WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

ANDREAS BUJA
DIANE DUFFY
BELLCORE
445 SOUTH STREET
BOX 1910
MORRISTOWN, NEW JERSEY 07962-1910

TREVOR HASTIE
AT & T BELL LABORATORIES
600 MOUNTAIN AVENUE
MURRAY HILL, NEW JERSEY 07974-2070

ROBERT TIBSHIRANI
DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO
CANADA M5S1A8

FINBARR O'SULLIVAN[1]
University of Washington

This article reviews a set of key developments in nonparametric function estimation, many of them due in part or in large to Professor Friedman, which have radically changed the scope of modern statistics. MARS is an impressive addition to this set. There is a growing practical interest in innovative adaptive function estimation techniques. For example, I am aware of the need for sophisticated covariate adjustment in connection with survival analysis of a large clinical trail, where $N = 27{,}000$ and $n \geq 200$; the thought of sending these data to MARS for analysis will have undoubted appeal!

1. General comments. With any adaptive regression technique, it is of interest to know the kinds of functions which cause greatest difficulty. MARS is coordinate-sensitive. A rotation of the coordinate axes in the examples in Sections 4.2 and 4.3 will destroy the simple additive and low-order interactive structure. Will this substantially degrade the performance (ISE) of MARS? Perhaps the effect could be ameliorated by allowing linear combination splits in the algorithm. A natural set of split coordinates would be those obtained by successive orthogonally restricted regression of residuals r at the Mth order model on the covariates: The linear combination c_1 determining the first split coordinate solves the least-squares regression of r on covariates, the linear combination c_2 determining the second split coordinate solves the least-squares regression of r on covariates but subject to the orthogonality constraint $c_2'c_1 = 0$ and so on. The relevant formulas are available in Seber ([4], pages 84–85). Algorithm 2 only requires a minor change to incorporate consideration of linear combination splits. Obviously it would no longer make sense to have a

[1] Research supported in part by Dept. of Energy Grant FG0685-ER2500 and by National Cancer Institute Grant 2P01-CA-42045.

constraint on the order of interaction $k \leq K_m$, but it would perhaps be natural to put a constraint on the number of split coordinates to be examined. The rapid updating formulae in equation (52) does not apply but for split coordinate and knot optimization it should be adequate to compute the lack of fit in the innermost loop of Algorithm 2 by leaving $a_1, a_2, \dots, a_{M-1}$ provisionally fixed and minimizing only over a_M and a_{M+1}. Optimal coefficients can be evaluated after completing the inner loop.

With a modification of this type and with more elaborate function estimation algorithms, the problem arises of how to interpret/visualize the nonparametric regression surface $\hat{f}$. The output will not be a simple sum of first, second and higher order interaction terms, so the attractive decomposition in equation (24) will not be available. However, numerical integration can of course be used to obtain a decomposition in terms of variables of interest. For example, if the x-variables are split as $x = (x_1, x_2)$, then

$$\hat{f}(x) = \hat{f}_1(x_1) + \hat{f}_2(x_2) + \hat{f}_{12}(x_1, x_2),$$

where $\hat{f}_1(x_1) = \int_{x_2} \hat{f}(x)\,dx$, $\hat{f}_2(x_2) = \int_{x_1} [\hat{f}(x) - f_1(x_1)]\,dx$ and $\hat{f}_{12}(x_1, x_2) = \hat{f}(x) - \hat{f}_1(x_1) - \hat{f}_2(x_2)$. The percent variance explained by these orthogonal components would be of interest.

A further visualization tool, focusing on isolating local collinearity-type effects, could be obtained by applying multivariate statistical density exploration procedures, such as clustering and principal component projections, to the x-distribution associated with specified levels of $\hat{f}$. For example, the analysis of the distribution of x-values for which $a \leq \hat{f}(x) \leq b$ would be of interest. Function visualization is an area where there is a growing need for better statistical tools.

The MARS algorithm offers considerable power particularly in situations where there are additive low-order nonlinear interactions. One of the motivations for MARS given in the paper is dissatisfaction with the lack of continuity in CART. I will finish by briefly describing an alternative continuous modification to CART which retains some of its algorithmic and interpretative simplicity.

2. Smoothed CART by finite elements. The CART model in (17) is represented as

$$\hat{f}(x) = \sum_{m=1}^{M} a_m B_m(x),$$

where $B_m = I_m$, the indicator function for an n-dimensional rectangular region R_m. Replace the indicator function by a smooth element $I_m(x, s) \geq 0$ whose support is allowed to extend beyond R_m and define a smoothed CART model by

$$\hat{f}(x; s) = \sum_{m=1}^{M} a_m B_m(x; s).$$

Here $B_m(x; s)$ is forced to satisfy a local partition of unity by setting

$$B_m(x;s) = \frac{I_m(x;s)}{\Sigma_{m'} I_{m'}(x;s)},$$

so $\Sigma_{m=1}^{M}(B_m(x; s) = 1$. I require that $I_m(x, s) > 0$ for $x \in R_m$. I have introduced a parameter s which gives control over smoothness. Models of the previous form connect with mixture models used in the interpretation of multichannel image data; see Smith [5], Choi, Haynor and Kim [1] and O'Sullivan [2], for example.

Laplacian finite elements based on triangular grids have been extensively analyzed in the approximation theory literature; see the references in Schumaker [3]. With the rectangular grids of CART, a reasonable choice for $I_m(x; s)$ is defined by tensor products of coordinate functions.

$$I_m(x;s) = \prod_{j=1}^{n} w_{mj}(x_j;s),$$

where $w_{mj}(x_j; s)$ is a smooth nonnegative function whose support for $s > 0$ will extend beyond the projection of R_m onto the jth coordinate. Specifically, suppose the set of split points on the jth variable are $t_j^{(k)}$ for $k = 1, 2, \dots, K_j$ and the projection of R_m is $[t_j^{(k_m)}, t_j^{(k_m+1)})$. Then $w_{mj}(\cdot; s)$ can be a B-spline basis element, of any specified order, supported on $[t_j^{(k_m-[s])}, t_j^{(k_m+1+[s])}] \cap [t_j^{(1)}, t_j^{(K_j)}]$. Here $[s]$ is the closest integer to s. If we use cubic order elements, then $\hat{f}$ will have continuous second-order mixed partial derivatives.

The smoothed version of CART, call it SCART ("scairt" is the Irish word for a bush or bushy place!), is easily computed. Let $0 \leq p \leq 1$ be given. The algorithm applies partitioning and pruning as in CART with a couple of minor modifications: (i) For M fixed, the tree predictions are $\hat{f}(\cdot; s)$ with $s = pK_j$ and the coefficients a_m optimized by least squares (likelihood can also be used). (ii) At stage M, the selection of the potential split point for R_m is done to improve the local fit. Thus if r are the residuals from the Mth order model, then the algorithm just applies the CART splitting rule to components of these residuals lying in R_m. The local support of $I_m(x; s)$ must be exploited for rapid computation of $\hat{f}$. Cross-validation is used to compare trees for different values of p. A preliminary least squares version of SCART with piecewise linear elements was developed and applied to some of the examples in the paper—those used to compute Tables 4, 7, 9 and 11. The ISE was evaluated and compared to that achieved by MARS. MARS is a clear winner for the additive model in equation (56) and the additive model with the single low-order interaction in equation (61). For example, with $N = 200$, the ISE obtained by SCART was on the order of 0.17 so MARS is 90% better here. SCART wins on the alternating current impedance example in equation (63a) with a 50% or better improvement in the ISE at all sample sizes. A smaller improvement between 10–30% is achieved by SCART on phase angle data in equation (63b).

I suppose the message here is that no single adaptive regression technique can perform uniformly best on all examples, which echoes the point made by Professor Friedman in Section 2.

REFERENCES

[1] Choi, H. S., Haynor, D. R. and Kim, Y. (1989). Multivariate tissue classification of MRI images for 3-D volume reconstruction—a statistical approach. *SPIE Medical Imaging III: Image Processing* **1092** 183–193.

[2] O'Sullivan, F. (1990). Mixture models for multi-channel image data. Technical report, Dept. Statist., Univ. Washington.

[3] Schumaker, L. L. (1976). Fitting surfaces to scattered data. In *Approximation Theory II* (G. G. Lorentz, C. K. Chui and L. L. Schumaker, eds.) 203–268. Academic, New York.

[4] Seber, G. A. F. (1976). *Linear Regression Analysis*. Wiley, New York.

[5] Smith, M. O., Adams, J. B. and Johnson, P. (1986). Spectral mixture modelling, a new analysis of rock and soil types at the viking lander 1 site. *J. Geophys. Res.* **91** 8098–8112.

Department of Biostatistics and Statistics
University of Washington
Seattle, Washington 98195

Art Owen

Stanford University

I like MARS. It looks like a good tool for pulling out the most useful parts of large interaction spaces. Most of my comments are directed at accounting issues: How many degrees of freedom are used in knot selection? How can the cost be lowered? At the end, there are some comments on how one might apply MARS to models for which fast updating is not available.

My main interest in MARS stems from work in computer experiments. In these applications, smooth functions of fairly high complexity are evaluated over high dimensional domains with no sampling error. I plan to use MARS on such functions evaluated over Latin hypercube designs [McKay, Conover and Beckman (1979)]. Some theory for linear modeling of nonrandom responses over such designs is given in Owen (1990).

When there is no noise, one expects that a larger number of knots might be warranted. It then becomes worthwhile to lower the price of a knot somehow.

Degrees of freedom in broken line regression. Consider the broken line regression model

$$Y_i = b_0 + b_1 t_i + \beta(t_i - \theta)_+ + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where $t_1 \le t_2 \le \cdots \le t_n$ are nonrandom with $\sum t_i = 0$ and $\sum t_i^2 = n\sigma_t^2$, ε_i are independent $N(0,1)$ and b_0, b_1, β and θ are parameters. Taking $\beta = 0$ in (1) yields a one-segment model. Taking $\beta \ne 0$ and $t_1 < \theta < t_n$ yields a two-segment model. This model has been studied by Feder (1967), Hinkley (1969),