[40] YAKOWITZ, S. and FISHER, L. (1975). Experiments and developments on the method of potential functions. *Proc. of Computer Science and Statistics*: 8th *Annual Symposium on the Interface* 419–423. Health Science Computer Facility, UCLA.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90024

DISCUSSION

PETER J. BICKEL

University of California at Berkeley

As Professor Stone has pointed out, over the years a large variety of methods have been proposed for the estimation of various features of the conditional distributions of Y given X on the basis of a sample $(X_1, Y_1), \cdots, (X_n, Y_n)$. The asymptotic consistency of these methods has always been subject to a load of regularity conditions. In this elegant paper, Professor Stone has given a unified treatment of consistency under what seem to be natural necessary as well as sufficient conditions.

His work really reveals the essentials of the problem. He has been able to do this by defining the notion of consistency properly from a mathematical point of view in terms of L_r convergence. However, the notions of convergence that would seem most interesting practically are pointwise notions. An example is uniform convergence on (x, y) compacts of the conditional density of Y given $X = x$. The study of this convergence necessarily involves more regularity conditions. At the very least there must be a natural, unique choice of the conditional density. However, such a study and its successors, studies of speed of asymptotic convergence, asymptotic normality of the estimates of the density at a point, asymptotic behavior of the maximum deviation of the estimated density from its limit (see [1] for the marginal case), etc., would seem necessary to me and to Professor Stone too! (He informed me, when I raised this question at a lecture he recently gave in Berkeley, that a student of his had started work on such questions.)

One important question that could be approached by such a study is, how much is lost by using a nonparametric method over an efficient parametric one? If density estimation is a guide, the efficiency would be 0 at the parametric model for any of the nonparametric methods surveyed by Professor Stone. However, even if this is the case, it seems clear that one can construct methods which are asymptotically efficient under any given parametric model and are generally consistent in Stone's sense. This could be done by forming a convex combination of the best parametric and a nonparametric estimate, with weights

depending upon some measure of distance of the empirical distribution of the sample from the postulated parametric model. How do such estimates perform short of $n = \infty$? Both analytic and Monte Carlo studies might be worthwhile.

Clean results for uniform convergence of estimates would presumably also be applicable to the large class of situations where the X_i are not random but selected by the experimenter, i.e., the classical regression problem.

I'll imitate the format of an R.S.S. meeting and thank the writer for a most stimulating paper.

LEO BREIMAN

University of California at Berkeley

Charles Stone's work is a significant addition to the few small bits and pieces of known theory regarding nonparametric regression. In part, its existence and publication reflects the influence of computers on statistical theory. Twenty years ago it would have been interesting but academic. Currently, the reason for this and other stirring of interest in nonparametric regression is that the research is "relevant." That is, it can be implemented in a computer program and used.

From the point of view of intelligent use, what we need badly now are studies of what happens for large but not infinite sample size. This will almost certainly be difficult. The behavior of $\hat{E}_n(Y \mid X)$ depends on an intricate interplay between sample size, the curvature of the regression surface, and the variability of Y about its regression.

What adds to this difficulty is that in actual use the number of nearest neighbors used in the estimate is calibrated by the "leave-one-out" method. Thus, the sequence of probability weights used is not predetermined, but is a function of the sample sequence.

I have two suggestions for investigating this complicated large sample behavior. The first is to look at some examples where the joint distribution of Y and X is very simple. For instance, assume that Y is a linear function of X plus an additive normal error. The second is to carry out a series of Monte Carlo experiments trying to separate out the effects of sample size, curvature and variability.

Nonparametric regression methods can be very useful tools when Y and X are related in some unknown but nonlinear fashion. Perhaps the most important application is variable selection. Here, nonparametric regression is used to compute the residual sum of squares taking for X any candidate subset of independent variables. These RSS values are then used to rank and evaluate various subsets.

Other interesting problems can be tackled. For instance, suppose we suspect that there is a good deal of nonlinear dependency between the independent variables. Then use a nonparametric regression program to estimate the proportion

of variance of X_i, the ith component of X, explained by the other components. These generalized multiple R^2 can be used to get a picture of the dependency structure.

Or suppose we want to get an estimate of the extent of the nonadditive interaction between two groups of variables, say X_1 and X_2. This problem is sticky in multiple linear regression. A possible resolution, using nonparametric methods, goes as follows: first, estimate the percent of variance explained using the "best" predictor of the general form $c(X_1) + g(X_2)$. Second, compare this to the value gotten by using the "best" predictor of the form $h(X_1, X_2)$.

All considered, it is conceivable that in a minor way, nonparametric regression might, like linear regression, become an object treasured both for its artistic merit as well as usefulness.

DAVID R. BRILLINGER

University of California at Berkeley and University of Auckland

The Bayes rule introduced by Professor Stone in Section 8 would appear to be useful for the construction of conditional M-estimates. Suppose that one is interested in estimating $\delta(X)$ of the model $Y = \delta(X) + \varepsilon$ with ε a variate statistically independent of X and with density function $f(\varepsilon)$. Then

$$-\int \log f(y - d(X))f(y - \delta(X))\,dy \tag{1}$$

is minimized by $d(X) = \delta(X)$. This suggests the estimation of $\delta(X)$ by $\hat{\delta}_n(X)$, the $d(X)$ that minimizes

$$-\sum_i \log f(Y_i - d(X))W_{ni}(X) = -\hat{E}_n(\log f(Y - d(X)) \mid X) \tag{2}$$

as, following Theorem 1, expression (2) tends to (1). Such a procedure corresponds to the Bayes rule with $\mathscr{L}(Y, a) = -\log f(Y - a)$. Robust estimates may be produced by requiring that $\mathscr{L}(Y, a)$ not give too much weight to extreme values of Y. Just as Huber did, one could equally take as estimate the solution of the equation

$$\sum_i \psi(Y_i, \hat{\delta}_n(X))W_{ni}(X) = 0$$

for some function ψ with $E(\psi(Y, \delta(X)) \mid X) = 0$. Can Professor Stone suggest some conditions, analogous to those set down for the consistency of maximum likelihood or M-estimates, under which $\hat{\delta}_n(X)$ converges to $\delta(X)$ in probability?

It is important that some measure of sampling variability be attached to the estimates of the paper. On many occasions there are strong arguments for considering variability conditional on the observed X_i values. Is there a simple analog of Theorem 1 for the case of fixed X values? Important information concerning variability is clearly contained in the residuals $g(Y_i) - \hat{E}_n(g(Y) \mid X_i)$, $i = 1, \cdots, n$. Can Professor Stone suggest a reasonable estimated based on these values? Tukey's jackknife procedure could clearly be used in many situations.

Finally, because the proposed estimate smooths across X-space, the more

nearly constant $E(g(Y) \mid X)$ the better. Transformations should be employed to make the relationship more nearly constant whenever possible, in the manner of the prewhitening operation of power spectral analysis.

H. D. BRUNK AND DONALD A. PIERCE

Oregon State University

Charles Stone has skillfully attacked an important problem and predictably has obtained interesting and useful results. He characterizes weight functions having desirable consistency properties and describes a family of uniformly consistent weight functions. Of course it is conceivable that in a particular situation an estimator not obtained from a consistent weight function could be better in some appropriate sense for moderate sample sizes. Still the classes Stone describes would seem to offer promise of being able to furnish estimators that are good in practice.

In the related problem of density estimation, a great many kernel estimators are available that have interesting asymptotic properties. Whittle's approach (1958) points the way to a method for selecting some that can be expected to work well in practice. And his basic idea is applicable in the present context as well.

For simplicity of exposition, let x and Y both be real valued. For fixed real x let Y denote an observation on a univariate distribution associated with x. Denote the regression function by $R(\cdot)$:

$$R(x) \equiv EY .$$

This regression function is assumed unknown and is to be estimated. We assume that the variance of the distribution is known:

$$v(x) \equiv \operatorname{Var} Y .$$

Let $Y_1, \cdots, Y_n$ be independent observations on the associated distributions:

$$EY_j = R(x_j) , \qquad j = 1, 2, \cdots, n .$$

The integer n and the reals $x_1, \cdots, x_n$ are fixed throughout. Let W be a weight function; the estimator $\hat{R}$ under consideration is

$$\hat{R}(x) \equiv \sum_{j=1}^{n} Y_j W_j(x) .$$

For greater clarity in the ensuing discussion we use a tilde underline to indicate a quantity conceived (modeled) as random. Since $\underset{\sim}{Y}_1, \cdots, \underset{\sim}{Y}_n$ are random variables, so is $\underset{\sim}{\hat{R}}(x)$ for each x, and we may consider, for fixed x,

$$E_S[\underset{\sim}{\hat{R}}(x) - R(x)]^2 ,$$

where S stands for "sample" and E_S is expectation according to the joint distribution of $\underset{\sim}{Y}_1, \cdots, \underset{\sim}{Y}_n$. Following Whittle we impose also a prior probability structure on $\{\underset{\sim}{R}(t) : t \in \mathbb{R}\}$ and now may consider, for fixed x,

$$E_P(E_S[\underset{\sim}{\hat{R}}(x) - \underset{\sim}{R}(x)]^2) \tag{1}$$

where E_P denotes expectation according to the (prior) joint distribution of $\{\underset{\sim}{R}(t): t \in \mathbb{R}\}$. One then hopes to choose a weight function W so as to minimize this expected squared discrepancy.

The weights $\{W_j(x), j = 1, 2, \cdots, n\}$ that minimize (1) may be identified also as coefficients of the linear expectation of $\underset{\sim}{R}(x)$ (recall x is fixed and $\underset{\sim}{R}(x)$ a random variable) given $\underset{\sim}{Y} \equiv (\underset{\sim}{Y}_1, \underset{\sim}{Y}_2, \cdots, \underset{\sim}{Y}_n)'$:

$$\sum_{j=1}^n W_j(x)\underset{\sim}{Y}_j = \hat{E}(\underset{\sim}{R}(x) \mid \underset{\sim}{Y}) . \tag{2}$$

The term "linear expectation" is used in the sense given it by Hartigan (1969): the linear expectation of a random variable $\underset{\sim}{T}$ given a random vector $\underset{\sim}{U} = (\underset{\sim}{U}_1, \cdots, \underset{\sim}{U}_n)'$ is defined to be the linear function $L(\underset{\sim}{U}) \equiv a_0 + a_1\underset{\sim}{U}_1 + \cdots + a_n\underset{\sim}{U}_n$ that minimizes $E[L(\underset{\sim}{U}) - \underset{\sim}{T}]^2$. That is, in the Hilbert space of random variables with finite variances, $\hat{E}(\underset{\sim}{T} \mid \underset{\sim}{U})$ is the projection of $\underset{\sim}{T}$ on the span of $\{\underset{\sim}{1}, \underset{\sim}{U}_1, \cdots, \underset{\sim}{U}_n\}$. If $\underset{\sim}{T}$ is a random vector, $\underset{\sim}{T} = (\underset{\sim}{T}_1, \cdots, \underset{\sim}{T}_k)'$, then $\hat{\underset{\sim}{T}} \equiv E(\underset{\sim}{T} \mid \underset{\sim}{U})$ is the random vector whose rth component is $\hat{E}(\underset{\sim}{T}_r \mid \underset{\sim}{U})$, $r = 1, 2, \cdots, k$.

We shall attempt to select prior distributions for $\{\underset{\sim}{R}(t): t \in \mathbb{R}\}$ which express an opinion that the regression function is "smooth." To this end, let $\{\phi_r(x): r = 1, 2, \cdots, k, x \in \mathbb{R}\}$ be a system of functions $\mathbb{R} \to \mathbb{R}$, orthonormal with respect to a prescribed measure ν:

$$\int \phi_r(x)\phi_s(x)\nu(dx) = \delta_{rs} , \qquad r, s = 1, 2, \cdots, k .$$

We assume that $R(\cdot)$ has an expansion in terms of these functions. That is, there are $\beta_1, \cdots, \beta_k$ such that

$$R(x) = \sum_{r=1}^k \beta_r\phi_r(x) = [\phi(x)]'\beta , \qquad x \in \mathbb{R} ,$$

where $\beta \equiv (\beta_1, \cdots, \beta_k)'$, $[\phi(x)] \equiv (\phi_1(x), \cdots, \phi_k(x))'$.

The prior distribution of $\{\underset{\sim}{R}(t): t \in \mathbb{R}\}$ will be specified by describing a joint prior distribution for $\underset{\sim}{\beta}_1, \cdots, \underset{\sim}{\beta}_k$. After subtraction of a likely candidate for the prior mean, $R_0(x) \equiv E_P\underset{\sim}{R}(x)$, we may assume we should like to specify the prior distribution so that $E_P\underset{\sim}{R}(x) = 0$. This can be achieved by setting $E\underset{\sim}{\beta} = 0$.

For the further specification of the distribution of $\underset{\sim}{\beta}$, it is useful to consider its "best fit" interpretation. Not only is the integrated squared error $\int [R(x) - \sum_{r=1}^k c_r\phi_r(x)]^2\nu(dx)$ minimized by setting $c_r = \beta_r$, $r = 1, 2, \cdots, k$, but also for fixed r, $c_r = \beta_r$ minimizes $\int [R(x) - c_r\phi_r(x)]^2\nu(dx)$. Thus each coefficient β_r has an interpretation that is independent of β_s for $s \neq r$. This makes it seem reasonable to give $\underset{\sim}{\beta}$ a prior distribution according to which $\underset{\sim}{\beta}_1, \cdots, \underset{\sim}{\beta}_k$ are independent. Since only first and second moments of $\underset{\sim}{\beta}_1, \cdots, \underset{\sim}{\beta}_k$ are involved in the determination of the posterior linear expectation of $\underset{\sim}{R}(x)$ given $\underset{\sim}{Y}$, the problem of specifying, for present purposes, the prior distribution of $\underset{\sim}{\beta}$ is now reduced to that of specifying the precisions

$$\tau_r \equiv (\text{Var } \underset{\sim}{\beta}_r)^{-1} , \qquad r = 1, 2, \cdots, k .$$

For appropriate choices of systems $\{\phi_r(\cdot): r = 1, 2, \cdots, k\}$, one may express a prior opinion that $R(\cdot)$ is smooth by letting τ_r increase rapidly as r increases.

When the prior distribution of $\underset{\sim}{\beta}$ has been specified, the weights $W_j(x)$, $j = 1, 2, \cdots, n$, that are optimal in Whittle's sense are given by (2). Hartigan (1969) provides formulas for calculation of a linear expectation as follows.

We have

$$E(\underset{\sim}{\beta}) = 0\,, \qquad V(\underset{\sim}{\beta}) = \Sigma_0\,,$$
$$E(\underset{\sim}{Y} \mid \underset{\sim}{\beta}) = A\underset{\sim}{\beta}\,, \qquad V(\underset{\sim}{Y} \mid \underset{\sim}{\beta}) = \Sigma\,,$$

where $\Sigma_0^{-1} = \operatorname{diag}(\tau_i)$, $\Sigma = \operatorname{diag}(v(x_i))$, and $A = \{\alpha_{ij}\}$ with $\alpha_{ij} = \phi_j(x_i)$. Then, using Hartigan's formula, the linear expectation of $\underset{\sim}{\beta}$ given $\underset{\sim}{Y}$ is

$$\hat{E}(\underset{\sim}{\beta} \mid \underset{\sim}{Y}) = (A'\Sigma^{-1}A + \textstyle\sum_0^{-1})^{-1}A'\Sigma^{-1}Y\,,$$

and so

$$\hat{E}(\underset{\sim}{R}(x) \mid \underset{\sim}{Y}) = [\phi(x)]'\hat{E}(\underset{\sim}{\beta} \mid \underset{\sim}{Y})$$
$$= [W(x)]'Y\,,$$

where

$$[W(x)]' = [\phi(x)]'(A'\Sigma^{-1}A + \Sigma_0^{-1})^{-1}A'\Sigma^{-1}\,.$$

These optimal weights and the corresponding estimator $\hat{R}(x)$ take particularly simple forms when ν is that probability measure on the finite set $\{x_1, \cdots, x_n\}$ that assigns probability

$$p_i \equiv \pi(x_i)/K$$

to $\{x_i\}$, $i = 1, 2, \cdots, n$, where $\pi(\cdot)$ is the precision:

$$\pi(x) \equiv 1/v(x)\,,$$

and where

$$K \equiv \textstyle\sum_{j=1}^{n} \pi(x_j)\,.$$

In this case

$$A'\Sigma^{-1}A = KI$$

where I is the $k \times k$ identity matrix. We have then

$$W_i(x) = p_i \textstyle\sum_{r=1}^{k} \dfrac{K}{K + \tau_r}\, \phi_r(x_i)\phi_r(x)\,,$$

and

$$\hat{R}(x) = \textstyle\sum_{r=1}^{k} \lambda_r \hat{\phi}_r \phi_r(x)\,,$$

where

$$\hat{\phi}_r \equiv \textstyle\sum_{i=1}^{n} p_i \phi_r(x_i) Y_i$$

and

$$\lambda_r = K/(K + \tau_r)\,, \qquad r = 1, 2, \cdots, k\,.$$

Note that $\sum_j W_j(x) = 1$ if $\phi_1(\cdot) \equiv 1$ and if $\tau_1 = 0$.

One of us has been studying the use of these estimators in certain applications with the support of the National Science Foundation through Grant MCS76-02166.

REFERENCES

[1] HARTIGAN, J. A. (1969). Linear Bayes methods. *J. Roy. Statist. Soc. Ser. B* **31** 446–454.

[2] WHITTLE, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. Ser. B* **20** 334–343.

HERMAN CHERNOFF

Massachusetts Institute of Technology

This paper is remarkable in achieving rather deep generality of results with great efficiency of presentation at very little cost expressed in terms of strength of conditions. One exception is condition (2) of Theorem 1 which appears unnecessarily strong if one were to confine attention to problems where the regression function were subject to adequate regularity conditions. On the other hand the trimming techniques discussed in Section 4 could be applied to modify weights for which (2) is not satisfied to those where they are and to establish desired results.

Consistent nonparametric regression has considerable potential value in applications involving complex relations and several independent variables. Then the use of least squares regression applied to polynomials or other simple finite expansions has potential disadvantages. If the polynomial or functional form used is not theoretically meaningful, the parameters estimated are not easily interpreted. Neighborhoods in the X-region where the regression fluctuates rapidly have a very large influence on the estimates of the parameters of the regression being fitted. Consequently, it is possible, and indeed likely, that over large regions of the X-region where the regression is stable, the estimated regression will be consistently biased. This bias is a consequence of the parametric approximation and not of limits on the information available and the nonparametric methods will not be subject to this difficulty.

In the section on trend removal Stone indicates how a linear trend can be removed in a way which makes more use of global behavior than does the method of local linear regression. The same idea can be applied to testing the adequacy of parametric models of regression.

Suppose that theory calls for a regression with a specified functional form f, i.e.,

$$Y_i = f(X_i, \theta) + u_i \qquad i = 1, 2, \cdots, n$$

where least squares can be used to estimate θ by $\hat{\theta}$. Then we have the calculated residuals

$$\hat{u}_i = Y_i - f(X_i, \hat{\theta})$$

which should behave like random residuals under appropriate conditions. If X were one dimensional, or if the X_i were preselected with sufficient regularity, visual inspection or the application of the Durbin–Watson statistic could detect signs of systematic behavior of the $\hat{u}_i$ which would indicate inadequacy of the model.

Without such regularity, one could apply simple local linear regression to the residuals to fit

$$\hat{E}_n(\hat{u}_i \mid X_i) = v_i = \hat{a}_n(X_i) + \hat{b}_n(X_i) \cdot X_i \,.$$

Let $w_i = \hat{u}_i - v_i$, $i = 1, 2, \cdots, n$ be the residuals from the locally linear regression of the $\hat{u}_i$ on X. Then the relative magnitudes of the $\hat{u}_i$ and w_i indicate

how well the theory fits. If the original model fits well, the regression of the residuals should do very little at reducing the residuals and the w_i should be close to the u_i. If the original model did not fit, the regional bias would easily be eliminated and the w_i would tend to be small compared to the $\hat{u}_i$.

In its simplest form the local linear regression could regress $\hat{u}_i$ on 1, giving v_i as the locally weighted average of the $\hat{u}_i$. One could go to the other extreme and use high order polynomial expansions although it is expected that simple linear expansions would ordinarily be effective with the use of local linear weights.

THOMAS M. COVER

Stanford University

Stone's paper has theoretical and practical importance in regression and classification when the underlying joint distribution of the observed and unknown random variables is unknown. The nearest neighbor principle on which these estimators rely might be stated as, "Objects that look alike are likely to be alike." I shall discuss this idea and attempt to describe why the weighted nearest neighbor methods are consistent.

The essence of Stone's investigation can be perceived as the use of one random variable to estimate the value of an independent copy. Consider, for example, independent identically distributed random variables Y_0, Y_1. Suppose we observe Y_1 and wish to say something about Y_0. I shall examine three cases to show that Y_1 contains much of the usually required information about the as yet unobserved random variable Y_0.

(1) *Estimation.* Assume Y_i takes values in R^d and assume a squared error loss criterion. The optimal estimate of Y_0 is simply the mean $\mu = EY_0$, assuming, of course, that the underlying distribution of Y_0 is known. The incurred risk is $R^* = E(Y_0 - \mu)^2$. However, if the distribution is unknown, so μ cannot be computed, Y_1 is a reasonable estimate of Y_0 in the following sense:

$$E(Y_0 - Y_1)^2 = E((Y_0 - \mu) + (Y_1 - \mu))^2 = 2R^* .$$

(2) *Estimation.* Suppose that the risk criterion is given by a metric ρ on $\mathscr{Y} \times \mathscr{Y}$. Suppose also that μ^* minimizes $E\rho(\mu, Y_0)$. We observe that

$$E\rho(Y_1, Y_0) \leqq E\rho(Y_1, \mu^*) + E\rho(\mu^*, Y_0) = 2R^* .$$

Again, the risk is within a factor of two of the minimal risk.

(3) *Classification.* Now suppose that Y is atomic, taking on m values with probabilities $p_1, p_2, \cdots, p_m$. Assume a probability of error loss criterion. Thus, the minimal risk for a known $p_1, p_2, \cdots, p_m$ is $R^* = 1 - p_{\max}$. On the other hand [2],

$$P\{Y_0 \neq Y_1\} = \sum p_i(1 - p_i) \leqq R^*(2 - (m/m - 1)R^*) \leqq 2R^* , \quad \forall p .$$

Yet again the risk is less than twice the minimal risk.

Thus if we can get our hands on a similarly drawn random variable, we can achieve a risk less than twice the Bayes risk.

If n independent copies $Y_1, Y_2, \cdots, Y_n$ are available, the obvious good estimators for Y_0 are the values of μ_n minimizing $(1/n) \sum_{i=1}^{n} L(\mu, Y_i)$ for (1) $L(\mu, y) = (\mu - y)^2$; (2) $L(\mu, y) = \rho(\mu, y)$; and (3) $L(\mu, y) = I_{(\mu \neq y)}$, respectively. Then $EL(\mu_n, Y_0) \to R^*$.

In [1, 2, 3, 4] and the current paper, one is not given an independent copy of Y_0, nor is one given an observation X_0 and a joint distribution on (X_0, Y_0). Instead, one is provided with a collection of pairs of random variables (X_1, Y_1), $(X_2, Y_2), \cdots, (X_n, Y_n)$, independently distributed as (X_0, Y_0), where the underlying joint distribution is unknown. Given X_0, what is a good estimate of Y_0? It is natural, given $(X_0, \cdot)$, to estimate the unknown variable Y_0 by referring to the random sample $\{(X_i, Y_i)\}_1^n$. This is where the problem comes in, because usually none of the X_i's will be precisely equal to X_0. Thus the distribution of Y_0 and the distribution of the Y coordinate of a selected $X \in \{X_1, \cdots, X_n\}$ will generally not be the same. One assumes that nearby X_i's will have nearby conditional distributions. Treating the nearest neighbor X_i as if it were equal to X_0 and following the previous procedure, one would expect to get a good estimate for Y_0. This is precisely what is done in rules of the nearest neighbor type. Stone weights the neighbors according to their rank in distance.

It would seem at first that some continuity of the joint distribution on (X, Y) is required, and indeed the consideration of the previous publications on the subject were limited to joint distributions which had no singular part. Stone has extended the discussion to joint distributions without restriction, while at the same time adding much to the knowledge of the asymptotic behavior of such procedures. Moreover, Stone finds necessary and sufficient conditions on the weighting functions that yield consistency. The extension of Stone's theorem to separable metric spaces X is a natural open question.

We see that continuity of the joint distribution is not the essential assumption necessary for believing that nearest neighbors have nearby distributions. Perhaps the heuristic reason for this is that the event that X is a point for which the continuity properties fail has probability measure zero.

REFERENCES

[1] COVER, T. (1968). Estimation by the nearest neighbor rule. *IEEE Trans. on Information Theory* **IT-14** 50–55.

[2] COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. on Information Theory* **IT-13** 21–27.

[3] FIX, E. and HODGES, J., JR. (1951). Discriminatory analysis, nonparametric discrimination. I. Consistency properties. USAF School of Aviation Medicine, Texas Project 21-49-004, Report No. 4, Contract AF41(128)-31.

[4] FIX, E. and HODGES, J., JR. (1952). Discriminatory analysis, nonparametric discrimination. II. Small sample performance. USAF School of Aviation Medicine, Texas, Project 21-49-004, Report No. 11, Contract AF41(128)-31.

D. R. Cox

Imperial College, London

Dr. Stone has given some very interesting results. The following brief comment concerns not the results themselves so much as one circumstance under which methods of this type are likely to be useful. This is in the preliminary analysis of sets of data, leading towards some simple parametric formulation of the systematic part of the regression relation. Such parametric formulations are highly desirable for concise summarization; of course, if the objective is prediction in the narrow sense, parametric formulations are by no means essential. Now one may hope for a preliminary analysis to suggest a parametric form, and consistency of the smoothed data with that form will need checking. This 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. The analogy is with simple Daniell smoothing in spectral estimation. It would be very useful to have Dr. Stone's comments on the effect of introducing this admittedly vaguely formulated constraint into the problem.

WILLIAM F. EDDY

Carnegie-Mellon University

Professor Stone has given a general set of conditions (Theorem 1), and a set which are independent of the distribution of (X, Y) (Theorem 2), under which an unknown conditional regression function $E(Y \mid X)$ can be consistently estimated. It is impressive indeed that he was able to derive such general results but practical considerations suggest the generality has some drawbacks.

Nearest neighbor methods usually assign weights 0 or 1 depending only on the rank of $\rho_n(X, X_i)$ and are thus discontinuous in X. Kernel methods, on the other hand, assign weights depending only on the value of $\rho_n(X, X_i)$ and are thus independent of $E(Y \mid X)$ near $X = x_0$. Professor Stone has made a sensible compromise between these two methods in defining nearest neighbor weights. His definition (8) requires that the weights be monotonic nonincreasing in the ranks of the $\rho_n(X, X_i)$. This is apparently needed in the proof of Proposition 11 but is not obviously necessary otherwise. It makes sense to consider weight functions that are not monotone; in fact it may even be sensible to allow negative weights. By analogy with spectral analysis of time series, weight functions with negative side lobes may reduce the error of the estimate, particularly if $E(Y \mid X)$ changes a great deal in the vicinity of $X = x_0$.

Implementation of Professor Stone's k-NN procedure for large numbers of observations in high dimensions will require formidable amounts of computation. The expensive portion of the computation is identification and ranking of those X_i which are nearest X. The usual nearest-neighbor problem is merely to identify those k (out of n) of the X_i which are nearest X; here, the k points must be ordered by $\rho_n(X, X_i)$. The usual version of the problem has been attacked by

computer scientists with some success by allowing preprocessing. In R^2, Shamos and Hoey (1975) find the k closest points to a new point X in $O(\max(k, \log n))$ time. For R^d, Friedman, Baskett, and Shustek (1975) gave an algorithm with expected time (when the X_i are uniformly distributed) $O(n(k/n)^{1/d})$ for each new point. Neither of these algorithms solves the problem of ordering the k nearest points.

As mentioned above, kernel weights do not depend on $E(Y \mid X)$ near $X = x_0$ and thus, to achieve consistency, some restrictions must be made on the distribution of (X, Y) when using them. Nadaraya (1970) has shown that if X has a positive continuous marginal density and the regression function $m(x) = E(Y \mid X = x)$ is continuous then $\hat{E}_n(Y \mid X = x)$ converges to $m(x)$. The computational advantage of kernel weights occurs when the weights are chosen to be zero whenever $\rho_n(X_i, X) > a_n$ for some positive decseasing sequence $\{a_n\}$. The advantage can be further increased by an alternative definition of kernel weights. Let $X_i = ({}_1X_i, \cdots, {}_dX_i)$ and $X = ({}_1X, \cdots, {}_dX)$ and then let the weights be given by

$$W_{ni}(X) = \frac{\prod_{j=1}^{d} K_j(\rho_{nj}({}_jX_i, {}_jX))}{\sum_{i=1}^{n} \prod_{j=1}^{d} K_j(\rho_{nj}({}_jX_i, {}_jX))}$$

where K_j is a one-dimensional kernel and ρ_{nj} is a metric on R^1. This definition simplifies the distance calculations by separating the dimensions; it is especially useful when ${}_1X$ and ${}_2X$ measure variables which are not commensurate so that nearest-neighbor methods may be inappropriate.

The asymptotic mean-square error of $\hat{E}_n(Y \mid X)$ depends on the joint distribution of (X, Y) so it is unreasonable to hope that a fixed sequence of weights $\{W_{ni}\}$ could minimize this error for all distributions of (X, Y). A complex adaptive scheme to generate the weights could probably be concocted so as to minimize this asymptotic mean-square error but it would require considerable effort and the moderate sample-size behavior might not be good. A computationally simpler scheme would be to generate weights whose degree of smoothing depends on a single parameter related to the sample size.

REFERENCES

FRIEDMAN, J. H., BASKETT, F. and SHUSTEK, L. J. (1975). An algorithm for finding nearest neighbors. *IEEE Trans. Comp.* **C-24** 1000–1006.

NADARAYA, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theor. Probability Appl.* **10** 134–137.

SHAMOS, M. I. and HOEY, D. (1975). Closest-point problems. *Sixteenth Symp. on the Foundations of Computer Science IEEE Conf. Rec.* 151–162.

FRANK HAMPEL

Swiss Federal Institute of Technology, Zürich

The approach is nonparametric in the strong sense that not only the distribution of errors is arbitrary, but also the shape of the regression function or

fitted model. Thus, e.g., no assumption of linearity, or of maximum degree of a polynomial serving as the regression function of Y on X has to be made. A closer look, however, reveals that there is still some assumption lurking in the background, an assumption weaker than any parametric model for the fit, yet implying some redundance: namely an assumption about some sort of smoothness or local linearity of the regression function. Perhaps even sufficient continuity of the conditional distributions seems desirable. It is true that, formally, the definition of consistency allows for a (variable) exceptional set of "discontinuity"; however, on such a set, one would not expect meaningful practical results. Moreover, the fact that locally and globally linear trends are taken out, with resulting "improved performance," seems to show that the author does not really believe in the usefulness or frequent occurrence of completely arbitrary nonparametric models. Finally, the basic idea itself assumes at least continuity of the expectations considered.

This basic idea says that since we usually do not have enough information about the conditional distribution of Y at a fixed value of X, let us "borrow strength" from neighboring values of X, by smoothing the model locally. There is, as usual, an interplay between variance and bias at each x as $n \to \infty$; and for random X_i there is also an interplay between variance and bias for fixed n and varying x. The k_n-nearest neighbor weight functions considered in the paper fix essentially the variance reduction while allowing variable window width for fixed n; for $n \to \infty$, any sequence such that variance and bias both tend to zero is permitted. There are the usual problems of the meaning of an asymptotic sequence which for every n does something else, and of imbedding a procedure for a fixed n into an asymptotic sequence. It should be kept in mind, though, that for each fixed n not the true regression function is estimated, but rather the regression function smoothed by some random window.

The resulting estimated regression function will still be rather wiggly locally, even if the true regression function happens to be very smooth. This is well known for moving averages and running medians, for example. One may, however, use these estimators as a starting point for fitting a "smoother" model.

To talk about robustness is meaningless or, rather, hopeless in the case of a completely arbitrary model; for a model with wild spikes and a nice model with some distant gross errors superimposed are indistinguishable. If we believe in a "smooth" model without spikes, however, then some robustification is possible. In this situation, a clear outlier will not be attributed to some sudden change in the true model, but to a gross error, and hence it may be deleted or otherwise made harmless. Obviously, many estimators discussed in the paper, notably the estimators of first and second order quantities including the trimmed local linear weight functions, are not robust in this sense: a single outlying Y can arbitrarily change the estimate. On the other hand, such nonlinear methods as the estimators of quantiles are more or less robust, depending on the particular quantile and the weight function considered.

RICHARD A. OLSHEN

University of California at San Diego

With the paper under discussion Professor Stone has made a fundamental contribution to the theory of nonparametric regression. Whereas previous work on weighted nearest neighbor procedures has invariably been laced with superfluous regularity conditions on the joint distribution of Stone's (X, Y), Theorem 1 gets to the very heart of what is needed for consistency and for the famous results of Cover [2] and of Cover and Hart [3]. Moreover, once Stone has pointed the way, it is clear that the function $\beta(\cdot, \cdot)$ figures in the condition (1) of Theorem 1, and thus the importance of Propositions 11 and 12 is manifest.

The independence of $(X, Y), (X_1, Y_1), \cdots$ is crucial to Theorem 1. Yet that independence is used only sparingly in the proof, which is basically an L^2 argument. Indeed, an implicit application of Fubini's theorem in the paragraph following (13) seems to be the only real use of independence. There is at least one situation of practical importance to which the results of this paper do not apply precisely because of the stated assumption of independence. It occurs in problems of Stone's Model 3 (classification), which is of special interest to me, and which is the subject matter of virtually all of my subsequent remarks. For convenience, suppose in what follows that the range of Y has only two values.

It often happens that an experimenter has available to him large sets of descriptive data on members of the two populations—call them I and II. He fixes two numbers, say n_1 and n_2, in advance, and records data on n_1 members of population I and n_2 members of population II. It seems intuitively clear that if, for example, $k = k(n_1 + n_2)$ nearest neighbor weights are used in determining $\hat{\delta}$ then with appropriate choices of k, n_1, and n_2 the Bayes classification rule should be arbitrarily well approximated, and yet this is a scenario to which the theorems of the present paper do not apply. (If instead the composition of the data by population is determined by i.i.d. tosses of a coin, then consistency obtains as the size of the data set increases without bound.)

Weighted nearest neighbor rules for classification have one interesting and possibly important shortcoming in the Model 3 scenario of the present discussion. For the classification problem is invariant under all strictly monotone transformations of the coordinate axes; the maximal invariants are the coordinatewise ordered population labels of the training sets (see [1]). And the rules of Professor Stone's paper are not, as they stand, invariant rules. I think it is important that two scientists engaged in classification based on otherwise identical data should not utilize different rules only because one is given the weights of the patients, and the other is given the logarithms of the weights. (When the range of Y is the real line instead of a finite set, it may be more important that $\hat{\delta}$ be a smooth function of the data than that it be invariant in the sense described.)

It is easy to mimic rules of the paper with invariant rules. Simply coordinatize the data by the indices of their marginal order statistics, and apply any of

the given rules to the "transformed" data. The question of consistency of the "transformed" rules remains to be investigated. I cite a simple, very preliminary example of what can be proved: if the l_∞ metric is employed on the transformed data, and if uniform, consistent k-nearest neighbor weights are used, then when the true marginal distributions of the training samples contain no atoms, consistency obtains. Notice that when the l_∞ metric is used, neighbors no more than a given distance from an observation lie in a rectangular parallelopiped with sides parallel to the coordinate axes, and center at the observation.

In work which Stone has cited, Louis Gordon and I study universally consistent (in Bayes risk) rules for classification, rules which also depend on certain rectangular parallelopipeds, or boxes. The rules we discuss, which are derived largely from those of Anderson [1], of Morgan and Sonquist (see [6]), and especially of Friedman [4], all involve successive partitioning of boxes by hyperplanes parallel to the coordinate axes. The rules of Friedman, for example, partition a box on that axis and at a point so as to effect the greatest reduction in the Kolmogorov–Smirnov distance between the two within-box marginal distributions. All three classes of rules must be supplemented so as to guarantee that ultimately, arbitrarily often each box is partitioned on each axis near the center of the box. All the rules Gordon and I discuss are invariant rules, and our proofs cover the case where the sizes of the two training samples are chosen by the experimenter.

Friedman shows [4] that for a variety of problems, his rules are computationally preferable to simple nearest neighbor classification in terms of average decision time, error rate and amount of memory used to store information needed to implement the rule.

REFERENCES

[1] ANDERSON, T. W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks. *Multivariate Analysis*, (P. R. Krishnaiah, ed.) Academic Press, New York.

[2] COVER, T. M. (1968). Estimation by the nearest neighbor rule. *IEEE Trans. Information Theory* **14** 50–55.

[3] COVER, T. M. and HART, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Information Theory* **13** 21–27.

[4] FRIEDMAN, J. H. (1976). A variable metric decision rule for nonparametric classification. *IEEE Trans. Computers* **25**. To appear.

[5] GORDON, L. and OLSHEN, R. A. (1976). Asymptotically efficient solutions to the classification problem. Unpublished.

[6] SONQUIST, J. (1970). *Multivariate Model Building: The Validation of a Search Strategy*. Institute for Social Research, The Univ. of Michigan, Ann Arbor.

EMANUEL PARZEN

State University of New York at Buffalo

In my discussion of Charles Stone's significant paper on consistent estimators of conditional expectations and conditional quantiles, I would like to introduce

an approach which emerges out of my recent work on "time series theoretic nonparametric statistical methods."

Let (X, Y) be a pair of continuous random variables of which one has observed a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$. One desires to estimate the conditional expectation $E(Y \mid X = x)$, the conditional distribution $F_{Y|X}(y \mid x) = P(Y \leqq y \mid X = x)$, and the conditional quantile function $Q_{Y|X}(p \mid x) = F^{-1}_{Y|X}(p \mid x) = \inf\{y : F_{Y|X}(y \mid x) \geqq p\}$.

The intuitive approach to the estimation of these parameters is what could be called the "histogram" approach; to the X_j's in a neighborhood of x (satisfying, say, $|X_j - x| \leqq h$ for a suitably determined "bandwidth" h) there is a corresponding set of Y_j's obtained as the second component of the observations (X_j, Y_j). The mean and distribution function of this set of Y_j values is a "histogram" estimator of the conditional mean and distribution function of Y given $X = x$. This approach has two basic drawbacks: how to chosen h, and the estimated functions may not be as smooth functions of x as we may have reason to believe the true functions are. To help overcome these problems, Stone considers estimators of the form $\hat{E}(Y \mid X = x) = \sum_{i=1}^n W_{ni}(x) Y_i$. However, it is not clear to me whether Stone's suggestions for the construction of the weights $W_{ni}(x)$ are useful in practice.

More importantly, I do not believe that "universally consistent weights" are what is wanted in practice. I believe that what is desired are weights that are chosen adaptively by the sample to provide "asymptotically efficient" estimators. Many theorems remain to be proved before this goal can be rigorously attained but I believe I can propose a formula for estimators which will have such properties.

Let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ be the order statistics of the Y values, and $\tilde{F}_X(x)$ denote the empirical distribution function of the X-values. I propose the estimator

$$\hat{E}(Y \mid X = x) = \sum_{j=1}^n Y_{(j)} \hat{w}_j(\tilde{F}_X(x)) ;$$

the weight $\hat{w}_j(u)$, $0 \leqq u \leqq 1$, is a ("time series theoretic") estimator, based on the entire sample, of

$$w_j(u) = H_1\left(u, \frac{j}{n}\right) - H_1\left(u, \frac{j-1}{n}\right),$$

where $H_1(u_1, u_2)$ is a distribution function defined in the next paragraph. In the case that $X = (X_1, \cdots, X_d)$ is a d-vector, there are functions $\hat{w}_j(u_1, \cdots, u_d)$ estimated from the entire sample such that the proposed estimator is of the form

$$\hat{E}(Y \mid X_1 = x_1, \cdots, X_d = x_d) = \sum_{j=1}^n Y_{(j)} \hat{w}_j(\tilde{F}_{X_1}(x_1), \cdots, \tilde{F}_{X_d}(x_d)) .$$

Let $F(x, y)$, $F_X(x)$, $F_Y(y)$, $Q_X(u)$, $Q_Y(u)$ denote respectively the joint distribution function of X and Y, the individual distribution functions of X and Y, and the quantile functions of X and Y where $Q_X(u) = F_X^{-1}(u)$. Define new random variables U_X and U_Y satisfying

$$U_X = F_X(X) , \qquad U_Y = F_Y(Y) , \qquad X = Q_X(U_X) , \qquad Y = Q_Y(U_Y) .$$

U_X and U_Y are individually uniformly distributed over the unit interval $0 \leqq u \leqq 1$; denote their joint distribution and density functions by $H(u_1, u_2)$ and $h(u_1, u_2)$ respectively. Explicitly

$$H(u_1, u_2) = F(Q_X(u_1), Q_Y(u_2))$$
$$h(u_1, u_2) = \frac{f(Q_X(u_1), Q_Y(u_2))}{f_X(Q_X(u_1))f_Y(Q_Y(u_2))} .$$

The conditional probability density of U_Y given U_X satisfies

$$f_{U_Y|U_X}(u_2 \mid u_1) = h(u_1, u_2) ;$$

therefore

$$E[Y \mid X = x_1] = E[Q_Y(U_Y) \mid U_X = u_1 = F_X(x_1)]$$
$$= \int_0^1 Q_Y(u_2)h(F_X(x_1), u_2)\, du_2$$
$$F_{Y|X}(y \mid x_1) = \int_0^{F_Y(y)} h(F_X(x_1), u_2')\, du_2'$$
$$Q_{Y|X}(p \mid x_1) = Q_Y H_1^{-1}(F_X(x_1), p)$$

defining

$$H_1(u_1, u_2) = \int_0^{u_2} h(u_1, u_2')\, du_2' = \frac{\partial}{\partial u_1} H(u_1, u_2) .$$

The distribution function H and its derivative can be "optimally" estimated using time series theoretic methods, starting with the raw estimators

$$\tilde{H}(u_1, u_2) = \tilde{F}(\tilde{Q}_X(u_1), \tilde{Q}_Y(u_2))$$
$$\tilde{\varphi}(v_1, v_2) = \int_0^1 \int_0^1 \exp\{2\pi i(u_1 v_1 + u_2 v_2)\}\, d\tilde{H}(u_1, u_2)$$

for $0 \leqq u_1, u_2 \leqq 1$, $v_1, v_2 = 0, \pm 1, \pm 2, \cdots$. A naive "$k$-nearest neighbor" estimator of $H_1(u_1, u_2)$ is

$$\hat{H}_1(u_1, u_2) = \frac{n}{2k} \left\{ \tilde{H}\left(u_1 + \frac{k}{n}, u_2\right) - \tilde{H}\left(u_1 - \frac{k}{n}, u_2\right) \right\} .$$

M. ROSENBLATT

University of California at San Diego

The results that Stone has obtained on consistency of regression estimates and estimates of conditional quantities are certainly very interesting and relate to many problems that are currently under study. It would seem to be important to get more detailed insight into the local and global behavior of some of these estimates, particularly in terms of their asymptotic distribution and bias. Results of this type have been obtained for a variety of density and regression estimates (the paper of Bickel and myself [1] contains a few of these results). The nearest neighbor regression estimates have attractive features in terms of consistency in view of Stone's Theorem 2. However, nearest neighbor density estimates appear to have disadvantages under certain circumstances (see the paper of Fukunaga and Hostetler [3] and comments on their work in Friedman's paper [2]). It is

suggested that possible difficulties are due to the bias of the estimate in the tail of the distribution. One hopes that there will be further work on the large sample behavior of the class of estimates that Stone has discussed as well as on the computational ease of using such estimates and their stability.

REFERENCES

[1] BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann Statist.* **1**, 1071–1095.

[2] FRIEDMAN, J. H. (1974). Data analysis techniques for high energy particle physics. SLAC Report No. 176.

[3] FUKUNAGA, K. and HOSTETLER, L. D. (1973). Optimization of kth nearest neighbor density estimates. *IEEE Trans. Information Theory* **IT-19** 320–326.

JEROME SACKS

Northwestern University

The requirement in Theorem 2 that the weights be nonnegative may not be a drawback when no smoothness is assumed about $f(x) = E(Y \mid X = x)$ but it is often restrictive when some smoothness can be assumed. For example, when $d' = d = 1$, X has compact support and x_0 is an endpoint of the support then the use of nonnegative weights results in weighting values of f at x's which lie on one side of x_0 and there is no way of effectively using the smoothness of f to reduce the resulting bias. Indeed, it is shown in Sacks and Ylvisaker [1] that, if

$$|f(x) - f(x_0) - f'(x_0)(x - x_0)| \leqq M_1(x) \tag{1}$$

where M_1 is specified and $M_1(x) = o(|x - x_0|)$ near x_0 then the set of weights which minimizes the maximum (over all f's satisfying (1)) of the mean-square-error will usually contain some negative ones. Nonnegative weights will often suffice if x_0 lies closer to the center of the support of X and will always suffice if (1) above is replaced by the assumption $|f(x) - f(x_0)| \leqq M_0(x)$ for some specified M_0.

The rate of convergence of the estimators treated by Professor Stone will depend on the smoothness of f and reasonable rates cannot be expected without smoothness (e.g., it is roughly true (from [1]) that $n^{\frac{4}{5}}E(\hat{f}_n(x) - f(x))^2$ is bounded in n and f if (1) holds for the optimum $\hat{f}_n$). It is possible that the type of modification proposed by Professor Stone in Section 4 may be particularly valuable when the f's involved have some smoothness. The modification in Section 4 also creates weights which depend on the location x which is an advantage not possessed by the nearest-neighbor weights of Theorem 2.

REFERENCE

[1] SACKS, J. and YLVISAKER, D. (1976). Linear estimation for approximately linear models. Discussion Paper 9, Center for Statistics and Probability, Northwestern Univ.

GRACE WAHBA

University of Wisconsin at Madison

I am sure all the discussants join me in thanking Professor Stone for an interesting and thought provoking paper. I will restrict my remarks to the problem of estimating $E(Y \mid X = x) \equiv f(x)$, certainly an important problem. It is commendable that Professor Stone was able to obtain convergence properties of $\hat{E}_n(Y \mid X = x)$ under very weak assumptions. By making regularity assumptions on f and the distribution of X, one can go much further—one can obtain (quadratic mean) convergence rates, and furthermore can obtain empirical Bayes estimates for the (minimum integrated mean square error) bandwidth parameter when the estimates $\hat{E}_n(Y \mid X = x) = \sum W_{n,i}(x)Y_i$ turn out to be kernel-type. A modest example of this was kindly referenced by Professor Stone [37], but I would like to indicate some of the more general results that can be obtained.

We may write, for any $X = x$,

$$(Y \mid X = x) \equiv Y(x) = f(x) + \varepsilon(x)\,,$$

where, for each fixed x, $f(x) = E(Y \mid X = x)$, $E\varepsilon(x) = 0$ and the $\varepsilon(x)$ are independent for distinct x. I will assume that Y is R^1-valued, that X has a density $h(x)$ which is strictly positive on a known, closed, bounded subset T of R^d and 0 elsewhere, and that $E\varepsilon^2(x) \equiv E\{(Y \mid X = x) - E(Y \mid X = x)\}^2 = \sigma^2\delta(x)$ where $\delta(x)$ is a known sufficiently nice function. The parameter σ^2 may be unknown. A general regularity condition that allows extension of Professor Stone's results, is, that $f \in \mathscr{H}_Q$, where $\mathscr{H}_Q$ is a reproducing kernel Hilbert space of real valued functions on T, with continuous reproducing kernel $Q(s, t)$. With these assumptions families of estimates $\hat{E}_n(Y \mid X = x)$ of Professor Stone's form

$$\hat{E}_n(Y \mid X = x) = \sum_{i=1}^n W_{n,i}(x; X_1, \cdots, X_n)Y_i \tag{1}$$

can be generated by letting $\hat{E}_n(Y \mid X = x) = f_{n,\lambda}(x)$, where $f_{n,\lambda}$ is the solution to the problem: Find $f \in \mathscr{H}_Q$ to minimize

$$\frac{1}{n}\sum_{i=1}^n \frac{(Y_i - f(X_i))^2}{\delta(X_i)} + \lambda\|f\|_Q^2\,, \tag{2}$$

where $\|\cdot\|_Q$ is the norm in $\mathscr{H}_Q$, and λ is the "smoothing" or "bandwidth" parameter. The solution $f_{n,\lambda}$ is given [4] by

$$f_{n,\lambda}(x) = (Q(x, X_1), \cdots, Q(x, X_n))(Q_n + n\lambda D_n)^{-1}(Y_1, \cdots, Y_n)'\,, \tag{3}$$

where Q_n is the $n \times n$ matrix with jkth entry $Q(X_j, X_k)$, and D_n is the $n \times n$ matrix with diagonal entries $\delta(X_j)$. The right hand side of (3) clearly is of the form (1). If ε were Gaussian, a Bayesian could construct $f_{n,\lambda}(x) \equiv \hat{E}_n(Y \mid X = x)$ of (3) as $f_{n,\lambda}(x) = E(f(x) \mid Y(x_i) = Y_i, i = 1, 2, \cdots, n)$ by adopting the Gaussian prior on f with $Ef(x) = 0$, $\operatorname{Cov} f(u)f(v) = bQ(u, v)$, $\lambda = \sigma^2/nb$.

Returning to a fixed, unknown f, the parameter λ controls the bias-variance

tradeoff for the mean square error $R(\lambda)$,

$$R(\lambda) = E\left\{\frac{1}{n}\sum_{i=1}^{n}(f_{n,\lambda}(X_i) - f(X_i))^2 \mid X_i = x_i,\ i = 1, 2, \cdots, n\right\}$$

$$\equiv E\left\{\frac{1}{n}\sum_{i=1}^{n}(\hat{E}_n(Y \mid X = x_i) - E(Y \mid X = x_i))^2\right\}.$$

and, roughly speaking, λ plays the same role as k in the k-NN examples cited by Professor Stone. To have a practical method, one must have a prescription for choosing k (or λ). (The correct choice of the "bandwidth" parameter is more important than the choice of the "shape" provided the "shape" is in an appropriate class.) It can be deduced from hypothesis (3) of Professor Stone's Theorem 1 that rather weak requirements on the "bandwidth" parameter suffice to insure consistency; but with the correct choice, sharper results can be obtained, as I shall show, and furthermore, λ in the estimate (3) can be chosen by empirical Bayes methods from the data.

The problem of choosing λ in (3) is essentially the same problem as choosing the ridge parameter λ in a ridge estimate β_λ of β in the standard regression model $y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$, where β_λ is the solution to the problem: Find $\beta \in E_p$ to minimize $(1/n)||y - X\beta||_n^2 + \lambda||\beta||_p^2$ (Euclidean n and p norms). See [8] and just about any recent issue of JASA, Technometrics, Communications in Statistics or JRSS-B for a discussion of this issue! To avoid inessential complications, I now let $\delta(x) \equiv 1$ and condition on $X_i = x_i$ where the sample c.d.f. of the x_i's coincides with the true c.d.f. at $x = x_i$. Then $R(\lambda)$ may be written

$$R(\lambda) = E\frac{1}{n}||A(\lambda)\bar{Y} - \bar{f}||_n^2 \equiv E\frac{1}{n}||(I - A(\lambda))\bar{f} - A(\lambda)\bar{\varepsilon}||_n$$

$$= \frac{1}{n}||(I - A(\lambda))\bar{f}||_n^2 + \frac{\sigma^2}{n}\operatorname{Trace} A^2(\lambda),$$

where $A(\lambda) = Q_n(Q_n + n\lambda I)^{-1}$, $\bar{Y} = (Y_1, \cdots, Y_n)'$, $\bar{f} = (f(x_1), \cdots, f(x_n))'$, and $\bar{\varepsilon} = (\varepsilon(x_1), \cdots, \varepsilon(x_n))'$. An unbiased estimate $\hat{R}(\lambda)$, for $R(\lambda)$ may be obtained from Mallows [10] or Hudson [3] if σ^2 is known, and is

$$\hat{R}(\lambda) = \frac{1}{n}||(I - A(\lambda))\bar{Y}||_n^2 - \frac{2\sigma^2}{n}(\operatorname{Tr}(I - A(\lambda)) + \sigma^2.$$

If σ^2 is known, it is reasonable to take the minimizer of $\hat{R}(\lambda)$ as a good choice of λ. If σ^2 is not known, my favorite estimate of λ is the generalized cross-validation estimate, which is the minimizer of

$$V(\lambda) = \frac{\bar{Y}'(I - A(\lambda))^2\bar{Y}}{(\operatorname{Tr}(I - A(\lambda))^2}.$$

See [8] for the source of this estimate. It can be shown [2, 7] that for any $f \in \mathscr{H}_Q$, the minimizer of $EV(\lambda)$, call it $\tilde{\lambda}$, satisfies $\tilde{\lambda} = \lambda^*(1 + o(1))$, where λ^* is the minimizer of $R(\lambda)$, and $o(1) \to 0$ as $n \to \infty$. The convergence result that is available concerns the convergence rate of the mean square error $R(\lambda)$ at its

minimizer $\lambda = \lambda^*$. Suppose $f \in \mathscr{H}_{Q^*Q}$, the reproducing kernel Hilbert space with reproducing kernel $(Q^*Q)(s, t) = \int_T Q(s, u)Q(t, u)\, du$, and h is a constant; then it can be shown (see [7]) that

$$R(\lambda) \leqq \lambda^2 \|f\|^2_{Q^*Q} + \frac{\sigma^2}{n} \sum_{\nu=1}^{\infty} \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right)^2 \tag{4}$$

where $\{\lambda_\nu\}$ are the eigenvalues of the Hilbert–Schmidt operator with Hilbert–Schmidt kernel Q. For example, if $T = [0, 1]$ and $\mathscr{H}_Q$ is a space of functions $\{g : g, g', \cdots, g^{(m-1)}$ abs. const., $g^{(m)} \in \mathscr{L}_2[0, 1]\}$ then, roughly, $f \in \mathscr{H}_{Q^*Q}$ entails that $f^{(2m)} \in \mathscr{L}_2[0, 1]$, and $\lambda_\nu = O(\nu^{-2m})$ and the second term on the right of (4) is $O(1/n\lambda^{1/2m})$. The right-hand side of (4) is then minimized for $\lambda^{**} = \text{const}\,((\sigma^2/\|f\|^2_{Q^*Q})(1/n))^{2m/(4m+1)}(1 + o(1))$ and it follows that $R(\lambda^*) \leqq R(\lambda^{**}) = O(n^{-4m/(4m+1)})$. (It can be shown that $R(\lambda^*) = R(\lambda^{**})[1 + o(1)]$.) See [7] for details. This kind of argument also appears in [1]. It appears that $R(\lambda^*) = O(n^{-4m/(4m+1)})$ can be obtained if h is any "nice" strictly positive density, see [2]. For $T = [0, 1] \times [0, 1] \times \cdots \times [0, 1]$, d times, one can let $\mathscr{H}_Q$ be the d-fold tensor product of d one dimensional spaces (see [6]); more interesting spaces can be found in the approximation theory literature. The eigenvalues associated with tensor product spaces are the tensor products of the one dimensional eigenvalues ($\lambda_{\mu\nu} = \lambda_\mu \lambda_\nu$).

The estimates of the form (3) generally do not give us k-NN type estimates, since, loosely speaking, the weight given to Y_i in $\hat{E}_n(Y \mid X = x)$ in (3) depends on the distance x is from X_i, rather than how many neighbors are "between" x and X_i. Loosely speaking, it can be shown (see [6, 8]) that

$$f_{n,\lambda}(x) \approx \sum_{\nu=1}^{n} \hat{f}_\nu \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right) \phi_\nu(x)$$

where $\{\phi_\nu\}$ are the eigenfunctions associated with the eigenvalues $\{\lambda_\nu\}$ and $\hat{f}_\nu$ is an estimate of $f_\nu = \int_T f(x)\phi_\nu(x)\, dx$. Then, roughly,

$$\hat{f}_\nu \simeq \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_\nu(x_i) h^{-1}(x_i)\,,$$

where h is the density (or an empirical density) of the $\{x_i\}$. Then

$$f_{n,\lambda}(x) \simeq \sum_{i=1}^{n} Y_i K_\lambda(x, x_i)\,,$$

where

$$K_\lambda(x, y) = \frac{1}{n} \sum_{\nu=1}^{n} \left(\frac{\lambda_\nu}{\lambda_\nu + \lambda}\right) \phi_\nu(x)\phi_\nu(y) h^{-1}(y)\,.$$

If $\mathscr{H}_Q$ is the Hilbert space of periodic functions on $[0, 1]$ with, for example, $\|f\|_Q^2 = [\int_0^1 f(u)\, du]^2 + \int_0^1 [f^{(m)}(u)]^2\, du$, and $h(u) = 1$, then the eigenvalues are $\lambda_0 = 1$, $\lambda_\nu = (2\pi\nu)^{-2m}$, and for large n, it can be shown that

$$\begin{aligned} K_\lambda(x, y) &\simeq \frac{1}{n}\frac{1}{(1 + \lambda)} + \frac{1}{n} \sum_{\nu=-\infty;\nu\neq 0}^{\infty} \frac{1}{(1 + (2\pi\nu)^{2m}\lambda)} e^{2\pi i\nu(x-y)} \\ &\simeq \frac{1}{n}\frac{1}{(1 + \lambda)} + \frac{1}{n\lambda^{1/2m}} k\left(\frac{x - y}{\lambda^{1/2m}}\right) \end{aligned}$$

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.