where

$$k(\tau) = \frac{1}{\pi} \int_0^\infty \frac{1}{(1 + y^{2m})} \cos \tau y \, dy ,$$

illustrating the "bandwidth" role of λ. (See [1].)

Moore and Yackel [5] have made a detailed comparison of window vs. k-NN type density estimates and conclude (not surprisingly) that one does better with k-NN estimates near x where $h(x)$ is small (and presumably vice-versa). A direct comparison of practical k-NN type estimates vs. window type estimates for $E(Y \mid X = x)$ must of course include the prescription for choosing k or λ as well as for choosing the shape, e.g., uniform, triangular or quadratic examples as given by Professor Stone, or as determined by Q here. Any Q within the same equivalence class (in the sense of [9]) will give the same (asymptotic) results, so within a class, computational ease can be the criteria. To choose from among a finite number of representatives of equivalence classes compute $\min_\lambda V(\lambda)$ or $\min_\lambda \hat{R}(\lambda)$ for each representative and take the minimizer over the representatives tried.

REFERENCES

[1] COGBURN, R. and DAVIS, H. T. (1974). Periodic splines and spectra estimation. *Ann. Statist.* **2** 1108–1126.

[2] CRAVEN, P. and WAHBA, G. (1976). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Unpublished.

[3] HUDSON, H. M. (1974). Empirical Bayes estimation. Technical Report 58, Dept. Statist., Stanford Univ.

[4] KIMELDORF, GEORGE and WAHBA, GRACE (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95.

[5] MOORE, D. S. and YACKEL, J. W. (1976). Large sample properties of nearest neighbor density function estimators. Mimeo series 455, Dept. Statist., Purdue Univ.

[6] WAHBA, G. (1975). A canonical form for the problem of estimating smooth surfaces. Technical Report 420, Dept. Statist., Univ. of Wisconsin-Madison.

[7] WAHBA, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Num. Anal.* **14**, No. 4. To appear.

[8] WAHBA, G. (1976). A survey of some smoothing problems and the method of generalized cross-validation for solving them. Technical Report 457, Dept. Statist., Univ. of Wisconsin-Madison. *Proc. Symp. Appl. Statist.* (P. R. Krishnaiah, ed.). To appear.

[9] WAHBA, G. (1974). Regression design for some equivalence classes of kernels. *Ann. Statist.* **2** 925–934.

[10] MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

REPLY TO DISCUSSION

First I wish to thank an Associate Editor handling the paper for suggesting that it be used for discussion. I also wish to express my gratitude to him and the other discussants for the wide variety of interesting, thought provoking and uniformly constructive comments and to the Editor, Richard Savage, for his help in improving the accuracy, style and readability of the paper.

Cover wonders why continuity requirements are not needed for consistency.

Perhaps the best explanation is that a function in $L^p(\mathbb{R}^d, \mu)$ can be arbitrarily well approximated in norm by a continuous function having compact support. Cover also mentions possible extensions to separable metric spaces. But it is not at all clear how to eliminate the Euclidean nature of the proof of Proposition 11.

Breiman refers to the "leave one out" (coined PRESS by Allen) method of computing the residual sum of squares from a fit. In the present context let ${}_i\hat{E}_n(Y \mid X)$ denote an estimator of $E(Y \mid X)$ based on the data (X_j, Y_j), $1 \leqq j \leqq n$, and $j \neq i$. Then PRESS $= \sum_i (Y_i - {}_i\hat{E}_n(Y \mid X_i))^2$. As Breiman points out, the method of minimizing PRESS can be used to determine the proper amount of smoothing, to select subsets of independent variables and even to choose between, say, kernel and nearest neighbor weighting systems. For large sample sizes PRESS should be calculated as i ranges over only a subset of the cases. It may be desirable to carefully select which cases to use, intentionally including (or perhaps excluding, depending on the application) cases where X_i is in the tail of its distribution or Y_i has a large residual from a linear fit.

The method of minimizing PRESS can be combined with the method discussed by Chernoff for detecting nonlinearity. Specifically PRESS should be obtained and minimized as indicated above when the Y_i's are replaced by their residuals from a linear fit. The original residual sum of squares RSS should be compared to PRESS. If RSS is smaller or only negligibly larger than PRESS, linearity is confirmed. Otherwise nonlinearity is indicated and the estimated regression function can be obtained as suggested in Section 5 on trend removal.

Bickel wonders how much is lost by using a nonparametric method over an efficient parametric method. For reasonably large sample sizes, the combined method just described may be highly efficient in comparison to linear regression even if the true regression function is linear. Bickel proposes a different and imprecisely formulated method for achieving this goal. Parzen also proposes a method of estimation which he believes will be "asymptotically efficient."

Cox suggests that the test of a parametric form "is most easily done if the 'smoothed' estimators are calculated at an isolated set of points using nonoverlapping data sets, so that independent errors result." This could be done by using statistically equivalent blocks (see the reference to Anderson in Olshen's discussion).

Bickel, Brillinger and Olshen mention various contexts in which $(X_1, Y_1), \cdots, (X_n, Y_n)$ are not i.i.d. Sampling schemes such as the one described by Olshen are particularly interesting. Let the population on which (X, Y) is defined be divided into $J < \infty$ subpopulations representing known proportions of the original population. Let random samples of sizes $n_1, \cdots, n_J$ be drawn from these subpopulations. It should be possible to obtain consistent estimators of the quantities discussed in this paper from the combined sample as $n_1, \cdots, n_J$ each tend to infinity.

Several of the discussants describe possible advantages of various weighting systems. Wahba states in her context that "the correct choice of the 'bandwidth'

parameter is more important than the choice of the 'shape' provided the 'shape' is in an appropriate class." Eddy mentions the computational advantage of kernel weights if the weights are chosen to be zero whenever $\rho_n(X_i, x) > a_n$ for some constant a_n. A disadvantage of such weights is that the estimator must be given a special definition on the nonempty set $\{x: \min_i \rho_n(X_i, x) > a_n\}$. Eddy and Sacks point out the desirability of using weights which may be negative. Rosenblatt and Sacks are particularly concerned about possible bias when X is in the tail of its distribution. The local linear regression estimator discussed in Section 4 does yield negative weights and was designed to reduce bias in the tails. Trend removal as discussed in Section 5 is also useful in this regard. Estimators obtained from kernel weights can be expected to have smaller bias but larger variance in the tails than estimators obtained from nearest neighbor weights.

Rosenblatt refers to work which indicates that nearest neighbor *density* estimates appear to have disadvantages under certain circumstances. He suggests that possible difficulties are due to bias of the estimate in the tail of the distribution. On the other hand, Wahba states that "Moore and Yackel have made a detailed comparison of window vs. k-NN type density estimates and conclude (not surprisingly) that one does better with k-NN estimates near x where [the density] $h(x)$ is small."

Olshen suggests replacing the coordinates of the independent variables by their order statistics, thereby obtaining rules which are invariant under all strictly monotone transformations of the coordinate axes. It might be possible to modify the techniques of the present paper to verify the consistency of the resulting procedures (Olshen indicates having obtained some positive results for the classification problem). Parzen proposes some rather complicated estimators involving order statistics of the dependent variable as well as those of the independent variables. Much work seems required to justify his belief that these estimators will be asymptotically efficient.

Brunk and Pierce describe in detail a novel method of generating weight functions. Their estimate of the regression function has an expansion in terms of a finite system $\{\varphi_1, \cdots, \varphi_k\}$ of functions which are orthogonal with respect to a prescribed measure ν. No suggestions are given for choosing k or the functions $\varphi_1, \cdots, \varphi_k$.

Bickel, Eddy, Parzen, Rosenblatt and Wahba all touch upon the important problem of determining asymptotic rates of convergence of the estimators studied in the present paper. Work is clearly needed in this direction. It should be pointed out that local and global rates of convergence need not be the same. Set $m(x) = E(Y \mid X = x)$, $\hat{m}_n(x) = \hat{E}_n(Y \mid X = x)$, $r_n(x) = E((\hat{m}_n(x) - m(x))^2 \mid X = x)$ and $R_n = Er_n(X) = E(\hat{m}_n(X) - m(X))^2$. I conjecture that typically the main contribution to R_n comes from $r_n(X)$ with X in the tail of its distribution and specifically that $\lim_n r_n(x)/R_n = 0$ for $x \in \mathbb{R}^d$.

Brillinger and Hampel both point out the need for robust estimators.

Estimators of conditional quantiles are fairly robust as Hampel indicates. Also robust are *conditional L-estimators* of the form $\int_{[p_1,p_2]} J(p)\hat{Q}_n^Y(p \mid X)\,dp$, where J is continuous on $[p_1, p_2] \subset (0, 1)$. Consistency of such estimators is described in Corollary 6.

Brillinger suggests the use of what might be called *conditional M-estimators*. In comparison with conditional L-estimators they have the disadvantage of requiring an iterative solution or approximation thereto at each value of x for which an estimate is desired. These estimators also have problems of lack of uniqueness and difficulties of interpretation when $F^Y(\cdot \mid X)$ is asymmetric.

Consistency of conditional M-estimators under appropriate conditions is not hard to obtain. Let $\mathscr{F}$ denote the collection of probability distribution functions on $\mathbb{R}$. This collection can be made into a metric space by using the Prohorov metric $\mathscr{P}$, convergence in this metric being equivalent to weak convergence. An M-estimator can be thought of as a particular functional T on $\mathscr{F}$. A point $F_0 \in \mathscr{F}$ is a continuity point of T if $T(F_n) \to T(F_0)$ whenever $F_n \to F_0$ in the Prohorov metric. Let $\hat{F}_n^Y(\cdot \mid X)$ be the estimator of $F^Y(\cdot \mid X)$ based on a consistent sequence of weights. Then $\hat{F}_n^Y(y \mid X) \to F^Y(y \mid X)$ in probability for every $y \in \mathbb{R}$. Consequently $\hat{F}_n^Y(\cdot \mid X) \to F^Y(\cdot \mid X)$ in probability in the sense that $\mathscr{P}(\hat{F}_n^Y(\cdot \mid X), F^Y(\cdot \mid X)) \to 0$ in probability. The conditional M-estimator can be written as $T(\hat{F}_n^Y(\cdot \mid X))$. Suppose that $F^Y(\cdot \mid X)$ *is a continuity point of* T *with probability one*. Then $T(\hat{F}_n^Y(\cdot \mid X)) \to T(F^Y(\cdot \mid X))$ in probability. This established the consistency of the conditional M-estimator and thereby answers one of Brillinger's several interesting questions.

The Bayes and empirical Bayes formulation of Wahba's approach makes her estimator most appealing, at least for moderate sample sizes. It is computationally difficult, however, for large values of the sample size n since evaluation of her formula (3) requires $O(n^3)$ operations. The parameter λ occurring in this formula must also be determined. Wahba's suggestion of using the generalized cross-validation estimate of λ seems to yield a very good estimator. But in this connection she states on page 11 of reference [8] of her discussion that "Data sets of 50 can be handled for a few dollars." If the cost is $O(n^3)$, data sets of size $n = 500$ could be handled for a few thousand dollars. Wahba's estimator shares with the estimator $\hat{E}_n(Y \mid X)$ of the present paper the property of being sensitive to outlying values of the dependent variable.

Hampel points out that estimators such as those studied in the present paper can be used "as starting values for fitting a 'smooth' model." Wahba's discussion suggests a particular smooth model to use. Specifically let N and points $\tilde{X}_1, \cdots, \tilde{X}_N$ in $\mathbb{R}^d$ be determined by the user and set $\tilde{Y}_i = \hat{E}_n(Y \mid \tilde{X}_i)$ for $1 \leqq i \leqq N$. If the $\tilde{X}$'s are not too close together and the weight function used to determine the $\tilde{Y}$'s is appropriately selected by minimizing PRESS, then Wahba's formula (3) with $\lambda = 0$ could be applied to the data $(\tilde{X}_i, \tilde{Y}_i)$, $1 \leqq i \leqq N$, yielding

$$f(x) = (Q(x, \tilde{X}_1), \cdots, Q(x, \tilde{X}_N))Q_N^{-1}(\tilde{Y}_1, \cdots, \tilde{Y}_N)' . \tag{1}$$

Here the covariance function Q is determined by the user. Note that $f(\tilde{X}_i) = \tilde{Y}_i$ for $1 \leqq i \leqq N$. To robustify this estimator, the $\tilde{Y}$'s could be replaced by conditional medians, L-estimates, or M-estimates.

The computation of the right side of (1) can be simplified considerably by choosing the $\tilde{X}$'s and Q appropriately. Suppose first that $d = 1$ and let ν denote a positive integer parameter whose role will become clear shortly. Let Q_ν be a covariance function on $\mathbb{R} \times \mathbb{R}$, let $a_\nu \in \mathbb{R}$, let $b_\nu > 0$, let m denote a positive integer, and set $\tilde{X}_{\nu i} = a_\nu + (i - 1)b_\nu$ and $\tilde{Y}_i = \hat{E}_n(Y \mid \tilde{X}_{\nu i})$ for $1 \leqq i \leqq m$. Then (1), applied to the data $(\tilde{X}_{\nu i}, \tilde{Y}_i)$, $1 \leqq i \leqq m$, can be written in the form

$$f_\nu(x) = \textstyle\sum_i W_{\nu i}(x)\tilde{Y}_i\,, \tag{2}$$

where evaluation of $W_{\nu i}(x)$ requires inversion of the $m \times m$ matrix $(Q_\nu(\tilde{X}_{\nu i}, \tilde{X}_{\nu j}), 1 \leqq i, j \leqq m)$. If

$$Q_\nu(x, y) = (1 - b_\nu^{-1}|x - y|)^+\,, \qquad x, y \in \mathbb{R}\,, \tag{3}$$

then f_ν represents linear interpolation between successive $\tilde{X}$'s and its form outside the range of the $\tilde{X}$'s is equally clear. If $Q_\nu(x, y) = \exp(-c_\nu|x - y|)$ for $x, y \in \mathbb{R}$, where $c_\nu > 0$, then f_ν represents a simple form of nonlinear interpolation between the successive $\tilde{X}$'s.

Suppose now that $d > 1$. Let Q be the covariance function defined by $Q(x, y) = Q_1(x_1, y_1) \cdot \cdots \cdot Q_d(x_d, y_d)$, where $x = (x_1, \cdots, x_d)$ and $y = (y_1, \cdots, y_d)$. (To see that this is indeed a covariance function, let $\xi_1(\cdot), \cdots, \xi_d(\cdot)$ be independent stochastic processes having covariance functions $Q_1, \cdots, Q_d$ respectively. Then $\xi(x) = \xi_1(x_1) \cdot \cdots \cdot \xi_d(x_d)$ defines a stochastic process having covariance function Q.) For $1 \leqq i_1, \cdots, i_d \leqq m$ set $\tilde{X}_{i_1,\cdots,i_d} = (\tilde{X}_{1,i_1,\ldots}, \tilde{X}_{d,i_d})$ and $\tilde{Y}_{i_1,\cdots,i_d} = \hat{E}_n(Y \mid \tilde{X}_{i_1,\cdots,i_d})$. When applied to the data $(\tilde{X}_{i_1,\cdots,i_d}, \tilde{Y}_{i_1,\cdots,i_d})$, $1 \leqq i_1, \cdots, i_d \leqq m$, (1) takes the form

$$f(x) = \textstyle\sum_{i_1} \cdots \sum_{i_d} (\prod_{\nu=1}^d W_{\nu i_\nu}(x_\nu))\tilde{Y}_{i_1,\cdots i_d}\,. \tag{4}$$

If (3) holds, (4) reduces to a weighted average of 2^d or fewer of the $\tilde{Y}$'s. Suppose, for example, that (3) holds, that $d = 2$, and that $x_1 = s\tilde{X}_{1i} + (1 - s)\tilde{X}_{1,i+1}$ and $x_2 = t\tilde{X}_{2j} + (1 - t)\tilde{X}_{2,j+1}$, where $0 \leqq s, t \leqq 1$ and $1 \leqq i, j \leqq m - 1$. Then (4) simplifies to

$$\begin{aligned} f(x) = {} & st\tilde{Y}_{ij} + (1 - s)t\tilde{Y}_{i+1,j} + s(1 - t)\tilde{Y}_{i,j+1} \\ & + (1 - s)(1 - t)\tilde{Y}_{i+1,j+1}\,. \end{aligned} \tag{5}$$

This particular method of interpolation was suggested in Stone (1975) as being convenient for obtaining contour plots.

Equation (4) and trend removal together suggest another method for implementing Hampel's suggestion in a robust manner. Let the regression function be approximately by a function f of the form

$$f(x) = a + \textstyle\sum_{\nu=1}^d b_\nu x_\nu + \sum_{i_1} \cdots \sum_{i_d} c_{i_1,\cdots,i_d} (\prod_{\nu=1}^d W_{\nu i_\nu}(x_\nu))\,, \tag{6}$$

$$x = (x_1, \cdots, x_d)\,.$$

This function is linear in the $m^d + d + 1$ unknown parameters. If these parameters are not too numerous, they can be estimated robustly from the data by means of M-estimators. For starting values of a and b's one could use a robust linear fit. For the starting value of $c_{i_1,\ldots,i_d}$ one could use a robust estimator $\tilde{Y}_{i_1,\ldots,i_d}$ applied to the residuals from that fit.