the modified basis function are not understood. Do they approximate with the same power as linear splines? Surely they not do as well as quadratic ones. My next remark relates to the basis functions being used. As noted in Section 3.9, the one-sided truncated power basis is well known to be very badly conditioned whereas the classical B-splines are very well-conditioned. Why not use the latter? Updating might even be easier.

The idea of simplifying the model by removing knots (recombining pieces) strikes me as very important. This idea has recently been discovered by approximation theorists in connection with general spline fitting. The papers [8]–[10] are representative.

REFERENCES

[1] Birman, M. S. and Solomiak, M. E. (1966). Approximation of the functions of the classes W_p^α by piecewise polynomial functions. *Soviet Math. Dokl.* **7** 1573–1577.

[2] Brudnyi, Ju. A. (1971). Piecewise polynomial approximation and local approximations. *Soviet Math. Dokl.* **12** 1591–1594.

[3] de Boor, C. and Rice, J. R. (1979). An adaptive algorithm for multivariate approximation giving optimal convergence rates. *J. Approx. Theory* **25** 337–359.

[4] DeVore, R. A. and Popov, V. A. (1987). Free multivariate splines. *Constructive Approx.* **3** 239–248.

[5] Franke, R. and Schumaker, L. L. (1987). A bibliography of multivariate approximation. In *Topics in Multivariate Approximation* (C. K. Chui, L. L. Schumaker and F. Utreras, eds.) 275–335. Academic, New York.

[6] Light, W. A. and Cheney, E. W. (1985). *Approximation Theory in Tensor Product Spaces. Lecture Notes in Math.* **1169**. Springer, New York.

[7] Light, W. A. and Cheney, E. W. (1989). On the approximation of a bivariate function by the sum of univariate functions. *J. Approx. Theory* **29** 305–322.

[8] Lyche, T. and Mørken, K. (1987). Knot removal for parametric B-spline curves and surfaces. *Computer-Aided Geometric Design* **4** 217–230.

[9] Lyche, T. and Mørken, K. (1987). A discrete approach to knot removal and degree reduction algorithms for splines. In *Algorithms for Approximation* (J. C. Mason and M. G. Cox, eds.) 67–82. Oxford Univ. Press, New York.

[10] Lyche, T. and Mørken, K. (1988). A data reduction strategy for splines. *J. Numer. Anal.* **8** 185–208.

[11] Popov, V. (1989). Nonlinear multivariate approximation. In *Approximation Theory VI* (C. K. Chui, L. L. Schumaker and J. D. Ward, eds.) 519–560. Academic, Boston.

Department of Mathematics
Vanderbilt University
Nashville, Tennessee 37240

Charles J. Stone

University of California, Berkeley

This pioneering paper successfully combines creative breakthroughs (especially, *not* removing the parent basis function) with numerous techniques developed over the years by the author and his collaborators and others (especially Patricia Smith). The tactics of MARS are specific to least-squares estimation of a regression function, but the general strategy is much more widely applicable.

In Algorithm 2, the residual sum of squares could be used for LOF. This would allow the entering of new basis functions to be broken up into two steps: decide which basis function next to enter into the model; enter the selected basis function. Using the "t to enter" algorithm from stepwise regression would then eliminate the need to fit the various candidate models in the course of determining the next basis function to enter.

Algorithm 3 could be divided into three tasks: decide which basis function next to remove from the model; remove the selected basis function; determine the final model. In the first two of these steps, the residual sum of squares could be used for LOF. Using the "t to remove" algorithm would then eliminate the need to fit the various candidate models in the course of determining the next basis function to remove.

With these changes, MARS could be extended to handle logistic regression in a more natural manner. The "t to enter" step in Algorithm 2 could be replaced by an algorithm based on Rao's score test and the "t to remove" step in Algorithm 3 could be replaced by an algorithm based on Wald's test. The actual maximum likelihood fitting (based, say, on the Newton–Raphson algorithm and taking advantage of the concavity of the log-likelihood function) would be applied $M_{\max}$ times in Algorithm 2, once after each application of the score test and $M_{\max}$ times in Algorithm 3. The final model selection could be based on a variant of AIC, modified along the lines of (32).

Consider now the estimation of an unknown density or probability function f on a set $\mathscr{Y}$. In order to guarantee positivity, we can model $\log(f)$ as a member of some adaptively selected space $\mathscr{S}$ that does not contain the constant functions. Letting $B_1, \dots, B_K$ be a basis of $\mathscr{S}$, we can write the estimate of f as $\hat{f} = \exp(\Sigma_k \hat{\theta}_k H_k - c(\hat{\theta}))$, where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_K)^t$ and $c(\hat{\boldsymbol{\theta}})$ is the normalizing constant. This has the form of an exponential family. If $\mathscr{Y} = \mathbb{R}$ and the functions in $\mathscr{S}$ have linear tails, then $\hat{f}$ has exponential tails. The selection of $\mathscr{S}$ could be done by applying the general strategy of MARS. Ignoring the model selection, we can choose $\hat{\boldsymbol{\theta}}$ by maximum likelihood. The asymptotic theory of such estimates, with $\mathscr{Y}$ a compact interval in $\mathbb{R}$ and $B_1, \dots, B_K$ consisting of B-splines and without model selection, has been treated in Stone (1990). The numerical behavior of such estimates, modified to incorporate stepwise knot deletion based on Wald's test and a variant of AIC, is studied in Kooperberg and Stone (1990).

Consider next a random pair $(\mathbf{X}, Y)$, where $\mathbf{X}$ is an n-dimensional random vector and Y is a $\mathscr{Y}$-valued random variable. Let $f(y|\mathbf{x})$ denote the conditional density or probability function of Y given that $\mathbf{X} = \mathbf{x}$. Consider an estimate $\hat{f}(y|\mathbf{x})$ of the form $\hat{f}(y|\mathbf{x}) = \exp(\Sigma_k \hat{\theta}_k(\mathbf{x}) B_k(y) - c(\hat{\boldsymbol{\theta}}(\mathbf{x})))$, where $B_1, \dots, B_K$ are suitable basis functions of a possibly adaptively selected space $\mathscr{S}$ of functions on $\mathscr{Y}$. [The functions in $\mathscr{S}$ should be piecewise linear for $c(\cdot)$ to be computed rapidly.] We could model $\theta_1(\cdot), \dots, \theta_K(\cdot)$ in turn as members of possibly adaptively selected spaces $\mathscr{H}_1, \dots, \mathscr{H}_K$, respectively. Letting H_{jk}, $1 \le j \le J_k$, be a basis of $\mathscr{H}_k$, we can write $\hat{\theta}_k(\mathbf{x}) = h_k(\mathbf{x}; \hat{\boldsymbol{\beta}}_k) = \Sigma \hat{\beta}_{jk} H_{jk}(\mathbf{x})$.

This leads to an estimate of $f(y|\mathbf{x})$ having the form

$$(1)\qquad \hat{f}(y|\mathbf{x}) = \exp\Bigl(\sum_{k}\sum_{j}\hat{\beta}_{jk}H_{jk}(\mathbf{x})B_k(y) - c\bigl(\mathbf{h}(\mathbf{x};\hat{\boldsymbol{\beta}})\bigr)\Bigr), \qquad y \in \mathscr{Y},$$

where $\hat{\beta}$ is the JK-tuple consisting of $\hat{\beta}_{jk}$, $1 \le k \le K$ and $1 \le j \le J_k$, in some order. This estimate has the form of a multiparameter exponential family, so the corresponding log-likelihood function is again concave. The asymptotic theory of such estimates, with $\mathscr{Y}$ a compact interval in $\mathbb{R}$, $\mathscr{H}_1 = \cdots = \mathscr{H}_K$ and bases consisting of B-splines and without model selection, has been treated in Stone (1989). It remains to investigate the numerical behavior of such estimates, especially as modified to incorporate the strategy of MARS. Perhaps the resulting technology should be referred to as multivariate adaptive response splines (MARES).

Suppose, in particular, that $\mathscr{Y} = \{0, 1\}$. Then we can let $\mathscr{S}$ be the one-dimensional space having basis $B_1(y) = y$. In this context, (1) reduces to logistic regression. Similarly, by letting $\mathscr{Y}$ be a finite set of size 3 or more, we can apply the strategy of MARS to the polytomous extension of logistic regression.

The more general setup given by (1) allows for the estimation of the conditional variance and conditional quantiles of an arbitrary random variable Y given $\mathbf{X}$ as well as estimation of the conditional mean of Y given $\mathbf{X}$, which is treated in the present paper.

The general strategy of MARS is also applicable to time series.

REFERENCES

KOOPERBERG, C. and STONE, C. J. (1990). A study of logspline density estimation. *Comput. Statist. Data Anal.* To appear.

STONE, C. J. (1989). Asymptotics for doubly-flexible logspline response models. *Ann. Statist.* To appear.

STONE, C. J. (1990). Large-sample inference for log-spline models. *Ann. Statist.* **18** 717–741.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

CHONG GU[1] AND GRACE WAHBA[2]

Purdue University and University of Wisconsin-Madison

We would like to begin by thanking Professor Friedman for a very interesting and thought-provoking paper. The idea of combining splines with recursive

[1] Research supported by National Sciences and Engineering Council of Canada while the author was at the University of British Columbia.

[2] Research supported in part by Air Force Office of Scientific Research Grant AFOSR 90-0103 and in part by NSF Grant DMS-90-02566.