\frac{d(d+2)}{2(d+4)}} \left(\frac{4(d^2-4)}{d^2-6d+16}\right)^{\frac{d}{d+4}} \to \frac{4}{e}$$

as $d \to \infty$ (see Table 6.6). Thus the nearest-neighbor estimator, which overadapts to the tails in the univariate and bivariate cases, is seen to perform better than the fixed kernel estimator when $d \geq 5$, at least for normal data. This superiority is reassuring since the algorithm has a proven track record in high-dimensional applications such as clustering.

6.6.4 Practical Adaptive Algorithms

We conclude this topic with a discussion of two algorithms that use the unbiasedness property of UCV in a fundamental manner. PI approaches cannot compete with UCV because of the unwieldly asymptotics encountered.

6.6.4.1 Zero-Bias Bandwidths for Tail Estimation One common truism about nonparametric density estimates is that their quality degrades in the tails where data are sparse. The estimates are generally "lumpy" around whatever data points may occur. However, there is an interesting phenomena for fixed bandwidth estimators that can provide remarkably accurate and useful estimates in the tails. The visual presentation can be much improved as well.

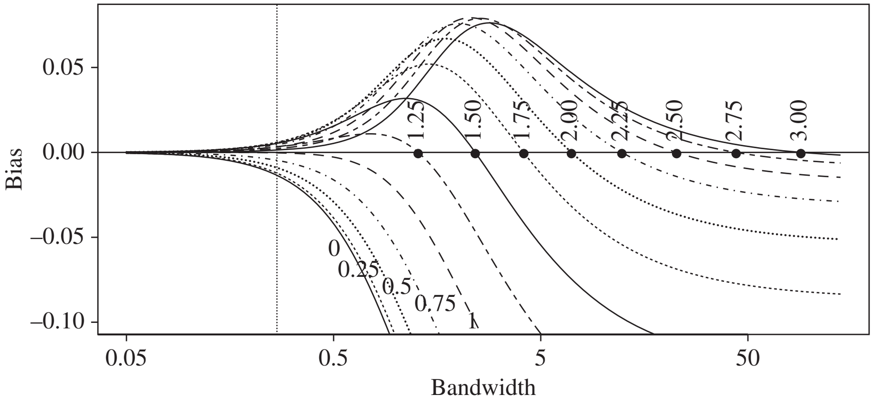

FIGURE 6.27 Bias of normal kernel density estimator for a standard normal sample of size $n = 1000$ as function of bandwidth.

Let us focus on the general behavior of the bias of a positive kernel estimator as the bandwidth h ranges from 0 to ∞. Recall the introduction to data-sharpening in Section 6.6.2.4, where it was observed that peaks are underestimated and valleys are overestimated. This is the direct result of the fact that the bias of a kernel estimate is proportional to $h^2 f''(x)$. In Figure 6.27, the exact bias (Fryer, 1976) of a normal kernel estimator for a normal sample is essentially 0 for small bandwidths, then becomes negative or positive depending whether $0 < x < 1$ or $x > 1$, respectively. (Since the $N(0,1)$ density is symmetric, the bias curves for negative values of x are identical; the density is concave for $-1 < x < 1$ and convex for $|x| > 1$.) At the other extreme, as $h \to \infty$ the density estimate $\hat{f}(x) \to 0$ for all x; hence, the bias$(x) \to -f(x)$, which is negative everywhere for a $N(0,1)$ density (see Figure 6.27).

Recall from Equation (6.13) that the expected value of a kernel estimate does not directly depend on the sample size n, except possibly through h. Since the expectation is a continuous function of x, we may conclude that when $x > 1$, there must be a fixed a fixed bandwidth h_x where $\mathrm{E}\hat{f}_{h_x}(x) = f(x)$ exactly. This bandwidth is highlighted in Figure 6.27. We shall refer to this bandwidth as the *zero-bias bandwidth* and denote it by $h_0(x)$; see Sain (1994, 2003), Sain and Scott (2002), and also Hazelton (1998) who independently investigated what he called the bias-annihilating bandwidth.

The existence of the zero-bias bandwidth can result in more than one local minima in the pointwise MSE(x) and more than one "optimal" bandwidth (see Figure 6.7). For n sufficiently large, there will be a local minimum in the MSE(x) near $h_0(x)$, which we will denote by $h_0^*(x)$. The smaller optimal bandwidth corresponds to the usual asymptotic theory, that is, $h_x^* \propto n^{-1/5} \to 0$ as $n \to \infty$. However, if we choose to use the zero-bias bandwidth $h_0(x)$, then the squared bias is 0 and the MSE(x) is simply the variance $R(K)f(x)/(nh_0(x))$ to good approximation (see Equations (6.11)) and (6.15). This parametric rate $O(n^{-1})$ is not achievable in practice. Both Hazelton (1998) and Sain (2003) show the rate is actually the familiar $O(n^{-4/5})$ with real data.

In Figure 6.28, the values of the bandwidths that (locally) minimize the MSE(x) are shown for the standard normal density and for the two-component mixture density

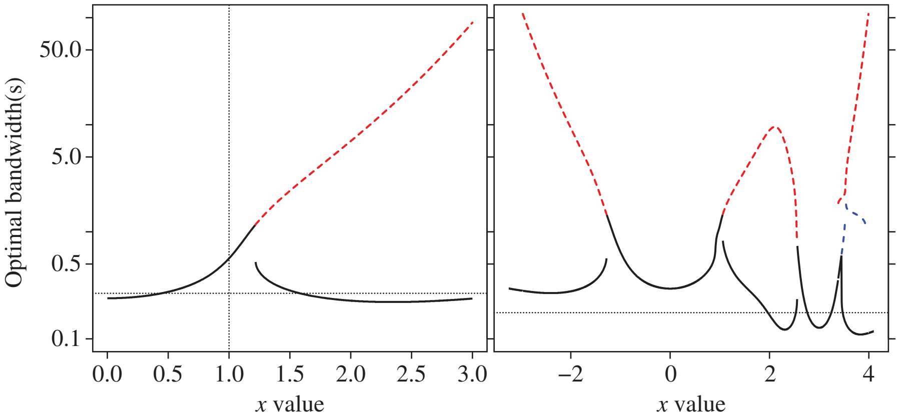

FIGURE 6.28 (Left) Asymptotically optimal bandwidths and zero-bias bandwidths for normal sample of size $n = 1000$. There are two optimal bandwidths when $x > 1.218$. The dashed line shows bandwidths close to zero-bias ones. (Right) Same for normal mixture but $n = 500$.

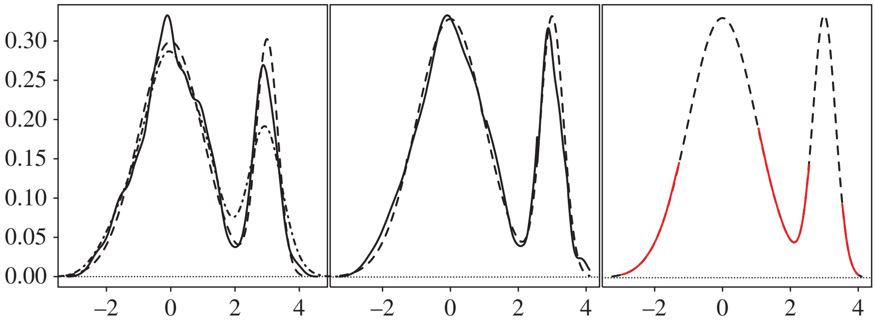

FIGURE 6.29 Several density estimates (solid lines) of a sample of size $n = 500$ from the mixture density in Equation (3.78), which is shown as a dashed line in each frame. (Left) A fixed kernel estimate with $h^* = 0.176$ and the normal reference rule $h = .473$ (dotted line). (Middle) Kernel estimate using $h^*(x)$, which is the smallest bandwidth in the right frame of Figure 6.28. (Right) Kernel estimate using $h_0^*(x)$ where it exists.

in Equation (3.78). In the right frame, there is even a region near $x \approx 3.7$ where there are three such bandwidths.

In Figure 6.29, we compare the various available kernel estimates for a sample of size $n = 500$ from the two-component mixture density. The left frame uses the asymptotically optimal fixed bandwidth $h^* = 0.176$, which undersmooths on the left and oversmooths on the right. The middle frame shows the use of $h^*(x)$ pointwise. This is much improved over the fixed kernel estimate, except for some minor anomolies near the inflection points. Finally, in the right frame, we show the zero-bias estimates where they exist. To the naked eye, there is no error in this estimate (even with a sample of only 500 points). Of course, without knowledge of the exact zero-bias bandwidths, such a result should not be expected. Note that for the fixed bandwidth estimator, the normal reference rule, UCV, BCV, and Sheather-Jones bandwidths are 0.473, 0.130, 0.222, and 0.218, respectively. The $h = 0.130$ UCV bandwidth (not shown) is well-calibrated for the narrow right component, leaving the left component quite undersmoothed.

Finally, to implement a data-based procedure for zero-bias estimation, the UCV formulation in Equation (3.52) can simply be modified by including an indicator function for data in an assumed known interval (a,b),

$$\int_{-\infty}^{\infty} \left[\hat{f}(x) - g(x)\right]^2 I(x \in (a,b)) = \int_a^b \hat{f}(x)^2 - 2\,\mathrm{E}\left[\hat{f}(X) I(X \in (a,b))\right] + c,$$

leading to

$$\mathrm{UCV}_{(a,b)}(h) = \int_a^b \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i)\, I(x_i \in (a,b)). \qquad (6.101)$$

The simplest implementation is to assume that the function $h_0(x)$ can be approximated by a constant over the interval (a,b). Obviously, other functional forms can be chosen, from polynomials to splines.

Returning to the mixture example, consider the interval $(a,b) = (1.35, 2.51)$ in the convex region around the anti-mode. The left frame in Figure 6.30 displays the data-based $\mathrm{UCV}(h)$ function in Equation (6.101) assuming a constant bandwidth over (a,b). Two local minima are observed, as hoped for. The larger bandwidth, $\hat{h}_0^* = 5.71$, is in the zero-bias region, and might be used at the midpoint of (a,b) rather than over the entire interval. Note, however, that the predicted error is not smaller than that at the asymptotically optimal bandwidth, $\hat{h}^* = 0.149$. The middle frame displays the $\mathrm{UCV}(h)$ bandwidth fits, where the logarithm of bandwidth h is taken to be a polynomial of order 0, 1, or 2 on (a,b). Depending on the starting values, there are

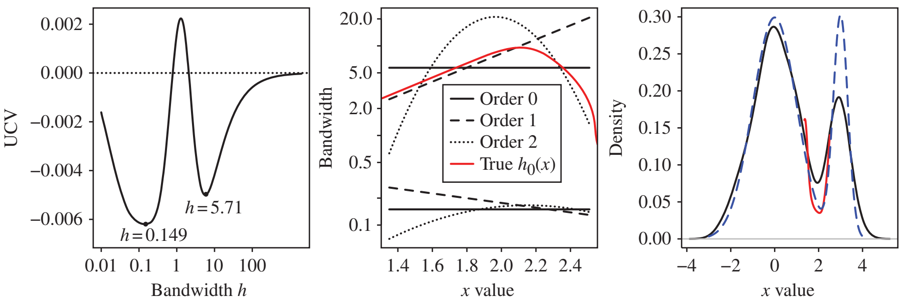

FIGURE 6.30 (Left) The UCV function over the interval $(a,b) = (1.35, 2.51)$ for the normal mixture dataset. (Middle) Log-polynomial fits to $h_0(x)$ that minimize UCV over (a,b). (Right) The zero-bias estimate over (a,b) together with the true mixture density and normal reference rule kernel estimate.

two solutions, corresponding to the two bandwidth types. For reference, the exact zero-bias $\text{MSE}(x)$ minimizing bandwidth, $h_0^*(x)$, is shown as the thick solid curved line. Notice that none of the polynomial fits has sufficient degrees of freedom to approximate that curve. The order-1 (log-linear) fit is good for $x < 2.1$, and the order-2 (log-quadratic) fit good for $x > 2.3$. The right frame shows the zero-bias density estimate using quadratic fit, which has the smallest UCV score (20% lower than a constant bandwidth). The overestimation of the fixed kernel estimate over (a,b) is corrected, and the zero-bias estimate is almost perfect for $x > 2$. (The log-linear fit, not shown, is also almost perfect for $x < 2$.) A spline model for $h_0^*(x)$ is indicated.

We conclude by demonstrating a simple approximation for the zero-bias bandwidth due to Sain (2003). Since the bandwidth h is very large, we take a Taylor series of the *kernel* about 0 in Equation (6.12) (rather than the change of variables) and obtain

$$\begin{aligned}
\text{E}\hat{f}(x) &= \int \frac{1}{h}\left[K(0) + \left(\frac{x-t}{h}\right)K'(0) + \left(\frac{x-t}{h}\right)^2 K''(0) + \cdots\right] f(t)\,dt \\
&= \frac{K(0)}{h} + 0 + \frac{K''(0)}{2h^3}\int (x-\mu_f+\mu_f-t)^2 f(t)\,dt + \cdots \\
&= \frac{K(0)}{h} + \frac{K''(0)}{2h^3}\left[(x-\mu_f)^2 + \sigma_f^2\right] + \cdots,
\end{aligned}$$

since $K'(0) = 0$ for symmetric kernels. If $\text{E}\hat{f}(x) = f(x)$, that is, is unbiased, then $f(x) \approx K(0)/h_0(x)$ or $h_0(x) \approx K(0)/f(x)$, which is quite accurate in the tails. Observe that finding the zero-bias bandwidth function is related to estimation of the inverse density.

To summarize, zero-bias bandwidths hold interesting potential for estimating the density nonparametrically in areas typically deemed too difficult for quality estimation. The unbiasedness property of UCV permits investigation of these unusually large bandwidths.

6.6.4.2 UCV for Adaptive Estimators Consider the sample-point estimator in Equation (6.92), and its obvious multivariate extension. With n parameters, there are too many parameters for practical cross-validation. There are a number of strategies for reducing the number of smoothing parameters, such as grouping the data into bins, in which the data share the same smoothing parameter, or rounding the data to bin centers. Sain (2002) and Duong and Hazelton (2005a) consider such options as well as full smoothing matrix parametrizations H in the bivariate case.

For the purpose of illustration, we group the data not by binning but by using the k-means algorithm (MacQueen, 1967), and allow each cluster to have its own smoothing parameter. The usual UCV data-based criterion is optimized numerically over the vector of smoothing parameters, for example, using R function *nlminb*. Figure 6.31 displays the fixed and adaptive kernel estimates. Seven clusters were selected using

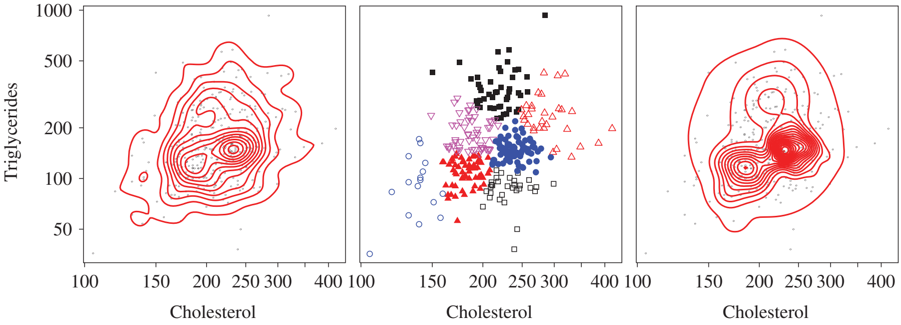

FIGURE 6.31 (Left) Twelve contours of the UCV-calibrated ($\hat{h} = 0.276$) bivariate Gaussian fixed-kernel estimate of the standardized log cholesterol and triglyceride data. (Middle) Seven clusters from k-means. (Right) The adaptive kernel estimator. The seven bandwidths range from 0.174 to 2.36. The mode is 54% greater than in the left frame. The 19 contour levels are the same as in the left frame plus seven more at higher levels.

k-means and are shown in the middle frame. To reduce the number of parameters, the data were standardized and a spherical gaussian kernel was selected that has only one smoothing parameter (instead of 2). The fixed bandwidth is 0.174 while the adaptive bandwidths are 0.17, 0.25, 0.47, 1.02, 1.48, 1.50, and 2.36. The adaptive estimate enhances the two primary modes as well as the smaller third bump, and reduces noise in the tails.

However, other choices of k lead to quite different estimates, some with fewer, some with more modes. Artifacts are also a potential problem with overparameterized estimates. This one was selected as more faithful to the basic structure in the fixed kernel estimator. When experimenting with a fully parameterized elliptical kernel, Sain (2002) and Duong and Hazelton (2005a) produced estimates that could have even more extreme anomolies. Users of normal mixture models and the expectation-maximization (EM) algorithm are familiar with the potential for singularities when choosing a fully parameterized covariance matrix (when the fitted covariance matrix is nearly singular). UCV is also susceptible to such singularities. A well-written mixture package, Mclust, was applied to these data (see Fraley and Raftery (2002)). The BIC criterion evaluated mixtures with k in the range of 1–10; $k = 1$ was indicated by BIC. The single bivariate normal fit misses the bimodal feature altogether.

6.7 ASPECTS OF COMPUTATION

We conclude our survey of kernel methods by revisiting several options for computing a kernel estimator on an equally spaced mesh $\{a = t_0, t_1, \ldots, t_m = b\}$ based on a sample of size n. How can this be made more efficient, as alternatives to the histogram or ASH?

6.7.1 Finite Kernel Support and Rounding of Data

The kernel estimate is to be evaluated at the m bin midpoints, $\{m_1, m_2, \ldots m_m\}$. If the kernel, K, has infinite support, then the direct approach requires $n \cdot m$ kernel evaluations, n kernel evaluations for each of m estimation points. However, if a kernel with finite support is chosen, then only $f \cdot m$ kernel evaluations are required for each data point, where $f = 2h/(b-a)$, assuming the support of the kernel is $(-1,1)$, w.l.o.g. The calculations should be reversed so that the outer loop is over the data points, rather the inner loop, as this pseudo R code shows:

```
$\delta = (b-a)/m;$   $m_k = \mathrm{seq}(a+\delta/2, b-\delta/2, \delta);$   $y = \mathrm{rep}(0, m)$
for ($i$ in $1\!:\!n$) { $k_0 = 2 + \lfloor (x_i - h - m_1)/\delta \rfloor$ ; $k_1 = 1 + \lfloor (x_i + h - m_1)/\delta \rfloor$
    for ($k$ in $\max(1, k_0) : \min(m, k_1)$) { $y_k = y_k + K_h(x_i - m_k)$ } };$y = y/n$.
```

However, the work is still proportional to the sample size n. Prebinning or rounding the data to the same mesh reduces the work from $f \cdot m \cdot n$ to $f \cdot m^2$. The portion of the code that needs to modified is

```
for ($\ell$ in $1\!:\!m$) { $k_0 = \ell + 1 - \lfloor h/\delta \rfloor$ ; $k_1 = \ell + \lfloor h/\delta \rfloor$
    for ($k$ in $\max(1, k_0) : \min(m, k_1)$) { $y_k = y_k + K_h(m_\ell - m_k)$ } }.
```

The use of rounded data in a kernel density estimator attracted much attention with regards to loss of accuracy (see Scott (1981), Scott and Sheather (1985), and Silverman (1982); see also Jones (1989)). Härdle and Scott (1988) coined the phrase *weighted averaging of rounded points* or the WARPing method to describe the general approach and applied the WARPing method to other multivariate algorithms. With modern computers, m can be quite large and δ quite small so that the approximation error introduced is negligible. In multiple dimensions, however, keeping an entire grid in core must be considered when $d \geq 3$.

6.7.2 Convolution and Fourier Transforms

In Section 6.1.2, we explored one motivation for the fixed kernel estimator via convolution. This engineering approach also leads to efficient computational algorithms using Fourier methods for performing the convolution. We start with three definitions:

$$\text{Convolution} \qquad (f * g)(x) = \int_{-\infty}^{\infty} f(t)\, g(x-t)\, dt \tag{6.102}$$

$$\text{Fourier Xfm} \qquad \tilde{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i x \xi}\, dx \tag{6.103}$$

$$\text{Inverse Fourier Xfm} \qquad f(x) = \int_{-\infty}^{\infty} \tilde{f}(\xi)\, e^{2\pi i \xi x}\, d\xi. \tag{6.104}$$

For example, representing the data $\{x_1,\ldots,x_n\}$ via the epdf, $f_n(x)$, in (2.2),

$$\begin{aligned}\tilde{f}_n(\xi) &= \int\left[\frac{1}{n}\sum_{j=1}^{n}\delta(x-x_j)\right]e^{-2\pi ix\xi}\,dx\\ &= \frac{1}{n}\sum_{j=1}^{n}\int\delta(x-x_j)\,e^{-2\pi ix\xi}\,dx = \frac{1}{n}\sum_{j=1}^{n}e^{-2\pi ix_j\xi}. \qquad (6.105)\end{aligned}$$

An alternative to performing the direct convolution operation (6.102), which can be very slow, is to perform three Fourier Transforms to obtain $c(x)=(f*g)(x)$ via:

$$\tilde{c}(\xi)=\tilde{f}(\xi)\cdot\tilde{g}(\xi) \quad\Longrightarrow\quad c(x)=\widetilde{\tilde{c}(\xi)}=\widetilde{\tilde{f}(\xi)\cdot\tilde{g}(\xi)}.$$

6.7.2.1 Application to Kernel Density Estimators Applied to a kernel estimator, Equation (6.102) becomes

$$\begin{aligned}\tilde{\hat{f}}(x) &= \int\hat{f}(x)\,e^{-2\pi ix\xi}\,dx \qquad (6.106)\\ &= \int\left[\frac{1}{n}\sum_{j=1}^{n}\frac{1}{h}K\left(\frac{x-x_j}{h}\right)\right]e^{-2\pi ix\xi}\,dx\\ &= \frac{1}{n}\sum_{j=1}^{n}\left[\int K(s)\,e^{-2\pi i(x_j+hs_j)\xi}\,ds_j\right]\\ &= \frac{1}{n}\sum_{j=1}^{n}\left[e^{-2\pi ix_j\xi}\int K(s)\,e^{-2\pi i(hs_j)\xi}\,ds_j\right]\\ &= \frac{1}{n}\sum_{j=1}^{n}e^{-2\pi ix_j\xi}\,\tilde{K}(hs_j) \qquad (6.107)\\ &= \tilde{f}_n\cdot\tilde{K}(hs),\end{aligned}$$

using the change of variables $s_j=(x-x_j)/h$ and noting that $\tilde{K}(hs_j)$ is the same for all j.

Now $\tilde{K}$ is known for many kernels, including the normal kernel:

$$\tilde{\phi}_h(\xi)=e^{-2h^2\pi^2\xi^2}.$$

Therefore, (6.107) becomes

$$\tilde{\hat{f}}(\xi)=\frac{1}{n}\sum_{j=1}^{n}e^{-2\pi ix_j\xi}\,e^{-2h^2\pi^2\xi^2}. \qquad (6.108)$$

We may check the correctness of (6.108) by computing its inverse Fourier Transformation using Mathematica (2012):

$$\tilde{\tilde{\hat{f}}}(x) = \hat{f}(x) = \int \tilde{\hat{f}}(\xi)\, e^{2\pi i \xi x}\, d\xi \tag{6.109}$$

$$= \frac{1}{n}\sum_{j=1}^{n} \int e^{-2\pi i x_j \xi}\, e^{-2h^2\pi^2\xi^2}\, e^{2\pi i \xi x}\, d\xi \tag{6.110}$$

$$= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}h} e^{-(x-x_j)^2/2h^2}, \tag{6.111}$$

which is the Gaussian kernel density estimate exactly.

6.7.2.2 FFTs In practice, the discrete Fourier Transform and the FFT are employed to compute a very good approximation to $\hat{f}(x)$. The FFT is applied to a finite interval, say $(-\frac{1}{2}, \frac{1}{2})$ w.l.o.g., and the resulting estimate is periodic. To avoid end effects, the data are rescaled to a subinterval such as $(-0.3, 0.3)$. Silverman (1986) recommends a buffer of at least $3h$ on each side. In the time series context, padding with zeroes in spectral density estimation is a well-known requirement to avoid aliasing, that is, frequencies overlapping (Blackman and Tukey, 1958).

Binning of the data is critical to the approach, that is,

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \approx \frac{1}{n}\sum_{k=1}^{m} \nu_k K_h(x - m_k),$$

and if we compute $\hat{f}(x)$ at the same bin midpoints

$$\hat{f}(m_\ell) = \frac{1}{n}\sum_{k=1}^{m} \nu_k K_h(m_\ell - m_k) = \sum_{k=1}^{m} \nu_k \cdot \frac{1}{n} K_h\big(\delta(\ell - k)\big), \tag{6.112}$$

which is precisely a discrete convolution. In R, this may be implemented via the *convolve* function, with arguments given by the bin counts and the kernel values at $t = m_\ell$ divided by n. So that the final values are properly aligned, the kernel values should be circularly shifted by half.

Multivariate extension of the FFT approach is discussed by Wand (1994) and Wand and Jones (1995). These authors and Silverman (1986) advocate replacing simple binning by linear binning, although with modern computing and very fine meshes, the impact may be rather minor.

6.7.2.3 Discussion The default density function in R now uses the FFT algorithm in the univariate case. In situations where a boundary kernel is called for, then other algorithms will need to be employed. The use of the FFT in more than one dimension is also available, but the mesh grows exponentially so that three dimensions is a practical limit.

The ASH approach goes beyond three dimensions by computing slices or conditional densities. It is also possible to look at subsets of the domain. In any case, there are many more opportunities to try kernels of ASH estimators in the multivariate setting, and the reader is encouraged to try these on their data. In addition to the ASH and *ashn* software available in R and from the author, the *ks* package is recommended (Duong, 2014). The package uses the *rgl* animation toolbox.

6.8 SUMMARY

This chapter has covered kernel methods, which is often the sole topic of a book. The kernel approach was motivated in different ways that might appeal to a statistician, numerical analyst, engineer, or mathematician. Issues related to the choice of kernel—its order, boundary properties, and design—were surveyed. Extensions to multivariate data and the use of the product kernel reviewed. Cross-validation ideas discussed in earlier chapters were extended to kernels. Finally, the important topic of locally adaptive methods was introduced. Adaptive methods hold much promise, but usually introduce many new parameters that are difficult to estimate, and frequently introduce artifacts of the sample (rather than the underlying density). An interesting survey is provided by Jones and Signorini (1997), who cover these adaptive algorithms and others in the context of higher order bias strategies. They observe there are many technical details to overcome. We have examined but a few. A conservative strategy that may be recommended is to stick with a fixed bandwidth kernel estimate and be aware of and sensitive to its limitations.

PROBLEMS

6.1 Compute the IV of estimator (6.2) and compare to the histogram result. What is the kernel if $\hat{f}(x) = [F_n(x+h) - F_n(x)]/h$ is used?

6.2 Compute the bias and variance directly for estimator (6.3). Demonstrate that the equivalent kernel $K = U(-0.5, 0.5)$.

6.3 Check that the ASH results are obtained from Theorem 6.1 when the isosceles triangle kernel is specified.

6.4 Examine the graphs of Wahba's equivalent kernel (6.9) for several choices of the two parameters p and λ.

6.5 Following the discussion of the rootgram for the histogram, show that the square root of the kernel estimate is variance stabilizing compared to (6.15). Show that the variance of the root-kernel estimate is $R(K)/(4nh)$.

6.6 Verify Equations (6.21). Find the optimal bandwidths for estimating the first and second derivatives of f. Evaluate these for the $N(\mu, \sigma^2)$ density.

6.7 Consider kernel $K_4(t)$ in Table 6.1. What if you make this kernel smoother by increasing the power on the first factor, such as $c(1-t^2)^k(a+bt^2)$, where k is 2 or 3, and the constants c, a, and b are chosen so the kernel is order-4 and integrates to 1. How do these alternative kernels behave theoretically? Visually?

6.8 Show that two kernel estimates with sample sizes in the ratio given in Equation (6.26) have the same AMISE.

6.9 Show that the IV of the fixed kernel estimate equals (exactly)

$$\mathrm{IV}(h) = \frac{1}{n}\int \mathrm{E}K_h(x-X)^2 dy - \frac{1}{n}\int [\mathrm{E}K_h(x-X)]^2 \, dx.$$

Show that the first term in the IV is *exactly* $R(K)/(nh)$. Show that the next terms are

$$-\frac{R(f)}{n} + \frac{h^2\mu_2 R(f')}{n} - \frac{h^4}{n}\left(\frac{\mu_2^2}{4} + \frac{\mu_4}{24}\right) R(f'') + \cdots,$$

where μ_k is the kth moment of the kernel.

6.10 Show that the ISB of a fixed kernel estimate is

$$\mathrm{ISB}(h) = \int [\mathrm{E}K_h(x-X) - f(x)]^2 \, dx.$$

Suppose that the kernel K is symmetric around zero so that $\int w^i K(w)dw = 0$ for i odd. Show that the first few bias terms equal

$$h^4 \frac{\mu_2^2}{4} R(f'') - h^6 \frac{\mu_2\mu_4}{24} R(f''') + h^8 \left(\frac{\mu_4^2}{476} + \frac{\mu_2\mu_6}{720}\right) R(f^{vi}).$$

Hint: Watch the cross-product terms and note that $\int f'' f^{\mathrm{iv}} = -\int (f''')^2$, for example.

6.11 Verify the bias expressions for the higher order finite difference estimators in Equation (6.32). Devise your own higher order boxcar kernels using different spacings than integer multiples of h.

6.12 Compare the AMSE(x) values for four combinations of pointwise kernel estimates—at 0 and 1 with a second-order and fourth-order kernel.

6.13 Empirically compare the order-4 kernel method to the Terrell–Scott fourth-order ratio estimator for simulated normal data as well as the snowfall data.

6.14 Derive the Terrell–Scott fourth-order kernel ratio estimator. Extend the procedure to a sixth-order estimator.

6.15 Compute the equivalent kernel of the parametric estimator $\phi(x|0,s^2)$.

6.16 Compute the "theoretical" kth moment of $\hat{f}$, treating the kernel estimator as a "true" density.

6.17 Find the indifferent frequency polygon kernel mentioned in Table 6.2.

6.18 Consider the class of shifted-Beta kernels, $c_k(1-t^2)_+^k$. Find their variance and rescale so that each has variance 1. Show that these rescaled kernels converge to a standard normal kernel as $k \to \infty$.

6.19 Derive the product kernel AMISE results from the general multivariate kernel formulas.

6.20 Verify the equivalent-bandwidth formula for higher order kernels in Equation (6.31). Check that the last factor is approximately equal to 1 for several kernels.

6.21 Using a Taylor's series on $\Delta F_n(x,kh)$, verify that the finite difference estimates in Equation (6.32) are of higher order. Derive some additional estimates based on the spacings $h, 2h, 4h, 8h, \ldots$.

6.22 Check by direct integration that the "variance" of the kernel in Equation (6.36) is 0.

6.23 Find boundary modification kernels based on the Epanechnikov kernel. Compare with kernels supported on $[c,\, 1]$. Investigate the increase in $R(K_c)$. How much wider (on the right) should the kernel be so that the roughness is the same as for the biweight kernel? Does such an "equivalent roughness" always exist?

6.24 Verify Equation (6.40).

6.25 Recall estimator (6.45). Show that it is functionally equivalent to choose K to be $N(\mathbf{0}_d, \Sigma)$ with $H = I_d$, or to choose K to be $N(\mathbf{0}_d, I_d)$ with $H = \Sigma^{1/2}$. Thus the linear transformation may be applied to either the data or the kernel as a matter of preference.

6.26 Finite support kernels need not have only a finite number of derivatives. For example, consider

$$K(t) \propto e^{-1/(1-t^2)} I_{(-1,1)}(t).$$

Show that the normalizing constant is $\sqrt{\pi}/e \times \text{Hypergeometric U}\left[\frac{1}{2}, 0, 1\right]$. Plot the kernel. *Hint:* Use the change of variable $t = (1-x^2)^{-1} - 1$ for $x \in (0,1)$.

6.27 Show that $c(x-a)^4/144$ is the solution to the null adaptive density differential Equation (6.90).

6.28 Verify the equivalent kernel for the parametric estimator $N(\bar{x}, 1)$. Plot $K(x,t)$. Compute the equivalent kernel for the parametric estimator $N(0, s^2)$ and plot it.

6.29 Recall that $\sigma_{\hat{f}}^2 = s_x^2 + h^2\sigma_K^2$. From this result, argue that if $s_x > \sigma_f$, then $\hat{h}_{\mathrm{ISE}} > \hat{h}_{\mathrm{MISE}}$ is the likely result. Is the converse true?

6.30 Try the simple orthogonal series estimator on some Beta data with $0 \le m \le 6$. With a larger sample, is there sufficient control on the estimate with m alone?

6.31 Using the simple estimate for the Fourier coefficients in Equation (6.6), find unbiased estimates of the two unknown terms in Equation (6.62).

6.32 Show that if a correction factor of the form $(1+bn^{-\delta})$ is applied to h^* in (6.79), then the best choice is $\delta = 1/5$.

6.33 Derive Equation (6.85). *Hint*: The fraction of points, *k*/*n*, in the ball of radius h centered on $\mathbf{x}$ is approximately equal to $f(\mathbf{x})$ times the volume of the ball.

6.34 Show that the k-NN estimator using the d_k distance function has infinite integral. *Hint*: In the univariate case, look at the estimator for $x > x_{(n)}$, the largest order statistic.

6.35 Use Jensen's inequality to show that the ratio in Equation (6.89) is ≤ 1. *Hint*: The integral in brackets in the denominator is $\mathrm{E}[f^{(p)}(X)^2/f(X)]$.

6.36 (Research). Rather than abandoning UCV in favor of more efficient asymptotic estimators, consider adjusting the data before computing the UCV curve. The noise in UCV is due in part to the fact that the terms $(x_i - x_j)/h$ are very noisy for $|x_i - x_j| < h$. Imagine placing small springs between the data points with the result that the interpoint distances become very smoothly changing. The resulting UCV seems much better behaved. A locally adaptive method replaces $x_i \leftarrow (x_{i-1} + x_{i+1})/2$, assuming that the data have been sorted. Try this modification on simulated data.

6.37 Prove that the average bias of a kernel estimate is *exactly* 0. *Hint*: The pointwise bias is $\int K(w) f(x-hw)dw - f(x)$.

6.38 Wand and Jones (1995) use two matrix identities, which may be found in Neudecker and Magnus (1988), to derive Equation (6.51). Suppose B and C are symmetric $d \times d$ matrices. Then hvec(B) vectorizes the lower triangular portion, while vec(B) vectorizes the entire matrix. These two vectors are connected by the $d^2 \times \frac{d(d+1)}{2}$ *duplication matrix*, D_d, whose entries are all 0's and 1's; that is, $D_d \,\mathrm{vech}(B) = \mathrm{vec}(B)$. The identities and relationships are

$$\begin{aligned} \mathrm{vec}(B) &= D_d\,\mathrm{vech}(B) \\ D_d^T\,\mathrm{vec}(B) &= \mathrm{vech}[\,2B - \mathrm{Diag}(B)\,] \qquad\qquad \text{and} \\ \mathrm{tr}\,(B^T C) &= \mathrm{vec}(B)^T\,\mathrm{vec}(C) = \mathrm{vec}(C)^T\,\mathrm{vec}(B). \end{aligned}$$

Complete the following argument. Let $B = AA^T$ in Equation (6.50). Write the integrand as

$$\begin{aligned} \mathrm{tr}\big[B\nabla^2 f(\mathbf{x})\big]^2 &= (\mathrm{vec}\,B)^T[\mathrm{vec}\,\nabla^2 f(\mathbf{x})]\cdot[\mathrm{vec}\,\nabla^2 f(\mathbf{x})]^T(\mathrm{vec}\,B) \\ &= \mathrm{vech}(B)^T D_d^T\,[\mathrm{vec}\,\nabla^2 f(\mathbf{x})]\cdot[\mathrm{vec}\,\nabla^2 f(\mathbf{x})]^T D_d\,\mathrm{vech}(B). \end{aligned}$$

The derivation may be completed by using the second identity on $D_d^T\,[\mathrm{vec}\,\nabla^2 f(\mathbf{x})]$.