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 $x_{12} = x_1x_2$, 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?

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.

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

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.

infinity so as to achieve the optimal rate of convergence (see Stone (1985, 1986)). Asymptotically optimal rules for selecting N based on the data have been obtained by Burman (1985). In practice $N = 5$ has proven sufficient. This is not surprising, since the standard linear approach allows only 1 degree of freedom per component function. Allowing 4 degrees of freedom should provide enough flexibility to fit the regular departures from nonlinearity that are likely to occur in practice, especially when linear constraints are used in the tails. For the flexibility is then highest in that portion of the axis that contains the bulk of the data. Linear restrictions on splines lead to tail behavior very similar to that of the linear smoothers (local linear regression) recommended by Stone (1975, 1977), Cleveland (1979), Friedman and Stuetzle (1981), and this paper. As Hastie and Tibshirani and others have pointed out, it is desirable to have a reasonable automatic default rule. The rule that has emerged from Stone and Koo (1986b) is this: given a specific covariate, order its observed values as $x_{(1)}, \cdots, x_{(n)}$; put knots at the minimum value $x_{(1)}$ and maximum value $x_{(n)}$; put additional knots at $x_{(i)}$, $i = i_2, i_3, i_4$, chosen so that the logits of $1/(n+1)$, $i_2/(n+1)$, $i_3/(n+1)$, $i_4/(n+1)$, $n/(n+1)$ are approximately equally spaced.

In the few cases where the two approaches have been applied to the same data, the resulting curves appeared visually to be quite similar (see Stone and Koo, 1986a, and Devlin and Weeks, 1986), except that the approach of Hastie and Tibshirani leads to small scale roughness not present in curve estimates obtained by the parametric spline approach. The approaches seem equally feasible numerically and equally automatic. But the parametric spline approach has several conceptual advantages. In particular, the standard maximum likelihood method can be used to estimate the parameters and obtain confidence intervals that are asymptotically valid, at least when N is fixed. The χ^2 approximation to the asymptotic distribution of the logarithm of likelihood ratio statistics is also asymptotically valid with an integral number of degrees of freedom. The theory is analytically tractable even when $N \longrightarrow \infty$ as $n \longrightarrow \infty$, provided that the covariates are restricted to a compact set. Undoubtedly, Hastie and Tibshirani could site advantages for their approach.

In order to carry out the asymptotics for the parametric spline approach when $N \longrightarrow \infty$ as $n \longrightarrow \infty$, it seems necessary that the log likelihood function be strictly concave. Such concavity holds in generalized additive models when $\eta = \theta$ and for some other choices of the link function (such as that corresponding to probit models), but it is not true for an arbitrary link function. Strict concavity is desirable even when N is fixed, for it guarantees that the log likelihood function have at most one local maximum and that a local maximum, if it exists, be the unique global maximum. Hastie and Tibshirani do not explicitly mention strict concavity, which in their notation amounts to the requirement that $d^2l/d\eta^2 < 0$. Even without this requirement, it is true that $E(d^2l/d\eta^2 \mid x) < 0$, but perhaps the algorithm in (23) of this paper is more reliable when the log likelihood function is strictly concave.

Generalized additive modeling as studied by Hastie and Tibshirani, by Burman, and by myself is an extension of the generalized linear models (GLMs) introduced by Nelder and Wedderburn (1972). But, so far at least, one limitation of GLMs has been preserved; namely, the restriction to exponential models that involve a one-dimensional parameter θ. The most obvious practical advantage of considering a multidimensional parameter θ is that the setup would then include multinomial models for conditional distributions and thereby allow for categorical response variables Y having more than two possible categories. Once covariates are included we have a natural setup for developing reasonable and flexible multiple classification procedures. In the linear form of the model, each coordinate of θ would be a linear function of the covariates. In the additive extension, each coordinate would be an additive function of the covariates. Ideally, the fitting procedure should be such that the estimated conditional probabilities of the various categories are positive and sum to one. This can undoubtedly be done with the parametric spline approach. Can it also be done with the approach of Hastie and Tibshirani?

In the present paper, Hastie and Tibshirani treat Cox's proportional hazards model as being outside the framework of GLMs. However, the logarithm of the partial likelihood is of the form log-PL $= \sum_{i\in D} [\beta x_i - \log(\sum_{j\in R_i} e^{\beta x} j)]$. For each i, the expression enclosed by brackets is exactly in the form of a multinomial model, there being as many categories as there are elements in the risk set R_i. Thus the setup is essentially that of independent but not identically distributed multinomial response experiments. In particular, log-PL is a strictly concave function of the unknown parameters, so the parametric spline approach should also be viable. But the asymptotics, especially when $N \longrightarrow \infty$ as $n \longrightarrow \infty$, have yet to be worked out.

ADDITIONAL REFERENCES

BURMAN, P. (1985). Estimation of generalized additive models. To appear in *J. Multivariate Anal.*

DEVLIN, T. F. and WEEKS, B. J. (1986). Spline functions for logistic regression modeling. In *Proceedings of the Eleventh Annual SAS Users Group International Conference*, 646–651. SAS Institute, Inc., Cary, N. C.

STONE, C. J. (1975). Nearest neighbor estimators of a nonlinear regression function. In *Proceedings of Computer Science and*

Statistics: 8th Annual Symposium on the Interface, 413–418. Health Sciences Computer Facility, UCLA.

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

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

STONE, C. J. and KOO, C.-Y. (1986b). *Function Estimates.* AMS Contemporary Math. Ser., Amer. Math. Soc., Providence, R. I.

Comment

Peter McCullagh

Hastie and Tibshirani are to be congratulated for presenting the theory and methodology of generalized additive models in a form that keeps incidental mathematical details at an acceptably low level. I have little to add and my single comment is therefore brief.

The whole thrust of the authors' development seems to be based implicitly on the following assumption, here reduced to the bare essentials: zero interaction is fundamentally more plausible than componentwise linearity in the covariates. Has there been any attempt to justify this point of view, either philosophically or empirically by examining a large number of examples or by any other means? A closely related question concerning statistical strategy is the following: at what stage of analysis does the assumption of zero interaction come under scrutiny?

Peter McCullagh is Professor of Statistics, Department of Statistics, The University of Chicago, 5734 University Avenue, Chicago, Illinois 60637.

Rejoinder

Trevor Hastie and Robert Tibshirani

1. THE GENERAL PROBLEM

In Section 5 of the paper, we motivated the local scoring and local likelihood estimation procedures as empirical methods for maximizing $E(l(\eta(X), Y))$. In the two procedures, the maximization problem is approached in different ways. In the *local likelihood* method, an estimate of $E((l(\eta(X), Y) \mid X = x))$ is constructed (for each x) and this has the form $(1/k_n)\cdot\sum_{j\in N_i} l(\eta(x_j), y_j)$ given in (26) of the paper. As Brillinger notes (his Section 2), one can generalize this and hence include robust estimates and many others.

On the other hand, the local scoring procedure maximizes $E(l(\eta(X), Y))$ by estimating the quantities in the update expressions (22) and (36). Note, however, that this procedure is not expressible as a maximation of the kind that Brillinger describes, i.e., a maximization of a function of the form $\sum_i \rho(Y_i \mid \hat{\eta}) W_{ni}(X)$. However, it is possible to write down a finite sample justification of local scoring (to answer a question of Brillinger's) based on the notion of penalized likelihood. This justification applies only in the special case in which the local scoring algorithm uses linear smoothers. Recall that a linear smoother is one for which the result of smoothing a vector $\mathbf{z}$ can be written simply as $\hat{\mathbf{z}} = S\mathbf{z}$, for some matrix S, called a "smoother matrix." Now suppose we have data $(y_1, x_{11}, x_{12}, \cdots, x_{1p}), \cdots, (y_n, x_{n1}, x_{n2}, \cdots, x_{np})$ and let S_j be the smoother matrix for the jth variable. Let $\mathbf{s}_j = (s_1(x_{1j}), s_2(x_{2j}), \cdots, s_n(x_{nj}))^t$, $j = 1, 2, \cdots, p$ and consider the following problem. Find $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_p$ to maximize

$$l(\boldsymbol{\eta}) - \frac{1}{2}\sum_{1}^{p} \mathbf{s}_j^t(S_j^- - I)\mathbf{s}_j \tag{1}$$

where $\boldsymbol{\eta} = \alpha + \sum_1^p \mathbf{s}_j$ and S_j^- is a generalized inverse of S_j. Then it is easy to show that the local scoring procedure is a Fisher scoring step for maximizing (1) (see Hastie and Tibshirani, 1986a, for details). Now a typical smoother matrix is close to symmetric, has eigenvectors that are close to polynomials, and has eigenvalues that tend to decrease with increasing order of the eigenvector. Hence, the penalty term in (1) puts greater penalty on the higher order polynomial components of each $\mathbf{s}_j$. There is also a close tie here to smoothing splines. If we start with a penalty of the form $\sum_1^p \lambda_j \mathbf{s}_j^t K_j \mathbf{s}_j$, where K_j is an appropriate quadratic penalty matrix, we derive a local scoring procedure that uses cubic spline smoothers. Hence, there is close relation of local scoring to the work of O'Sullivan, Yandell, and Raynor (1986), Green (1985), and Green and Yandell (1985). These authors consider a