## REFERENCES


AHMAD, I. A. and LIN, P.-E. (1984). Fitting a multiple regression function. *J. Statist. Plann. Inference* **9** 163–176.

BECKER, R., CHAMBERS, J. and WILKS, A. (1984). *The New S Language.* Wadsworth, Pacific Grove, CA.

BELLMAN, R. E. (1961). *Adaptive Control Systems.* Princeton Univ. Press.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555.

BUTA, R. (1987). The structure and dynamics of ringed galaxies. III. Surface photometry and kinematics of the ringed nonbarred spiral NGC7531. *Astrophys. J. Supplement Ser.* **64** 1–37.

CHU, C.-K. and MARRON, J. S. (1991). Choosing a kernel regression estimator (with discussion). *Statist. Sci.* **6** 404–436.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

CLEVELAND, W. S. and DEVLIN, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 596–610.

CLEVELAND, W. S., GROSSE, E. and SHYU, W. M. (1991). Local regression models. In *Statistical Models in S* (J. Chambers and T. Hastie, eds.) 309–376. Wadsworth, Pacific Grove, CA.

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998–1004.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21** 196–216.

FRIEDMAN, J. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

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

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

GASSER, TH. and MÜLLER, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171–185.

HALL, P. and WEHRLY, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimates. *J. Amer. Statist. Assoc.* **86** 665–672.

HASTIE T. and TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

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

MÜLLER, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231–238.

MÜLLER, H.-G. (1988). Nonparametric regression analysis of longitudinal data. *Lecture Notes in Statist.* **46** Springer, Berlin.

MÜLLER, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika* **78** 521–530.

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

PRIESTLEY, M. B. and CHAO, M. T. (1972). Nonparametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 384–392.

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

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–620.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

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


# Comment

**J. Fan and J. S. Marron**

### 1. GENERAL COMMENTS

We would like to thank the authors for a useful and informative article on the state of the art in nonparametric regression. Especially enjoyable were the novel and imaginative graphical methods that were developed to illustrate the points being made. These reveal more intuition behind the theoretical results of Stone (1977, 1982) and Fan (1992, 1993). It contains a nice summary of many points which have already been made and justified (theoretically and intuitively) by the recent papers of Chu and Marron (1991) and the discussions therein and of Fan (1992, 1993).

The main contribution of the paper is a very accessible introduction to a point which is becoming quite clear to insiders in the field of nonparametric regression: local (i.e., moving window) polynomial regression estimators have a number of compelling advantages over the more widely used and studied kernel estimators.

In view of the very large literature on kernel regression estimators, an interesting issue is why it took so long for the smoothing community at large to understand fully the benefits of local polynomials. We speculate that this was because of "equivalence results," the best known being Müller (1987) but see also Lejeune (1985), whose main intuitive message was for *equally*

*J. Fan is Assistant Professor and J. S. Marron is Associate Professor, Department of Statistics, University of North Carolina, Chapel Hill, North Carolina 27599-3260.*

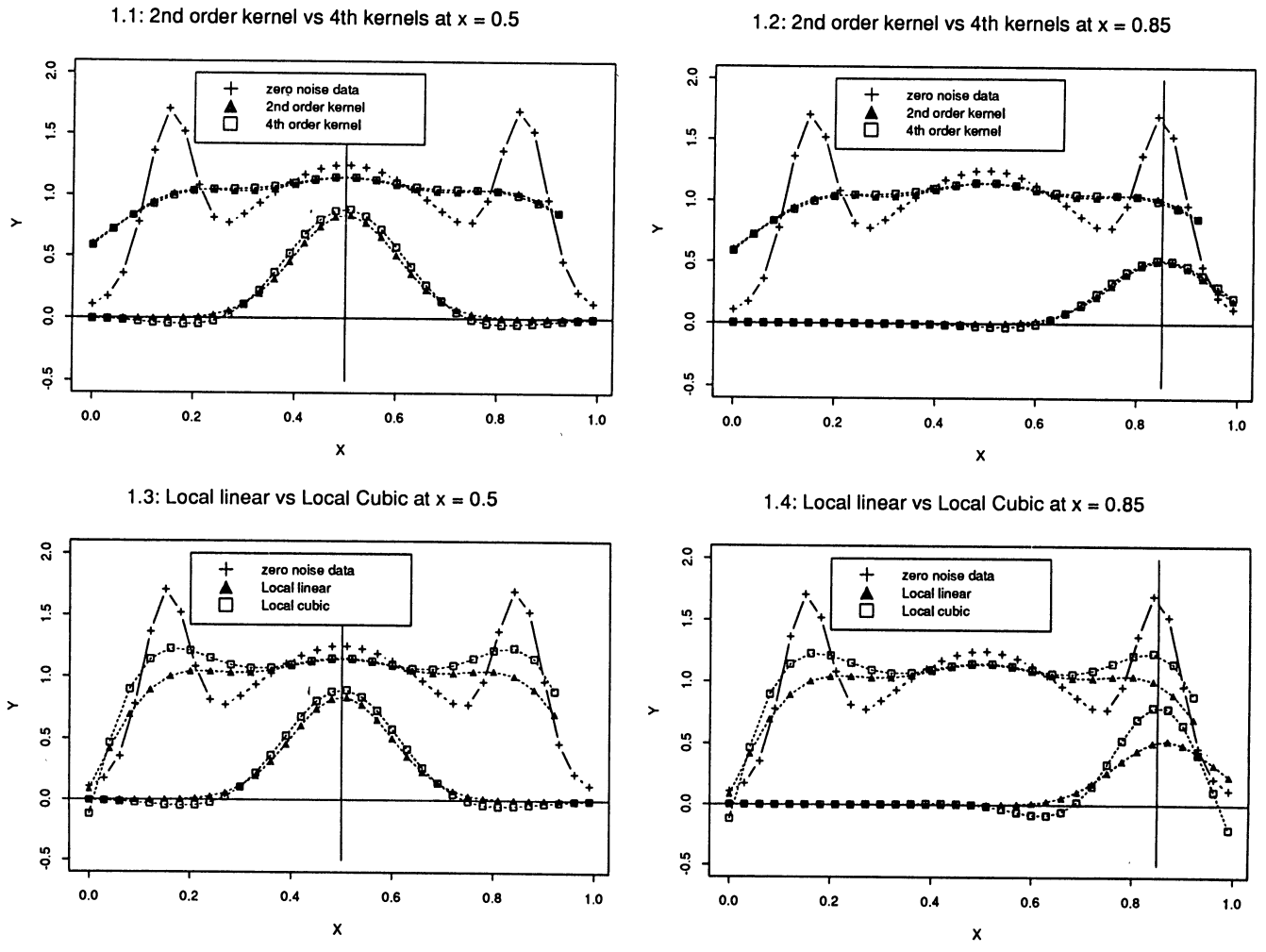


FIG. 1. *Comparison of high- and low-degree methods. Equally spaced design, with regression curve demonstrating failure of higher-order methods. Effective weights shown at the bottom of each, for estimation at* $x = 0.5$ *in figures* 1.1 *and* 1.3 *and at* $x = 0.85$ *in Figures* 1.2 *and* 1.4.

*spaced* $x_j$, when studying behavior *away from the boundaries,* there is essentially no difference between local polynomial and traditional kernel estimators (with a nice correspondence between degree of polynomial and kernel order). Since many people's intuition seemed to indicate that there were only "technical" differences between the lessons learned there and the more general case, local polynomials did not receive the widespread attention they deserved. A key to the recent renaissance was the observation by Fan (1993) that in the random design case neither the Nadaraya–Watson (NW) nor the Gasser–Müller (GM) could attain the minimax lower bound but that the local linear could (with the correct choice of kernel function).

## 2. BIAS ADJUSTMENTS

About higher-order kernels, considerably more is known than is indicated in this paper. In particular Marron and Wand (1992) have shown, in the closely related density estimation setting, that higher-order kernels are rarely worth the loss in interpretability inherent to a local average which uses negative weights. The reasons behind this are made clear in a visual sense in Marron (1992). The main idea is indicated in Figure 1. Figures 1.1 and 1.2 show NW estimators, using the normal kernel $\varphi(x)$ with a bandwidth $h_2 = 0.1$ and the fourth-order kernel $(3 - x^2)\varphi(x)/2$ with a bandwidth $h_4 = \sqrt{2}h_2$, together with effective weight functions. Figures 1.1 and 1.3 demonstrate effective weights for estimation at the interior point $x = 0.5$. Figure 1.1 shows clearly why the second-order kernel estimate is biased for estimation at $x = 0.5$: data points roughly in the intervals [0.3, 0.4] and [0.6, 0.7] are significantly too low. The higher-order kernel attempts to adjust for this by adding more observations to the local average, whose means are even lower, but applying negative weights to cancel the bias effect. If the underlying regression had "constant curvature" including the areas where the higher-order kernel puts negative weight,

then there would be a big reduction in the bias. But in this case, the curvature is not constant where the weights are negative, and the sharp peaks in that area actually make the bias slightly worse.

Figures 1.3 and 1.4 show local linear and cubic fits and effective weights, using the normal kernel and the bandwidths $h_2$ for the linear and $h_4$ for the cubic. It is evident that in this case the gain of the higher-order fit is again very limited, for the same reason. The lessons of the usual asymptotic theory of bias gains for higher-order methods are not effective in this situation. The reason is that the underlying regression does not have constant curvature on large enough neighborhoods. Figures 1.2 and 1.4 show a similar effect at the boundary $x = 0.85$, except that the local cubic fit has a smaller bias because of the very sharp decrease of the second derivative of the true regression function. The benefit of the local cubic is less when the data are noisy – the local cubic will have a big increase in the variance.

While we heartily endorse the statement in the last paragraph of Section 2 that local polynomials provide an attractively simple and intuitive way to correct biases, we find the remarks in the previous paragraph to be less on target. In particular, we do not feel it is useful to view choice of order of the polynomial (and the same applies for choice of kernel order) as simply a bias–variance trade-off. This is because that viewpoint holds up only when the bandwidth is held fixed, which is inappropriate when one uses a method with, for example, less bias and more variance. When comparing across orders, effective performance (measured either intuitively or else asymptotically) depends on using a sensible (and hence different) bandwidth for each order.

Choice of order is an interesting open problem, of deep interest from both practical and theoretical viewpoints. While higher orders offer advantages in terms of bias reduction, there is a price to be paid in terms of increase of variance [see Fan and Gijbels (1992b) for quantification of this], computational speed, ease of implementation, and interpretability (analogous, although not so severe as for higher-order kernels). We suspect that the final resolution will be a recognition that there exist different situations where each of 0, 1, 2, 3 could be preferable (but we doubt there is much call for more than 3). In particular, orders 0 and 1 are preferable when there is a premium on speed and/or interpretability, while orders 2 or 3 can possibly yield significant gains for larger samples (assuming bandwidths for each order are chosen properly). The choice between even and odd (i.e., 0 vs. 1 and 2 vs. 3) depends on the setting. Odd orders are usually preferable at the boundaries, and in the interior they have the same variance as the corresponding even order, but they have potential for very substantial bias reduction. But for interior points in an equally spaced design, odd orders mean additional complication, with no benefits. See Section 2.3 of Fan and Gijbels (1992b) for discussion. We believe that serious statisticians will eventually want all four in their toolbox (and good software packages will provide this). Despite the noted limitations of higher-order fits (see Figures 1.1–1.4), considerable gains can still be made by using variable order approximation where the local linear is used at a sloped region and the local cubic is used at peaks and valleys. An adaptive procedure for this has recently been proposed by Fan and Gijbels (1992b).

## 3. BOUNDARY EFFECTS

There is an intuitive way of understanding the trade-off at boundaries that can arise between the NW estimator and the local linear. Note that if the underlying regression is relatively "steep" near the boundary, for example, as in Figures 3 and 5 in the original paper, then the bias problems of the NW method make it clearly inferior. However, if the true regression is relatively "flat," the gain from fitting a local line is less, and the greater variability can drown it out. More precisely, an increase of variance of a factor ranging from 1 (no increase) to 4 has to be paid when a point runs from the interior (no increase) to the boundary [see Figures 1–3 of Fan and Gijbels (1992a) and discussion therein].

The benefits of the local polynomial fits are even greater in higher dimensional settings, where the boundary problem is more severe. For example, if the bandwidth is such that 20% of the data on each end may be considered "boundary points" in a one-dimensional problem, then the same bandwidth in the corresponding $d$ dimensional problem will result in about $(1 - 0.6^d)$ 100% of the data being in the "boundary region." For $d = 2$, this means that 64% of the data are in the boundary. This quantifies and supports the first conclusion made in Section 7.

## 4. DERIVATIVE ESTIMATION

The clever graphical devices developed in Figures 1–5 suggest a similar investigation of the problem of estimating derivatives of the regression function. There are three possible estimators: the derivative of the NW estimator, the derivative of the GM estimator and the linear coefficient of the local quadratic fit. All of these are linear estimators. Figure 2 gives a visual comparison of the estimators. Clearly, the NW derivative estimator has larger bias, exhibited by the zero crossing of the weights being too far to the left, while the GM derivative estimator has larger variance, indicated by the wildly fluctuating heights, in this nonuniform design. The local polynomial fit clearly overcomes both problems.

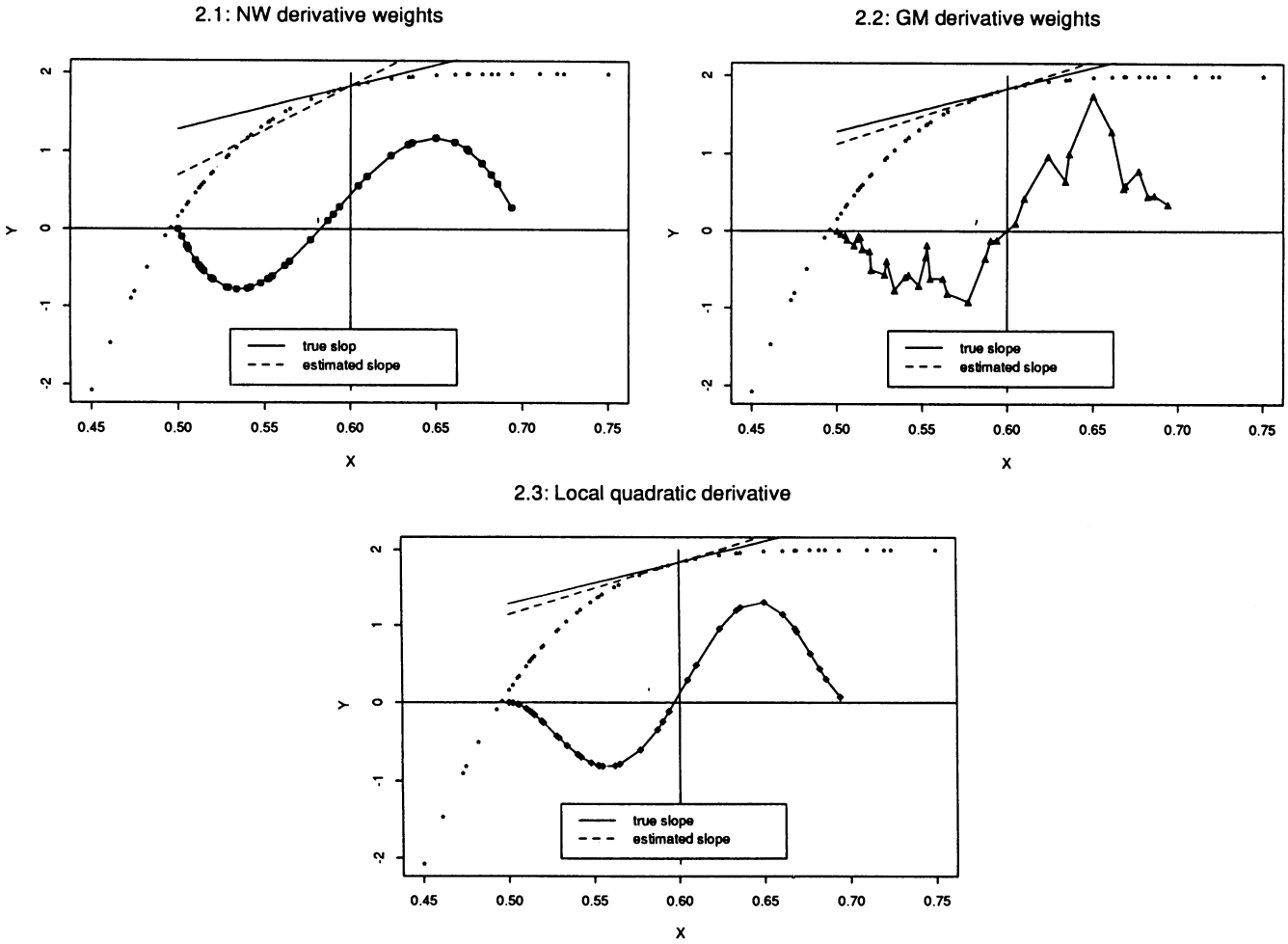


FIG. 2. *Effective weights in derivative estimation. Zero noise regression data from random design shown as stars. Effective weights, for estimation at* $x = 0.6$ *shown as other symbols.*

## 5. COMPUTATIONAL ISSUES

An important point not considered carefully in the Hastie and Loader paper is computational speed. The standard folklore in smoothing, apparently reflected in the discussion at the end of Section 1, is that smoothing splines are much faster computationally (although we disagree that fast implementations are so simple) but much less interpretable than kernel methods (with the asymptotic equivalence of Silverman (1984) being the most that seems available in this direction). However, some much faster implementations of kernel and local polynomial methods have recently been developed, including the binned approximations described in Härdle and Scott (1992), and the "extended updating algorithm" of Gasser and Kneip (1989). We are currently doing a careful comparison of these methods and find them to be far faster than the usual naive implementations (speed factors in the 100s are available even for moderately large sample sizes), and they are at least competitive with smoothing splines.

## 6. KERNELS VERSUS NEAREST NEIGHBORS

Another relevant issue not yet discussed deeply is the comparison between "fixed width" and "nearest neighbor width" window methods. [See Cleveland (1979) for a local polynomial estimator of the latter type.] The simplest formulation of these involves using a "uniform window," that is, an *unweighted* fit of a polynomial to the points in some neighborhood of $x$. In the one-dimensional case, there are two methods for determining this neighborhood. The first involves taking the smallest neighborhood, whose endpoints are equidistant from $x$, which contains a given number, say $k$, of the observations. We call this the "total nearest neighbor" method. The other uses a neighborhood which contains $k/2$ points on each side of $x$, which we call the "split nearest neighbor" method. Here we give examples only in the local average case, but the main ideas clearly extend to local polynomials.

Visual impression for how these estimators compare with each other, and with the NW estimator using a

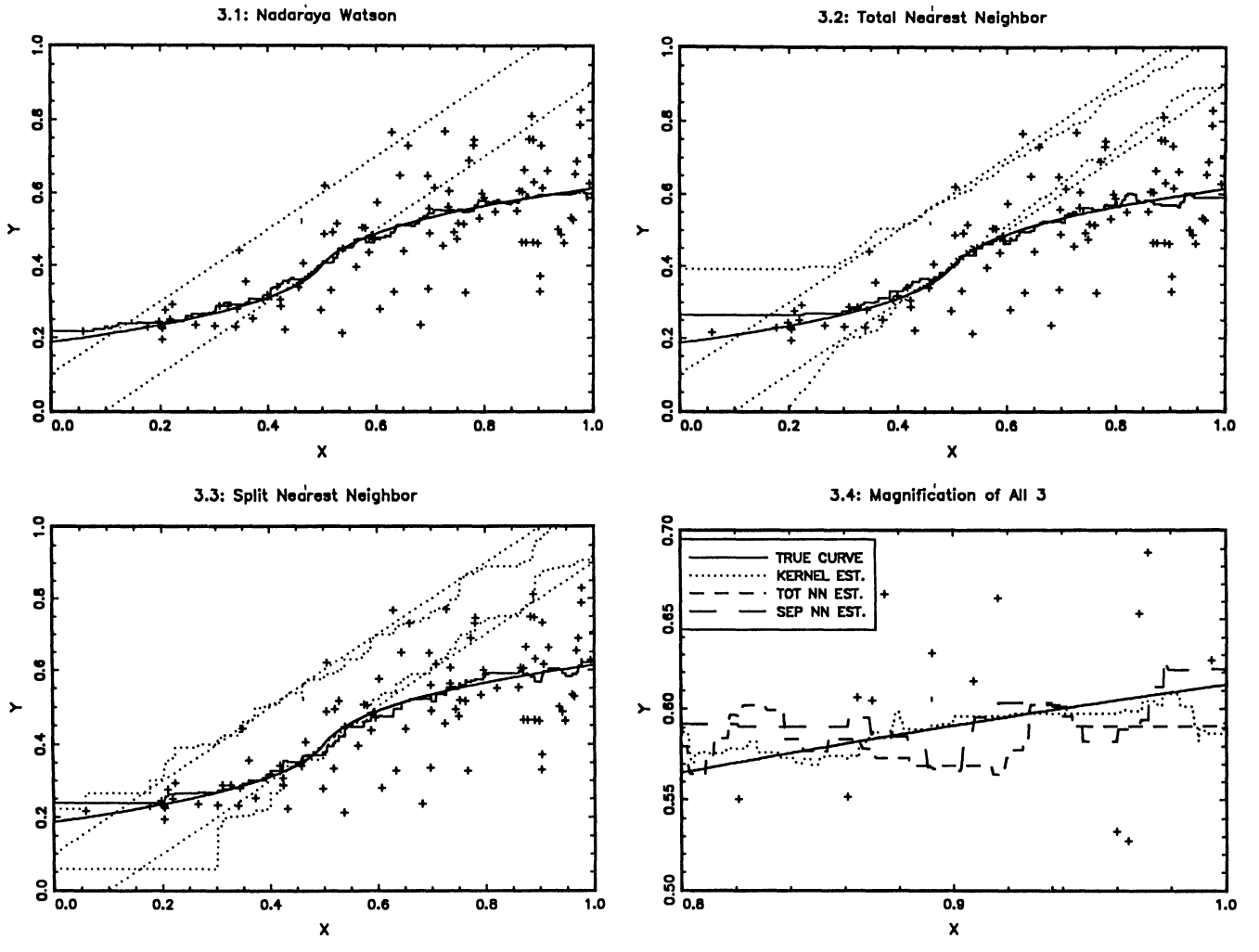


FIG. 3. *Simulated regression example, true curve is solid, data are plusses. Diagonal dotted lines in* 3.1–3.3 *show window widths for NW, jagged dotted lines in* 3.2 *and* 3.3 *show window widths for nearest neighbor methods. Estimates are step functions.*

uniform kernel, is given in Figure 3. All estimators are step functions, because of the kernel shape. The two diagonal dotted lines in Figures 3.1–3.3 show, for each $x$, the window width of the NW estimator (chosen to represent roughly the same amount of overall smoothing). For each $x$, the point on the upper jagged curve in Figures 3.2 and 3.3 shows the right endpoint of the neighborhood over which averaging is performed. Similarly the lower curve indicates the left endpoint of the window. All estimates are quite "jagged and wobbly" because of the uniform kernels, but the NW estimate is perhaps "more controlled" in its wobbliness. This is most clear in Figure 3.4, which shows a magnified comparison of the 3 estimators in the region $x \in [0.8, 1]$. Note that the two nearest neighbor estimators have more "wide flat spots occurring at random" together with "larger random jumps." The jagged dotted curves in Figures 2.2 and 2.3 explain why these happen: for some reasonably wide intervals, the neighborhoods remain constant in $x$, but in some places the neighborhoods change quite abruptly. These effects occur in a random, uncontrolled fashion because of the chaotic nature of the design points.

We view this as a major drawback of nearest neighbor methods because they lack the simple interpretability of kernel methods. The simple "moving average" intuition becomes harder to accept when the window changes in such a noninterpretable way.

## 7. ADDITIONAL COMMENTS

### 7.1 Unpublished Related Work

Interesting proposals which address the crucial problem of bandwidth selection can be found in Fan and Gijbels (1992b) and Ruppert, Sheather and Wand (1992). The local polynomial fits provide an easily implemented approach to assess the bias and the variance.

Useful results concerning analysis of both high-degree and also high-dimension local polynomial estimators are in Ruppert and Wand (1992).

A forthcoming manuscript by Fan, Gasser, Gijbels, Brockmann and Engel shows that the efficiency of the local polynomial regression fit is good even for the estimation of quite high derivatives and that the local polynomial fits yield minimax efficient linear smoothers for estimating the regression function as well as its derivatives. Moreover, it is seen that for all polynomial degrees and estimation of any derivative, the optimal kernel is still the familiar Epanechnikov kernel. This answer is much simpler than the complicated case wise solutions developed for kernel estimation in Gasser, Müller and Mammitzsch (1985), for example.

### 7.2 Open Questions

Here is a summary of the open problems discussed above.

1. What is the best way to compute local polynomial estimators?
2. Are local polynomials competitive with smoothing splines in terms of speed?
3. Which degree of polynomial should be used?
4. Is it really better to estimate derivatives by the appropriate coefficient, rather than by differentiating an estimator of the regression?

### 7.3 Closing Quote

As Theodor Gasser has said (in private conversation): "We have not found any disadvantages of the local polynomial method as yet. It should become a golden standard nonparametric technique."

### ACKNOWLEDGMENT

Research of both authors was supported in part by NSF Grant DMS-92-03135.

# Comment

**Hans-Georg Müller**

## 1. INTRODUCTION

The article by Hastie and Loader (H&L) clearly demonstrates the importance of choosing a good smoothing method in the nonparametric regression context. The authors provide important insights and further strengthen the case for the "Local Weighted Least Squares" (LWLS) method. This article is a continuation of the extensive discussion of Chu and Marron (1991) who compared various aspects of different kernel regression smoothers but did not include LWLS.

It can be argued that LWLS is a third type of kernel method, generalizing the Nadaraya–Watson (NW) approach. When discussing kernel smoothing, one may want to refer to a broader perspective which includes not only nonparametric regression as probably the most important application but also the estimation of density, spectral density, hazard, intensity, quantile density and other functions. It is then useful to have a general framework available which provides for the construction of kernels, boundary kernels and bandwidth selectors for a whole range of smoothing problems. Such a framework can be provided for "explicit" kernel methods, including Parzen–Rosenblatt kernel estimates in the density and Nadaraya–Watson, Priestley–Chao or Gasser–Müller kernel estimates in the regression context. It is not clear whether LWLS could be included in such a framework, as it is uniquely geared toward regression.

The LWLS method is of particular interest for change-point modeling (Section 5), owing to its extraordinary flexibility which allows, for instance, constructing local fits satisfying linear constraints within the local regression model. Moreover, many well-studied features of (global) linear model fits can be extended to local linear models, like testing of linear hypotheses, diagnostics, local goodness-of-fit, modeling of correlation structure and heteroscedasticity and so on. Along with the many desirable features demonstrated by H&L, this makes LWLS a very attractive option for smoothing.

We should not, however, overly rely on a single method for all possible nonparametric regression problems. It is clear that a fixed bandwidth LWLS method has problems with smoothing data like those presented in Figure 6 of H&L: the "holes" in the data may lead to inappropriate zero valued or undefined regression estimates. Window and bandwidth choices adapting to design nonuniformities are needed in such cases. This may lead to a fairly complicated smoother, so that some of the initial simplicity is lost for highly nonuni-

*Hans-Georg Müller is Professor, Division of Statistics, University of California, Davis, California 95616.*