Rejoinder

J. S. Marron

Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA

1 Comments on "the debate"

Many people are aware of some substantial recent controversy in smoothing. The current discussion is partly a consequence of this. I have observed a number of differing ideas as to "what the debate is all about". Of these various ideas, I view the main ones as being:

I1 Should one use local polynomial (degree higher than 0) or local constant (i.e. conventional kernel methods) smoothers?

I2 Should one choose "smoothing windows" using the same width everywhere, or a width based on nearest neighbor considerations?

I3 Is LO(W)ESS the best possible way to do smoothing?

I4 Is mathematical analysis a useful tool in statistics?

I5 Are there effective methods of choosing the bandwidth from the data? Are these good enough to become defaults in software packages?

Here are some personal comments on each of these:

1.1 Kernels vs. local polynomials

There is less disagreement here than some people seem to expect. My personal current default, e.g. what I use when faced with a new data set to smooth, is the local linear. This choice is based on its excellent handling of boundary problems and random design issues, as made clear in a number of ways by several authors.

However, I understand and respect those who select any of local constant, quadratic or cubic (each has its reasons for being used, and situations where it is "best"). Since local constant is the one that seems to be a favorite target, I will comment further on that.

While it has some serious weaknesses (which is why I currently don't use it myself as a general rule), there are also situations where its use is appropriate, and even "best". In particular, when there is an equally spaced design, and boundary issues are not important (e.g. a "circular design", or the regression is flat at the boundaries, or boundaries are not relevant to the problem being addressed), then such an estimator can be "best" in terms of: simplicity, in terms of interpretability (i.e. "what is it doing to the data"), and in terms of being "close to the true regression" (e.g. will have the best performance in a simulation, or best theoretical properties).

Another context in which such estimators are sensible is when doing mathematical analysis. This is because simple moving averages are very tractable to analyze, with the benefit that one can analyze more deeply (e.g. tackle harder problems). A first class analysis would go on to show that the lessons learned hold also for the more useful higher degree local polynomials. But in many contexts it is sensible to first look at the simplest case: degree 0, and in some instances such work by itself would be publishable in my view.

1.2 Width of smoothing windows

Actually there are at least three general classes of ideas here:

1. fixed width,
2. nearest neighbors,
3. location adaptive bandwidths that use more of the data than simply the design.

Each of these has its strengths, its weaknesses, and its advocates.

The appeal of the fixed width approach comes from its simplicity and interpretability. Its weaknesses are that some data sets can show structure at different amounts of smoothing (usually in different locations), and it can behave erratically (even to the point of numerical instability) in regions where the data are sparse.

I view nearest neighbor methods as a quick and dirty fix to the data sparsity problem. However, they are very weak in terms of interpretability (just how does the resulting curve relate to the data?), and adapt only to the design (the X's, not the Y's) , while ignoring other reasons for varying the amount of smoothing, such as curvature in the regression function and heteroscedasticity. This point is discussed by Hall and Turlach, and made quite clear in section 2.2 of the discussion of Sheather, Wand, Smith and

Kohn, where the nearest neighbor based method LO(W)ESS has difficulties because it ignores curvature in the regression function.

Higher tech methods of location adaptive smoothing have more potential for "getting the amount of smoothing right", than either of the others. The example in section 2.4 of the discussion of Sheather, Wand, Smith and Kohn makes this point very clear. This has motivated useful research on this topic (and I am involved in this). However, higher tech methods have their attendant costs. The obvious costs are in terms of complexity, and interpretability. A less well understood drawback is the ease with which spurious features can be introduced, and important ones be eliminated. See Marron and Udina (1995) for discussion of this issue, and an environment which allows user choice of a location adaptive bandwidth function in kernel density estimation.

When I analyze data, I place a very high priority on interpretability, so my personal default method is a fixed width approach.

Any problems with data sparsity are readily apparent when one tries several bandwidths. When the sparsity is so severe it causes numerical problems, my current personal preference is for a simple ridge regression approach, which can often work well for surprisingly small ridge values, see Hall and Marron (1995), However, I plan to rethink this in view of the ideas in the contribution of Gasser and Seifert and the discussion of Hall and Turlach.

In my view, the problem of different important features being visible at different smoothing levels is more important than data sparsity. I do not believe that any location adaptive method will be able to completely replace the data analyst who looks interactively at different amounts of smoothing. However, insights gained from this process are also available from the "family approach" discussed in Marron (1995a). The main idea is to present a "one plot summary", which shows different smoothing levels, through overlaying smoothers with several bandwidths. Nearest neighbor methods are hopeless in this context as "continuity in the smoothing direction" is crucial to gaining visual insight.

While I do not believe that good location adaptive methods have much future in exploratory data analysis, I consider them worth more research, for the goal of display of final conclusions. An important open problem in this area is to find a device to show that the bandwidth function used is not hiding important features or highlighting spurious ones.

1.3 Is LO(W)ESS best?

LO(W)ESS includes a number of really excellent ideas, many of which were implemented well before many "experts in smoothing" even understood the issues. These ideas include local polynomials instead of local constants, really clever computational approaches, and adding robustness in smoothing. For these reasons, LO(W)ESS is recommended in the popular software package

S, and its variant Splus. This has resulted in this method being very broadly used for effective data analysis.

However, the fact that this is a useful method does not imply that it is the only useful method, or that it is the best method. As noted in my original contribution, there are many other effective approaches to data analysis by smoothing (M1 - M6). Furthermore, there are many different criteria by which smoothers can be judged (F1 - F12, including those of Thomas-Agnan). A commonly agreed upon answer to this question is impossible, because different people weight these criteria differently (all with good reason).

A more important issue is that I have (and hope others will gain) respect for:

1. the many clever ideas that went into LO(W)ESS, many of which were far ahead of their time.
2. the idea that many different viewpoints on how to choose a smoother are allowable, acceptable and rational.

1.4 Mathematics in statistics

Cleveland and Loader take issue with the use of mathematical methods in curve estimation. Their basic premise seems to be that "there are a lot of not very informative papers out there". I agree with this, but am not willing to accept a blanket condemnation of the usefulness of mathematical approaches to understanding smoothing. As noted in section 4 of my original contribution mathematics has played, and will continue to play an important role in methodological statistics. Moreover, the ability to bring to bear an array of methods (including real data examples, mathematics and simulation) is becoming increasingly important as we tackle harder and harder methodological problems.

1.5 Does bandwidth selection "work"?

In my original contribution it was made quite clear, from all of the important viewpoints of real data examples, asymptotic analysis, and simulations, that "modern" bandwidth selection methods do work quite well. As detailed in the specific comments section below, I take issue with Cleveland and Loader on this point, and several variants of it.

In fact, modern bandwidth selection is at a point that it is worth implementing as the default in software packages. This would be much more useful than an arbitrary choice based on experience with a few examples.

The issue of defaults for smoothing parameters is an important one. Cleveland and Loader convey the impression that it doesn't matter so much, because a proper analysis will involve looking at several bandwidths. While I agree with the latter in an abstract theoretical sense (e.g. I always try several myself, despite that fact that a modern bandwidth selection method will be effective), I find it contrary to much of the actual practice of smoothing. I have attended many public talks by, and have had much private discussion with, people who have used smoothing. A large proportion of these use LOWESS because of its strong recommendation in S. When I ask what smoothing parameter they use, a too common answer is "the default in S". While it is easy (and theoretically correct) to say "they shouldn't do that", I submit that software developers should take into account how users choose to use their product, as opposed to attempting to dictate how it should be used. From this point of view, the use of modern data based methods as defaults would make smoothing a much more useful tool for the very large number of non-expert user who do not choose to (or think they have no time to) try out several bandwidths.

An interesting approach to the problem of users attaching too much significance to defaults is implemented in the Software package R-code discussed in Cook and Weisberg (1994). In this package the bandwidth is controlled by a slide bar, which is manipulated by a mouse. This is not a new idea, but the clever contribution is that there is no initial default smooth presented. To get a smooth, the user must start moving the bandwidth slider. This is useful because then the user automatically scans through several bandwidths immediately, and sees how important it is to try several values of this parameter. With this user interface, problems with the user thinking "the default is good enough for me" are avoided.

2 Specific Comments

Here are some additional comments on the original papers and the resulting discussion

2.1 Cleveland and Loader

These authors are to be congratulated: some excellent contributions are made here.

My favorite among these is the fascinating historical discussion in Section 2. The level of scholarly work done in this section, with detailed attention to early references of which I (and surely many others) had been unaware, is truly outstanding. This is an important contribution that deserves to be frequently referenced.

Another important contribution is the interesting and convincing data analysis. This well written section makes it clear that the methods advocated by the authors can be effective tools for data analysis. It shows that in the hands of an expert, who has a lot of time to devote to the problem, methods that I personally find less intuitive and easy to work with, such as those based on nearest neighbors, and higher degree local polynomials, can give good results. However, I have not yet been convinced of how these tools will work in the hands of a user with less expertise, and less time to devote to the project.

Section 10.1 is also very useful. It is a simple and elegant mathematical way of seeing why local polynomials are better than local constants.

I wish to take issue with some of the other aspects of the paper.

Fix and Hodges (1951) (a preprint that was finally published as Fix and Hodges (1989)!), and Akaike (1954) predate the famous work of Rosenblatt and Parzen as the origins of the "modern view of density estimation" discussed at the beginning of Section 2.2.

The remarks in section 5 are generally correct, but other preferences for choice of kernel function are reasonable as well. My personal preference is for the Gaussian, because it gives very smooth pictures (smoother than piecewise polynomials with three derivatives, such as the tricube, as noted in Marron and Nolan (1989)). It also has some useful "continuity in smoothness properties" when the bandwidth changes. For example this seems to be the only kernel where the number of bumps in the smooth is a monotone function of the bandwidth, see Silverman (1981), and the kernel with the nicest "mode tree" which provides another way of understanding how smooth the Gaussian kernel is, see Minnotte and Scott (1993). Cleveland and Loader's objection to the Gaussian on speed grounds is not important when a fast "binned implementation" is used, for example as described in Fan and Marron (1994).

The examples in Section 8.2 were interesting and informative, but not convincing. The main reason is doubts about Mean Squared Error (and its many variants) as an error criterion. As noted in Marron and Tsybakov (1995), such error criteria too often "feel aspects of the problem" that are quite different from those that are relevant for data analysis. In my view it is more sensible to assess smoothers from the viewpoint of "how well do they convey the important lesson in the data?", than from the viewpoint of "how close to the true curve is the estimate in the sense of MSE?". From this viewpoint, the very interpretable global bandwidth smoother is more attractive than suggested here. Another important point to keep in mind is that this type of presentation is necessarily anecdotal in nature, with lessons depending on the example chosen. Examples are useful for making points, but in my view it would have been more useful to show examples where both were in turn better, thus illustrating the respective strengths and weaknesses of each method. Cleveland and Loader's examples together with those of Sheather, Wand, Smith and Kohn make a reasonable collection of this type.

The last paragraph of Section 8.2 doesn't say much, other than the obvious "don't take asymptotics too literally". It appears intended to suggest that asymptotics are useless, but is not convincing. It carries no more weight than an assertion that simple linear regression is useless on the basis of an example with silly results from extrapolation far outside the range of the data.

Lejeune (1984, 1985) probably deserves explicit mention at the beginning of section 10.2.

The issue of when bias approximations work and when they don't, discussed in the ninth paragraph of Section 10.2 is certainly an interesting and important one. However the suggestion that mathematical statisticians are unaware of this is incorrect. For example, see Marron and Wand (1992) and Marron (1995b) for theoretical work and insights on this issue.

The last paragraph of section 10.2 is not a fair representation of how competent mathematical statisticians work and think. Such workers do not draw "sweeping conclusions", they report what they find. Mathematical results are of the form "if A, then B". Accusations of this type do not apply as long as A is made clear, and discussed in a relevant statistical light. I agree that the term "optimal" gets overused, and is often used for things which don't turn out as well as one might hope. However, when the term is used in a paper about asymptotics, it is a short form of "asymptotically optimal". In this form, the objections of Cleveland and Loader do not apply.

In my view, the most dissappointing part of the paper comes near the beginning of section 10.3, with the statement "much effort has been expended, ... with little regard for the possibility that the optimal bandwidth may be very inadequate". As noted in Section 3 of my original contribution, there has been quite a lot of work (much of it already published), on assessing how bandwidth selectors actually work (as opposed to merely their asymptotic properties) in terms of both simulation and real data examples.

The criticisms of plug-in bandwidth selection methods in the third and fourth paragraphs of Section 10.3 are well known, and have had an important effect on how the field has developed. It has been interesting to think about why such methods work as well as they do when they are actually used, in view of this type of consideration. Some personal thoughts on precisely this "paradox" are given in the appendix of my original contribution.

Cleveland and Loader's attitude on bandwidth selection is mystifying. Through most of the paper, they are pragmatic, with effective data analysis as their primary goal. However they diverge dramatically from this principle at this point in their discussion, and "sweeping conclusions are drawn solely from asymptotics, ... without any real justification". Also the asymptotics are not quite right, in that selection of a global bandwidth depends on things more like the integral $\int (f'')^2$ than like the second derivative $f''(x)$ itself. The asymptotics are different for estimation of the integral (e.g. estimators with rate of convergence $n^{-1/2}$ are available under strong enough assumptions for the former, but not the latter). But this is a minor issue. The important point is that there is a convincing case for the effectiveness of modern bandwidth selectors . See Section 3 of my original contribution for references.

I agree with Cleveland and Loader's position in Section 10.4, and have also not been convinced by the suggestions that odd order local polynomials are "better". However, I do not agree with the suggestion that this is an example where asymptotics give the "wrong answer".

2.2 Gasser and Seifert

This is a good exposition of how local polynomial methods work, including important numerical issues. The analysis of data sparsity is deep and interesting. The probability of having three points in some neighborhood, displayed in Figure 12, turns out to be important in the asymptotic analysis of ridge regression done by Hall and Marron (1995).

2.3 Hall and Turlach

The alternative approach to data sparsity looks interesting. I agree with the addition of the factor F7, for the choice of the smoothing parameter, although my F1 - F6 were intended for the choice of smoother, e.g. choice among M1 - M6.

2.4 Jones

I like the point about many good methods being rather similar in many cases. Much of the controversy has been about only some of the cases that one encounters. E.g. most smoothers work quite well (and similarly) in the important and simple case of a fixed equally spaced design.

2.5 Kohn, Sheather, Smith and Wand

This discussion provides some thought provoking examples. I also liked the additional reasons given as to why modern bandwidth selection methods should be considered for implementation in software packages.

The point of Figure 4 is a thought provoking one.

2.6 Thomas-Agnan

I agree with the addition of F8 - F12 to the list of factors that go into the choice of a smoother. I am not so sure about adding M7 and M8 to the list of methods, on the grounds that I view M7 as essentially the same as M4, and M8 as a special case of M1. However, this is personal, and there are certainly other ways of classifying estimators.

Thomas-Agnan raises an interesting point about effective kernels and my factor F2. I agree that one can make a plot of the *large number of different* effective kernels resulting from say a nearest neighbor estimator (or many others as noted). But for what I call "less interpretable methods", there will be different effective kernels for each point where the estimate is constructed. Because it is hard to keep all these different kernel functions in my mind, (note the contrast to the *single* effective kernel for a fixed bandwidth method), it is hard to understand what the estimator tells about the data. However, this whole discussion underlines one of my main points: there are many different viewpoints on smoothing. Factors that are important to some (e.g. F2 is quite important to me), are less important to others.

The points on comparability are well taken. Comparisons with analogs in numerical analysis need to be interpreted with caution for two reasons. First they tend to deny the presence of variance, which is crucial in statistical smoothing. Second a "good smooth" is much harder to define in statistics. I am not prepared to accept "the smooth should be as close as possible to the true regression curve" as a criterion, because it is not always the same as the more important "the curve should tell us as much as possible about the data".

I am not sure that the importance of F1 will recede in time, because no software can make everything easily available. Choices have to be made about what is easy, and what takes more effort to use (e.g. menus can only be so long, and things that are on submenus are anticipated to get less use).

The "entropy" analogy is thought provoking.

3 Acknowledgement

The author is grateful to many people for input on these issues. N. I. Fisher, R. Kohn and M. P. Wand have made especially many helpful comments, and spent substantial time discussing these issues.

References

[1] Akaike, H. (1954) An approximation to the density function, *Annals of the Institute of Statistical Mathematics*, 6, 127-132.

[2] Cook, R. D. and Weisberg, S. (1994) *Regression Graphics*, Wiley, New York.

[3] Fix, E. and Hodges, J. L. (1951) Discriminatory analysis - nonparametric discrimination: consistency properties. *Report No. 4, Project No. 21-29-004*, USAF School of Aviation Medicine, Randolph Field, Texas.

[4] Fix, E. and Hodges, J. L. (1989) Discriminatory analysis - nonparametric discrimination: consistency properties, *International Statistical Review*, 57, 238-247.

[5] Fan. J. and Marron, J. S. (1994) Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics*, **3**, 35-56.

[6] Hall, P. and Marron, J. S. (1995) On the role of the ridge parameter in local linear smoothing, unpublished manuscript.

[7] Lejeune, M. (1984) Optimization in nonparametric regression, *Compstat 1984 (Proceedings in Computational Statistics)*, eds. T Havranek, Z. Sidak, and M. Novak, Physica Verlag, Vienna, 421-426.

[8] Lejeune, M. (1984) Estimation non-paramétrique par noyaux: régression polynomiale mobile, Revue de Statistiques Appliquées, 33, 43-67.

[9] Marron, J. S. (1995a) Presentation of smoothers: the family approach, unpublished manuscript.

[10] Marron, J. S. (1995b) Visual understanding of higher order kernels, to appear in *Journal of Computational and Graphical Statistics.*

[11] Marron, J. S. and Nolan, D. (1989) Canonical kernels for density estimation, *Statistics and Probability Letters*, 7, 195-199.

[12] Marron, J. S. and Udina, F. (1995) Interactive local bandwidth choice, unpublished manuscript.

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

[14] Minnotte, M.C. and Scott, D. W. (1993). The Mode Tree: A Tool for Visualization of Nonparametric Density Features, *Computational and Graphical Statistics*, 2, 51-68.

[15] Silverman, B. W. (1981) Using kernel estimates to investigate modality, *Journal of the Royal Statistical Society, Series B*, 43, 97-99.