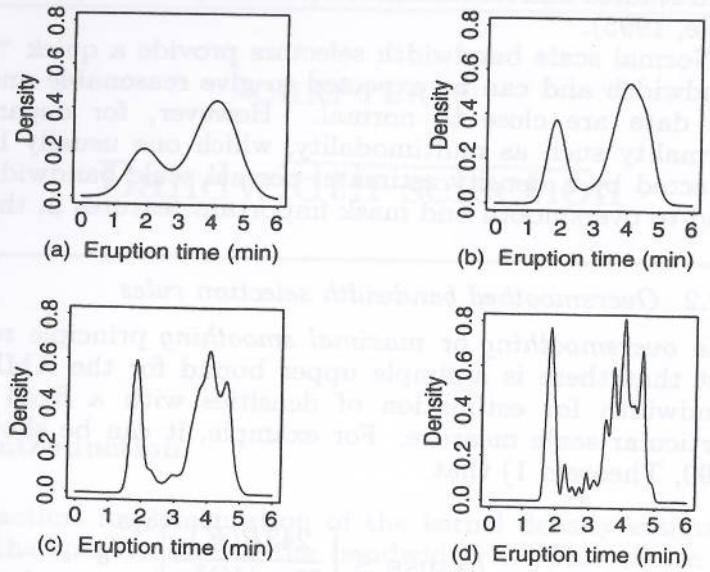

Figure 3.1. *Density estimates based on 107 eruption times (minutes) of the Old Faithful Geyser. The normal kernel is used. Bandwidths are (a) $\hat{h}_{\mathrm{OS}}$, (b) $\hat{h}_{\mathrm{OS}}/2$, (c) $\hat{h}_{\mathrm{OS}}/4$ and (d) $\hat{h}_{\mathrm{OS}}/8$.*

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

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

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

$$\mathrm{MISE}\{\hat{f}(\cdot;h)\} - \int f(x)^2\,dx = E\left[\int \hat{f}(x;h)^2\,dx - 2\int \hat{f}(x;h)f(x)\,dx\right].$$

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

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

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.

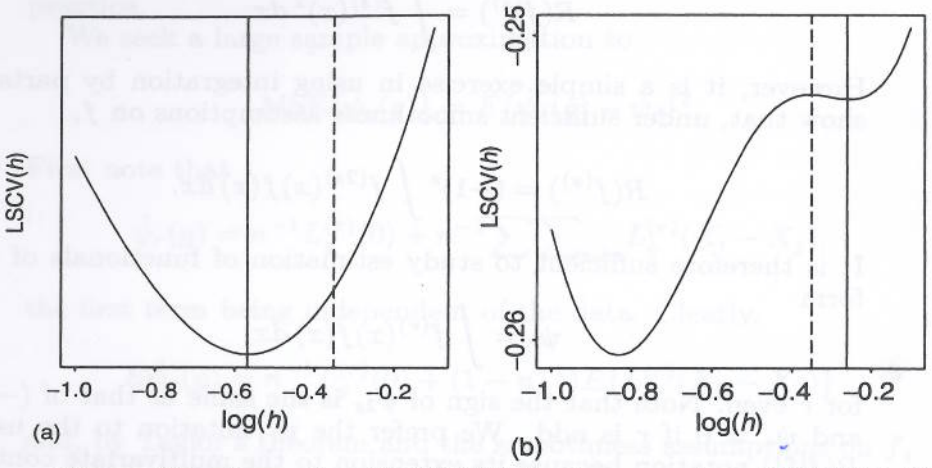

Figure 3.2. *Examples of* LSCV(h) *for two samples of* 100 $N(0,1)$ *observations. A* $\log_{10}$ *scale is used on the horizontal axis. The dashed vertical line shows the position of* $\log_{10}(h_{\text{MISE}})$. *The solid vertical lines show the position of* $\log_{10}(\hat{h}_{\text{LSCV}})$ *if*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

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:

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

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

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

The BCV bandwidth selector, which we denote by $\hat{h}_{\text{BCV}}$, is the minimiser of 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 $\hat{h}_{\text{BCV}}$ is that it is more stable than $\hat{h}_{\text{LSCV}}$, 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 $\hat{h}_{\text{BCV}}$ tending to be larger than the MISE-optimal bandwidth. This is illustrated in Figure 3.3, which shows kernel density estimates based on $\hat{h}_{\text{LSCV}}$ and $\hat{h}_{\text{BCV}}$ 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 $\log_{10}(h_{\text{MISE}})$ 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 $\hat{h}_{\text{LSCV}}$ and $\hat{h}_{\text{BCV}}$ 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.