The Annals of Statistics
1999, Vol. 27, No. 2, 415–438

BANDWIDTH SELECTION: CLASSICAL OR PLUG-IN?

BY CLIVE R. LOADER

Lucent Technologies

Bandwidth selection for procedures such as kernel density estimation and local regression have been widely studied over the past decade. Substantial "evidence" has been collected to establish superior performance of modern plug-in methods in comparison to methods such as cross validation; this has ranged from detailed analysis of rates of convergence, to simulations, to superior performance on real datasets.

In this work we take a detailed look at some of this evidence, looking into the sources of differences. Our findings challenge the claimed superiority of plug-in methods on several fronts. First, plug-in methods are heavily dependent on arbitrary specification of pilot bandwidths and fail when this specification is wrong. Second, the often-quoted variability and undersmoothing of cross validation simply reflects the uncertainty of bandwidth selection; plug-in methods reflect this uncertainty by oversmoothing and missing important features when given difficult problems. Third, we look at asymptotic theory. Plug-in methods use available curvature information in an inefficient manner, resulting in inefficient estimates. Previous comparisons with classical approaches penalized the classical approaches for this inefficiency. Asymptotically, the plug-in based estimates are beaten by their own pilot estimates.

1. Introduction. The problem of automatic choice of smoothing parameters has been widely studied. The work has been most predominantly in the setting of kernel density estimation with a single fixed bandwidth, so we initially focus on that setting. The bandwidth selection methods studied in the literature can be divided into two broad classes.

Classical methods. Cross validation, Mallows' C_p, Akaike's information criterion and the like. These are more or less natural extensions of methods used in parametric modeling.

Plug-in methods. The bias of an estimate $\hat{f}$ is written as a function of the unknown f, and usually approximated through Taylor series expansions. A pilot estimate of f is then "plugged in" to derive an estimate of the bias and hence an estimate of mean integrated squared error. The "optimal" h minimizes this estimated measure of fit.

More complete descriptions of these approaches are given in Section 3 for density estimation and Section 6 for local regression.

In the context of kernel density estimation, the plug-in approach appears to predate "classical" approaches, dating to Woodroofe (1970). However, more

Received December 1995; revised December 1998.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62:07, 62-09, 62G20.

Key words and phrases. Akaike's information criterion, bandwidth, cross validation, density estimation, local fitting, local likelihood, plug-in.

specific algorithms and the strong promotion of the approach began in the mid-1980s, and continues with increasing vigor. Proponents of the plug-in approach have been strongly critical of classical approaches. For example, Park and Marron (1990) state:

> In many simulation studies and real data examples, however, the performance of (least squares cross-validation) has been often disappointing ...

and continue:

> Because of the limitations of least squares cross-validation, there has been serious investigation made into other methods of bandwidth selection. The most appealing of these are plug-in rules and biased cross-validation.

Similarly strong comments are made by other authors; for example, Ruppert, Sheather and Wand (1995) and Marron [(1996), Section 3]. The evidence presented to back up these claims is threefold: real data examples [Sheather (1992), Jones, Marron and Sheather (1996)], simulation studies [Park and Marron (1990), Park and Turlach (1992), Scott and Terrell (1987), Gasser, Kneip and Köhler (1991)] and asymptotic theory [Hall, Sheather, Jones and Marron (1991), Chiu (1991), Härdle, Hall and Marron (1992)].

In this paper we take a detailed look at this evidence. We find the evidence for superior performance of plug-in approaches is far less compelling than previously claimed. In turn, we consider real data examples, simulation studies and asymptotics. Among the findings are that plug-in approaches are tuned by arbitrary specification of pilot estimators and are prone to oversmoothing when presented with difficult smoothing problems.

The purpose of this paper is not simply *comparison*, but *understanding* bandwidth selectors. Thus we concentrate on a fairly small number of examples and investigate how the bandwidth selectors perform in relation to the datasets and difficulty of the problems at hand. A complete understanding of this paper requires careful interpretation of the following question: What makes bandwidth selection difficult?

Consider for example Figure 1. This shows a simulated dataset, fitted with a locally quadratic smooth and two different bandwidths; the small bandwidth on the left is selected by a classical approach; the larger bandwidth on the right by a plug-in approach. The model and bandwidth selectors will be described in Section 6.

Visually, the plug-in fit in Figure 1 is preferable; it captures the main trend in the data and has far less spurious noise. However, there is also another possible feature; near $x = 0.6$, there are several successive large observations that do not fit the underlying pattern.

Usually, we hope that nature isn't too nasty, and faced with a real dataset of this type, most statisticians would conclude that the blip at $x = 0.6$ is due to random chance, and the left panel of Figure 1 is seriously undersmoothed. However, a bandwidth selector has to make its decision purely from the data: are these observations sufficient to represent a real feature? Clearly, this is a difficult decision, and any bandwidth selector is occasionally going to make

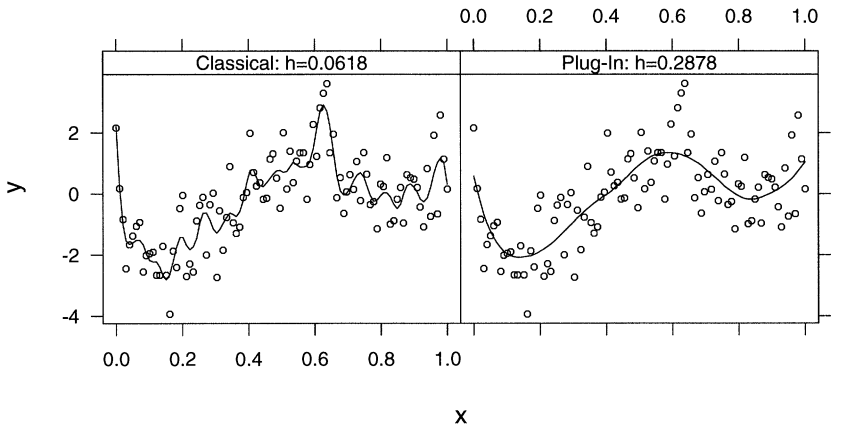

FIG. 1. *Local quadratic smooths of a simulated dataset, with two different bandwidths.*

mistakes. What we can expect is a meaningful assessment of the dataset at hand; the bandwidth selector should say that there is little to choose between the two fits in Figure 1.

An often repeated criticism of classical approaches is that they are too variable and frequently undersmooth. If repeated samples are drawn from the same model, cross validation (and other classical approaches) can select bandwidths that are very different from sample to sample. See, for example, Section 3 of Marron (1996) or Gasser, Kneip and Köhler [(1991), page 643]. But in light of examples such as Figure 1, this behavior is to be expected; a bandwidth selector has to make a decision as to what features in the dataset are real. We argue that variability of cross validation is not a problem but a symptom of the difficulty of bandwidth selection. Less variable bandwidth selectors display this difficulty in another way: consistently oversmoothing when presented with problems with small and difficult to detect features.

This paper is organized as follows: Some density estimates and bandwidth selectors are introduced in Sections 2 and 3, respectively. Some examples, both real and simulated, are presented in Section 4. The relevance of asymptotic theory is discussed in Section 5. Local regression examples are presented in Section 6. The main conclusions of the paper are summarized in Section 7. The simulations and fits in this paper were obtained using the author's LOCFIT software package; further details can be found at `http://cm.bell-labs.com/stat/project/locfit`.

2. Some density estimates. Let $X_1, \ldots, X_n$ be an independent sample from an unknown density $f(x)$. The kernel estimate [Rosenblatt (1956)] of $f(x)$ is

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} W\left(\frac{X_i - x}{h}\right) \tag{1}$$

for a suitable weight function $W(u) \geq 0$, with $\int_{-\infty}^{\infty} W(u)\,du = 1$. An alternative representation is that $\hat{f}(x)$ is the solution $\hat{a}$ of the equation

$$\frac{1}{n}\sum_{i=1}^{n} W\left(\frac{X_i - x}{h}\right) = \int_{-\infty}^{\infty} W\left(\frac{u-x}{h}\right) a\,du.$$

Thus the kernel estimate is a locally constant approximation, matching a weighted zeroth order sample moment with the corresponding moment of the local estimate.

Better estimates can be obtained by replacing the locally constant approximation with local parametric approximations; see Hjort and Jones (1996). As a specific example, let $A(v) = (1 \quad v \quad \frac{1}{2}v^2)^T$, and consider solutions $\hat{a} = \hat{a}(x)$ of the equations

$$\frac{1}{n}\sum_{i=1}^{n} W\left(\frac{X_i - x}{h}\right) A(X_i - x) = \int_{-\infty}^{\infty} W\left(\frac{u-x}{h}\right) A(u-x)\tilde{f}(u,a)\,du$$

where $\tilde{f}(u,a)$ is a locally three-parameter approximation. The density estimate is $\hat{f}(x) = \tilde{f}(x, \hat{a}(x))$. Using the locally quadratic approximation

$$\tilde{f}(u,a) = \langle a, A((u-x))\rangle = a_0 + a_1(u-x) + \frac{a_2}{2}(u-x)^2 \tag{2}$$

gives rise to the fourth-order kernel estimate [Lejeune and Sarda (1992)].

Another approximation is the locally log-quadratic $\tilde{f}(u,a) = \exp(\langle a, A(u-x)\rangle)$, which leads to the local likelihood method of Loader (1996a). This method has some significant advantages over the higher order kernel methods; it necessarily produces nonnegative estimates and is better for estimating the tails of the density.

3. Some bandwidth selectors. The original cross-validation criterion, proposed by Habbema, Hermans and Van Der Broek (1974) and Duin (1976), selects the bandwidth h that maximizes

$$\mathrm{LCV}(h) = \prod_{i=1}^{n} \hat{f}_{h,-i}(X_i),$$

where $\hat{f}_{h,-i}(X_i)$ denotes the density estimate with X_i deleted. For kernel density estimation, this is equivalent to minimizing

$$\mathrm{LCV}(h) = -\sum_{i=1}^{n} \log \hat{f}_h(X_i) - \sum_{i=1}^{n} \log\bigl(1 - W(0)/(nh\hat{f}_h(X_i))\bigr).$$

An approximation to the LCV criterion is the Akaike-style criterion,

$$\mathrm{AIC}(h) = -\sum_{i=1}^{n} \log \hat{f}_h(X_i) - \sum_{i=1}^{n} \mathrm{infl}(X_i)$$

where $\mathrm{infl}(X_i)$ measures the sensitivity of the density estimate when X_i is changed; $\mathrm{infl}(x) = W(0)/(nh\hat{f}(x))$ for kernel density estimation. See Loader (1996b) for more details.

The use of LCV/AIC with kernel density estimation tends to be unsatisfactory. The reason is that the likelihood criteria are very tail sensitive, where kernel estimates perform particularly poorly. The advantage of the LCV and AIC criteria is their completely general definition; they extend almost immediately to better density estimates and to other settings such as local regression and likelihood models.

For kernel density estimation, most bandwidth selectors target the integrated squared error loss function

$$\begin{aligned}\mathrm{ISE}(h) &= \int_{-\infty}^{\infty} (\hat{f}_h(x) - f(x))^2\,dx \\ &= \int_{-\infty}^{\infty} \hat{f}_h(x)^2 dx - 2\int_{-\infty}^{\infty} \hat{f}_h(x) f(x)\,dx + \int_{-\infty}^{\infty} f(x)^2\,dx.\end{aligned}$$

The term $\int_{-\infty}^{\infty} \hat{f}_h(x)^2\,dx$ depends solely on the density estimate and can be evaluated numerically. The third term $\int_{-\infty}^{\infty} f(x)^2\,dx$ does not depend on h and can be ignored. Least squares cross validation [Rudemo (1982), Bowman (1984)] then estimates the central term $\int_{-\infty}^{\infty} \hat{f}_h(x) f(x)\,dx$ by leave-one-out cross validation, giving the criterion

$$\mathrm{LSCV}(h) = \int_{-\infty}^{\infty} \hat{f}_h(x)^2\,dx - \frac{2}{n-1}\sum_{i=1}^{n}\left(\hat{f}_h(X_i) - \frac{W(0)}{nh}\right).$$

To describe plug-in selectors, we use the bias and variance approximations

$$E(\hat{f}_h(x)) - f(x) \approx \frac{h^2}{2} f''(x) \int v^2 W(v)\,dv,$$

$$\mathrm{var}\,((\,)\hat{f}_h(x)) \approx \frac{f(x)}{nh}\int W(v)^2\,dv - \frac{f(x)^2}{n};$$

see Scott (1992), page 130. The mean integrated squared error is then approximately

$$\mathrm{MISE}(h) \approx \frac{a_0(W)^2 h^4}{4}\int_{-\infty}^{\infty} f''(x)^2\,dx + \frac{a_1(W)}{nh}, \tag{3}$$

where $a_0(W) = \int_{-\infty}^{\infty} v^2 W(v)\,dv$ and $a_1(W) = \int_{-\infty}^{\infty} W(v)^2\,dv$. The asymptotically optimal bandwidth is obtained by minimizing (3),

$$h_{\mathrm{opt}} = \left(\frac{a_1(W)}{n a_0(W)^2 \int_{-\infty}^{\infty} f''(x)^2\,dx}\right)^{1/5}. \tag{4}$$

To estimate h_{opt}, one substitutes an estimate of $\int_{-\infty}^{\infty} f''(x)^2\,dx$. Usually this is derived from a "pilot" kernel estimate of the second derivative,

$$\hat{f}_k''(x) = \frac{1}{nk^3}\sum_{i=1}^{n} W''\left(\frac{X_i - x}{k}\right),$$

$$\int_{-\infty}^{\infty} \hat{f}_k''(x)^2\,dx = \frac{1}{n^2 k^6}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty} W''\left(\frac{X_i - x}{k}\right) W''\left(\frac{X_j - x}{k}\right) dx.$$

Using the standard normal kernel $\phi(x)$, this becomes

$$\int_{-\infty}^{\infty} \hat{f}_k''(x)^2\,dx = \frac{1}{n^2(\sqrt{2}k)^5} \sum_{i=1}^{n}\sum_{j=1}^{n} \phi^{(4)}\left(\frac{X_i - X_j}{\sqrt{2}k}\right). \tag{5}$$

Figure 2 shows an example for the Old Faithful geyser dataset, discussed more in Section 4. A pilot bandwidth k selects a bandwidth $h = h(k)$; this relation is shown by the solid line in Figure 2. Clearly the plug-in step alone doesn't solve much; by varying the pilot bandwidth, a wide range of bandwidths can be selected.

The most common solution to the pilot bandwidth problem is through an "assumed" relation between the pilot bandwidth and bandwidth, $k = k(h)$. For example, the fixed point iterations of Gasser, Kneip and Köhler (1991) (hereafter GKK) implicitly assume $k = n^{1/10}h$. The Sheather–Jones method [Sheather and Jones (1991)] (SJPI) uses a more complicated relation based on a reference normal model. These relations are shown by the dashed lines in Figure 2. The selected bandwidth is determined by the intersection of the actual and assumed relations.

There are many variants of the plug-in idea in the literature. First, several different ideas for specifying the pilot bandwidth k have been suggested. Second, alternative estimates of $\int_{-\infty}^{\infty} f''(x)^2\,dx$ have been considered. Third, more accurate bias approximations can be used in (3); for example, "smoothed bootstrap" and "smoothed cross-validation" selectors effectively substitute pilot estimators into the exact bias expression. These changes are relatively minor in the context of the issues raised in this paper, so we refer to Jones, Marron and Sheather (1996) for more discussion and references.

Another variant is biased cross validation [Scott and Terrell (1987)] (BCV), which takes $k = h$, and substitutes (5) (modified by deleting the $i = j$ terms)

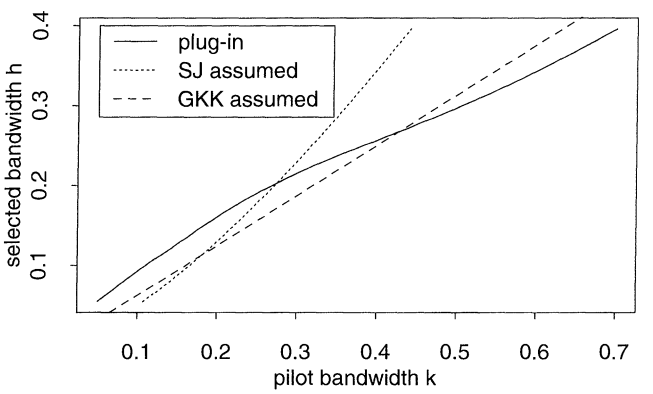

FIG. 2. *Plug-in bandwidth selection. A pilot bandwidth k selects a bandwidth h; this relation is shown by the solid lines. The dashed lines show the assumed relations for the Sheather–Jones and Gasser–Kneip–Köhler selectors.*

directly into the MISE expansion (3). The bandwidth is selected to minimize the estimated MISE.

4. Some density estimation examples. A widely used dataset in the density estimation literature consists of 107 eruption durations of the Old Faithful geyser. Azzalini and Bowman (1990) note that there are several different Old Faithful datasets and provide an interesting Markov chain analysis, as well as discussing some geological background. The dataset used here is given by Silverman (1986), Scott (1992) and others. Silverman smoothed the data using a kernel density estimate and the standard Gaussian kernel, and visually selected $h = 0.25$. Classical selectors select $h = 0.101$ (LSCV); $h = 0.0649$ (AIC) and $h = 0.126$ (LCV).

Plug-in selectors reported by Sheather (1992) include $h = 0.206$ (SJPI); $h = 0.228$ using the method of Park and Marron (1990) and $h = 0.494$ using the method of Hall, Sheather, Jones and Marron (1991). BCV selects $h = 0.282$. Chiu (1991) selects $h = 0.215$. The GKK assumed relation in Figure 2 produces $h = 0.268$.

Figure 3 shows the density estimates produced by five of these selectors. The results seem fairly clear cut: the three classical approaches (AIC, LCV and LSCV) all undersmooth and produce estimates that are far too noisy. The plug-in approaches (BCV and SJPI) are far better, smoothly reproducing the two peaks that are supported by the data. From the list above, nearly all the plug-in approaches produce h between 0.2 and 0.3; this agrees with the visual smooth of Silverman. Thus there appears to be a fairly clear consensus as to what works and what does not work on this dataset.

Of course, with real data we can't be completely sure. Instead, consider simulations from the Gaussian mixture distribution

$$f_\sigma(x) = \frac{1}{107\sigma} \sum_{i=1}^{107} \phi\left(\frac{x - X_i}{\sigma}\right),$$

where X_i are the observations in the Old Faithful dataset. Samples of size 107 are drawn, and two values of σ are considered: $\sigma_0 = 0.219$ and $\sigma_1 = 0.070$. The advantage of considering normal mixtures is that, with a normal kernel

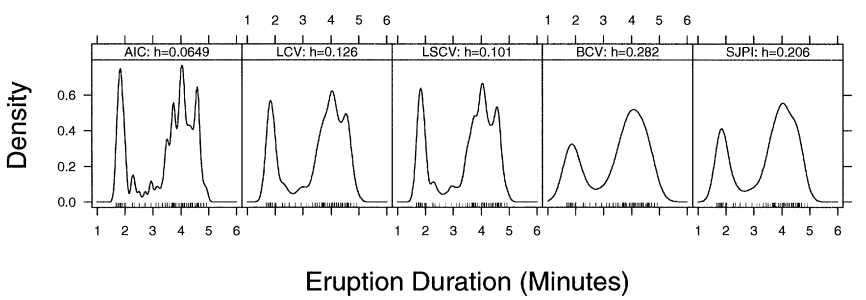

FIG. 3. *Density Estimates for the Old Faithful dataset. Each estimate is computed using a standard Gaussian kernel estimate; the bandwidths are selected by five different methods.*

density estimate, the mean integrated squared error has a closed form expression [Taylor (1989), Marron and Wand (1992)], and the MISE-minimizing bandwidth is easily computed. For $\sigma = \sigma_0$, we obtain the bandwidth $h = 0.206$, that selected by the SJPI method on the original data. For $\sigma = \sigma_1$, the bandwidth is $h = 0.101$, that selected by LSCV on the original data.

The results of 1000 simulations at each value of σ are shown in Figure 4. Each of the five bandwidth selectors is applied to each dataset. From the results of the 1000 simulations, we display estimated densities of the selected bandwidths.

AIC tends to undersmooth, often producing no bandwidth at all, as the criterion may be monotone. LCV and LSCV are quite variable, although on average they get close to the desired bandwidths. BCV oversmooths substantially. The Sheather–Jones plug-in method is indeed the least variable selector; unfortunately, it shows only modest response to the data and is quite incapable of selecting the smaller bandwidth, even when the small bandwidth is correct.

The conclusion with regard to Figure 3 is quite straightforward. The apparent better performance of the BCV and SJPI methods has nothing to do with asymptotic superiority enabling the method to reject a poor bandwidth. Rather, it is prior assumptions implicitly made by the selectors, and there is nothing in the data that caused the selection of $h = 0.206$ in preference to $h = 0.101$.

The simulations lead us to reconsider the original data; have the plug-in methods really performed better, or are they missing something? Plotting the fits, as in Figure 3, gives a very one-sided view of the bias-variance trade-off. Variance is easily seen, since it translates into spurious bumps and wiggles. Bias is much more difficult to see, since it requires very careful comparison of the fit with the data. Thus, just looking at the fitted curves may lead one to oversmooth.

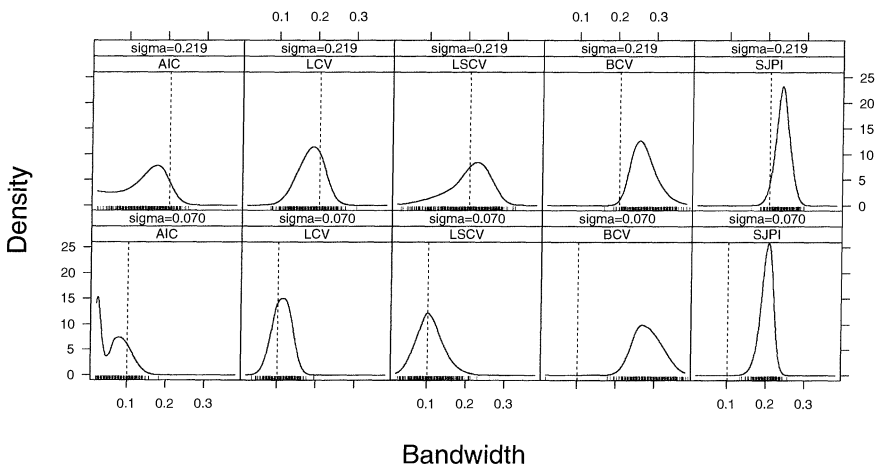

FIG. 4. *Densities of the bandwidths selected for* 1000 *resamples of the Old Faithful dataset. In the top row, we take* $\sigma = 0.219$, *with a target bandwidth of* $h = 0.206$. *In the bottom row,* $\sigma = 0.070$, *with a target bandwidth of* $h = 0.101$.

In the regression setting, the use of residual plots to look for bias is well established; see, for example, the extensive discussion in Cleveland (1993). In density estimation, the use of residuals is less well established, in part because of the difficulty of defining residuals. One approach is to convert density estimation into a local likelihood regression [Tibshirani and Hastie (1987)], either by considering spacings and fitting a local exponential regression, or by rounding the data and considering a local Poisson regression. One can then use any of the residuals used in generalized linear models; see Section 2.4 of McCullagh and Nelder (1989).

Figure 5 shows two such residual plots for the Old Faithful dataset, at $h = 0.101$ and $h = 0.206$. The residual plots are enhanced by adding a smooth local regression fit. Notice that in the right panel, the smooth shows a sharp peak just to the left of duration = 2; this lines up perfectly with the left peak in the original data. This provides a clear indication the peak has been oversmoothed. At the smaller $h = 0.101$, there is still some suggestion of oversmoothing, but far less severe.

The evidence is beginning to suggest that LSCV may have selected the correct bandwidth for the original data, and the plug-in approaches have oversmoothed. This does not imply that every bump shown by LSCV in Figure 3 is real. Rather, it must be remembered that the LSCV, BCV and SJPI are attempting to minimize MISE and not to produce the correct number of peaks. The residual plots of Figure 5 suggest the left peak is trimmed by SJPI. In retrospect, this oversmoothing can also be seen in Figure 3; both the BCV and SJPI estimates place substantial mass to the left of the smallest observation.

For our second example, consider the claw density from Marron and Wand (1992), which consists of five peaks superimposed on a standard normal density. Under a theoretical MISE criterion, the claws show up at $n = 54$ (see page 726 of Marron and Wand); at $n = 193$, they should be detectable in practice with some reliability.

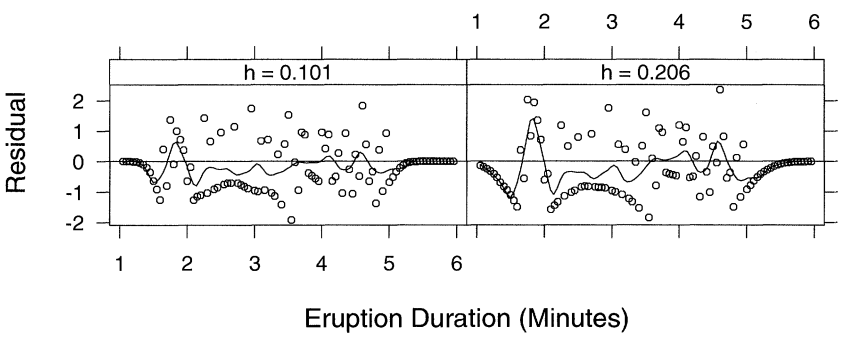

FIG. 5. *Deviance residual plots for the Old Faithful Dataset, at* $h = 0.101$ *and* $h = 0.206$. *The data is rounded to the nearest* 0.05 *and fitted using a local constant Poisson regression. The residuals are smoothed by a local quadratic regression with span covering* 15% *of the data.*

Figure 6 shows one sample, with $n = 193$. Both BCV and SJPI completely miss the structure. LSCV does find the claws, although with some noise; this is to be expected since each claw represents on average only 19.3 observations.

Figure 7 displays the bandwidths selected for 1000 simulations from the claw density, at three different sample sizes. At the smallest sample size, $n = 54$, the MISE function has two local minima; $h = 0.126$ represents claws and $h = 0.394$ represents the global structure. Since both local minima have approximately equal height, a reasonable selector should have the behavior shown by LSCV: targeting each minimum about half the time. SJPI nicely models the global structure, but completely misses the claws. At $n = 193$, the claws should be easier to detect, but only LSCV does so reliably. At $n = 400$, the problem should be getting easy. But BCV only sometimes finds the claws, and SJPI is always oversmoothing.

If a bandwidth selector is to be useful, it must perform reliably in difficult cases. In the claw density, this means $54 \leq n \leq 193$, when the claws should be detectable with some reliability, but will be far from obvious. In Figure 7, it is quite clear that only LSCV delivers.

Our final example consists of an equal mixture of ten normal distributions,

$$f(x) = \frac{1}{10} \sum_{i=1}^{10} \phi(x - (10i - 5)).$$

The sample size is $n = 100$. One such sample, along with the density estimates produced by four selectors, is shown in Figure 8. While the ten-modal structure is quite obvious in the data, only one selector, LSCV, gets close to the MISE-minimizing bandwidth $h = 0.809$. The plug-in approaches produce estimates that obviously don't fit the data.

Figure 9 summarizes the selected bandwidths for 1000 simulations. The plug in selectors never find the structure; BCV finds the structure (with a *local* minimum) in just 3 of the 1000 simulations.

Figure 10 displays the information provided by the various bandwidth selectors. Although only one simulation is reported, the picture was fairly con-

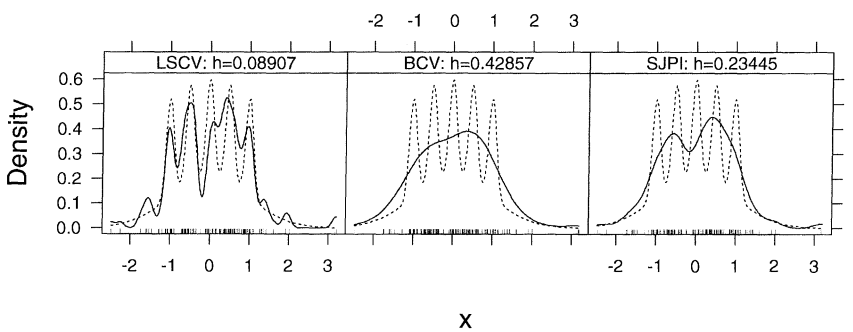

FIG. 6. *Density estimates for the claw density, with* $n = 193$. *The true density is shown with dashed lines, the estimates with solid lines.*

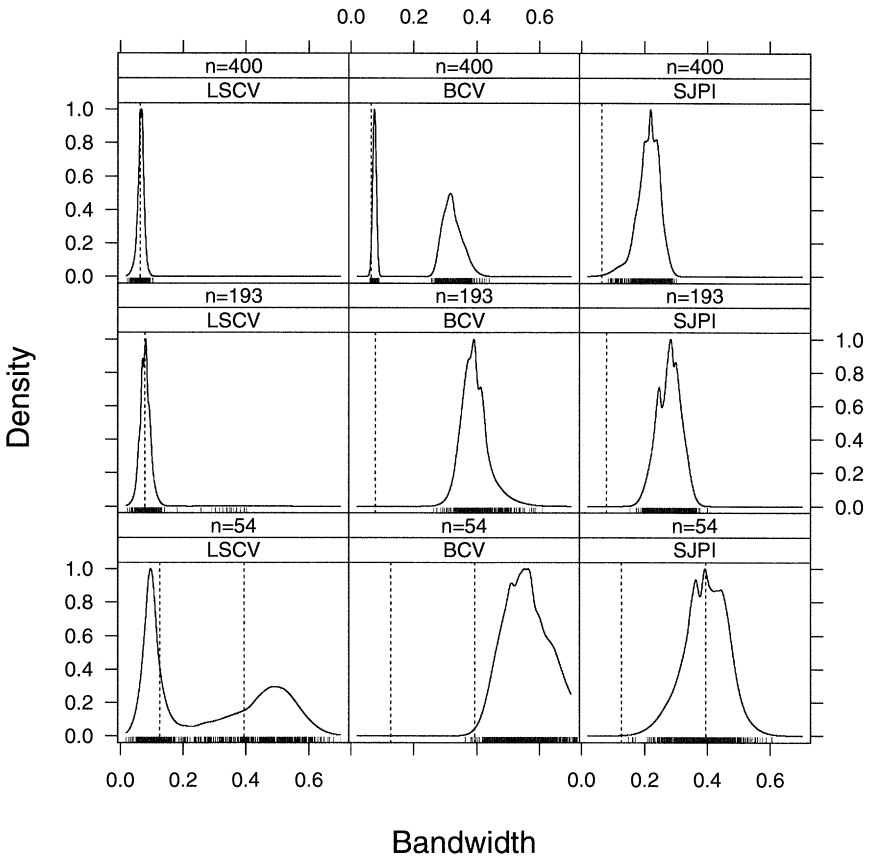

FIG. 7. *Bandwidth simulations for the claw density, at three different sample sizes. The densities of the selected bandwidths are deliberately undersmoothed, to ensure any bimodality is displayed. Dashed lines represent the true MISE-minimizing bandwidth(s). Note the densities have all been rescaled to have height* 1.0.

sistent over other replications. LSCV shows a sharp minimum around $h = 0.8$, and larger bandwidths are strongly rejected.

BCV is very flat for $h > 5$ and strongly rejects all smaller bandwidths. The problem here is BCV's use of the second derivative in the expansion (3): At large bandwidths, there is almost no curvature in the estimate, so the bias is significantly underestimated.

The plug-in curve in the right panel of Figure 10 clearly shows some response to the ten-modal structure. But the arbitrariness of the assumed relations is clear: both SJPI and GKK produce very oversmoothed solutions. Although all curves converge [to (0, 0)] on the left, neither method produced a satisfactory solution in any of the 1000 simulations.

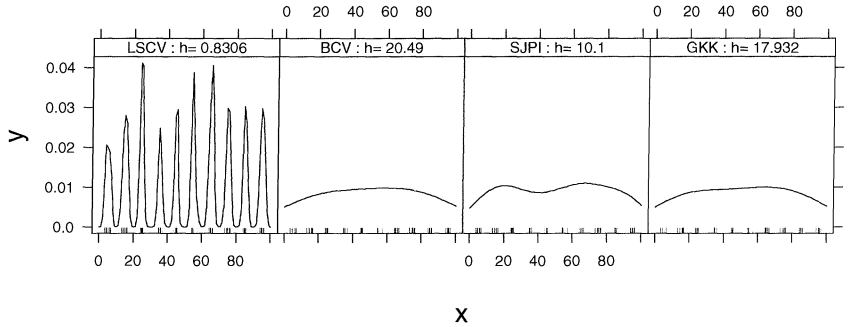

FIG. 8. *Density estimation for a ten-modal normal mixture, using four bandwidth selectors.*

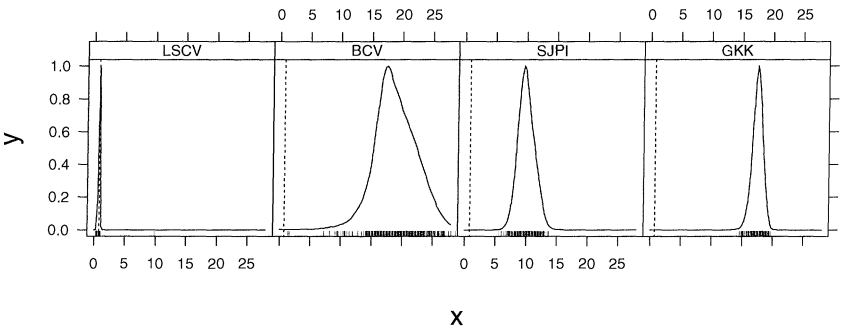

FIG. 9. *Selected bandwidths for* 1000 *simulations from the ten-modal example. The density estimates have been rescaled to have maximum* 1.0.

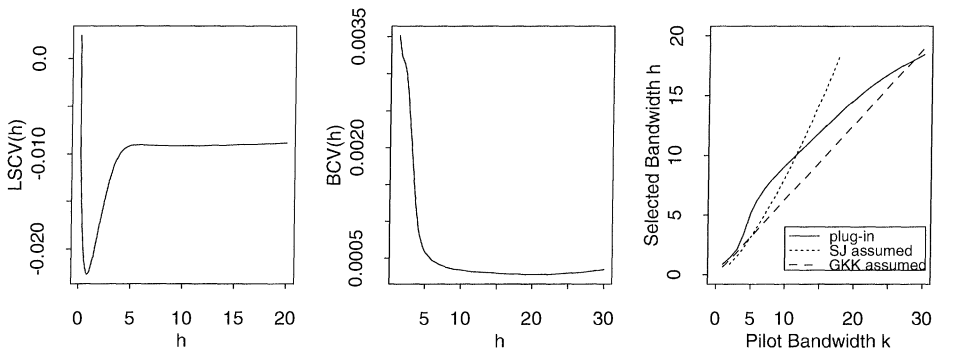

FIG. 10. *Informative bandwidth selection: The LSCV criterion (left); BCV (middle) and plug-in (right).*

5. Do plug-in selectors have better asymptotic performance? Some of the strongest arguments in favor of plug-in bandwidth selectors have been based on asymptotic studies. In particular, the rates of convergence (in a sense defined later) of cross validation and similar selectors is $O_p(n^{-1/10})$, while plug-in selectors achieve much faster rates. The SJPI method achieves a rate $O_p(n^{-5/14})$, and other plug-in algorithms achieve the rate $O_p(n^{-1/2})$ [Hall, Sheather, Jones and Marron (1991)]. At a glance, these results appear to provide compelling evidence that plug-in selectors must be better, at least asymptotically.

For asymptotic comparisons such as this to have any meaning, one has to follow a simple two stage procedure.

1. Formulate a set of assumptions.
2. Let each method do as well as possible under the assumptions made.

However, for the bandwidth selection results, this procedure has not been followed. In this section we discuss the crucial difference between the assumptions under which the rates are derived and their relation to local quadratic fitting. We argue that when the above procedure is followed, the existing asymptotics in fact favor the cross-validation selectors.

First, how should we assess the asymptotic performance of bandwidth selectors? We can consider either of two questions.

1. How close is the selected bandwidth to a target bandwidth? For example, let $h_0 = h_0(n)$ be the minimizer of the mean integrated squared error, and ask how fast does

$$\frac{\hat{h} - h_0}{h_0}$$

converge to 0? The bandwidth asymptotics stated above measure the rate of convergence of this quantity.
2. How well does the estimate $\hat{f}_h(x)$, using bandwidth $h = \hat{h}$, estimate the true $f(x)$? This can be measured by rates of convergence, or loss and risk measures such as mean integrated squared error.

Most comparisons of bandwidth selectors are based on the first type of measure, since it more directly measures the performance of the selector and is much more sensitive to differences between selectors. However, we must remember that usually measures of the second type address the real question of interest. In particular, if $\hat{f}_h(x)$ is an asymptotically inefficient estimate, it doesn't matter how good the bandwidth selector is.

What assumptions are made in the analysis of bandwidth selectors? Most crucially, one needs assumptions about the smoothness of the underlying density. For asymptotic analysis of bandwidth selectors, this amounts to assuming a sufficient number of derivatives. For the asymptotic results for LSCV, two

derivatives are required. For the asymptotic results for plug-in selectors, at least four derivatives are required.

Now, recall the optimal rates of convergence for density estimation, discussed for example in Stone (1980). Under the two derivative assumption, the best possible rate (in a minimax sense) is $\hat{f}_h(x) - f(x) = O_p(n^{-2/5})$, obtained by a kernel estimate with bandwidth $h = O(n^{-1/5})$. But under the four derivative assumption, the best possible estimates achieve a convergence rate of $O_p(n^{-4/9})$. Thus the kernel estimate is asymptotically inefficient, even when the best bandwidth is known.

Why make a big deal of the difference between $O_p(n^{-4/9})$ and $O_p(n^{-2/5})$? The answer comes in two parts: first, considering estimates that attain the $O_p(n^{-4/9})$ rate, and second studying how these methods relate to plug-in bandwidth selectors. Achieving the $O_p(n^{-4/9})$ rate is straightforward; locally quadratic approximations, and other asymptotically equivalent methods, such as fourth-order kernels, achieve this rate. See, for example, Stone (1980).

How do locally quadratic estimates relate to plug-in bandwidth selectors? From (4), it is clear that a good plug-in selector must use good estimates of $\int f''(x)^2\,dx$. However, this requires second derivative estimation and extending equivalence results for kernel and local regression estimates [Henderson (1916), Scott (1992), Section 6.2.3.3], one can show that essentially any second-derivative estimate is (at least asymptotically) the curvature term of a locally quadratic estimate. For a normal kernel density estimate, one has exact equivalence,

$$\frac{1}{nh^3}\sum_{i=1}^{n}\phi''\left(\frac{X_i - x}{h}\right) = \hat{a}_2,$$

where $\hat{a}_2$ is the curvature coefficient from the locally quadratic estimate (2). Thus, plug-in methods such as SJPI are making implicit use of locally quadratic estimates to estimate the curvature of the density. This curvature information is used in an asymptotically inefficient manner, namely, to estimate the bias of the kernel estimate.

What does all this mean in practice? In Figure 11, we compute the AIC criterion for the Old Faithful dataset, for both kernel (deg = 0) and locally log-quadratic (deg = 2) estimates. Clearly, AIC prefers the log-quadratic model, choosing a bandwidth $h = 0.37$. The resulting estimate, shown in the right panel of Figure 11, appears to achieve the good aspects of all the kernel estimates in Figure 3; we have the sharp left peak as shown by the cross-validation methods and a broad right peak without the apparently spurious noise.

Now let's do a plug-in step. The locally log-quadratic estimate in Figure 11 is used as a pilot estimate; the second derivative estimate is

$$\hat{f}''(x) = e^{\hat{a}_0}\left(\hat{a}_2 + \hat{a}_1^2\right),$$

where $(\hat{a}_0, \hat{a}_1, \hat{a}_2)$ are the coefficients of the local polynomial. Now, substitute this estimate into the plug-in formula (4). This produces $h = 0.124$; thus we

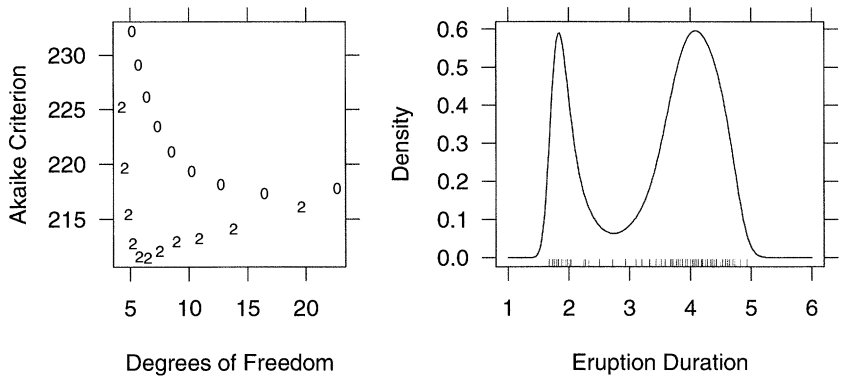

FIG. 11. *Fitting a local quadratic model to the Old Faithful dataset. At left, we plot Akaike's criterion, for local constant fitting* (0) *for h ranging from* 0.05 *to* 0.25, *and local log quadratic fitting* (2) *for h ranging from* 0.1 *to* 0.6 *(h increases from right to left). The selected smoother is local log-quadratic with* $h = 0.37$; *the corresponding fit is shown on the right.*

have a plug-in selector producing results quite comparable to the classical selectors in Figure 3. But clearly the plug-in step doesn't achieve anything useful: first, it just reproduces the features (the sharp left peak) in the pilot estimate, and second, it adds noise in the right peak.

The relative merits of kernel or locally constant fitting versus locally quadratic and cubic smoothing has been central to the smoothing literature for over a hundred years. In early work, predominantly in actuarial applications, locally quadratic and cubic methods were nearly universal, since they are better at modeling peaks in data. See Cleveland and Loader (1996) for more discussion, references and examples. Despite this enormous experience, there have been attempts in recent years to argue against locally quadratic smoothing; for example, Marron and Wand (1992) claim that enormous sample sizes are required for higher order methods to have practical value.

However, the question of usefulness of locally quadratic methods at practical sample sizes is surprisingly irrelevant to the present discussion of bandwidth selectors. The important point is that both locally quadratic smoothing and the second derivative estimates used in a plug-in bandwidth selector rely on the success of a locally quadratic approximation. Thus we can expect similar sample sizes to be required for both approaches.

This point is illustrated by the simulations reported in Figure 6. The middle sample size, $n = 193$, is the break-even point for the second-order kernel and fourth-order kernel (locally quadratic) estimates; see Table 2 of Marron and Wand (1992). For $n < 193$, there is insufficient data for a locally quadratic approximation to be successful, and locally constant or locally linear fitting beat locally quadratic fitting. In these cases, Figure 6 shows the classical selectors outperforming the plug-in.

From Figure 6, n must be much larger than 193 for the second-derivative estimation to be successful and the plug-in selectors to work. Even $n = 400$

is insufficient. If n were increased sufficiently (such simulations become computationally prohibitive), the asymptotics would eventually take over, and the plug-in selectors would be less variable than LSCV, when both are restricted to locally constant fitting. But this clearly is not relevant: at larger sample sizes, the plug-in selectors are beaten by their own pilot locally quadratic estimates. By allowing plug-in, but not LSCV, to use locally quadratic estimates, we penalize LSCV for the inefficiency of the plug-in estimate.

6. Local regression. So far we have studied the density estimation problem. Bandwidth choice also arises in other smoothing problems such as local regression [Henderson (1916), Cleveland and Devlin (1988)]. Most of the methods used in density estimation have analogs in the regression problem, and vice versa.

The regression model we consider is $Y_i = \mu(x_i) + \varepsilon_i$, and use constant bandwidth locally polynomial estimates. The bandwidth selectors we consider target the squared error loss function

$$L(\hat{\mu}, \mu) = \sum_{i=1}^{n} \big(\hat{\mu}(x_i) - \mu(x_i)\big)^2$$

and the risk function

$$R(\hat{\mu}, \mu) = \mathrm{E}\,(L(\hat{\mu}, \mu)) = \sum_{i=1}^{n} \mathrm{E}\,\big((\hat{\mu}(x_i) - \mu(x_i))^2\big)\,.$$

For equally spaced data, $n^{-1}L(\hat{\mu}, \mu)$ is a quadrature approximation to integrated squared error. Analagously to (4), the risk-minimizing bandwidth is asymptotically

$$h_{\mathrm{opt}} = \left(\frac{\sigma^2}{n}\,\frac{a_1(W)}{a_0(W)^2}\,\frac{\int f(x)^{-1}\,dx}{\int \mu''(x)^2\,dx}\right)^{1/5}, \tag{6}$$

where $f(x)$ is the design density and $W(v)$ the weight function; we use $W(v) = (1-|v|^3)^3 I_{[0,1]}(v)$. Formula (6) was given incorrectly by equation (3.1) of GKK and correctly by Fan (1993).

The four bandwidth selectors we consider are the following.

1. Generalized cross validation. Choose h to minimize

$$\mathrm{GCV}(h) = \frac{n\big\|(\mathbf{I}-\mathbf{H})Y\big\|^2}{(n-\mathrm{tr}\,(\mathbf{H}))^2}.$$

   Here, $Y = (Y_1 \cdots Y_n)^T$ and $\mathbf{H}$ is the hat matrix $(\hat{\mu}(x_1)\cdots\hat{\mu}(x_n))^T = \mathbf{HY}$.
2. C_p [Mallows (1973), Rice (1984), Cleveland and Devlin (1988)]. Choose h to minimize

$$C(h) = \frac{1}{\sigma^2}\big\|(\mathbf{I}-\mathbf{H})Y\big\|^2 - n + 2\,\mathrm{tr}\,(\mathbf{H})\,.$$

3. The plug-in algorithm of GKK. Estimate $\mu''(x)$ using a pilot locally quadratic estimate with bandwidth k, and select the bandwidth $h = h(k)$ plugging this into (6). Select k by solving the pilot relation $k = h(k)n^{1/10}$. GKK originally proposed their method for use with certain kernel estimates; with equally spaced x_i the change to local polynomials makes essentially no difference.
4. A hybrid of C_p and plug-in methods proposed by Ruppert, Sheather and Wand (1995) (RSW). The pilot estimate is a blocked locally quartic fit with the number of blocks, p, chosen by C_p. The bandwidth is selected by plugging the local curvature of this fit into (6).

Two other points need noting. First, all selectors except GCV require an estimate of σ^2. In our simulations, the same variance estimate is used in all selectors, namely, the normalized residual sum-of-squares based on a locally quadratic fit with $h = 0.05$. Second, the asymptotic arguments leading to (6) are not valid in boundary regions. Both GKK and RSW modify the loss function by truncating boundary regions; in our implementation of these methods, the integrals in (6) are taken over $[0.1, 0.9]$ rather than $[0, 1]$.

We consider regression problems with $n = 100$ equally spaced points on $[0, 1]$; $\mu(x) = 4(x - 0.5)^2 + c\sin(10\pi x)$ and $\varepsilon_i \sim N(0, 1)$. For $c = 0$, $\mu(x)$ is quadratic, with a small, but detectable, curvature. The risk function for a locally linear smooth has a single minimum. For moderate c, the risk has two local minima, one corresponding to the quadratic structure and the other to the sinusoidal structure. The crossover, when the sinusoidal minimum dominates, is about $c = 0.4$. For large c, the risk again has a single minimum.

Figure 12 displays the results of 1000 simulations. At $c = 0$, both C_p and GCV are centered exactly where they should be. The GKK and RSW methods are less variable, but consistently undersmooth. At $c = 0.4$, the $C(h)$, GCV and RSW are all bimodal, representing the two local minima of the true risk. GKK has got this completely wrong, being distributed in between the two minima.

Since RSW uses C_p at the pilot stage and plug-in at the second stage, shows the bandwidths, Figure 13 split by the number of blocks selected at the C_p stage. The three panels are visually almost identical; that is, the plug-in step does almost no adapting to the different regression functions. What has changed is the initial C_p step.

For our second regression example, we use locally quadratic regression since this frequently beats locally linear in practice. The risk-minimizing bandwidth is asymptotically

$$h_{\text{opt}} = \left(\frac{\sigma^2}{n} \frac{a_1(W^*)}{a_0(W^*)^2} \frac{\int f(x)^{-1}\,dx}{\int \mu^{(iv)}(x)^2\,dx} \right)^{1/9}, \tag{7}$$

where W^* is the "equivalent" kernel [Henderson (1916)],

$$W^*(v) = A(0)^T \left(\int A(u)A(u)^T W(u)\,du \right)^{-1} A(v)W(v)$$

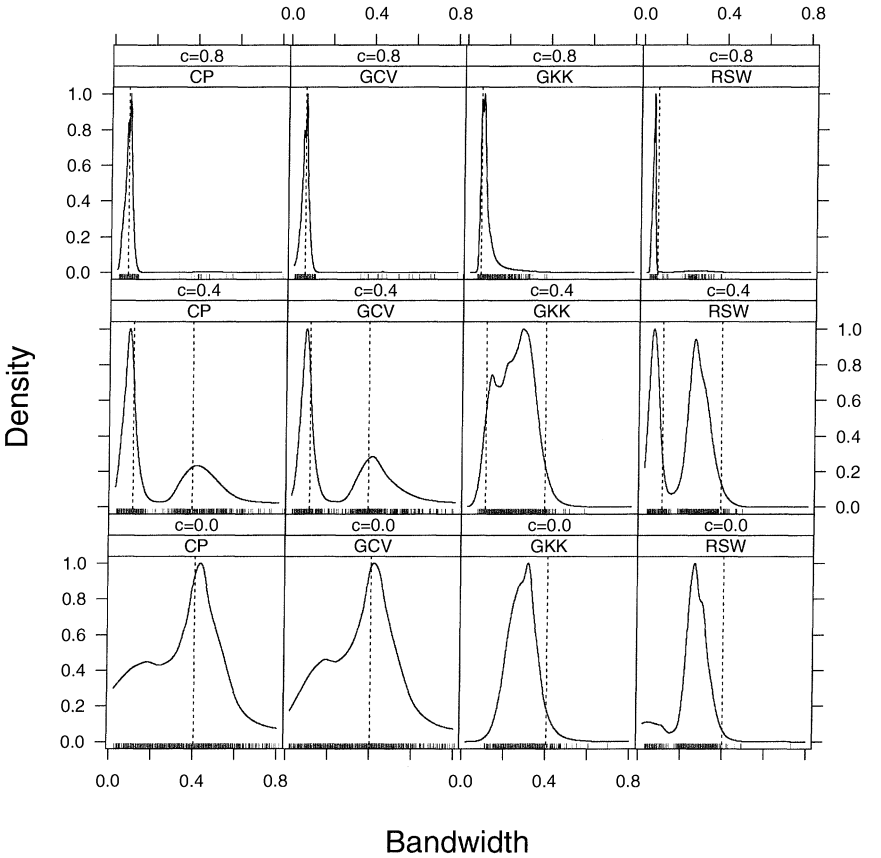

FIG. 12. *Selected bandwidths for* 1000 *local linear regression simulations.*

with $A(v)^T = (1 \quad v \quad v^2)$. To modify the GKK algorithm for the locally quadratic estimate, we simply use a locally quartic fit at the pilot stage to estimate $\mu^{(iv)}(x)$. For RSW, we again use the blocked locally quartic fit as the pilot estimate, but plug into (7).

We take $\mu(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4 + c\exp(-1000(x-0.62)^2)$. Figure 14 displays the results of 1000 simulations for $c = 0$ and $c = 3$. At $c = 0$, this should be favorable to plug-in selectors, since the pilot locally quartic estimate has no bias, and large pilot bandwidths can be used. This is reflected in Figure 14, where GKK and RSW are substantially less variable.

Are $C(h)$ and GCV too variable? In Figure 14 with $c = 0$, the bandwidths selected range from 0.05 (the programmed lower bound) to over 0.4. In Figure 12 with $c = 0$, the variability is even larger.

Let's take a closer look at one of the "worst" samples generated in the simulations of Figure 14. The dataset, that used in Figure 1, selected $h = 0.0618$ ($C(h)$); $h = 0.0616$ (GCV) and $h = 0.288$ (GKK). The RSW method

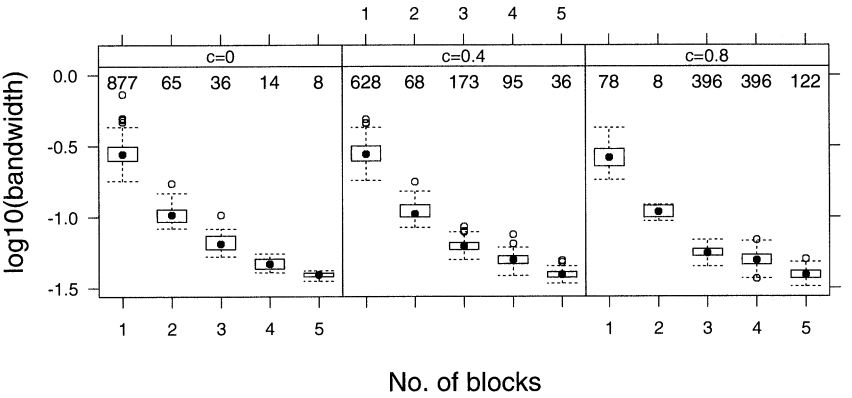

FIG. 13. *Effect of initial C_p step on the Ruppert–Sheather–Wand selector. The numbers at the top represent the number of times the C_p selected p blocks. The box plots show the distribution of the selected bandwidth, conditional on p.*

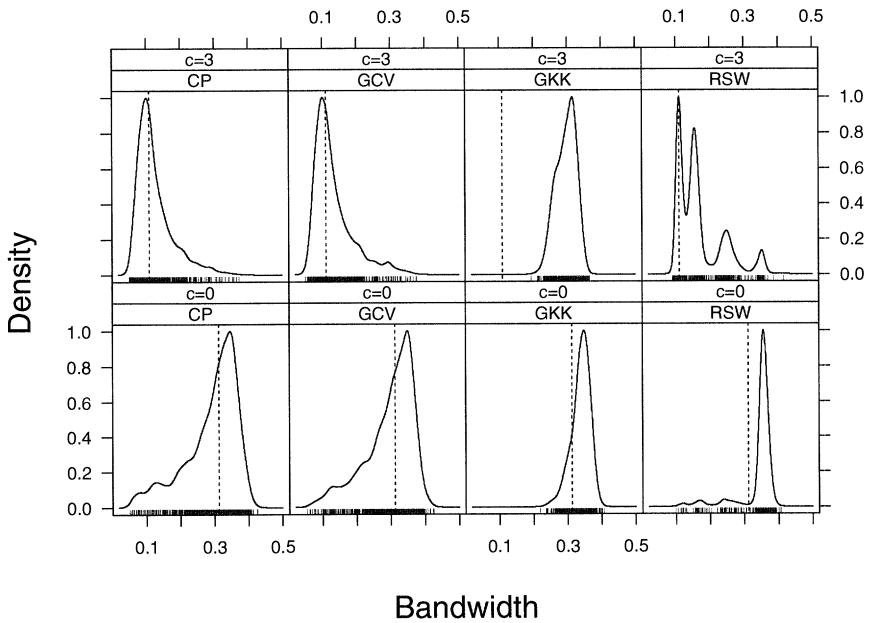

FIG. 14. *Selected bandwidths for local quadratic regression.*

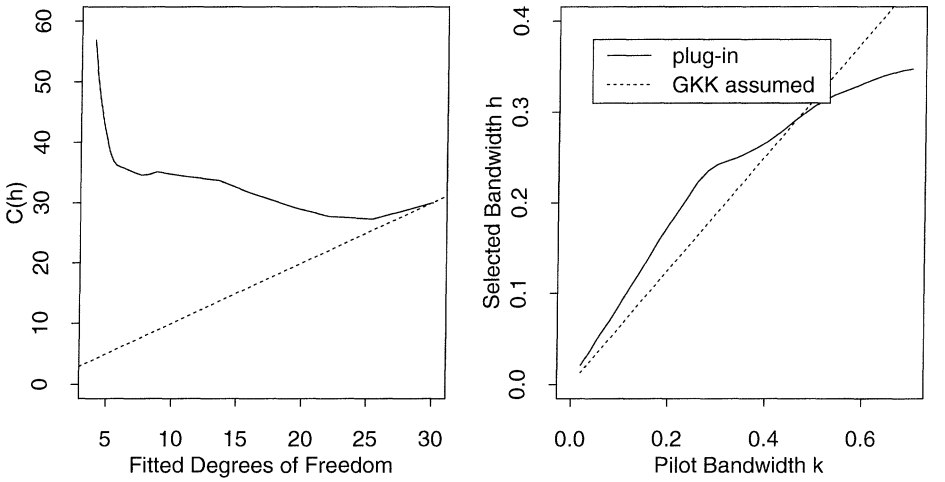

FIG. 15. *Informative bandwidth assessment for the "bad" dataset. The flatness of the $C(h)$ plot (left) reflects the uncertainty in the data. The GKK method selects the larger bandwidth (right) without any suggestion of the uncertainty.*

selected one block at the C_p step, leading to the bandwidth $h = 0.353$. The $C(h)$ and GKK fits were shown in Figure 1.

If one looks just at the selected bandwidth, the GKK method and RSW (by virtue of its initial C_p step) have got this dataset right, and $C(h)$) and GCV have got it wrong. But the plug-in methods (particularly GKK) pay a price for this: oversmoothing and missing the structure on the more difficult problem in Figure 14, when $c = 3$. Looking at the whole criterion, rather than just the selected bandwidth, produces a much more valuable assessment. The $C(h)$ plot in the left panel of Figure 15 correctly reflects the uncertainty in the dataset, with two local minima and a nearly flat plot from 5 to 30 degrees of freedom. GCV (not shown) produces a similar flat plot. GKK selects the larger bandwidth, with no hint of uncertainty at the smaller bandwidth. The result is catastrophic failure when the bump is real, as demonstrated in Figure 14.

The conclusion here is simple. Variability of $C(h)$ and GCV is not the problem, but a symptom of how difficult purely data-based bandwidth selection is. It is easy to "fix" the variability of $C(h)$ to give better results on the dataset in Figure 1, for example, by taking the left-most local minimum rather than the global minimum. However, this type of fix fails to address the difficulty of bandwidth selection and will lead to failure in difficult problems, similar to GKK in Figure 15.

7. Conclusions. We have studied a wide range of bandwidth selectors, on both real and simulated data. When the results are analyzed carefully, the

much touted plug-in approaches have fared rather poorly, being tuned largely by arbitrary specification of pilot bandwidths and being heavily biased when this specification is wrong. We do not claim that classical approaches such as AIC and cross validation will always produce the best estimates, but rather that, used properly, the results will often be far more informative than other recent work in bandwidth selection suggests.

Much of the criticism directed at cross validation and classical approaches to bandwidth selection would be better directed at kernel estimation and fixed bandwidth methods. We see this in the Old Faithful dataset: the small bandwidths are being selected by LSCV because that is the only way the sharp left peak can be modeled. Another criticism of classical approaches, particularly LCV, is that they can oversmooth when used with heavy tailed distributions. If an outlier is left out of the dataset, then smoothing the remaining observations may produce no estimate at that point, forcing a larger bandwidth to be selected. A reference for this point is Schuster and Gregory (1981), who point out that LCV produces inconsistent fixed bandwidth kernel estimates when the underlying distribution has heavy tails. Schuster and Gregory then correctly conclude a fixed bandwidth estimate is inadequate for heavy tails and use this to motivate variable bandwidth kernels. Subsequent authors [e.g., Scott (1992), page 163] incorrectly conclude there is a problem with LCV.

With the Old Faithful dataset, simulations based on a smoothed bootstrap approach, residual diagnostics and higher order fits have all suggested the classical approaches are correct in choosing small bandwidths, and the plug-in approaches incorrectly oversmooth, with regard to the integrated squared error loss function. This point has been missed by previous authors applying kernel methods, who rely exclusively on bandwidth selectors and looking at the fitted curves to determine an acceptable fit and do not perform any diagnostics to detect lack of fit. While the statistician may still prefer the oversmoothed estimate, it is hardly fair to praise plug-in methods (and criticize LSCV), since these methods target MISE and not smoothness of the estimate. If smoother estimates are preferred, then the MISE criterion should be acknowledged as inadequate and bandwidth selectors directed towards a more appropriate criterion.

The comparisons between classical and plug-in approaches presented in the literature have several weaknesses. First, plug-in approaches, through the specification of tuning parameters for pilot estimates, effectively make substantial prior assumptions about the required bandwidth and will fail if this information is wrong. Second, the plug-in approaches obtain much of their information from the data through the use of higher order pilot estimates; if classical approaches are also allowed to consider higher order methods, better *estimates* result. Third, plug-in methods are not rescued by asymptotic analysis showing better rates of convergence; assumptions about the underlying function make the resulting estimate asymptotically inefficient, regardless of how good the bandwidth selector is.

We have emphasized the importance of not relying blindly on any bandwidth selector to produce the right bandwidth automatically. If one just ap-

plies a bandwidth selector plots the fit, one gets a very one-sided view of the bias-variance trade-off, seeing variance, but not bias. It is extremely important to use appropriate residual diagnostics to look for lack of fit. Likewise, plotting the cross validation or AIC criteria provides valuable diagnostic information as to how difficult the bandwidth selection is; a flat plot suggests that different features of the data may be competing for attention at different bandwidths. Plug-in approaches, which arbitrarily impose an $Ah^4 + B/(nh)$ profile on the integrated squared error in such cases, discard this information.

The importance of using carefully designed graphical displays in conjunction with bandwidth selectors cannot be overemphasized. Even relatively mundane points, such as showing the data along with the fit, are of considerable importance. For example, the oversmoothing of the left peak by SJPI and BCV can be seen in Figure 4, but is quite invisible in Figure 6.17 of Scott (1992) or Figure 2.2 of Sheather (1992).

We conclude by mentioning some important issues that have not been discussed in this paper, since they have little bearing on the points discussed.

1. Is ISE the right loss function? Our almost exclusive use of ISE and MISE should not be considered an endorsement, but rather a reflection of the literature. Most bandwidth selectors target these measures, so it is these measures by which bandwidth selectors are judged.
2. Bandwidth schemes: fixed versus nearest neighbor versus locally adaptive choices? Most bandwidth selection literature centers on the single fixed bandwidth, and so this paper is restricted to that setting. Often a fixed bandwidth is inadequate, and both classical and plug-in selectors have locally adaptive variants, where the bandwidth is chosen separately for each fitting point x. Most of the issues in this paper also arise in the locally adaptive setting; however, this adds little to the comparison of classical versus plug-in approaches.

REFERENCES

AZZALINI, A. and BOWMAN, A. W. (1990). A look at some data on the Old Faithful geyser. *Appl. Statist.* **39** 357–365.

BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

CHIU, S. T. (1991). Bandwidth selection for kernel density estimation. *Ann. Statist.* **19** 1883–1905.

CLEVELAND, W. S. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.

CLEVELAND, W. S. and DEVLIN, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 596–610.

CLEVELAND, W. S. and LOADER, C. R. (1996). Smoothing by local regression: principles and methods. In *Statistical Theory and Computational Aspects of Smoothing* (W. Härdle and M. G. Schimek, eds.) 10–49. Physica, Heidelberg.

DUIN, R. P. W. (1976). On the choice of smoothing parameter for Parzen estimators of probability density functions. *IEEE Trans. Comput.* **C-25** 1175–1179.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21** 196–216.

GASSER, T., KNEIP, A. and KÖHLER, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86** 643–652.

HABBEMA, J. D. F., HERMANS, J. and VAN DER BROEK, K. (1974). A stepwise discriminant analysis program using density estimation. In *COMPSTAT 1974, Proceedings in Computational Statistics, Vienna* (G. Bruckman ed.) 101–110. Physica, Heidelberg.

HALL, P., SHEATHER, S. J., JONES, M. C. and MARRON, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78** 263–270.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1992). Regression smoothing parameters that are not far from their optimal. *J. Amer. Statist. Assoc.* **87** 227–233.

HENDERSON, R. (1916). Note on graduation by adjusted average. *Trans. Actuarial Soc. America* **17** 43–48.

HJORT, N. L. and JONES, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24** 1619–1647.

JONES, M. C., MARRON, J. S. and SHEATHER, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91** 401–407.

LEJEUNE, M. and SARDA, P. (1992). Smooth estimators of distribution and density functions. *Comput. Statist. Data Anal.* **14** 457–471.

LOADER, C. R. (1996a). Local likelihood density estimation. *Ann. Statist.* **24** 1602–1618.

LOADER, C. R. (1996b). *Local Regression and Likelihood*. Electronic book, http://cm.bell-labs.com/stat/project/locfit/.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MARRON, J. S. (1996). A personal view of smoothing and statistics. In *Statistical Theory and Computational Aspects of Smoothing* (W. Härdle and M. G. Schimek eds.) 1–9. Physica, Heidelberg.

MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

PARK, B. U. and MARRON, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85** 66–72.

PARK, B. U. and TURLACH, B. A. (1992). Practical performance of several data driven bandwidth selectors. *Comput. Statist.* **7** 251–270.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270.

SCHUSTER, E. F. and GREGORY, G. G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (W. F. Eddy, ed.) 295–298. Springer, Berlin.

SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York.

SCOTT, D. W. and TERRELL, G. R. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **82** 1131–1146.

SHEATHER, S. J. (1992). The performance of six popular bandwidth selection methods on some real datasets. *Comput. Statist.* **7** 225–250.

SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

TAYLOR, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika* **76** 705–712.

TIBSHIRANI, R. J. and HASTIE, T. J. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559–567.

WOODROOFE, M. (1970). On choosing a delta sequence. *Ann. Math. Statist.* **41** 1665–1671.

LUCENT TECHNOLOGIES
ROOM 2C-279
600 MOUNTAIN AVENUE
MURRAY HILL, NEW JERSEY 07974
E-MAIL: clive@bell-labs.com