intercept). One can therefore remedy the mismatch of these lines by simply correcting for the variance inflation. However, this discussion is very closely tied to this particular situation: Variance correction is by no means a panacea, and its effects away from the normal design are (a) less considerable and (b) not necessarily beneficial (Jones, 1991) in other cases (such as the remainder of C&M's Figure 11).

5. CONCLUSIONS

It is not so long ago that the version of the "folklore" that I was contented with (without much thought!) was that one used G-M for fixed designs and N-W in the random case (e.g., Cheng, 1990). This now seems somewhat dubious.

I have a particular liking for (1) in the fixed uniform design context. So far as N-W and G-M go, however, I am happy that one could afford to use either of these instead in this case without really changing anything. A verdict on the fixed but nonuniform design case is given in Jones and Davies (1991). But none of the existing versions of kernel regression are the last word in the random design case. There, both N-W and G-M/P-C have disadvantages, as C&M make clear, yet it does not appear to be impossible to get the best of both internal and external estimation worlds with new—but not overly sophisticated—methods; it is also sensible to apply such estimators back to the fixed design case. Hopefully, the authors might agree that thinking in such a framework helps to clarify the issues involved and illuminate a way forward.

I am very pleased to have been afforded the opportunity to append some comments on this most interesting paper.

ACKNOWLEDGMENTS

I am very grateful to Steve Marron for providing me with various drafts of the current paper and to him and C.-K. Chu for providing other unpublished papers. I also greatly appreciate Steve Davies' considerable contribution to the computational backup to these comments.

Comment: Should We Use Kernel Methods at All?

B. W. Silverman

I would like first of all to thank the authors for a most interesting, thoughtful and provocative paper. I think it is important to broaden out the discussion to consider other possible estimators in more detail. The authors' attempt to be even-handed is particularly to be welcomed, and if my own contribution does not immediately appear to be in the same vein it is only because the authors have already themselves dealt with the two kernel estimators.

1. SOME PHILOSOPHICAL REMARKS

The authors have set out an interesting dichotomy between two different viewpoints, P1 and P2, that might be adopted. I wonder, though, whether a synthesis of these approaches gives the real clue to what smoothing methods might ideally be aiming at. Certainly my own view would be more like a philosophy *P4*: *We are looking for structure in this set of numbers, without imposing rigid parametric assumptions, but still within a statistical framework of some sort.*

The statement P1 is very much along the lines of the "exploratory data analysis" approach of Tukey (1977). This was a very welcome reaction to the overemphasis on uncritical model fitting as exemplified by P2, and in order to clear the air it needed to turn its back on several decades of statistical thinking. For example, Tukey's original book—always intended as an introductory text—nowhere even mentioned the idea of calculating the average of the data set. But, of course, the classical statistics that had become so constraining had itself originally developed in order to answer questions raised by data analytic approaches. Thus, in setting out a dichotomy of the P1/P2 kind, we can either give ourselves two different extremes between which to oscillate or else two different ingre-

B. W. Silverman is Professor of Statistics, School of Mathematical Sciences, University of Bath, Bath BA2 7AY, England.

dients that have both been found to be valuable and that might, if combined properly, reinforce one another. I think that the authors agree with me in taking the latter view.

I would disagree strongly with the assertion that nonparametric regression estimators fall into the P2 philosophy. They ought to be attractive in practice because they can provide what one might call a *model-based exploratory approach* that therefore has elements of both P1 and P2. So I very much welcome the authors' reminder that both P1 and P2 should be borne in mind, and would add that a combination of the two is both desirable and possible.

2. KERNEL REGRESSION ESTIMATES

The authors have made it very clear that they are trying to be even-handed in setting out the competing advantages of the two different kernel-based approaches. However, they point out some properties of the convolution-weighted estimator that seem to be very disturbing, notably the unpleasant behavior demonstrated in their Figure 5. One of the important properties of almost every statistical technique is reproduceability, so that the same analysis carried out on the same data gives the same results. Reproduceability at least assures the user that if the analysis of data from two experiments produces different results, the experiments haven't produced exactly the same data. (Bootstrappers had better stick to repeatable random number generators!) At an elementary level, this a strong argument in favor of techniques like linear regression over fitting lines "by eye."

The authors' Figure 5 and its accompanying discussion highlight the desirability of a property that somewhat extends the idea of reproduceability, that of continuity in all the data inputs. This property would ensure that if the y or the x component of any data point was perturbed a small amount, the resulting curve would only be altered by a small amount.

To see why this is related to reproduceability, and to Figure 5, suppose that three statisticians were given a scatter plot of the data points plotted in Figure 5. Statistician A makes careful measurements and records the points in exactly the configuration shown (i.e., in the sequence low-high-high-low in the region near 0.5). Statistician B rounds the data a little and records the x-coordinates near 0.5 as two coincident pairs, at each of which there is one high and one low data point. Finally, Statistician C, who only has a dirty and small scale picture to work on, moves the points around a little into the order high-low-low-high. Continuity would imply that all three statisticians got almost exactly the same result; this example demonstrates that in order to be truly reproducible in practice, a technique should as far as possible be continuous in all its inputs. Figure 5 quite clearly shows that convolution weighted kernel regression doesn't have this property.

Let us now turn to the authors' Figures 3 and 9, because these highlight another property, that of being able to detect simple structure when it is present. If one were fitting a parametric regression model to data with abscissae as shown in these figures, there might be some argument as to whether the values should be equally weighted. But I find it hard to believe that a scientist would be satisfied by a method that failed to draw an exact straight line though these points! Scientific models are often based on linear relationships—perhaps achieved by suitable transformations—so if an experiment is so good that it produces a linear relationship without noticeable error the least one could ask of a fitting method is that this should be picked up.

One way of saving the situation would be to fit an ordinary linear regression, to apply kernel regression to the residuals and to add the resulting curve to the linear regression fit. This would seem a very natural precaution to take to ensure that linear or near-linear structure was always detected. But Figures 5 and 9 taken together cast doubt on both forms of kernel regression and demonstrate a need to consider other approaches.

3. SPLINE SMOOTHERS AND THE CONCERNS DISCUSSED IN THE PAPER

The authors of course mention spline smoothing and refer to Eubank (1988) and Wahba (1990). For a more succinct survey, see Silverman (1985). The terminology "spline smoothing" is unfortunate because it is in a sense fortuitous that spline smoothers are splines, and confusion with certain other aspects of splines may have led to the relative unpopularity of spline smoothing. A better terminology in the statistical setting would have been *penalized least-squares smoothing* since the spline smoother in its usual form is just the minimizer of

$$S(m) = \sum_i \{Y_i - m(x_i)\}^2 + \alpha \int_a^b m''(x)^2 \, dx,$$

where $[a, b]$ is any interval containing all the x_i. The minimizing curve $\hat{m}_S$ is easily calculated and, just as is the case for the kernel estimators described by the authors, $\hat{m}_S(x)$ is for each x a weighted linear combination of the Y_i.

The spline smoother will behave perfectly in both the respects discussed above. It can be shown that perturbing the x_i (even if this results in values coinciding or crossing over) will affect $\hat{m}_S$ in a continuous fashion, so in this respect $\hat{m}_S$ shares with $\hat{m}_E$ the property of dealing properly with the situation of Figure 5. On the other hand (no matter what value of the smoothing parameter α is used) $\hat{m}_S$ will yield an exact straight line when applied to the data of Figures 3 and 9. Spline smoothers have other advantages too, and we shall briefly consider these in the next section.

4. OTHER PROPERTIES OF SPLINE SMOOTHERS

4.1 Adaptivity to Uneven Data

An important matter to consider when dealing with unequally spaced data is the question of the relative amount of smoothing applied in different parts of the same sample. For the kernel regression estimators, this question is addressed by the plots given by the authors in their Figure 10. Suppose we have a relatively large sample, and that the points x_i are not evenly spaced. Suppose that x_{thin} is a point in a region where there are not many points x_i while x_{thick} is in a region with a large number of x_i. How many data points Y_i would we wish to have a substantial influence on $\hat{m}(x_{\text{thin}})$ and $\hat{m}(x_{\text{thick}})$, respectively?

For both the estimators discussed in the paper, the estimator at any given point x will be mainly based (for a normal kernel) on the data falling within $2h$ of x. Using a broad definition of the design density f, the number of influential data points will be approximately proportional to $f(x)$. Thus, $\hat{m}(x_{\text{thin}})$ will be based on far fewer data points than $\hat{m}(x_{\text{thick}})$. There are approaches, not discussed in detail by the authors or in the present comment, that always base the estimate on about the same number of data points, but these can be substantially biased.

It is shown in Silverman (1984) that the spline smoother steers an attractive middle course between these two extremes and automatically uses weights that allow substantial influence from a number of data points approximately proportional to $f(x)^{1/4}$. Working from an equation essentially the same as the authors' (5.1), Silverman (1984) showed that this kind of adaptation to different local densities of data is for practical purposes ideal. I do not think it is helpful to place too much emphasis on the very detailed asymptotic behavior of the various estimators, but the automatic adaptivity of the spline smoother—combined with its linear dependence on the data—certainly seems worth bearing in mind.

4.2 Applicability to Generalized Linear Model Dependence

One of the major developments in statistics in recent years has been the systematic approach to generalized linear models pioneered by McCullagh and Nelder (1983). This extends enormously the whole idea, and applicability, of linear regression. Because the spline smoothing method works by explicitly trading off a measure of fidelity to the data, the residual sum of squares, against a measure of roughness, $\int m''^2$, it can immediately be extended in a corresponding way to problems where there is more general dependence of the observation Y_i on $m(x_i)$ than $\mathrm{E}Y_i = m(x_i)$. If each Y_i is assumed to depend on its predictor $m(x_i)$ through a model with log likelihood $L(y; \theta)$, then a very natural approach is to estimate m by maximizing the penalized log likelihood

$$S(m) = \sum_i L\{Y_i; m(x_i)\} - \alpha \int_a^b m''(x)^2 \, dx.$$

A typical example is logistic regression. Suppose at time (or dose level) x_i we observe Y_i 'successes' out of n_i trials. Then a natural way of constructing the penalized log likelihood is to write

$$L\{Y_i; m(x_i)\} = Y_i \log \pi_i + (n_i - Y_i) \log(1 - \pi_i),$$

where

$$\pi_i = \frac{\exp\{m(x_i)\}}{1 + \exp\{m(x_i)\}}.$$

An early (and in many ways defective) paper along these lines is Silverman (1978). A much more satisfactory reference is Green (1987), who investigates many operational matters and pursues detailed connections with generalized linear models. It is hard to see how one would even start to tackle this sort of problem in any natural and general way using kernel regression methods.

4.3 Unusual Problems

The penalized likelihood idea can be extended to deal with problems that are very nonstandard. For example, Silverman and Wood (1987) consider an experiment in which the treatment groups remain identical up to various known points in time and then diverge. A natural way to investigate and present the data is to draw a branching or bifurcating curve, where each branch represents a particular treatment and the points at which the curve bifurcates are precisely those at which the treatments begin to differ. Again, it would be hard to estimate such a bifurcating curve using kernel methods, but a way of thinking based on penalized least squares and spline smoothing gives a natural way to proceed.

In conclusion, it would be facile to suggest that any particular method will yield answers to all the problems one might encounter. But I do believe that if we are making detailed comparisons between different approaches, the spline smoothing method as a general approach has a great deal to commend it.

Rejoinder

C.-K. Chu and J. S. Marron

We are very grateful to the discussants for their interesting and thoughtful additions to the points we made in this paper. We also thank the editor for many helpful comments, and for the nontrivial task of organizing the discussion.

Our response is organized into sections, with the first three concerning topics raised by more than one discussant, followed by some individual responses in alphabetical order.

1. PHILOSOPHICAL ISSUES

Useful elaboration of our P1-P2 formulation of the viewpoints that have been adopted to consider smoothing, has been provided by Grund and Härdle and by Silverman.

We agree with Grund and Härdle that computational issues are very important and welcome the addition of their P3 as a general principle. However, in the present context, we do not view proper incorporation of this factor as having a major impact on the ideas indicated here. The reason is that both $\hat{m}_C$ and $\hat{m}_E$, when properly implemented, for example, as described in Section 3 of Grund and Härdle, have roughly comparable computation time. On the other hand, this P3 could easily become vital in, for example, a comparison of splines versus kernels as suggested by Silverman.

We also find Silverman's P4 and the surrounding discussion very useful. This is a very nice extension of the points we were trying to make. One small point we would like to clear up is that when we attached the phrase "nonparametric regression estimation" to P2, we were referring to the phrase, not the methodology. Our intention was to convey the point that most people who have used this phrase in the literature tend to lie in the P2 camp. However, we wholeheartedly endorse Silverman's main point (also expressed well by Grund and Härdle) that there needs to be more combined use of P1 and P2.

2. ADDITIONAL COMPARISON

Some interesting and carefully considered alternative ways of comparing $\hat{m}_C$ and $\hat{m}_E$ are presented by Grund and Härdle and by Hart. We welcome these deeper analyses and are happy that the main conclusions are not much different from what we saw by simpler methods.

The figures of Grund and Härdle are very informative and provide excellent visual quantification of the points we were driving at in the paper. However, we caution against trying to infer too much from these examples. We are hesitant to make a recommendation as to which estimator is better on the basis of the size of the region where the ratio is bigger than one, because this is only one example. Even if one looked at several such examples, there are doubtless other examples that give the opposite conclusion. Furthermore, even in the presented example caution is indicated, because these sizes of regions are also dependent on the parameterization that has been chosen for the example. For example, the regions seen in Figure 4 could look quite different if this picture were based on the logarithms of these two parameters. But, of course, the main point here is that $\hat{m}_C$ and $\hat{m}_E$ are not really comparable in terms of one being always preferable, and the figure illustrates this in a compelling fashion.

Hart's idea of looking at the joint probability structure of $\hat{m}_C$ and $\hat{m}_E$ is excellent. He admirably illustrates that this is an important issue, and things are not as one might expect at first guess. This clearly needs to be borne in mind in future comparisons of estimators.

3. OTHER KERNEL ESTIMATORS

Hart and Jones discuss alternative kernel smoothers to $\hat{m}_C$ and $\hat{m}_E$ and make some interesting cases for their serious consideration. We were