lyzed in Cressie (1991), where data are only available averaged over counties, with counties of quite irregular shapes. In this case the data are aptly described by the "empirical regression function" $\Sigma Y_j 1_{A_j}$, where sets A_j represent the counties. Obviously, GM type estimators are directly applicable to such data, while NW and LWLS estimators are not.

A generalized version of LWLS which is capable of handling data like this is as follows: consider $\tilde{f}_{LS}(x) = \tilde{\beta}_0(x)$, where

$$(\tilde{\beta}_0, \tilde{\beta}_1) = \arg\min_{(\beta_0, \beta_1)} \int [y - (\beta_0 + \beta_1^t(t-x))]^2 \cdot G\left(\frac{t_1 - x_1}{b}, \ldots, \frac{t_p - x_p}{b}\right) dF(x,y) \tag{7}$$

for any probability measure dF on $\mathcal{R}^p \times \mathcal{R}$, $p \geq 1$.

If F is the joint distribution of measurements (X_i, Y_i), $X_i \in \mathcal{R}^p$, $Y_i \in \mathcal{R}$, this yields the convolution of the true regression function with the bandwidth-scaled kernel K_{EQ} of (2). If $dF \equiv dF_n \equiv n^{-1} \Sigma \delta_{(X_i, Y_i)}$, the empirical measure, then $\tilde{f}_{LS}(x)$ is the ordinary LWLS estimator.

Let now (A_i) be a partition of the compact domain A into measureable sets $A_i \subset A$, $1 \leq i \leq n$. Choosing $dF = \Sigma[\lambda(A_i)/\lambda(A)]\{dF_{U,A_i} \times \delta_{Y_i}\}$, where F_{U,A_i} is the uniform distribution on A_i, δ_y an atom of mass 1 at y and λ the Lebesgue measure, then one obtains

$$(\tilde{\beta}_0, \tilde{\beta}_1) = \arg\min_{(\beta_0, \beta_1)} \frac{1}{\lambda(A)} \sum_{i=1}^{n} \int_{A_i} [y_i - \beta_0 + \beta_1(t-x)]^2 \cdot G\left(\frac{t_1 - x_1}{b}, \ldots, \frac{t_p - x_p}{b}\right) dx.$$

This special case of (7) yields the proposed LWLS type estimator $\tilde{f}_{LS}(x) = \tilde{\beta}_0(x)$ for this spatial smoothing problem.

7. OTHER SMOOTHING METHODS

As mentioned by H&L, other important and popular nonparametric regression estimators are smoothing and regression splines (see Eubank, 1988, or Wahba, 1990), and it would be of interest to see how these compare with the other methods. It follows from results of Silverman (1984) that for regular designs smoothing splines are asymptotically equivalent to GM estimators with certain "equivalent" kernels with noncompact supports and bandwidths which vary locally according to the design density; compare also Messer and Goldstein (1993) who investigated corresponding boundary kernels. We may therefore expect that at any fixed point, even for finite n, the behavior of smoothing splines will be closer to GM estimators than to the other kernel methods. The design-adaptive local bandwidth variation feature of smoothing splines is a bonus. For other smoothing methods, such local bandwidths variation can be implemented as well, but only at the expense of substantial additional conceptual and numerical complexity.

One advantage of controlled local bandwidth variation for kernel and LWLS estimates, however, is that this may include adaptations not only to locally varying design density but also to curvature and heteroscedasticity (Müller and Stadtmüller, 1987; Fan and Gijbels, 1992a).

Considering the random design case, the NW estimator which is a special case of LWLS has serious drawbacks as pointed out by H&L and Chu and Marron (1991). One of the more serious problems is that this estimator cannot reproduce straight lines as regression functions (in the univariate case) when the marginal density of the predictors X_i is nonuniform. An identity reproducing transformation, which is applicable to any nonparametric regression estimator, was introduced in Müller and Song (1991) to address this problem. A corresponding identity reproducing nonparametric regression estimator (IRENORE) derived from the NW estimator then has the same asymptotic MSE properties as LWLS in random designs. It is thus another approach which achieves the desirable MSE properties of LWLS.

ACKNOWLEDGMENT

Research supported by Air Force Grant AFOSR 89-0386.

Rejoinder

Trevor Hastie and Clive Loader

If the "smoothing community" includes the users of smoothers, then local regression has been popular for more than 10 years. In particular, Cleveland's (1979) implementation is widely used across a broad spectrum of disciplines and is appreciated for its simple but effective approach to boundary bias, local bandwidth

adaptation and resistance to outliers. The S implementation lowess and more recently the multidimensional loess are largely responsible for this widespread use, although FORTRAN source code is also available from NETLIB. That local regression has excellent theoretical properties has also been clear for more than 10 years: Stone (1977, 1980, 1982) has established consistency and shown that local polynomials obtain optimal rates of convergence under minimal design assumptions. Yet in the subsequent years local regression has largely been ignored by the theoretical community. Instead much effort has been expended developing modified kernel type methods that require unnecessary restrictive assumptions and have many undesirable features. A culmination of this work was Chu and Marron (1991) who compare asymptotic bias and variance for some of these kernel estimates. Among many other good features, local linear regression achieves both the desirable asymptotic bias and asymptotic variance expressions, yet Chu and Marron decided not to consider local regression methods, discarding them as "obscure alternative methods" (p. 434). We are therefore thankful for the recent refreshing breakthroughs of Fan, which will hopefully nudge this community along more realistic directions.

The discussants have chosen to ignore the issues we raised, instead focusing on issues such as bandwidth selection and computation. These arise whether one adopts a local regression or modified kernel approach, and so we view them as irrelevant to comparisons between the approaches; however, they are important issues that practitioners must face up to. We comment briefly on these additional issues below.

The discussants also raise questions related to other nonparametric function estimation problems such as density estimation and regression in non-Gaussian families. Although these extensions were also not a focus of our paper, we emphasize below that the bias-correcting properties of local polynomial regression can be extended in a natural way via local-likelihood techniques to cover all these problems.

1. EQUIVALENT KERNELS

Effective use of nonparametric regression requires a clear understanding of what the fitting procedures are doing, how they work and when they might fail. It seems that our most important messages were lost on the discussants:

It is not helpful to think about the bias correcting properties of a smoother via the equivalent kernel (we don't appreciate the properties of linear regression by plotting the rows of the hat matrix!).

Figures 1.1–1.4 of Fan and Marron's discussion and the accompanying explanation should support this claim! Our pictures of local regression show separately the kernel that assigns the local weights (this is its natural role) and the polynomial that is fit locally to reduce the bias. It is because of this nice separation that local regression has had such intuitive appeal to practitioners. Applying this separation to Fan and Marron's Figure 1.3, one is attempting to fit a cubic polynomial over a window that includes much of the data. Clearly this cannot be a good approximation; if the data are good enough for this type of structure to be detected at all, then a smaller bandwidth should be used.

It is not intuitive to tamper with kernels in order to achieve particular bias properties.

Müller's Sections 3 and 4 support this claim!

We seriously doubt anyone would consider fitting a parametric model using modified kernels, high-order kernels and the like. It is quite possible to do so: letting $b \to \infty$ in the Gasser–Müller (GM) estimate results in a "boundary kernel" method for fitting parametric models (Hart and Wehrly, 1992). For randomly scattered designs, this limit is very inefficient and cannot be considered a serious competitor to least squares. For some of the simplest designs, such as an equidistant design, this boundary kernel method is almost indistinguishable from the least squares estimate; however, most would agree this is a miserably weak reason for introducing new estimates.

By contrast, if one lets $b \to \infty$ in the local polynomial estimate, all observations receive equal weight and the parametric least squares fit results. This is exactly how one should expect their smoother to behave; spline smoothers have a similar property.

The comparison between modified kernel and local regression methods is identical. We need the local regression method for many of the complicated designs encountered in practice: multiple predictors, randomly scattered and nonuniform designs and the like. Methods which work well but only for very special classes of designs are not worth the space they take up in the practicing statistician's toolbox.

We do not understand Müller's defense of the optimality of $K_{SO,q}$. For example, setting $q = 0$ in Müller's expression gives $K_{SO,0}(x) = -6x(1 + x)(6 - 10x)$. When fitting at an endpoint, this kernel gives zero weight to observations at the endpoint: the most informative observations are discarded! This cannot be optimal under any sensible conditions; there is no reason to enforce the boundary constraints imposed by Müller. A dramatic example of the effect of this inefficiency is described in the discussion of change points below.

2. BANDWIDTH AND ORDER SELECTION

How should one decide on the order of fitting to use, and how should the bandwidth be determined – locally or globally; fixed width; nearest neighbor or somewhere

in between? For bandwidth selection, much literature has been written on automatic methods that attempt to minimize a lack-of-fit criterion such as integrated squared error.

From a practical viewpoint, this is discarding much of the power of nonparametric regression. A major strength of nonparametric regression is flexibility in the scope of models that can be fit; we should also want flexibility in the way we specify the smoothing parameters. By fitting models of different bandwidths and orders, the user can observe how the fit changes, and diagnostic tools such as the M_f plot (Cleveland and Devlin, 1988) can be used to view the bias–variance trade-off and help determine whether observed structure is real or random. One can then choose a final fit which shows any interesting structure found, but is minimally cluttered by noise.

Müller suggests that the boundary problem becomes a bandwidth problem for local regression. This is misleading; the question of how to vary bandwidth in boundary regions arises whether one uses kernel or local regression methods. Indeed, under some assumptions on the design density, local linear regression with a constant bandwidth (but decreasing as sample size increases) achieves the same rate of convergence at boundary and interior regions, suggesting that asymptotically the boundary bandwidth problem is less important for local linear regression than for kernel methods. This uniform rate-of-convergence property continues to hold for any odd-order polynomial and in the multidimensional case; Fan and Gijbels (1992a) and Ruppert and Wand (1992) discuss these issues further.

Fan and Marron object to our statement that order selection is part of the bias–variance trade-off. We see no harm in having two smoothing parameters—order and bandwidth—and in selecting the most suitable pair for the data at hand, locally or globally.

They also raised objections to nearest neighborhoods, illustrating the rough estimates obtained using a uniform kernel. In Cleveland's (1979) implementation of local regression, a tricube kernel (with compact support) is wedged into the interval defined by the nearest neighborhood. As the neighborhood slides along, this allows observations to enter smoothly, and the visual roughness disappears. On the positive side, windows defined by near neighborhoods adapt to the local design density (in data-dense regions the windows can afford to be narrower) and lead to estimates with roughly constant variance in the interior.

3. LOCAL LIKELIHOOD

Müller defends boundary kernels by arguing that they are still required for "other models" such as Poisson regression, via local-likelihood techniques. In fact, local polynomial fitting extends to many other models through the same local-likelihood techniques. Tibshirani and Hastie (1987) studied local-likelihood estimation in models such as logistic regression and the proportional hazards model. Examples there show the value of local linear fitting in bias reduction at endpoints. Although they use local linear fitting and a rectangular weight function, these can easily be generalized. Staniswalis (1989) made the generalization to nonrectangular kernels but used local constants rather than local linear fits.

Local-likelihood methods can also be used in problems such as density and hazard-rate estimation. Supposed that $Y_1, \ldots, Y_n$ are an independent sample from a density $f(t)$. By considering a limit of multinomial models the appropriate local log likelihood for local linear fitting is

$$l(t) = \sum_{i=1}^{n} K(Y_i - t)(a + bY_i) - n\int_{-\infty}^{\infty} K(y - t)e^{a+by}dy.$$

Let $\hat{a}(t)$ and $\hat{b}(t)$ be the values of a and b which maximize $l(t)$. The density estimate is $\log \hat{f}(t) = \hat{a}(t) + \hat{b}(t)t$.

This type of density estimate has not to our knowledge been studied previously and has the correct correspondence to the penalized likelihood approach in Silverman (1986, Equation 5.30). Modeling $\log f(t)$ rather than $f(t)$ has some potential advantages over traditional kernel methods. First, the density estimate must be positive, in contrast to kernel estimates with high-order or boundary kernels. Second, when the density has unbounded support, the tails of $\log f(t)$ will often be much more polynomial-like than the tails of $f(t)$, suggesting that a local-likelihood approach may result in better estimates of the tails of densities.

4. COMPUTATION

Direct fitting of a local regression surface at individual points may be computationally prohibitive with large data sets. This is particularly true in the multidimensional setting, since the fitted surface must be evaluated at a large number of points to get an adequate representation of the surface. However, fast computational algorithms have been developed, and computational issues should not be seen as hindering the use of local regression.

A fast algorithm, loess, has been implemented in local regression software of the current release of S and is described in Cleveland, Devlin and Grosse (1988) and Cleveland and Grosse (1991). The local regression surface is computed exactly at a small but carefully chosen set of points, and a fast interpolation scheme is then used to approximate the surface at remaining points.

Another approach (especially in the one-dimensional case) is via the Fast Fourier Transform (FFT) (Silverman, 1982; Jones and Lotwick, 1984; Härdle, 1987). We outline the essential idea for the Nadaraya–Watson

(NW) estimator for uniform data. The sequence of numerators and denominators for the vector of fits at the observed design points are a discrete convolution of a kernel and, respectively, the response vector (Y) and a vector of 1's. Additional bells and whistles fix up the boundary truncation of the kernel and nonuniform designs. These FFT techniques are easily extended to local regression; for local linear regression, for example, we need additional convolutions with the vectors X, X^2 and XY (Hastie and Shirey, 1988).

Fan and Marron ask whether local polynomial methods are competitive with smoothing splines in terms of computational speed. In one dimension the answer is yes; fast $O(n)$ algorithms exist for both methods. In multiple dimensions, Cleveland and Devlin (1988) see computation as the most serious problem with thin plate splines, since a large optimization problem must be solved, taking $O(n^3)$ time. Some fast approximations have recently been developed by O'Sullivan (1991) and others; however, for large data sets these are not competitive with the local regression algorithms of Cleveland and Grosse (1991) which are $O(n)$ with a fixed smoothing parameter.

5. CHANGE POINTS AND SPATIAL ESTIMATION

Local polynomial fitting is designed for smooth functions; with other basis functions, nonsmooth functions can be obtained. For example, using the basis functions $b(x) = \{1, I(x > \tau), x, (x - \tau)^+\}$ for fixed τ induces a discontinuity at τ; $b(x) = \{1, x, (x - \tau)^+\}$ gives a discontinuous first derivative at τ. This differs slightly from Müller's proposal in the second case; in particular, data lying exactly on a broken line $a + b_1x + b_2(x - \tau)^+$ will be exactly reproduced by our method.

Often change points are unknown and must be estimated from the data. Müller (1992a) has proposed using the "smooth optimum" boundary kernels to estimate left and right limits at any point t, estimating τ to be the point where the difference is maximized. These kernels give absolutely terrible change point estimates, with rates of convergence way below the n^{-1} expected in change point models. Under a more sensible weighting scheme, Loader (1992) has established estimates (either kernel or local regression-based) that attain the n^{-1} convergence, with the same asymptotic distribution familiar from likelihood estimates in parametric change point models.

Müller suggests spatial smoothing as an application where GM estimates may be useful. However, considerable care must be used here to avoid inefficiencies. For example, consider two neighboring counties of similar size; one predominately urban with high population density, and the second predominately rural with low population density. Then the observed rate of Sudden Infant Death Syndrome will be subject to much larger random fluctuations in the second county, and hence this county should be downweighted proportionately. The direct implementation of the GM estimate will not account for this. However, variance information can easily be incorporated into the local regression approach.

ADDITIONAL REFERENCES

Azari, A. S. and Müller, H. G. (1992). Preaveraged localized orthogonal polynomial estimators for surface smoothing and partial differentiation. *J. Amer. Statist. Assoc.* **87** 1005–1017.

Cleveland, W. S., Devlin, S. J. and Grosse, E. (1988). Regression by local fitting: Methods, properties and computational algorithms. *J. Econometrics* **37** 87–114.

Cleveland, W. S. and Grosse, E. (1991). Computational methods for local regression. *Statistics and Computing* **1** 47–62.

Cressie, N. (1991). *Statistics for Spatial Data.* Wiley, New York.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Dekker, New York.

Eubank, R. L. and Speckman, P. (1991). A bias reduction theorem with applications in nonparametric regression. *Scand. J. Statist.* **18** 211–222.

Eubank, R. L. and Speckman, P. (1992). Nonparametric estimation of functions with jump discontinuities. Technical Report, Dept. Statistics, Univ. Missouri, Columbia.

Fan, J. and Gijbels, I. (1992a). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20** 2008–2036.

Fan, J. and Gijbels, I. (1992b). Design and spatial adaptation: Variable order approximation in function estimation. North Carolina Inst. Statistics Mimeo Series 2080.

Gasser, T. and Kneip, A. (1989). Discussion of "Linear smoothers and additive models" by A. Buja, T. Hastie and R. Tibshirani. *Ann. Statist.* **17** 532–535.

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. (1987). Resistant smoothing using the fast Fourier transform. *J. Roy. Statist. Soc. Ser. C* **36** 104–111.

Härdle, W.-K. and Scott, D. W. (1992). Smoothing by weighted averaging of rounded points. *Computational Statistics* **7** 97–128.

Hart, J. and Wehrly, T. (1992). Kernel regression when the boundary region is large, with an application to testing the adequacy of polynomial models. *J. Amer. Statist. Assoc.* **87** 1018–1024.

Hastie, T. and Shirey, C. (1988). A variable bandwidth kernel smoother. Unpublished AT&T technical memorandum.

Jones, M. C. and Lotwick, H. W. (1984). A remark on Algorithm AS 176. Kernel density estimation using the fast Fourier transform. *J. Roy. Statist. Soc. Ser. C* **33** 120–122.

Lejeune, M. (1985). Estimation non-paramétrique par noyaux: régression polynomiale mobile. *Rev. Statist. Appl.* **33** 43–67.

Loader, C. (1992). Change point estimation using nonparametric regression. Technical report, Dept. Statistics, AT&T Bell Laboratories.

Marron, J. S. (1992). Graphical understanding of higher order kernels. North Carolina Inst. Statistics Mimeo Series 2082.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

Messer, K. and Goldstein, L. (1993). A new class of kernels for nonparametric curve estimation. *Ann. Statist.* **21** 179–195.

Müller, H. G. (1992a). Change-points in nonparametric regression analysis. *Ann. Statist.* **20** 737–761.

Müller, H. G. (1992b). On the boundary kernel method for

nonparametric curve estimation near endpoints. *Scand. J. Statist.* To appear.

MÜLLER, H. G. and SONG, K. S. (1991). Identity reproducing multivariate nonparametric regression. *J. Multivariate Anal.* To appear.

MÜLLER, H. G. and STADTMÜLLER, U. (1987). Variable bandwidth kernel estimators of regression functions. *Ann. Statist.* **15** 610-625.

MÜLLER, H. G. and WANG, J. L. (1992). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics.* To appear.

MÜLLER, H. G. and ZHOU, H. (1991). Discussion of "Transformations in density estimation" by M. Wand, J. S. Marron and D. Ruppert. *J. Amer. Statist. Assoc.* **86** 356-358.

O'SULLIVAN, F. (1991). Discretized Laplacian smoothing by Fourier methods. *J. Amer. Statist. Assoc.* **86** 634-642.

RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1992). An effective bandwidth selector for local squares regression. Technical report, Australian Graduate School of Management, Univ. New South Wales.

RUPPERT, D. and WAND, M. (1992). Multivariate locally weighted least squares regression. Technical Report 92-4, Operations Research and Industrial Engineering Dept., Cornell Univ.

SCHUSTER, E. F. (1985). Incorporating support constraints into nonparametric estimates of densities. *Comm. Statist. Theory Methods* **14** 1123-1126.

SILVERMAN, B. (1982). Algorithm AS 176. Kernel density estimation using the fast Fourier transform. *J. Roy. Statist. Soc. Ser. C* **31** 93-99.

SILVERMAN, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.* **12** 898-916.

SILVERMAN, B. (1986). *Density Estimation.* Chapman and Hall, London.

STANISWALIS, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *J. Amer. Statist. Assoc.* **84** 276-283.

TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559-568.

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