Statistical Science
1991, Vol. 6, No. 4, 404–436

Choosing a Kernel Regression Estimator

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

Abstract. For nonparametric regression, there are two popular methods for constructing kernel estimators, involving choosing weights either by direct kernel evaluation or by the convolution of the kernel with a histogram representing the data. There is an extensive literature concerning both of these estimators, but a comparatively small amount of thought has been given to the choice between them. The few papers that do treat both types of estimator tend to present only one side of the pertinent issues. The purpose of this paper is to present a balanced discussion, at an intuitive level, of the differences between the estimators, to allow users of nonparametric regression to rationally make this choice for themselves. While these estimators give very nearly the same performance in the case of a fixed, essentially equally spaced design, their performance is quite different when there are serious departures from equal spacing, or when the design points are randomly chosen. Each of the estimators has several important advantages and disadvantages, so the choice of "best" is a personal one, which should depend on the particular estimation setting at hand.

Key words and phrases: Asymptotic variance, design points, kernel estimators, nonparametric regression.

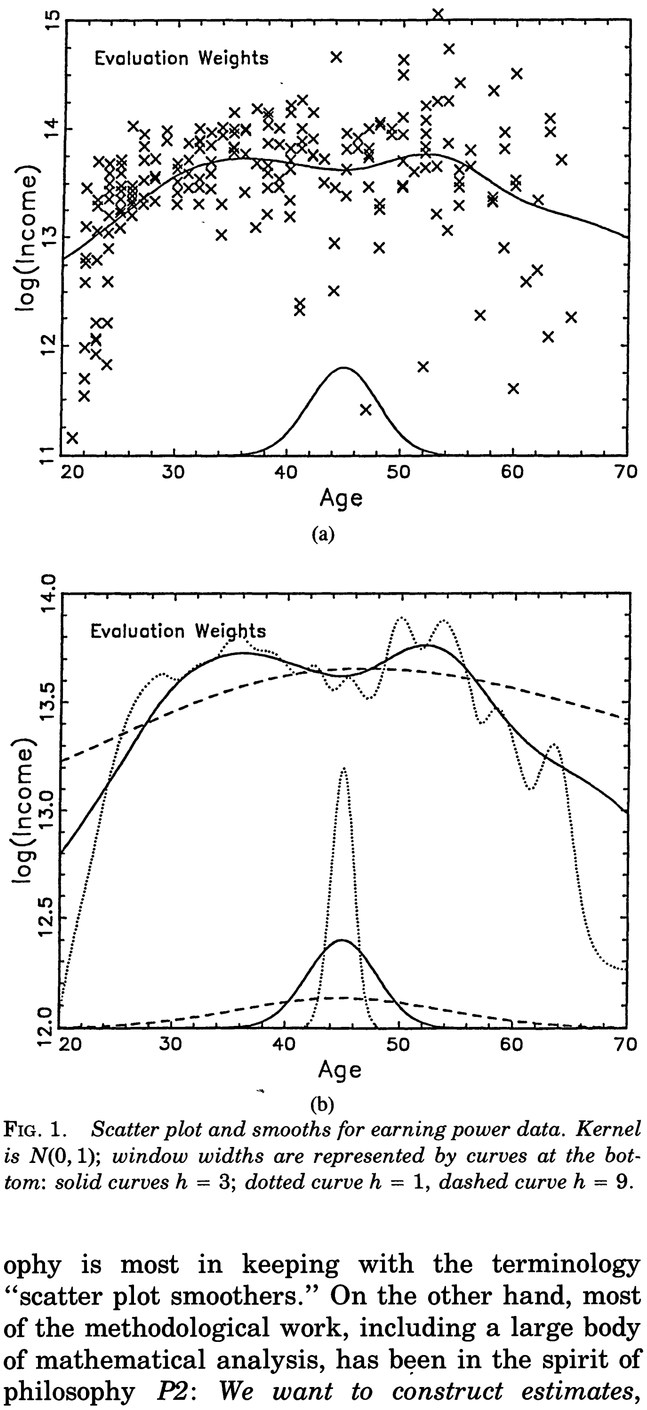
1. INTRODUCTION

Nonparametric regression, by smoothing methods, has been well established as a useful data-analytic tool. Figure 1a shows an example of this. The data here are from Ullah (1985). The scatter plot shows 205 pairs of log(income) versus age, as described in that paper. The solid curve is a "scatterplot smoother," which is a moving weighted average of the points in the scatter plot, where the weights are proportional to the curve at the bottom. Scatter plot smoothers are often also called nonparametric regression estimators, because the points can be viewed as coming from a bivariate probability structure, for which the local average provides an estimate of the conditional expectation. As expected, there is a general increase in earning power with increasing age, over the younger years, with a tendency to level off as age increases. Perhaps not too surprising is the fact that average income actually decreases in the latter years, as more and more people retire. However, the dip in average income in the middle ages is certainly unexpected and, if not merely a consequence of the sampling error, is clearly of strong interest because it represents the discovery of a new economic phenomenon. A. Ullah, C. Robinson and K. Ball have found a strong indication in favor of the existence of this dip, by noticing its presence also in other data sets of this type. Two important questions are the subject of ongoing research by these workers: (a) Is this part of the underlying structure or only a sampling artifact? (b) What mechanism causes this phenomenon? Deeper discussion of this example goes beyond the scope of this paper. Its purpose here is to illustrate the point: Nonparametric regression is a simple and useful tool for obtaining insight into the structure of data.

See the monographs by Eubank (1988), Müller (1988), Härdle (1990) and Wahba (1990) for a large variety of other interesting real data examples where applications of such methods have yielded analyses essentially unobtainable by other techniques, and also for access to the literature concerning these estimators (see also Cheng, 1991).

For intuitive understanding of smoothers, there are two common viewpoints. Those who focus most on data analysis, often have uppermost in their minds the philosophy we will call *P1*: *We are looking for structure in this set of numbers*. This philos-

C.-K. Chu is Associate Professor, Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan 30043, Republic of China. J. S. Marron is Associate Professor, Department of Statistics, University of North Carolina, Chapel Hill, North Carolina 27707.



(a)

(b)

FIG. 1. *Scatter plot and smooths for earning power data. Kernel is* $N(0, 1)$; *window widths are represented by curves at the bottom*: *solid curves* $h = 3$; *dotted curve* $h = 1$, *dashed curve* $h = 9$.

ophy is most in keeping with the terminology "scatter plot smoothers." On the other hand, most of the methodological work, including a large body of mathematical analysis, has been in the spirit of philosophy *P2*: *We want to construct estimates, based on some data from an underlying probability structure.* This second philosophy is in the spirit of the name "nonparametric regression estimators." We feel that for proper understanding of this subject, both points of view provide important insights, and both need to be kept in mind. We believe that P2 is very useful for learning many important things, however the relevance of the lessons learned from that point of view should always be assessed from the P1 side as well.

From either point of view, the most important thing to know about any type of regression smoothing is that the amount of smoothing needs to be specified. This is demonstrated in Figure 1b, which shows three different smooths, for the same data set as in Figure 1a. The three different curves there represent three different choices of the width of the window used in the local average. Note that the shape of the resulting smooth is highly dependent on the choice of this width. When the width is too small, the curve feels sampling variation too strongly. From the viewpoint P2, there is too much "variance" present because each part of the moving average is making use of too few observations. On the other hand, when the width is too large, important features disappear, because points from too far away are being used in each local average. Those who adopt the viewpoint P2, will say there is too much "bias" present.

This dependence of the result on the window width makes it clear why the problem (a) mentioned above is indeed a challenging one, because, with a large enough window width, the dip in the middle incomes can be made to disappear. This shows that there is a price to be paid for the use of smoothers: Important insights can be easily obtained because of their great flexibility, but this same flexibility is also a curse because it requires choice of the degree of smoothing. An important subject is the use of the data to choose the window width, however there are still no universally accepted methods for this; see Marron (1988) for a survey. In this paper, we take the approach that is currently used most frequently in data analysis: The window width is chosen visually by a trial and error approach.

The simplest and most widely used regression smoothers are based on kernel methods, although strong reasons for considering other possibilities, especially splines, are discussed in Eubank (1988) and Wahba (1990). The name "kernel" comes from the fact that these smoothers are local weighted averages of the response variables, whose weights are somehow based on a "kernel function." The precise way that this kernel function is used can make quite a difference, and indeed comparison of the major ways of doing this is the main point of this paper. On the other hand, shape of the kernel function (i.e., the curves appearing at the bottoms of Figures 1a and b) makes very little difference; see, for example, the monographs Müller (1988) or Härdle (1990). In this paper, a Gaussian kernel is always used, because we like its visual appearance slightly more, but this choice is personal.

Nadaraya (1964) and Watson (1964) proposed choosing the weights by evaluating a kernel function at the design points, and then dividing by the sum of the weights, so that they add up to 1. These weights are called "evaluation" weights here. The essential idea is illustrated in Figure 2, which demonstrates the construction of this estimator, for a data set that is contrived to make a certain point later, but also suffices for this purpose. The solid curve comes from a moving average of the data, which is constructed by sliding the "window function" (i.e., "kernel function") shown at the bottom along the x axis, and simply calculating the average of the points in the window, using weights proportional to the height of the kernel above each corresponding x coordinate. These heights are shown as solid vertical bars in the figure, for the current kernel location. A drawback to this approach is that the estimator becomes tricky to analyze from the technical view point, that is, under philosophy P2. Indeed, for some kernel functions, Härdle and Marron (1983) have shown that the moments of such an estimator can fail to exist, when the predictor variables are also random.

An alternative, which overcomes this problem, is to consider kernel smoothers based on "convolution" of a kernel function with some function representing the raw data in an absolutely continuous sense. The first version of this that we are aware of was proposed by Clark (1977, 1980), although here we will study a version which is currently more strongly advocated, due to Gasser and Müller (1979). See Jennen-Steinmetz and Gasser (1988) for literature on other closely related estimators. Intuitive insight into how these convolution estimators are constructed is given in Figure 3, which again shows a data set contrived to make a point later. The histogram is that picture is one way of representing the data in a "continuous sense," that is, the "mass" of each data point is represented by one step in this step function. Now the convolution smoother, shown as the curve going through the data points, is simply the convolution (this can be viewed as a "continuous moving average") of this step function with the kernel function, which is the curve at the bottom of the picture. The Clark version of this estimator replaces the histogram with a function which linearly interpolates the observations.

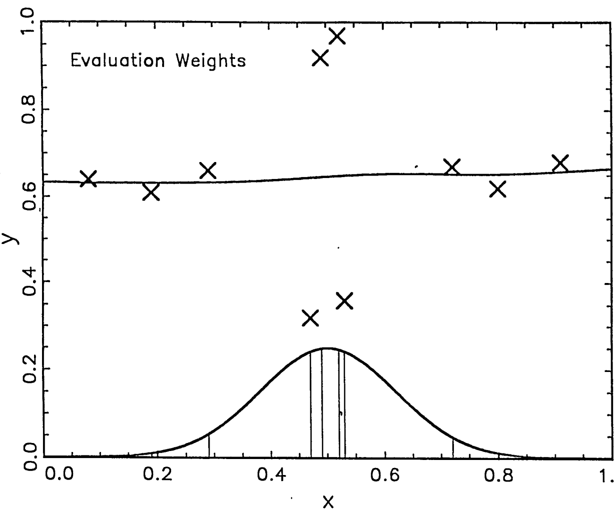


FIG. 2. *Intuition behind evaluation weighted kernel smoother. Solid curve is moving average of scatter plot, with weights chosen proportional to heights of kernel function (curve at bottom) evaluated at ordinates of data points.*

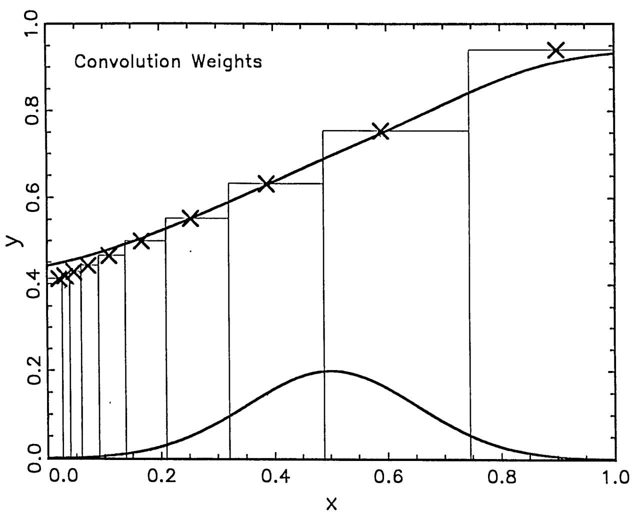


FIG. 3. *Intuition behind convolution weighted estimator. Solid curve is convolution (i.e., continuous moving average) of step function representing observations, with kernel function (curve at bottom).*

Because the evaluation weighted estimator is in fact a discrete version of the continuous convolution used in Figure 3, one should intuitively expect that quite often there will be little practical difference between the evaluation and convolution weighted estimators, which is indeed often the case. If the design points are equally spaced (or essentially so, as defined below), then the evaluation and convolution weighted estimators are very nearly the same, by the integral mean value theorem. However, when the design points are not equally spaced, or when they are iid random variables, there are very substantial and important differences in these estimators. The main point of this paper is to make clear both sides of the issues involved. While opinions have been expressed in both directions, in fact there can never be an absolute resolution of which is "best." For some purposes one will be better, for other purposes the other is superior.

The advantages of the evaluation weighted estimator in the random design case are: (1) superior variance qualities (discussed in Section 3) and (2) superior performance in high dimensions (discussed in Section 7). Advantages of the convolution weighted estimator are: (1) superior interpretability and "null space" properties, in the nonuniform case (discussed in Section 4) and (2) more straightforward generalization to the estimation of derivatives and easier adjustment for boundary effects (discussed in Section 7).

Section 2 introduces mathematical notation, including precise formulation of the estimators, and various setups commonly used under philosophy P2.

Section 3 illustrates, both through simple examples, and also through asymptotic analysis, situations in which the evaluation weighted estimator is superior to the convolution weighted estimator. The main idea is that because the bars in the histogram of Figure 3 can have unequal width, some of the observations can be severely down weighted by the convolution estimator, which can lead to severe inefficiencies, in terms of increased variance, in several different ways.

Section 4 discusses the other side of the choice of which estimator to use. In particular, examples and also asymptotic analysis are used to show when the convolution estimator is superior. The key idea here is that precisely this same adjustment in the weighting scheme that can lead to inefficiency in some situations can also be of large practical and intuitive benefit in others, of a type important to the other side of the smoothing problem, bias.

Because there are situations in which each estimator will clearly be superior, the choice of which should be used must depend on the particular context at hand. In particular, this choice should be made on the basis of whether one is willing to pay the price in terms of increased variance, shown in Section 3, for the bias advantages discussed in Section 4. Section 5 shows how variance and bias should be considered together in making this necessarily personal trade-off.

Section 6 discusses two possibilities for combining the best aspects of both estimators.

Other issues that can affect the choice between evaluation and convolution weighted estimators are discussed in Section 7. In particular, it is known that the convolution weighted estimator has severe problems with the bias approximation in the case of a high-dimensional design, while the evaluation weighted estimator is unaffected. On the other hand, because it is a fraction of sums, it is more complicated to do boundary adjustments and to construct derivative estimates using the evaluation weighted estimator.

2. THE ESTIMATORS

In this section, mathematical models, under philosophy P2, for studying scatter plot smoothers are given. There are two such models commonly considered. One is M1: "fixed design," where the ordinates of the data points in the scatter plot are thought of as being deterministic values. These ordinate values are usually chosen by the experimenter, as in a designed experiment. Since in the absence of other information it is intuitively sensible to take such points to be equally spaced, this is frequently the case. The other is M2: "random design," where the data points in the scatter plot are thought of as being realizations from a bivariate probability distribution. These ordinates are usually not chosen by the experimenter, as in sample surveys, or other types of observational studies.

The fixed design model, M1, is given by

$$Y_j = m(x_j) + \varepsilon_j,$$

for $j = 1, 2, \ldots, n$, where $m(x)$ is the "mean" or regression function, the x_j's are nonrandom design points with $0 \le x_1 \le x_2 \le \cdots \le x_n \le 1$ (without loss of generality), and the ε_j's are independent random variables with mean 0 and variance σ^2. The random design model, M2, is given by

$$Y_j = m(X_j) + \varepsilon_j,$$

for $j = 1, 2, \ldots, n$, where the (X_j, Y_j)'s are independent identically distributed random variables, with ε_j defined by $\varepsilon_j = Y_j - m(X_j)$ and assumed to have mean 0 and variance σ^2. In each case, the goal, under P2, is to use $Y_1, \ldots, Y_n$ to estimate the curve $m(x)$. In both cases, the technical assumptions can be weakened substantially in several directions, with no major changes in the key ideas. However, weaker assumptions are technically cumbersome, so we stick with these for simplicity of presentation.

At first glance, one might think there is little practical difference between these models, because the regression function (i.e., conditional expected value), only depends on the *conditional* distribution, where the ordinate values are given. While this is correct, it will be seen that the particular configuration of these ordinate values is very relevant to the performance of the two types of estimators. Because the way in which this happens can be intuitively understood quite well by thinking of these two models, both are considered here.

The evaluation weighted estimator of $m(x)$, for $0 < x < 1$ (defined here for the fixed design, for the random design case replace x_j by X_j), motivated intuitively in Figure 2, is given by

$$\hat{m}_E(x) = \frac{n^{-1} \sum_{j=1}^{n} K_h(x - x_j) Y_j}{n^{-1} \sum_{j=1}^{n} K_h(x - x_j)},$$

where $K_h(\cdot) = h^{-1} K(\cdot / h)$ (if the denominator is 0, take $\hat{m}_E(x) = 0$). For $0 < x < 1$, the convolution weighted estimator is given by (again explicitly defined only in the fixed design case)

$$\hat{m}_C(x) = \sum_{j=1}^{n} Y_j \int_{s_{j-1}}^{s_j} K_h(x - t)\, dt,$$

where $s_0 = -\infty$, $s_n = \infty$, $x_j \le s_j \le x_{j+1}$ for $j = 1, \ldots, n-1$, and K is a density function (in the random design case x_j is taken to be the jth order statistic $X_{(j)}$ among the X's and Y_j is replaced by $Y_{(j)}$, which is the corresponding Y). The choice given here for s_0 and s_n ensure that the sum of the weights is one. This will create a strong "boundary effect," because near either end the observation at the end will receive a very large weight. This should usually be adjusted for; see, for example, Section 4.3 of Müller (1988). However, the best of such adjustments tend to be rather complicated, so again for simplicity of presentation, this is not done here. Another way of handling these boundaries is to take $s_0 = 0$ and $s_n = 1$, however this gives pictures with even more severe boundary effects, because near the edges the weights do not sum to 1 (so instead of giving large weight to the outermost data point, it is essentially given to the arbitrary value of 0).

A convenient structure for displaying most of the choices of the s_j that have appeared in the literature is $s_j = \beta x_j + (1 - \beta) x_{j+1}$, where β is a parameter allowed to range between 0 and 1. A β that is easy to work with is $\beta = 1$, which would put the vertical parts of the steps in the histogram in Figure 3 at the observations. However, from that picture it is clear that the step function better represents the data if one uses $\beta = 1/2$. In a number of early papers, little distinction has been made between these two, because in the fixed and essentially equally spaced case (this includes designs which satisfy the asymptotic condition $x_i = i/n + o(n^{-1})$), the practical difference is negligible. However, it will be seen in Section 3 that for the random design case there is quite a large difference, with $\beta = 1/2$ being clearly superior. In all examples constructed in this paper, $\beta = 1/2$, unless otherwise noted.

While it may not be immediately obvious, it is a straightforward calculation to check that $\hat{m}_C$ is indeed the convolution with the histogram illustrated in Figure 3. The mathematical form given above motivates another way of thinking about this estimator. Note that the integral of K over the subinterval is providing a weight for Y_j in the moving weighted average. This idea is demonstrated in Figure 4, which gives a feeling for how these weights work, by representing them as areas between s_j's under the kernel function K, for a particular choice of x_j's and for $\beta = 1/2$.

The monographs Eubank (1988), Müller (1988) and Härdle (1990) are excellent sources for introduction to, and detailed discussion of many aspects of, these and related estimators.

3. EFFICIENCY ISSUES

One way of seeing how the convolution weighted estimator can be inefficient is already being shown in Figure 4. Note that because the x_j's there are not equally spaced, the relative weight assigned to Y_5 will be much less than that assigned to either Y_4 or Y_6. The fact that this can have a very strong effect on the resulting smoother is demonstrated in Figure 5. The construction of the estimator shown in this picture is the same as in Figure 3, except that it is now applied to the artificial data set of

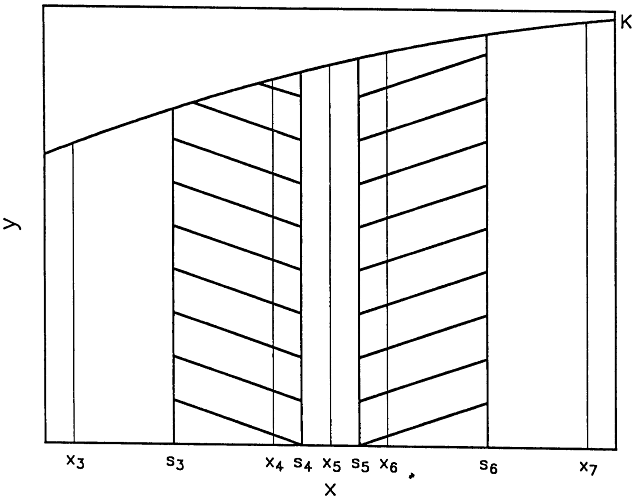


FIG. 4. *Macroscopic view of relative weights (shaded areas, and area between) given to observations by convolution weighted estimator.*

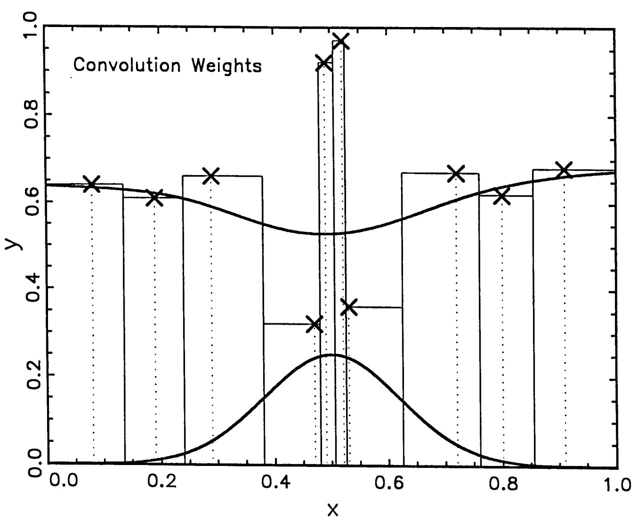
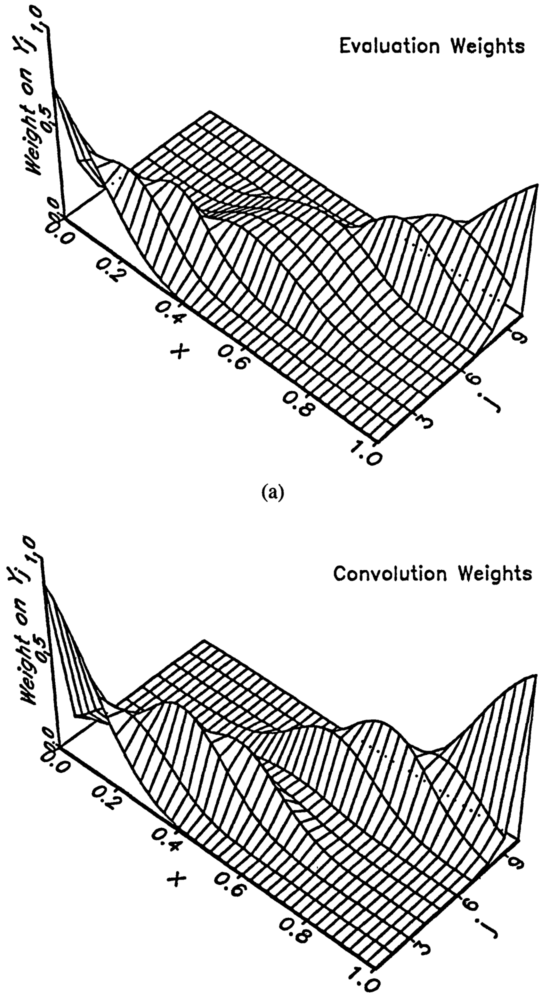


FIG. 5. *Convolution weighted estimator applied to same data, and also using same kernel and window width as in Figure 2.*

Figure 2. Note that the resulting smooth is now substantially different. In particular, near the center, the curve is pulled downwards by the two low observations in the center. This is disconcerting, since the visual impression of "average behavior of the data" is much different, because the downward influence of the low observations should be canceled by the two nearby high observations. However, because weights are proportional to the histogram widths, the convolution weighted estimator fails to make this cancellation. Simply because the ordinates of these high observations arbitrarily happen to be closer to other data points, they receive less weight. If the high observations were to trade ordinates with the low ones, the convolution estimator would then be pulled upward by roughly the same amount, but the visual impression of the data would still be essentially the same. On the other hand, note that the evaluation weighted estimator in Figure 2 is behaving in a more intuitively reasonable fashion, recovering conditional structure in a way that is much closer to what can be seen by eye.

Another means of viewing this down weighting effect of the convolution smoother is given in Figure 6. Recall that in both Figures 2 and 5, the smoothers are constructed, by calculating at each location on the axes, a weighted average of the Y_j's. Figure 6 summarizes the local structure of those weights by showing, for each value of x on the axes in Figures 2 and 5, and for each data point at site x_j, the effective weights for the jth observation. In Figure 6a, the heights of the surface at each point are calculated by evaluating the kernel function and dividing by the sum of the weights (i.e., those for the evaluation estimator), while in Figure 6b the heights are calculated by integrating K over a subinterval (recall this is equivalent to taking the convolution of the kernel with the data histogram). For a fixed j, the curves as a function of x show the weight applied to Y_j, as the kernel is moved along horizontally. Note each of these curves is a "bump" with its highest point at x_j, and tapering off elsewhere, which reflects the local averaging character of these estimators. In both figures, note the weights on the last observation become quite large, although this effect is less drastic for the evaluation weights. This is generally true, because the unadjusted convolution estimator puts so much weight on the observation closest to the edge. For this reason, boundary adjustments are much more important for the convolution estimator than they are for the evaluation estimator.

The main lesson from Figure 6 comes from looking at the weights for the Y_j in the interior. Note that, for the evaluation estimator, these are very nearly the same for the four interior points. On the other hand, the two center points (representing the high observations in Figures 2 and 5) are drastically down weighted by the convolution estimator, compared to their two nearest neighbors (for the low observations).



(a)

(b)

FIG. 6. *Surface plots, showing for each data point Y_j the weights as a function of location x, used in constructing the kernel smoothers in*: (*a*) *Figure* 2; (*b*) *Figure* 5.

Another way of looking at this down weighting effect, which allows some mathematical quantification of the inefficiency of the convolution weighted estimator, will be considered in the next example. The intuitive idea here is that illustrated in Figure 4 (and mathematically understood through the integral mean value theorem), the essential weight given to each observation by $\hat{m}_C$ is proportional to the length of the corresponding subinterval $[s_{j-1}, s_j]$.

Start with an equally spaced design (i.e., $x_j = j/n$), and consider consecutive triples of points. For each triple, move the first and the third towards the center. The points x_4, x_5, x_6 (locations shown by the heavy vertical lines) in Figure 4 gives an example of what is meant here. The amount of shift to the center can be parameterized by a value $\alpha \in [0,1]$, which results in the design

$$\begin{aligned} &\vdots \\ x_4 &= (5 - \alpha)/n \\ x_5 &= 5/n \\ x_6 &= (5 + \alpha)/n \\ &\vdots \end{aligned} \qquad .$$

Note that $\alpha = 1$ gives the usual equally spaced design, while $\alpha = 0$ gives a design that is also essentially equally spaced but with three replications at each point. We are not claiming this design is important in practice, however it is considered here because it provides a clear and simple illustration of the points being made, and the numerical answer appears in a surprising way later. The effects described will obviously also be present in more realistic unequally spaced designs, and, most important, they will provide an intuitive basis for understanding the causes of the same effects in the more complicated random design case.

The inefficiency of the convolution type estimator can now be seen at an intuitive level by considering the effect on the weight given to the center observation of each triple (i.e., x_5 in Figure 1), as α varies between 0 and 1. Note that the weights given to these points by the kernel evaluation estimator, $\hat{m}_E$, are nearly independent of α, while the weights assigned by the kernel convolution estimator, $\hat{m}_C$, are roughly proportional to α (this weight is the unshaded area between the two shaded areas in Figure 1). Hence, for α close to 0, the weight on the center observation is essentially 0, so the weighted average, $\hat{m}_C$, is making use of only 2/3 of the available observations. The extent of the inefficiency caused by this can be measured in terms of the asymptotic variance, which in view of this intuition should be expected to be larger by a factor of about 3/2 for α close to 0.

For simple mathematical quantification of these ideas, we will consider some simple asymptotic analysis. A very useful type of asymptotics in nonparametric regression has been to study the behavior of a sequence of estimators, in the limit as $n \to \infty$, with $h \to 0$ and $nh \to \infty$. The last two assumptions ensure that, as more information is added, only successively nearer points are used in each local average and that the local averages are taken over an increasing number of points, respectively. An important philosophical point is that asymptotics are not done because we feel that n is large, but instead because they provide an analytical tool, which enables us to see the simple main structure that underlies the rather complicated quantities being studied.

To facilitate the analysis, we will make the following technical assumptions (which again can be weakened in many ways, but with additional effort, which will tend to obscure the main ideas): (A.1) m is twice continuously differentiable on a neighborhood of the point x; (A.2) K is a symmetric, probability density supported on $[-1,1]$, bounded above 0 on $[-1/2,1/2]$, with a bounded derivative; and (A.3) $n \to \infty$, with $n^{-1+\delta} \le h \le n^{-\delta}$, for some $\delta \in (0,1/2)$.

Under these assumptions, for $0 < x < 1$,

$$\text{(3.1)} \quad \begin{aligned} &\operatorname{Var}(\hat{m}_E(x)) \\ &\quad = n^{-1}h^{-1}\sigma^2 \int K^2 + O(n^{-2}h^{-2}), \end{aligned}$$

$$\text{(3.2)} \quad \begin{aligned} &\operatorname{Var}(\hat{m}_C(x)) \\ &\quad = C(\alpha) n^{-1}h^{-1}\sigma^2 \int K^2 + O(n^{-2}h^{-2}), \end{aligned}$$

where $C(\alpha) = 1 + (\alpha - 1)^2/2$, $0 \le \alpha \le 1$. This can be shown by standard methods; see, for example, Section 4.1 of Müller (1988) (details in this case may be found in equations (2.2.2) and (2.2.3) of Chu (1989)). The main idea of the proof is the usual formula for the variance of a sum of independent random variables, together with a Riemann approximation of the resulting sum by $\int K^2$.

In addition to allowing simple comparison of these estimators, through the function $C(\alpha)$, the representations (3.1) and (3.2) demonstrate clearly the usefulness of asymptotic analysis, because they provide simple and insightful quantification of other aspects of nonparametric regression. For example, note that this shows how the estimation becomes more accurate as either n increases, or else σ^2 (which measures the magnitude of the variability in the errors) decreases. In addition, the dependence of h^{-1} quantifies an important effect visible in Figure 1b: As h is decreased, the estimator becomes more wiggly, that is, variable.

Observe that for $\alpha = 1$ (the minimizer of $C(\alpha)$), the two estimators have essentially the same performance, which, as remarked above, is to be expected from the integral mean value theorem, because in this case the x_j are equally spaced. However, in other cases, the variance of $\hat{m}_C$ will be larger. In the extreme case of $\alpha = 0$, note that $\hat{m}_C$ will have 3/2 times the variance of $\hat{m}_E$, which, in view of the above intuition, is also to be expected, because then $\hat{m}_C$ only using 2/3 of the available data (this effect is even worse when $\beta = 1$). Of course, if one really had three replications at each design point (as we have when $\alpha = 0$), the obvious thing to do is to pool, by working with the average of the observations at these points. It is a compelling feature of $\hat{m}_E$ that it makes this adjustment automatically, as $\alpha \to 0$, while $\hat{m}_C$ has a disturbing tendency to delete an observation. These appear to be extreme examples, but, as remarked above, the basic ideas carry over to more natural nonequally spaced examples, such as random designs as discussed in the next section, and the inefficiency of 2/3 appears in an interesting way quite soon.

While it is the variance that quantifies the inefficiency of the convolution weighted estimators, as with any smoothing method attention must also be paid to the bias. In the present example, both methods are, at least asymptotically, the same in the following sense. Under the above technical assumptions, it can also be shown (equations (2.2.4) and (2.2.5) of Chu, 1989) that

$$\begin{aligned}&\operatorname{Bias}(\hat{m}_E(x))\\ (3.3)\qquad &\quad = \int K_h(x-t)(m(t)-m(x))\,dt\\ &\qquad + O(n^{-1}),\end{aligned}$$

$$\begin{aligned}&\operatorname{Bias}(\hat{m}_C(x))\\ (3.4)\qquad &\quad = \int K_h(x-t)(m(t)-m(x))\,dt\\ &\qquad + O(n^{-1}).\end{aligned}$$

Proofs of these are proofs of the equations (2.2.4) and (2.2.5) of Chu (1989). The essential ideas are that the expected values of the estimators give the convolution of K with m, and the integral comes from a Riemann sum approximation.

This example can be generalized in a straightforward fashion, to the case of forming clusters of k points, instead of just three as done above. When this is done, all of the above results remain the same, except the constant $C(\alpha)$ in (3.2) becomes $C_k(\alpha) = 1 + (\alpha - 1)^2(k-2)/2$, $k \geq 2$ and $0 \leq \alpha \leq 1$. Note that the down weighting effect of the convolution weighted estimator can be made arbitrarily bad here, subject of course to the fact that these asymptotics describe only the situation where $nh \gg k$ (by this we intend to convey the intuitive idea that nh is much bigger than k, but it can be mathematically formulated as $\lim_{n\to\infty} nh/k = \infty$).

Both the above example and that represented in Figures 4 and 5 can be criticized on the grounds that they are quite artificial and contrived. However, they have been included here because they illustrate well what typically occurs in the very important case of the random design model M2.

The down weighting property of the convolution estimator has a very strong effect in the random design case. It is clear that, just by chance, some design points will certainly have nearest neighbors closer than the others. The magnitude of this effect on the variance of the convolution weighted estimator is much more than we had expected, in fact being *on average* as bad as in the worst $\alpha = 0$ case of the deliberately pathological example just discussed.

As illustrated in Figure 4, the relative weights assigned to each observation $Y_{(j)}$ by $\hat{m}_C$ are proportional to

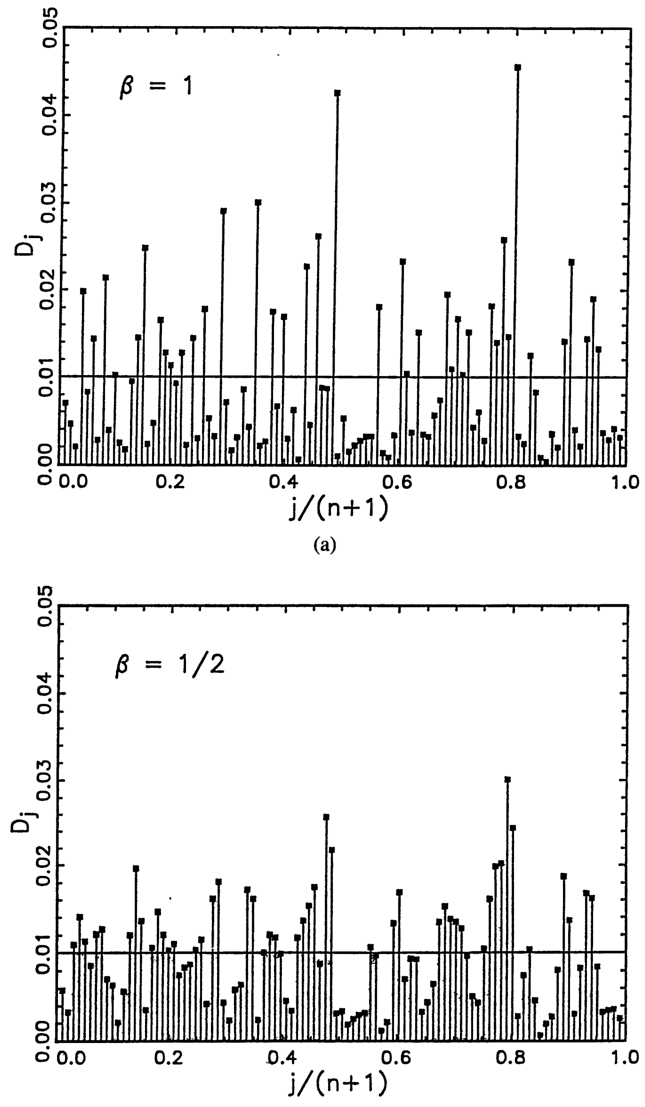
$$\begin{aligned}D_j &= s_j - s_{j-1}\\ &= -\beta X_{(j-1)} - (1-2\beta)X_{(j)} + (1-\beta)X_{(j+1)}.\end{aligned}$$

For an intuitive feeling as to just how much variability there is among the D_j for a typical random sample, consider Figure 7. Here $X_1, \ldots, X_n$ are simulated Uniform[0, 1] random variables. The relative weight given to the observation $Y_{(j)}$ in the construction of $\hat{m}_C$, that is, D_j, is plotted as a function of $j/(n+1)$. Figure 7a shows the case $\beta = 1$ and Figure 7b shows $\beta = 1/2$. Note that, in both cases, the relative weights differ across observations to a surprisingly large degree, with a substantial number of the points significantly down weighted, which means that the convolution weighted estimator is making very inefficient use of the data. At an intuitive level, it seems clear that inefficient use of the data can be expected to give dramatically increased variance of $\hat{m}_C$ with respect to $\hat{m}_E$, whose relative weights are essentially given by the horizontal line. Another way to view this is to consider the shape of surface plots analogous to Figure 6 for these data. For each j, the curve in the variable x will again be a bump centered at $X_{(j)}$, and the height of these bumps at the peak will correspond to the heights shown in Figure 7b. Hence the weights will vary wildly for $\hat{m}_C$, but be nearly constant for $\hat{m}_E$.

Note that the inefficient use of the data appears worse when $\beta = 1$ than when $\beta = 1/2$. This is because there is some slight averaging of consecutive order statistics being done in the latter case, which gives greater stability to D_j. A means of increasing this averaging effect, to give an improved modification of the convolution weighted estimator, is discussed in Section 7.

To quantify the above ideas, we now consider some more asymptotic analysis. In addition to the assumptions (A.1) to (A.3) above, add: (A.4) The marginal density f of X_j has a bounded and continuous first derivative and is bounded above zero, on a neighborhood of x and (A.5) X_j and ε_j are uncorrelated.

Again, these can be weakened; see Chu (1989). Under the assumptions (A.1) to (A.5), it can be shown (see equations (2.3.1) and (2.3.2) of Chu, 1989, for details, and for closely related results see



(a)

(b)

FIG. 7. *Heights of bars represent relative weights (actually gaps between end points of subintervals) given to observations by convolution weighted estimator, for one simulated data set, in uniform random design case. Horizontal line represents corresponding weighted estimator. Figure 7a is for β = 1; Figure 7b is for β = 1/2.*

Collomb, 1981; Jennen-Steinmetz and Gasser, 1987; and Mack and Müller, 1989a),

$$\operatorname{Var}(\hat{m}_E) = n^{-1}h^{-1}f(x)^{-1}\sigma^2\int K^2 + o(n^{-1}h^{-1}), \tag{3.5}$$

$$\operatorname{Var}(\hat{m}_C) = 2(1-\beta+\beta^2)n^{-1}h^{-1}f(x)^{-1}\sigma^2\int K^2 + o(n^{-1}h^{-1}). \tag{3.6}$$

Observe that (3.5) and (3.6) are rather similar in form to (3.1) and (3.2), which is not surprising given the strong intuitive connection between random and fixed designs discussed in Section 2. In fact, n, h, σ^2 and K appear in exactly the same way. An important difference is that f now appears, in a way that makes sense, because at points where $f(x)$ is bigger there will be more observations, and hence less variability. The other important difference is the coefficient $2(1-\beta+\beta^2)$ for $\hat{m}_C$. This quantifies the intuition, as provided by Figures 7a and 7b, about the relative variabilities of the two estimators. In particular, because of the variability of the spacings, the variance of $\hat{m}_C$ is bigger than that of $\hat{m}_E$ by a factor of $2(1-\beta+\beta^2) \geq 3/2$.

Note that, as intuitively expected, the choice $\beta = 1/2$ is optimal in the sense of minimizing the leading coefficient in (3.6). In the other cases, the performance of $\hat{m}_C$ is substantially worse. In particular, the choice of $\beta = 1$, made in Jennen-Steinmetz and Gasser (1988) and Mack and Müller (1989a), seems definitely inadvisable in practice, although it seems clear that this was done in those papers only for technical convenience.

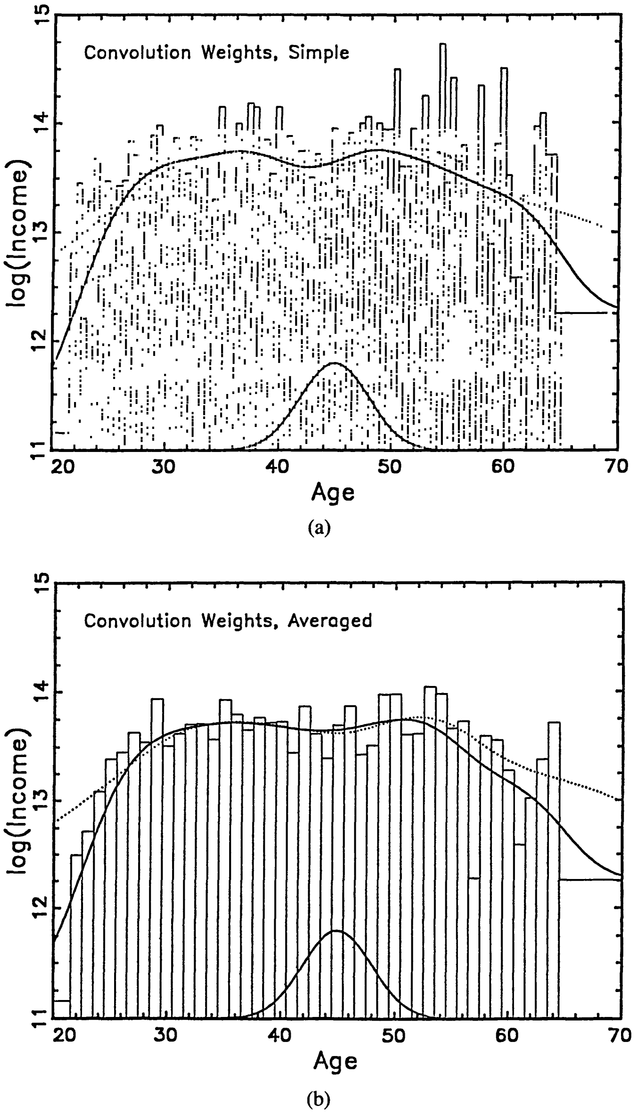
Even with the best possible choice of β, note that $\hat{m}_C$ is still only 2/3 as efficient as $\hat{m}_E$. This and Figure 7 make it clear that the issues raised in the previous examples were not idle pathologies, but indeed the effective deletion of 1/3 of the data is a situation that arises *on the average*, in random sampling. The opinion has been expressed by Mack and Müller (1989a) and by Gasser and Engel (1990) that this is not a large lack of efficiency. The latter authors in particular seem to feel that variability is not a major issue, apparently basing their feelings on the premise that it is always easy to simply gather more data. We are not convinced by this. In particular, we feel that when any reasonable scientist is given a choice between two estimators, one with a given accuracy for 100 observations and another which requires 150 observations for the same accuracy, he will always select the former when all other factors are equal. In real life, data cost money, and hence need to be utilized as efficiently as possible. However, we stress that this is only one side of this issue. While strong reasons are certainly essential to justify discarding 1/3 of the data, in fact such strong reasons do indeed come up in certain important situations, as discussed in Section 4.

There is one other important area where the inefficiency of the convolution estimator can make a real difference in data analysis. This is when there are replications among the ordinates, as one can see in Figure 1a. The reason for the replications is that, while age itself is a continuous variable, the values in this data set have been truncated to the next smallest integer (it is very common to record ages in this way). Such roundings happen quite frequently in real data, especially those in sample surveys. In fact, given that we must always work with only digitized values, there must be at least some rounding during the analysis of any data set. Of course, if there is not much rounding, then the continuous model is very useful and effective. However rounding to the point where replications appear can have a strong effect on $\hat{m}_C$.

An example of this is shown in Figure 8a. This shows what happens when the $\beta = 1/2$ version of the convolution weighted estimator is applied in the simplest possible way to the data of Figure 1a. For each distinct x_j, there are now two bars in the histogram, which represent all of the points having that ordinate. This representation is very poor, because the height of the bar to the left of x_j represents the first such Y_j in the raw data, and the height of the bar to the right of x_j represents the last such Y_j. The remaining Y_j do not appear at all in this picture because for them $s_{j-1} = s_j$. Figure 1a shows that in the present case this amounts to deleting in fact the majority of the data. For this present data set, this deletion of most of the data was not terribly disastrous, although the increased variability of the result (the solid curve in Figure 8a) can be seen in terms of more oscillation around the much more stable $\hat{m}_E$ (the dotted curve in Figure 8b). Since the estimator is in effect randomly choosing only two of the observations, sometimes the result is above average, and sometimes below. This estimator also goes down too sharply at the ends, but this is a boundary effect that looks especially bad for these data because the final observations on each end happen to be unusually low. This edge effect is not an important issue, because it can be substantially mitigated, using for example the methods discussed in Section 4.3 of Müller (1988). Again this is not done here, because the best of these is rather complicated to describe and implement, and off our main points.

When there are replications among the design points, it is clearly inappropriate to use the Section 2 definition of $\hat{m}_C$ in this simple-minded manner. We do this to make the point that great care needs to be taken with this estimator. In particular, we feel it would be a fundamental mistake to design any software package implementations of this estimator without this effect firmly in mind.

It is intuitively clear that, when there are replications among the x_j, one should pool observations, by replacing the values with the same x_j by a single point representing their average. However, the ideas discussed above, concerning relative weights of observations, can still be seen to apply even when this is done, in Figure 8b. In that picture, except at the boundaries, which again are not discussed here, the estimator $\hat{m}_C$ is now closer to $\hat{m}_E$. However, there are some differences which do make this point. Note that near age 46, $\hat{m}_C$ is higher than $\hat{m}_E$. We feel this to be inappropriate, and only due to the arbitrary and unnatural way that the convolution estimator chooses weights. In particular, a look at the scatterplot in Figure 1a shows only two observations for each of the ages 46 and 49. Furthermore, all of these are higher than usual. The evaluation weighted estimator is not seriously affected by this, but the convolution esti-



(a)

(b)

FIG. 8. *Convolution weighted estimators applied to data in Figure 1a, using same kernel and window width as there. Figure 8a shows the straightforward implementation, as defined in Section 2, that is, as in Figure 3. Figure 8b shows the estimator applied to the pooled data where replications are replaced by the average over points having the same ordinate.*

mator is, because it gives each of these higher values the same weight as the much more representative values at other points that represent averages of more values. In the terminology of robustness, these are "leverage points" for $\hat{m}_C$. However they do not have such "high influence" on $\hat{m}_E$, so it is much more robust in this sense. A visually more distressing occurrence of this same phenomenon occurs for ages between 50 and 60, where $\hat{m}_C$ is now quite a bit too low. This is caused by the low single values at 57 and 61. The histogram in Figure 8b shows visually how these values pull down $\hat{m}_C$, compared to $\hat{m}_E$, which is affected only to the extent seen visually in the scatter plot of Figure 1. We feel it is an important property of $\hat{m}_E$ that it behaves more like the eye in scatter plot smoothing.

4. BIAS ISSUES

Variance, which was the main theme underlying the difficulties illustrated in the previous section, comprises only half of the smoothing problem. For a balanced assessment of the situation, one also needs to consider the bias.

A simple, but illustrative, example of how the down weighting effect of the convolution estimator can be very beneficial is given in Figure 9, which shows the performance of $\hat{m}_E$ on the same artificial data as used for $\hat{m}_C$ in Figure 3. In both of these figures, the observations lie on a straight line. Note that away from the boundaries (as discussed above this is the only relevant area for the points being discussed here), $\hat{m}_C$ runs nicely through the data as one would hope. On the other hand, $\hat{m}_E$ is quite disturbing because it lies always below the data. This is caused by the fact that $\hat{m}_E$ uses the data in a symmetric fashion, but they are highly asymmetric, so the greater density of lower observations on the left side tends to pull down each local average. On the other hand, because $\hat{m}_c$ assigns weights as shown in Figure 3, each observation is weighted exactly as needed to cancel this effect. This occurrence is not a special artifact of this example, but in fact happens quite generally. It is straightforward to make this effect appear much stronger, but such examples require more data points, which were not added here because they would tend to obscure the other purpose of Figure 3. This is why the efficiency issues of the previous section are not a sufficient basis for sensible choice between $\hat{m}_E$ and $\hat{m}_C$.

Another way of looking at this is in the surface plots of Figure 10. Note that, except for the boundary effect at $j = 1$, the convolution weights fall off faster for smaller j than the evaluation weights. The convolution estimator also puts more weight on the bigger values for x large. These effects

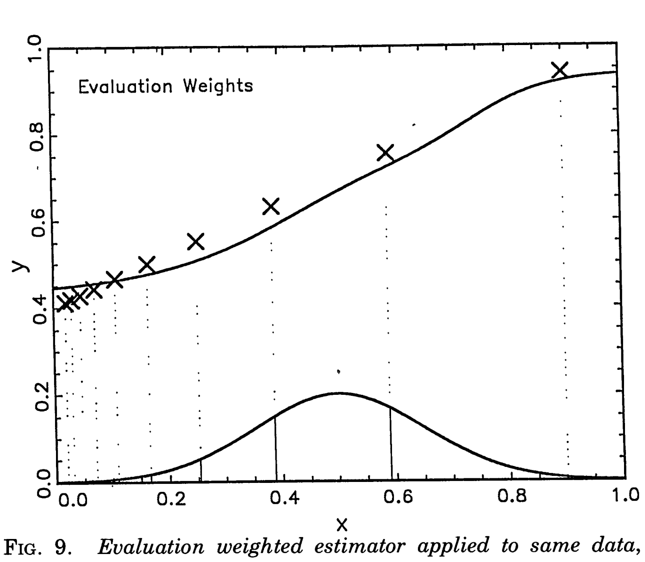
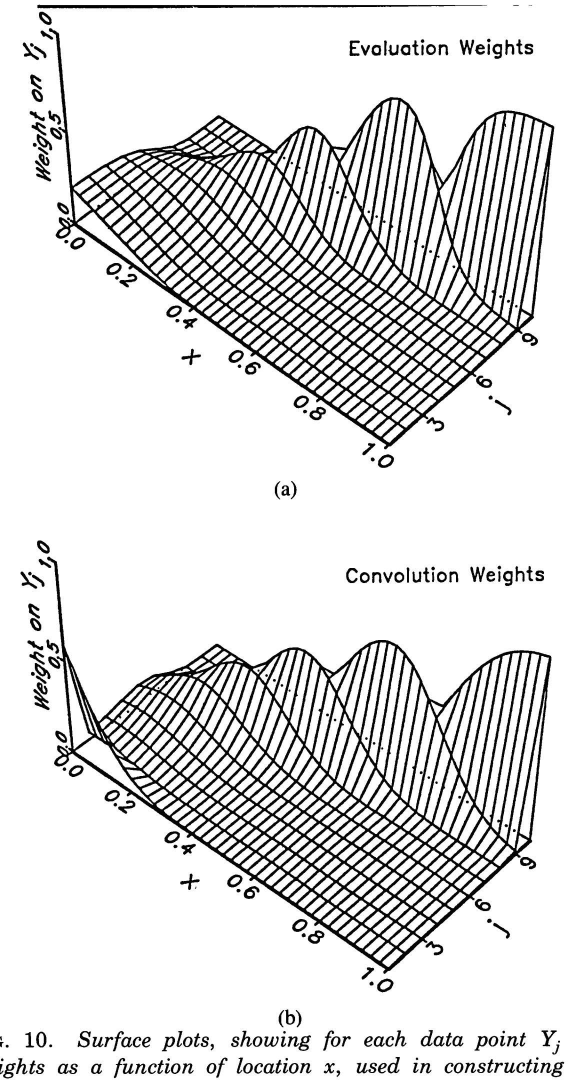


FIG. 9. *Evaluation weighted estimator applied to same data, and also using same kernel and window width, as in Figure 3.*



(a)

(b)

FIG. 10. *Surface plots, showing for each data point Y_j the weights as a function of location x, used in constructing the kernel smoothers in: (a) Figure 3; (b) Figure 8.*

cause $\hat{m}_C$ in Figure 3 to run through the data points, while $\hat{m}_E$ in Figure 9 is too low.

Once again asymptotic analysis is useful for simple and intuitive quantification of these ideas. It can be shown that for bias considerations, the random design case is essentially the same as for fixed designs satisfying the asymptotic property $x_i = G^{-1}(i/n) + o(n^{-1})$ (for some cdf G), so we explicitly treat only the former. Under the assumptions (A.1) to (A.5), one may also show (same references as at (3.5) and (3.6) above)

$$\begin{aligned} &\operatorname{Bias}(\hat{m}_E) \\ (4.1)\qquad &= \frac{\int K_h(x-t)(m(t)-m(x))f(t)\,dt}{\int K_h(x-t)f(t)\,dt} \\ &\qquad + O(n^{-1/2}h^{1/2}), \end{aligned}$$

$$\begin{aligned} &\operatorname{Bias}(\hat{m}_C) \\ (4.2)\qquad &= \int K_h(x-t)(m(t)-m(x))\,dt \\ &\qquad + O(n^{-1}). \end{aligned}$$

Note that, in the important special case of a uniform design, that is, f is constant on some interval, these two expressions are the same. However, when the design is nonuniform, the bias is much more simple for $\hat{m}_C$. This yields benefits in two forms. Gasser and Engel (1990) point out that this gives a large advantage in terms of interpretability. In particular, it is much easier to explain to a nonexpert (especially one who is not very mathematically inclined) how bias is entering for $\hat{m}_C$. The other advantage comes in terms of when the estimator will be unbiased (at least to the level in (4.1) and (4.2)). Mack and Müller (1989a) point out that $\hat{m}_C$ is unbiased in the case that m is linear (and $\hat{m}_E$ is generally not, except in the important case of m constant). This is especially important because the linear case comes up when nonparametric regression is being used to address such questions as: Is the regression function linear or not? Of course, combinations of f and m can be found that make $\hat{m}_E$ unbiased when $\hat{m}_C$ is not, but they do not include this important special case.

To gain an intuitive feeling for these issues, consider Figure 11. The idea for this was given to us in private conversation by Hans-Georg Müller. It shows the bias of $\hat{m}_E$, as given in (4.1) in the case where m is linear, the kernel K is Gaussian and f is taken to be: (a) the standard normal(0, 1) density, (b) the standard exponential(1) density, (c) the piece wise parabolic density, $f(x) = (3/4)(1 - x^2) \cdot 1_{[-1,1]}(x)$, and (d) the mixture density, $0.5N(2,1) + 0.5N(-2,1)$. In each part, $m(x)$ is the diagonal line, $f(x)$ is the solid curve and the dashed curve represents $E[\hat{m}_E(x)]$ by the sum of the dominant part of the right side of (4.1) and $m(x)$. Feeling for the bandwidth and the effective amount of smoothing being done is given by the dotted curve at the bottom of each part, which is a vertical rescaling of K_h.

Observe that, for all but the normal mixture, there is some curvature, which can be quite disturbing when one is trying to decide if the regression is linear or not (recall the actual estimate has a distribution centered around this curve). However, for the exponential, this curvature appears only near the edges, and is entirely due to boundary effects caused by the fact that $f(x)$ is taken to be 0 outside the interval shown. In the important special cases of the normal and the exponential (away from the boundaries), while there is definitely a good deal of bias present, the dashed curve is still essentially linear. The reason for this will be discussed later in this section, but at this point observe that in these cases bias will not affect one's visual impression as to whether or not the regression is linear. Observe that the magnitude of all these effects is an increasing function of h. We chose these h's because speaking visually they are among the largest we have worked with.

Certainly the curvatures in Figures 11c and 11d are of major importance and cannot be ignored. We were pleased that Gasser and Engel (1990) were sufficiently impressed with a version of Figure 11d, appearing in an earlier version of this paper, that they used it as their Figure 1. See their Figure 2 for another variation on this idea.

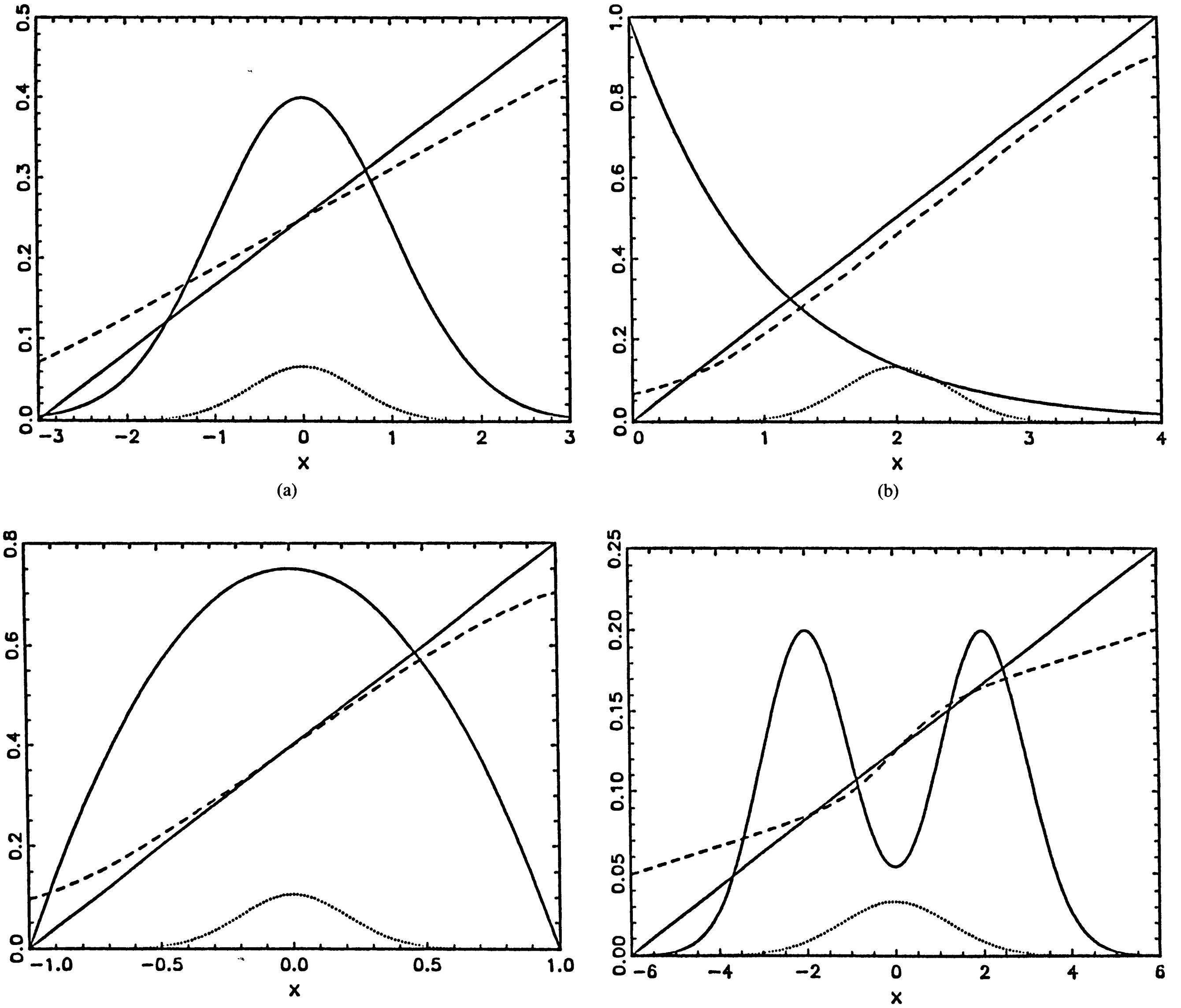
While the bias of $\hat{m}_C$ is certainly more simple, it is not clear that it is "more natural." A case can be made for the bias of $\hat{m}_E$ being the more natural one. In particular, when the data are being used efficiently (recall from Section 3 that $\hat{m}_C$ does not do this), the design density $f(x)$ is an important entity, which intuitively should affect how well one can estimate $m(x)$. The fact that it nearly disappears in the analysis of $\hat{m}_C$ can thus be considered to be an unattractive feature of that estimator. An interpretation of this is that, in order to make the design density $f(x)$ essentially disappear from the bias, one must pay some price, which in this case is increased variance, as quantified in the last section.

Note that by (A.1)-(A.5) and by Taylor's theorem, (4.1) and (4.2) admit the further expansions:

$$\begin{aligned} &\operatorname{Bias}(\hat{m}_E) \\ (4.3)\qquad &= h^2(m''f + 2m'f')\left(\int u^2K\right)/(2f) \\ &\qquad + O(n^{-1/2}h^{1/2}) + o(h^2), \end{aligned}$$

$$\begin{aligned} (4.4)\qquad \operatorname{Bias}(\hat{m}_C) &= h^2 m''\left(\int u^2K\right)/2 \\ &\qquad + O(n^{-1}) + o(h^2), \end{aligned}$$

where m', m'' denote first and second derivatives of



(a) (b)

(c) (d)

FIG. 11. *Bias in evaluation weighted estimator. Diagonal line is $m(x)$ and also essentially $E\hat{m}_C(x)$; dashed curve is essentially $E\hat{m}_E(x)$, dotted curve is a vertical rescaling of the kernel function $K_h(x)$ and solid curve is $f(x)$, which is (a) standard normal, (b) exponential, (c) piecewise parabola and (d) normal mixture.*

$m(x)$, etc. These representations again show the usefulness of asymptotic analysis. In particular, the simple idea that, when m has more curvature, it is harder to estimate (because nearby observations contain less information about $m(x)$) is nicely quantified by (4.4), which measures this effect in terms of the "curvature," $m''(x)$. When $m(x)$ is curved upward, $\hat{m}_C$ will be too big, and vice versa for $m''(x) < 0$. Furthermore, this reflects the point made intuitively in Figure 1b that bias effects are worse when the window width h is large.

It is unfortunate that, for comparison purposes, (4.3) and (4.4) are not comparable: For some choices of m, f and x one will be bigger in magnitude (depending on the signs of $m''(x)f(x)$ and $m'(x)f'(x)$), while for other choices the other will be. It would be nice to find some way to resolve this completely, say by some finding some average sense in which one bias is bigger than the other, but we do not see how to do this. The next section discusses the relative effects of variance and bias on the mean squared error.

The representation (4.3) provides considerable insight into the effects observes in Figures 11a–d. In particular, the biases exhibited in Figures 11a and b are so surprisingly close to linear because, when m is linear, (4.3) shows that $E[\hat{m}_E(x)]$ is roughly proportional to

$$(\text{slope of } m)\, f'(x)/f(x).$$

In the normal and exponential cases, these functions are linear (and essentially only in those cases by an elementary differential equation argument). The curvatures in Figures 11c and d are also easily understood by this method. For example, observe that the normal mixture bias is linear in those regions where one peak or the other is dominant, and curved in between.

5. MEAN SQUARE ERROR

In Section 3, it was seen that the variance of $\hat{m}_E(x)$ is substantially better than for $\hat{m}_C(x)$. However, Section 4 discusses several bias related reasons why one may be prepared to pay the price of increased variance entailed by use of $\hat{m}_C(x)$. In private conversation, Jeff Hart has pointed out that one should use "sensitivity analysis" ideas, developed at the end of Section 3 of Scott (1979) and in Corollary 2.1 of Scott (1985), to properly account for the relative effects of variance and bias.

This is done as follows. From (3.5), (3.6), (4.3) and (4.4), it is clear that under the assumptions (A.1)–(A.5), for $\hat{m}(x)$ representing either $\hat{m}_E(x)$ or $\hat{m}_C(x)$, the mean square error is

$$\text{(5.1)} \quad \begin{aligned} \mathrm{MSE}(\hat{m}(x)) &= E[\hat{m}(x) - m(x)]^2 \\ &\sim vn^{-1}h^{-1} + b^2h^4, \end{aligned}$$

where $\sim$ means the ratio tends to one in the limit and where the specific values taken on v and b are easily seen from (3.5), (3.6), (4.3) and (4.4). Simple calculus shows that the right hand side of (5.1) is minimized by $h_{\mathrm{AOPT}} = (v/4b^2n)^{1/5}$, from which it follows that

$$\text{(5.2)} \quad MSE(h_{AOPT}) \sim 5 \cdot 4^{-4/5} v^{4/5} b^{2/5} n^{-4/5}.$$

Hence, for reasonable values of the bandwidth (there are a number of ways to ensure that the bandwidth is asymptotically the same as h_{AOPT}; see Marron (1988) for access to the bandwidth selection literature), the effect of the variance on the MSE is much stronger than that of the bias. In particular, the factors of either 2 or 1.5 (which come up when comparing variances as in (3.5) and (3.6)), should really be squared when comparing with the various derivatives involved in (4.3) and (4.4). We feel that this, together with the fact that $\hat{m}_E(x)$ is clearly superior when the design density is uniform, is a weak indication that, in many situations where MSE is considered to be most important, $\hat{m}_E(x)$ may turn out to be marginally better. However, it must be kept in mind that this is only personal opinion, and in fact the estimators are not generally comparable in this sense. In the nonuniform design case, the other considerations pointed up in Section 5 can easily outweigh MSE considerations in the choice of an estimator.

Another means of attempting to assess the relative performance of these estimators is through formulation of minimax results. One result of this type may be found in Gasser and Engel (1990), who prove a theorem, of which the main intuitive content is that, if one first fixes the regression function m, then the worst case over a class of f's is worse for $\hat{m}_E$ than for $\hat{m}_C$. They go on to assert that they feel this provides strong motivation for general choice of the convolution estimator. We are unconvinced by this for several reasons. One is that their result is dependent on assuming the design density f is bounded below. This does not seem reasonable, for example in observational studies where the X_j could easily be normally distributed. Their theorem falls apart when this assumption is deleted. Hence we are left with the same conclusion given above: The estimators are not comparable.

A second reason is that their result is poorly formulated, because information containing important intuitive content is buried away in the case "$\infty = \infty$." If this case is analyzed properly, using *ratios* of the given quantities, then a different picture appears. In particular, note that, for any h, in the case of $m(x)$ constant, letting F denote the class given in Gasser and Engel (1990) (despite the inappropriateness of this class argued above), and defining IMSE to be the integrated (over x) MSE as there,

$$\sup_{f \in F} \frac{\mathrm{IMSE}(\hat{m}_C, h)}{\mathrm{IMSE}(\hat{m}_E, h)} = \frac{2}{3},$$

for any bandwidth h. Now taking the minimum over h, as done in the comparison of Gasser and Engel, reveals that in fact this ratio can be either bigger or smaller than 1, depending on the curve $m(x)$. This provides a second way to see that these estimators should not be considered comparable, even in this special sense.

A third reason is that, for honest and relevant minimax comparison, suprema should be taken over *both* f and also m.

We feel a much more relevant and unbiased minimax comparison is made through the study of

$$\sup_{f,m} \frac{\mathrm{IMSE}(\hat{m}_E, h_{\mathrm{IMSE}})}{\mathrm{IMSE}(\hat{m}_C, h_{\mathrm{IMSE}})}$$

and of

$$\inf_{f,m} \frac{\mathrm{IMSE}(\hat{m}_E, h_{\mathrm{IMSE}})}{\mathrm{IMSE}(\hat{m}_l C, h_{\mathrm{IMSE}})},$$

where the supremum and infimum are taken over some suitable class, and where h_{IMSE} is the bandwidth to minimize IMSE. It is straightforward to use calculations along the lines in Gasser and Engel (1990) to check that the former is ∞ and the latter is 0. Once again we arrive at the same conclusion derived intuitively in Section 4: These estimators are not comparable in any reasonable minimax sense.

6. IMPROVED ESTIMATORS

The results of Section 3 show clearly that the increased variance of $\hat{m}_C$ in the random design case is caused by the instability of the s_j. A means of reducing this instability is to average together more order statistics in the definition of the s_j. One means of doing this is, given a nonnegative integer k, to define

$$s_j = \sum_{i=1}^{2k+2} \gamma_i X_{(j-k+i-1)},$$

for $j = k + 1, \dots, n - k - 1$, where the weights γ_i are nonnegative and satisfy $\sum_{i=1}^{2k+2} \gamma_i = 1$, and where the values near the boundaries $s_0, \dots, s_k$ and $s_{n-k}, \dots, s_n$ are defined in any reasonable manner (again the exact definition not affecting our main ideas). Note that (3.5) is the special case $k = 0$. Once again using the assumptions (A.1)–(A.5), equations (2.4.2) and (2.4.3) of Chu (1989) show that

$$\begin{aligned}\operatorname{Var}(\hat{\mathrm{m}}_I)& \\ (6.1) \quad &= \left(1 + \sum_{1=1}^{2k+2} \gamma_i^2\right) n^{-1} h^{-1} \sigma^2 f(x)^{-1} \int K^2 \\ &\quad + o(n^{-1}h^{-1}),\end{aligned}$$

$$\begin{aligned}\operatorname{Bias}(\hat{\mathrm{m}}_I)& \\ (6.2) \quad &= \int K_h(x-t)(m(t) - m(x))\, dt \\ &\quad + O(n^{-1}).\end{aligned}$$

It is easy to see that the best choice of the γ_i is $\gamma_i = 1/(2k+2)$, $i = 1, \dots, 2k+2$. In this case, $\operatorname{Var}(\hat{m}_I) = O(n^{-1}h^{-1}(1 + (2k+2)^{-1}))$, so the amount by which the variance of $\hat{m}_E$ improves over $\hat{m}_I$ can be made arbitrarily small in the limit, by taking k large. Of course a practical limitation is that these asymptotics are only meaningful when $nh \gg k$.

Deeper analysis of this estimator and several obvious modifications of it go beyond the intent of this paper, but provide interesting topics for future research.

Another means of combining the best properties of both estimators has been found by Fan (1990), who shows that this happens for the old idea of replacing local averages by local linear fits (i.e., instead of taking local advantages, doing local weighted linear least squares). It will be interesting to see how this variation fits into the ideas of this paper.

7. OTHER ISSUES

There are other aspects to the problem of choice between $\hat{m}_E$ and $\hat{m}_I$ that can sometimes be important. We have not highlighted these in the above discussion, because they all pertain to modifications of the very basic nonparametric regression settings considered here.

In Härdle (1990), it is pointed out that one of these is the extension to the case where the real valued X becomes a d-dimensional vector. One can still use kernel estimators to estimate the d-dimensional regression function, and many of the same lessons still apply. For $\hat{m}_E$ there are appropriate analogs of (3.5), (4.1) and (4.3). However, the situation becomes more difficult for $\hat{m}_I$. In particular, the negligible error in (4.2) rapidly becomes dominant. For example, using Theorem 6.1 of Müller (1988), observe that, in our setting (Müller's $\nu = 0$, $k = 2$, $m = d$), this breakdown occurs at $d = 4$. It is an interesting open problem to find an adaptation of $\hat{m}_C$ that shares its nice bias properties and technical tractability, without having this high dimensional breakdown problem.

Mack and Müller (1989a) point out that $\hat{m}_E$ is much harder to work with for the estimation of derivatives. This is because its derivatives take on a very complicated form because of the quotient structure. The result suffers both in being messy to analyze, and also in losing insight and interpretability.

Jeff Hart, in private correspondence, and also Mack and Müller (1989a) point out that for proper adjustment for boundary effects (see Rice, 1984; Gasser, Müller and Mammitzsch, 1985), the form of $\hat{m}_C$ is again far more convenient.

ACKNOWLEDGMENTS

C.-K. Chu was partially supported by Grant NSC80-0208-M007-32. J. S. Marron was partially supported by NSF Grant DMS-87-01201.

REFERENCES

Cheng, P. E. (1991). Applications of kernel regression estimation: A survey. *Comm. Statist. Theory Methods* **19** 4103–4134.

Chu, C.-K. (1989). Some results in nonparametric regression. Ph.D. dissertation, Univ. North Carolina, Chapel Hill.

CHU, C.-K. and MARRON, J. S. (1988). Comparison of kernel regression estimators. North Carolina Inst. Statistics, Mimeo Series 1754.

CLARK, R. M. (1977). Nonparametric estimation of a smooth regression function. *J. Roy. Statist. Soc. Ser. B* **39** 107-113.

CLARK, R. M. (1980). Calibration, cross-validation and carbon-14, II. *J. Roy. Statist. Soc. Ser. A* **143** 177-194.

COLLOMB, G. (1981). Estimation non-paramétrique de la regression: Revue bibliographique. *Internat. Statist. Rev.* **49** 75-93.

EUBANK, R. A. (1988). *Spline Smoothing and Nonparametric Regression.* North-Holland, Amsterdam.

FAN, J. Q. (1990). A remedy to regression estimators and nonparametric minimax efficiency. North Carolina Inst. Statist, Mimeo Series 2028.

GASSER, T. and ENGEL, J. (1990). The choice of weights in kernel regression estimation. *Biometrika* **77** 377-381.

GASSER, T. and MÜLLER, H. G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23-68. Springer, New York.

GASSER, T., MÜLLER, H. G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238-252.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press.

HÄRDLE, W. and MARRON, J. S. (1983). The nonexistence of moments of some kernel regression estimators. North Carolina Inst. Statistics, Mimeo Series No. 1537.

JENNEN-STEINMETZ, C. and GASSER, T. (1988). A unifying approach to nonparametric regression estimation. *J. Amer. Statist. Assoc.* **83** 1084-1089.

MACK, Y. P. and MÜLLER, H. G. (1989a). Convolution type estimators for nonparametric regression estimation. *Statist. Probab. Lett.* **7** 229-239.

MARRON, J. S. (1988). Automatic smoothing parameter selection: A survey. *Empirical Economics* **13** 187-208.

MÜLLER, H. G. (1988). *Nonparametric Analysis of Longitudinal Data. Lecture Notes in Statist.* **46**. Springer, New York.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141-142.

RICE, J. (1984). Boundary modifications for kernel regression. *Comm. Statist. Theory Methods* **13** 893-900.

SCOTT, D. W. (1979). An optimal and data-based histograms. *Biometrika* **66** 605-610.

SCOTT, D. W. (1985). Frequency polygons: Theory and application. *J. Amer. Statist. Assoc.* **80** 348-354.

ULLAH, A. (1985). Specification analysis of econometric models. *Journal of Quantitative Economics* **2** 187-209.

WAHBA, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359-372.

Comment

Theo Gasser, Christine Jennen-Steinmetz and Joachim Engel

Nonparametric curve estimation is coming of age, and it is thus timely to study the merits of various approaches. Two weighing schemes have been proposed in the kernel estimation literature, called "evaluation weights" and "convolution weights" by Chu and Marron. The goal of their paper is to give a balanced discussion of their merits, based on two complementary philosophies P1 and P2. We feel that the paper falls short of presenting a balanced discussion and often disregards philosophy P1, that is, looking for structure in a set of numbers. For many years the evaluation weights (due to Nadaraya and Watson) have been studied primarily for random design, the convolution weights for fixed design. Random design is defined and treated adequately by the authors, while fixed design is represented by rather peculiar examples (see below). As is common (see, e.g., Silverman, 1984), we define a regular fixed design as $x_i = F^{-1}((i - 0.5)/n)$, $f = F'$, where F is some distribution function with density f. Under standard assumptions, the asymptotic bias and variance for the two weighting schemes are as in Table 1, where $M_2(K) = \int u^2K(u)\,du$ and $V(K) = \int K(u)^2\,du$.

VARIANCE

The factor C in the variance of the convolution estimator is 1 for fixed and 1.5 for random design. Thus, we have an increase in variance for convolution weights with respect to the random design only; variances are asymptotically identical for regular fixed design. There is one fixed but not regular design of importance, that is, when we have multiple points, for example, due to rounding. It is easy to modify convolution weights for this design appropriately, and this has been done in our programs.

We are puzzled by the frequent use of the word *efficiency* in Section 3, when in fact only variance is

Theo Gasser is a Professor, and Christine Jennen-Steinmetz is a Ph.D., Department of Biostatistics, Zentralinstitute für Seelische Gesundheit, 68 Mannheim 1, Germany. Joachim Engel is a Research Fellow, University of Heidelberg, Sonderforschungsberichte 123, 69 Heidelberg, Germany.