MORGAN, J. N. and SONQUIST, J. A. (1963). Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.* **58** 415–434.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

SHEPARD, D. (1964). A two-dimensional interpolation function for irregularly spaced data. *Proc. 1964 ACM Nat. Conf.* 517–524.

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

SHUMAKER, L. L. (1984). On spaces of piecewise polynomials in two variables. In *Approximation Theory and Spline Functions* (S. P. Singh, J. H. W. Barry and B. Watson, eds.) 151–197. Reidel, Dordrecht, Holland.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

SMITH, P. L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. Report NASA 166034, Langley Research Center, Hampton, Va.

STONE, C. J. (1977). Nonparametric regression and its applications (with discussion). *Ann. Statist.* **5** 595–645.

STONE, C. J. and KOO, C.-Y. (1985). Additive splines in statistics. *Proc. Ann. Meeting Amer. Statist. Assoc. Statist. Comp. Section* 45–48.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictors (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

WAHBA, G. (1986). Partial and interaction splines for the semiparametric estimation of functions of several variables. In *Computer Science and Statistics: Proc. Eighteenth Symp. on the Interface* (T. J. Boardman, ed.) 75–80. Amer. Statist. Assoc., Alexandria, Va.

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

DEPARTMENT OF STATISTICS
SEQUOIA HALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DISCUSSION

ANDREW R. BARRON AND XIANGYU XIAO

University of Illinois

1. Introduction. We describe the multivariate adaptive polynomial synthesis (MAPS) method of multivariate nonparametric regression and compare it to the multivariate adaptive regression spline (MARS) method of Friedman (1990). Both MAPS and MARS are specializations of a general multivariate regression algorithm that builds hierarchical models using a set of basis functions and stepwise selection. We compare polynomial and spline bases in this context. Our experience is that there is no substantial difference in the statistical accuracy for the data sets that we have investigated, provided that some care is taken in the choice of the model selection criterion. It is argued that the polynomial methods, with a smaller set of basis functions to select from at each step, should yield a computationally faster algorithm.

A potential difficulty of either polynomial methods with high-order terms or spline methods with closely spaced knots is the high sensitivity of the estimated response to slight changes in the inputs. We advocate the use of a roughness penalty in the performance criterion that forces smoother models to be accepted. A consequence of this roughness penalty in the polynomial case is that large coefficients in high-order terms are avoided. We believe that the MARS algorithm could also benefit from the use of the roughness penalty.

Polynomial networks are synthesized by a simple variant of the algorithm. In particular, an option of the MAPS algorithm is to allow the outputs of tentatively selected models to be considered along with the original explanatory variables as inputs for subsequent steps of the algorithm. This algorithm exhibits many of the properties of more complicated polynomial networks and other "neural" networks for multivariate regression. For an overview of statistical learning networks, see Barron and Barron (1988).

The advantage of adaptively synthesized model structure compared to fixed model structure is the opportunity to seek out accurate lower dimensional nonlinear models in the high dimensional space of functions of several variables. MARS, MAPS and some adaptive network techniques have the potential for selecting accurate yet parsimonious models in these high dimensional settings.

2. Adaptive regression. Multivariate adaptive regression is a stepwise procedure for the automatic selection of basis functions from observed data. The selected basis functions $B_m(\mathbf{x})$ yield models of the form

$$f_M(\mathbf{x}, \theta) = \sum_{m=1}^{M} \theta_m B_m(\mathbf{x})$$

for $\mathbf{x}$ in $\mathbb{R}^n$. These models are fit to observed data $(\mathbf{x}_i, y_i)_{i=1}^N$.

Briefly, the forward algorithm takes the following form. The initial function estimate is taken to be a constant by setting $B_1(\mathbf{x}) = 1$. Then the following steps are repeated until a model selection criterion stops the growth of the model. From a list of candidate basis functions Γ_M, we choose one or two new basis functions $B_{M+1}(\mathbf{x})$, $B_{M+2}(\mathbf{x})$ to append to the current list of basis functions $\{B_1(\mathbf{x}), B_2(\mathbf{x}), \ldots, B_M(\mathbf{x})\}$ and then regress the data onto the span of the new list. At each step we adopt the choice of basis functions that provide the best improvement in a model selection criterion.

The key to the algorithm is the specification of a parsimonious yet flexible set Γ_M of candidate basis functions (or pairs of candidate basis functions) from which the new terms $B_{M+1}(\mathbf{x})$ and in some cases $B_{M+2}(\mathbf{x})$ are selected. Naive choices of Γ, such as the set of all polynomial terms in n variables up to a given degree, are exponentially large sets in the dimension n and, consequently, would be computationally prohibitive in moderately high dimensions.

Friedman's strategy is to adapt the set Γ_M of candidate basis functions to the current list of terms, by taking the set of products of terms in the current list with one-dimensional basis functions. In particular, MARS (with $q = 1$)

takes Γ_M to be the set of pairs of candidate terms $B_m(\mathbf{x})[\pm(x_j - t)]_+$ for $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, n$ with the restrictions that x_j is not already in a factor of $B_i(\mathbf{x})$ and t is in a list of candidate knot locations determined by the sample quantiles of x_j.

The MAPS algorithm, in its simplest form, takes $\Gamma_M = \{B_i(\mathbf{x})x_j\colon\ i = 1, \ldots, M,\ j = 1, \ldots, n\}$. Thus each step of the MAPS procedure yields a polynomial, where in one coordinate the degree is incremented by one. Experimentation has revealed that in some cases it is better to introduce pairs of basis functions $B_i(\mathbf{x})x_j$ and $B_i(\mathbf{x})x_j^2$ on each step. This has two benefits: first, it tends to orient the search in early steps toward models with low interaction order that are more reliably estimated and second, it avoids some of the traps of stepwise selection. [For instance, if $y = x_1^2$ and x_1 is symmetrically distributed about the origin, then the linear regression of y onto the span of $\{1, x_1\}$ is constant and, consequently, without forced consideration of higher-order terms, the stepwise algorithm would stop with a constant as its best estimate.] With this modification, each step of the MAPS algorithm selects $B_{M+1}(\mathbf{x})$ and $B_{M+2}(\mathbf{x})$ from the set

$$\Gamma_M = \left\{B_i(\mathbf{x})x_j,\ B_i(\mathbf{x})x_j^2\colon i = 1, \ldots, M,\ j = 1, \ldots, n\right\}.$$

Here the second new term may be rejected from inclusion at the current step if it does not yield an improvement in the performance criterion, whereas the first new term may be rejected only after consideration of the performance with both new terms.

As in Friedman's MARS algorithm, MAPS provides an option to restrict the order of interaction of candidate terms to be not greater than a specified limit mi. Here $mi = 1$ yields an additive polynomial, $mi = 2$ allows cross terms $B(\mathbf{x}) = x_i^{r_i}x_j^{r_j}$ and $mi = n$ allows general polynomial terms $B(\mathbf{x}) = x_1^{r_1}x_2^{r_2} \cdots x_n^{r_n}$ to eventually be synthesized by the algorithm. Even with $mi = n$, the algorithm often stops before such high-order interactions are considered.

Following the completion of the forward stepwise algorithm, it is advisable to perform a backward stepwise pass that at each step removes the term that permits the best improvement in the performance criterion for the remaining terms. This backward pass is implemented by Friedman. With the backward pass implemented, one is free to allow the forward pass to proceed past the point at which the performance criterion is optimized with respect to forward selection. Extraneous terms are left to be removed by the backward pass provided their removal is determined to be beneficial. At the present time, we have only implemented the forward stepwise synthesis in the MAPS program.

We hasten to point out that stepwise selection procedures cannot in general be guaranteed to provide the best set of terms of a given size; see, for instance, Cover (1974). To get the best set of terms essentially exhaustive subset selection procedures would be needed. Such exhaustive procedures are feasible for certain linear regression problems, but they are not feasible for multivariate nonlinear regression, because of the exponential explosion of number of terms from which the subsets are selected. Therefore, stepwise selection is a necessary compromise in the multivariate nonlinear case.

In the related context of iterative L^2 approximation of functions, a recent result of Jones (1990) states that if the target function is in the closure of the convex hull of a given set of functions with bounded norm, then the squared norm of the error is of order $O(1/m)$ for an m-step forward stepwise selection. That result can be adapted to the context of adaptive polynomial regression with Γ equal to all polynomial terms up to a given degree, to yield bounds on the statistical risk. Thus forward stepwise selection can yield a reasonably accurate set of terms even if it is not, strictly speaking, the best set of terms. It is not clear whether theory analogous to that provided by the result of Jones can be developed that applies to the smaller sets of candidate terms used by MARS or MAPS.

The approximation capabilities of polynomials and splines are known in the case of complete bases (in which all polynomial terms up to a prescribed order are included). These results show, for instance, that for any given $k \geq 1$, if $f(x)$ is a k times differentiable function on $[0, 1]^n$, then there exists a polynomial $f_k(\mathbf{x}) = \sum \theta_r x_1^{r_1} \cdots x_n^{r_n}$, where the sum is for all $r = (r_1, \ldots, r_n)$ with $0 \leq r_j \leq k$, for which the L^2 approximation error is bounded by

$$\int_{[0,1]^n} (f(\mathbf{x}) - f_k(\mathbf{x}))^2 \, d\mathbf{x} \leq \frac{1}{(2k+1)!4^k} \sum_{d=1}^{n} \int \left(\frac{\partial^k}{\partial x_d^k} f(\mathbf{x}) \right)^2 d\mathbf{x}.$$

This particular bound is a specialization of the multivariate extension in Sheu (1989) of a bound due to Cox (1988). It shows the exponential convergence rate of polynomial approximation for analytic functions, assuming that the norm of the k partial derivatives is bounded by a multiple of $k!$. Spline approximations of a fixed order q, which are chosen to have roughly the same number of basis functions k^n, are not capable of the same accuracy. The integrated squared error saturates at the slower polynomial rate $(1/k)^{2q}$; see for instance, Schumaker (1981).

Unfortunately, there is not yet an analogous theory for the approximation with subsets of the complete set of basis functions. An exception is the case of additive approximation. For instance, to approximate the additive part of a function, the best additive polynomial approximation [which uses only $nk + 1$ terms instead of $(k + 1)^n$ terms], achieves the same accuracy as derived above by Sheu, whereas spline approximation again saturates at the slower rate. It also may be possible to use the theory to characterize the error of approximation of the second-order interaction component of a function. Such a theory would hopefully show that parsimonious approximation is possible for all function with negligible higher-order interactions.

3. Complexity penalties for adaptive regression selection. At each step of the adaptive regression algorithms, terms are chosen to optimize a statistical performance criterion. The criterion depends on the average-squared residuals (ASR) but incorporates a modification to penalize the number of parameters.

There is a proliferation of criteria that have been proposed for model selection. They can be roughly categorized into two groups. The first group seeks to estimate the mean-squared error of prediction $\text{MSEP}_{M,N} = E(Y - f_M(X, \hat{\theta}))^2$ or related quantities of cross-validation, where X, Y denotes a sample drawn independently of the training data. The idea is that the best model is the one with the minimum $\text{MSEP}_{M,N}$. Criteria that estimate the MSEP can be interpreted as adding a penalty to ASR which is roughly equal to $2(M/N)$ times an estimate of the variance of the error incurred by the best function of $\mathbf{x}$, where M is the number of parameters and N is the sample size. A representative criterion in this group is the generalized cross-validation (GCV), a modification of which is used by Friedman in his MARS program. For models of the form $f_M(\mathbf{x}, \hat{\theta}) = \sum_{m=1}^{M} \hat{\theta}_m B_m(x)$, the generalized cross-validation (not accounting for the selection bias) takes the form

$$\text{GCV} = \frac{\text{ASR}}{(1 - M/N)^2},$$

where $\text{ASR} = (1/N)\sum_{i=1}^{N}(y_i - f_M(\mathbf{x}_i, \hat{\theta}))^2$ is the average-squared residual. See, for instance, Eubank (1988) for properties of the generalized cross-validation and its relationship to other criteria, including Mallows' C_p, Akaike's final prediction error FPE and Akaike's information criterion AIC. The predicted-squared error PSE criterion studied in Barron (1984) is defined as $\text{PSE} = \text{ASR} + 2(M/N)\sigma^2$, where σ^2 is either the known error variance $E(Y - E(Y|X))^2$ or a rough estimate of it provided prior to the stepwise selection process. Under certain formulations, it is equivalent as a selection criterion to C_p and AIC. The final predication error $\text{FPE} = \text{ASR}(1 + M/N)/(1 - M/N)$, is the minimum variance unbiased estimator of the mean-squared error of prediction, in the case of a correctly specified model and Gaussian errors. A surprising fact, shown in Barron (1984), is that for reasonable choices of σ^2, PSE is a more accurate estimate of the mean-squared error of prediction for a given model.

All of these criteria that estimate $\text{MSEP}_{M,N}$ for each model M have the problem that they are not necessarily uniformly accurate if too large a set of candidate models is considered. This leads to the problem of selection bias. The minimized criterion value may be significantly smaller than the value for the best model, and the selected model may be overfit.

At first we ran our MAPS program with the GCV criterion, to facilitate comparison with the MARS procedure, believing that more conservative criteria would probably not be necessary in our case. However, as the results show in Section 5, we encountered persistent problems of overfit. This overfit invariably occurred whenever the selection was taken over a very large set of candidate terms, a large fraction of which are spurious. Friedman avoids the overfit problem by modifying the GCV criterion. He replaces the number of parameters M with an associated cost $C(M)$ in the expression $\text{GCV} = \text{ASR}/(1 - C(M)/N)^2$, where $C(M)$ is between $3M$ and $5M$. After encountering the overfit problems with the ordinary GCV, we found greater success

using criteria which are specifically designed for selection and not just for the estimation of risk.

This second group of criteria, to which we turned our attention, includes criteria designed to approximate the test statistic that minimizes the overall probability of error in a Bayesian formulation of the model selection problem [such as the BIC of Schwarz (1978)] or seeks to approximate the length of an asymptotically optimal information-theoretic code that describes the observed response values given the explanatory variables [such as the MDL criterion of Rissanen (1983)]. For either case, there is a formulation in which the dominant terms of the statistic define a criterion equivalent to the following:

$$\text{BIC} = \text{MDL} = \text{ASR} + \frac{M}{N}\sigma^2 \ln(N).$$

This criterion is similar to the first group of criteria, but it incorporates a penalty which is a factor $(\frac{1}{2})\ln N$ greater. For N between 50 and 400, the factor $(\frac{1}{2})\ln N$ is between 2 and 3. So for conservative values of σ^2 (values believed to be not smaller than the true error variance), the BIC/MDL criteria and Friedman's modification of the GCV criterion should give similar results. Indeed, it appears that in practice, Friedman's modified GCV is closer to the BIC/MDL criterion than the original GCV criterion upon which the modification is based.

The MAPS algorithm is set up to compute any of the previous criteria, GCV, FPE, PSE and MDL, as well as the AIC and BIC criteria that obtain when the error variance σ^2 is regarded as an unknown parameter. An option selects which criterion is used for the minimization.

For theoretical properties, the work of Shibata (1981) and Li (1987) demonstrates an asymptotic optimality property satisfied by any of the criteria in the first group [with penalty equal to $2(M/N)$ times a consistent estimate of σ^2]. In particular, Li (1987) gives conditions such that if $\text{MSE}_{M,N} = E(\hat{f}_M(X) - f(X))^2$ denotes the mean-squared error in the estimation of $f(X) = E(Y|X)$ by a linear model M fit by ordinary least squares, then the mean-squared error $\text{MSE}_{\hat{M},N}$ incurred by the selected model $\hat{M}$ satisfies $\text{MSE}_{\hat{M},N}/\text{MSE}_{M_N,N} \to 1$ in probability as $N \to \infty$, where $\text{MSE}_{M_N,N} = \min_M \text{MSE}_{M,N}$. The minimizations are assumed to be taken for a fixed sequence of lists of models H_N, rather than for an adaptively determined list. The theory assumes a condition that effectively limits the asymptotic number of candidate linear models that may be considered. Namely, the quantity $\Sigma_{M \in H_N}(\text{MSE}_{M,N}N)^{-r}$ must be negligible (as $N \to \infty$), for some r for which the $2r$th moment of the distribution of the error $(Y - E(Y|X))$ is finite. For very large sets of candidate models, this quantity is not negligible and the theory is not applicable. Indeed, in this case, significant selection biases can occur that are characterized by a tendency to overfit. Another implication of Li's condition is the requirement that the mean-squared error for the best sequence of models $\text{MSE}_{M_N,N}$ tends to zero slower than the rate $(1/N)$ that is achieved if the true function were finite-dimensional. In the case that the true function $f(X)$ is in one of the

finite-dimensional families, it is known that models selected by criteria in the first group have nonzero probability of asymptotically selecting an overfit model.

Asymptotic theory for model selection by the more conservative BIC or MDL criteria is given in Barron and Cover (1990) and Barron (1989, 1990). This theory gives conditions such that the mean-squared error of the selected model $\mathrm{MSE}_{\hat{M},N}$ converges to zero at rate bounded by $\mathrm{MSE}_{M_N,N} + (M_N/N)\ln N$. Convergence at this rate holds in both parametric and nonparametric cases and holds without restriction on the number of candidate models. As for the Shibata and Li theory, it is assumed that the criterion is optimized for a fixed sequence of lists of models, indexed by the sample size, rather than optimized stepwise for an adaptively determined set of models. Nevertheless, it suggests useful guidelines that might also be appropriate in the adaptive context. Chief among these is the need for care in the choice of criteria when a very large number of candidate models are considered. Somewhat larger penalties are required for accurate model selection in this case.

4. Roughness penalty for polynomial smoothing. Essential to polynomial methods of regression in the presence of noise and/or model uncertainty is the use of a criterion which incorporates a roughness penalty. In particular, the MAPS algorithm chooses the parameters of each model so as to minimize

$$\mathrm{ASR} + \mathrm{RP},$$

where ASR is the average-squared residual

$$\mathrm{ASR} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f(\mathbf{x}_i,\theta)\right)^2$$

and RP is the roughness penalty

$$\mathrm{RP} = \delta^2\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla_{\mathbf{x}} f(\mathbf{x}_i,\theta)\right\|^2.$$

Then ASR + RP is used in place of the average-squared residuals ASR in the model selection criteria discussed in the preceding section.

Here δ is a parameter that controls the smoothness of the model. One interpretation of the roughness penalty is that it captures the sensitivity of the average-squared error to slight changes in the input variables. If the inputs are permitted to be changed from x_{ji} to $x_{ji} \pm \delta$, then ASR + RP is an estimate of the average squared error that would be incurred by the function when perturbed inputs. The δ may often be set by the scientist or engineer who supplies the data as quantifying the size of changes in the input that should not be accompanied by significant changes in the response. Or it may be set by the statistician by inspection of a few runs to determine the one with which he is most satisfied. The selection of δ may also be automated by generalized cross-validation, but at considerable additional computational expense.

There is a relationship between polynomial smoothing and smoothing splines. The second order smoothing spline arises as the solution to the problem of minimization of

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2 + \delta^2 \int \|\nabla f(\mathbf{x})\|^2.$$

For smoothing splines, the minimization is taken over all continuously differentiable functions $f(\mathbf{x})$. In contrast, for polynomial smoothing, the integral is replaced, for convenience, with the sample average and the minimization is taken over the restricted class of polynomial functions with specified bases. It is our experience that polynomial smoothing approximates the capabilities of spline smoothing, while providing advantages of speed of computation, due to the reduced size of the dimension of the linear system that is solved to obtain the polynomial approximation.

It is seen that polynomial smoothing with a roughness penalty is a generalized form of ridge regression. For a linear model

$$f(\mathbf{x}, \theta) = \sum_{j=1}^{m} \theta_j B_j(\mathbf{x}),$$

the roughness penalty is a positive-definite quadratic function of the parameters $\theta = (\theta_1, \ldots, \theta_m)^T$,

$$\text{RP} = \theta^T R \theta,$$

where R is the m by m matrix with entries

$$R_{jk} = \delta^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{d=1}^{n} \frac{\partial B_j(\mathbf{x}_i)}{\partial x_d} \frac{\partial B_k(\mathbf{x}_i)}{\partial x_d}.$$

The roughness penalty prevents the estimation of models with large coefficients for terms that contribute large derivatives. With polynomial basis functions $B_j(\mathbf{x})$, each derivative $\partial B_j(\mathbf{x}_i)/\partial x_d$ is also a polynomial but with the degree reduced by one in the coordinate x_d. The larger derivatives are typically associated with the higher-order terms.

The set of equations to be solved for θ is obtained as follows. Let $V = (1/N)\mathbf{B}^T\mathbf{B}$ and $c = (1/N)\mathbf{B}^T\mathbf{y}$ be defined as in Friedman, equation (49), where $\mathbf{y} = (y_1, \ldots, y_N)^T$. Assume that the sample average of the response variable has been subtracted off so that $\bar{y} = (1/N)\Sigma_{i=1}^{n} y_i = 0$ and let $\hat{\sigma}_y^2 = (1/N)\|\mathbf{y}\|^2$ denote the sample variance of y. The penalized average-squared residual may then be expressed as the following quadratic function of θ,

$$\begin{aligned}\text{ASR} + \text{RP} &= (1/N)\|y - B\theta\|^2 + \theta^T R \theta \\ &= \hat{\sigma}_y^2 + \theta^T V \theta - 2c^T\theta + \theta^T R \theta \\ &= \hat{\sigma}_y^2 + \theta^T \tilde{V} \theta - 2c^T\theta,\end{aligned}$$

where $\tilde{V} = V + R$. The parameter vector $\hat{\theta}$ which minimizes this expression is

found by solving the modified normal equations

$$\hat{V}\hat{\theta} = c.$$

Each new basis function introduces a new row and column of R and V, so the solution may be updated by Cholesky decomposition in the same manner as explained by Friedman.

In its simplest form, the roughness penalty treats each of the variables equally. More generally, it may be desirable to associate a parameter $\delta_{d,i}$ for each value $x_{d,i}$ of the explanatory variables. The roughness penalty then takes the form

$$\mathrm{RP} = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{n}\delta_{d,i}^2\left(\frac{\partial f(\mathbf{x}_i,\theta)}{\partial x_d}\right)^2.$$

In like manner, the expression for the matrix R is modified to bring $\delta_{d,i}^2$ inside the double summation. The choice $\delta_{d,i} = \hat{\sigma}_{x_d}^2\delta$ allows the penalty to be scaled to the observed spread of the explanatory variables as measured by the standard deviation $\hat{\sigma}_{x_d}^2$. Alternatively, it may be desirable to let $\delta_{d,i}$ be proportional to the magnitude of the observed values of the variables, that is

$$\delta_{d,i} = |x_{d,i}|\delta.$$

The latter choice helps to mitigate the effect of extreme observations. Also, in the case of polynomial basis functions, it allows the entries in the matrix R to be determined as a weighted sum of entries of the matrix V, thereby avoiding much of the additional computational expense otherwise associated with the use of the roughness penalty. Specifically, R may be expressed as

$$R_{jk} = \delta^2\sum_{d=1}^{n} r_{jd}r_{kd}V_{jk},$$

where r_{jd} denotes the exponent of variable x_d in the basis function $B_j(\mathbf{x}) = x_1^{r_{j1}}\cdots x_n^{r_{jn}}$. Note that the off-diagonal entries R_{jk} of the matrix R are zero for those pairs of basis functions B_j and B_k that share no common factors. The largest entries are typically on the diagonal and correspond to the terms with large exponents.

Some of the characteristics of the roughness penalty are incorporated in Friedman's MARS algorithm. He adds a small multiple of the diagonal entries of V to numerically stabilize the resulting modified normal equations.

5. Experimental results. First, we took 10 replications of the simulated data from Friedman's Section 4.2, each with a sample of size $N = 200$. In this example, the dependence of y on $x_1,\ldots,x_{10}$ is additive, as given in Friedman's equations (56) and (57) and the inputs are drawn uniformly over the unit cube $[0,1]^{10}$. For each replication, the observed response was scaled to have sample mean zero and sample variance one, but the input variables are left unscaled. The parameter for the roughness penalty was set to be $\delta = 0.01$, a moderately small value that allows for the high gradients of the response near $x_1 = 1.0$

TABLE 1
Summary of the results of MAPS modeling with the unmodified GCV criterion on Friedman's additive data (Section 4.2)

mi	ISE	MSEP	GCV	TERMS
1	0.030(0.010)	0.16(0.01)	0.15(0.01)	12.0(1.4)
2	0.053(0.018)	0.18(0.02)	0.14(0.01)	17.3(2.9)
10	0.086(0.037)	0.21(0.31)	0.13(0.01)	20.7(5.2)

and near $x_2 = \frac{1}{2}$ in equation (56). The generalized cross-validation GCV (without modification) was used as the selection criterion. The results of the 10 MAPS runs for each interaction limit $mi = 1, 2, 10$ are summarized in Table 1. Depicted are the averages and standard deviations based on the 10 runs of the standardized integrated-squared error (ISE), the standardized mean-squared error of prediction (MSEP), the generalized cross-validation (GCV) and the number of selected terms (TERMS). In accordance with the definitions in Friedman, the integrated-squared error and the mean-squared error of prediction are computed using knowledge of the true function and 5,000 new sample points for Monte Carlo integration.

The results in Table 1 may be compared with those reported in Friedman's Table 4 in the $N = 200$ case. It shows that when the unmodified GCV criterion is used, if the model is forced to be additive, the polynomial method is nearly as good as the spline method; however, in the case that interaction terms are considered, $mi = 2, 10$, we have noticeably worse integrated-squared error. The large average numbers of terms 17.3 and 20.7 reveal that the polynomial models are overfit using the unmodified GCV criterion. Indeed, the first 11 or 12 terms were almost exclusively additive terms in the meaningful variables (x_1 through x_5), but the additional terms chosen in the $mi = 2$ and $mi = 10$ case were almost exclusively spurious cross product terms and terms involving the nuisance variables x_6 through x_{10}. This large number of spurious models contribute to the large selection bias that results in overfit with the unmodified GCV criterion.

We then repeated the experiment using the more conservative BIC/MDL criterion. In the definition of the BIC/MDL, we used the known variance of the noise $\sigma^2 = 1$. The results are summarized in Table 2. The results in Table 2 for adaptive polynomial modeling compare quite favorably with those for adaptive splines in Friedman's Table 4. Indeed, the averages and standard deviations of the integrated-squared error and the mean-squared error of prediction are either equally good or slightly better in every case. The most noticeable improvement is in the case that arbitrary interactions are allowed ($mi = 10$). With the BIC/MDL criterion, almost all of the spurious interaction terms are rejected.

Next, we drew one sample of size 100 in accordance with Friedman's example 4.3. Again there are 10 variables. The first two variables contribute to the response through a term which is a sinusoidal function of the product

TABLE 2
Summary of the results of MAPS modeling with the BIC/MDL criterion on Friedman's additive data (Section 4.2)

mi	ISE	MSEP	BIC, MDL	TERMS
1	0.025(0.006)	0.155(0.006)	0.16(0.01)	11.1(1.3)
2	0.030(0.007)	0.159(0.007)	0.16(0.01)	11.5(1.6)
10	0.031(0.007)	0.160(0.007)	0.16(0.01)	11.3(1.6)

x_1x_2. The remaining contributions are additive, specifically, a quadratic term in x_3 and linear terms in x_4 and x_5. (Admittedly, the polynomial terms in the true response give polynomial modeling an unfair advantage for this example.) The other variables x_6 through x_{10} are not used by the true response. We set $\delta = 0.01$, $\sigma = 1.0$, $mi = 2$ and standardized the observed response. With the BIC/MDL criterion, the terms and coefficients for the normalized model are given in Table 3 in the order in which they are selected. (The model is expressed in standardized form; it is unitized by multiplying by $\hat{\sigma}_y = 5.8189$ and then adding $\bar{y} = 15.0549$.) The criteria values suggest that the MAPS model and the MARS model are roughly equally accurate for this sample from Section 4.3.

Table 3 shows that apparently meaningful terms were selected by the MAPS algorithm with the BIC/MDL criterion, with the exception of the third term, which is a quadratic in x_4 while the true response is linear in x_4. It is anticipated that this term would be removed by a backward stepwise selection. When the unmodified GCV is used as the criterion, six more terms are

TABLE 3
Coefficients and exponents of the selected polynomial terms for the model in Example 4.3

Term	Coefficient	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	−1.6809	0	0	0	0	0	0	0	0	0	0
2	2.0970	0	0	0	1	0	0	0	0	0	0
3	−0.3138	0	0	0	2	0	0	0	0	0	0
4	0.2363	0	1	0	0	0	0	0	0	0	0
5	−0.2508	0	2	0	0	0	0	0	0	0	0
6	−0.3203	1	0	0	0	0	0	0	0	0	0
7	0.3347	2	0	0	0	0	0	0	0	0	0
8	−4.0523	0	0	1	0	0	0	0	0	0	0
9	3.9369	0	0	2	0	0	0	0	0	0	0
10	0.7853	0	0	0	0	1	0	0	0	0	0
11	−0.0794	2	1	0	0	0	0	0	0	0	0
12	−6.9382	2	2	0	0	0	0	0	0	0	0
13	6.9397	1	1	0	0	0	0	0	0	0	0
				GCV	BIC, MDL						
				0.041	0.049						

TABLE 4
Portuguese olive oil

Method	# Variables	GCV	CV	Error CV
MAPS ($mi = 2$)	7	0.15	0.22	0.036

selected, all of which involve interactions with the extraneous variables x_6, x_7, x_9 and x_{10}. Evidently, the extraneous variables introduce two many diverse candidate terms for the GCV to provide a uniformly accurate criterion.

Finally, we considered the Portuguese olive oil data, a copy of which was obtained from Friedman. We standardized the response variable to have sample mean zero and sample variance one. Inspection of the data set shows that the variables are rounded to the nearest one-tenth, so we set $\delta = 0.05$, accordingly. Using a 10-fold cross-validation as explained in Friedman, we ran MAPS 10 times each with 41 or 42 observations removed, using the unmodified GCV criterion and a maximum term limit of 30. The resulting standardized average-squared residual (CV) and relative frequency of misclassifications (error CV) on the cross-validated data is given in Table 4. Also shown is the unmodified GCV and number of variables obtained from the MAPS procedure with all observations included. With the GCV criterion, the MAPS procedure hit the maximum term limit of 30. (Subsequent runs with an increased term limit stopped at 42 terms, fueling suspicion of overfit). Despite the excessive number of terms selected with the unmodified GCV criterion, the cross-validation results in Table 4 for MAPS are as good as obtained by Friedman in Table 13 with MARS and the least-squares criterion. Nevertheless, improved cross-validation results may be possible with MAPS using a more conservative performance criterion.

In Table 5, we show the selected model, when MAPS is run with the BIC/MDL criterion and $\sigma^2 = 0.03$ (which corresponds to a standardized variance of 0.18). Here we standardized both the observed explanatory variables and the response variables to have sample mean zero and sample variance one. In this case a more parsimonious model (16 terms) is selected. The values of the BIC/MDL and GCV criteria for this selected model appear to be reasonable, but we have not yet completed the 10-fold cross-validation of models selected by BIC/MDL to provide additional confirmation of the apparent accuracy.

6. An adaptive network example. In this last section, we illustrate a simple feedforward polynomial network. This is created using an option of the MAPS algorithm. With the feedforward network option, each step of the algorithm augments the X matrix with the current model output for consideration at subsequent steps. The set of candidate new terms, when submodel outputs are fed forward, is determined as before. An exception is that terms which are linear in a previous model output are not permitted, since such a term would introduce a linear dependence with other terms still on the term

TABLE 5
Coefficients and exponents of the selected polynomial terms for the Portuguese olive oil data

Term	Coefficient	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
1	−0.5731	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0.2923	0	1	0	0	0	0	0	0	0	0	0	0	0
3	0.2829	0	2	0	0	0	0	0	0	0	0	0	0	0
4	0.3247	0	2	0	0	0	0	0	0	0	0	0	0	0
5	−0.0088	0	0	0	0	0	1	0	0	0	0	0	0	0
6	0.0588	0	0	0	0	0	2	0	0	0	0	0	0	0
7	−0.2483	0	1	0	0	0	0	0	0	0	0	0	1	0
8	−0.0794	0	2	0	0	0	0	0	0	0	0	0	2	0
9	−0.0417	0	2	0	0	0	0	0	0	0	0	0	3	0
10	−0.0417	0	2	0	0	0	0	0	0	0	0	0	0	1
11	−0.7796	0	3	0	0	0	0	0	0	0	0	0	0	0
12	0.0527	0	4	0	0	0	0	0	0	0	0	0	0	0
13	0.1896	0	5	0	0	0	0	0	0	0	0	0	0	0
14	−0.0520	0	2	0	0	0	0	0	0	1	0	0	0	0
15	−0.0264	0	0	0	0	0	2	0	0	0	0	0	1	0
16	0.0440	0	4	0	0	0	0	0	0	0	0	0	0	1

GCV	BIC, MDL
0.159	0.188

list. Consequently, we require previous model outputs to be initialized in the list as a nonlinear product with itself or with other variables.

Table 6 depicts the results using the Portuguese olive oil data with the network option and the BIC/MDL criterion. Here the outputs from model 3 and model 6 were selected by the criterion to be input to subsequent models. The bases functions for models 3 and 6, respectively, are the same as the first

TABLE 6
Coefficients and exponents of the selected polynomial terms with the feedforward network option, for the Portuguese olive oil data

Term	Coefficient	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	m_3	m_4	m_6	m_8	m_{10}
1	−0.71278	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	−0.25350	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0.07698	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0.18919	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
5	−0.00770	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	0.02824	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
7	1.89370	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
8	0.49567	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
9	−1.26261	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
10	0.32184	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0
11	−0.05317	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Effective #Parameters	GCV	BIC, MDL
14	0.14	0.174

3 and first 6 terms in Table 6. However, the coefficients for these submodels are somewhat different than for the final linear combination in model 11.

The effective number of parameters of the network for use in the performance criteria is defined as the number of coefficients directly computed in the present model plus α times the number of parameters in submodels that feed into the present model. For the results in Table 6, we are using $\alpha = 0.3$. When a step of the algorithm selects a pair of terms, only the output with both included is passed as an input to subsequent models; this explains why there is no m_2, m_5, m_7, or m_9 in Table 6.

The results with the network option for the olive oil data show a slight improvement in the values of the criteria. It is interesting that in Table 6, the algorithm proved eager to include successive powers of an intermediate output, rather than successive powers of a single variable as in Table 5. Cross-validation values have not yet been computed for the models in Tables 5 and 6. We suspect that the CV will be slightly better but not substantially better than reported in Table 4.

The network option also was tried on the data from Sections 4.2 and 4.3. In both of these cases, no factors were selected from the list of previous models. So in these cases, the result of the network synthesis algorithm reverts to the results obtained with the conventional adaptive selection algorithm. This is not surprising since the synthetic examples in Sections 4.2 and 4.3 are defined directly as a sum of terms rather than indirectly through a composition. Experience has shown, as reported in Barron, Mucciardi, Cook, Craig and Barron (1984) and Barron and Barron (1988) that network methods frequently provide useful answers for large dimensional data from real engineering and scientific problems for which conventional linear techniques have not been as successful.

7. Conclusions. Adaptive synthesis of nonlinear models is essential in those empirical modeling contexts where scientific or engineering considerations do not provide a complete parametric solution, and where the high dimensionality of the space of candidate inputs prohibits the use of other nonparametric smoothing techniques.

The techniques that gain widest acceptance among empirical modelers are those that are statistically accurate, computationally reasonable, flexible to use in diverse contexts and (sometimes most importantly) well understood by the scientist or engineer and his clients. The paper by Jerry Friedman goes a long way toward making a powerful technique clearly understood. Also, the thoroughness of his methodological and experimental studies of a statistical modeling technique provide an excellent prototype for what, hopefully, will be many more such papers in the field.

The preliminary comparison provided here of adaptive regression splines with adaptive polynomial smoothing on several data sets suggests that the spline method does not provide any substantial gain in accuracy over the polynomial method. This should provide some pause for the empirical modeler

who is debating whether to switch from the more customary polynomial models to the sometimes less familiar splines.

In addition to the known approximation capabilities, polynomial models have in their favor the relative ease of interpretation in many scientific and engineering contexts. Against polynomial models has been the fact that least squares polynomials are prone to wild extrapolative behavior in the high-order case. Here we have pointed out that the simple device of a roughness penalty, familiar to spline smoothers, can be used for polynomial smoothing to mitigate this wild behavior. We recommend properly smoothed and adaptively synthesized polynomial modeling as a serious competitor to adaptively synthesized splines.

Notes added in proof. Some additional computer runs were completed after first submission of this discussion. In particular, the experiments using Friedman's Example 4.2 and a sample size of $N = 200$ were repeated using 100 replications (instead of just 10) to obtain a more reliable assessment of performance of the MAPS algorithm. This time no upper limit was imposed on the number of terms to be selected by the criteria. Again, substantial overfit problems occurred with the use of the unmodified GCV criterion. It selects an average of 15.2, 23.2 and 32.3 terms in the cases of $mi = 1$ (no interaction), $mi = 2$ (second order interaction) and $mi = 10$ (arbitrary interactions permitted), respectively. The average integrated-squared error is excellent in the $mi = 1$ case ($\overline{\text{ISE}} = 0.025$), as predicted by the theory of Li, but breaks down in the $mi = 2$ case ($\overline{\text{ISE}} = 0.058$) and $mi = 10$ case ($\overline{\text{ISE}} = 0.111$) due to the excessive number of spurious terms that the GCV criterion accepts in these cases. In contrast, the BIC/MDL criterion consistently selects a moderate number of terms (an average of 11.7, 11.7 and 11.8 terms with a standard deviation of 1.6, 1.6 and 1.7 terms in the $mi = 1$, $mi = 2$ and $mi = 3$ cases, respectively). Moreover, no breakdown in performance occurs as we increase the number of candidate terms. The average and standard deviation of the integrated-squared error is 0.031 (0.012) for $mi = 1$, 0.032 (0.012) for $mi = 2$ and 0.033 (0.014) for $mi = 10$. To compare with Friedman's spline method, the corresponding values that he reports in Table 3 are 0.026 (0.011), 0.033 (0.021) and 0.037 (0.017), respectively, which show a somewhat greater divergence of ISE values for $mi = 1$, 2 and 10. In this example, the polynomial and spline methods achieve about the same overall performance, provided a suitable model selection criterion is used (BIC/MDL or modified GCV). These results confirm the preliminary conclusions reached using the smaller number of replications.

We also had opportunity to try a tenfold cross-validation experiment on the Portuguese Olive Oil data using models selected by the BIC/MDL criterion instead of by the GCV criterion as in Table 4. We use the CV statistic to give an estimate of the mean-squared error of prediction. In this case, the CV statistic improved to 0.17 instead of 0.22 and the GCV of the selected model is 0.16 instead of 0.15. This illustrates the fact that while the selection of models and the estimation of mean-squared error of prediction are essentially separate

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