CHAPTER 3

Bandwidth selection

3.1 Introduction

The practical implementation of the kernel density estimator requires the specification of the bandwidth h. This choice is very important as was shown graphically in Section 2.2 and through the AMISE analysis in Section 2.5. There are many situations where it is satisfactory to choose the bandwidth subjectively by eye. This would involve looking at several density estimates over a range of bandwidths and selecting the density that is the "most pleasing" in some sense. One such strategy is to begin with a large bandwidth and to decrease the amount of smoothing until fluctuations that are more "random" than "structural" start to appear. This approach is more viable when the user has reasons to believe that there is certain structure in the data, such as knowledge of the position of modes. However, there are also many circumstances where it is very beneficial to have the bandwidth automatically selected from the data. One reason is that it can be very time consuming to select the bandwidth by eye if there are many density estimates required for a given problem. Another is that, in many cases, the user has no prior knowledge about the structure of the data and would not have any feeling for which bandwidth gives an estimate closest to the true density. When kernel estimators are used as components of larger statistical procedures, automatic bandwidth selection is usually necessary.

A method that uses the data $X_1, \ldots, X_n$ to produce a bandwidth $\hat{h}$ is called a *bandwidth selector*. The bandwidth selection problem is present in all types of kernel estimation, including the scatterplot smoothing problem dealt with in Chapter 5, although kernel density estimation provides a convenient setting for the development of many of the key ideas.

Currently available bandwidth selectors can be roughly divided into two classes. The first class consists of simple easily computable formulae which aim to find a bandwidth that is "reasonable" for a wide range of situations, but without out any mathematical guarantees of being close to the optimal bandwidth. We will call such bandwidth selectors *quick and simple* and discuss some proposals in Section 3.2. Quick and simple bandwidth selectors are motivated by the need to have fast automatically generated kernel estimates for algorithms that require many curve estimation steps as well as providing a reasonable starting point for subjective choice of the smoothing parameter.

The second type of bandwidth selector will be labelled as *hi-tech* since such selection procedures are based on more involved mathematical arguments and require considerably more computational effort, but aim to give a good answer for very general classes of underlying functions. Each of the hi-tech bandwidth selectors that we discuss in this chapter can be motivated through aiming to minimise $\mathrm{MISE}\{\hat{f}(\cdot;h)\}$ and can be shown to attain this goal *asymptotically* to some extent. Such a bandwidth selector is said to be *consistent* with respect to MISE. This chapter is devoted to a presentation and comparison of such MISE-driven bandwidth selectors. It should be pointed out that there exist approaches to bandwidth selection based on other loss criteria. However, since their analysis is more difficult we will restrict attention to the simpler MISE-driven selectors.

At the time of writing the field of bandwidth selection remains fairly unsettled, with new selectors being developed and several unresolved issues. Despite this, there has been a considerable amount of path-breaking research on the problem in recent years.

This chapter aims to reach a compromise between simple presentation of the main ideas and representative coverage of current approaches to practical bandwidth selection.

Throughout this section we will assume that $\hat{f}(\cdot;h)$ is a kernel density estimator based on a random sample $X_1,\ldots,X_n$ having density f. Furthermore, we will assume that $\hat{f}(\cdot;h)$ uses a second-order kernel K and that f is sufficiently well-behaved for all arguments involving differentiability and integrability assumptions to be valid.

3.2 Quick and simple bandwidth selectors

In this section we describe two commonly used quick and simple ideas for selecting the bandwidth of the kernel density estimator $\hat{f}(\cdot; h)$. Quick and simple rules also play an important role in the implementation of several hi-tech bandwidth selectors, as we will see further on.

3.2.1 *Normal scale rules*

A *normal scale* bandwidth selector simply involves using the bandwidth that is AMISE-optimal for the normal density having the same scale as that estimated for the underlying density. As shown in Section 2.5, the bandwidth that minimises $\text{MISE}\{\hat{f}(\cdot; h)\}$ asymptotically is

$$h_{\text{AMISE}} = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')n}\right]^{1/5}.$$

If f is normal with variance σ^2 then it is easily shown (Exercise 3.1) that

$$h_{\text{AMISE}} = \left[\frac{8\pi^{1/2}R(K)}{3\mu_2(K)^2 n}\right]^{1/5}\sigma. \tag{3.1}$$

A normal scale bandwidth selector is obtained from (3.1) by simply replacing σ by $\hat{\sigma}$:

$$\hat{h}_{\text{NS}} = \left[\frac{8\pi^{1/2}R(K)}{3\mu_2(K)^2 n}\right]^{1/5}\hat{\sigma} \tag{3.2}$$

where $\hat{\sigma}$ is some estimate of σ (e.g. Silverman, 1986, pp.45–47). Common choices of $\hat{\sigma}$ are the sample standard deviation s and the standardised interquartile range

$$\hat{\sigma}_{\text{IQR}} = (\text{sample interquartile range})/\{\Phi^{-1}(\tfrac{3}{4}) - \Phi^{-1}(\tfrac{1}{4})\}$$

where Φ^{-1} is the standard normal quantile function. Note that the normalising factor in the denominator of $\hat{\sigma}_{\text{IQR}}$ is the population interquartile range of the standard normal density and is approximately equal to 1.349. Use of $\hat{\sigma}_{\text{IQR}}$ guards against outliers if f has heavy tails. It is sometimes recommended that the smaller of s and $\hat{\sigma}_{\text{IQR}}$ be used (Silverman, 1986, p.47) to lessen the chances of oversmoothing. More sophisticated scale estimates have also

been studied and recommended (Janssen, Marron, Veraverbeke and Sarle, 1995).

Normal scale bandwidth selectors provide a quick "first guess" bandwidth and can be expected to give reasonable answers when the data are close to normal. However, for departures from normality such as multimodality, which one usually hopes to be detected by a density estimate, normal scale bandwidth selectors tend to oversmooth and mask important features in the data.

3.2.2 *Oversmoothed bandwidth selection rules*

The *oversmoothing* or *maximal smoothing* principle relies on the fact that there is a simple upper bound for the AMISE-optimal bandwidth for estimation of densities with a fixed value of a particular scale measure. For example, it can be shown (Terrell, 1990, Theorem 1) that

$$h_{\text{AMISE}} \leq \left[\frac{243R(K)}{35\mu_2(K)^2 n}\right]^{1/5} \sigma \tag{3.3}$$

for all densities having standard deviation σ and that this bound is attained by the beta(4,4) or triweight density, pictured in Figure 2.11. Similar results can be shown to hold for other scale measures (Terrell and Scott, 1985, Terrell, 1990). The above bound on h_{AMISE} motivates the *oversmoothed* bandwidth selector

$$\hat{h}_{\text{OS}} = \left[\frac{243R(K)}{35\mu_2(K)^2 n}\right]^{1/5} s$$

where s is the sample standard deviation. It is also possible to base $\hat{h}_{\text{OS}}$ on other common scale measures (Terrell, 1990). While $\hat{h}_{\text{OS}}$ will give too large a bandwidth for optimal estimation of a general density f it provides an excellent starting point for subjective choice of the bandwidth. A sensible graphical strategy is to plot an estimate with bandwidth $\hat{h}_{\text{OS}}$ and then successively look at plots based on convenient fractions of $\hat{h}_{\text{OS}}$ to see what features are present in the data for various amount of smoothing.

Figure 3.1 illustrates this idea for the *Old Faithful* data set, consisting of 107 eruption times in minutes for the Old Faithful Geyser in Yellowstone National Park (source: Silverman, 1986, p.8).

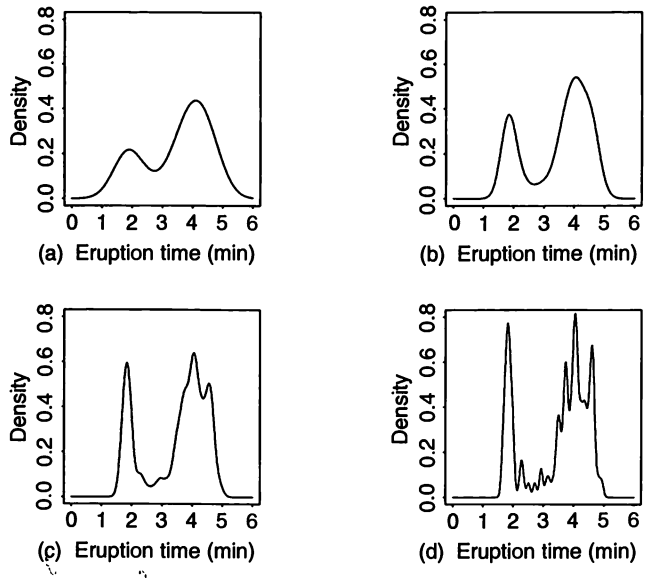

Figure 3.1. *Density estimates based on 107 eruption times (minutes) of the Old Faithful Geyser. The normal kernel is used. Bandwidths are (a)* $\hat{h}_{\mathrm{OS}}$*, (b)* $\hat{h}_{\mathrm{OS}}/2$*, (c)* $\hat{h}_{\mathrm{OS}}/4$ *and (d)* $\hat{h}_{\mathrm{OS}}/8$.

Figure 3.1 (a) shows the density estimate based on the normal kernel with bandwidth $\hat{h}_{\mathrm{OS}} = 0.467$. An important point to note from this estimate is that it is bimodal, despite the fact that it is oversmoothed. This provides very strong evidence in favour of eruption times exhibiting a bimodal distribution, with distinct clusters centred around about 1.9 minutes and 4.2 minutes. Decreasing the bandwidth by a factor of 2 results in Figure 3.1 (b). This density estimate retains the bimodal structure of the oversmoothed estimate, but resolves these features more sharply. Halving the bandwidth again leads to Figure 3.1 (c). In this case five modes are present, the smallest three almost certainly an artifact of having too small a bandwidth. Figure 3.1 (d), with bandwidth $\hat{h}_{\mathrm{OS}}/8$, leads to a very undersmoothed estimate which is far too wiggly to be a serious contender for modelling eruption times. Of these four estimates, (b) is the most pleasing since it appears to reach a good compromise between highlighting features in the data and containing its variability.

The oversmoothed and normal scale bandwidth selectors based on standard deviation are closely related in the sense that

$$\hat{h}_{\mathrm{NS}}/\hat{h}_{\mathrm{OS}} = (280\pi^{1/2}/729)^{1/5} \simeq 0.93.$$

This is because the normal density is close to obtaining the upper bound in (3.3).

3.3 Least squares cross-validation

We will now begin our description of a selection of hi-tech bandwidth selectors. Among the earliest fully automatic and consistent bandwidth selectors were those based on cross-validation ideas. *Least squares cross-validation* (LSCV) (Rudemo, 1982, Bowman, 1984) is the name given to a conceptually simple and appealing bandwidth selector. Its motivation comes from expanding the MISE of $\hat{f}(\cdot;h)$ to obtain

$$\begin{aligned}\text{MISE}\{\hat{f}(\cdot;h)\} &= E\int \hat{f}(x;h)^2\,dx - 2E\int \hat{f}(x;h)f(x)\,dx \\ &\quad + \int f(x)^2\,dx.\end{aligned}$$

Notice that the $\int f(x)^2\,dx$ term does not depend on h, so minimization of $\text{MISE}\{\hat{f}(\cdot;h)\}$ is equivalent to minimization of

$$\begin{aligned}&\text{MISE}\{\hat{f}(\cdot;h)\} - \int f(x)^2\,dx = \\ &\qquad E\left[\int \hat{f}(x;h)^2\,dx - 2\int \hat{f}(x;h)f(x)\,dx\right].\end{aligned}$$

The right-hand side is unknown since it depends on f. However, it can be shown (Exercise 3.3) that an unbiased estimator for this quantity is

$$\text{LSCV}(h) = \int \hat{f}(x;h)^2\,dx - 2n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}(X_i;h)$$

where

$$\hat{f}_{-i}(x;h) = (n-1)^{-1}\sum_{j\neq i}^{n} K_h(x-X_j)$$

is the density estimate based on the sample with X_i deleted, often called the "leave-one-out" density estimator. This is the reason for the term "cross-validation" which refers to the use of part of

a sample to obtain information about another part. It therefore seems reasonable to choose h to minimise LSCV(h). We denote the bandwidth chosen according to this strategy by $\hat{h}_{\text{LSCV}}$. It is sometimes the case that LSCV(h) has more than one local minimum (Hall and Marron, 1991a).

Figure 3.2 shows LSCV(h) versus $\log_{10}(h)$ for two particular samples of size $n = 100$ from the standard normal density using the normal kernel.

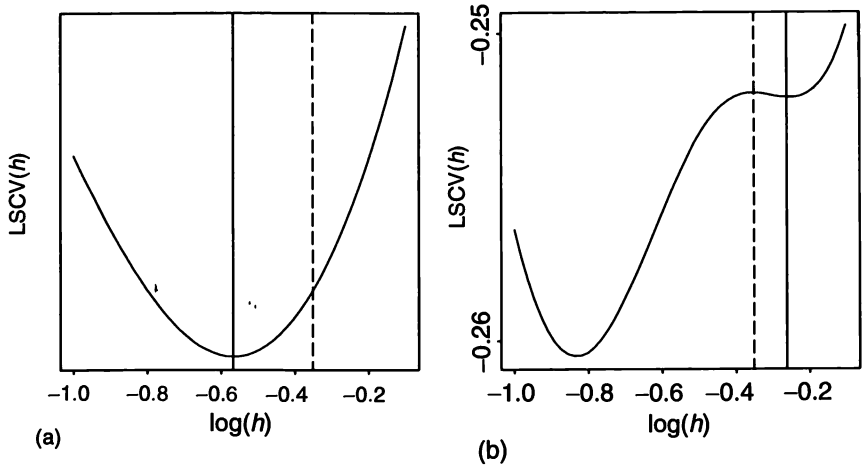

Figure 3.2. *Examples of* LSCV(h) *for two samples of* 100 $N(0,1)$ *observations. A* $\log_{10}$ *scale is used on the horizontal axis. The dashed vertical line shows the position of* $\log_{10}(h_{\text{MISE}})$. *The solid vertical lines show the position of* $\log_{10}(\hat{h}_{\text{LSCV}})$ *if* $\hat{h}_{\text{LSCV}}$ *is taken to correspond to the largest local minimum. The kernel is the standard normal density.*

Figure 3.2 (b) is an example of LSCV having two minima. The $\log_{10}$ of the MISE-optimal bandwidth $h_{\text{MISE}} \simeq 0.445$ is shown by the dashed vertical line. Notice that the actual minimum is much smaller than h_{MISE}, while the larger minimiser is considerably closer to h_{MISE}. This phenomenon has led to the suggestion that $\hat{h}_{\text{LSCV}}$ be taken to correspond to the *largest* local minimiser of LSCV(h) (Marron, 1993). The multiple minima phenomenon also means that care needs to be taken when finding $\hat{h}_{\text{LSCV}}$ in practice.

Studies have shown (e.g. Hall and Marron, 1987a, Park and Marron, 1990) that the theoretical and practical performance of this bandwidth selector are somewhat disappointing. In particular, $\hat{h}_{\text{LSCV}}$ is highly variable (see Figure 3.2). This has since led to the

proposal of several other hi-tech bandwidth selectors that aim to improve upon $\hat{h}_{\mathrm{LSCV}}$.

3.4 Biased cross-validation

Instead of the exact MISE formula used by least squares cross-validation, *biased cross-validation* (BCV) (Scott and Terrell, 1987) is based on the formula for the asymptotic MISE:

$$\mathrm{AMISE}\{\hat{f}(\cdot;h)\} = (nh)^{-1}R(K) + \tfrac{1}{4}h^4\mu_2(K)^2R(f''). \qquad (3.4)$$

The BCV objective function is obtained by replacing the unknown $R(f'')$ in (3.4) by the estimator

$$\begin{aligned}\widetilde{R(f'')} &= R(\hat{f}''(\cdot;h)) - (nh^5)^{-1}R(K'') \\ &= n^{-2}\sum\sum_{i\neq j}(K_h'' * K_h'')(X_i - X_j)\end{aligned}$$

to give

$$\mathrm{BCV}(h) = (nh)^{-1}R(K) + \tfrac{1}{4}h^4\mu_2(K)^2\widetilde{R(f'')}.$$

The BCV bandwidth selector, which we denote by $\hat{h}_{\mathrm{BCV}}$, is the minimiser of $\mathrm{BCV}(h)$. This selector is really a hybrid of cross-validation and "plug-in" bandwidth selection, as described in Section 3.6, since it involves replacement of the unknown $R(f'')$ by the cross-validatory kernel estimator $\widetilde{R(f'')}$.

The main attraction of $\hat{h}_{\mathrm{BCV}}$ is that it is more stable than $\hat{h}_{\mathrm{LSCV}}$, in the sense that its asymptotic variance is considerably lower (see Section 3.8). However, this reduction in variance comes at the price of an increase in bias, with $\hat{h}_{\mathrm{BCV}}$ tending to be larger than the MISE-optimal bandwidth. This is illustrated in Figure 3.3, which shows kernel density estimates based on $\hat{h}_{\mathrm{LSCV}}$ and $\hat{h}_{\mathrm{BCV}}$ bandwidths obtained from 500 simulated samples of size $n = 100$ from the normal mixture density f_1 defined at (2.3). The estimates are on a $\log_{10}$ scale with $\log_{10}(h_{\mathrm{MISE}})$ subtracted. The bandwidths for these estimates were obtained using the normal scale rule based on the sample standard deviation. This is a reasonable choice since $\hat{h}_{\mathrm{LSCV}}$ and $\hat{h}_{\mathrm{BCV}}$ each have asymptotically normal distributions (see Section 3.8). The vertical line at 0 shows the position of h_{MISE}. The normal kernel is used throughout.

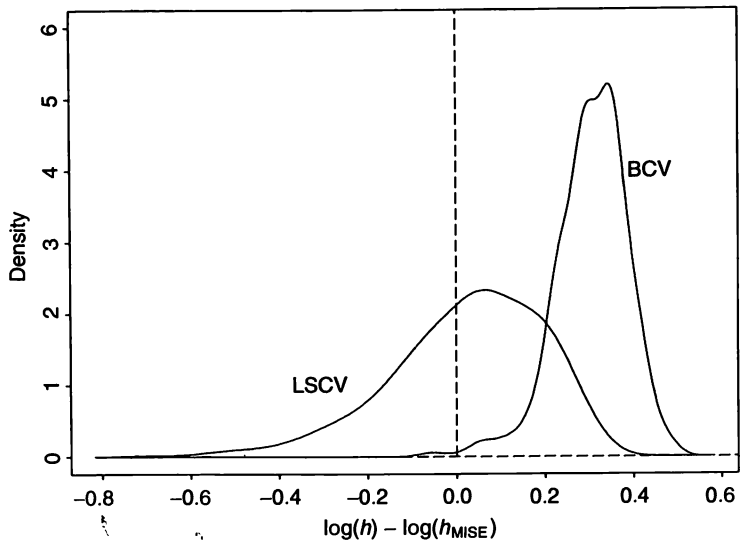

Figure 3.3. *Density estimates of* $\log_{10}(\hat{h}_{\mathrm{LSCV}}) - \log_{10}(h_{\mathrm{MISE}})$ *and* $\log_{10}(\hat{h}_{\mathrm{BCV}}) - \log_{10}(h_{\mathrm{MISE}})$. *Selected bandwidths are based on 500 simulated samples of size* $n = 100$ *from the normal mixture density* f_1 *defined at (2.3).*

The relative stability of BCV is manifest through the tightness of the distribution of the density of the $\hat{h}_{\mathrm{BCV}}$'s compared to those of the $\hat{h}_{\mathrm{LSCV}}$'s. However, the fact that the BCV distribution is situated to the right of 0 indicates the positive bias present in the $\hat{h}_{\mathrm{BCV}}$'s, while there is no noticeable bias in the $\hat{h}_{\mathrm{LSCV}}$'s. This variance-bias trade-off for bandwidth selectors is also present for other types of rules and is mathematically quantified in Section 3.8.

Like LSCV, the BCV criterion function occasionally has more than one local minimum, as well as being globally minimised at $h = 0$. Rules for choosing among local minima have been suggested in the literature (Scott, 1992, p.167, Marron, 1993).

3.5 Estimation of density functionals

An important component of many of the current hi-tech univariate bandwidth selectors is the estimation of integrated squared density derivatives. This is because they arise in various expressions for optimal bandwidths. The general integrated squared density derivative functional is

$$R(f^{(s)}) = \int f^{(s)}(x)^2 \, dx.$$

However, it is a simple exercise in using integration by parts to show that, under sufficient smoothness assumptions on f,

$$R(f^{(s)}) = (-1)^s \int f^{(2s)}(x) f(x) \, dx.$$

It is therefore sufficient to study estimation of functionals of the form

$$\psi_r = \int f^{(r)}(x) f(x) \, dx$$

for r even. Note that the sign of ψ_{2s} is the same as that of $(-1)^s$ and $\psi_r = 0$ if r is odd. We prefer the ψ_r notation to the usual $R(f^{(r)})$ notation because its extension to the multivariate context is more straightforward.

Note that

$$\psi_r = E\{f^{(r)}(X)\}.$$

This motivates the estimator

$$\hat{\psi}_r(g) = n^{-1} \sum_{i=1}^{n} \hat{f}^{(r)}(X_i; g) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} L_g^{(r)}(X_i - X_j)$$

(Hall and Marron, 1987b, Jones and Sheather, 1991) where g and L are, respectively, a bandwidth and kernel that are possibly different from h and K.

The asymptotic MSE properties of $\hat{\psi}_r$ are of fundamental importance for the bandwidth selectors described in the following two sections. We will give their derivation here under the following assumptions:

(i) the kernel L is a symmetric kernel of order k, $k = 2, 4, \ldots$, possessing r derivatives, such that

$$(-1)^{(r+k)/2+1} L^{(r)}(0) \mu_k(L) > 0.$$

(ii) the density f has p continuous derivatives that are each ultimately monotone, where $p > k$.

(iii) $g = g_n$ is a positive-valued sequence of bandwidths satisfying

$$\lim_{n\to\infty} g = 0 \quad \text{and} \quad \lim_{n\to\infty} ng^{2r+1} = \infty.$$

Condition (i) is satisfied by most of the kernels that are used in practice.

We seek a large sample approximation to

$$\text{MSE}\{\hat{\psi}_r(g)\} = E\{\hat{\psi}_r(g) - \psi_r\}^2.$$

First note that

$$\hat{\psi}_r(g) = n^{-1}L_g^{(r)}(0) + n^{-2}\sum\sum_{i\neq j} L_g^{(r)}(X_i - X_j),$$

the first term being independent of the data. Clearly,

$$E\hat{\psi}_r(g) = n^{-1}L_g^{(r)}(0) + (1 - n^{-1})E\{L_g^{(r)}(X_1 - X_2)\}$$

and, by Taylor's theorem and the smoothness assumptions on f,

$$\begin{aligned}
E\{L_g^{(r)}(X_1 - X_2)\} &= \int\int L_g^{(r)}(x - y)f(x)f(y)\,dx\,dy \\
&= \int\int L_g(x - y)f(x)f^{(r)}(y)\,dx\,dy \\
&= \int\int L(u)f(y + gu)f^{(r)}(y)\,du\,dy \\
&= \int\int L(u)f^{(r)}(y) \\
&\qquad \times \left\{\sum_{\ell=0}^{k}(\ell!)^{-1}(ug)^{\ell}f^{(\ell)}(y) + O(g^{k+1})\right\} du\,dy \\
&= \psi_r + (k!)^{-1}\mu_k(L)g^k\psi_{r+k} + O(g^{k+2}).
\end{aligned}$$

The bias can therefore be expressed as

$$\begin{aligned}
E\hat{\psi}_r(g) - \psi_r = {} & n^{-1}g^{-r-1}L^{(r)}(0) \\
& + (k!)^{-1}g^k\mu_k(L)\psi_{r+k} + O(g^{k+2}).
\end{aligned}$$

It follows from Exercise 3.4 and the symmetry of $L^{(r)}$ for r even that

$$\begin{aligned}\mathrm{Var}\{\hat{\psi}_r(g)\} &= 2n^{-3}(n-1)\mathrm{Var}\{L_g^{(r)}(X_1 - X_2)\} \\ &\quad + 4n^{-3}(n-1)(n-2)\mathrm{Cov}\{L_g^{(r)}(X_1 - X_2), L_g^{(r)}(X_2 - X_3)\}.\end{aligned} \tag{3.5}$$

We will treat each component of the variance and covariance in turn. Firstly,

$$\begin{aligned}E[\{L_g^{(r)}(X_1 - X_2)\}^2] &= \int\int \{L_g^{(r)}(x-y)\}^2 f(x)f(y)\,dx\,dy \\ &= g^{-2r-1}\int\int L^{(r)}(u)^2 f(y+gu)f(y)\,du\,dy \\ &= g^{-2r-1}\psi_0 R(L^{(r)}) + o(g^{-2r-1})\end{aligned}$$

while

$$\begin{aligned}&E\{L_g^{(r)}(X_1 - X_2)L_g^{(r)}(X_2 - X_3)\} \\ &= \int\int\int L_g^{(r)}(x-y)L_g^{(r)}(y-z)f(x)f(y)f(z)\,dx\,dy\,dz \\ &= \int\int\int L_g(x-y)L_g(y-z)f^{(r)}(x)f(y)f^{(r)}(z)\,dx\,dy\,dz \\ &= \int\int\int L(u)L(v)f^{(r)}(y+gu)f(y)f^{(r)}(y-gv)\,du\,dv\,dy \\ &= \int f^{(r)}(y)^2 f(y)\,dy + o(1).\end{aligned}$$

Lastly, from above we have

$$E\{L_g^{(r)}(X_1 - X_2)\} = \psi_r + o(1).$$

Combining each of these approximations with (3.5) leads to

$$\begin{aligned}&\mathrm{Var}\{\psi_r(g)\} = 2n^{-2}g^{-2r-1}\psi_0 R(L^{(r)}) \\ &+ 4n^{-1}\left\{\int f^{(r)}(x)^2 f(x)\,dx - \psi_r^2\right\} + o(n^{-2}g^{-2r-1} + n^{-1}).\end{aligned}$$

The asymptotic MSE is therefore

$$\begin{aligned}&\mathrm{MSE}\{\hat{\psi}_r(g)\} = \\ &\{n^{-1}g^{-r-1}L^{(r)}(0) + (k!)^{-1}g^k\mu_k(L)\psi_{r+k}\}^2 \\ &+ 2n^{-2}g^{-2r-1}R(L^{(r)})\psi_0 + 4n^{-1}\left\{\int f^{(r)}(x)^2 f(x)\,dx - \psi_r^2\right\} \\ &+ O(g^{2k+2}) + o(n^{-2}g^{-2r-1} + n^{-1}).\end{aligned}$$

Notice that, because of our assumption about the sign of $L^{(r)}(0)\mu_k(L)$, the main bias term can be made to vanish by choosing g to equal

$$g_{\text{AMSE}} = \left[\frac{k!L^{(r)}(0)}{-\mu_k(L)\psi_{r+k}n}\right]^{1/(r+k+1)}. \tag{3.6}$$

While this choice reduces the squared bias of the MSE to be of order $n^{-(2k+4)/(r+k+1)}$ we also need to check the orders of the variance terms. Since $g_{\text{AMSE}} = O(n^{-1/(r+k+1)})$ we obtain the leading variance terms to be of orders $n^{-(2k+1)/(r+k+1)}$ and n^{-1} respectively. The first of these variance terms dominates the remaining squared bias term so it is clear that the rate of convergence of the minimum MSE depends only on the leading variance terms. It is easy to check that for $k < r$,

$$\inf_{g>0} \text{MSE}\{\hat{\psi}_r(g)\} \sim$$
$$2R(L^{(r)})\psi_0 \left[\frac{\mu_k(L)\psi_{r+k}}{-L^{(r)}(0)k!}\right]^{(2r+1)/(r+k+1)} n^{-(2k+1)/(r+k+1)}$$

while for $k > r$,

$$\inf_{g>0} \text{MSE}\{\hat{\psi}_r(g)\} \sim 4\left[\text{Var}\{f^{(r)}(X)\}\right] n^{-1}.$$

If $k = r$ then the two leading terms in the above expression are of the same order and the leading term of the minimum mean squared error is the sum of these terms. Therefore, the parametric rate of convergence of order n^{-1} is achievable, provided the kernel is at least of order r. Of course, choosing k to be higher means that p, the number of continuous derivatives that we assume for f, is higher as well.

Finally, we point out that the computation of $\hat{\psi}_r(g)$ can be very expensive if a direct algorithm is used. This is because it involves $O(n^2)$ operations. In practice it is recommended that binned approximations, as described in Appendix D, be used instead.

3.6 Plug-in bandwidth selection

3.6.1 *Direct plug-in rules*

Plug-in bandwidth selectors are based on the simple idea of "plugging in" estimates of the unknown quantities that appear in formulae for the asymptotically optimal bandwidth. In terms of the ψ_r functionals studied in the previous section the AMISE-optimal bandwidth is

$$h_{\text{AMISE}} = \left[\frac{R(K)}{\mu_2(K)^2 \psi_4 n}\right]^{1/5}.$$

Replacement of ψ_4 by the kernel estimator $\hat{\psi}_4(g)$ leads to the *direct plug-in* (DPI) rule

$$\hat{h}_{\text{DPI}} = \left[\frac{R(K)}{\mu_2(K)^2 \hat{\psi}_4(g) n}\right]^{1/5}.$$

Unfortunately, this rule is not fully automatic since $\hat{h}_{\text{DPI}}$ depends on the choice of the *pilot bandwidth* g. One way of choosing g is to appeal to the formula for the AMSE-optimal bandwidth for estimation of $\hat{\psi}_4(g)$. If the same second-order kernel K is used in $\hat{\psi}_4(g)$ then from (3.6) the AMSE-optimal bandwidth is

$$g_{\text{AMSE}} = \left[\frac{2K^{(4)}(0)}{-\mu_2(K)\psi_6 n}\right]^{1/7}.$$

However, this rule for choosing g has the same defect as $\hat{h}_{\text{DPI}}$ above: it depends on an unknown density functional, namely ψ_6. We could estimate ψ_6 using another kernel estimate, but its optimal bandwidth depends on ψ_8. This problem will not go away since it is apparent from (3.6) that the optimal bandwidth for estimating ψ_r depends on ψ_{r+2}.

The usual strategy for overcoming this problem is to estimate a ψ_r functional with a quick and simple estimate, such as a version of the normal scale rule described in Section 3.2.1. This means that we really have a family of direct plug-in bandwidth selectors that depend on the number of stages of functional estimation before a quick and simple estimate is used. Suppose that a direct plug-in rule involves ℓ successive kernel functional estimations, with the initial bandwidth chosen via a quick and simple procedure. We will call such a rule an *ℓ-stage direct plug-in* bandwidth selector

and denote it by $\hat{h}_{\mathrm{DPI},\ell}$. Note that the normal scale rule (3.2) can be thought of as being a zero-stage direct plug-in bandwidth selector.

Appendix C contains results that are very useful for computing quantities required for certain bandwidth selection strategies, especially those that involve normal scale rules and normal kernels. The following result is particularly useful and follows from Fact C.1.12 and Fact C.1.6 in Appendix C. If f is a normal density with variance σ^2 then, for r even,

$$\psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1}(r/2)!\pi^{1/2}}. \tag{3.7}$$

EXAMPLE. A version of the two-stage plug-in bandwidth selector follows (Sheather and Jones, 1991). Note that we are using $L = K$, where K is a second-order kernel.

Step 1 Estimate ψ_8 using the normal scale estimate $\hat{\psi}_8^{\mathrm{NS}} = 105/(32\pi^{1/2}\hat{\sigma}^9)$ where $\hat{\sigma}$ is an estimate of scale. The formula for $\hat{\psi}_8^{\mathrm{NS}}$ is obtained from (3.7).

Step 2 Estimate ψ_6 using the kernel estimator $\hat{\psi}_6(g_1)$ where $g_1 = [-2K^{(6)}(0)/\{\mu_2(K)\hat{\psi}_8^{\mathrm{NS}} n\}]^{1/9}$.

Step 3 Estimate ψ_4 using the kernel estimator $\hat{\psi}_4(g_2)$ where $g_2 = [-2K^{(4)}(0)/\{\mu_2(K)\hat{\psi}_6(g_1) n\}]^{1/7}$.

Step 4 The selected bandwidth is

$$\hat{h}_{\mathrm{DPI},2} = [R(K)/\{\mu_2(K)^2\hat{\psi}_4(g_2) n\}]^{1/5}.$$

∎

Another selection problem should now be apparent: how should one choose the value of ℓ, the number of stages of functional estimation? To give an idea of the effect of ℓ on the performance of $\hat{h}_{\mathrm{DPI},\ell}$, density estimates of 500 simulated bandwidths (on a $\log_{10}$ scale with $\log_{10} h_{\mathrm{MISE}}$ subtracted) are plotted in Figure 3.4 for the direct plug-in rule with $\ell = 0, 1, 2$ and 3 and sample standard deviation used in the normal scale rule for the initial bandwidth.

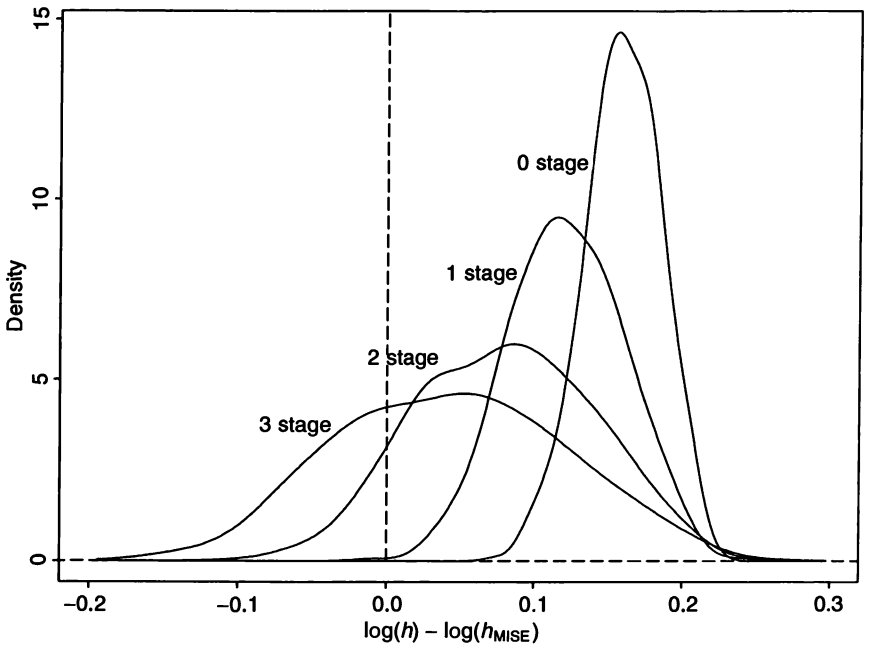

Figure 3.4. *Density estimates based on values of* $\log_{10}(\hat{h}_{\mathrm{DPI},\ell}) - \log_{10}(h_{\mathrm{MISE}})$ *for* $\ell = 0, 1, 2, 3$. *Selected bandwidths are based on 500 simulated samples of size* $n = 100$ *from the normal mixture density* f_1 *defined at (2.3).*

The selected bandwidths were obtained from samples of size $n = 100$ from the normal mixture density f_1 defined at (2.3). The vertical line shows the position of h_{MISE}. The estimated density of the zero-stage bandwidth $\hat{h}_{\mathrm{DPI},0}$, which we have also denoted by $\hat{h}_{\mathrm{NS}}$, is represented by the right-most curve. Since these bandwidths are based on a normal distribution assumption it is not surprising that they tend to be larger than h_{MISE} (since the optimal bandwidth for normal data is also larger). Also note the relative tightness of the distribution of the $\hat{h}_{\mathrm{DPI},0}$. This is because the variability between different bandwidths is due only to different standard deviation estimates. As ℓ increases we see that the bandwidth selector becomes less biased, since the dependence on the normal scale rule diminishes. However, the extra functional estimation steps for larger ℓ lead to the selector being more variable. This stage selection problem is also present in related hi-tech rules, such as smoothed cross-validation described in Section 3.7. At the time of writing, there was no method for objective choice of ℓ. However, theoretical considerations (Aldershof, 1991; Park and

Marron, 1992) favour taking ℓ to be at least 2, with $\ell = 2$ being a common choice.

3.6.2 *Solve-the-equation rules*

Also motivated by the formula for the AMISE-optimal bandwidth, *solve-the-equation* (STE) rules (Scott, Tapia and Thompson, 1977, Sheather, 1986, Park and Marron, 1990, Sheather and Jones, 1991, Engel, Herrmann and Gasser, 1995) require that h be chosen to satisfy the relationship

$$h = \left[\frac{R(K)}{\mu_2(K)^2 \hat{\psi}_4(\gamma(h))n}\right]^{1/5},$$

where the pilot bandwidth for the estimation of ψ_4 is a function γ of h. The choice of γ can be motivated by noting the relationship

$$g_{\text{AMSE}} = \left[\frac{2L^{(4)}(0)\mu_2(K)^2}{R(K)\mu_2(L)}\right]^{1/7} (-\psi_4/\psi_6)^{1/7} h_{\text{AMISE}}^{5/7}.$$

This suggests taking

$$\gamma(h) = \left[\frac{2L^{(4)}(0)\mu_2(K)^2}{R(K)\mu_2(L)}\right]^{1/7} \{-\hat{\psi}_4(g_1)/\hat{\psi}_6(g_2)\}^{1/7} h^{5/7}$$

where $\hat{\psi}_4(g_1)$ and $\hat{\psi}_6(g_2)$ are kernel estimates of ψ_4 and ψ_6 (Sheather and Jones, 1991). The choice of g_1 and g_2 can be made using (3.6), although this leads to another "stage selection" problem, as in the direct plug-in case.

EXAMPLE. A two-stage solve-the-equation bandwidth selector that uses $L = K$, which we will denote by $\hat{h}_{\text{STE},2}$, is given below (Sheather and Jones, 1991). It requires a numerical routine to implement Step 4; simulation experience suggests uniqueness of the solution and a relatively easy computational problem to overcome.

Step 1 Estimate ψ_6 and ψ_8 using $\hat{\psi}_6^{\text{NS}} = -15/(16\pi^{1/2}\hat{\sigma}^7)$ and $\hat{\psi}_8^{\text{NS}} = 105/(32\pi^{1/2}\hat{\sigma}^9)$.

Step 2 Estimate ψ_4 and ψ_6 using the kernel estimators $\hat{\psi}_4(g_1)$ and $\hat{\psi}_6(g_2)$ where

$$g_1 = \{-2K^{(4)}(0)/(\mu_2(K)\hat{\psi}_6^{\text{NS}} n)\}^{1/7}$$

and

$$g_2 = \{-2K^{(6)}(0)/(\mu_2(K)\hat{\psi}_8^{\mathrm{NS}} n)\}^{1/9}.$$

Step 3 Estimate ψ_4 using the kernel estimator $\hat{\psi}_4(\gamma(h))$ where

$$\gamma(h) = \left[\frac{2K^{(4)}(0)\mu_2(K)\hat{\psi}_4(g_1)}{-\hat{\psi}_6(g_2)R(K)}\right]^{1/7} h^{5/7}.$$

Step 4 The selected bandwidth is the solution to the equation

$$h = \left[\frac{R(K)}{\mu_2(K)^2\hat{\psi}_4(\gamma(h))n}\right]^{1/5}.$$

■

3.7 Smoothed cross-validation bandwidth selection

Smoothed cross-validation (SCV) (Müller, 1985, Staniswalis, 1989a, Hall, Marron and Park, 1992) is similar to plug-in bandwidth selection in that it uses a kernel estimator with pilot bandwidth g to estimate the integrated squared bias component of $\mathrm{MISE}\{\hat{f}(\cdot;h)\}$. Because of this, the methods have similar theoretical properties. The difference is that SCV is based on the exact integrated squared bias rather than its asymptotic approximation. This has the intuitively appealing feature of having less dependence on asymptotic approximations. On the other hand, SCV is not as easy to implement as DPI and is somewhat more difficult to analyse.

In Section 2.3 we showed that

$$\begin{aligned}\mathrm{MISE}\{\hat{f}(\cdot;h)\} = (nh)^{-1}R(K) + (1-n^{-1})\int (K_h * f)^2(x)\,dx \\ - 2\int (K_h * f)(x)f(x)\,dx + \int f(x)^2\,dx.\end{aligned}$$

Ignoring the asymptotically negligible n^{-1} in the second term, we obtain

$$\mathrm{MISE}\{\hat{f}(\cdot;h)\} \simeq (nh)^{-1}R(K) + \int (K_h * f - f)(x)^2\,dx.$$

The second term is exactly the integrated squared bias of $\hat{f}(\cdot;h)$ while the first is a good approximation to the integrated variance. The smoothed cross-validation objective function is obtained by replacing f by a pilot estimator

$$\hat{f}_L(x;g) = n^{-1}\sum_{i=1}^{n} L_g(x - X_i)$$

where $L_g(x) = L(x/g)/g$ for a possibly different kernel L and bandwidth g. This gives us

$$\mathrm{SCV}(h) = (nh)^{-1}R(K) + \widehat{\mathrm{ISB}}(h)$$

where

$$\widehat{\mathrm{ISB}}(h) = \int \{K_h * \hat{f}_L(\cdot;g) - \hat{f}_L(\cdot;g)\}^2(x)\,dx \tag{3.8}$$

is an estimate of integrated squared bias (ISB). The SCV bandwidth $\hat{h}_{\mathrm{SCV}}$ is defined to be the largest local minimiser of $\mathrm{SCV}(h)$. It is a straightforward exercise (Exercise 3.8) to show that $\widehat{\mathrm{ISB}}(h)$ has the more explicit formulation:

$$\begin{aligned}\widehat{\mathrm{ISB}}(h) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(&K_h * K_h * L_g * L_g \\ &- 2 * K_h * L_g * L_g + L_g * L_g)(X_i - X_j).\end{aligned} \tag{3.9}$$

The SCV bandwidth selector has some interesting alternative motivations (Hall, Marron and Park, 1992). The first is that SCV is an adjustment of least squares cross-validation that allows "presmoothing" of the pairwise differences $X_i - X_j$. This is the reason for the method's name and can be seen by noting that if there are no replications among the data then

$$\begin{aligned}\mathrm{LSCV}(h) &= (nh)^{-1}R(K) \\ &+ n^{-1}(n-1)^{-1}\sum\sum_{i\neq j}(K_h * K_h - 2K_h + K_0)(X_i - X_j)\end{aligned}$$

where K_0 is the Dirac delta function. A variation of $\mathrm{SCV}(h)$ is one where the integrated squared bias is instead estimated by

$$\widetilde{\mathrm{ISB}}(h) = n^{-1}\sum_{i=1}^{n}\int \{K_h * \hat{f}_{L,-i}(x;g) - \hat{f}_{L,-i}(x;g)\}^2$$

where $\hat{f}_{L,-i}(\cdot; g)$ is the leave-one-out kernel density estimator based on kernel L and bandwidth g. Then

$$\widetilde{\mathrm{ISB}}(h) = \{n(n-1)\}^{-1} \sum \sum_{i \neq j} [\{K_h * K_h - 2K_h + K_0\} * L_g * L_g](X_i - X_j).$$

Thus, this version of SCV estimates ISB in the same way as LSCV except that the function $L_g * L_g$ is applied to the $X_i - X_j$. This also makes it clear that LSCV is a special case of the leave-one-out version of SCV with $g = 0$.

SCV can also be motivated through the smoothed bootstrap. Let $X_1^*, \ldots, X_n^*$ be a sample from the density $\hat{f}_L(\cdot; g)$ conditional on the original sample $X_1, \ldots, X_n$. The bootstrap estimate of $\mathrm{MISE}\{\hat{f}(\cdot; h)\}$ is then

$$\mathrm{MISE}^*\{\hat{f}^*(\cdot; h)\} = E^* \int \{\hat{f}^*(x; h) - \hat{f}_L(x; g)\}^2 \, dx$$

where

$$\hat{f}^*(x; h) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i^*)$$

and E^* denotes expectation conditional on $X_1, \ldots, X_n$. However, one can show (Exercise 3.8) that

$$\begin{aligned} \mathrm{MISE}^*\{\hat{f}^*(\cdot; h)\} &= \mathrm{SCV}(h) \\ &\quad + n^{-1} \int \{K_h * \hat{f}_L(\cdot; g)(x)\}^2 \, dx. \end{aligned} \tag{3.10}$$

The second term is asymptotically negligible with respect to the first, so choosing h to minimise $\mathrm{MISE}^*\{\hat{f}^*(\cdot; h)\}$ is virtually equivalent to choosing the bandwidth that minimises $\mathrm{SCV}(h)$.

Alternative approaches to SCV, based on the frequency domain version of MISE, have also been proposed and analysed (Chiu, 1991, 1992).

Considerable theory has been devoted to the choice of the pilot bandwidth g in the SCV objective function (Jones, Marron and Park, 1991, Park and Marron, 1992), and this involves precisely the same considerations as were necessary for the selection of a pilot bandwidth of a plug-in bandwidth selector. There are good asymptotic reasons for allowing g to have dependence on h of the form

$$g = Cn^p h^m$$

for constants C, p and m. Optimal choice of these constants, aimed at enhancing the asymptotic performance of $\hat{h}_{\mathrm{SCV}}$, will be discussed in Section 3.8. (In parallel with this, direct plug-in rules take $(m,p) = (0, \frac{1}{7})$ and solve-the-equation rules take $(m,p) = (\frac{5}{7}, 0)$.) Asymptotically optimal choice of C depends on f through ψ_r functionals, so these can be estimated in the same way as was done for plug-in bandwidth selection algorithms. A smoothed cross-validation bandwidth selector that involves ℓ successive functional estimation steps, with an initial bandwidth chosen by a quick and simple rule, will be denoted by $\hat{h}_{\mathrm{SCV},\ell}$.

The following example describes one particular version of $\hat{h}_{\mathrm{SCV},2}$.

EXAMPLE. Let $K = L = \phi$, the normal kernel.

Step 1 Compute kernel estimates $\hat{\psi}_6(g_1)$ and $\hat{\psi}_{10}(g_2)$ where

$$g_1 = \{2/(7n)\}^{1/9} 2^{1/2} \hat{\sigma} \quad \text{and} \quad g_2 = \{2/(11n)\}^{1/13} 2^{1/2} \hat{\sigma}$$

are based on normal scale estimates of ψ_8 and ψ_{12}; see (3.7).

Step 2 Compute kernel estimates $\hat{\psi}_4(g_3)$ and $\hat{\psi}_8(g_4)$ where

$$g_3 = [-6/\{(2\pi)^{1/2} \hat{\psi}_6(g_1) n\}]^{1/7}$$

and

$$g_4 = [-210/\{(2\pi)^{1/2} \hat{\psi}_{10}(g_2) n\}]^{1/11}.$$

Step 3 Choose h to minimise

$$\mathrm{SCV}(h) = (nh)^{-1}(2\pi^{1/2})^{-1} + \sum_{i=1}^{n} \sum_{j=1}^{n} \{\phi_{(2h^2+2g^2)^{1/2}} - 2\phi_{(h^2+2g^2)^{1/2}} + \phi_{(2g^2)^{1/2}}\}(X_i - X_j)$$

where

$$g = \hat{C} n^{-23/45} h^{-2}$$

and

$$\hat{C} = \{441/(64\pi)\}^{1/18} (4\pi)^{-1/5} \hat{\psi}_4(g_3)^{-2/5} \hat{\psi}_8(g_4)^{-1/9}.$$

∎

In the above example we took both K and L to be the normal kernel to keep the presentation simple. Note that the identity

$$(\phi_\sigma * \phi_{\sigma'})(x) = \phi_{(\sigma^2+\sigma'^2)^{1/2}}(x)$$

(Appendix C) has been used to simplify the convolution. An explanation for the particular choices of $\hat{C}$, p and m used in this example is given in the next section.

3.8 Comparison of bandwidth selectors

As the previous four sections indicate, there now exist several hi-tech rules for selecting the bandwidth of a kernel density estimator from the data. This leads to the natural question: how do these methods compare? Some insight into this can be obtained through asymptotic analysis of each bandwidth selector which typically leads to a "rate of convergence" of the selected bandwidth to some optimal bandwidth. While such a theoretical result gives some indication of the relative merits of competing bandwidth selectors, it does not necessarily indicate what happens in practice since the asymptotics often do not take effect for smaller sample sizes. Therefore, computer simulation has also become an important tool for the comparison of bandwidth selectors.

3.8.1 *Theoretical performance*

The usual way of quantifying the theoretical performance of a particular bandwidth selector $\hat{h}$ is through an asymptotic distribution result, typically of the form

$$n^\nu(\hat{h}/h_0 - 1) \to_{\mathrm{D}} N(\mu, \sigma^2) \tag{3.11}$$

where μ and $\sigma^2 > 0$ depend only on f and K (but not on n) and h_0 is some "optimal" bandwidth. The notation

$$A_n \to_{\mathrm{D}} N(\mu, \sigma^2)$$

means that the sequence of random variables A_n converges in distribution to a $N(\mu, \sigma^2)$ random variable (see Appendix A). The quantity $\hat{h}/h_0 - 1$ is called the *relative error* of $\hat{h}$ and takes into account the fact that the bandwidth is a scale parameter. If $\hat{h}$ satisfies (3.11) then we will say that $\hat{h}$ has a *relative rate*

of convergence to h_0 of order $n^{-\nu}$ with asymptotic variance σ^2. Therefore, larger values of ν correspond to the bandwidth selectors converging to the optimum at faster rates. In addition, smaller values of σ^2 correspond to the bandwidth selectors being more stable.

An important question is that of the most appropriate choice for the "optimal" bandwidth h_0. Since each of the bandwidths in this section is based on minimising an estimate of $\mathrm{MISE}\{\hat{f}(\cdot;h)\}$, a natural first answer might be $h_0 = h_{\mathrm{MISE}}$. Observe, however, that h_{MISE} is the bandwidth minimising a quantity that is *averaged over all possible samples* while the optimal bandwidth for the *sample at hand* (with respect to squared error loss) is h_{ISE}, the bandwidth that minimises

$$\mathrm{ISE}\{\hat{f}(\cdot;h)\} = \int \{\hat{f}(x;h) - f(x)\}^2 \, dx.$$

In asymptotic distribution terms, the choices $h_0 = h_{\mathrm{MISE}}$ and $h_0 = h_{\mathrm{ISE}}$ give quite different results. One notable manifestation of this is through minimax results that state that the relative rate of convergence to h_{ISE} of *any* bandwidth selector cannot be faster than $n^{-1/10}$ while the relative rate of convergence of bandwidth selectors to h_{MISE} can be as fast as $n^{-1/2}$ (Hall and Marron, 1991b). These results indicate that the goal of minimising $\mathrm{ISE}\{\hat{f}(\cdot;h)\}$ is much more difficult than that of minimising $\mathrm{MISE}\{\hat{f}(\cdot;h)\}$, due to the fact that h_{ISE} is, itself, a random variable. For essentially this reason (see also Jones, 1991a, and Grund, Hall and Marron, 1995) we will take $h_0 = h_{\mathrm{MISE}}$ for our theoretical comparison.

Least squares cross-validation and biased cross-validation

Under certain regularity conditions the LSCV bandwidth selector satisfies

$$n^{1/10}(\hat{h}_{\mathrm{LSCV}}/h_{\mathrm{MISE}} - 1) \rightarrow_{\mathrm{D}} N(0, \sigma^2_{\mathrm{LSCV}})$$

(Hall and Marron, 1987a; Scott and Terrell, 1987) while the BCV bandwidth selector satisfies

$$n^{1/10}(\hat{h}_{\mathrm{BCV}}/h_{\mathrm{MISE}} - 1) \rightarrow_{\mathrm{D}} N(0, \sigma^2_{\mathrm{BCV}})$$

(Scott and Terrell, 1987). Here σ^2_{LSCV} and σ^2_{BCV} depend on functionals of f and K, but not on n. The ratio of these two asymptotic variances for the standard normal kernel is

$$\sigma^2_{\mathrm{LSCV}}/\sigma^2_{\mathrm{BCV}} \simeq 15.7.$$

This indicates that we should expect $\hat{h}_{\mathrm{LSCV}}$ to be considerably more variable than $\hat{h}_{\mathrm{BCV}}$, which is a theoretical explanation of the phenomenon observed in Figure 3.3.

The above results also show that both $\hat{h}_{\mathrm{LSCV}}$ and $\hat{h}_{\mathrm{BCV}}$ have a relative rate of convergence to h_{MISE} of order $n^{-1/10}$, which is considerably slower than the $n^{-1/2}$ lower bound. An understanding of what causes this poor asymptotic performance can be most easily gained by a comparison of BCV and a direct plug-in bandwidth selector. Recall that $\hat{h}_{\mathrm{BCV}}$ minimises

$$\mathrm{BCV}(h) = (nh)^{-1}R(K) + \tfrac{1}{4}h^4\mu_2(K)^2\widetilde{\psi}_4(h),$$

while a slight variant of the direct plug-in bandwidth selector, which we will call $\tilde{h}_{\mathrm{DPI}}$, minimises

$$\mathrm{DPI}(h) = (nh)^{-1}R(K) + \tfrac{1}{4}h^4\mu_2(K)^2\widetilde{\psi}_4(g),$$

where g is independent of h. Here

$$\widetilde{\psi}_4(g) = n^{-1}(n-1)^{-1}\sum\sum_{i\neq j}(K_g'' * K_g'')(X_i - X_j)$$

is a "leave-out-diagonals" estimate of ψ_4. The two methods differ in their choice of the pilot bandwidth g. The selector $\hat{h}_{\mathrm{BCV}}$ overcomes this choice by setting $g = h$ while $\tilde{h}_{\mathrm{DPI}}$ allows for arbitrary choice of g. Theory similar to that presented in Section 3.5 shows that optimal choice of g, with respect to asymptotic $\mathrm{MSE}\{\widetilde{\psi}_4(g)\}$, is of order $n^{-2/13}$ whereas the h used by BCV is asymptotically of order $n^{-1/5}$. This suboptimal choice of h for estimation of the integrated squared bias approximation is the reason for the asymptotic deficiency of BCV. Similar arguments can be used to explain the poor asymptotic performance of LSCV.

Plug-in and smoothed cross-validation selectors

As the preceding discussion suggests, the extra flexibility due to having a pilot bandwidth for estimation of integrated squared bias allows for considerable improvement of the asymptotic relative error of plug-in and smoothed cross-validation bandwidth selectors. We will now study the effect of g on the relative error of $\hat{h}_{\mathrm{DPI}}$. This will be followed by similar results for $\hat{h}_{\mathrm{SCV}}$.

The direct plug-in bandwidth selector can be written

$$\hat{h}_{\mathrm{DPI}} = \alpha(K)\hat{\psi}_4(g)^{-1/5}n^{-1/5}$$

where $\alpha(K) = \{R(K)/\mu_2(K)^2\}^{1/5}$. An analysis of the errors involved in the approximation of h_{MISE} by h_{AMISE} leads to

$$h_{\text{MISE}} = \alpha(K)\psi_4^{-1/5} n^{-1/5} + O(n^{-3/5}).$$

The relative error of $\hat{h}_{\text{DPI}}$ is therefore

$$\hat{h}_{\text{DPI}}/h_{\text{MISE}} - 1 = \psi_4^{-1/5}\{\hat{\psi}_4(g)^{-1/5} - \psi_4^{-1/5}\} + O_P(n^{-2/5})$$

where O_P denotes *order in probability* (see Appendix A) and takes into account the randomness of $\hat{\psi}_4(g)$. The $O_P(n^{-2/5})$ term is due to the relative error of h_{AMISE} as an approximation to h_{MISE}. Formal application of Taylor's theorem leads to

$$\hat{h}_{\text{DPI}}/h_{\text{MISE}} - 1 = -\tfrac{1}{5}\psi_4^{-1}\{\hat{\psi}_4(g) - \psi_4\} + \ldots + O_P(n^{-2/5}).$$

This expression shows that the relative error of $\hat{h}_{\text{DPI}}$ depends on the functional estimation error $\hat{\psi}_4(g) - \psi_4$ as well as the relative approximation error of h_{AMISE}. Either term can dominate, depending on the choice of L and g. For the bandwidth selector $\hat{h}_{\text{DPI},2}$ given in the example of Section 3.6.1, which is based on $L = K$ being a second-order kernel with g chosen to optimise $\text{AMSE}\{\hat{\psi}_4(g)\}$, it can be shown that $\hat{\psi}_4(g) - \psi_4$ is $O_P(n^{-5/14})$ which dominates the $O_P(n^{-2/5})$ term. More involved and rigorous arguments can be used to show that we in fact have

$$n^{5/14}(\hat{h}_{\text{DPI},2}/h_{\text{MISE}} - 1) \to_D N(0, \sigma^2_{\text{DPI}})$$

so this particular version of $\hat{h}_{\text{DPI}}$, with a relative convergence rate of order $n^{-5/14}$, is considerably closer to the $n^{-1/2}$ lower bound than the $n^{-1/10}$ rate of cross-validation methods. The same result can also be shown to hold for $\hat{h}_{\text{STE},2}$.

The role of the pilot bandwidth g for $\hat{h}_{\text{SCV}}$ is analogous to that for $\hat{h}_{\text{DPI}}$, although theory for optimal choice of g is not as simple to derive. Asymptotic distributional results can be used to provide insight into appropriate choice of g. Let $\hat{h}_{\text{SCV}}$ be the minimiser of $\text{SCV}(h)$, where $g = Cn^p h^m$. Also, let $g_{\text{MISE}} = Cn^p h^m_{\text{MISE}}$. If K and L are both second-order kernels then, under appropriate regularity conditions, we have

$$\begin{aligned}\hat{h}_{\text{SCV}}/h_{\text{MISE}} - 1 &= (n^{-4/5}h^6_{\text{MISE}}g^{-9}_{\text{MISE}}C_1 + n^{-1}C_2)^{1/2}Z \\ &\quad + (-n^{3/5}h^3_{\text{MISE}}g^2_{\text{MISE}}C_3 + n^{3/5}h^3_{\text{MISE}}g^4_{\text{MISE}}C_4 \\ &\quad + n^{-2/5}h^3_{\text{MISE}}g^{-5}_{\text{MISE}}C_5)\end{aligned} \tag{3.12}$$

(Jones, Marron and Park, 1991), where Z is asymptotically $N(0,1)$. An expression of this type also exists for $\hat{h}_{\mathrm{DPI}}$ and $\hat{h}_{\mathrm{STE}}$, which confirms that the choice of pilot bandwidth is analogous. The constants $C_1, \ldots, C_5$ depend on functionals of K and L, and on f only through ψ_r density functionals, an exception being C_2 which also depends on $\int \{f^{(4)}(x)\}^2 f(x)\,dx$. Of particular relevance is the expression for C_3:

$$C_3 = \tfrac{1}{10}(m+2)\left[\frac{-\mu_2(K)^6\mu_2(L)^5\psi_6^5}{R(K)^3\psi_4^2}\right]^{1/5}.$$

This is because the choice $m = -2$ leads to $C_3 = 0$. The first term on the right hand side of (3.12) can be thought of as representing the variation of $\hat{h}_{\mathrm{SCV}}$, while the second represents bias. This shows how choice of g represents a type of variance-bias trade-off. Bias is decreased by having $g \to 0$, but this decrease should not be too rapid, otherwise the variance is increased. A simple means of determining the optimal choice of g is to combine the asymptotic variance and bias terms to obtain an "asymptotic relative mean squared error (ARMSE)",

$$\begin{aligned}\mathrm{ARMSE} &= n^{-4/5}h_{\mathrm{MISE}}^6 g_{\mathrm{MISE}}^{-9} C_1 + n^{-1}C_2 \\ &\quad + (-n^{3/5}h_{\mathrm{MISE}}^3 g_{\mathrm{MISE}}^2 C_3 + n^{3/5}h_{\mathrm{MISE}}^3 g_{\mathrm{MISE}}^4 C_4 \\ &\quad + n^{-2/5}h_{\mathrm{MISE}}^3 g_{\mathrm{MISE}}^{-5} C_5)^2.\end{aligned}$$

If $m \neq -2$ then the term involving C_3 dominates that involving C_4. Choosing g to cancel this term asymptotically with the term involving C_5 leads to

$$g = \{(C_5/C_3)^{1/7}/C_0^m\}n^{m/5-1/7}h^m$$

being the asymptotically optimal choice, where

$$C_0 = [R(K)/\{\mu_2(K)^2 R(f'')\}]^{1/5}$$

and arises from the fact that $h_{\mathrm{MISE}} \sim C_0 n^{-1/5}$. If $m = 0$ (so g is independent of h) we simply have

$$g = (C_5/C_3)^{1/7}n^{-1/7}$$

which is essentially the same as the optimal pilot bandwidth g_2 for the direct plug-in bandwidth selector described in the example of

Section 3.6.1. The resulting rate of convergence of $\hat{h}_{\mathrm{SCV}}$ is given by

$$n^{5/14}(\hat{h}_{\mathrm{SCV}}/h_{\mathrm{MISE}} - 1) \rightarrow_{\mathrm{D}} N(0, \sigma^2_{\mathrm{SCV}}).$$

Root-n bandwidth selection

It is possible to construct plug-in and smooth cross-validation rules that achieve the optimal $n^{-1/2}$ relative rate of convergence. Such rules are called *root-n bandwidth selectors.*

One of the simplest root-n bandwidth selectors is based on the following two term approximation to h_{MISE},

$$h_{\mathrm{AMISE},2} = \left[\frac{R(K)}{\mu_2(K)^2\psi_4 n}\right]^{1/5} + \frac{\mu_4(K)\psi_6}{20}\left[\frac{R(K)^3}{\mu_2(K)^{11}\psi_4^8 n^3}\right]^{1/5}.$$

The relative error of $h_{\mathrm{AMISE},2}$ is of order $n^{-3/5}$, compared to the $n^{-2/5}$ relative rate of the usual one term approximation, h_{AMISE}. Thus, if higher-order kernels are used to estimate ψ_4 and ψ_6 with mean squared error of $O(n^{-1})$ then the resulting plug-in bandwidth selector achieves root-n performance (Hall, Sheather, Jones and Marron, 1991).

The root-n relative rate can also be achieved by judicious choice of the parameters via the $g = Cn^p h^m$ factorization of the pilot bandwidth. We will describe this in the SCV context, although similar results are obtainable for plug-in bandwidth selection. If $m = -2$ then C_3 as defined above vanishes and the bias term is asymptotically zeroed by

$$g = (-C_5/C_4)^{1/9} C_0^2 n^{-23/45} h^{-2}.$$

This leads to

$$n^{1/2}(\hat{h}_{\mathrm{SCV}}/h_{\mathrm{MISE}} - 1) \rightarrow_{\mathrm{D}} N(0, \sigma^2_{\mathrm{SCV}})$$

(with a different value for σ^2_{SCV} than above) (Jones, Marron and Park, 1991). An appealing feature of this result is that the $n^{-1/2}$ rate can be obtained without the use of higher-order kernels at any stage. The example of Section 3.7 describes the rule given by this strategy when the normal kernel is used at all stages.

Root-n performance has also been established for the frequency domain approach to SCV (Chiu, 1991, 1992).

3.8.2 *Practical advice*

The asymptotic results of the previous section need to be viewed with some caution. Apart from requiring that the sample size be sufficiently large they also have the defect of often masking the choice of various auxiliary parameters, such as the choice of scale estimate for a normal scale rule, or the number of stages of a plug-in strategy. These parameters can have a significant effect on the performance of a bandwidth selector in practice.

The main tool for assessing the practical performance of a bandwidth selector is simulation. Figure 3.4, for example, shows that important insight into the effect of the number of stages of a plug-in rule can be obtained from simulation. Figure 3.5 provides a similar comparison of the selectors $\hat{h}_{\mathrm{LSCV}}$, $\hat{h}_{\mathrm{BCV}}$ (with K equal to the standard normal kernel) and the versions of $\hat{h}_{\mathrm{DPI},2}$ and $\hat{h}_{\mathrm{SCV},2}$ given in the examples of Section 3.6.1 and Section 3.7 respectively. The sample size is $n = 100$ and the underlying density is f_1 as defined at (2.3).

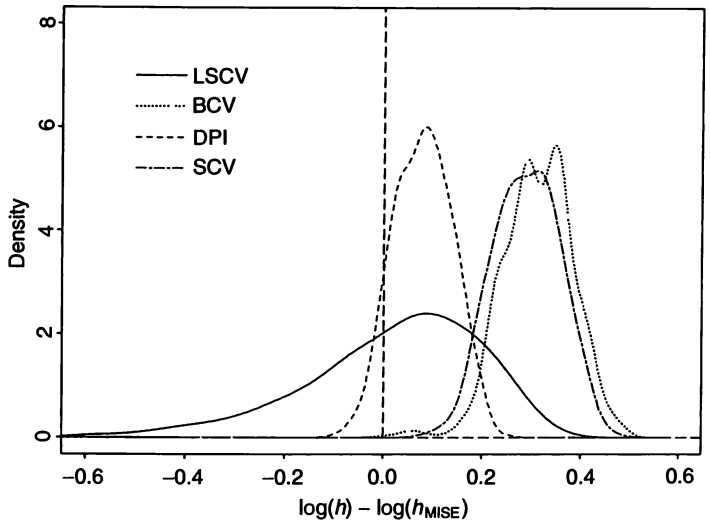

Figure 3.5. *Density estimates based on values of* $\log_{10}(\hat{h}) - \log_{10}(h_{\mathrm{MISE}})$ *for several bandwidth selectors (described in the text). Selected bandwidths are based on 500 simulated samples of size* $n = 100$ *from the normal mixture density* f_1 *defined at (2.3).*

For this particular setting we see that $\hat{h}_{\mathrm{DPI},2}$ provides the best compromise between bias and variance among these four selectors.

It is difficult to give a concise summary of simulation results

in general since the rankings of the selectors change for different densities (see e.g. Park and Turlach (1992), Cao, Cuevas and González-Manteiga (1994), Jones, Marron and Sheather (1992)), although certain patterns have emerged. As the asymptotics suggest, versions of hi-tech selectors involving pilot bandwidths, such as $\hat{h}_{\mathrm{DPI}}$ and $\hat{h}_{\mathrm{SCV}}$, perform quite well for many density types (e.g. Park and Turlach, 1992). However, for densities with sharp features, such as several modes, the asymptotics that these methods rely upon tend to be less accurate and the performance can be worse than a non-asymptotic method. LSCV does not depend on asymptotic arguments, but its sample variability is usually considered to be too high to be of reliable practical use (e.g. Jones, Marron and Sheather, 1992). The simulation performance of BCV has also been disappointing and we cannot recommend this bandwidth selector for general use.

In summary, while considerable recent progress has been made in the development towards high-performance bandwidth selectors, no rule comes with a guarantee that it will work satisfactorily in all cases. A sensible data analytic strategy is that of obtaining estimates for a variety of bandwidths, perhaps obtained from a variety of bandwidth selectors and choices of auxiliary parameters. If a single objective bandwidth selector is required then our recommendation, based on simulation evidence, is to use a version of $\hat{h}_{\mathrm{DPI}}$, $\hat{h}_{\mathrm{STE}}$ or $\hat{h}_{\mathrm{SCV}}$, rather than $\hat{h}_{\mathrm{LSCV}}$ or $\hat{h}_{\mathrm{BCV}}$.

3.9 Bibliographical notes

3.2 For a discussion of normal scale rules based on simple scale measures see Silverman (1986, pp.45–47). Bowman (1985) showed their usefulness for normal-like densities. Janssen, Marron, Veraverbeke and Sarle (1995) developed improved scale measures for use in bandwidth selection. Oversmoothed bandwidth selection methods were proposed and discussed in Terrell and Scott (1985) and Terrell (1990). Sheather (1992) applied several bandwidth selectors to the Old Faithful data.

3.3 Rudemo (1982) and Bowman (1984) independently derived LSCV. Notable contributions to the theory of LSCV include Hall (1983), Stone (1984) and Hall and Marron (1987a). Related cross-validatory smoothing parameter selectors for Fourier series density estimates were proposed by Kronmal and Tarter (1968) and Wahba (1981).

3.4 The BCV bandwidth selector was proposed by Scott and