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

penalized likelihood approach, with emphasis on quadratic penalties leading to spline smoothing, as above. None of these authors use backfitting type algorithms, however, because their models contain only a single smooth function or surface (in addition to parametric terms) and hence backfitting is not required.

Following Brillinger's comment, we note that the local scoring procedure can also be used for robust estimation. For a general ψ function, the local scoring step equivalent to (22) is

$$\eta^1(x) = E\left[\eta(x) - \frac{\psi(Y \mid \eta)}{E(d\psi/d\eta \mid x)} \,\middle|\, x\right]. \tag{2}$$

As before, the conditional expectation is estimated by a smoother, and for multiple covariates, η might be an additive function.

Stone's parametric splines can be cast in the same setting. The spline fit on each covariate can be written as a linear operation and hence his model could be fit with a local scoring algorithm performing the appropriate linear fit at each smoothing step. By the results quoted in Section 7, this procedure would converge to the maximum likelihood estimates of the functions; this would, however, be a very inefficient way to fit the model, since it can be solved efficiently via the usual iterative methods.

A disadvantage of this global minimization framework is that it doesn't incorporate nonlinear smoothers. These include variable span smoothers (e.g., "supersmoother," Friedman and Stuetzle, 1982) and the "split-linear smoother" (MacDonald and Owen, 1984) for capturing discontinuities. These and other nonlinear smoothers would be useful for capturing the irregular or discontinuous functional behavior that Brillinger mentions, but we've had only a limited experience with them so far. We have incorporated a cubic spline smoother into the lastest version of GAIM (thanks to Finbarr O'Sullivan for his code) and have been happy with the results. As a final point, we reiterate another nice feature of local scoring: one is free to choose different smoothers for different covariates. Hence, one could use a spline smoother for one covariate, a parametric spline for a second covariate, a variable span smoother for a third covariate, and so on. Of course, straight line fits and categorical variables can also be used, resulting in a rich class of models.

2. ADDITIVITY AND INTERACTION

Drs. Brillinger, McCullagh, and Nelder bring up the question of additivity and interaction. It is difficult to come up with a clear definition for the latter; one possibility is to define interaction as being the lack of fit in a standard (componentwise) linear model. Phrased in this way, nonlinearity in a covariate is a kind of interaction. We suspect that McCullagh is referring to a more restrictive form of interaction, something like product interactions of two or more variables. We don't feel that zero interaction (of this latter sort) is "fundamentally more plausible" than componentwise linearity; instead, we view the method that we have presented for estimating nonlinearities as just another tool for detecting departures from the linear model. McCullagh's question concerning when (in an analysis) interaction should come under scrutiny is a deep one that we don't know how to answer. To further stress the difficulty of this question, we note that transformations of Y are another way to model certain kinds of product interactions on the original Y scale. The overall goal of all these tools is to find simple departures from a componentwise linear model; developing an effective strategy for this is a challenging and important problem.

We do want to emphasize that simple interactions can be incorporated in a generalized additive model. These include interactions of the form $\beta x_1 x_2$, $s(x_1 x_2)$, and $\beta \hat{s}_1(x_1) \cdot \hat{s}_2(x_2)$ (suggested by Nelder), where $\hat{s}_1(x_1)$ and $\hat{s}_2(x_2)$ are known functions, possibly obtained from an additive fit. We don't feel (as Nelder does) that convergence of local scoring will be a problem in these cases; it will simply take longer to converge as the constructed variables become more correlated.

More recently, we have experimented with the use of two-dimensional smoothers to fit surfaces more general than an additive one. Figure 1B illustrates such a surface.

The variables income (of the head of household) and age are two of a number of variables used to model the proportion of families having a telephone at home (the data is part of a telephone survey kindly furnished by Ed Fowlkes). The terms $s_1(\text{Inc}) + s_2(\text{Age})$ were included in an additive logistic model, together with several other variables. Figure 1A gives the additive surface defined by these two fitted functions. Figure 1B shows the estimated interaction surface $s_{1,2}(\text{Inc}, \text{Age})$. This was estimated by using a (kernel) surface smoother within the local scoring algorithm for this pair. The single function for income was quadratic, whereas in Figure 1B we see that that the income effect appears mostly monotone (except for a dip around the middle ages) and levels off at higher ages (and thus higher incomes). This leveling off goes unnoticed in the additive function model; rather it simply dampens the overall effect. This example illustrates the fact that an additive model can give us a reasonable idea of what is going on, while finer details can be discovered by fitting more general models.

3. COMPUTATIONAL CONSIDERATIONS

Brillinger reports many problems with iteratively reweighted least squares algorithms, and while we don't doubt that better procedures will be developed,

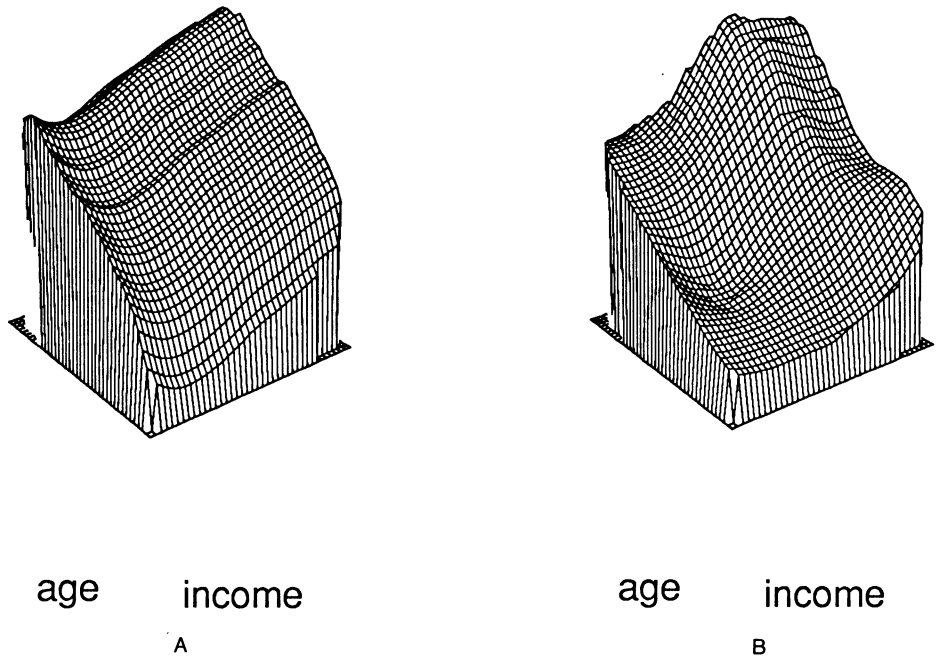

FIG. 1. *(A) Additive surface defined by* $s_1(\text{Inc}) + s_2(\text{Age})$. *This gives an idea of the types of surfaces produced by additive models. (B) Interactions surface* $s_{1,2}(\text{Inc, Age})$ *reveals two-dimensional features not captured by the additive model. The surface was estimated using a two-dimensional kernel smoother within the local scoring algorithm.*

we have had few difficulties with the present algorithm. Brillinger later told us that this problems occurred in special models that incorporated random effects, and perhaps this added complexity caused some of the difficulty.

In the local scoring algorithm, many variations are possible, in terms of the order of the smoothing and updating operations. In some early experimentation we had convergence problems with one variant. We chose the present method because it converged well in practice and because it reduces to Fisher scoring when linear fitting is used.

More recently, we have developed (with Andreas Buja, 1986) a new version of backfitting in which the linear components for all the variables are all fit in a separate projection. We have been able to prove convergence of this modified algorithm when a practical smoother like a cubic spline is used, something that neither we nor Breiman and Friedman (1985) have been able to do for the present algorithm. This algorithm should also be much more efficient computationally.

We note that in general, one has a choice of observed or expected information in the local scoring procedure; these correspond to Newton-Raphson and Fisher scoring, respectively. (In the exponential family with canonical link, they are the same.) In the paper, we used observed information for the general case, but we haven't yet studied this choice. Nor have we thought enough about concavity of the log likelihood, as mentioned by Stone.

Brillinger's points about an "automatic" algorithm are well taken. We were referring to the fact that our procedures eliminate some of the detective work necessary for finding nonlinearities from partial residual plots.

Finally, in response to Nelder's request for a GLIM version of the algorithms, we think we have a simple method for implementation via the new PASS facility that he is alluding to, but neither of us have the 3-77 version of GLIM and are looking forward to receiving it (for UNIX machines).

4. DIAGNOSTICS AND TOOLS FOR INFERENCE

Brillinger mentions the need for fit/validation procedures, diagnostics, and measures of influence. As mentioned in Section 9 and demonstrated in the examples, we have developed a notion degrees of freedom or "effective number of parameters," following that of Cleveland (1979). This is useful for assessing the importance of model terms. We also have a fairly simple way of estimating pointwise confidence bands for the estimated functions, if the smoothers used are linear. These are based on $\pm$ twice a local measure of standard deviation. See Hastie and Tibshirani (1984, 1985c) for further details of both the above techniques. Resampling methods, as suggested by Brillinger, would be another approach. An example of this is given in Efron and Tibshirani (1986), but we haven't yet studied this problem in detail.

The local scoring algorithm is not very robust to

outliers and making the smoothers robust would not solve the problem completely, if more than one covariate is present. What is needed is another outer loop in which points are downweighed based on the current fit; however, this may be computationally formidable. As far as diagnostics are concerned, Buja, Donnell, and Stuetzle (1986) have studied the analogous problem to collinearity in additive models (they call it "cocurvity"). This and much more work is needed to develop for additive models a diagnostic "black bag" like the one available for linear models.

5. THE DATA ANALYSES

In the two examples of our paper, we are guilty, as Brillinger points out, of brushing over the scientific aspects of the problem at hand. We will briefly try to make amends here. In the first example, the smooth in Figure 3 is interesting because it shows a plateau around age 50, something oncologists call the "Clemmenson hook." In the Cox model example, an interesting result was the disagreement, between the parametric and nonparametric analyses, as to whether the relative risk dropped or was about constant between ages 10 and 40. Further investigation (see Efron and Tibshirani, 1986) suggested that there was insufficient data in this age range to decide the issue.

To answer Nelder's questions on the first example, the addition of a term $\beta x_1 x_2$ to the generalized additive model did not significantly reduce the deviance, although it was significant when added to the parametric model. On Nelder's suggestion, we tried adding the term $\beta \hat{s}_1(x_1) \cdot \hat{s}_2(x_2)$ to the model, $\hat{s}_1(x_1)$ and $\hat{s}_2(x_2)$ being the functions from the generalized additive fit. This produced a drop in deviance of only 1.4. We also tried replacing each smooth by the corresponding parametric fit, as suggested by Nelder. The drops in deviance were 5.7, 3.6, and .01 on 1.7, 1.5, and 1.4, respectively. Hence, only the function for age is significantly better than its parametric fit.

For a more thorough data analysis using generalized additive models, we refer the reader to Hastie and Tibshirani (1985a and 1985c).

6. RELATED WORK AND EXTENSIONS

Stone discusses another approach to generalized additive model estimation, namely the use of fixed knot "parametric splines." His method does have the conceptual and mathematical advantages that he mentions, but practically speaking, we worry about the task of picking the number and position of the knots. How much does this choice effect the appearance of the final estimate? When many covariates are present, should the knots be chosen in some way to account for the other variables in the model? Another closely related approach, is that of smoothing splines, mentioned in Section 3 of this discussion. A comparative study of all these methods would be very useful.

Stone mentioned multiparameter models. We have, in fact, generalized the logistic model to incorporate ordered categorical responses (Hastie and Tibshirani, 1986b). We adapted the proportional odds model of McCullagh (1980):

$$\text{logit}[P(Y \le k \mid \mathbf{x})] = \alpha_k - \sum_{j=1}^{p} f_j(x_j), \quad k = 1, 2, \cdots, K-1, \tag{3}$$

where the response Y has K categories. The model essentially says that the histogram for the response categories shifts with the covariates according to $\eta(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j)$. We use the multinomial likelihood for estimation. The appropriate local scoring algorithm has an additional loop; we alternate between estimating the $K-1$ constants by weighted averages of $K-1$ adjusted dependent variates, and the additive functions by backfitting on a *scalar* linear combination of adjusted dependent variates. The model (3) can also be used when a continuous response has been categorized, and thus fills the gap between the extreme 0–1 response logistic regression model and the *continuous* response ordinary regression model.

Brillinger's final question concerning ACE and generalized additive models is a fascinating one. We would like to take this opportunity to clarify the relationship between the methods and report some current research. First note that, as alluded to in Section 9 of the paper, the local scoring algorithm can be used to estimate any function that appears in a model, not just a function of a covariate. We simply add a step like (22) to the algorithm for that function. Thus for example, we can estimate a link function (see Hastie and Tibshirani, 1984) or a variance function (Hastie and Pregibon, 1986). This fact will become important below.

Now consider the Gaussian additive model $E(Y \mid \mathbf{X}) = \alpha + \sum_1^p s_j(X_j)$. (We'll relate our comments to nonGaussian generalized additive models as we go along.) Two ways to extend this model are to allow a transformation of the mean, i.e., $E(Y \mid \mathbf{X}) = f(\alpha + \sum_1^p s_j(X_j))$ or a transformation of the response, i.e., $E(\theta(Y) \mid \mathbf{X}) = \alpha \sum_1^p s_j(X_j)$. The former has been looked at by Friedman and Owen (1986) and is a special case of link function estimation for generalized additive models via local scoring. The second model is the transformation model, for which Breiman and Friedman's (1985) ACE algorithm provides a method for estimation. The two models are not the same, even if $\theta(\cdot)$ is forced to be monotone. That is, we should not expect that $\hat{\theta}^{-1}(\cdot)$ will be close to $\hat{f}(\cdot)$ for a given data set.

Brillinger's question concerns two possible methods for estimating the functions of the transformation model. The ACE algorithm maximizes the correlation of the transformed variables. Brillinger's suggestion is to instead maximize the likelihood of the untransformed variables, by direct analogy to the parametric method of Box and Cox (1964). This likelihood would include a Jacobian, as Brillinger states, to account for the transformation $\theta(\cdot)$. One can carry through Brillinger's suggestion using the local scoring algorithm: unfortunately, the resultant algorithm requires estimates of the second and third derivatives of $\theta(\cdot)$. While we haven't tried it yet, our guess is that the algorithm might be unstable because of this.

Another approach to this problem, similar to Brillinger's suggestion, is given by Tibshirani (1986). He proposes an algorithm in which a (nonparametric) variance stabilizing transformation is used to estimate $\theta(\cdot)$. The procedure is called "RACE" for regression ACE. In both simulated and real data examples, he demonstrates that RACE eliminates many of the anomalies of ACE, in particular, sensitivity to the marginal distribution of the X's. RACE is likely to produce similar results (qualitatively) to Brillinger's suggestion, because the effect of the Jacobian is mainly to force $\theta(Y)$ to have constant variance (see Box and Cox (1964) and Tibshirani (1984, Remark F)).

A transformation of the response might also be useful in other generalized additive models, such as a Poisson model for categorical data. Marhoul (1984) looks at a related technique.

ACKNOWLEDGMENTS

We were fortunate to receive the comments of four such highly respected statisticans as Drs. Brillinger, McCullagh, Nelder, and Stone. They raise a number of important issues, some that we've thought about since the writing of the paper, but many others that we haven't yet resolved. We would like to thank them for their efforts and we would also like to thank Morris DeGroot for his editorial work. In our rejoinder we have tried to clarify and expand on some of these questions.

ADDITIONAL REFERENCES

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–252.

BUJA, A., DONNELL, D. and STUETZLE, W. (1986). Additive principal components. Manuscript in preparation.

EFRON, B. and TIBSHIRANI, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy (with discussion). *Statist. Sci.* **1** 54–77.

FRIEDMAN, J. H. and OWEN, A. (1986). Predictive ACE. Unpublished manuscript.

GREEN, P. J. (1985). Penalized likelihood for general semi-parametric regression models. Tech. Rept. 2819, Dept. Statistics, Univ. Wisconsin, Madison.

HASTIE, T. and PREGIBON, D. (1986). Manuscript in preparation.

HASTIE, T. and TIBSHIRANI, R. (1985c). Generalized additive models: some applications. Tech. Rept. 14, Dept. Statistics, Univ. Toronto.

HASTIE, T. and TIBSHIRANI, R. (1986a). Generalized additive models, cubic splines and penalized likelihood. Tech. Rept., Biostatistics Group, Univ. Toronto.

HASTIE, T. and TIBSHIRANI, R. (1986b). Nonparametric logistic and proportional odds regression. To appear in *Appl. Statist.*

HASTIE, T., TIBSHIRANI, R., and BUJA, A. (1986). Manuscript in preparation.

MACDONALD, J. and OWEN, A. (1984). Smoothing with split linear fits. Report LCS07, Dept. Statistics, Stanford Univ.

MARHOUL, J. (1984). A model for large sparse contingency tables. Report LCS13, Dept. Statistics, Stanford Univ.

O'SULLIVAN, F., YANDELL, B. and RAYNOR, W. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–103.

TIBSHIRANI, R. (1986). Estimating optimal transformations for regression: a variation on ACE. Tech. Rept. 1986-001, Biostatistics Group, Univ. Toronto.