Tech. Rept. 734, Dept. of Statistics, Univ. of Wisconsin, Madison.

OWEN, A. (1983). The estimation of smooth curves. Unpublished manuscript.

REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **10** 177–183.

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

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

STONE, C. J. and KOO, C.-Y. (1986a). Additive splines in statistics. *Proc. Statist. Comp. Sect. Amer. Statist. Assoc.*

TIBSHIRANI, R. (1984). Local likelihood estimation. Stanford technical report and unpublished Ph.D. dissertation, Dept. of Statistics, Stanford Univ.

WAHBA, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Proc. Conf. on Approximation Theory in Honour of George Lorenz, Jan. 8–10, Austin, Texas* (W. Chaney, ed.). Academic, New York.

WAHBA, G. and WOLD, S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation. *Comm. Statist.* **4** 1–7.

YOUNG, F. W., TAKANE, Y. and DE LEUUW, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika* **43** 279–282.

Comment

David R. Brillinger

> "All considered, it is conceivable that in a minor way, nonparametric regression might, like linear regression, become an object treasured for both its artistic merit as well as usefulness."
>
> L. Breiman (1977)

This paper by Hastie and Tibshirani lays bare the insight of the above remark of Leo Breiman made in the course of the discussion of a seminal work on regression with smooth functions (Stone, 1977). Here Hastie and Tibshirani increase the store of both artistic merit and usefulness by plugging nonparametric regression into the generalized linear model and by alluding to a variety of possible further extensions. It all makes being a statistician these days a joy—it seems approaches are now available to attack most any applied problem that comes to hand. (Understanding the operational performance of those approaches is clearly another matter however.)

It was nice to be asked to comment on such a stimulating paper. I have divided my comments into several sections, striving to focus on individual strains present in the paper, believing that future research on those strains will proceed at different rates.

1. STRUCTURE OF A BASIC PROBLEM

One has data $(Y_i, \mathbf{X}_i)$, $i = 1, \cdots, n$, with n moderately large. One is willing to consider a model for the individual Ys wherein: i) the conditional distribution of Y given $\mathbf{X}$ belongs to an exponential family, ii) it involves $\mathbf{X}$ only through $\eta = \sum s_j(X_j)$ with the $s_j(\cdot)$ unknown, but smooth, and iii) $E\{Y \mid \mathbf{X}\} = h(\sum s_j(X_j))$, with $h(\cdot)$ known. The parameter of the model is $\theta = \{s_j(\cdot), j = 1, \cdots, p\}$, and possibly a scale. The two key elements of the model are a) that the $s_j(\cdot)$ are smooth and b) that $\sum s_j(X_j)$ is additive.

It is to be noted that this model continues the contemporary statistical trend to eliminate distinctions between the cases of finite and infinite dimensional θ or between discrete and continuous data.

The problem is of interest, for one may wish to make inferences from the data via the model or one may wish to validate a model with a low dimensional parameter by imbedding it in a broader model, for example.

2. CONSTRUCTION OF ESTIMATES

To begin, focus on estimating $\eta = \eta(\mathbf{X})$, via a relationship that characterizes the true value η_0. Suppose one has a function $\rho(Y \mid \eta)$ such that $E_0\{\rho(Y \mid \eta) \mid \mathbf{X}\}$ is maximized at $\eta = \eta_0$. An example would be $\log f(Y \mid \eta)$, $f(\cdot)$ denoting the conditional density of Y. Alternately, suppose one has a function $\psi(Y \mid \eta)$ such that $E_0\{\psi(Y \mid \eta) \mid \mathbf{X}\} = \mathbf{0}$ at $\eta = \eta_0$. An example would be $\partial \log f(Y \mid \eta)/\partial\eta$. Estimates of the true η_0 may be constructed by paralleling these relations on the data. For example, given weights $W_{ni}(\mathbf{X})$ such as in Stone (1977) one might take $\hat{\eta}$ to maximize

$$\sum_i \rho(Y_i \mid \hat{\eta}) W_{ni}(\mathbf{X})$$

or to satisfy

$$\sum_i \psi(Y_i \mid \hat{\eta}) W_{ni}(\mathbf{X}) = \mathbf{0}.$$

The estimate of Hastie and Tibshirani based on (26) takes this form. One can expect such estimates to be

David R. Brillinger is Professor of Statistics, University of California, Berkeley, California 94720.

consistent under regularity conditions. Stone (1977, page 643) gives some simple conditions. Such estimates were called conditional M-estimates by Brillinger (1977) and it was remarked there that one could form robust estimates directly (by limiting the influence of individual observations for example). It is further clear that partial likelihood estimates, censored data estimates, and unequal probability of selection estimates are particular cases.

The critical advance of Hastie and Tibshirani is to look for extrema with η of the form $\sum s_j(X_j)$. They limit consideration to likelihood- and partial likelihood-based estimates, but it is clear that they could go on to form for example robust-resistant ones by choice of ρ or ψ.

It is further apparent that were the dimension of $\mathbf{X}$, p, unclear one could add an Akaike type term in p and estimate p as well. Continuing, this makes it apparent that penalized maximum likelihood estimates also may be fit into this general setup. We have here a type of inverse unstable problem. These are often solved by forms of regularization (smoothing). It is perhaps worth remarking that the first approach above is a form of Courant regularization, while penalized likelihood would correspond to Tihonov regularization. (These techniques are discussed in Allison (1979).)

There is much insight in Hastie and Tibshirani's remark that because of the additivity of η in the $s_j(\cdot)$, the smoothing need not be local (in the $\mathbf{X}$-space).

3. COMPUTATIONS

In the next few years, the structure set out in the preceding section may not be expected to change too much. This is probably not true for the algorithms numerically determining the extrema.

Hastie and Tibshirani propose an iteratively reweighted least squares solution, as in GLIM, interwoven with a stepwise selection procedure as in Breiman and Friedman (1985). My experience with such algorithms is that they are troubled by initial values, precision/round-off, convergence criterion, underflow/overflow, and instability among other things. Nonlinear iterations can do strange things. In particular I expect better algorithms for determining the components of $\sum s_j(\cdot)$ to be developed.

4. SOME QUIBBLES

I do have some disagreements with the paper. In the abstract, it is stated: "It has the advantage of being completely automatic . . ." I see this as both a disadvantage and not true. A disadvantage because surely one wants flexible analyses. Not true because someone (the programmer?) has made many choices: machine precision, convergence criterion, smoother, ... The analyst will not know these choices at his peril.

In Section 7 it is stated: "This is the chief motivation for the additive model." The reason given is a statistical one. To my mind the motivation is substantive. Additivity is basic to science (see Luce and Tukey, 1964, particularly the references therein).

Two medical data sets are analyzed, but no inferences are made. Can the authors not set down some (biological) insight or understanding that has been gained from the analyses? Otherwise they might have just as well presented the results of simulations.

5. FURTHER ISSUES AND PROBLEMS

In this section I am not complaining about possible omissions from the present paper, rather I am interested in the authors' thoughts regarding future directions of work. The paper certainly stands on its own.

The statistical properties of the estimates need to be understood. What are they actually estimating in the case of a finite sample? In time series we know that the conventional spectrum estimate is estimating an average of the power spectrum, albeit concentrated near the frequency of concern. Is that the case here or are remote values influential? The time series case further suggests the possible utility of pretransforming the Xs to reduce bias.

The sampling variability of the estimates need to be assessed. Could the authors indicate their preferred technique. Mine would be a jackknife variant, because of its bias reducing properties and nonmodel dependence. There is a need for goodness of fit/validation procedures, diagnostics, measures of influence.

In power spectrum estimation, I do not generally take the same bandwidth for all frequencies in the conventional estimate (and more complex estimators have a similar effect). Here the span is taken to be the same. Have the authors thought of making it variable?

Smoothness is essential in the development in the paper. Yet many natural relationships are discontinuous and even multivalued. It would seem appropriate to develop techniques for such situations. For the former, perhaps one would smooth only when an estimate of the derivative is small.

6. A QUESTION

In Section 1, the authors refer to the ACE procedure of Breiman and Friedman (1985) as a means of determining a transform of the *dependent* variable. This involves maximizing a correlation. In Section 9, they refer to the use of local scoring (i.e., a likelihood-based technique) for the analogous problem of determining the link function. The two criteria are quite different seemingly. Can the authors comment? I wonder about yet another alternative, namely picking the transformations to maximize a nonparametric estimate of the

mean information in $\eta(\mathbf{X})$ about $\theta(Y)$. (This does not involve a Jacobian.)

ADDITIONAL REFERENCES

ALLISON, H. (1979). Inverse unstable problems and some of their applications. *Math. Sci.* **4** 9–30.

BREIMAN, L. (1977). Discussion of Consistent nonparametric regression, by C. J. Stone. *Ann. Statist.* **5** 621–622.

LUCE, R. D. and TUKEY, J. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *J. Math. Psych.* **1** 1–27.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

Comment

J. A. Nelder

I congratulate the authors on a fascinating piece of work and offer three comments.

1. In order to make smoothing work it is necessary to restrict it to one-dimensional covariate spaces, hence the strong assumption of additivity. In principle one could introduce cross-terms, e.g., have $x_{12} = x_1 x_2$, as well as x_1 and x_2, in the model; however, I suspect the convergence of the algorithm might now become immensely slow or even nonexistent because of the functional relations between the covariates. An alternative might be to include a term of the form $s_1(x_1) \cdot s_2(x_2)$, with coefficient to be estimated. Have the authors any comments on this problem?

J. A. Nelder is Visiting Professor, Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7, England.

2. To me it seemed intuitively surprising that the figures in Table 2 show the generalized additive model to have one parameter more than the original parametric one, but a deviance nearly 6 higher. I then realized that the latter has a cross-term in it, and this appears to be important. What would be the effect of adding a term in $s_1(x_1) \cdot s_2(x_2)$ to the former? Also it would help interpretation if the difference in deviance were given when each term in their model was replaced by a parametric form. This would give summary statistics for differences visible in Figures 3, 4, and 5.

3. The new version of GLIM (3-77) now available has a facility for inserting new code. I very much hope that the authors can be persuaded to exploit this in order to make available the fitting of generalized additive models in GLIM.

Comment

Charles J. Stone

Hastie and Tibshirani deserve commendation for the originality, significance, and interest of their approach and the excellent expository review in the present paper.

Recently I have been working on a different approach to fitting more or less the same class of models, but using polynomial cubic splines to model the component functions $s_j(\cdot)$ and the Newton–Raphson method to calculate the ordinary maximum likelihood estimate. In order to avoid artificial end effects of polynomial fits such as those shown in Figures 2 and 3, the splines are constrained to be linear to the left of the first knot and to the right of the last knot. To avoid multiple representations of the constant term, zero sum constraints are imposed on the individual terms (when $p \geq 2$), as is done in this paper. Thus, if there are N knots, there are $N + 4$ degrees of freedom for the unconstrained spline and $N - 1$ degrees of freedom for the constrained spline. There is also 1 degree of freedom for the constant term; so there are $(N - 1)p + 1$ degrees of freedom in total. This approach will be referred to as the parametric spline approach to distinguish it from the smoothing spline approach favored by Wahba and others in which smoothing is achieved by a roughness penalty instead of by confining attention to spline models with a modest number of degrees of freedom. In theory, N should tend to infinity as the sample size n tends to

Charles J. Stone is Professor of Statistics, University of California, Berkeley, California 94720.