HUTCHINSON, M. and BISCHOF, R. (1983). A new method for estimating the spatial distribution of mean seasonal and annual rainfall applied to the Hunter Valley, New South Wales. *Australia Meteorology Magazine* **31** 179–184.

NYCHKA, D. (1988). Confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.* **83** 1134–1143.

NYCHKA, D., WAHBA, G., GOLDFARB, S. and PUGH, T. (1984). Cross-validated spline methods for the estimation of three-dimensional tumor size distributions from observations on two-dimensional cross sections. *J. Amer. Statist. Assoc.* **79** 832–846.

O'SULLIVAN, F. (1990). An iterative approach to two-dimensional Laplacian smoothing with application to image restoration. *J. Amer. Statist. Assoc.* **85** 213–219.

POGGIO, T. and GIROSI, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science* **247** 978–982.

STEWART, G. W. (1987). Collinearity and least squares regression (with discussion). *Statist. Sci.* **2** 68–100.

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.

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

WAHBA, G. and WENDELBERGER, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review* **108** 1122–1145.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706

## REJOINDER

JEROME H. FRIEDMAN

*Stanford University*

I thank the editors for inviting such a distinguished group of researchers to discuss this paper and the discussants for their valuable contributions. The discussants are among the leaders in the field of function approximation and estimation; it is therefore no surprise that their comments are so perceptive and stimulating. Many important suggestions are made for improving the MARS procedure. These discussions provide a clearer and deeper understanding of both the strengths and limitations of the MARS approach. Each of them raises many very important issues, some of which I respond to here. Space limitations preclude a more thorough discussion of all of the cogent points and innovative ideas presented.

**Schumaker.** I thank Professor Schumaker for providing the additional references, especially the recent ones that were not available in 1987 when I performed the main body of this work. Of the five that relate to multivariate adaptive approximation, only one [de Boor and Rice (1979)] presents a procedure that could possibly lead to a practical method in high dimensions. Their

procedure is essentially recursive partitioning (15) using polynomials for the parametric functions $g_m(\mathbf{x}|\{a_j\}_1^p)$ in each subregion. The limitations of recursive partitioning methods are discussed in Section 2.4.2. It is perhaps interesting to note that in none of these papers was the method proposed therein ever actually applied to any (real or synthesized) multivariate problems.

In the asymptotic limit with no noise it may be true that only the continuity properties of the function to be estimated and the degree of the polynomials in each subregion limit the accuracy of a recursive partitioning scheme. As discussed in Section 2.2, all practical sample sizes are far from being asymptotic in high dimensional problems. In such cases there is a trade-off between the number of subregions and the maximum degree of the approximating polynomial in each one. A $q$-degree polynomial in $n$ dimension is characterized by more than $\binom{q+n-1}{n-1}$ parameters. Thus, even globally, there is a severe restriction on the degree of a polynomial fit. One can use the available degrees of freedom to fit moderate degree polynomials in a few regions or very low degree ones in many regions. Which is the best strategy in a particular application will depend upon the true underlying function $f(\mathbf{x})$ (1), but a great body of experience indicates that the latter approach gives the best results in general. In fact, the most popular choice with recursive partitioning methods is zero degree [AID, Morgan and Sonquist (1963) and CART, Breiman, Friedman, Olshen and Stone (1984)]. With this rather strong limitation on the allowable (and desirable) polynomial degree in each subregion, the lack of continuity of approximations derived from recursive partitioning methods does have an adverse effect on their accuracy, especially when compared to similar methods that produce continuous approximations. This is provided, of course, that the function to be approximated is relatively smooth.

The discussion concerning end effects in Section 3.7 centers on the bias-variance trade-off of estimates near the edges. In the mathematical approximation literature, the value of the function is generally assumed to be measured without error [$\varepsilon = 0$ (1)]; that is, all relevant variables that contribute to the variation of the response are available to help construct the approximation. In this case, there is no variance and only bias considerations are relevant. When there is error, the variance of the function estimate near the edges is a dominant concern. Having the estimate smoothly match a linear function near the edges removes only the first-order bias there, but has a strong moderating influence on the variance. In fact, even this precaution is sometimes inadequate and additional measures are required (see Section 3.8). As discovered by both Breiman (1989a) and Stone and Koo (1985), restricting the dependence of the approximation near the edges to be (at most) linear is an essential ingredient for controlling variance.

I agree that if one does not need $C^1$ functions, it is often better to use the approximation based on linear splines (see second paragraph of Section 3.7). The continuous derivative approximation often does better, but the effects is seldom dramatic, whereas the converse is not true. (See the example in Section 4.7 and the comments of Buja, Duffy, Hastie and Tibshirani.) When the

piecewise-linear approximation is doing dramatically better, this is evident from the estimated lack-of-fit (relative GCV scores). Again the converse is not necessarily true. Since the model optimization is based on the linear spline basis, the piecewise-cubic fit to the data is generally slightly (to moderately) worse than the corresponding piecewise-linear fit for comparable (future prediction) accuracy. I generally use the piecewise-cubic model if its optimized GCV score is no more than 10–20% worse than that for the linear spline fit. Otherwise, I prefer the piecewise-linear model. Of course one could use cross-validation to more directly judge the relative quality of the two models.

I also agree that the approximation properties of the modified (piecewise-cubic) basis functions are not yet well-understood. Whether they approximate as well as linear, quadratic or other splines will depend on the particular application. It seems reasonably clear that they will perform better than quadratic or higher order splines in noisy situations (with a smooth underlying function) owing to the end effects discussed earlier [Breiman (1989a) and Stone and Koo (1985)].

The reason for using the truncated power spline basis is given in Section 3.9; there is a one-to-one correspondence between knots and basis functions. This was the central ingredient in the (univariate) adaptive regression spline strategy originally proposed by Smith (1982). Lyche and Mørken (1988) propose essentially the same idea but use a $B$-spline implementation. As described in their paper, severe computational complexities are involved with the $B$-spline approach, as compared to the simple and elegant method proposed by Smith (1982). In the multivariate case, the analog of these increased computational complexities are likely to lead to insurmountable problems.

**Owen.** I am grateful to Professor Owen for contributing his valuable piece of research to the discussion of this paper. He has set down a theoretical framework that explains many of the results that I obtained empirically through Monte Carlo studies. More importantly, his work provides the understanding and insight necessary to suggest improvements to the simple model selection approach that I implemented.

The results contained in his last paragraph concerning the relative insensitivity of the fit to the precise locations of the placed knots are surprising and encouraging. I observed this phenomenon during empirical studies on varying the minimum span (43). The quality of the fit appeared to be remarkably insensitive to the value used provided it was not too small (in high-noise situations). Also, I implemented an algorithm for post knot optimization (model polish) given the topology of the model (23) derived form the forward/backward stepwise strategy. This algorithm minimized the lack-of-fit criterion simultaneously with respect to all the knot parameters $\{t_{km}\}$ in (23). This can be done surprisingly rapidly using a minor variation of an algorithm originally proposed by Breiman (private communication) in the context of recursive partitioning. Given the suboptimal way in which the knots are placed with the stepwise strategy, the expectation was that this model polish would

provide distinctly superior fits. This was not the case. In all situations in which I tried model polish, the resulting fits were changed very little from that provided originally by the stepwise algorithm. It should be noted, however, that the examples that I used in both studies (minimum span variation and model polish) involved very smooth functions with no really sharp local structure. For functions with such (sharp) structure, model polish may provide substantial improvement. Routinely applying model polish would then provide insurance against this possibility.

Owen's suggestion is to deliberately use a very coarse grid during the initial stepwise knot placement and then to rely on model polish to produce the final refined model. This will not improve speed for least-squares fitting since the updating formulae (52), in conjunction with suggestions in Stone's discussion, allow all possible (data point) knot locations to be evaluated with nearly the same computation as any subset of them. As Owen points out, however, for lack-of-fit criteria that do not admit fast updating, this strategy will produce dramatic computational savings, perhaps enough to make their routine application feasible.

I agree with Owen that MARS may have potential as a method for modeling noiseless data such as that produced by computer simulation experiments. Although it may produce adequate results in its present formulation, it is quite likely that its performance can be improved for this purpose by changing aspects of the implementation to take advantage of the knowledge that there is no noise (as noted by Owen). For example, the strong concern about end effect variance that motivated several important design decisions, could be relaxed and different strategies may be more appropriate in this case. Also, different model selection strategies are likely to help improve performance in these settings.

**Stone.** Stone's suggestions in the second and third paragraphs of his discussion turn out to be quite useful. The principal idea in the "$t$ to enter" algorithm is to orthogonalize both the response and the yet to be entered (predictor) variables with respect to each predictor as it is entered into the regression equation. In the context of MARS, the corresponding predictors are the basis functions derived during the forward stepwise procedure (Algorithm 2). Of course, during the execution of the forward stepwise algorithm all of the yet to be entered basis functions are not known. However, one can orthogonalize each basis function as it is entered with respect to all of those already selected, thereby keeping all currently entered basis functions orthonormal to each other. Let $B_M(\mathbf{x})$ be one of two (centered) next basis functions selected as a result of executing the outer loop of Algorithm 2. Then $\tilde{B}_M(\mathbf{x})$ and $c_M$ are saved, where

$$\tilde{B}_M(\mathbf{x}) = \left[B_M(\mathbf{x}) - \sum_{m=1}^{M-1} b_{Mm}\tilde{B}_m(\mathbf{x})\right] \Big/ b_{MM}^{1/2},$$

with

$$b_{Mm} = \sum_{i=1}^{N} B_M(\mathbf{x}_i)\tilde{B}_m(\mathbf{x}_i), \qquad 1 \le m \le M-1, \tag{75}$$

$$b_{MM} = \sum_{i=1}^{N}\left[B_M(\mathbf{x}_i) - \sum_{m=1}^{M-1} b_{Mm}\tilde{B}_m(\mathbf{x}_i)\right]^2$$

and

$$c_M = \sum_{i=1}^{N} y_i \tilde{B}_M(\mathbf{x}_i).$$

The denominator $b_{MM}$ in (75) is zero if and only if $B_M(\mathbf{x})$ has an exact linear dependence on the previously entered basis functions $\{B_m(\mathbf{x})\}_1^{M-1}$. In this case $\tilde{B}_M(\mathbf{x})$ and $c_M$ need not be saved. $B_M(\mathbf{x})$ must still be retained however to serve as a candidate parent for future basis functions.

Consider now the search for the next basis function $B_{M+1}(\mathbf{x})$. The (least-squares) lack-of-fit criterion to be evaluated in the innermost loop of Algorithm 2 (line 7) at each potential knot location $t$ is proportional to $-I(t)$, where

$$I(t) = \frac{\left[c_{M+1}(t) - \sum_{i=1}^{M} c_i V_{i,M+1}(t)\right]^2}{V_{M+1,M+1}(t) - \sum_{i=1}^{M} V_{i,M+1}^2(t)} \tag{76}$$

with $\{c_i\}_1^M$ given in (75) and the other quantities given by the updating formulae (52) with the orthonormal basis functions $\{\tilde{B}_i\}_1^M$ replacing the $\{B_i - \tilde{B}_i\}_1^M$ appearing there. The quantity $I(t)$ (76) is the improvement in the residual sum-of-squares resulting from adding the corresponding basis function with knot location $t$. This must be computed at every eligible knot location.

A computational advantage results from the fact that $I(t)$ (76) can be computed rapidly in $O(M)$ time. The strategy described in Section 3.9 required a partial Cholesky decomposition $[O(M^2)]$ and a full back-substitution [also $O(M^2)$] at each eligible knot location. Since the computation associated with the updating (52) is $O(M)$ (as it was before), the total computation for the new approach is

$$C \sim nNM_{\max}^3$$

in general and $C \sim nNM_{\max}^2$ for additive modeling ($mi = 1$). Here $M_{\max}$ is the maximum number of basis functions. Basically the second term in the sum of (53) has been eliminated by this approach reducing the computation by a factor of roughly $M_{\max}$ for (very) large values. This will enable the MARS procedure to be applied to much larger and more complex problems than with the initial implementation described in Section 3.9.

I have implemented this new approach into the latest version of the MARS software. In order to get an idea of the computational savings involved, I reran the example of Table 1 (this time on a DECstation 3100) contrasting the

TABLE 18
*Ratio of running times (sec.) for the new versus old (new/old) implementations of the least-squares updating algorithm on the example of Table 1. Computations were performed on a* DEC*station* 3100

| | $M_{max}$ | | | | |
|---|---|---|---|---|---|
| ***mi*** | **5** | **10** | **20** | 40 | 50 |
| 1 | 0.3 / 0.4 | 0.5 / 0.8 | 1.0 / 2.3 | 2.4 / 8.4 | 3.1 / 14.1 |
| 2 | 0.4 / 0.6 | 1.4 / 2.1 | 3.7 / 7.8 | 16.0 / 55.2 | 22.9 / 107.3 |
| 4 | 0.5 / 0.7 | 1.4 / 2.3 | 5.3 / 11.4 | 16.4 / 114.7 | 24.8 / 228.2 |

running times of the new versus the old implementation. Table 18 shows the ratio of running times (new/old) in the same format as Table 1.

The last two columns especially of Table 1, representing the higher $M_{max}$ values, indicate that the computational savings associated with the new approach are nontrivial. For example, with $mi = 4$ and $M_{max} = 40$ or 50, one can now do a tenfold cross-validation to access the quality-of-fit in the same time that the old program could do just a single fit. Buja and Duffy (see discussion of Buja, Duffy, Hastie and Tibshirani) consider very large $M_{max}$ values, for which the computational savings ought to be even more dramatic.

Stone's suggestions for extending MARS to logistic regression parallel those made by Buja, Duffy, Hastie and Tibshirani. I think this idea has substantial potential and that it should lead to a better method than that described in Section 4.5. His ideas for density and conditional density (MARES) estimation are quite clever and intriguing and hold practical promise. The computational burden, especially in the multivariate case, is likely to be heavy unless special tricks can be found. Perhaps the ideas mentioned by Owen in the last paragraph of his discussion might be helpful here.

Lewis and Stevens (1990) have been studying the application of MARS to the autoregressive modeling of time series. They report considerable success.

**O'Sullivan.** Professor O'Sullivan presents some valuable ideas for enhancing the power and interpretability of MARS-like approaches. MARS is indeed coordinate sensitive as are most of the commonly used regression techniques (linear regression with variable subset selection, CART, additive modeling). The very notions of main effects and interactions are coordinate sensitive. Coordinate sensitive procedures will outperform their affine equivariant counterparts in situations for which the dependence of the true underlying function is simplest in the coordinate system chosen (usually the original measured predictor variables). Here simplicity is defined in terms of

the ease with which a given procedure can approximate the function. Introducing adjustable linear combinations in place of the original coordinates removes the coordinate sensitivity and allows the approximation procedure to search for linear combinations that provide the simplest representation. This potential reduction in bias comes at the expense of increased variance (optimizing with respect to the linear combination coefficients) and usually substantially increased computational complexity. Interpretation of the fitted models is most easily achieved in terms of the original measured variables so that special interpretation/visualization tools are necessary for affine equivariant procedures. O'Sullivan presents some nice ideas along these lines in his discussion.

Experience with the linear combination split option in CART [Breiman, Friedman, Olshen and Stone (1984)] has yielded somewhat surprising results. Employing linear combination splitting seems to only rarely give substantially improved performance over axis-oriented splitting and surprisingly often it does worse. This may be a reflection of the types of problems to which CART has been applied. It also might be a reflection of the local variable subset selection property of axis-oriented recursive partitioning (see Section 2.4.2) which, of course, is a coordinate sensitive concept.

As discussed in Section 3.3, recursive partitioning procedures produce basis function sets that involve high-order interactions. This explains their poor performance in situations where main effects and low-order interactions dominate. On the other hand, in converse situations where the true underlying function happens to predominantly involve high-order interactions, this aspect of recursive partitioning becomes an asset. The alternating current examples (Sections 4.4–4.4.2) were included because they involve strong interaction effects to all orders. (Such examples involving real situations are not easy to find.) The MARS results indicate that it does a credible, if less than perfect, job in this case. CART with its bias toward high-order interactions ought to do better here except that lack of smoothness limits its accuracy. A smoothed version of CART, such as O'Sullivan's SCART, mitigates this limitation and, as reported in his discussion, it does somewhat better than MARS. The two examples where he reports MARS does substantially better are ones that involve predominantly main effects and/or only low-order interactions.

O'Sullivan suggests a clever approach based on finite element techniques (SCART) for smoothing CART models. Another way to implement a smoothed version of CART would be to directly follow the paradigm outlined in Section 3.2. That is, replace the CART step functions by corresponding higher order (univariate) spline functions, without employing the strategy described in Section 3.3, that is, the parent basis function would be removed as in CART. Repeated factors involving the same variable would be allowed in the same basis function product with this strategy. Also, the backward stepwise procedure would follow that used in CART. A possible advantage of this approach over one based on finite elements is that no additional global smoothing (really smearing) parameter $s$ (or $p$) need be introduced and adjusted to optimize the fit. All (local) smoothing is controlled directly by the forward/backward knot placement algorithm as in CART.

**Breiman.** I share Professor Breiman's wonderment at this article appearing in the *Annals of Statistics*. I was surprised when the journal solicited this paper and astonished when I discovered that they were serious. I hope that as a result traditional readers of the *Annals* will not cancel their subscriptions, but instead swallow hard and wait for the next issue where things ought to be back to normal.

Breiman's remarks center on the issue of model selection and specifically on the use of a penalized least-squares criterion for this purpose. I agree that the (historical) name generalized cross-validation for the GCV criterion (30) is misleading and that, especially with nonlinear fitting, it bears little resemblance to ordinary leave one out cross-validation. This fact, in and of itself, does not indicate its superiority or lack thereof. Ordinary cross-validation has its detractors and the issue of model selection is still being hotly debated in what has become a vast literature on the subject. I wholeheartedly agree with Breiman that the use of model selection criteria derived for linear fitting in nonlinear contexts (such as the use of $C_p$ or $F$-testing with variable subset selection procedures) represents a long-standing abuse in our field. In fact, Breiman has been one of the (few) leaders in pointing this out and suggesting remedies [Breiman (1989c) and Breiman and Spector (1989)]. The modification to the (linear) GCV criterion proposed in Section 3.6 is an (admittedly crude) attempt to account for the nonlinear aspects of MARS fitting. As noted by Breiman, the motivation for this approach was largely computational. The extent of its statistical success seems remarkable given the crudeness of the approximation. There is nothing intrinsic in the MARS approach to the use of any particular model selection criterion. If a better criterion can be found, this will improve the performance of MARS and so the results obtained with it so far represent lower bounds on what may be possible with this approach in the future. With the increased speed of the MARS algorithm obtained as a result of Stone's suggestion (Table 18), model selection through ordinary cross-validation is certainly computationally feasible except for large problems. (For these an independent test set can be used.) I intend to implement this approach and compare its performance to that of the current implementation. To the extent that Breiman's speculations are correct, this should lead to improved statistical performance for MARS.

I also found the packing problem to be an important concern. This was the motivation for the introduction of a minimum span described here in Section 3.8 and in Friedman and Silverman (1989), Section 2.3. Requiring a minimum number of observations between successive knot locations limits the number of candidate models in precisely the same way as Breiman's strategy of placing a limited number $K$ of initial knots in his backwards stepwise method. They both limit the number of eligible knots locations and prevent nearby models from becoming too closely packed in the space of candidate models. The analogy in MARS to Breiman's strategy of choosing $K$ through model selection, would be to adjust the parameter $L$ that controls the minimum number of observations between knots, to optimize a model selection criterion such as cross-validation. The generic default value for $L$ given by (43) is conservative

in the sense that it is set as small as possible consistent with some resistance to runs in the noise. The motivation is to keep the procedure as sensitive as possible to potential sharp structure in the function. When MARS is used in an initial exploratory mode, this seems like a reasonable choice. This of course has the collateral effect of increasing the variance of the estimates (possibly reacting to noise masquerading as sharp structure). When the true underlying function is very smooth with no sharp structure anywhere [such as $f(x) = 0.667\sin(1.3x_1) - 0.465x_2^2$], then there is no bias increase in increasing $L$, but there is a corresponding variance decrease. Figures 1 and 2 of Breiman's discussion compare MARS using its generic default value (no adjustment) to a procedure for which the corresponding parameter has been adjusted to do best on this particular example (through cross-validation). Such adjustment is a good idea and I recommend it with MARS modeling, especially in the latter stages of obtaining a final model estimate.

Restricting the minimum number of observations between knots (or the number of initial knots $K$ in a backward strategy like Breiman's) is an indirect way of trying to limit the global (absolute) second derivative of the final estimate. In his discussion of TURBO [Friedman and Silverman (1989)], Hastie (1989) suggested that the sometimes wild behavior of adaptive spline estimates (seen in low signal to noise, small sample situations) could be mitigated by directly placing a mild bound (or penalty) on the average-squared second derivative. (As can be seen in Figures 1 and 2 of Breiman's discussion, the wild estimates tend to have much higher absolute second derivatives.) Such a bound or penalty is global in nature, so that if one wanted to retain the flexibility of adaptive regression splines to adapt the degree of smoothing locally, the penalty should be made just large enough to inhibit possible wild behavior. In their discussion of this paper, Buja, Duffy, Hastie and Tibshirani show how to extend this idea to the more general context of MARS and, in addition, suggest the possibility using this approach for general model selection in place of backward basis function deletion. Both ideas are straightforward to implement in MARS and are currently under investigation (jointly with Hastie). We expect this approach to produce the biggest improvements in small sample, low signal-to-noise situations where the underlying function is very smooth.

Breiman points to some of the simulation study results as indicating a general failure of the (modified) GCV criterion (30) when used in the context of MARS. As noted before, the MARS procedure in no way requires the use of this criterion and one that can be shown to work better would be enthusiastically welcomed. However, I feel an obligation to those who support GCV to respond to some of Breiman's concerns. Cross-validation produces an estimate of average predicted-squared error. It can be shown to be an unbiased estimate but it surely has variance [sometimes rather high—see Efron (1983)]. For this reason it will not select the best model every time. I view the pure noise results in Tables 2 and 3 to be highly encouraging. While the GCV score estimated the MARS model as doing (slightly) better than the response mean about half the time, the actual distribution (percent points) show that it hardly ever claims

that the MARS model is very much better, making it unlikely that one would be led to embarrassing conclusions. Whether ordinary cross-validation would do better, given the high variance of its estimates, is by no means certain. In any case, the results in Tables 2 and 3 are really more a test of the smoothing parameter choice $d = 3$ (32), rather than the criterion (30), since one can make the procedure arbitrarily conservative by increasing the value of $d$. The fact that on the data of Section 4.3 (Table 7), full MARS modeling does as well as when interactions are restricted to be bivariate ($mi = 2$), represents a (fairly dramatic) success for the model selection procedure.

When the number of basis functions that can potentially enter is very large, beginning with forward stepwise selection is the only viable option. Sample size limitations, if nothing else, do not permit fitting with the full (astronomically large) basis set. In fact, Breiman's (1989b) implementation of the Π-method employs a forward then backward stepwise approach (using GCV for model selection). It is my guess that in low dimensional applications ($n \lesssim 3$), the Π-method will emerge as a strong competitor to other procedures intended for those settings, especially for certain very highly structured functions. [More detailed comments on the Π-method appear in the discussion of Breiman (1991)].

**Golubev and Hasminskii.** Golubev and Hasminskii raise the important issue of the theoretical properties of MARS when applied to various classes of smooth functions. I appreciate their inclusion of the known general results in this area. Gaining any theoretical understanding of the performance properties of MARS would likely be an enormous help in improving it.

**Buja, Duffy, Hastie and Tibshirani.** The discussion by Buja, Duffy, Hastie and Tibshirani describes important enhancements to MARS (and related procedures) that will likely improve both performance and interpretability. I am especially grateful for the discussion of their experiences with MARS on an important real problem. One tends to learn much more about a procedure when it is applied in a setting where the actual answer is the important thing, rather than serving simply as a test bed for the procedure.

A diagnostic tool indicating when a future covariate vector is outside the range of the training data is important. With flexible fitting procedures like MARS, the notion of outside needs to be broadened. Simply being inside (say) the convex hull of the design may not be enough to be safe. If the density of the design contains extended sparse or empty regions (holes), predictions within those regions may be suspect, especially if they are dramatic (far from the response mean). A diagnostic procedure that could indicate such situations would be an invaluable companion to any flexible fitting procedure in practical applications.

Buja and Duffy's experiences with using MARS raised a series of questions put forward in their discussion. I have observed some of the phenomena they relate in my experiences with MARS. I will try to address the issues they raise

in their points 1–5 of Section 1 of their discussion.

1. The recommendation at the end of Section 3.6 for choice of $M_{\max}$ (maximum number of basis functions) was intended to serve as a starting guide. It will not likely be the final best choice in all situations. As was done by Buja and Duffy, I recommend some experimentation with several values, using cross-validated performance as a guide, before a final model is chosen.
2. The cost parameter $d$ (32) is the primary smoothing parameter of the MARS procedure. [Secondary less influential smoothing parameters are $M_{\max}$ and $L(\alpha)$ (43)]. Increasing its value will cause fewer basis functions (knots) to be entered, thereby increasing smoothness. The penalty increase (in the GCV criterion) that it regulates is intended to compensate for the increased degree of data fitting associated with the stepwise basis function selection. As such, its value should depend on the degree of optimization performed. This in turn has a mild dependence on other parameters that limit the optimization procedure such as $M_{\max}$, $L$ (43) and $mi$ (maximum interaction order). Its greatest dependence can be, however, on the underlying function $f(\mathbf{x})$ (1) when it is highly structured. The default value $d = 3$ [$d = 2$ for additive modeling ($mi = 1$)] was chosen (mainly through simulation studies) to be appropriate for the null situation, $f(\mathbf{x}) =$ constant. This is a conservative choice and is appropriate as a starting value in order to avoid the embarrassment (discussed by Breiman) of fitting pure noise with a structured approximation. The results shown in Tables 2 and 3 indicate that it works reasonably well for this purpose. Theoretical results quoted in Owen's discussion, however, suggest that this choice might be too conservative for highly structured functions (far from null situations). The intuitive reason is that such highly structured functions give rise to highly preferred knot locations and the sampling functions induced by the noise are not strong enough to cause (initial) knots to be placed far from these preferred locations. The optimization procedure is thereby (indirectly) restricted as to where it can place knots, reducing the variance of the estimates. In the null case the true underlying (constant) function has no preference at all for where knots are located and knot placement is totally driven by the noise, inducing the most variance. The results quoted in Section 3.6 were based on simulation studies involving very smooth weakly structured functions, where the default values motivated by the null case seemed to work reasonably well. The situation faced by Buja and Duffy seems to be one involving a highly structured function with sharp threshold effects. The default values may not be appropriate in this case.
3. In the present implementation of MARS, $d$ (32) is not adjusted for $M_{\max}$ although it would be reasonable to do so. The oddities observed by Buja and Duffy may stem from implementation details associated with the backward stepwise strategy like those pointed out in Owen's discussion.
4. The anomalous behavior associated with the piecewise-cubic fits observed by Buja and Duffy is (as they concluded) due to the (apparently) highly

structured nature of their function. The same effect was observed in the semiconductor design example in Section 4.7. It also has sharp threshold effects that cause problems for the derivative smoothing. (Note that Section 4.7 was added to a later version of the manuscript than was supplied to the discussants.) This anomalous behavior is in some sense a blessing in disguise. It stems from the ability of the piecewise-linear basis to approximate sharp thresholding with a small number of basis functions; a single threshold can be captured with at most two knots. Imposing a higher level of smoothness (say by using higher order splines) would require placing more knots in the vicinity of each threshold to allow the derivative to change very rapidly there. The piecewise-cubic approximation used in MARS has an even greater disadvantage since it attempts derivative smoothing using only the knot locations derived from the piecewise-linear fitting. It therefore does not have additional knots near each threshold to help it rapidly adjust the derivative. Thus it even more dramatically oversmooths the derivative in such regions. Approximations that impose a higher level of smoothness of course perform best when the underlying function is very smooth; that is, functions that nowhere have locally high second derivatives. For such functions, piecewise-linear splines will require several knots over the data interval to adequately approximate the smoothly changing (first) derivative. The derivative smoothing strategy (Section 3.7) can then use these to fashion a piecewise cubic basis to remove the resulting mild derivative discontinuities. As observed by Buja and Duffy, anomalous behavior of the piecewise cubic fit (when it occurs) is readily diagnosed by simply examining the respective GCV scores of the two models. Simulation studies on relatively smooth functions indicate that the continuous derivative approximation tends to fit the data at hand slightly worse (since the knots are optimized for the linear basis) but has slightly better predictive performance. Experience with sharply structured functions (such as those involving threshold effects) indicate that the piecewise-cubic basis fits both the data at hand and future data badly. This suggests a strategy of accepting the cubic fit if its GCV score is at most slightly (10–20%) worse than the piecewise-linear fit; otherwise use the linear basis fit. Considerably worse GCV scores for piecewise cubic fits can serve as a useful diagnostic indicating sharp structures in the response. One (but not the only) source for such sharp structure can be single or multiple response outliers.

5. The evenness of the growth of the selected model size on $M_{\max}$ (maximum number of basis functions) likely depends on a variety of different characteristics associated with each particular problem. These include sample size, properties of the design and the true underlying function.

**Gu and Wahba.** Gu and Wahba present a lovely concise descriptive summary of the smoothing spline approach to fitting functions of several variables. Researchers at the "Madison spline school" led by Wahba have been pioneers in this important area of statistics. Thin plate and interaction splines represent important contributions that are theoretically attractive and are, in

addition, highly competitive in practical settings involving low-to-moderate dimensionalities and relatively small sample sizes.

In their discussion, Gu and Wahba point out some of the aspects that MARS and interaction splines hold in common. There are also important differences. These differences basically stem from differing motivations associated with intended applications. Applications motivating MARS are similar to those that motivated CART [Breiman, Friedman, Olshen and Stone (1984)], namely (relatively) large complex data sets where little is known about the true underlying function $f(\mathbf{x})$ (1). In such settings one needs a general procedure that requires as its only input specification the training data, from which it produces a reasonably accurate and interpretable approximation $\hat{f}(\mathbf{x})$. If such a procedure is successful in this, the user may wish to use the derived information to further refine the model. This can be done by reapplying the general procedure in various restricted modes, where the restrictions are guided by the results of the earlier general application. For example, if a MARS run indicates that certain variables enter additively or participate only in limited interactions, further tuning of the model would be done in the presence of these constraints. This refinement can also be done by applying a different less general procedure that requires a more detailed input specification, with these details provided by the output of the previously applied general procedure.

Application of interaction splines requires as part of its user specified input an ANOVA decomposition (10), (24). That is, the user must tell the procedure what variables enter the model, which specific variables (if any) they interact with and the levels of those interactions. Since these reflect properties of the true underlying function $f(\mathbf{x})$ (1), the quality of the approximation can depend strongly on this input in any given situation. Statistical and computational considerations strongly limit the number of ANOVA functions that can be entered into an interaction spline model. In low dimensional settings this problem is mitigated by the (relatively) small number of ANOVA functions that can potentially enter. The user can simply enter them all or experiment with different possible subsets. With increasing dimension of the predictor variable space, the number of possible ANOVA functions grows very rapidly and this is no longer a viable strategy.

MARS does not require an ANOVA decomposition as part of its input specification. Using only the data, it provides an ANOVA decomposition as part of its output. As noted in the discussions, this can be a valuable tool for interpreting the approximating function and (to the extent it reflects the true underlying function) also the system under study that generated the data. If it turns out that MARS produces an ANOVA decomposition with a small number of ANOVA functions each involving only a few of the variables (as is often the case), this information can be used as input for an interaction spline model. Whether this will result in improved accuracy will largely depend on the smoothness properties of the underlying function.

Other differences between the two approaches also largely center on issues of generality. In its most general application, interaction splines associate a

single smoothing parameter with each specified ANOVA function (10), (12), (24), that constrains its overall global smoothness. As a consequence of its adaptive knot placement algorithm, MARS attempts to adapt its smoothness constraint locally within each ANOVA function that it produces. Note that this will only provide a potential benefit if there is sharp structure present, as was the case in the semiconductor design example (Section 4.7) and in the problem encountered by Buja and Duffy (discussion of Buja, Duffy, Hastie and Tibshirani). In cases where the (true underlying) ANOVA functions are all globally very smooth, this local adaptability can be counterproductive in that the associated increased variance is not offset by decreased bias. For these cases, model refinement using an interaction spline approach might provide a real benefit. (Applying MARS with a strong global second derivative penalty may also help in these situations.)

The method for incorporating categorical variables into interaction spline models described by Gu and Wahba basically encodes them into real-valued (0/1) dummy variables. Regularization is provided by grouping together the coefficients associated with each original categorical variable and shrinking them towards a common mean by penalizing the variance of their solution estimates. This is a clever idea. Categorical variables are incorporated into MARS using a different strategy [Friedman (1990)]. This strategy is motivated by that of CART, which does not use dummy variables, but instead attempts to find subgroups of categorical values within each variable over which the response (conditioned on the rest of the model) is roughly constant (small variance). The differences between these two approaches for categorical variables reflect the respective differences between the two methods for ordinal variables. Interaction splines apply a global smoothness constraint on each variable or ANOVA function, whereas adaptive spline strategies like MARS attempt to adjust the smoothing constraint locally within each variable or ANOVA function, to adapt to possible sharp local structure.

**Barron and Xiao.** Barron and Xiao suggest several clever modifications to the MARS procedure, some of which will likely improve its performance. Like Hastie (1989) and Buja, Duffy, Hastie and Tibshirani (discussion herein), they propose the imposition of a global roughness penalty on the solution. Their criterion penalizes increasing (squared) first, rather than second, derivatives but it should produce similar results. In fact, the Barron–Xiao penalty has an especially nice intuitive appeal. As discussed in the rejoinder to Breiman, a global roughness penalty will likely provide substantial benefit when the true underlying function varies quite smoothly over the predictor space with no (relatively) sharp local structure anywhere. There is, however, no free lunch. While helping with such very gentle functions, the price paid for using a strong (global) roughness penalty will be to inhibit the ability of MARS to capture local structure when it is present (without degrading the fit elsewhere). At least for (initial) exploratory applications, the penalty should be set just large enough to inhibit only possible wild behavior (see Breiman rejoinder) and not limit the flexibility of the procedure. Considerations are different for proce-

dures based on global polynomials (such as MAPS). Since polynomials (unlike splines) already have inherent difficulty dealing with local sharp variation of the underlying function (see later), nothing (additional) is lost by imposing a moderate to strong global roughness penalty on them.

Barron and Xiao also raise the issue of model selection and suggest alternatives to the GCV criterion used in the present implementation of MARS. As mentioned in the rejoinder to Breiman, MARS is in no way wedded to the GCV criterion. More effective model selection can only help improve performance. The BIC/MDL criterion suggested by Barron and Xiao has strong intuitive appeal and, along with ordinary cross-validation (proposed by Breiman), will be tested in the context of MARS. The evidence provided for its superiority in Table 2 of their discussion, however, is only partially convincing. This is due to the fact that the criterion was used there in conjunction with the true underlying error variance ($\sigma^2 = 1$, known only because this is a simulated example) rather than trying to estimate it from the data at hand. Knowing that the true error is homoscedastic and the value of its variance provides a strong advantage to any model selection procedure. If an estimated value of $\sigma^2$ happens to be too large, the model selection criterion will tend to include too few terms whereas conversely, too many will be entered. The variance of estimates of $\sigma^2$ can be quite high, especially in conjunction with (nonlinear) flexible fitting procedures. There may also be considerable bias unless one knows how to correct for the (basis function) selection aspect. There is no obvious analog here for obtaining an unbiased estimate based on the largest possible model.

The MAPS procedure introduced by Barron and Xiao closely follows the MARS strategy and as a consequence inherits many of its characteristics. The only basic difference is the substitution of global polynomials in place of (adaptive) splines. If one is constrained to produce only polynomial models this represents an attractive approach. If not, the difficulties associated with polynomial approximations can impose (sometimes rather strong) limitations.

Global polynomials are held in somewhat low esteem as general tools for function approximation (or estimation) in both the mathematical approximation and statistical curve and surface estimation literatures. Nearly all the recommended procedures that have emerged so far possess in common a locality property; the estimate at a point is most strongly influenced by (training) observations close to that point and observations further away have little or no influence. This gives rise in large part to their flexibility. They can respond to (sharp) local properties of the function without affecting the fit everywhere else. Global polynomial fits do not share this property; the function estimate at a point can be strongly influenced by data points very far away from it in the predictor space. As a consequence, locally sharp structure anywhere can influence the fit everywhere. This is the likely motivation for the quote from J. W. Tukey, "polynomials cut peoples' throats" and the observation by de Boor (1978), "If the function to be approximated is badly behaved anywhere in the interval of approximation, then the approximation is poor everywhere. This global dependence on local properties can be avoided using

piecewise polynomials (splines)." [See de Boor (1978), Chapter II for a nice discussion of the limitations of polynomial approximations.]

If the true underlying function $f(\mathbf{x})$ (1) is everywhere gently varying with no sharp structure anywhere, then approximations based on global polynomials perform very well. In fact, if $f(\mathbf{x})$ happens to actually be a polynomial (or very close to one), then they will give the best performance. However, local methods such as splines also do quite well in these settings. Thus, if one expects to only encounter situations such as this, there is (as noted by Barron and Xiao) little to choose between them. If, however, one wants to maintain the additional ability to adequately deal with more structured functions, then local methods might be preferred.

The MARS procedure is based on (adaptive) spline functions because they emerge naturally as a generalization of recursive partitioning. It thereby inherits the attractive properties of the recursive partitioning approach discussed in Sections 2.4.2 and 6. These include local variable subset selection and automatic local adjustment of the degree of smoothing within each ANOVA function produced. Substituting polynomials for the adaptive spline functions sacrifices the local aspects of both these properties; only global variable subset selection and automatic adjustment of global smoothing on each ANOVA function are retained. If in a particular situation the nature of the underlying function happens to be such that this additional flexibility of MARS gives rise to no advantage, then there is little to choose between MARS and MAPS. Therefore, issues of generality will (as before) likely guide the choice.

Using adaptive splines also causes MARS to inherit the ability to isolate local sharp structure and deal with it separately without affecting the fit in other regions of the predictor space. Buja and Duffy (discussion by Buja, Duffy, Hastie and Tibshirani) report this as a crucial advantage in their application. Also, the semiconductor component example of Section 4.7 likely benefits from this property.

The examples of Sections 4.2 and 4.3 were chosen largely because they do not play to the particular strengths of MARS. They are globally very smooth quite gentle functions and (as noted by Barron and Xiao) global low-order polynomials make up a substantial part of their definitions. One would expect approximations based on low-order polynomials to do very well here. This is partially verified by the results presented by Barron and Xiao (Tables 2 and 3). Perhaps somewhat surprising is the degree of competitiveness displayed by adaptive splines in these situations. For these examples, local variable subset selection and adaptive local smoothing provide little or no advantage. The MAPS procedure on the other hand has the additional benefit of a global roughness penalty constraint (not yet incorporated into MARS) which as noted before is a real help with very smooth functions. It also has the advantage (Tables 2 and 3) of being supplied with the true underlying error variance. As discussed before, this provides the model selection criterion with an important advantage. This can be seen by comparing Tables 1 and 2 of Barron and Xiao's discussion.

TABLE 19
*Accuracy of* MARS *applied to the example of Section* 4.2 ($N = 200$) *with the smoothing parameter* $d$ (32) *selected through cross-validation*

| *mi* | $\overline{\text{ISE}}$ | $\overline{\text{PSE}}$ | $\bar{d}$ |
|---|---|---|---|
| 1 | 0.015 (0.013) | 0.15 (0.01) | 4.9 (1.2) |
| 2 | 0.030 (0.015) | 0.16 (0.01) | 5.7 (1.3) |
| 10 | 0.031 (0.016) | 0.16 (0.01) | 5.8 (1.3) |

The results presented for MARS in Tables 4 and 7 were obtained using its default smoothing parameter value $d = 3$ (32) without any attempt to even estimate a best value from the data. The best value is controlled by the underlying error variance which for the purpose of illustration was assumed not to be known. As discussed in the rejoinder to Breiman, using this default value may be reasonable for initial exploratory work, but one may wish to refine the fit in the later stages of the analysis by estimating a better value through cross-validation. To get an idea of whether this can give rise to substantial improvements, the simulation study on the example of Section 4.2 was rerun ($N = 200$ only) but this time using cross-validation to estimate the smoothing parameter rather than using the default value. Table 19 shows the results based on 100 replications. Also shown are the average estimated smoothing parameter values. (The quantities in parentheses are the standard deviations over the 100 trials.)

Table 19 (column 4) shows that the cross-validation procedure was choosing (on average) considerably larger smoothing parameter values than the default ($d = 3.0$) in this case. This reflects the very smooth nature of the underlying function. Comparing Tables 4 and 19 shows substantial improvement [24% in $\sqrt{\text{ISE}}$ (58)] only for $mi = 1$ (additive model). However, as Table 5 indicates, this is the one most likely to be chosen. Comparing Table 19 with Table 2 of the discussion of Barron and Xiao shows that even without a global roughness penalty and knowledge of the true underlying error, adaptive splines compare favorably with global polynomials in this setting.

The motivations put forward by Barron and Xiao for substituting global polynomials for adaptive splines in MARS are mainly computational and (to a lesser extent) interpretability. An implementation based on polynomials did gain a computational advantage over one based on adaptive splines when the old implementation strategy described in Section 3.9 was used for the latter. This is, however, no longer the case with the new implementation discussed in the rejoinder to Stone. The computation for both global polynomials and adaptive splines (new implementation) is proportional to $nNM_{\max}^3$. The actual relative computing speeds will depend on implementation details but are not likely to be very much different. Adaptive splines may have a slight advantage since fewer basis functions $M_{\max}$ are often needed for comparable accuracy. This is a consequence of the fact that each basis function included in a MARS model is adapted to the data through the adjustment of its knot locations.

Interpretability of the resulting approximation was an important design goal motivating the MARS approach. The ANOVA decomposition (Section 3.5) and slicing (Section 4.7) are intended as important interpretational aids. The MAPS procedure by closely following the MARS strategy inherits these aspects. The most powerful interpretational aids in understanding each ANOVA function are likely to be graphical representations (curves or surface displays). For these, whether the approximation is internally represented by a polynomial or spline function is of little consequence. If the polynomial representation involves more than a very few terms and/or high degrees, some may argue that spline representations are more interpretable, especially since they are likely to be more parsimonious (see discussion of Buja, Duffy, Hastie and Tibshirani). On the other hand, if a function can be approximated by a few low-degree polynomials, those familiar with polynomials may feel more comfortable interpreting them. Such interpretations can be misleading, however, since (except for artificially constructed examples) the corresponding true underlying function is seldom really a polynomial.

One very important aspect in interpreting these approximations is a small number of resulting ANOVA functions. In this respect adaptive splines may have a distinct advantage (as noted earlier). In the Portuguese olive oil example (Section 4.5), the MARS model resulted in four ANOVA functions, involving only three variables (Table 14), that can be represented by only two (surface) plots (Figure 8). While achieving comparable accuracy in this case, the polynomial approximation produced by MAPS was far more complex as reflected in Table 4 (and 5) of the discussion by Barron and Xiao. The underlying function for this example is also globally quite smooth as can be seen in Figure 8. [Note that Figure 8 displays the log-odds. The underlying function estimate $1/(1 + e^{-\hat{f}(x)})$ is correspondingly much more gentle.]

For those who happen to have a strong preference for approximations based on polynomials (or other global basis functions) but would still like to retain the flexibility of the adaptive spline approach, there are several possibilities. One simple possibility would be to apply both to the data. If they provide comparable estimated accuracy, then the user could interpret the one he/she finds to be most understandable. Perhaps a more elegant approach would be to include both global functions (such as polynomials) and adaptive splines, in a hybrid MARS modeling strategy. To enhance interpretability, one could (optionally) forbid interactions between the global and spline basis functions. An incremental penalty might be attached for including adaptive splines to reflect the user's preference for the global basis functions. The global basis function part of the resulting approximation could then be interpreted as reflecting the very smooth aspects of the underlying function whereas the spline part (in this case) would reflect possible local sharp structure. (Such a hybrid strategy is easily added to the MARS program as a user option.)

The feed-forward idea proposed by Barron and Xiao is quite intriguing. It would allow the procedure (MARS or MAPS) to more rapidly build up (synthesize) higher order interaction terms but in a constrained manner. The result-

ing models (if feed-forward inputs are selected) will be far more complex and difficult to interpret. Whether this approach will give rise to substantially improved prediction accuracy in some situations awaits the results of further investigation.

## REFERENCES

BREIMAN, L. (1989a). Fitting additive models to data. Technical report 210, Dept. Statist., Univ. California, Berkeley.

BREIMAN, L. (1989b). The Π-method for estimating multivariate functions from noisy data. Technical report 231, Dept. Statist., Univ. California, Berkeley.

BREIMAN, L. (1989c). Submodel selection and evaluation in regression I. The $X$-fixed case and little bootstrap. Technical report 169, Dept. Statist., Univ. California, Berkeley.

BREIMAN, L. and SPECTOR, P. (1989). Submodel selection and evaluation in regression $x$-random case. Technical report 191, Dept. Statist., Univ. California, Berkeley.

DE BOOR, C. and RICE, J. R. (1979). An adaptive algorithm for multivariable approximation giving optimal convergence rates. *J. Approximation Theory* **25** 337–359.

FRIEDMAN, J. H. (1990). Estimating functions of mixed ordinal and categorical variables using multivariate adaptive regression splines. Technical report LCS 107, Dept. Statist., Stanford Univ.

HASTIE, T. (1989). Discussion of "Flexible parsimonious smoothing and additive modeling" by J. H. Friedman and B. W. Silverman. *Technometrics* **31** 23–29.

LEWIS, P. A. W. and STEVENS, J. G. (1990). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). Technical report, Naval Postgraduate School, Monterey, Calif.

LYCHE, T. and MYØRKEN (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.) Oxford Univ. Press, New York.

DEPARTMENT OF STATISTICS
SEQUOIA HALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305