and Marron's (A.1)–(A.5),

$$(3)\quad \operatorname{Var}\big(\hat{m}_T(x)\big) = \frac{\sigma^2}{nh}\int K^2 + o\big((nh)^{-1}\big).$$

This implies that $\hat{m}_T$ with bandwidth h has the same asymptotic variance as $\hat{m}_E$ with the bandwidth $h_x = h/f(x)$. In particular, the limiting variances of $\hat{m}_T$ and $\hat{m}_E$ are the same in a case highlighted by Chu and Marron, that is, when $X_1, \ldots, X_n$ are a random sample from a $U(0,1)$ distribution.

The bias of $\hat{m}_T(x)$ has the representation (again under assumptions akin to (A.1)–(A.5))

$$(4)\quad \begin{aligned}&\operatorname{Bias}\big(\hat{m}_T(x)\big)\\ &\quad = \frac{h^2}{2}(mQ)''\big(F(x)\big)\int u^2K + o(h^2)\\ &\quad = \frac{h^2}{2}\left\{\frac{m''(x)f(x) - m'(x)f'(x)}{f^3(x)}\right\}\int u^2K\\ &\qquad + o(h^2).\end{aligned}$$

In general, $\operatorname{Bias}(\hat{m}_T)$ is different from both $\operatorname{Bias}(\hat{m}_E)$ and $\operatorname{Bias}(\hat{m}_C)$; this is true even if one allows the bandwidths of $\hat{m}_E$ and $\hat{m}_C$ to vary with x a la $h_x = h/(f(x))^a$. By considering (3) and (4) above, and Sections 3 and 4 of Chu and Marron, one finds, not surprisingly, that $\operatorname{MSE}(\hat{m}_T)$ is not comparable with either $\operatorname{MSE}(\hat{m}_C)$ or $\operatorname{MSE}(\hat{m}_E)$. It is worth noting, though, that when $X_1, \ldots, X_n$ are iid $U(0,1)$, the asymptotic MSEs of $\hat{m}_T$ and $\hat{m}_E$ are identical when the two estimators use the same bandwidth.

Introducing the estimator $\hat{m}_T$ certainly does not settle the mean squared error issue. However, $\hat{m}_T$ is attractive in that it avoids both the random denominator problem of $\hat{m}_E$ and the down weighting pathology of $\hat{m}_C$. Another nice feature of $\hat{m}_T$ is that, like $\hat{m}_C$, it has a convenient form for estimating m', so long as $\hat{F}$ is differentiable. Considering $\hat{m}_T$ also brings into light the question of estimating the regression-quantile function mQ, an object whose importance has been stressed by Parzen (1981). Since it is natural to use a fixed, evenly spaced design on $[0,1]$ to estimate mQ, the convolution estimator seems ideally suited for estimating regression-quantile functions.

My final point concerns the use of kernel methods to test the adequacy of linear models. I was glad that Chu and Marron mentioned the problem of testing for linearity, and the attendant importance of how $\hat{m}_C$ and $\hat{m}_E$ perform when m is a straight line. I prefer $\hat{m}_C$ over $\hat{m}_E$ for purposes of testing linearity, since, as Chu and Marron point out, $\hat{m}_C$ has smaller bias than $\hat{m}_E$ in the straight line case. Indeed, Hart and Wehrly (1991) show that a boundary-corrected version of $\hat{m}_C$ (with bandwidth h) tends to a straight line as h tends to infinity. The limiting line is a consistent estimator of m when $m(x) = \beta_0 + \beta_1 x$. Higher-order kernels can be used to obtain kernel estimates that are polynomials (of any given degree) for large h. Such kernel estimates are a crucial part of a test proposed by Hart and Wehrly (1991) for checking the fit of a polynomial.

Comment

M. C. Jones

It is a great pleasure to congratulate the authors on a most informative, thought-provoking and, above all, *balanced* investigation of the issues involved in choosing between versions of the kernel regression estimator.

Chu and Marron (henceforth C&M) understandably concentrate on comparing and contrasting the two kernel estimators probably most widely employed in the literature: the Nadaraya-Watson (N-W) estimator, $\hat{m}_E$, and the Gasser-Müller

M. C. Jones is Lecturer, Department of Statistics, The Open University, Milton Keynes MK7 6AA, United Kingdom.

(G-M) estimator, $\hat{m}_C$. I was, nonetheless, surprised to see no mention at all of the Priestley and Chao (1972) (P-C) estimator, which also gets cited as being *the* kernel regression estimator sometimes, and which will figure later in my comments. Conceptually, however, I would like to widen the discussion to that of competition between two underlying *classes* of kernel regression estimators. Earlier drafts of C&M's paper called the G-M estimator $\hat{m}_I$ (I for Integral; why C for convolution when $\hat{m}_E$ is based on convolutions, as indeed are all kernel estimators, too?) and this suited my purpose well. For I would like to draw a distinction between methods that are E for *external*, in their treatment of f during the "main" kernel smoothing, and those that are I for *internal* in this respect. And fortuitously, in the random design case, $\hat{m}_E$ is a representative of the E class and, *in the same random design case*, the G-M estimator is, slightly indirectly, a representative from the I class.

My personal preferences in different situations will emerge as these comments progress. Interestingly, I never plump for exactly N-W or G-M, although I do recognize that often one or both of these would be a perfectly satisfactory alternative to my suggestion. I was originally going to suggest that, as in the fixed design case one knows f and in the random design case one doesn't (usually), so the requirement for two different kernel estimators for the two situations, one using f itself, the other incorporating estimation of it, seems to be very natural. However, it then occurred to me that the quest for a single version suitable for all cases was not so unreasonable because of the high quality of (kernel) estimated f's based on their quantiles, that is, use of random design case choices for fixed designs is actually perfectly sensible.

A much more detailed version of these comments, together with certain generalizations of this work, is given in Jones and Davies (1991).

1. THE FIXED UNIFORM DESIGN CASE

There is little to say about the (important) fixed uniform design case beyond stressing the word "very" in C&M's statement in their abstract that the N-W and G-M estimators "give very nearly the same performance" in this case. Both these and other versions of the kernel regression estimator are essentially indistinguishable in practice as well as in theory (apart perhaps from boundary effects, which, like C&M, I do not consider here).

A particularly natural and simple "other version" is the "naive" kernel formula

$$(1) \qquad \hat{m}_U(x) = n^{-1} \sum_{j=1}^{n} Y_j K_h(x - x_j).$$

Why not save oneself a little complication and use this formulation when the fixed uniform design case is the only one of interest? It has the advantages of being extendible to derivative estimation, boundary correction and the multivariate case all together. I can think of one objection that I suspect has contributed to (1) being viewed with suspicion, but that I claim is a red herring, namely that the weights in the weighted average of the Y's don't add exactly to one. While this matters at the boundary, preboundary correction, otherwise the difference from unity is asymptotically and practically negligible (witness the similarity with its "corrected" version N-W: the uniformity of f drives this). This estimator is precisely that of Priestley and Chao (1972).

2. THE FIXED NONUNIFORM DESIGN CASE

If there is one main point that I would wish to emerge clearly from these comments, it is this: $Y_1, Y_2, \ldots, Y_n$ yield information directly about the function $r(x) \equiv m(x)f(x)$, and not about $m(x)$ itself, in fixed or random nonuniform design cases. This is so in the sense that $\hat{m}_U(x)$ has expectation (essentially) $(K_h{*}r)(x) = r(x) + O(h^2)$, where $*$ denotes convolution. I guess f is actually G' in the notation of the paper for fixed designs. Hence, we have to do something about the nuisance function f to be able to get at m itself.

Well, with f known in the current fixed design context, that should be easy: Simply divide $\hat{m}_U(x)$ by $f(x)$, that is, use

$$(2) \qquad \mu_E(x) = \{nf(x)\}^{-1} \sum_{j=1}^{n} Y_j K_h(x - x_j).$$

This is my prototype E for *external* method of coping with f: $1/f(x)$ appears externally to the summation over datapoints. Johnston (1979) is an early reference on this although Härdle (1990) gives other references including earlier work, in Polish, by W. Greblicki.

Immediate as this is, there is an alternative. As $\{Y_j, j = 1, 2, \ldots, n\}$ pertains directly to r, so $\{Y_j/f(x_j), j = 1, 2, \ldots, n\}$ pertains directly to m. Kernel smooth this adjusted dataset to get

$$(3) \qquad \mu_I(x) = n^{-1} \sum_{j=1}^{n} \{f(x_j)\}^{-1} Y_j K_h(x - x_j).$$

And here, f^{-1} appears I for *internally* to the summation. It is Mack and Müller (1989b) who explicitly proposed this and gave me much inspiration for these comments. Formula (3) is more attractive than is formula (2) as a basis for estimating derivatives of m.

In Jones and Davies (1991), we look at the properties of (2) and (3) in the fixed nonuniform design case. Here, though, I will move on to the case of random designs without any further ado.

3. THE RANDOM DESIGN CASE

As well as replacing x_j by X_j in formulas (2) and (3), the challenge now is to estimate the unknown f-dependent quantities there. There is but one representative of the E approach with estimated f in the literature, and it is, of course, the N-W estimator, $\hat{m}_E$. C&M make a number of fine points about the comparison of N-W's bias with that of G-M. I might just add that the form of $\hat{m}_C$'s bias might be the easier of the two to estimate as part of an automatic bandwidth selection method.

So to internal estimators. G-M's relationship to this approach will be most clear if we first consider its relationship with P-C. Write $X_{(j)}$ for the jth-order statistic of the X's and $Y_{[j]}$ for its concomitant Y. The extension of (1) to "remove the restriction that the [X]'s are equally spaced" briefly suggested by Priestley and Chao (1972) was

$$\hat{m}_{PC}(x) \equiv \sum_{j=1}^{n} \left(X_{(j)} - X_{(j-1)}\right) Y_{[j]} K_h\left(x - X_{(j)}\right), \tag{4}$$

(with suitable definition of $X_{(0)}$). As Mack and Müller (1989a) and others have explicitly noted, $\hat{m}_{PC}$ is extremely close to $\hat{m}_C$ *using* $\beta = 1$ (in the notation of C&M). But C&M show that, in G-M, one should really use $\beta = 1/2$, and the equivalent P-C-type representation of that case is

$$\hat{m}_{PCv}(x) \equiv \sum_{j=1}^{n} \frac{1}{2}\left(X_{(j+1)} - X_{(j-1)}\right) Y_{[j]} K_h\left(x - X_{(j)}\right). \tag{5}$$

The apparent folklore that says that "P-C is not as good as G-M" is very largely based, it seems to me, on comparing (4) with $\hat{m}_C$, where from now on all references to $\hat{m}_C$/G-M revert to assuming $\beta = 1/2$. Compare like with like instead, that is, $\hat{m}_{PCv}$ with G-M, and we again have a pair of estimators that are not far from indistinguishable in practice.

Now, $\hat{m}_{PC}$ and $\hat{m}_{PCv}$ immediately fit into the framework of "estimated I class" methods, that is, they are of the form

$$n^{-1} \sum_{j=1}^{n} \tilde{q}_j Y_j K_h\left(x - X_j\right); \tag{6}$$

cf. (3). Here $\tilde{q}_j$ is shorthand for any estimator of $f(X_j)^{-1}$. From this viewpoint, then, the natural interpretation of G-M is as an approximation to $\hat{m}_{PCv}$ and not vice-versa! Indeed, I fail to see what *fundamental* role the integration in $\hat{m}_C$ serves. I have already argued that forcing weights to sum exactly to one is a minor consideration (and there is a more immediate way of doing so in $\hat{m}_{PCv}$). Kernel smoothing the smoothing afforded by the initial piecewise constant function employed by G-M is, very loosely, akin to a single kernel smoothing using some $K_h{}^*L_l$ (exactly so with uniform L in the fixed uniform design case) but where the other bandwidth l is of order n^{-1} and doesn't really have a noticeable effect. So, again, this time in $\hat{m}_{PCv}$, we have a simpler kernel estimator that performs very much like G-M.

All this business about coping with, and estimating, nonuniform f is, of course, a more formal way of looking at the insightful intuitive arguments of C&M's Section 3 and early Section 4.

If one takes the view that G-M's bias is the desirable one, it remains a nice question to obtain a kernel estimator that has this bias but at no expense in terms of variance, that is, one with MSE

$$h^4\{m''(x)\}^2\left(\int u^2K\right)^2 \Big/ 4 + (nh)^{-1} f(x)^{-1}\sigma^2 \int K^2.$$

This is the challenge taken up in Jones and Davies (1991), where we argue our way to a proposal that turns out to be fairly closely related to the main content of C&M's Section 6. Wu and Chu (1991) have independently provided an alternative method with this same property. Fan (1990), with a different idea again, provides a method with an advantage that, to me at the moment of writing, appears to be the most exciting of them all. By the way, following the comments towards the end of my introduction, Jones and Davies' estimator, for example, reduces exactly to (1) for fixed uniform designs, but only approximates (3) for fixed nonuniform.

4. A SPECIFIC REMARK

The bias in C&M's Figure 11a of the N-W estimator with Gaussian kernel applied to data relating to a linear mean, $m(x) = ax + b$ say, and based on a standard normal design (with density ϕ), is easily explained and corrected *in this particular case* as follows. The bias is (essentially) entirely due to the well-known variance inflation property of the kernel density estimate (see Jones, 1991, and references therein), that is, the variance that the estimate associates with the design is $1 + h^2$ instead of unity. In fact, N-W's expectation $(\phi_h{}^*r)(x)/(\phi_h{}^*\phi)(x) = (1 + h^2)^{-1/2}ax + b$ here, so the mean is (essentially) the line with slope shrunken by the factor $(1 + h^2)^{-1/2}$ (but the same

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.