Comments

by S.J. Sheather, M.P. Wand, M.S. Smith and R. Kohn

Australian Graduate School of Management, University of New South Wales, Sydney, 2052, Australia

1 Introduction

We welcome the opportunity to comment on a set of papers seeking to advance nonparametric regression techniques. Especially as the authors of all three papers have already made substantial and influential contributions in this field. We agree with the authors on many points, both in principle and in substance. However, we have chosen to concentrate our comments on those aspects of the papers that we disagree with or that we find ambiguous. We take this critical role as we think it will be more interesting for readers and elicit further clarification from the authors. Most importantly, it points out that there are enormous opportunities for research in this field. Despite the large volume of work in the 1980's and 1990's we believe that the subject is still in its infancy.

2 Comments on Cleveland and Loader

Cleveland and Loader provide a useful outline of the history of local regression, introduce several innovative ideas and give their views on the state of the field. They have chosen to discuss their local regression approach mainly for the one dimensional case but it is clear from their other work that the great strength of their approach is its ability to generalize to higher dimensions. Despite this generality we believe that some fundamental aspects of their approach can be improved. Our views on their procedures are outlined in the following sections.

W. Härdle et al. (eds.), *Statistical Theory and Computational Aspects of Smoothing*

2.1 Graphical diagnostic procedures

The examination of residuals is now widely accepted as an important step in parametric regression analysis. Cleveland and Loader demonstrate the importance of graphical diagnostic methods in the nonparametric regression setting and recommend that these tools be routinely used. We endorse this practice and the authors' views that "residual plots provide an exceedingly powerful diagnostic that nicely complements a (model) selection criterion" (Section 8) and that "when we have a final adaptive fit in hand, it is critical to subject it to graphical diagnostics to study the performance of the fit" (Section 1).

Both Cleveland (1993) and the current article use a loess fit to plots of residuals against fitted values to detect lack of fit. Though looking at the residuals from a nonparametric fit is an extremely useful process, we question the validity of judging the appropriateness of a nonparametric regression estimate by fitting another such curve to the residuals.

The problem is how to choose the fit for the residual plot. The same local polynomial fit is not appropriate for both the original data and the residuals because if this fit allowed structure to go undetected in the original data, then it is very unlikely to capture structure in the residuals. Cleveland (1993, Section 3.6) and Section 8.1 of the current paper suggest a local linear fit to the residuals using an arbitrary smoothing parameter such as $\alpha = 1/3$. The inadequacy of this recommendation as a general principle is illustrated using the authors simulated example.

We generated 100 observations from the model given in section 8.2 and used loess to obtain locally quadratic fits with smoothing parameters increasing in equal percentage steps as suggested by the authors. These are given on the left hand side of Figure 1, along with local linear fits to each set of residuals on the right using $\alpha = 1/3$. It can be seen that the fits to the data improve as α increases, whereas the fits to the residuals deteriorate. If such residual fits were used as a guide, the roughest fit (where $\alpha = 0.2$) would be regarded as the best, while the best fit (where $\alpha = 0.675$) would be classified as the very worst. This effect is more disturbing when one considers that in higher dimensions simple scatter plots of the data cannot be used to resolve this apparent contradiction.

An alternative way of comparing the local fits based on different values of α is based on F-tests. Cleveland, Grosse and Shu (1992, page 331) provide an example of the use of this procedure. What concerns us is that these F-tests are based on the assumption of no bias and in general this is not the case.

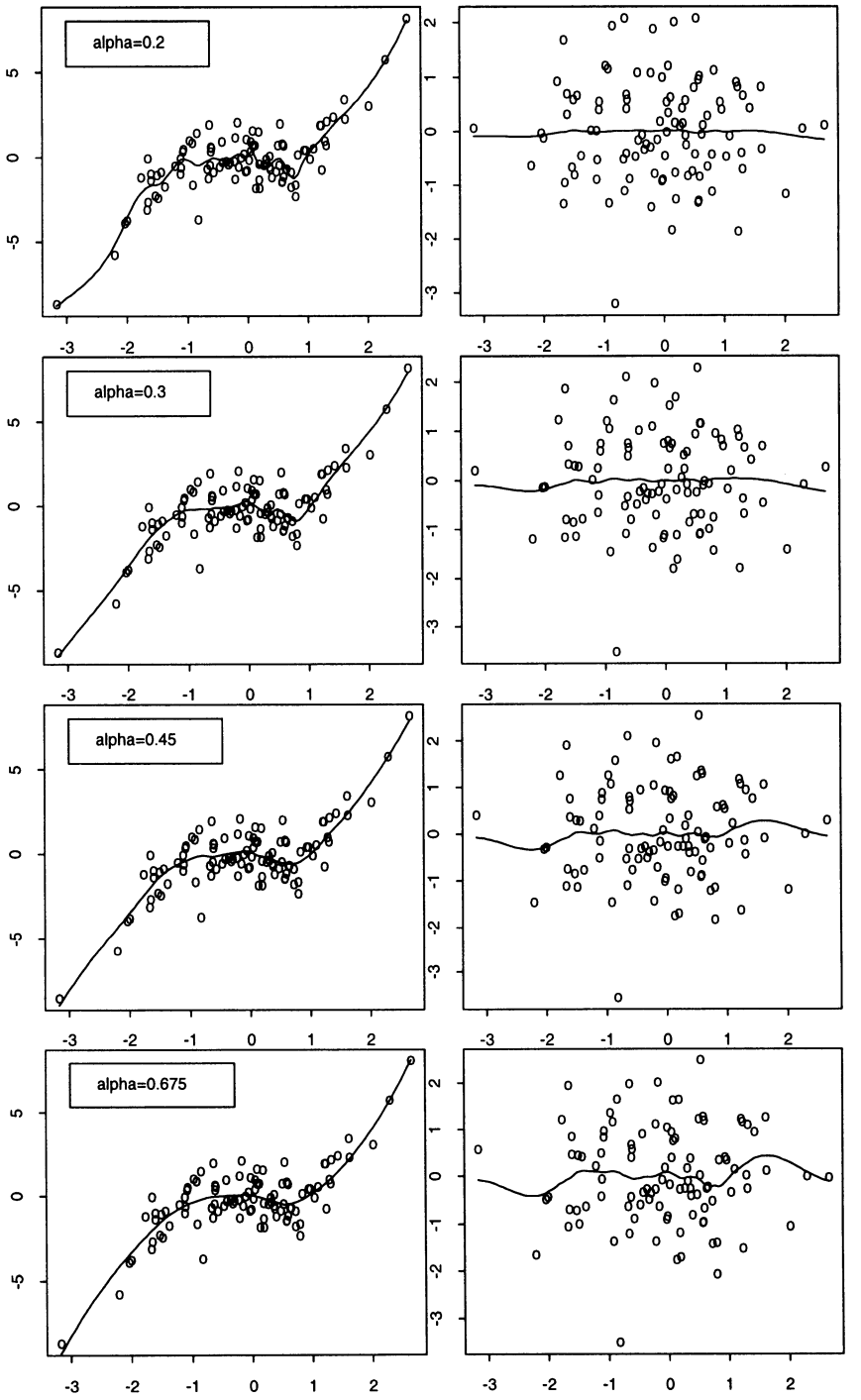

Figure 1. Local quadratic fits to data for various values of α (left) and local linear fits to the resulting residuals with α fixed at 1/3

We agree with Cleveland and Loader's practice of displaying fits based on different values of the smoothing parameter. The point at which we differ is that we would typically recommend as the default bandwidth a fixed plug-in bandwidth, h_{PI} that minimises an error criterion like estimated mean square error. The reason for this, as Cleveland and Loader mention in Section 10.3, is that plug-in bandwidths work quite well in the sense that they produce bandwidths quite close to the value that minimizes mean square error. Thus values of h smaller than h_{PI} will produce curve estimates with smaller bias but larger variance than the estimate based on h_{PI}. The opposite is true for values of h larger than h_{PI}. As Cleveland and Devlin (1988) point out, low bias is critical when the estimate is used for graphical exploration, since the eyes can tolerate some noise but cannot recover a missed effect. The curve estimate with $h = h_{PI}$ can be used as a reference point, since it balances the bias and variance of the curve estimate. This makes the fits with the different values of h straightforward to interpret. In addition, it is usually clear from a plot of the curve estimate with $h = h_{PI}$ whether an adaptive rather than a global bandwidth is needed for the given data set.

2.2 Nearest-neighbour bandwidths

The implicit adaptation of the bandwidth through the use of nearest-neighbour distances is a convenient and often successful means of addressing design sparseness. Nearest-neighbour bandwidths asymptotically stabilise the variance of the resulting curve estimate. However, this approach has a serious shortcoming that detracts from its use as a general principle (e.g. as a default in computer packages). The problem is that a global nearest-neighbour bandwidth pays no attention to bias, that is, curvature in the regression function. Since the adaptation is based solely on the positioning of the design points, for a given α and design the global nearest-neighbour bandwidth is the same whether the underlying curve is a straight line or a high frequency sine curve. This can sometimes lead to an unattractive curve estimate, especially when the design is skewed.

An example of such behaviour is given in Figure 2 where a loess estimate is plotted against a fixed bandwidth estimate. The (X_i, Y_i) regression data were generated according to

$$X_i = 1 - U_i^{0.7}; \quad Y_i = \sin(2\pi X_i^3) + 0.2Z_i, \quad i = 1, \ldots, 200,$$

where the U_i's are uniform $[0, 1]$ variates and the Z_i's are standard normal. The dashed curve is a local linear fixed bandwidth estimate, with bandwidth chosen using the rule of Ruppert, Sheather and Wand (1995). The solid curve is the estimate obtained using the `S-PLUS` `loess()` function with a span of 0.12, chosen so that the each smooth performed about the same in the valley on the right. Default values were used for the other `loess()` parameters. The cost of the reasonably good performance in the valley is the aberrant

behaviour of loess on the left hand side, despite the fact that there are more data here and less structure to resolve. Increasing the span value doesn't help since it leads to the valley being smoothed away. Such disconcerting behaviour is due simply to the fact that there are more data where there is less curvature, and vice versa.

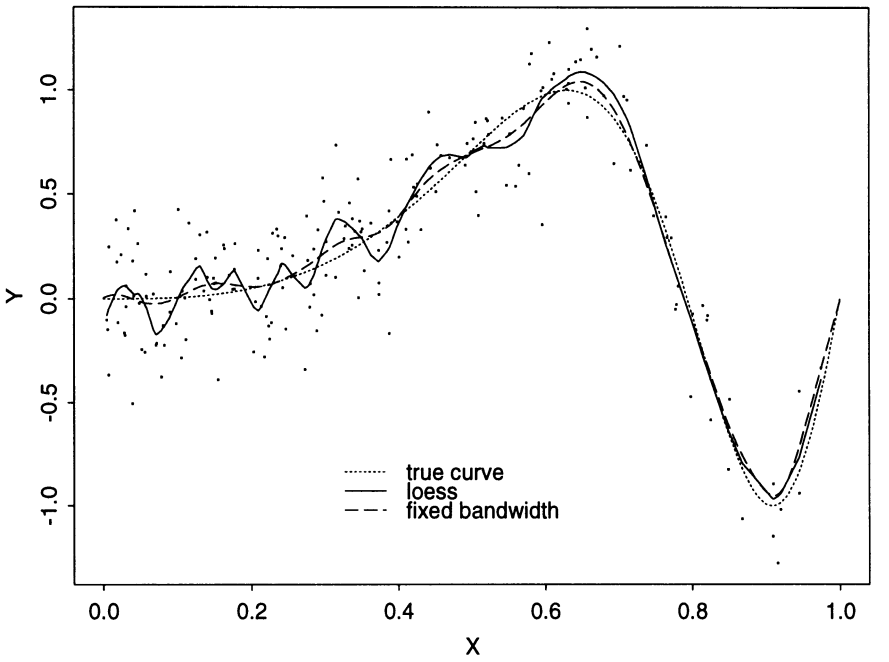

Figure 2. Nearest-neighbour (loess) and fixed bandwidth estimates for a simulated example.

The overall performance of fixed bandwidth estimate in this case is certainly more pleasing and leaves us somewhat skeptical about the authors' claim that "nearest neighbors appear to perform better overall in applications" compared to fixed bandwidths. There is also room for improvement in the fixed bandwidth estimate since it does not properly resolve the peak and valley and has some slight spurious bumps on the left hand side. Clearly some sort of local adaptation of the bandwidth is required for data such as these. We look forward to seeing research into further methodology for doing this that addresses both the sparsity and curvature issues.

2.3 Equivalent quadratic smoothers

In Section 7, Cleveland and Loader argue that local quadratic fitting does better than local linear and local constant fitting "when there is a rapid

change in slope, for example, a local minimum or maximum". Then in Section 10 they recommend that one compare the fit of a local constant or linear fit with that of a local quadratic regression based on a bandwidth with the same variance as the local constant or linear fit.

We undertook the suggested comparison using a local linear fit based on the plug-in bandwidth rule of Ruppert, Sheather and Wand (1995). The regression functions considered were $f_1(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ and $f_2(x) = \sin(10\pi x)$. These were chosen since they each have at least one minimum and maximum. The x's were generated as uniform (0,1) random variables and the errors as normal random variables with mean 0 and standard deviation equal to one quarter of the range of f. The goodness of fit criterion used was the mean square error summed over the x's, which Cleveland and Loader denote by $R(\hat{f}, f)$ (equation (1) in Section 8.2).

In order to compare the local linear fit ($\hat{f}^L$) and the local quadratic fit ($\hat{f}^Q$) we computed their relative efficiency, that is, the ratio of their mean square errors, $RE = R(\hat{f}^L, f)/R(\hat{f}^Q, f)$ for 100 samples of size $n = 100$ and 10,000. Estimates of the mean relative efficiency and associated 95% confidence limits are given in Table 1. The confidence limits were calculated using a normal approximation to the distribution of mean relative efficiency.

Function	Sample Size	Mean RE	95% CI for Mean RE
f_1	100	0.9729	(0.9707, 0.9752)
f_1	10000	0.9833	(0.9832, 0.9834)
f_2	100	1.1216	(1.1140, 1.1292)
f_2	10000	1.5428	(1.4862, 1.5993)

Table 1. Relative efficiency of local quadratic to local linear fits

In terms of mean square error, the local quadratic fit of the sine curve (f_2) provides an improvement over the local linear fit, especially for the larger sample size. However, this improvement is not evident for the quartic (f_1). On the other hand, the loss of efficiency due to using a local quadratic fit is small and thus the idea of converting a local linear bandwidth into a local quadratic one is worthy of further investigation.

The results for f_1 in Table 1 illustrate a relatively well-known phenomenon that in some cases very large sample sizes are needed in order to see the benefits of higher order methods, which improve rates of convergence from $O_p(n^{-2/5})$ to $O_p(n^{-4/9})$.

2.4 Local adaptability

Methodologies that employ a single global smoothing parameter are not appropriate for use on functions that take on highly dissimilar forms on different sections of the domain of the independent variable. What is required is a degree of *local adaptability* – something that can be achieved by varying the

bandwidth over the domain, as the authors outline in Section 9.2 using a local version of Mallows C_p. Nevertheless, it can be computationally difficult to locally estimate such bandwidths at each point in the design. Instead, some authors (e.g., Härdle and Marron, 1995) split up the design domain into a series of regions and a different bandwidth is used in each of these regions. The problem then becomes one of choosing these regions.

An alternative strategy that is computationally tractable and avoids ad hoc domain splitting rules is to link regression splines with Bayesian variable selection within the computational framework of the Gibbs sampler as outlined in Smith and Kohn (1994). Here, a large number of potential knots are introduced along the domain of the independent variable (for example, 1 knot every 5 design points) and a significant subset determined, resulting in a regression spline that is both locally adaptive and smooth. The following example clearly illustrates the importance of local adaptability in nonparametric regression.

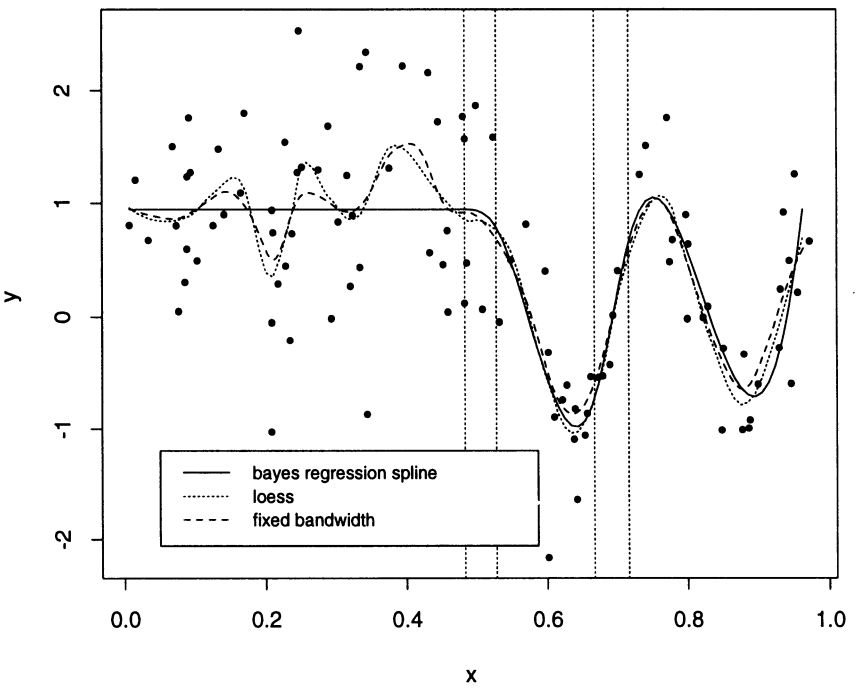

Figure 3. Nearest-neighbour (loess), fixed bandwidth and Bayes estimates for a simulated example.

One hundred observations were generated from a regression model with

the x's uniformly distributed on (0,1) and $f(x) = 1$ if $x < 0.5$ and $f(x) = \cos(8\pi x)$ if $x > 0.5$. The errors were distributed normally with mean zero and standard deviation 0.75. A kernel based local linear smooth with automatic bandwidth selection (Ruppert, Sheather and Wand, 1995) was fit to the data, along with a local quadratic fit given by loess with $\alpha = 0.2$. Figure 3 shows that both fits capture the cosine curve well, but are not at all smooth in on the rest of the domain. No global bandwidth value will successfully produce a fit that is both smooth and captures the oscillations of the cosine curve well. However, the locally adaptive Bayesian regression spline (the solid curve) is both smooth and possesses low bias. The four significant knots found for these data are marked as vertical lines in Figure 3. We look forward to the opportunity to compare the performance of this Bayesian regression spline with the locally adaptive procedure of Cleveland and Loader.

3 Comments on Marron

Marron stresses the importance of reliable automatic smoothing parameter selection. It seems to us that it is time that the current, often arbitrary, default choices for smoothing parameters common in software packages were replaced by what Marron describes as the "second generation of bandwidth selection methods". Other situations which we would add to the authors' list in Section 3 are:

- analysis of discrete regression data sets, such as those with binary responses, and
- multivariate smoothers.

In these cases it is not as straightforward to use graphical techniques to assess the fit.

We agree with the author's view regarding the importance of evidence for performance of smoothing parameter selection rooted in E1, E2 and E3. High quality smoothing parameter choice is a very subtle and challenging problem. It cannot be properly developed and evaluated without the use of a sensible combination of all of these research tools.

4 Comments on Seifert and Gasser

It is clear that fixed bandwidth local polynomial smoothers require modifications to guard against degeneracies in the local fitting process. Seifert and Gasser have made an important first step into the development of such modifications. The results from their bandwidth inflation and ridge regression ideas are very encouraging.

Another lesson that is apparent from their work (e.g. Figure 8) is that the Gaussian kernel weight can go a long way to alleviating degeneracy problems.

We are not so much concerned about claims by the authors that Gaussian weights are "computationally slow" because of the existence of fast computational methods other than theirs (see e.g. Fan and Marron, 1994) that do not impose restrictions on the kernel type.

An important question that arises from Section 5 of this article is: at what point is a design become so sparse that a smoothing technique should not be used at all? Figure 2 shows some contrived designs over $[0, 1]$ generated from the beta mixture density $\frac{1}{2}\text{Beta}(1, s) + \frac{1}{2}\text{Beta}(s, 1)$ for increasing values of s. The regression curve is $m(x) = \sin(8\pi x)$, shown without noise for clarity.

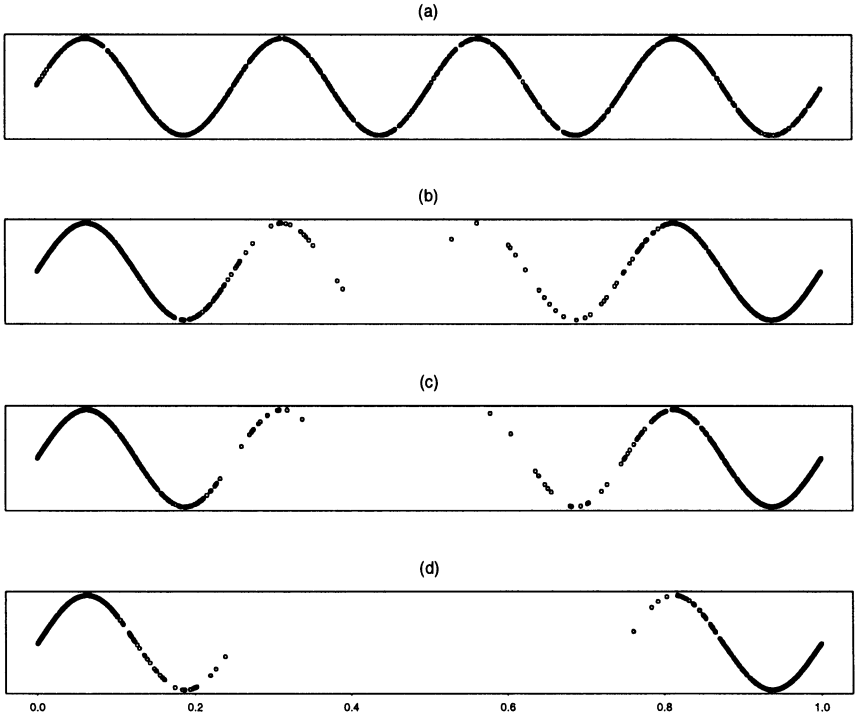

Figure 4. Regression designs with increasing sparsity in the centre.

For design (a) there is clearly no problem with application of a smooth over the whole of $[0, 1]$. At the other extreme most would agree that for design (d) the data is non-existent in the middle part of $[0, 1]$ and that it is pointless to try to use local regression here. Designs (b) and (c) are not as clear-cut. There are some data in the middle, but are there enough for one to expect reasonable recovery of the underlying regression function? It would seem that methodology for estimating when the design is thick enough to apply a smoothing technique be required. Density estimation could lead to effective solutions. Such technology would be even more important for two- and three-dimensional designs where there are more obscure ways in which a design can "peter out".

References

[1] Fan, J. and Marron, J.S. (1994). Fast implementations of nonparametric curve estimates. *J. Comput. Graph. Statist.*, **3**, 35–56.

[2] Härdle, W. and Marron, J.S. (1995). Fast and simple scatterplot smoothing. *Comput. Statist. Data Analysis*, **18**, to appear.

[3] Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression, *J. Amer. Statist. Assoc.*, **90**, to appear.

[4] Smith, M. and Kohn, R. (1994). Robust nonparametric regression with automatic data transformation and variable selection. Working Paper, 94-026, Australian Graduate School of Management, University of New South Wales.