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

$$(1) \qquad \hat{f}(y|\mathbf{x}) = \exp\biggl(\sum_k \sum_j \hat{\beta}_{jk} H_{jk}(\mathbf{x}) B_k(y) - c\bigl(\mathbf{h}(\mathbf{x};\hat{\boldsymbol{\beta}})\bigr)\biggr), \qquad y \in \mathscr{Y},$$

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

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

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

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

REFERENCES

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

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

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

Department of Statistics
University of California
Berkeley, California 94720

Chong Gu[1] and Grace Wahba[2]

Purdue University and University of Wisconsin-Madison

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

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

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

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

partitioning ideas is clearly an idea whose time has come and this paper is sure to generate much interest!

We have a number of comments that fall into several areas.

1. Nonparametric function estimation is a rich field. Of course, the general idea of nonparametric function estimation is a rich and growing one. The major methods in one variable are kernel methods, smoothing spline methods, regression spline methods, orthogonal series methods and nearest neighbor methods. When the data points are uniformly spaced, if these methods have their various tuning parameters matched up, they are quite similar for medium-sized data sets and can essentially be tuned so that they have the same convergence rates for functions with the same number of (square integrable, say) derivatives. Even when the data points are not uniform, kernel methods and spline methods can be shown to be similar under certain circumstances. As soon as we get into more than one dimension, however, choices proliferate and results are not necessarily so similar. Narrowing our consideration to kernel estimates, smoothing splines and regression splines, we may, at the outset, consider what might be called a tensor product structure versus a thin plate structure—we will define these by example. Considering two variables $\mathbf{x} = (x_1, x_2)$ and given a knot $\mathbf{t} = (t_1, t_2)$, a basis function of tensor product type is of the form $B_{t_1,t_2}(x_1, x_2) = H(x_1 - t_1)H(x_2 - t_2)$, where H stands for a generic function, usually depending on some order parameter q, whereas a basis function of thin plate type is of the form $B_{t_1,t_2}(x_1, x_2) = H(\|\mathbf{x} - \mathbf{t}\|)$, where $\|\mathbf{x} - \mathbf{t}\|$ is the Euclidean distance between $\mathbf{x} = (x_1, x_2)$ and $\mathbf{t} = (t_1, t_2)$. Thin plate splines do not know the difference between north and east, whereas tensor product splines do. Of course, kernels as well as regression splines come in both types. Friedman's splines are regression splines of tensor product type with a sophisticated procedure for choosing the number and order(s) of the spline basis functions and the knots. It is, of course, equally possible to do regression splines on thin plate basis functions. [See Poggio and Girosi (1990) who discuss regression thin plate splines with moveable knots in the context of neural nets for multidimensional function estimation.] Which type one might prefer would certainly be related to the nature of the variables one is dealing with. In principle, it is quite possible to mix the various types of basis functions. However, Friedman's recursive partitioning approach to knot selection fits naturally into the tensor product setup and probably not in the thin plate setup (since there are not natural cutting planes in that case), while Poggio and Girosi's approach appears to fit in to the thin plate case and not the tensor product setup. In both these cases, knot selection is a nontrivial operation, and clearly as time goes on further insight will be gained.

2. Smoothing splines with multiple smoothing parameters. We agree with Friedman that automatic selection of multiple smoothing parameters is inherently difficult and computationally consuming. Nevertheless, our experience is that it is feasible for relatively small number of covariates and medium sample sizes on modern workstations. Recall that given responses y_i

and covariates $\mathbf{x}_i$, where $y_i \sim p(y; \eta(\mathbf{x}_i))$, a smoothing spline regression fit with multiple smoothing parameters is the solution to the problem: Find $\eta \in \mathscr{H}$ to minimize

$$(2.1) \qquad -\sum_{1}^{n} l_i(\eta(\mathbf{x}_i)) + \frac{n}{2}\lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} J_\beta(f_\beta),$$

where $l_i(\eta) = \log p(y_i; \eta)$, $\eta = \sum_{\beta=0}^{p} f_\beta$ and $\mathscr{H} = \mathscr{H}_0 \oplus \mathscr{H}_1 \oplus \cdots \oplus \mathscr{H}_p$ is a reproducing kernel Hilbert space with reproducing kernel $R = R_0 + R_1 + \cdots + R_p$; see Wahba (1990). Here $\mathbf{x}_i$ may be quite general consisting of several components, in arbitrary index sets. $\mathscr{H}_0$, on which there is no penalty, is of necessity of finite dimension M, say, which is less than n, and as λ tends to infinity, the estimate tends to the maximum likelihood estimate in $\mathscr{H}_0$. f_β is the component (projection) of η in $\mathscr{H}_\beta$ and J_β is a suitable quadratic penalty, which we will take as the squared norm in $\mathscr{H}_\beta$. The elements in each $\mathscr{H}_\beta$ may actually depend on all or only a few of the components of $\mathbf{x}$. It can be shown that the solution of (2.1) has an expression

$$(2.2) \quad \eta(\mathbf{x}) = \sum_{\nu=1}^{M} \phi_\nu(\mathbf{x}) d_\nu + \sum_{i=1}^{n} \left(\sum_{\beta=1}^{p} \theta_\beta R_\beta(\mathbf{x}_i, \mathbf{x}) \right) c_i = \boldsymbol{\phi}^T(\mathbf{x})\mathbf{d} + \boldsymbol{\xi}^T(\mathbf{x})\mathbf{c},$$

where $\{\phi_1, \dots, \phi_M\}$ span $\mathscr{H}_0$, $\boldsymbol{\xi}^T(\mathbf{x}) = (\xi_1(\mathbf{x}), \dots, \xi_n(\mathbf{x}))$, $\xi_i(\mathbf{x}) = \sum_{\beta=1}^{p} \theta_\beta R_\beta(\mathbf{x}_i, \mathbf{x})$ and $\mathbf{c}$ and $\mathbf{d}$ are the minimizers of

$$(2.3) \qquad -\sum_{i=1}^{n} l_i\big(\boldsymbol{\phi}^T(\mathbf{x}_i)\mathbf{d} + \boldsymbol{\xi}^T(\mathbf{x}_i)\mathbf{c}\big) + \frac{n}{2}\lambda \sum_{\beta=1}^{p} \theta_\beta \mathbf{c}^T Q_\beta \mathbf{c},$$

where Q_β is an $n \times n$ matrix with (i, j)th entry $R_\beta(\mathbf{x}_i, \mathbf{x}_j)$. ξ_i's in the smoothing spline examples have knots at the data points $\mathbf{x}_i$ and under certain circumstances, they span the same space as commonly used local bases such as the B-splines on the real line. Let S be the matrix with (j, ν)th entry $\phi_\nu(\mathbf{x}_i)$, $Q = \sum_{\beta=1}^{p} \theta_\beta Q_\beta$, $W = \operatorname{diag}(w_1, \dots, w_n)$, where $w_i = -dl^2 l_i/d\eta_i^2$ and $\mathbf{u} = (u_1, \dots, u_n)$, where $u_i = -dl_i/d\eta_i$. The Newton iteration for minimizing (2.3) proceeds by solving

$$(2.4) \qquad \begin{aligned} W^{1/2}(Q + n\lambda W^{-1})\mathbf{c} + W^{1/2} S \mathbf{d} &= W^{1/2}\tilde{\mathbf{y}}, \\ S^T \mathbf{d} &= 0, \end{aligned}$$

where $\tilde{\mathbf{y}} = \boldsymbol{\eta}_0 - W^{-1}\mathbf{u}$, $\boldsymbol{\eta}_0$ is the fit at the previous step and W and $\mathbf{u}$ are evaluated at $\boldsymbol{\eta}_0$. See, for example, Gu (1990a). This setup provides a unified numerical treatment for broad varieties of nonparametric regression problems with either Gaussian or non-Gaussian responses and with all kinds of covariate structures; see the next section. Generic algorithms for solving (2.4) with automatic smoothing parameters λ and θ's appear in Gu, Bates, Chen and Wahba (1989) and Gu and Wahba (1991) mentioned by Friedman. Transportable code is available from the netlib under the name RKPACK; see Gu (1989). We feel comfortable with these algorithms for n up to 500 and p up to

6 on the contemporary workstations and with smaller p we can afford larger n.

3. ANOVA decomposition and varieties of subspaces. A nice feature of the MARS product is the ANOVA decomposition which greatly enhances the interpretability of the computed fit. It is known that the same structure can also be obtained via interaction smoothing splines, which are important specializations of (2.1). Recall that for a bivariate covariate $\mathbf{x} = (x_1, x_2)$ on a domain $\mathscr{T}_1 \times \mathscr{T}_2$, given reproducing kernel Hilbert spaces $\mathscr{H}^i = \mathscr{H}_0^i \oplus \mathscr{H}_1^i$ of functions on $\mathscr{T}_i$, $i = 1, 2$, the tensor product Hilbert space of functions on $\mathscr{T}_1 \times \mathscr{T}_2$ has a tensor sum decomposition $\mathscr{H} = \mathscr{H}^1 \otimes \mathscr{H}^2 = (\mathscr{H}_0^1 \otimes \mathscr{H}_0^2) \oplus (\mathscr{H}_1^1 \otimes \mathscr{H}_0^2) \oplus (\mathscr{H}_0^1 \otimes \mathscr{H}_1^2) \oplus (\mathscr{H}_1^1 \otimes \mathscr{H}_1^2)$. Assuming finite dimensional $\mathscr{H}_0^i$, $i = 1, 2$, an interaction spline is obtained by specializing (2.1) with $\mathscr{H}_0 = \mathscr{H}_0^1 \otimes \mathscr{H}_0^2$, $\mathscr{H}_1 = \mathscr{H}_1^1 \otimes \mathscr{H}_0^2$, $\mathscr{H}_2 = \mathscr{H}_0^1 \otimes \mathscr{H}_1^2$ and $\mathscr{H}_3 = \mathscr{H}_1^1 \otimes \mathscr{H}_1^2$. When $\mathscr{H}_0^i = \{1\}$, $i = 1, 2$, the setup provides an ANOVA decomposition of the estimate by construction. When we take the tensor sum or tensor product of any two reproducing kernel spaces, we just add or multiply their reproducing kernels to get the reproducing kernel of the resulting space. Obviously, the tensor sum on each coordinate can take more than two operands and so can the tensor product.

Examples of interaction splines in subspaces of tensor products of $W_2^m[0, 1]$'s appear to be the most popular in the existing literature. However, the foregoing general framework covers a much broader spectrum of model specifications. Consider a covariate x on a categorical domain $\mathscr{T} = \{1, \dots, C\}$. A real function on $\{1, \dots, C\}$ is just a real vector in the Euclidean space R^C. Adopting a roughness penalty proportional to the Euclidean norm of the projection of a vector onto $\{\mathbf{1}\}^\perp$, a smooth vector would then be one with a small variance. The corresponding Hilbert space decomposition is $R^C = \{\mathbf{1}\} \oplus \{\mathbf{1}\}^\perp$ with a reproducing kernel $R(x, x')$, $x, x' \in \{1, \dots, C\}$, representable as a $C \times C$ real matrix, $\mathbf{11}^T/C + [I - \mathbf{11}^T/C]$, where the term in brackets is the reproducing kernel for $\{\mathbf{1}\}^\perp$. Using R^C for categorical (nominal) covariates in the construction of tensor product Hilbert space, one can incorporate both continuous and categorical covariates simultaneously to build a model with a natural ANOVA decomposition. We understand from Friedman's talk at Interface '90 that categorical covariates can also be incorporated into the MARS framework. It would be interesting to compare the two approaches.

We have discussed the tensor product structure versus the thin plate structure in Section 1. In general a tensor product structure is appropriate for combining individually interpretable covariates and a thin plate structure is appropriate for dealing with rotation invariant problems. The example later indicates that certain problems require mixed structures. From the Eastern Lake Survey of 1984 implemented by the Environmental Protection Agency of the United States, a data set has been derived by Douglas and Delampady (1990) which contains geographic information, water acidity measurements and main ion concentrations of 1798 lakes in four regions in the eastern United States. An attempt is made to explore the dependence of the water acidity on the geographic locations and other information concerning the lakes. Preliminary analysis and consultation with a water chemist suggest that a model for the surface pH in terms of the geographic location and the calcium ion concentration is appropriate. Obviously, a thin plate structure is appropriate for the geographic location. To account for the joint effect of geographic location and the calcium concentration, however, a tensor product structure appears to be appropriate. A tensor product reproducing kernel Hilbert space with a thin plate space component for the geographic locations and a W_2^2 component for the calcium concentrations does the job simply. This example actually illustrates the fact that the general framework of interaction splines can paste up arbitrarily complicated components to provide an interpretable ANOVA decomposition.

To use a thin plate spline as a component of an ANOVA model as we have just described, one needs an explicit reproducing kernel. [Only a so-called semikernel is needed for the construction of a thin plate spline by itself; see the references in Wahba (1990).] An explicit reproducing kernel appears in Wahba and Wendelberger (1980), but it is not a natural one to use in an ANOVA model. We will provide a more natural one here. Let $\mathscr{X}$ be the thin plate function space (in two variables) consisting of linear functions plus all functions (modulo the linear functions) for which the thin plate penalty functional

$$J(f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(f_{uu}^2 + 2f_{uv}^2 + f_{vv}^2\right)du\,dv$$

is well-defined and finite. Now let $\mathbf{t}_j$, $j = 1,\dots,K$, be any set of $K \geq 3$ points in R^2 not falling on a straight line. Let $\phi_1(\mathbf{x}) = 1/\sqrt{K}$ and let ϕ_2 and ϕ_3 be linear functions satisfying $\sum_{j=1}^{K}\phi_\mu(\mathbf{t}_j)\phi_\nu(\mathbf{t}_j) = 1$, $\mu = \nu, 0$, otherwise, and let $w_j(\mathbf{x}) = \sum_{\nu=1}^{3}\phi_\nu(\mathbf{t}_j)\phi_\nu(\mathbf{x})$. If we endow $\mathscr{X}$ with the squared norm

$$\|f\|_{\mathscr{X}}^2 = \sum_{\nu=1}^{3}\left(\sum_{j=1}^{K}\phi_\nu(\mathbf{t}_j)f(\mathbf{t}_j)\right)^2 + J(f),$$

it can then be shown that the reproducing kernel for $\mathscr{X}$ is

$$\begin{aligned}R(\mathbf{x},\mathbf{x}') = {} & \sum_{\nu=1}^{3}\phi_\nu(\mathbf{x})\phi_\nu(\mathbf{x}') \\ & + \Bigg[E(\mathbf{x},\mathbf{x}') - \sum_{j=1}^{K} w_j(\mathbf{x})E(\mathbf{t}_j,\mathbf{x}') - \sum_{k=1}^{K} w_k(\mathbf{x}')E(\mathbf{t}_k,\mathbf{x}) \\ & \qquad + \sum_{j=1}^{K}\sum_{k=1}^{K} w_j(\mathbf{x})w_k(\mathbf{x}')E(\mathbf{t}_j,\mathbf{t}_k)\Bigg],\end{aligned}$$

where $E(\mathbf{x},\mathbf{x}') = (\|\mathbf{x}-\mathbf{x}'\|^2\log\|\mathbf{x}-\mathbf{x}'\|)/(8\pi)$. The term in brackets is the reproducing kernel for $\mathscr{H}_1$ if $\mathscr{H}_0$ is taken as the span of the linear functions. Furthermore, each element f of $\mathscr{H}_1$ satisfies the discrete orthogonality conditions $\sum_{j=1}^{K}\phi_\nu(\mathbf{t}_j)f(\mathbf{t}_j) = 0$, $\nu = 1,2,3$. It is natural to choose $K = n$ and the $\mathbf{t}_j$'s at the data points. With this construction, thin plate splines can be included in a nonparametric ANOVA model, possibly by moving span $\{\phi_2, \phi_3\}$ into $\mathscr{H}_1$. This construction extends to higher order and higher dimensional thin plate splines in a straightforward way.

4. Model selection and concurvity. Model selection in MARS is built into the stepwise estimation procedure by choosing bases and knots to minimize an intuitive GCV score. In a smoothing spline setup (2.1), model selection is by selective inclusion of subspaces and selection of smoothing parameters. Stepwise nonparametric estimation procedures are likely to be confused by concurvity (collinearity), for which Friedman proposes several cures in his algorithms. On the other hand, the direct approach of Section 2 estimates all terms simultaneously and is not bothered by negligible terms and concurvity. For the sake of parsimony and interpretability, however, one wishes to remove aliasing effects and noise terms in a fit. Recently, Gu (1990b) proposed simple geometric diagnostics to tackle the problem. Recall that the solution $\hat{\eta}$ of (2.1) also minimizes

$$\sum_1^n w_i(\tilde{y}_i - \eta(\mathbf{x}_i))^2 + n\lambda \sum_{\beta=1}^p \theta_\beta^{-1} J_\beta(f_\beta), \tag{4.1}$$

where the $\tilde{y}_i$'s and the w_i's are those of (2.4) evaluated at $\hat{\eta}$; see, for example, Gu (1990a). Evaluating the computed fit on the data points, one obtains a retrospective linear model

$$\begin{aligned} W^{1/2}\tilde{y} &= W^{1/2}\big(\tilde{\mathbf{f}}_0 + \tilde{\mathbf{f}}_1 + \cdots + \tilde{\mathbf{f}}_p + \tilde{\mathbf{e}}\big) \\ &= W^{1/2}S\mathbf{d} + W^{1/2}\tilde{F}_1 + W^{1/2}\tilde{\mathbf{e}}, \end{aligned} \tag{4.2}$$

where $\tilde{\mathbf{f}}_\beta = (f_\beta(\mathbf{x}_1), \dots, f_\beta(\mathbf{x}_n))^T$ and $\tilde{F} = (\tilde{\mathbf{f}}_1 \cdots \tilde{\mathbf{f}}_p)$. Removing the null model effect by projecting (4.2) onto the orthogonal space of $W^{1/2}S$, one gets

$$\mathbf{z} = F_1 + \mathbf{e}. \tag{4.3}$$

The collinearity indices of F [Stewart (1987)], which is equivalent to the cosines between the columns of F, measure the concurvity in the fit. The columns of F are supposed to predict the response $\mathbf{z}$ so a near orthogonal angle between a column of F and $\mathbf{z}$ indicates a noise term. Signal terms should be reasonably orthogonal to the residuals, hence a large cosine between a column of F and $\mathbf{e}$ makes a term suspect. $\cos(\mathbf{z}, \mathbf{e})$ and $R^2 = \|\mathbf{z} - \mathbf{e}\|^2/\|\mathbf{z}\|^2$ are informative ad hoc measures for the signal-to-noise ratio in the data. Finally, a *very* small norm of a column of F relative to that of $\mathbf{z}$ also indicates a negligible term. It could be argued that the cosine diagnostics can be treated as absolute measures provided an automatic smoothing parameter selection is adopted. These diagnostics can be used to sequentially delete redundant subspaces to build a parsimonious model in a backward fashion; see Gu (1990b) for details and examples.

5. Accuracy estimates. For any method, it would be nice to be able to say something about the accuracy of the estimate. Monte Carlo bootstrap is one method that has found use in the application of smoothing splines. In the Monte Carlo bootstrap, one generates a new set of data via a random number generator centered about the estimated model. In this setup the distribution of the y_i has to be assumed up to parameters which can be estimated, that is, p is Gaussian, Bernoulli and so on. From this new data set, one estimates a curve or surface and by repeating the operation, obtains a cloud of curves or surfaces which hopefully gives some idea of the accuracy of the estimate. Although the clouds one gets seem reasonable, it still appears that not much is known about their properties, although they have been around a while. In particular, since the estimate is generally a bit smoother than the truth, it is possible that these clouds give too rosy a picture. This method might be used with Friedman's approach here, if (assuming the ε_i are i.i.d. zero mean Gaussian) an estimate of σ^2 were available.

In the single smoothing parameter spline case, there are coverage bands (also called Bayesian "confidence intervals") which have the property that the expected number of true data points they cover is about $0.95n$; see Wahba (1983) and Nychka (1988). It would be interesting to see what happens in the multiple smoothing parameter case and also if there is anything like a counterpart for MARS. Some authors, in the case of (nonadaptive) regression splines, have suggested the usual parametric confidence intervals for the estimated basis coefficients, as though the estimate were really in the span of the basis functions. This has yet to be justified if the true function is not in this span and hence there is bias. There is, however, no free lunch in the case of nonparametric regression, since the good methods are all biased and we have to accept a somewhat weaker definition of confidence interval in the nonparametric regression case if we wish to remain honest.

6. Hybrid methods. Hybrid methods which combine regression spline and smoothing spline ideas have also been considered; see, for example, Nychka, Wahba, Goldfarb and Pugh (1984), Hutchinson and Bischof (1983) and O'Sullivan (1990). Let $\{B_l(\mathbf{x})\}_{l=1}^N$ be a set of N basis functions, where N is generally (much) smaller than n. The estimate of η is then that η is the span of the $\{B_l\}$ which minimizes (2.1). If N is closer to n, then the estimate will be a good approximation to the minimizer in $\mathscr{H}$ of (2.1) and the λ and θ_i's will be controlling the smoothing. In this case, the basis functions are mainly used to ease the computation and the location of the knots may not be very crucial. For example, if n is large, one might choose a set of basis functions with the knots a regular subset of the data points, as did Hutchinson and Bischof (1983). In their case the basis functions were formed from *representers of evaluation* at the knot points. That would amount to using $\phi_1, \dots, \phi_M$ and a subset of the ξ's of (2.2) in the present context; see also Wahba (1990), Chapter 7. O'Sullivan (1990) used a large basis of tensor products of B-splines, with a thin plate penalty functional.

If N is smaller, then the basis functions may be helping along the smoothing (and they might be explicitly used to eliminate the possibility of too much fine structure that is known not to be there), in this case their number and locations of their knots may bê more influential.

7. An omnibus GCV? We think that all of these methods and combinations will come into use and no one of the possibilities is going to turn out to be uniformly superior—which method is best is going to depend on the context. Of course, it would be lovely if there was one grand criterion for comparing among the different methods, given a particular data set. It would be nice to have something like an omnibus GCV criterion, which would compare different model building procedures (for example, a pure smoothing vs. a pure regression procedure), but this is something remaining to be done. As Friedman takes pains to note, all of the degrees of freedom for signal must be accounted for. In order to compare across different methods, this accounting should be done in comparable ways. In the case of smoothing splines, when a subspace is added, if this subspace carries an independent smoothing parameter, then the minimum GCV value is nonincreasing. To see this, note that if the smoothing parameter for the new subspace is estimated as infinity, then we have reverted to the model without the new subspace. It is not yet known how to charge for adding another free smoothing parameter in this context. The preceding suggestions on methods for choosing subspaces as a form of model selection were in part motivated by this lack.

Given a sufficiently large data set, in practice it is of course possible to do the classical double cross-validation, say, divide the set in half, fit and tune two or more models for comparison on the first half of the data and compare them as to their predictive ability on the second half. Naturally, this brings up the question of what is a "significant" difference—which, again, could be answered if one could partition the data into several subsets. Data sets are getting bigger and computers more powerful, so statisticians are not likely to be out of business soon!

REFERENCES

DOUGLAS, A. and DELAMPADY, M. (1990). Eastern Lake Survey—Phase I: Documentation for the data base and the derived data sets. SIMS Technical report, Dept. Statist., Univ. British Columbia, Vancouver.

GU, C. (1989). RKPACK and its applications: Fitting smoothing spline models. Technical report 857, Dept. Statist., Univ. Wisconsin-Madison.

GU, C. (1990a). Adaptive spline smoothing in non-Gaussian regression models. *J. Amer. Statist. Assoc.* **85** 801–807.

GU, C. (1990b). Diagnostics for nonparametric additive models. Technical report 92, Dept. Statist., Univ. British Columbia, Vancouver.

GU, C., BATES, D. M., CHEN, Z. and WAHBA, G. (1989). The computation of GCV functions through householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal. Appl.* **10** 457–480.

GU, C. and WAHBA, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.* **12**.

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