Rejoinder

William S. Cleveland and Clive Loader

AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974, USA.

1 Sound Premises

Theoretical work in any area of statistics can have a substantial impact on the statistical methods that we use to analyze data in that area. But to do so, the premises on which the theory is based must be sound. The premises must sensibly model the sources of variation in the data. And the premises must address the methodology as it is used in practice, and set criteria that are of genuine importance for that usage. Having set the premises, the investigator must then derive results. This requires command of the necessary technical tools.

The theoretical work by actuaries in the early smoothing literature from the 1880s to the 1920s led to important advances such as local polynomial fitting. The strength of the work derived, in part, from elegant and insightful application of the technical tools, which were based on the algebra of finite difference operators. But far more important was an incisive setting of premises which came from a deep understanding of the behavior of the data under study, mortality and sickness data. These theoreticians were practitioners as well. They did far more than simply plug in data from some remote source to test an already developed tool. They started with the data. They shaped the premises of their theory from the data. *They were as responsible for the subject matter conclusions that emanated from their tools as they were for the scientific validity of the tools.* The consequence of this grounding in the data was the construction of the major pieces of intuitive insight that still guide smoothing today:

1. The trade-off between bias and variance.
2. The need for smooth weight functions to produce smooth fits.

3. Local polynomial fitting.
4. The poor performance of local constant fits compared with higher order fits.
5. Optimal weight functions.
6. Penalty functions and smoothing splines.
7. Smoothing in likelihood models.

The research community in smoothing today, or any other statistics research community, cannot expect to develop sound premises for theoretical work without a similar grounding in data.

2 Very Small Bandwidths

A major premise of the Seifert and Gasser paper is that we can learn about smoothers with bandwidths that are exceedingly small. In some cases the bandwidths get so small that the definitions of the smoothers are indeterminate without further rules about what to do with when there are no data or just a single point in a neighborhood. For example, in Figure 8, they investigate properties of smoothers by artificially generated data with a sample size of $n = 50$. The explanatory variable is uniform over [0, 1]. Their smoothing neighborhoods range from 0.04 to 0.4 so that the *expected* number of points in these neighborhoods ranges from 2 to 20. They point out that for local linear fitting with a neighborhood size of 0.04, the MISE "increases steeply to values as large as 629". With a neighborhood this small the results depend heavily on what one decides to do when 0 or 1 point appears inside the neighborhood.

These neighborhoods are far too small to give us an understanding of the relative performance of smoothers in practice. Except for studies where the data have little little or no noise and interpolation is the goal rather than smoothing, the fits with such small bandwidths would be far too noisy. A single attempt to use such small bandwidths on a real set of noisy data where we were responsible for the conclusions would make this obvious. Jones puts it well: "Some days I think the deep study (apparently, studies) of Seifert & Gasser of problems due to sparse data is very valuable, some days less so. Perhaps in practice it is something like local bandwidths and a degree of common sense that is called for."

A theoretical treatise is not wrong when its premises are at odds with the data and with the sensible practice of the methodology. It is simply uninformative.

3 Pseudodata

We find the Hall and Turlach suggestion of pseudodata interesting, although we reserve final judgment until we see their full account. There is a precedent for this idea, although in a much more limited form. For smoothers targeting the case of one explanatory variable, some, such as Tukey (1977), have faced the boundary problem by predicting the data beyond the boundaries and then smoothing the data and the predictions.

4 The Golden Standard: Past and Present

In a number of cases, there has been mention of local polynomial fitting as an emerging new discovery that holds great promise for the future, a coming golden standard. But what is emerging is awareness, not discovery. Local polynomial fitting, as we have emphasized in earlier papers, in our paper in this collection, and just now in this rejoinder, began in the actuarial research community and has been under development for over a century (e.g., Woolhouse 1880, Spencer 1904a, Henderson 1916, Macaulay 1931, Stone 1977, Cleveland 1979, Hastie and Loader 1993, Fan 1993.)

5 Plug-In Estimates

It makes little sense to use local quadratic or cubic estimates solely to fine-tune less efficient local constant and local linear methods. This statement is one consequence of our Section 10.3 and of the simulations of Sheather, Wand, Smith and Kohn in their Section 2.3. If we need local quadratic or local cubic fitting to adequately describe the characteristics of a regression curve or surface, then we should use the local quadratic or cubic smoother to fit the data. (We find ourselves somewhat surprised that it has fallen to us to make a statement that is nearly a tautology.)

Plug-in methods should be considered an idea that failed, and allowed to die a natural death. For those attempting life support, consider the following. The pilot estimates used to tune plug-in methods can beat the plug-in methods. Table 1 of SWSK confirms our point nicely; in seven out of eight examples local quadratic has won, despite the tuning of the amount of smoothing to the local linear estimate. (N. B. We do not suggest doing this in practice; we have simply pointed this out to increase our understanding of the issues.)

6 Local Constant Fitting

Our skepticism about local constant fitting (over all x) is simply a matter of not having found cases where it convincingly models the data more satisfactorily than higher-degree fitting. We can imagine cases where local constant fitting might do better, for example, when the underlying pattern is constant at all boundaries and locally linear elsewhere. But we believe that in practice, cases of better modeling by degree zero at all x are at best rare. We hasten to emphasize that this was the conclusion of the early actuarial research community discussed above; in their work, cubic fitting became a standard.

Of course, if we use adaptive methods, then it is possible that at some special values of x where the surface is flat, we might well chose degree zero locally. But we do not expect to choose it at all x because that would imply an uninteresting selection of the explanatory variables.

Marron seems intent on holding on to local constant fitting. He writes: "Fixed bandwidth local constant kernel methods put the interested analyst in closest possible intuitive contact with the data, because they are simple, understandable, local averages. Note that I am not advocating this estimator as the solution to all problems ... but instead am merely pointing out it cannot be dismissed out of hand."

To the contrary, we do have every right to dismiss it out of hand until someone, perhaps Marron himself, provides data sets where local-constant smoothing with fixed bandwidths at all x does a better job of modeling the data than higher order fits. We have demonstrated the reverse, in our paper in this collection, and elsewhere.

And we disagree with the claim of greater intuition for local constant fits. It is no more acute than for higher order fits. In practice, our intuition about how well a smoother models the data is most acute when it is based on diagnostic and on our understanding of the broad performance characteristics of the smoother such as the class of functions it reproduces, the frequencies that is passes, and the frequencies that it suppresses.

Thomas-Agnan also disagrees with Marron's statement about intuition, pointing out that it overrates closed-form formulas and that the popular linear smoothers end up as linear combinations of the y_i. Clearly we agree with the discussant.

7 Residual Plots

Sheather, Wand, Smith, and Kohn have raised questions about the validity of smoothing residual plots as an aid to judging the performance of a smoother. They also implicate Cleveland (1993), but they should also implicate Tukey (1977), who introduced the systematic smoothing of residuals.

Unfortunately, the discussants have missed an important property of smoothers, and they have misinterpreted the process that is used to judge a fit from a smooth of the residuals.

They state: "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." But this is not so because smoother operators, unlike least-squares operators, are not idempotent. In fact, the use of the same smoother on the residuals that was used to produce the fit is given the name *twicing* by Tukey (1977).

The process is not to judge a fit to be adequate if a smooth curve on its residual plot is flat. A flat curve means simply that no systematic, reproducible lack of fit has been detected. The fit may well be too noisy. As Cleveland (1994) points out: "This method of graphing and smoothing residuals is a one-sided test; it can show us when [the smoothing parameter] is too large but sets off no alarm when [the smoothing parameter] is too small. One way to keep [the smoothing parameter] from being too small is to increase it to the point where the residual graph just begins to show a pattern and then use a slightly smaller value ... .

In Figure 1 of the discussion of Sheather, Wand, Smith and Kohn, one sees lack of fit only marginally at the largest bandwidth, and so one would usually select a parameter close to 0.675. A very similar pattern is shown in the residual plots of our paper — especially Figure 8 — and we chose the larger bandwidths.

One always looks at residual plots in conjunction with looking at plots of the fit. When a prominent feature in a residual plot corresponds to a feature in the fit that has a rapidly changing derivative, lack of fit is typically the cause. For example, in Figure 11 of our paper, the large residuals are seen to line up with the breaks in the fitted curve, indicating lack of fit rather than a heavy-tailed residual distribution.

8 What is loess?

We do not understand Marron's M-classification of smoothers. Lowess and loess are local polynomial methods and therefore belong to category M1.

9 Fixed vs. Nearest-Neighbor Bandwidth Selection

We have argued for the value of nearest-neighbor bandwidth selection as a reasonable default method.

In providing smoothing tools to data analysts it makes sense to have available, perhaps along with a reliable adaptive method, a simple bandwidth selection procedure that is based only on the x_i and that has one easy-to-understand parameter. Both fixed and nearest-neighbor selection would provide this. But of the two, we have found that nearest-neighbor typically does as well or better than fixed. Fixed selection can result in dramatic swings in variance; the greatest drama occurs when there are no data in the interior of a fixed neighborhood. Deciding how to protect users from a fit that is not well defined at some places where an evaluation is requested is a thorny problem, one that does not need solving for nearest-neighbor selection.

But this defense of nearest-neighbor selection is not meant to imply that fixed-bandwidth selection can never perform better than nearest-neighbor. In their Section 2.2, Sheather, Wand, Smith and Kohn provide one example, an artificially generated one, where fixed does better. Seifert and Gasser do the same in their Section 2.2, again with artificially data. Both examples have the same phenomenon that makes fixed selection do better. The curvature is greatest where the data are the sparsest. But this should not be construed as a statement that fixed selection in some sense reacts to changes in curvature in any way. Neither fixed nor nearest-neighbor do so. Note that in the example of Sheather, Wand, Smith and Kohn, if we alter the example and take X_i to be $U_i^{0.7}$ instead of $1 - U_i^{0.7}$, then nearest neighbor will perform better than fixed. But the curvature is the same in both cases. To react to changing curvature we need adaptive methods such as those we discuss in Section 9 of our paper.

There are claims that fixed selection is optimal and that nearest-neighbor is not. These claims result from an asymptotic approximation to the nearest neighbor bandwidth; In Section 8.2 of our paper we showed this approximation does not work. A claim of optimality is a strong assertion because a single a single counterexample suffices to disprove it. In the papers in this collecton, and in the discussions, there are examples where fixed is best, examples where nearest neighbor is best, and examples where both are inadequate.

In summary, we use nearest neighbor methods as a default in software for smoothing because (1) there is direct control over the number of points being smoothed, thus avoiding problems that arise from windows with a very small number of observations; and (2) it usually results in larger bandwidths at boundary regions, which is often desirable due to the one-sided nature of smoothing at boundaries oompounded by the tendency for data to be sparse at the boundaries.

10 Comparing Smoothers

Seifert and Gasser compare the performance of smoothers at the same values of the bandwidth. This is inappropriate because the degrees of freedom of two different smoothers can be radically different for the same bandwidth values. In fact, by a rescaling of the kernel of a smoother, we can effect a large change in its performance relative to others if we match bandwidth. For example, the performance of the smoothers in Figure 8 of Seifert and Gasser is almost wholly due to differing degrees of freedom. They scale the local linear fitting with Gauss weights to "reach the same asymptotically optimal bandwidth as Epanechnikov weights." Because the comparison is carried out as a function of h, the behavior of the this estimate could be made to look quite different by scaling by some other method. Clearly, in a study such as this we need to compare smoothers by matching degrees of freedom.

But do we fail to practice what we preach? Jones asks: "wouldn't it be better in comparative figures such as Figures 2, 3, 7 and 8 to have in each column smooths with equal equivalent degrees-of-freedom rather than equal, but not comparable, bandwidths?" The answer is "no", since our aim was not to make a blanket statement as to which smoother was best for the problem, but rather model selection, for which we want to consider a range of models with differing amouts of smoothing. If we fit all degrees with equal degrees of freedom, one would often be choosing between all undersmoothed or all oversmoothed fits, and the mixing would not work.

11 Software

We agree with the view expressed by Sheather, Wand, Smith and Kohn and some other papers and discussants — most notably Marron — that software should incorporate automated bandwidth selectors. But we strongly disagree with the views as to what, or how, to implement. No existing bandwidth selectors come close to being suitable for use as a default in general purpose software such as loess, either in terms of general applicability or reliability of performance. An adaptive method such as that used in Figure 9 is appropriate when there is plenty of data with low noise; it would be quite inappropriate for smoothing residual plots.

As with any sensible model choice criterion, any bandwidth selector will fit several models, and attempt to decide which will be best. As such, a bandwidth selector is not part of a basic smoothing algorithm, but something that can be built on top. The bandwidth selection should not be confounded as part of one's basic smoothing algorithm as Sheather, Wand, Smith and Kohn suggest. Rather, the basic smoothing algorithm should return sufficient quantities of diagnostic information to assess the fit. The user can then decide whether to automatically minimize a criterion, or to actually look at the fit using diagnostics as a guide.

Consider for example the C_p type statistics used by Cleveland and Devlin (1988) as part of loess. The use of C_p is not forced upon users. Rather, the loess method returns sufficient diagnostic information about each fit to enable the C_p statistics to be readily computed. Locfit software (Loader, 1995) is a recent development built on the loess model, and provides a wealth of additional diagnostic information that enables many different bandwidth selectors with short S scripts. The basic point here is a software design issue: different computations, such as bandwidth selection and the underlying local fitting, need to be kept separate.

12 Testing

Sheather, Wand, Smith and Kohn as well as Jones raise the issue of F-tests and t-intervals being contaminated by bias. But this issue arises in parametric fitting where these methods of inference are used routinely. In both parametric fitting and smoothing there must be at least one fit for which we have a reasonable assurance that no bias is present. For t-intervals we need this fit to estimate the variation in the errors. Bias in t-intervals has similarities to other bias problems and is discussed in Sun and Loader (1994) and Loader (1993). For an F-test we need such a fit to provide an alternative model. The null model will have larger bandwidth; if the F-test reveals substantial lack of fit, then clearly the null model is biased and inadequate.

13 Computation

Sheather, Wand, Smith and Kohn criticize our adaptive procedure on the grounds it is "computationally difficult to locally estimate such bandwidths at each point in the design." Their statement is correct, but this is not what we do. Rather, the bandwidth is computed at an adaptive sequence of knots, with the greatest knot density in regions of lowest bandwidth. For example, in the wavelet example of Figure 11 of our paper, the bandwidth is computed at 130 points, far less than the 2048 design points. Moreover, most of these knots are in regions where small bandwidths are used, and the evaluation is relatively cheap. Jones also seems to misinterpret our procedure; the local $C(h)$ statistics are computed over a wide range of bandwidths.

New References

Cleveland, W. S. (1994) *The Elements of Graphing Data.* Hobart Press, books@hobart.com.

Loader, C. R. (1993). Nonparametric regression, confidence bands and bias correction. *Proc. 25th Symposium on the Interface*, 131-136.

Loader, C. R. (1995). LOCFIT: A program for local fitting. `http://netlib.att.com/netlib/att/stat/prog/lfhome/home.html`

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, B*, **47**, 1-52.

Stoker, T. M. (1994). Smoothing bias in density derivative estimation. *Journal of the American Statistical Association* **88**, 855-871.

Sun, J. and Loader, C. R. (1994). Simultaneous confidence bands in linear regression and smoothing. *Ann. Statist.* **22**, 1328-1345.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Massachusetts.