Old Faithful Erupts: Bandwidth Selection Reviewed

Clive R. Loader

October 30, 1995

Abstract

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.

The claims of superior performance for plug-in approaches are found to be without foundation. The plug-in approach is heavily dependent on arbitrary specification of pilot bandwidths; if these are wrong, the selection is wrong. Previous authors have failed to do proper diagnostics on the resulting estimates; thus missing substantial oversmoothing of plug-in selectors. The plug-in approach uses available curvature information in an inefficient manner, resulting in inefficient estimates. Many 'comparisons' with classical approaches have been distorted to penalize the classical approaches for this inefficiency! Asymptotically - and often in finite samples - the plug-in based estimates are beaten by their own pilot estimates!

Key words: Adaptive smoothing; Akaike's information criterion; Bandwidth; Cross validation; Density estimation; Local fitting; Local likelihood; Plug-In.

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, although other settings are far from ignored. The bandwidth selection methods studied in the literature can be divided into two broad classes:

Classical Methods: Cross-validation; Mallow's 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 and other asymptotics. An estimate of f is then 'plugged in' to derive an estimate of the bias, and hence an estimate of measures of fit such as mean integrated squared error. One then chooses an 'optimal' h which minimizes this estimated measure of fit.

In the context of kernel density estimation, the plug-in approach appears to predate 'classical' approaches, dating to Woodroofe (1970). However, more specific algorithms and the strong promotion of the approach began in the mid 1980's, 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.

Ruppert, Sheather and Wand (1995) are even more forthright, stating on recent bandwidth selection research:

> One of the main findings of this research is that traditional smoothing parameter selection rules, such as those based on cross-validation exhibit very inferior asymptotic and practical performance.

and continue

> 'plug-in' bandwidth selection rules ... have been shown to perform more reliably, both theoretically and in practice.

We show this supposed superior performance of plug-in approaches is a complete myth. The 'plug-in' approach relies critically on the specification of a pilot estimate, which in turn relies critically on a pilot bandwidth. Most commonly, this pilot bandwidth is specified using a 'reference' distribution (usually the normal distribution); this results in plug-in approaches having much less variability than classical approaches. However, this only works in tuned examples where the reference distribution is not too wrong; otherwise, substantial oversmoothing results. Moreover, the pilot estimate uses substantial curvature information, which is used only to fine-tune less efficient estimates. In typical 'comparisons' with cross validation, this curvature information is hidden, thus penalizing cross-validation for inefficiencies of the plug-in method.

The problems of choice of pilot estimate and reference methods have been recognized in some recent papers in local regression; e.g. Fan and Gijbels (1994) and Ruppert, Sheather and Wand (1995). The solution? Use a classical approach, such as C_p, to tune the pilot estimate. Then add a couple of plug-in steps, and present the final method as a 'plug-in' selector. In fact, the initial C_p step does all the difficult work; the 'plug-in' steps contribute at most nothing, and can degrade the resulting estimate.

2 Some Density Estimates

Suppose we have an independent sample $X_1, \ldots, X_n$ from an unknown density $f(\cdot)$. The classical kernel density estimate (Rosenblatt, 1956) is

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

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

Better estimates can be obtained through the use of additional estimating equations and local parametric approximations (Hjort and Jones, 1994). Let $A(v) = (\,1 \quad v \quad \frac{1}{2}v^2\,)^T$, and consider solutions $\hat{a} = \hat{a}(x)$ equations of the form

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

where $\tilde{f}(u, a)$ is a local three-parameter approximation, and the density estimate is $\hat{f}(x) = \tilde{f}(x, \hat{a}(x))$. Using the local quadratic approximation $\tilde{f}(u, a) = \langle a, A(u - x)\rangle$ gives rise to the fourth order kernel estimate; this essentially follows from observations of Lejeune and Sarda (1992). The higher order coefficients of the local polynomial approximation provide estimates of the density derivatives; if a Gaussian weight function is used, the quadratic coefficient $\hat{a}_2$ is precisely the second derivative of the kernel estimate (1).

Another approximation is the local log-quadratic $\tilde{f}(u, a) = \exp(\langle a, A(u - x)\rangle)$, which leads to the local likelihood method of Loader (1993). This method has some significant advantages over the higher order kernel methods; it necessarily produces non-negative estimates, and is better for estimating the tails of the density.

An important point is that local quadratic approximations often provide vastly superior fits compared to local constant or local linear approximations. This has been widely appreciated for many decades in the regression setting; see for example Henderson (1916). Examples in this paper further support this point. However, this point seems to be lost in much recent work on bandwidth selection and kernel estimation; indeed, there are attempts to argue otherwise (Marron and Wand, 1991).

3 Some Bandwidth Selectors

Akaike's information criterion (Akaike, 1973) selects the model that minimizes

$$AIC(h) = -\sum_{i=1}^{n} \log \hat{f}_h(X_i) + "p".$$

For parametric models, "p" is the dimension of the model; usually, the number of parameters. This isn't directly meaningful for local models, hence the quotes. However, some definitions of degrees of freedom for local likelihood models are discussed for example by Tibshirani and Hastie (1987). For our purposes we use

$$p = \sum_{i=1}^{n} \text{infl}\,(X_i)$$

where for local likelihood fitting,

$$\text{infl}\,(X_i) = W\,(0)\, e_1^T \left(\int_{-\infty}^{\infty} W\left(\frac{u-x}{h}\right) A(\frac{u-x}{h}) A(\frac{u-x}{h})^T \tilde{f}(u, \hat{a}) du \right)^{-1} e_1.$$

For local constant (kernel) fitting, this reduces to

$$p = \sum_{i=1}^{n} \frac{W\,(0)}{nh\hat{f}(X_i)}.$$

The **likelihood cross-validation** method, proposed by Habbema, Hermans and Van Der Broek (1974) and Duin (1976), selects the bandwidth h that maximizes $\prod_{i=1}^{n} \hat{f}_{h,-i}(X_i)$, where $\hat{f}_{h,-i}(X_i)$ denotes the density estimate with X_i deleted. Some manipulation shows this is equivalent to minimizing

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

LCV and AIC thus appear very similar, although in practice the methods respond quite differently in the tails. With some exceptions (e.g. Kanazawa, 1993) likelihood based criteria such as these are rarely considered in recent literature on kernel estimation for reasons we discuss more later.

Least squares cross-validation (Rudemo, 1982; Bowman, 1984) attempts to minimize the integrated squared error:

$$ISE(h) = \int_{-\infty}^{\infty} (\hat{f}_h(x) - f(x))^2 dx = \int_{-\infty}^{\infty} \hat{f}_h(x)^2 dx - 2\int_{-\infty}^{\infty} \hat{f}_h(x) f(x) dx + \int_{-\infty}^{\infty} f(x)^2 dx.$$

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 the data, and so can be ignored. The central term $\int_{-\infty}^{\infty} \hat{f}_h(x) f(x) dx$ is estimated by leave-one-out cross-validation, giving the final criterion

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

Biased cross-validation (Scott and Terrell, 1987) uses the asymptotic approximation to mean integrated squared error,

$$AMISE(h) = \frac{a_0^2 h^4}{4} \int_{-\infty}^{\infty} f''(x)^2 dx + \frac{a_1}{nh}$$

where $a_0 = \int_{-\infty}^{\infty} v^2 W(v)\, dv$ and $a_1 = \int_{-\infty}^{\infty} W(v)^2\, dv$. Since $\int_{-\infty}^{\infty} f''(x)^2 dx$ is unknown, Scott and Terrell estimate this by substituting $\hat{f}_h''$ itself (hence, this algorithm should really be considered plug-in, rather than cross-validation). Note that $E(\hat{f}_h''(x)^2) = (E\hat{f}_h''(x))^2 + \mathrm{var}(\hat{f}_h''(x))$; the integral of the variance term is essentially a function of the weight function (and not f) so can be subtracted. This leads to the BCV criterion,

$$BCV(h) = \frac{a_0^2 h^4}{4} \left(\int_{-\infty}^{\infty} \hat{f}_h''(x)^2 dx - \frac{1}{nh^5} \int_{-\infty}^{\infty} W''(v)^2 dv \right) + \frac{a_1}{nh}.$$

The **Sheather-Jones plug-in** method (Sheather and Jones 1991) also targets the asymptotic approximation to mean integrated squared error. The development here follows that in Section 5 of Sheather and Jones for a standard normal kernel;

$$W(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

Using a pilot bandwidth α, Sheather and Jones estimate $\int_{-\infty}^{\infty} f''(x)^2 dx$ by

$$S(\alpha) = \frac{1}{n(n-1)\alpha^5} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi^{(4)} \left(\frac{X_i - X_j}{\alpha} \right);$$

note this is the integrated squared second derivative of the kernel density estimate with bandwidth $\alpha/\sqrt{2}$. Thus the optimal bandwidth is estimated by

$$\hat{h} = \left(\frac{a_1}{a_0^2 S(\alpha)} \right)^{1/5} n^{-1/5}.$$

with a_1 and a_0 given above.

It remains to specify α. Let $\hat{\lambda}$ be the interquartile range of the data. Based on a Normal reference model, and a standard normal kernel, Sheather and Jones argue the bandwidth and pilot bandwidth should satisfy the relation

$$\hat{\alpha}(h) = 1.357 \left(\frac{S(a)}{T(b)} \right)^{1/7} h^{5/7} \tag{2}$$

with $a = 0.920\hat{\lambda}n^{-1/7}$; $b = 0.912\hat{\lambda}n^{-1/9}$ and

$$T(b) = -\frac{1}{n(n-1)b^7}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{(6)}\left(\frac{X_i - x_j}{b}\right).$$

The final selected bandwidth solves the equation

$$\hat{h} = \left(\frac{a_1}{a_0^2 S(\hat{\alpha}(\hat{h}))}\right)^{1/5} n^{-1/5}. \tag{3}$$

Sheather and Jones claim a convergence rate $\hat{h}/h_0 = 1 + O_p(n^{-5/14})$. Faster rates of $O_p(n^{-1/2})$ are claimed for other variants of the plug-in approach; for example, Hall, Sheather Jones and Marron (1991). However, there seems general agreement among plug-in proponents that the version considered here is among the best from a practical point of view, and Sheather and Jones (1991) claim

> ... most importantly, it must be stressed that the bandwidth estimator that we recommend has a practical performance second to none in the existing literature ....

4 An Example

Let us give an example, using a dataset of eruption durations of the Old Faithful geyser. Azzalini and Bowman (1990) note there are several different Old Faithful datasets, and provide an interesting Markov chain analysis as well as discussing some geological background. The dataset we use consists of 107 eruptions in October, 1980, to which a number of authors have applied kernel density estimates (1) with a standard Gaussian kernel and a variety of bandwidth selectors. Silverman (1986) smoothed the data visually, selecting $h = 0.25$. Sheather (1992) reports $h = 0.101$ using LSCV; $h = 0.206$ using the SJPI method; $h = 0.228$ using the plug-in estimate of Park and Marron (1990), and $h = 0.494$ using the plug-in method of Hall, Sheather, Jones and Marron (1991). The BCV function has two local minima at 0.282 and 1.147; Sheather chooses the latter while Scott (1992) chooses the former. Chiu (1991) selects $h = 0.215$ using another plug-in based selector. We also consider AIC, which selects the smallest bandwidth; $h = 0.0649$, and LCV which selects $h = 0.126$. It should be remembered these last two approaches are not directly targeting mean integrated squared error.

Figure 1 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 we appear to have a fairly clear consensus as to what works, and what does not work, on this dataset.

Sheather (1992) presents six different real data examples, showing mostly poor performance for cross-validation; the performance in the Old Faithful example is average. The conclusion that plug-in promoters would like us to draw is that superior asymptotic rates of convergence have resulted in a greater ability to distinguish good from bad bandwidths in finite samples.

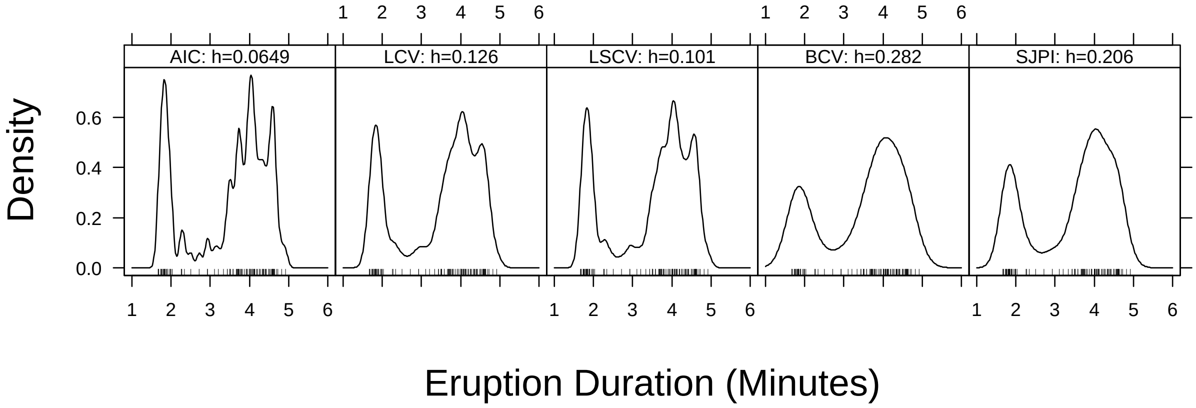

Figure 1: Density Estimates for the Old Faithful dataset. Each estimate is computed using a standard Gaussian kernel estimate; the bandwidths are selected by five different methods.

Of course, with real data we can't be completely sure. Let us consider some simulations from 'similar' distributions. Specifically, we sample from Gaussian mixture distributions of the form

$$G + \sigma Z$$

where G is sampled from the Old Faithful dataset, and $Z \sim N(0, 1)$. 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 density estimate, the mean integrated squared error has a closed form expression; see Taylor (1989) and Marron and Wand (1992). Hence, the MISE-minimizing bandwidth can be found through a fairly straightforward numerical computation. For $\sigma = \sigma_0$, we obtain the optimal bandwidth $h_0 = 0.206$; that selected by the SJPI method on the original data. For $\sigma = \sigma_1$, the optimal 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 2. Each of the five bandwidth selectors is applied to each dataset. From the results of the 1000 simulations, we display estimated densities of the bandwidth selectors.

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; interestingly, it tends to produce larger bandwidths under $\sigma = 0.070$ than under $\sigma = 0.219$! 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 1 is quite straightforward. The apparent better performance of the SJPI method has nothing to do with asymptotic superiority enabling the method to reject a poor bandwidth. Rather, it is prior information built into the selector, and there is nothing in the data that caused the selection of $h = 0.206$ in preference to $h = 0.101$.

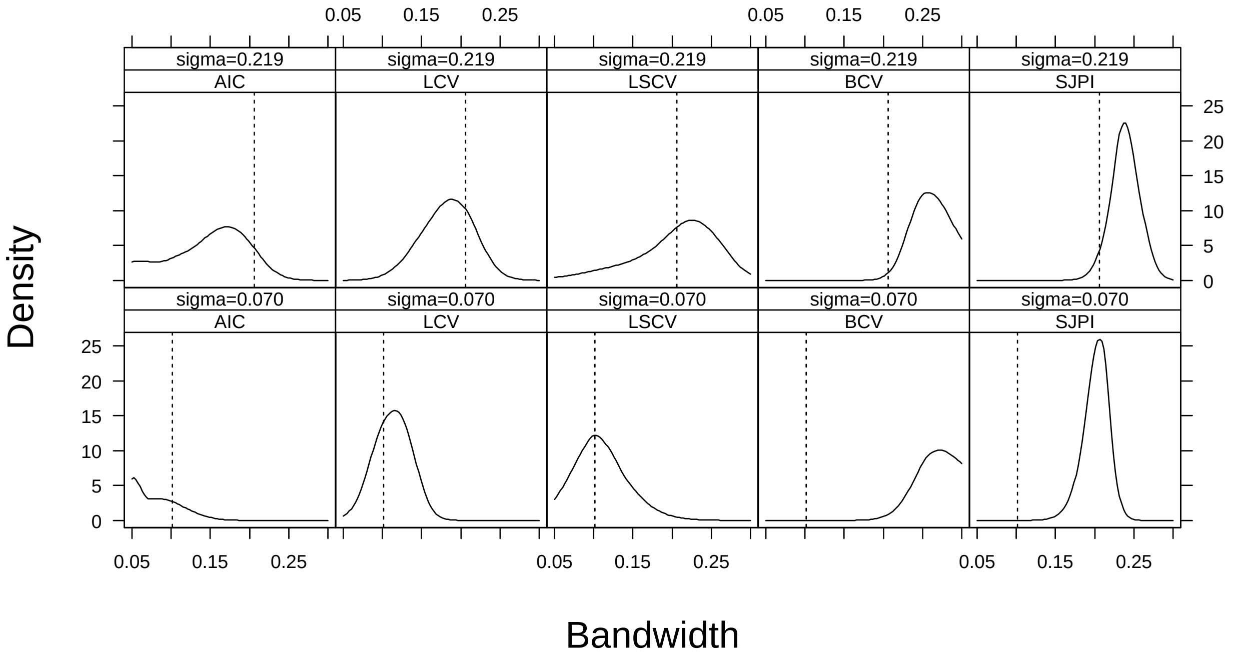

Figure 2: Densities of the bandwidths selected for 1000 resamples of the Old Faithful dataset. In the top row, we take $\sigma = 0.219$, with a target bandwidth of $h = 0.206$. In the bottom row, $\sigma = 0.070$, with a target bandwidth of $h = 0.101$.

Let us reconsider Figure 1. The selection of smoothing parameters is a bias-variance trade-off. But just plotting the fit, as in Figure 1, gives a very one-sided view of the trade-off. We easily see noise in the fit, since it translates into spurious bumps and wiggles. Bias, however, is much more difficult to discern, since it requires very careful comparison of the fit with the data. Thus, just looking at the fitted curves may lead one to oversmooth.

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).

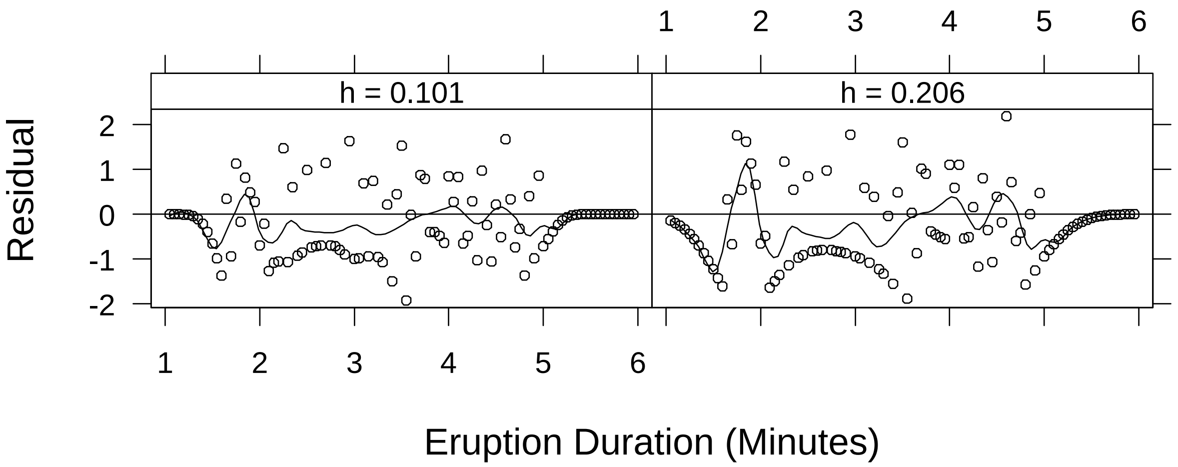

Figure 3: Residual plots for the Old Faithful Dataset, at $h = 0.101$ and $h = 0.206$. The data is rounded to the nearest 0.05, and fitted using a local constant Poisson regression. The residuals are the deviance residuals, and are smoothed by a local quadratic regression with span covering 15% of the data.

Figure 3 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 1 is real. Rather, it must be remembered that the LSCV, BCV and SJPI are attempting to minimize mean integrated squared error, and not to produce the correct number of peaks. The residual plots of Figure 3 suggest the left peak is trimmed by SJPI; a careful look at the fits in Figure 1 suggest the peak is also substantially broadened to regions where there is no data. This inflates the integrated squared error.

Let us give another example. We generate 500 observations from a distribution $10I+N(0,1)$, with I a random integer from $\{0,\ldots,9\}$. The results are shown in Figure 4. The SJPI and BCV methods oversmooth by a ridiculous extent, completely missing the obvious multimodality. The LCV and LSCV methods do far better, while AIC struggles, producing an undersmoothed estimate. Although only one simulation is reported here, similar patterns have been observed in other samples from this distribution; in particular, the SJPI and BCV method always dramatically oversmooth. Clearly, these methods are largely ignoring the data, and are quite incapable of selecting small bandwidths.

5 Do Plug-In Selectors have Better Asymptotic Performance?

The asymptotic performance of bandwidth selectors can be measured in terms of convergence to a theoretical bandwidth $h_0 = h_0(n)$ which minimizes a criterion such as MISE;

$$h_0 = \text{argmin} \int_{-\infty}^{\infty} \mathrm{E}\left((\hat{f}_h(x) - f(x))^2 \right) dx.$$

Since typically $h_0 \to 0$ as $n \to \infty$, it is usual to consider a relative rather than absolute measure of deviation for $\hat{h} = \hat{h}(n)$:

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

For classical approaches such as LSCV, the convergence rate is $O_p(n^{-1/10})$ (Hall and Marron, 1987). The BCV algorithm achieves the same rate, but with a smaller coefficient; see Scott (1992 p168). Other plug-in approaches achieve much faster rates of convergence: The SJPI method achieves a rate $O_p(n^{-5/14})$, while other more sophisticated variants achieve rate $O_p(n^{-1/2})$. Doesn't this provide strong evidence that the plug-in approach must be better, at least asymptotically?

The answer is no, since the results are derived under different assumptions which make the comparison meaningless. In particular, the plug-in results assume excessive smoothness about the true density, which make the comparison meaningless. The asymptotic results for SJPI and other plug-in approaches assume the underlying density has at least four derivatives. Under this assumption, the best possible rate of convergence of the density estimate is $O_p(n^{-4/9})$; see for example Stone (1980). This rate is achieved by local quadratic fitting.

However, the kernel density estimate achieves the slower rate $O_p(n^{-2/5})$, even when the theoretical best bandwidth is known, and hence is asymptotically inefficient under the four derivative assumption. Moreover, plug-in methods such as SJPI are making implicit use of local quadratic estimates to estimate the curvature of the density and hence the bias of the kernel estimate. These points have been made previously; for example Terrell (1992) and Cleveland and Loader (1995).

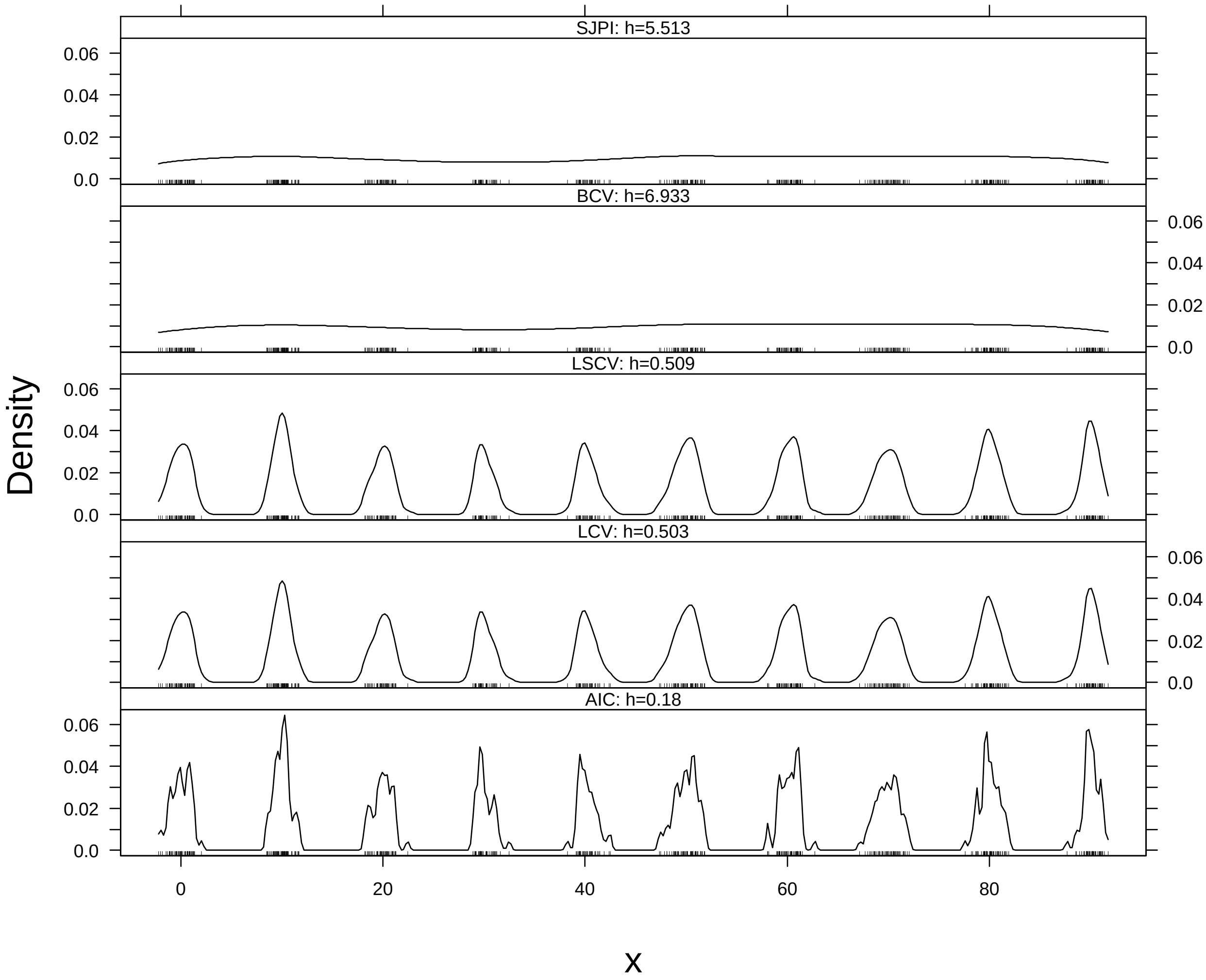

Figure 4: Density estimates for an equal mixture of ten normal distributions, using the AIC, LSCV and SJPI methods. The true peak height is about 0.04.

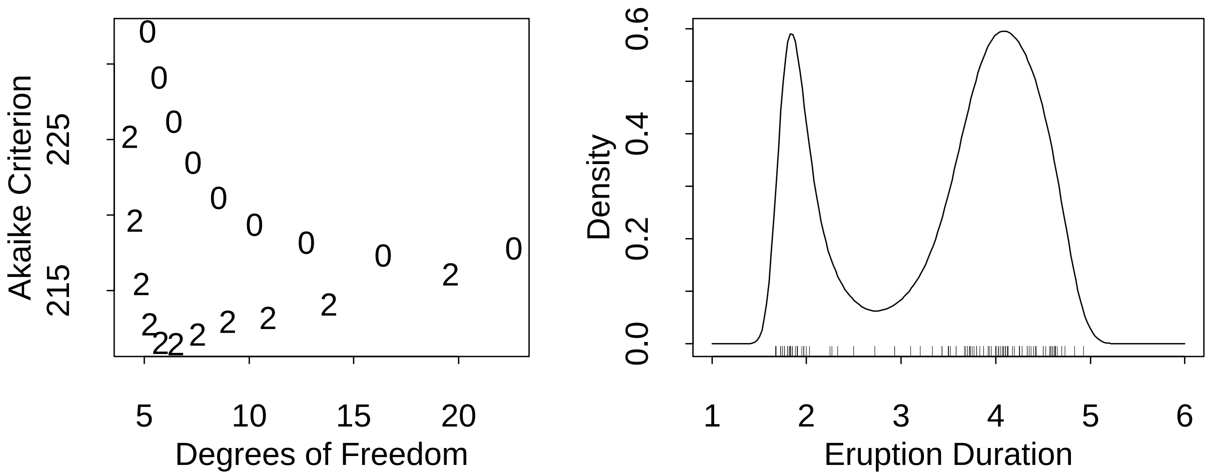

Figure 5: Fitting a local quadratic model to the Old Faithful dataset. At left, we plot Akaike's criterion, for local constant fitting (0) for h ranging from 0.05 to 0.25, and local quadratic fitting (2) for h ranging from 0.1 to 0.6 (h increases from right to left). The selected smoother is local quadratic with $h = 0.37$; the corresponding fit is shown on the right.

For meaningful comparison of plug-in approaches with cross-validation and AIC, the latter must also be allowed to consider local quadratic estimates. In Figure 5 we compare local constant and local log-quadratic fits for the Old Faithful dataset, and fit by the local likelihood method. In the left panel, we plot Akaike's criterion. The local log-quadratic method appears vastly superior. The criterion selects the local log-quadratic model with a bandwidth of $h = 0.370$; the fit obtained is shown in the right panel. Compared to the kernel fits obtained in Figure 1, this appears to wipe everything else off the map: we have more faithful reproduction of the left peak as it is supported by this dataset, without the apparently spurious noise.

Figure 6 shows a local log-quadratic fit to the ten-modal data of Figure 4. The smoothing parameter $h = 2.4$ is selected by AIC, and the resulting fit is far better than those obtained previously. There is a slight problem with filling in the troughs; this arises from a bounded parameter space in the log-quadratic fit when the Gaussian weight function is used. A compact weight function eliminates this problem.

In simulations, a local quadratic estimate with AIC cannot be guaranteed to beat a SJPI-equipped local constant fit, since AIC can be dominated by a few samples where it produces no estimate, or the problems can be well tuned to SJPI. Cleveland and Loader (1995) suggested matching the estimates on fitted degrees of freedom. Here we take an alternate approach, using the pilot estimate directly, which for the SJPI method is local quadratic (fourth order kernel) with bandwidth $\hat{\alpha}(\hat{h})/\sqrt{2}$, as determined by (2) and (3). The ISE of both the pilot estimate and the SJPI estimate can be computed exactly. For the simulations of Figure 2 with $\sigma = 0.219$, the 4th order pilot estimate has average ISE 0.0140; the SJPI estimate has average ISE 0.0151.

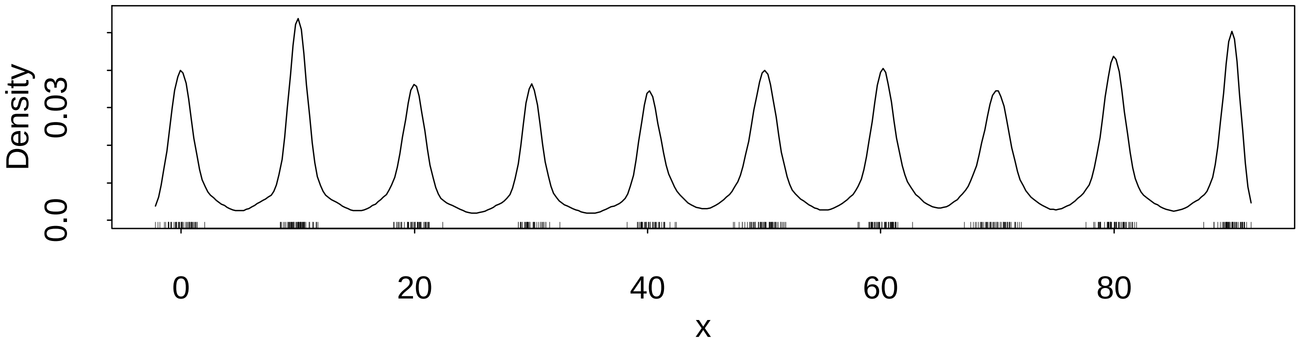

Figure 6: Local log-quadratic fit to the data of Figure 4. The smoothing parameter $h = 2.4$ is selected by AIC.

Under $\sigma = 0.070$, the average ISE is 0.0477 for the fourth order estimate, and 0.0501 for the SJPI estimate. In both cases, the SJPI estimate is beaten by it's own pilot 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. Although this has received less attention in the literature, most of the methods used in density estimation have analogs in the regression problem.

Plug-in methods for local polynomial estimates are considered in Fan and Gijbels (1994) and Ruppert, Sheather and Wand (1995). Both these papers use variants of Mallow's C_p to tune the 'pilot' estimates used in the plug-in steps. The use of plug-in is justified on the grounds of variance reduction; however, neither paper acknowledges that the pilot local cubic and local quartic estimates are usually better than the final local linear estimates, nor do they attempt to separate the contributions of the pilot and plug-in stages.

We consider regression problems with $n = 100$ and $x_i = (i-1)/99; i = 1\ldots,100$. The responses Y_i are generated as

$$Y_i = f(x_i) + \sigma Z_i$$

with $Z_i \sim N(0,1)$. The three regression functions are taken from Ruppert, Sheather and Wand (1995):

$$\begin{aligned} f_1(x) &= 1 - 48x + 218x^2 - 315x^3 + 145x^4; \sigma = 0.93 \\ f_2(x) &= \sin(5\pi x); \sigma = 0.5 \\ f_3(x) &= \sin(10\pi x); \sigma = 0.5 \end{aligned}$$

To measure the fit of local estimates, we consider the loss function

$$L(\hat{f}, f) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{f}(x_i) - f(x_i))^2$$

and the risk function

$$R(\hat{f}, f) = \mathrm{E}\left(L(\hat{f}, f)\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{E}\left((\hat{f}(x_i) - f(x_i))^2\right).$$

For equally spaced data, $n^{-1}L(\hat{f}, f)$ is a quadrature approximation to integrated squared error; this measure seems more appropriate for unequally spaced data since it automatically responds to sparseness of data. An adaptation of Mallow's (1973) C_p provides an unbiased estimate of risk; this was introduced by Rice (1984) for a variant of kernel (local constant) regression, and Cleveland and Devlin (1988) for local regression. Since the local regression is linear in the responses Y_i, the estimated values $\hat{f}(x_i)$ can be written in matrix form

$$\hat{f} = HY$$

for an $n \times n$ matrix H depending on the bandwidth. Then

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

is an unbiased estimate of $R(\hat{f}, f)$. Note that $\mathrm{tr}\,(H)$ is one definition of number of parameters of the smooth, corresponding to that used in density estimation earlier. To produce $C(h)$ plots, we we use the alternative definition $\mathrm{tr}\left(H^T H\right)$, which represents the variance component of $R(\hat{f}, f)$.

	f_1		f_2		f_3	
	h	$R(\hat{f}, f_1)$	h	$R(\hat{f}, f_2)$	h	$R(\hat{f}, f_3)$
deg=1	0.125	7.752	0.080	3.183	0.047	5.242
deg=2	0.313	6.034	0.192	2.534	0.104	4.352
deg=3	0.395	6.003	0.199	2.771	0.105	4.588
deg=4	∞	4.325	0.306	2.646	0.162	4.334
poly	4	4.325	10	2.763	17	4.678

Table 1: Optimal bandwidths and risks for estimating the three regression functions, for local linear, quadratic, cubic and quartic fitting. Also shown are the best fits that can be obtained by global polynomial regression, with the corresponding degree.

For the three regression functions, the true optimal bandwidths and risks are shown in Table 1, for fitting degrees 1 to 4. For f_1, the 'local' quartic fit becomes a global quartic fit, and this estimate beats the lower degrees quite substantially. For f_2 and f_3, there is little difference between local quadratic and local quartic; local cubic is slightly inferior due to boundary variability. The notable point in Table 1, however, is that local linear smoothing is substantially inferior to the higher degrees; it's optimal R is between 20% and 30% higher than for local quadratic. Also interesting is the comparison with high order polynomial fits; in all cases, local linear fitting is beaten by the optimal global polynomial.

Figure 7 shows the fit obtained to a dataset simulated from each model, with local linear an local quadratic fits at the theoretical risk-minimizing bandwidths. This shows real gains in using the higher order fits. While the absolute error isn't substantially reduced, the larger bandwidths used for the local quadratic fit translate into an estimate with less noise and fewer spurious wiggles.

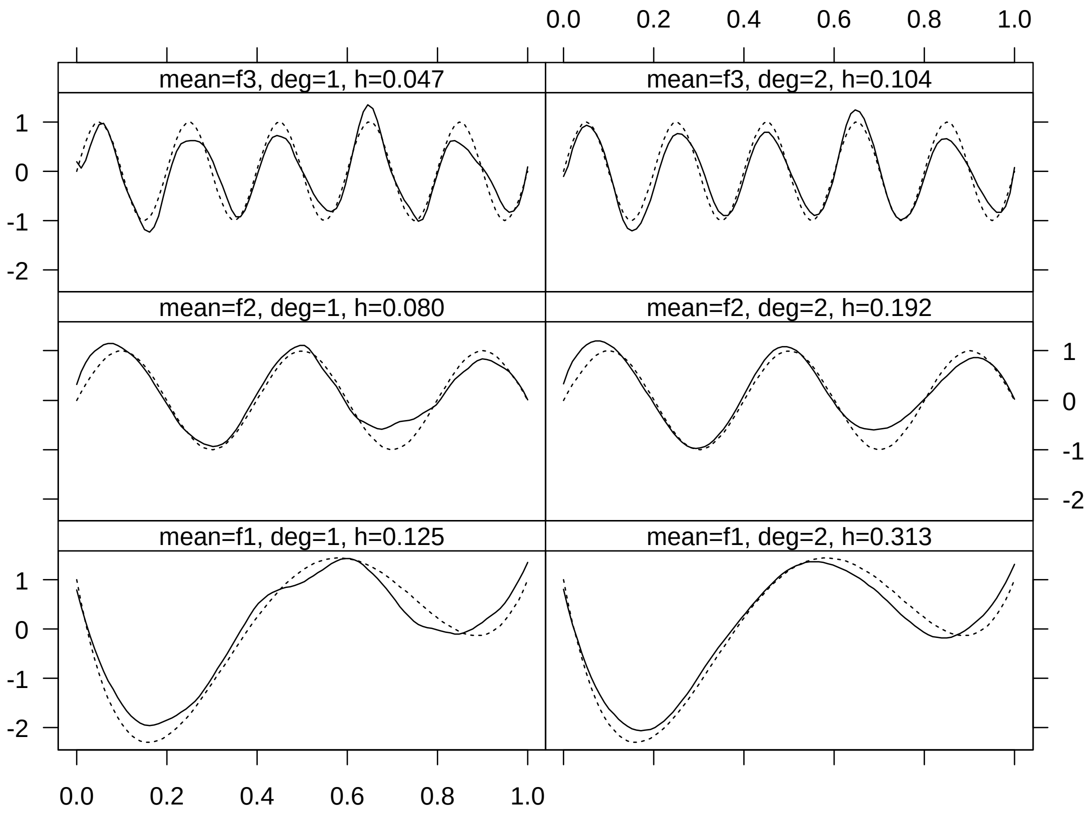

Figure 7: Estimates of the three regression functions for a random sample at the true risk-minimizing bandwidths. Local linear fits (left) and Local quadratic fits (right). True mean functions are shown by the dashed lines.

The average loss from 1000 simulations is shown in Table 2. For each simulation, the bandwidth is selected as the minimizer of $C(h)$. The compactly supported tricube weight function, $W(u) = (1 - |u|^3)^3 I(|u| < 1)$ is used. The variance σ^2 is estimated with $h = 0.05$; this gives 63.3 residual degrees of freedom. Comparing with Table 1, there is seen to be some loss compared with the theoretical minima for deg=2. However, the more important point is that the C_p equipped local quadratic fit still significantly beats the theoretical best local linear fit. Thus, *any* bandwidth selector for a local linear fit is beaten, regardless of it's rate of convergence to the true optimal bandwidth. Thus the plug-in steps of Ruppert, Sheather and Wand (1995), which are used to reduce a local quartic pilot estimate to a final local linear estimate, would

	f_1	f_2	f_3
$R(\hat{f}_{\hat{h}}, f)$	7.41	2.88	4.72
Std. Err.	0.13	0.04	0.05

Table 2: Fitting the three regression models using local quadratic fitting and $C(h)$. The average loss is estimated from 1000 simulations.

appear to be quite unhelpful.

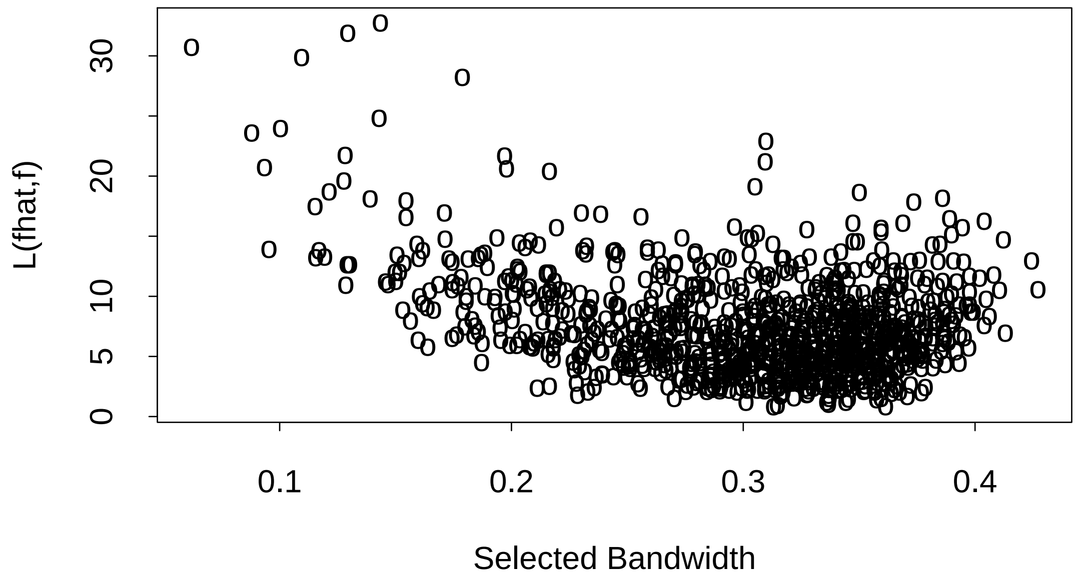

Figure 8: Selected bandwidths and actual losses for the 1000 simulations for f_1.

Figure 8 displays the selected bandwidths and actual values of $L(\hat{f}_{\hat{h}}, f)$ for the 1000 simulations for f_1. A similar picture, with different scales, results for f_2 and f_3.

To further understand the fairly high variability of the classical approach, let us take a closer look at a sample, corresponding to the worst case in Figure 8, which selected a bandwidth of 0.062, and resulted in a loss of 30.70. The dataset is shown in Figure 9. The dominant feature occurs between $x = 0.6$ and $x = 0.7$, with a sequence of several large positive residuals, compounded by a few mostly negative residuals on either side. A contributing problem - less obvious from the plot - is that the estimated standard deviation for this dataset is 0.842; rather less than the true 0.93.

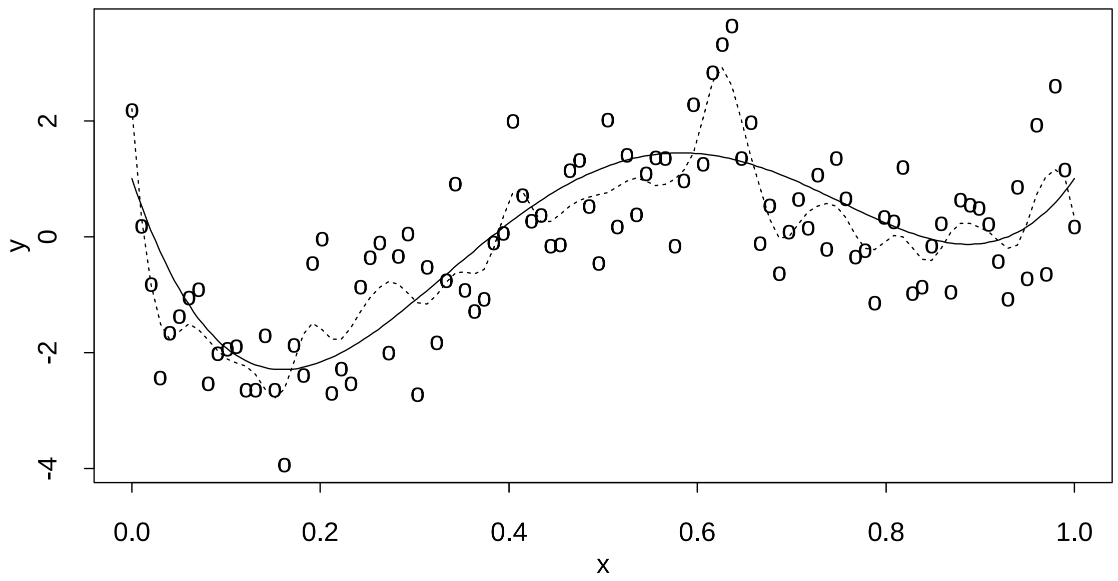

Figure 9: A selected dataset from the model $Y_i = f_1(x_i) + \epsilon_i$. The true mean $f_1(x)$ is shown by the solid line; the estimate with $h = 0.062$ is shown with the dashed line.

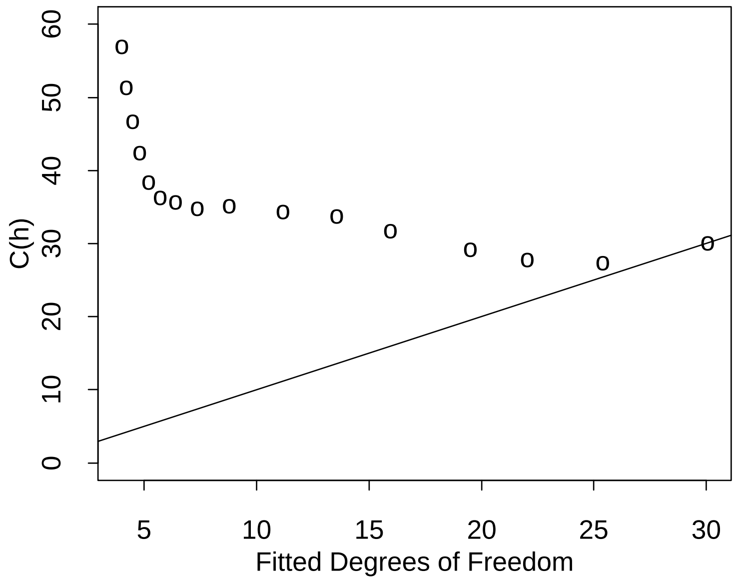

Figure 10: $C(h)$ plot for the data from Figure 9.

Figure 10 displays the $C(h)$ plot for this dataset. This immediately shows the bandwidth selection is having problems: Two local minima to the $C(h)$ function, and the $C(h)$ plot is very flat over a wide range of smoothing parameters. It is quite easy to design bandwidth selection algorithms which do much better on this type of dataset: for example, selecting the leftmost local minimum, or penalizing variance more heavily, or using less data-sensitive plug-in approaches, and essentially guaranteeing that larger bandwidths will be selected in problems such as this. This does not solve the problem: Such methods would be equally guaranteed to oversmooth if the peak was real. The real problem is that data-based bandwidth selection, like any model selection problem, is difficult; different 'true' models, requiring very different bandwidths, can give rise to very similar datasets.

7 Adaptive Fitting

The examples in this paper have used a single global bandwidth; thus following the course taken in the vast majority of work on bandwidth selection. This should not be considered an endorsement; it is easy to construct examples where a single global bandwidth is inappropriate; the Old Faithful dataset appears to fall into this category, especially when local constant fitting is used. Both 'classical' and 'plug-in' approaches to bandwidth selection can be performed in a locally adaptive manner, simply by targeting local, rather than global, goodness of fit criterion. The first implementation of locally adaptive local regression was Friedman and Steutzle (1982) using a local cross-validation approach.

The approach we take here is a development of that used in Cleveland and Loader (1995); in particular extending this to local likelihood regression models. Suppose Y_i has density $\exp(l(Y_i, \theta_i)$. We use the local AIC-type criterion

$$\frac{1}{\operatorname{tr}(W)} \sum_{i=1}^{n} w_i(x) \left(2(\sup_{\theta} l(Y_i, \theta) - l(Y_i, \hat{\theta}_i)) + 2\operatorname{tr}\left((X^T W \Psi'' X)^{-1} (X^T W^2 \Psi'' X) \right) \right) \qquad (4)$$

with X the $n \times p$ design matrix; W a diagonal matrix of the weights and Ψ'' a diagonal matrix with entries $\ddot{l}(Y_i, \hat{\theta}_i)$. At each fitting point the bandwidth is selected to minimize this criterion; the fitting points are selected according to a recursive tree algorithm as discussed in Loader (1994) and Cleveland and Loader (1995).

Figure 11 applies this method to the Old Faithful dataset, using both local linear and local quadratic fitting. This uses a local Poisson regression using the pooled version of the data, with prior weights 107×0.05 to correctly estimate the density. These fits support our earlier conclusions: The leftmost peak is quite sharp, and as tall or taller than the right peak. Under the log-quadratic fit, the left peak is substantially taller than any of the previous fits. The fitting points and bandwidths are shown inthe left panels of Figure 11; these are not comparable with the other bandwidths used for this dataset, since the compactly supported tricube weight function is used here.

As a second example, we consider the data of Figure 9. Using the Gaussian likelihood, the criterion (4) requires an estimate of the residual standard deviation; we consider both the previously estimated $\hat{\sigma} = 0.842$, and the true $\sigma = 0.93$. The results of locally adaptive local quadratic fits are shown in Figure 12.

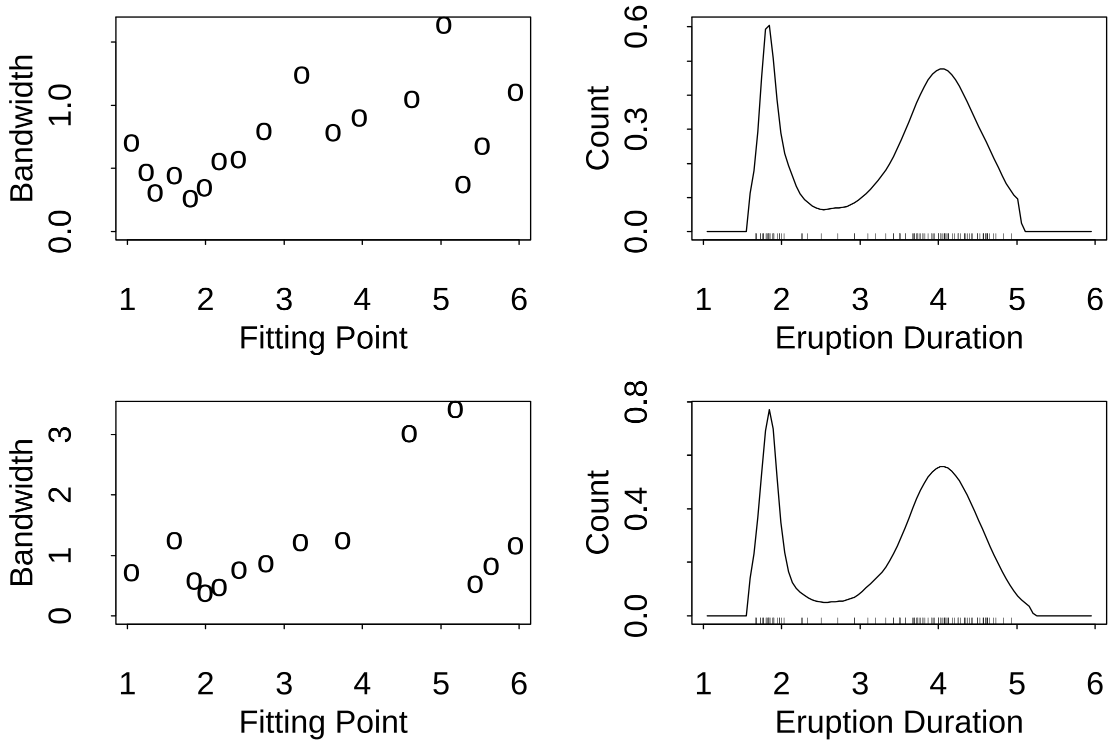

Figure 11: Locally adaptive Poisson regression for the Old Faithful dataset. Local log-linear fitting (top) and Local log-quadratic fitting (bottom).

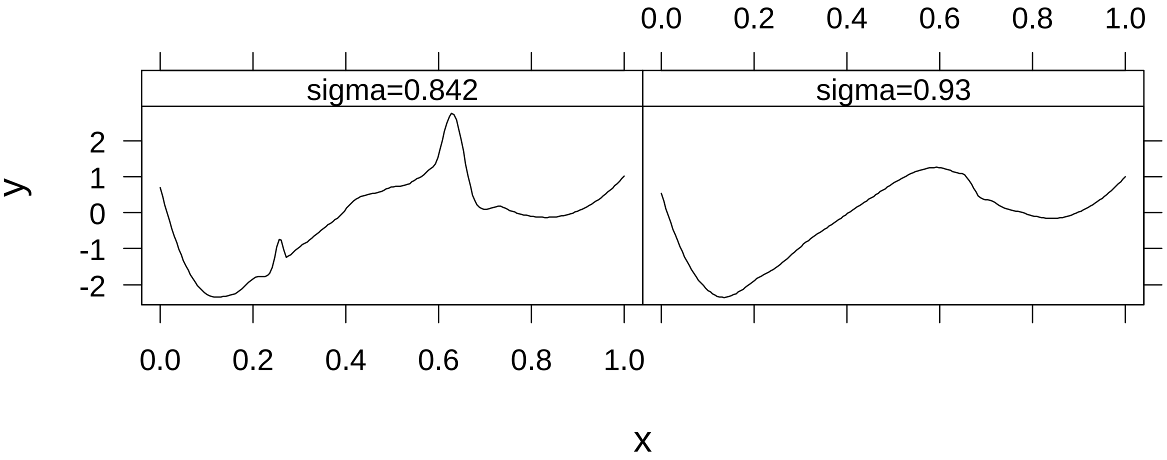

Figure 12: Locally adaptive fits to the data of Figure 9, using two values of σ: The estimated $\sigma = 0.842$, and the true $\sigma = 0.93$.

At the estimated σ, the false peak around 0.6 is faithfully reproduced; a smaller peak around 0.2 also shows up. At the slightly larger $\sigma = 0.93$, these false peaks are almost entirely obliterated. This illustrates a limitation of adaptive fitting; a fully adaptive algorithm will be quite sensitive to small changes in tuning parameters, and faced with a real dataset, we cannot conclude that structure is real or random based solely on adaptive smoothing procedures.

8 Conclusions

We have studied a wide range of bandwidth selectors, on both real and simulated data. When the results are carefully analyzed, the much touted plug-in approaches have fared rather poorly, being largely tuned 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 better 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 it 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 for 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 p163) incorrectly conclude there is a problem with LCV.

With the Old Faithful dataset, simulations based on a smoothed bootstrap approach, residual diagnostics, higher order fits and adaptive 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 not performing 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 prefered, 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 are typically meaningless. 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 estimate; if classical approaches are also allowed to consider higher order methods, better *estimates* results. Thirdly, plug-in methods are not rescued by asymptotic analysis showing better rates of convergence; these require assumptions about the underlying function that make the underlying estimate asymptotically inefficient, regardless of how good the bandwidth selector is.

We have emphasized the importance of not blindly relying on any bandwidth selector to automatically produce the right bandwidth. If one just applies 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, the C_p or AIC plot 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 a $Ah^4 + B/(nh)$ profile on the integrated squared error in such cases, discard this information.

A Appendix: Computation and Software Details

The simulations and local fits have all been performed in the author's LOCFIT software, available through the WWW at the address

`http://netlib.att.com/netlib/att/stat/prog/index.html`

The local fits can be performed through the S interface, while the simulations require running the C version of the code directly. A typical S call to `locfit` will look like:

```
> fit <- locfit(~geyser,alpha=2.5*0.101,bt="fix",kern="gauss",
+   deg=0,eval="grid",mg=100,flim=c(1,6))
> plot(predict(fit))
```

This produces the kernel density fit for the geyser data, with the bandwidth $h = 0.101$. The smoothing parameter, `alpha` includes a factor 2.5; this is internal to the `LOCFIT` code to make the Gaussian weight function look more like the default compactly supported weight functions. `bt="fix"` uses the same bandwidth everywhere. `deg=0` specifies local constant fitting. `eval="grid"` and `mg=100` force direct fitting on a grid of 100 points, rather than the faster approximations used by default. Adaptive fitting can be specified as

```
> fit <- locfit(count~duration,w=107*0.05,bt="ad",family="poisson",alpha=2)
> plot(fit)
```

Here, `count` and `duration` represent the pooled data. `family="poisson"` ensures the Poisson likelihood is used; `w=107*0.05` gives prior weights for the counts; effectively, this scales the fit to estimate the density. `bd="ad"` requests the locally adaptive fit; the `alpha` argument gives the variance penalty.

The simulations for the bandwidth selections are all performed by running the C version of LOCFIT directly; details of the commands will be made available through the web page above. The Sheather-Jones plug-in method naturally solves an equation (3) for the bandwidth h; for the other methods, we can differentiate the objective function to find such an equation. Our implementation searches for solutions over an interval $[h_0, h_1]$, first by evaluating the equation over a logarithmic grid of values, then applying a hybrid bisection-secant method on any interval where the equation changes sign. For the simulations in Figure 2, we took $h_0 = 0.05$ and $h_1 = 0.5$.

All methods can produce multiple solutions; some of the methods may produce no solution. We choose between local minima based on the smallest value of the objective function. For the SJPI method there isn't an objective function being directly minimized; instead, we use the estimated MISE at each candidate h as the tie-breaker.

For the local local regression simulations, differentiation of the $C(h)$ objective function is rather tricky. Hence our simulations directly minimize the objective function using the golden search technique.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csàki, F. , editors, *Second International Symposium on Information Theory* pages 267–281 Budapest. Akademia Kiadó.

[2] Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics* **39**, 357–365.

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

[4] Chiu, S. T. (1991). Bandwidth selection for kernel density estimation. *The Annals of Statistics* **19**, 1883–1905.

[5] Cleveland, W. S. (1993). *Visualizing Data.* Hobart Press, Summit, New Jersey.

[6] Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**, 596–610.

[7] Cleveland, W. S. and Loader, C. R. (1995). Smoothing by local regression: Principles and methods. *Computational Statistics* **11**.

[8] Duin, R. P. W. (1976). On the choice of smoothing parameter for Parzen estimators of probability density functions. *IEEE Transactions on Computing* **C-25**, 1175–1179.

[9] Fan, J. and Gijbels, I. (1994). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B* **57**, 371–394.

[10] Friedman, J. H. and Steutzle, W. (1982). Smoothing of scatterplots. Technical Report, Orion 3 Department of Statistics, Stanford Universtity.

[11] Habbema, J. D. F. ; Hermans, J. and Van Der Broek, K. (1974). A stepwise discriminant analysis program using density estimation. In Bruckman, G. , editor, *COMPSTAT 1974, Proceedings in Computational Statistics* pages 101–110 Vienna. Physica-Verlag.

[12] Hall, P. and Marron, J. S. (1987). Extent to which least-squares cross-validation minimizes integrated squared error in nonparametric density estimation. *Probability Theory and Related Fields* **74**, 567–581.

[13] 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.

[14] Henderson, R. (1916). Note on graduation by adjusted average. *Actuarial Society of America* **17**, 43–48.

[15] Hjort, N. L. and Jones, M. C. (1994). Locally parametric nonparametric density estimation. Technical report, University of Oslo.

[16] Kanazawa, Y. (1993). Hellinger distance and Kullback-Leibler loss for the kernel density estimator. *Statistics and Probability Letters* **18**, 315–321.

[17] Lejeune, M. and Sarda, P. (1992). Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis* **14**, 457–471.

[18] Loader, C. R. (1993). Local likelihood density estimation. Technical report, AT&T Bell Laboratories.

[19] Loader, C. R. (1994). Computing nonparametic function estimates. In *Computing Science and Statistics: Proceedings of the 26th Symposium on the Interface* pages 356–361.

[20] Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

[21] Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics* **20**, 712–736.

[22] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

[23] Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association* **85**, 66–72.

[24] Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics* **12**, 1215–1230.

[25] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics* **27**, 832–837.

[26] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9**, 65–78.

[27] Ruppert, D. ; Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90**.

[28] Schuster, E. F. and Gregory, G. G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In *Computing Science and Statistics: Proceedings of the 13th Symposium on the Interface* pages 295–298.

[29] Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley, New York.

[30] Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association* **82**, 1131–1146.

[31] Sheather, S. J. (1992). The performance of six popular bandwidth selection methods on some real datasets. *Computational Statistics* **7**, 225–250.

[32] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B* **53**, 683–690.

[33] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[34] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics* **8**, 1348–1360.

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

[36] Terrell, G. R. (1992). Comment on papers by sheather; and park and turlach. *Computational Statistics* **7**, 275–277.

[37] Tibshirani, R. J. and Hastie, T. J. (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82**, 559–567.

[38] Woodroofe, M. (1970). On choosing a delta sequence. *The Annals of Mathematical Statistics* **41**, 1665–1671.