goals, the former can have noticeable impact on the latter. For the model selected by MDL, the value of GCV = 0.16 is a reasonably good estimate of CV = 0.17; whereas, for the model selected by GCV, the minimum GCV value of 0.15 does not give as good an estimate of the corresponding CV = 0.22.

REFERENCES

Barron, A. R. (1984). The predicted squared error: A criterion for automatic model selection. In *Self-Organizing Methods in Modeling* (S. J. Farlow, ed.) 87–103. Dekker, New York.

Barron, A. R. (1989). Statistical properties of artificial neural networks. *Proc. Twenty-eighth Conf. on Decision and Control*. IEEE, New York.

Barron, A. R. (1990). Complexity regularization. *Proc. NATO Advanced Study Inst. Nonparametric Funct. Estimation Related Topics*. Kluwer, Boston.

Barron, A. R. and Barron, R. L. (1988). Statistical learning networks: a unifying view. *Computing Science and Statistics: Proc. Twentieth Symp. on the Interface* (E. J. Wegman, D. T. Gantz and J. J. Miller, eds.) 192–203. Amer. Statist. Assoc., Alexandria, Va.

Barron, A. R. and Cover, T. M. (1990). Minimum complexity density estimation. *IEEE Trans. Inform. Theory*. To appear.

Barron, R. L., Mucciardi, A. N., Cook, F. J., Craig, J. N. and Barron, A. R. (1984). Adaptive learning networks: Development and application in the United States of algorithms related to GMDH. In *Self-Organizing Methods in Modeling* (S. J. Farlow, ed.) 25–65. Dekker, New York.

Cover, T. M. (1974). The best two independent measurements are not the two best. *IEEE Trans. Systems Man Cybernet.* **4** 116–117.

Cox, D. D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16** 713–732.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–67.

Jones, L. (1990). A simple lemma on iterative sequences in Hilbert space and convergence rates for projection pursuit regression. Technical report 16, Dept. Math., Univ. Lowell, Lowell, Mass.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: Discrete index set. *Ann. Statist.* **15** 958–975.

Rissanen, Y. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** 416–431.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

Sheu, C.-H. (1989). Density estimation with Kullback–Leibler loss. Ph.D. dissertation. Dept. Statist., Univ. Illinois.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

Department of Statistics
University of Illinois
725 South Wright Street
Champaign, Illinois 61820

Leo Breiman

University of California, Berkeley

This is an exciting piece of methodology. The highest compliment I can pay is to express my feeling that "I wish I had thought of it." The basic idea is simple and powerful. The examples are interesting and illuminating. My sense of it is that this is a methodology that will become widely used in applications. Naturally, I have a few reservations and questions. But first, I want to express my sense of wonderment that this article is published in the *Annals of Statistics*.

There is not a single theorem, lemma or proposition in the whole paper. Have my senses taken leave of me? What, no asymptotics or results concerning the rate at which MARS approaches the "true model" as the sample size goes to infinity? For one of the few times in its history, the *Annals of Statistics* has published an article based only on the fact that this may be a useful methodology. All is ad hoc; there is no maximum likelihood, no minimax, no rates of convergence, no distributional or function theory. Is nothing sacred? What kind of statistical science is this? My thanks go to the editor and the others involved in this sacrilegious departure.

Now on to issues concerning Friedman's article:

If one fits a linear regression to data, then one is projecting the data onto a fairly small space—the set of all linear combinations of the x-variables. But now, suppose that one has 100 x-variables and perhaps 200 data cases and we are trying to find the best linear predictor of y based on a subset of the x-variables.

There are billions of different subsets of the x-variables. There are a few standard methods for choosing between these subsets. Forward variable addition can be used and so can backwards variable deletion, and with enough computing power, the minimum RSS least-squares regression equation using a specified number of variables can be found.

The procedure Friedman proposes is analogous to forward variable selection. There are a very large number of variables, consisting of all tensor products of spline functions. Forward stepwise addition is used up to a point, followed by stepwise deletion. The dimensionality of the final model is governed by what Friedman calls generalized cross-validation, but what is actually an adjusted residual sum-of-squares, with only a distant connection to true cross-validation.

MARS defines a very large class of candidate models by the specification of a large set of basis elements. Model selection is equivalent to selection of a subset of basis elements, since the coefficients are then defined by least-squares regression. For data with 10 variables and a sample size of 100, there are 1000 univariate splines and 450,000 bivariate spline products, where I am counting only splines zero to the left assuming that to each data value of each variable, there is a spline with knot at that point. The number of different candidate models using, say, 10 of these basis elements is staggering.

In principle, the way Friedman would get the best of all candidate models is to compute the PSE for all such models and select the one having the lowest PSE. Since this is not possible, the GCV is used as an estimate for the PSE and basis elements are found by a stepwise forward addition method. This procedure raises problems which are important not only to MARS, but to the entire venture of fitting more general multivariate models.

1. The set of candidate models in MARS may be too large.

A. *The packing problem*. The predecessor to MARS is TURBO. This is the program for fitting additive models reported on in Friedman and Silverman (1989). In the discussion of this paper, Trevor Hastie criticized TURBO for having high variability. He generated 50 data sets of sample size 100 from the model:

$$y = 0.667\sin(1.3x_1) - 0.465x_2^2 + \varepsilon,$$

where ε, x_1, x_2 are $N(0, 1)$ and x_1, x_2 have correlation 0.4. He ran TURBO on this data and plotted the resulting transformations. These graphs are given in Figure 1. Later, as I began working with additive models using different construction methods, I understood better what Hastie was driving at.

Consider the following simple method for constructing additive models—put K knots down on each predictor variable and using the power basis for splines, do stepwise backward spline deletion. Decide how many splines to leave in the model by finding the minimum value of the cross-validation estimate of PSE.

Initially, I had thought that the value of K would not be critical as long as it was large. For instance, I might typically begin with 15–20 knots per variable and then do the deletion. The reasoning for taking K large was to have plenty of knots around to fit the functions. I thought that having too many would not be a problem since all but a few would be deleted.

Much experimentation later, I realized that I was wrong. If the process was started with K too large, then the resulting models were noisy and could contain odd artifacts due to local quirks in the data. One way to think of this is that the deletion process forms a path through the space of all candidate models. The larger the space of candidate models, the more tightly they are packed together and the path will be forced to select between nearby models on the basis of small local properties. The result was that not only were noisy transformations produced, but also that prediction error increased as K got too large.

The procedure finally selected was this: For each value of K from one on up, set K initial knots on each variable, go through the deletion process and let PE(K) be the minimum cross-validated PSE estimate encountered in the deletion. Now select K to minimize PE(K). This process was carried out using Hasties data and resulted in the graphs in Figure 2. For more details see Breiman (1989a).

The lesson is that relative to some measure of the efficacy of the data the class of candidate models should not be packed too tightly together. Otherwise the results will be noisy, possibly containing local artifacts and with a loss in prediction accuracy. My concern is that the candidate models in MARS are tightly packed together. There are many more candidate models than in TURBO. The examples do not seem to show any signs of the packing problem, but we comment further on this later.

This is really not a criticism particular to MARS. It could apply also to CART and to ACE. It appears to me as a fundamental issue in model fitting. The larger the class one selects from, the more sensitive the procedure is to

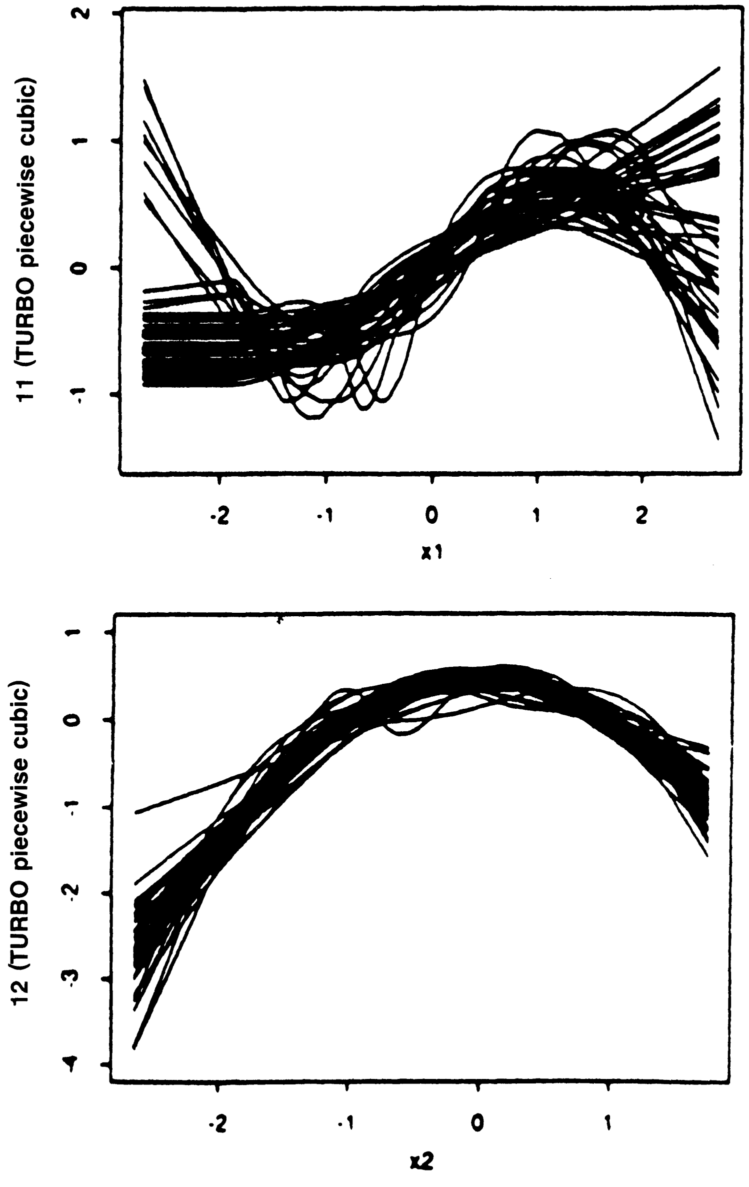


FIG. 1.

noise. This issue cries for some theoretical investigation and deserves at least as much energy and attention as one-dimensional density estimation.

B. *The Rashumon effect.* Suppose that we assume that we actually know or can compute the PSE for each model and can go along with the idea that the best model is the one with minimum PSE. From a predictive point of view,

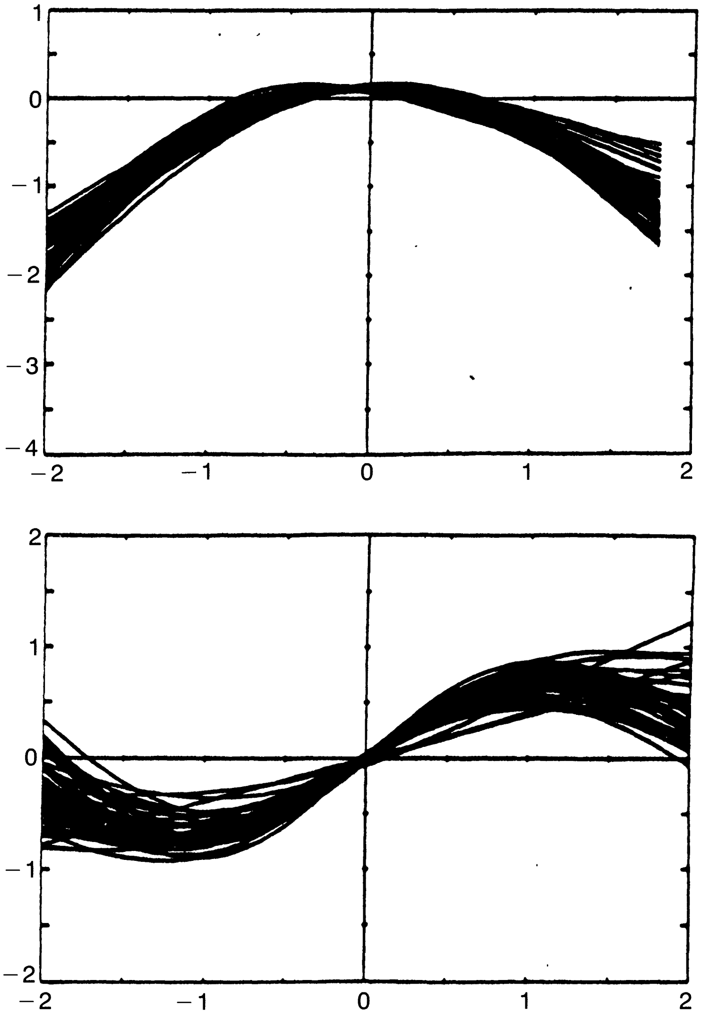


FIG. 2.

this is a perfectly defensible procedure. However, very often predictive procedures are carried out for the purposes of interpretation. Then the question posed is "how well does the estimated function mimic the TRUE function?" or implicitly, "how well can we recover the mechanism used for generating the data?"

Usually, this question is dealt with by setting up simulated data where the true function is known and seeing how well the estimation reproduces the known function. This is the strategy followed in Friedman's examples of Sections 4.2 and 4.3. That this works is almost always due to the simple structure of the simulated data. In most cases of complex real data, we are up against the Rashumon effect.

For instance, consider the best subsets procedure in regression for choosing the best regression equation depending, say, on five variables, out of 30. If one prints out the residual sum-of-squares for the, say, 10 lowest RSS equations depending on five variables, then most often the first few of these will have RSS values within a smidgen of each other. Yet the variables used may be quite different. The analogous effect would take place if we could compute the 10 lowest PSE equations.

The major cause of this lack of uniqueness lies in the sheer size of the class of candidate models and in the dependence between the basis elements. Now, if we assume that a model with low PSE gives a good picture of the data generating mechanism, then what we are getting is a multiplicity of equally good, but different, pictures of what goes on within the black box.

Thus, for complex data, there can be many different and equally valid (or equally invalid) pictures of the mechanism generating the data. Unfortunately, most procedures will produce only one picture: that is, running MARS on a data set will give only one picture. Yet there may be other models based on much different sets of basis elements that give either as low as or lower PSE.

Unfortunately, much of classical statistics is predicated on there being one unique and best answer. The data emanates from a black box, so the idea is to assume a stochastic model for the mechanism in the inside of the block box, estimate a few parameters and, bingo, we know what truth is. But for creative data analysis, the desideratum is to get as many different views as possible of what may be going on. Given this, if I were running MARS, then my predilection would be to run it on a number of bootstrap or leave-some-out samples and see what different results emerged.

2. RSS or GCV is not PSE. Another implicit assumption made in many model fitting procedures is that all other things being equal (for instance, in comparing two models both of which use the same number of parameters) that the model with lower RSS will have lower PSE. This is assumed in MARS, since for the same M, the lower RSS model will have lower GCV.

Unfortunately, this assumption is not valid. For the same dimensionality, the minimum RSS model may be quite different than the minimum PSE model and the PSE corresponding to the minimum RSS model may be considerably higher than the PSE of the minimum PSE model. Is this an inherent and unsurmountable difficulty, or is there some way around it?

MARS uses the GCV values to select dimensionality of the final model. No matter what you call it, the GCV criterion is not cross-validation. The reason for GCV is computational efficiency. Tenfold cross-validation would take about ten times as long, and MARS is not all that fast to begin with. Friedman has a number of examples showing that his version of GCV does a pretty good job. But I still have some reservations.

For instance, in the example modeling pure noise, about half of the time MARS produces a model that has a better GCV score than estimating the noise by its average. Would using cross-validation improve on this? Again, the problem is interpretation. Fitting noise with some structure can lead to embarrassing conclusions.

Near the end of Section 3.6, Friedman puts up a fight for GCV based on simulation results and claims that "the resulting model and its accuracy are seen to be fairly independent of the value chosen for the parameter d." He concludes that the best value for d is between 2 and 4, that 3 is fairly effective and that the accuracy of the result is not sensitive to the value of d in the 2–4 range. This is contrary to my experience in other contexts.

Selecting the dimensionality of the model used is critical. Selecting too large a model leads to inflated variance and too small to lack-of-fit bias. But simulations have shown that GCV in linear regression usually selects too large a model, whereas cross-validation or bootstrap do a good job in selecting the right-sized model [Breiman and Spector (1989)]. Therefore, even at increased computational cost, I would suggest that the author include a CV or bootstrap facility in the MARS program.

3. Data is not always high signal to noise. With the exception of the pure noise example, all of the examples given by Friedman had high signal-to-noise (s/n) ratios. For instance, the example of (56) had s/n = 3.28 (91% of variance explained). The example of (61) had s/n = 4.8 (96% of variance explained). The circuit examples of Section 4.4 had s/n = 3 (90% of variance explained). The olive oil example of Section 4.5 had a 3–5% misclassification rate. Finally, the example of equation (66) has s/n = 3.15 with 91% of the variance explained. For the other simulated example (67), the signal-to-noise ratio was not specified.

Any propensity of MARS to produce artifacts due to the noisy behavior referred to earlier will be most apparent in moderate-to-low signal-to-noise ratios. To the extent that Friedman has stayed away from such data, the impression given by the examples in the paper may be misleading.

Even so, there are some disturbing results in the examples. For instance, for the additive data of Section 4.2, the number of times that a nonadditive model is preferred by GCV increases as the sample size increases. For the data of Section (4.3) with one bivariate interaction, allowing an unlimited number of interactions is about as good as allowing only bivariate interactions. Can the author give explanations for these results?

4. Is stepwise forward the only way to go? Stepwise forward procedures make me a bit apprehensive. There is always the risk that with a poor step in the initial phases, it will produce a decidedly suboptimal fit. There is a similar problem in CART. While with tree-structured procedures we have been unable to come up with computationally effective alternatives to stepwise forward splitting, in fitting continuous functions to multivariate data there are other methods that have appeared in the literature.

For fitting additive equations, there is the backfitting method used in ACE, with continued research in the Buja, Hastie and Tibshirani (1989) article.

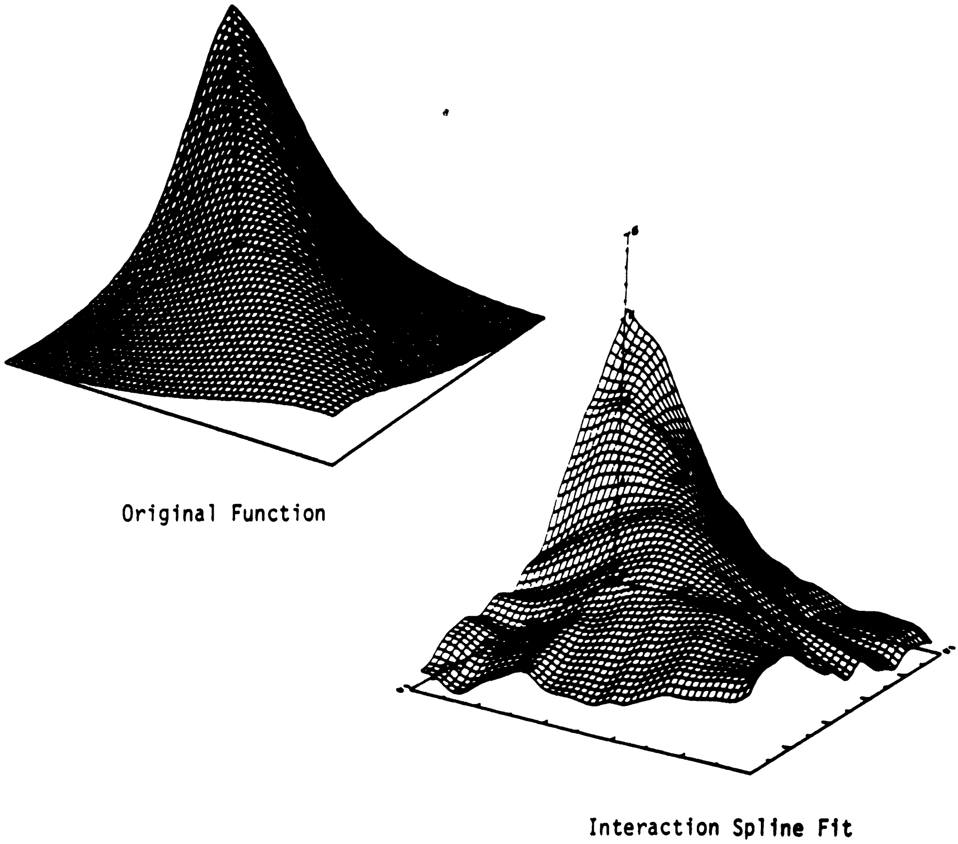


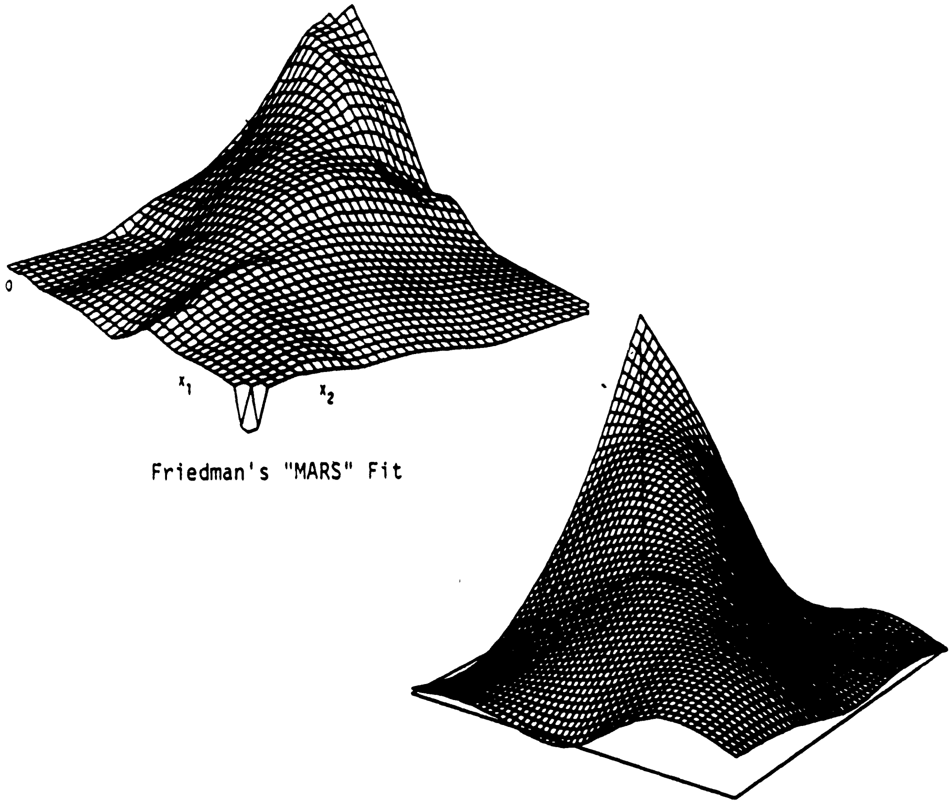
FIG. 3.

Another interesting method using backfitting was proposed by Hastie in his discussion of the Friedman and Silverman (1989) paper. There is also the work mentioned earlier doing backward knot deletion [Breiman (1989a)].

There has been less work on fitting interaction surfaces. This is where MARS breaks new ground in being the first published method that has an effective approach to the problem. However, as Friedman points out, the group at Wisconsin is making progress in the computation of interaction splines. There is also another method which depends on the decomposition of the function to be estimated into a sum of products of univariate functions [Breiman (1989b)].

This Π-method has given promising results. To illustrate this, we ran it on the example given in Section 4.6, equation (66), which originally appeared in the Chong, Bates, Chen and Wahba (1988) paper. Figure 3 shows the original function, the interaction spline fit, the MARS fit and the fit of the Π-method.

None of the alternative methods are as fully developed as MARS. The MARS algorithm, with the setting of a few parameters, produces a fit up to



Friedman's "MARS" Fit

Π - Method Fit

FIG. 3. (*Continued*)

whatever degree of interaction is wanted. Whether other methods can provide improved accuracy and comparable automation remains to be seen.

5. Quo vadis? The development and use of effective multivariate methods for fitting complex data is an endeavor largely carried on outside of statistics of diverse and active groups interested in results, rather than theorems. For instance, most of the CART applications that we know about have been done by nonstatisticians. The rapidly growing field of neural networks is built around a new class of algorithms for multivariate regression and classification with the principle protagonists being engineers and computer scientists. It was gratifying to find that at a recent neural network conference there was widespread knowledge of CART. I think that MARS will similarly become widely known and used in application areas.

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. *Technometrics*. To appear.

BREIMAN, L. and SPECTOR, P. (1989). Submodel selection and evaluation in regression *X*-random case. *Internat. Statist. Rev.* To appear.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555.

CHONG, G., BATES, D. M., CHEN, Z. and WAHBA, G. (1988). The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models. Technical report 823, Dept. Statist., Univ. Wisconsin.

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–39.

DEPARTMENT OF STATISTICS
367 EVANS HALL
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

GEORGE K. GOLUBEV AND RAFAEL Z. HASMINSKII

Institute for Problems of Information Transmission

J. H. Friedman presents the new recursive method of regression estimation for high dimensional data. This method is very interesting and has very good perspective. The main idea is an adaptive and recursive construction of the system of basis functions. The proposed estimation method has good flexibility and it is convenient for computer realization. We think that this approach is applicable for other nonparametrical estimation problems, for instance in the spectral density estimation for stationary Gaussian data.

The interesting problem connected with the proposed method is the theoretical study of quality of this method for different classes of smooth regression functions. (The reasons for consideration of the classes of smooth functions lie not only in practical importance of such constraints. From our point of view the most important theoretical results are established for these functional classes.) Let us recall some known results in this direction.

1. The best in minimax sense order of the rate of convergence of the L_p, $1 \leq p < \infty$, risks to zero for the regression function of the smoothness of β in $\mathbb{R}^k$ is equal to $n^{-\beta/(2\beta+k)}$ [Ibragimov and Hasminskii (1980) and Stone (1982)].
2. Speckman (1985) and Nussbaum (1985) found regression estimators which cannot be improved, not only in the sense of order of the rate of convergence but also in the sense of constant. Impossibility of improvement (in minimax sense) of this constant for special case ellipsoids in the Sobolev spaces and integrated mean-squared error was proved by Nussbaum (1985), who used the results of Pinsker (1980).