approximation, we need n operations to discretize the data into N_B nonempty bins. Thus, the numerical effort for this method is of order $O(n + N_B M)$.

Of course, the WARPing method introduces a discretization bias. The bias may be reduced by joining the obtained discrete step function (see (3.2)), via a polygon. Breuer (1990) has computed for $m(x) = x \sin(2\pi x) + 3x$ and uniform design the MSE as a function of x for both the $\hat{m}_E$ estimator and the WARPed estimator $\hat{m}_{M,K}$.

In Figure 5, the discretization bias is seen to be quite drastic, although we gained in speed of computation. The linear interpolant has a much better bias behavior, as is seen in Figure 6. For this estimator conservative bounds for the numerical discretization error and its effect on $MSE(x)$ can be given and are displayed in Figure 6 as long dashed lines.

ACKNOWLEDGMENT

The work of the first author was in part financially supported by CentER, Tilburg.

Comment

Jeffrey D. Hart

Chu and Marron have provided us with a clear and thorough account of the relative merits of evaluation and convolution type kernel regression estimators. One is left with the impression that neither type of estimator is to be preferred universally over the other. We learn, for example, that the weights of the convolution estimator sometimes have the unsettling behavior exhibited in Figures 6b and 7 of Chu and Marron. The authors make it clear that there are a number of factors, including type of design (fixed or random), design density and nature of underlying regression function, that need to be considered before choosing an estimator type. Having reading their article, I now have a slight preference for $\hat{m}_E$ over $\hat{m}_C$ in the random design case, at least in the absence of any information about the design density or regression curve. When the design points are nonrandom and evenly spaced, I prefer $\hat{m}_C$, since its convolution form appeals to me, and since boundary kernels are easy to construct with $\hat{m}_C$ (see Gasser and Müller, 1979). Below I will mention a modification of $\hat{m}_C$ that I feel is a viable competitor of $\hat{m}_E$ even in the random design case.

The authors' point about the down weighting phenomenon of the convolution estimator is certainly well taken. However, I would like to question an aspect of their comparison of the variances of $\hat{m}_E$ and $\hat{m}_C$. As the authors note in Section 4, the biases of the two estimators are not comparable, the bias of $\hat{m}_E$ being smaller in some cases and that of $\hat{m}_C$ smaller in other cases. It follows that "good" bandwidths for the estimators will generally be different. Why then is it sensible to compare $\mathrm{Var}(\hat{m}_E)$ and $\mathrm{Var}(\hat{m}_C)$ at the same value of h?

A little-used but informative way of comparing the errors of $\hat{m}_E$ and $\hat{m}_C$ is to consider the limiting distribution of

$$(1) \qquad \frac{|\hat{m}_E(x) - m(x)|}{|\hat{m}_C(x) - m(x)|}.$$

Unlike an MSE comparison, this approach takes into account the joint behavior of the two estimators. Suppose that Chu and Marron's assumptions (A.1)–(A.5) hold and that the design density is $U(0,1)$. Suppose further that the bandwidths of $\hat{m}_E$ and $\hat{m}_C$ minimize their respective MSEs. Then it can be shown that, for each x, the ratio (1) converges in distribution to

$$(2) \qquad \left(\frac{2}{3}\right)^{2/5} \frac{|Z_1 + 1/2|}{|Z_2 + 1/2|} = R$$

as $n \to \infty$, where (Z_1, Z_2) have a bivariate normal distribution with $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$ and

$$\mathrm{Corr}(Z_1, Z_2) = \left(\frac{2}{3}\right)^{3/5} \int K(z) K\left(\left(\frac{2}{3}\right)^{1/5} z\right) dz \Big/ \int K^2 = \rho_K.$$

Jeffrey D. Hart is Associate Professor of Statistics, Texas A&M University, College Station, Texas 77843.

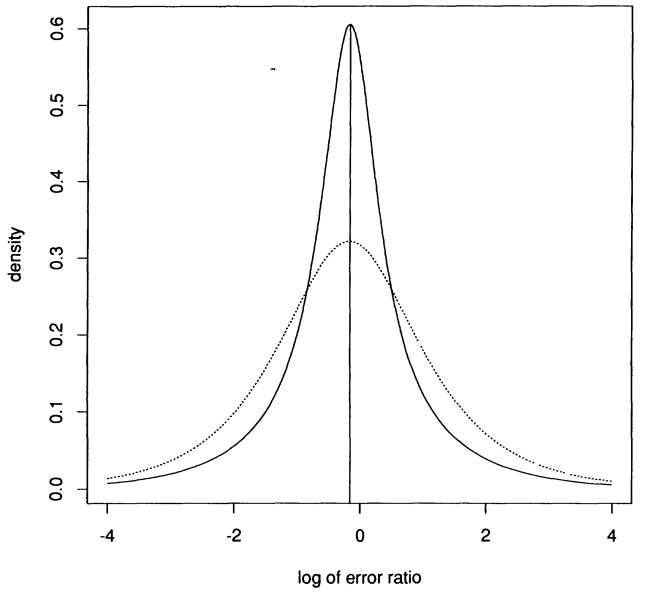

FIG. 1. *Limiting density functions of the logarithm of an error ratio. The solid curve is the actual limiting density of* $\log|\hat{m}_E(x) - m(x)|/|\hat{m}_C(x) - m(x)|$, *while the dashed curve is what the limiting density of the same quantity would be if* $\hat{m}_E(x)$ *and* $\hat{m}_C(x)$ *were asymptotically uncorrelated. The vertical line is drawn at the abscissa* $\log(0.85)$, *which is the limiting median of* $\log|\hat{m}_E(x) - m(x)|/|\hat{m}_C(x) - m(x)|$.

The quantity ρ_K is the limiting correlation between $\hat{m}_E$ and $\hat{m}_C$. For the kernels used in practice, $\rho_K \approx (2/3)^{3/5} = 0.784$. For the Epanechnikov kernel, that is, $K(z) = 0.75(1 - z^2)I_{(-1,1)}(z)$, $\rho_K = 0.813$. The density function of the random variable R (defined in (2)) is shown in Figure 1. It is noteworthy that the median of R, and hence the limiting median of $|\hat{m}_E(x) - m(x)|/|\hat{m}_C(x) - m(x)|$, is equal to

$$\lim_{n\to\infty}\left(\frac{\mathrm{MSE}(\hat{m}_E)}{\mathrm{MSE}(\hat{m}_C)}\right)^{1/2} = \left(\frac{2}{3}\right)^{2/5} = 0.850.$$

However, we also learn from this analysis that, for n large,

$$P(|\hat{m}_C(x) - m(x)| < |\hat{m}_E(x) - m(x)|) \approx 0.4.$$

To illustrate the effect of the high positive correlation between $\hat{m}_E$ and $\hat{m}_C$, Figure 1 also shows the density of

$$\left(\frac{2}{3}\right)^{2/5}\frac{|U_1 + 1/2|}{|U_2 + 1/2|},$$

where U_1, U_2 are iid $N(0,1)$. This would be the limiting distribution of, say,

$$\frac{|\hat{m}_1(x) - m(x)|}{|\hat{m}_2(x) - m(x)|}$$

on the assumption that $\hat{m}_1(x)$ and $\hat{m}_2(x)$ are asymptotically uncorrelated and

$$\lim_{n\to\infty}\frac{\mathrm{MSE}(\hat{m}_1(x))}{\mathrm{MSE}(\hat{m}_2(x))} = \lim_{n\to\infty}\frac{\mathrm{MSE}(\hat{m}_E(x))}{\mathrm{MSE}(\hat{m}_C(x))}.$$

At this point, I would like to suggest a simple variation of the convolution type estimator that I have found to be quite useful. This variation uses an idea proposed by Yang (1981) and studied by Stute (1984) and Carroll and Härdle (1989). Define

$$\hat{m}_T(x) = \frac{1}{h}\sum_{j=1}^{n} Y_{(j)}\int_{(j-1)/n}^{j/n} K\left(\frac{\hat{F}(x) - u}{h}\right)du,$$

where $\hat{F}$ is a $\sqrt{n}$-consistent estimator of the cdf F of X_1. The T in $\hat{m}_T$ stands for transformation, since this estimator uses transformed design points that are evenly spaced on $[0, 1]$. To ensure that $\hat{m}_T$ has a pleasingly smooth appearance, one should take $\hat{F}$ to be a slightly smoothed version of the usual empirical cdf F_n. Carroll and Härdle (1989) suggest using $\hat{F} = F_n$, but this often leads to estimates of m with a stair-step look. I have also noticed that $\hat{m}_T$ can be sensitive to fairly small changes in $\hat{F}$. This property needs to be investigated before $\hat{m}_T$ can be recommended for routine use.

One way to motivate $\hat{m}_T$ is to first introduce what Parzen (1981) calls the *regression-quantile* function mQ, where $mQ(t) = m[Q(t)]$, $0 < t < 1$, and Q is the quantile function of X_1. Define the estimator $\widehat{mQ}$ by

$$\widehat{mQ}(t) = \frac{1}{h}\sum_{j=1}^{n} Y_{(j)}\int_{(j-1)/n}^{j/n} K\left(\frac{t - u}{h}\right)du,$$

which is simply the convolution type estimator of $mQ(t)$ based on the *nonrandom* design points $(j - 1/2)/n$, $j = 1, \ldots, n$. Noting that $\hat{m}_T(x) = \widehat{mQ}(\hat{F}(x))$, $\hat{m}_T(x)$ estimates $m(x)$ inasmuch as $\widehat{mQ}(\hat{F}(x)) \approx mQ(F(x)) = m(x)$.

The estimator $\hat{m}_T$ does not have the down weighting pathology suffered by $\hat{m}_C$. Regardless of whether the design is fixed or random, the weight for $Y_{(j)}$ in $\hat{m}_T(X_{(j)})$ is (to a good approximation) $K(0)/(nh)$. One result of this weighting scheme is that the variance of $\hat{m}_T(x)$ tends to behave, in all cases, like that of $\hat{m}_C(x)$ in the case of a fixed, evenly spaced design. Arguing as in Stute (1984) and Carroll and Härdle (1989), it can be shown that, under technical assumptions much like Chu

and Marron's (A.1)–(A.5),

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

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))

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

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.