ANDREAS BUJA, DIANE DUFFY, TREVOR HASTIE AND ROBERT TIBSHIRANI

Bellcore, Bellcore, AT & T Laboratories and University of Toronto

We would like to congratulate Professor Friedman on this characteristically ingenious advancement in nonparametric multivariate regression modeling. MARS is a triumph of statistical computing and heuristics—the clever algorithmic and heuristic ideas make extensive searching computationally feasible. The resulting modeling technology offers the data analyst a remarkably flexible tool which we found very useful on a difficult real-world problem. We will address a few issues that arose in our reading of this excellent paper and our experience using the MARS program.

1. Some experience with MARS. Two of us (Buja and Duffy) acquired some experience with MARS in an extensive analysis of data concerning memory usage in electronic switches. The data comprised 241 observations on 27 variables. It was known from the onset that the available variables gave an incomplete description of the response. Careful and creative regression modeling yielded fits with good global properties ($R^2 \approx 0.995$), but there were still unacceptably large residuals and poor performance on cross-validation tests. The fits which we obtained from MARS, on the other hand, excelled in prediction and cross-validation. In addition, the robustness to influential points which MARS inherits from the local adaptivity of the selected basis functions was very advantageous in our context. Our set of observations was (purposely) chosen to include a subset consisting of notoriously difficult cases. These cases, as expected, wreaked havoc on regression models but MARS was able to adapt to them without degrading the fits to the rest of the cases. In addition, highly accurate MARS models could be built with fewer variables (13 as opposed to 18) which happens to be a true benefit in this situation. The MARS models involved several second and third order interactions which, while impossible to anticipate by subject matter experts, seemed reasonable in the sense that they involved variables which are expected to have large effects on the way memory is used.

An interesting aspect of this analysis is that the data exhibit genuine noise despite the fact that switches are basically deterministic systems. This is because the 27 predictors were selected from a complete set of about 300 predictors based on what information is available to engineers who operate these systems. The success of the MARS fits can only be explained as a result of strong interdependence within the large set of predictors, rendering most of them redundant. Thus, we are profiting from what we call "concurvity" which in most contexts is a cause for concern. Further, models based on the theory of the switching systems would necessarily involve many of the 300 predictors and would therefore be useless to the engineers. One danger in this, or any, data-driven as opposed to theory-driven approach is that the model may be misleading if future predictions occur in areas of the predictor space where data are sparse. It would be useful if MARS were accompanied by diagnostic

tools which indicated when a future set of covariate values is stepping dangerously outside the range of the training data. A first naive attempt at deriving such a tool would be to compute the Mahalanobis distance of test covariate vectors in the linear predictor space spanned by the basis functions of a given MARS model. However, such an approach may have problems since the constant zero stretches of the spline basis functions lead to clumps of data in the extended predictor space.

We found it quite useful that the first order truncated basis splines are of an exceedingly simple form. A fitted model is easily communicated to practitioners and it is trivial to implement on arbitrarily small machines. By comparison, we do not see a use for the enhanced cubic models in this (prediction) context. For graphical display and qualitative data analysis, they may have their advantages.

In using MARS to analyze our data the following questions and comments arose.

1. Friedman recommends running MARS with $M_{\max}$ (the maximum number of basis functions added in the forwards stepwise procedure) approximately equal to $2M^*$, where M^* is the GCV-minimizing choice for the number of basis functions in the model. In our context, M^* is in the neighborhood of 35–40. Based on our experience with honest cross-validation, this is too large for the sample size and it may indicate that, at least for these data, the default cost being charged for basis function optimization is probably too low.
2. It appears that no complete description of the heuristic choice of the cost parameter d is given. If there is no restriction on the degree of interaction ($mi = n$), we understand that the default value is $d = 3$. The question for which we were unable to find an answer was: How does d depend on mi, the maximal degree of interaction, if it is specified to be less than n? When the degree of interaction is limited ($mi < n$), d is, quite logically, decreased. We chose in one instance $mi = 5$, which seemed to result in a value of d closer to 2 than 3.
3. As mentioned in 2, the cost d is set to 3.0 when unlimited interactions are permitted. Based on our calculations, it appears that the value of d is also being adjusted based on $M_{\max}$. How exactly does it affect d? On one occasion, we observed an apparent oddity in the behavior of d which seems counterintuitive: d can decrease as $M_{\max}$ increases with mi (the maximum permitted degree of interaction) held fixed.
4. One of the startling features of our MARS runs is the fact that the piecewise cubic GCV values are often an order of magnitude larger than the corresponding piecewise linear GCV values. In addition, there is little correspondence between the piecewise linear and the piecewise cubic GCV's for these data. For example, the model which minimizes the piecewise linear GCV has an associated piecewise cubic GCV which is over three times larger than the piecewise cubic GCV of another model; this other model, however, has a piecewise linear GCV which is almost twice the minimal

value. Further, the two models are very different; the first has 35 basis functions and interactions up to the third degree whereas the second has only 21 basis functions and all are restricted to be main effects only. Hence, if a smooth (piecewise cubic) model had been our ultimate objective, we would have been led very far astray by basing the model choice on minimizing the piecewise linear GCV. We can see at least three reasons for the unpredictable behavior of the piecewise cubic modifications: First, the residual sum of squares is very nonrobust and responds dramatically to a few bad residuals. Second, in high dimensional predictor spaces and in the presence of higher order interactions, the seemingly inoccuous piecewise cubic modification is far from minor because of compounding effects in the products. And third, our data may very well be better described by piecewise linear functions due to threshold effects which we observed while performing graphical exploration of the data.

5. In fitting a series of models with increasing values of $M_{\max}$, the number of basis functions in the final model grew quite unevenly. While this might be expected for choices of $M_{\max}$ which produce poorer fitting models, it was surprising to us for a choice of $M_{\max}$ yielding near optimal models (i.e., models with piecewise linear GCV values near the minimum). We are unsure whether to interpret the widely varying numbers of basis functions as an artifact of our data or as a property of the MARS methodology.

2. Generalized MARS models. Friedman proposes using MARS for logistic models. This can, of course, be easily extended to include all generalized linear models. The standard method for fitting such models is to maximize the likelihood or, equivalently, to minimize the deviance. While it would be natural to use the penalized deviance as the criterion for knot inclusion or deletion in direct analogy to the penalized RSS or LOF used in the present paper, this is computationally impractical because iteration is required to estimate the parameters and the crucially important ability to rely on the updating formulae is lost. Consequently, Friedman offers an approximation and here we offer another.

Suppose the basis set has k members and we wish to find the $(k + 1)$st. The exact inclusion of a candidate b_{k+1} can be achieved by using an iteratively reweighted least squares algorithm, with the initial values and working response provided by the fit to the set of size k. Instead of iterating to convergence, we propose using one iteration and instead of using the deviance to evaluate the fit, we propose using the weighted residual sum of squares or chi-squared approximation to the deviance. Since the fits for all candidates for b_{k+1} use the same working response and the same weights, Friedman's entire updating approach carries over. Once the approximately optimal b_{k+1} has been selected, the corresponding iterations can be completed to estimate the associated coefficients.

This approximation for evaluating a candidate b_{k+1} is exactly that used in Rao's score test [Pregibon (1982)] with the additional advantage that we exploit the updating facility to simultaneously perform multiple score tests.

3. ANOVA decomposition. The ANOVA decomposition is achieved in MARS by grouping together all terms involving the same variables. Thus all the functions involving only X_1 comprise the main effect for X_1, all terms involving only X_1 or X_2, the interaction for X_1 and X_2 and so on. The usual ANOVA decomposition for categorical designs ensures that interactions are free of lower order effects by imposing suitable summation constraints. Note that the tendency for these surfaces to include lower order effects in MARS is exacerbated because MARS can destroy the hierarchical structure of its basis during knot deletion. It would be useful if MARS could produce an interaction surface which was free of lower order effects. One could then use this surface to assess the way in which the variables interact, without being distracted by the lower order effects.

Hastie and Tibshirani [(1990), page 266] propose a strategy for this which can be adapted to MARS. As in standard ANOVA, one needs only to uncouple the components in the fitted model, not during the fitting (unless one requires an a priori hierarchy in the terms). Let us focus first on a bivariate interaction term, say $f(X_1, X_2) = \Sigma_k \alpha_k b_{1k}(X_1) b_{2k}(X_2)$. We first identify all the univariate basis functions in each of the tensor products pairs. In this case they are the b_{1k} and b_{2k}. We then project the interaction surface onto the joint space defined by them and the other main-effect basis functions involving those two variables. This additive main effect component of the interaction is then removed and lumped together with the original main effects, leaving a residual component which can be interpreted as a pure interaction and which is orthogonal to these (new) main effects. It is important to stress that the fit of the model has not been changed during this operation, simply its ANOVA decomposition.

Of course, if higher order interactions are present, this procedure would have to be used in a top down fashion. It is not entirely obvious how this would proceed. For example, if the term in question is a third order interaction, then we should isolate all bivariate interaction basis pairs. These would be grouped with similar and lower order terms involving the same variables and the entire set used to remove the second order effects from the third order interaction.

Incidently, in the simple linear regression model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 X_2$, we would not need to do all of this to understand the interaction effect. Decomposing such a fit amounts to looking at the coefficients (their sign and magnitude). Although these are fitted jointly, we know from the Gram–Schmidt process that $\hat{\gamma}$ is also the coefficient of $X_1 X_2$ adjusted for X_1 and X_2. This would not be the case if the model were $Y = \alpha + \beta_1 X_1 + \gamma X_1 X_2$.

4. Shrinking versus knot deletion. In a discussion of the additive predecessor to MARS, called TURBO, Hastie (1989) outlined in some detail a method for shrinking a TURBO model. The idea is that the forward stepwise algorithm results in a rich set of knots/basis functions for each additive term. In particular, the knots will be denser on some variables and locally, within a variable, there may be clusters of knots in regions of high signal-to-noise ratio. At this point the model is (purposefully) overparametrized and some regular-

ization is needed. As an alternative to backward knot deletion, Hastie suggested regularizing by shrinking according to an appropriate smoothness penalty. He suggested the penalty $\sum_j \lambda_j \int (f_j''(x_j))^2\, dx_j$ and pointed out that the resulting procedure is a generalized ridge regression. Furthermore, since the second derivatives of the piecewise-cubic approximations to the piecewise-linear basis functions have local support, the ridge penalty matrix is diagonal.

With an appropriate set of penalty functionals, a similar approach can be taken with MARS. Wahba [(1990), Chapter 10] outlines in some detail an approach using tensor-product splines, which are exactly what MARS uses to build up its bases. In Wahba's setting, models are fit in subspaces of the tensor product space of all the univariate reproducing kernel Hilbert spaces and the penalty functionals (norms) of these subspaces are inherited from the univariate spaces. For example, functions involving only X_1 and X_2 would be penalized using $\int\int (\partial^4 f(x_1, x_2)/\partial x_1^2\, \partial x_2^2)^2\, dx_1\, dx_2$. In practice then, the terms are grouped according to their components (much like the ANOVA grouping in MARS), each group gets assigned an appropriate penalty (and potentially its own smoothing parameter) and then the fit is computed by penalized least squares. Thus suppose the MARS model after the stepwise inclusion stage can be written

$$f(x_1, x_2, \ldots, x_p) = \alpha + \sum_{k \in I_1} f_k(x_k) + \sum_{(l,m) \in I_2} f_{lm}(x_l, x_m) + \cdots,$$

where I_j denotes the sets of j-tuples corresponding to interactions of order j. Each of the f_* has a linear representation in terms of an appropriate set of tensor-product bases. Then the shrunken model is the minimizer of the penalized criterion

$$\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \sum_{k \in I_1} \lambda_k P_k(f_k) + \sum_{(l,m) \in I_2} \lambda_{lm} P_{lm}(f_{lm}) + \cdots,$$

where the P_* are the penalty functionals.

Without going into all the details, it is worth pointing out that each of the $P_*(f_*)$ evaluates to a quadratic form $\alpha_*^T M_* \alpha_*$ in the coefficients for the basis functions in f_*; if the cubic approximations to the piecewise-linear functions are used, then each of the M_* is diagonal and once again the problem is a generalized ridge regression. We have used a different regularization parameter λ_* for each of the previous components. In practice one could simply use a single global λ and trust that the forward knot selection will give some terms more importance than others. Alternatively, one could lump all terms of the same interaction order together with the same penalty and shrink them all at the same rate.

This is a current research project by Friedman and Hastie and we are experimenting with other strategies and penalty functionals.

5. MARS for classification. One of us (Tibshirani) has made some progress in the development of a methodology for classification that tries to

combine some of the features of MARS and CART. Consider a two-class problem with $Y = 0$ or 1. The working model is

$$\log \frac{\Pr(Y = 1|\mathbf{x})}{1 - \Pr(Y = 1|\mathbf{x})} = \alpha + \sum_{k \in I_1} f_k(x_k) + \sum_{(l, m) \in I_2} f_{lm}(x_l, x_m) + \cdots,$$

where the I_j's are as defined in the previous section, but where the f's are tensor products of order 0, that is, products of indicator functions of the form $(x_j - t)^+$ and $(x_j - t)^-$. The motivation for 0 order splines is, as in CART, the ease with which models can be substantively interpreted when terms are of this form.

The model estimation is carried out as follows:

1. The model is constructed in a forward stepwise manner, exactly as in MARS. A score test (analogous to that described in Section 2) is used to select the split point for each variable. In contrast to CART, a basis function is not removed from the model after it has been split, thereby encouraging main and lower order effects to appear.
2. The model is pruned in a backwards hierarchical fashion, similar to the pruning in CART. In detail, a pruning operation consists of deleting a pair of functions of the form $b(\mathbf{x})(x_j - t)^+$ and $b(\mathbf{x})(x_j - t)^-$ and any higher order terms in which they appear. k-fold cross-validation (of the entire estimation procedure) is used to determine the optimal amount of pruning. In contrast to CART, we favor the use of deviance rather than misclassification cost to guide the pruning; this enhances the interpretability of the final model.

The result of this process is an estimated binary logistic model with a linear predictor that is a sum of products of indicator functions. In addition, a global (cross-validation) estimate of its classification performance is available, which is hopefully a more accurate estimate of future performance than GCV estimates, such as those used by MARS, which do not involve cross-validating the term selection process. The model can be interpreted either in terms of its basis functions or of the binary partition of the predictor space that they define. Initial studies suggest that this procedure is more effective (as a descriptive tool) than CART in cases where main effects dominate. On the other hand, in general it does not seem to classify as well as CART does. Extensions to ordered and unordered multiple class problems are possible. Further details will appear in a forthcoming technical report.

REFERENCES

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

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

PREGIBONS, D. (1982). Score tests in GLIM with applications. In *Proc. GLIM 82 Conf. Lecture Notes in Statist* **14** 87–97. Springer, New York.

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.