able to determine precisely f. (When $l = 1$, the method is essentially the Aitken acceleration procedure.) Of course, one stops far short of $2l$ iterations to obtain a good approximation to f.

Therefore the extrapolation procedure consists of the following two steps:

1. Compute Gauss–Seidel iterations, $f^1, \ldots, f^{2p}$, forcing f^i to lie in the kernel of $\hat{P}$.
2. Extrapolate the iterates using the ε-algorithm to get an improved approximation to f.

We have used the above algorithm on various problems with relatively good success. A first data set which was given to us by the authors turned out to be incompatible and as a result of our calculations an error in a spline fitting program was detected! A second set of data was provided where Gauss–Seidel converged slowly and the application of the above algorithm yielded very satisfactory results.

Conclusion. Using the ε-algorithm to accelerate convergence of a basic iteration for linear systems with nonsymmetric matrices seems to be a very promising approach.

Acknowledgment. The authors wish to thank Mr. Will Sawyer for performing the indicated computations.

REFERENCES

[1] GEKELER, E. (1972). On the solution of systems of equations by the epsilon algorithm of Wynn. *Math. Comp.* **26** 472–436.

[2] GOLUB, G. H. and DE PILLIS, J. (1988). Toward an effective two-parameter SOR method. Presented at the Conference on Iterative Methods for Large Linear Systems, October 19–21, 1988, Austin, Texas.

[3] GOLUB, G. H. and VAN LOAN, C. F. (1983). *Matrix Computations*. Johns Hopkins Univ. Press, Baltimore, Md.

[4] ISAACSON, E. and KELLER, H. B. (1966). *Analysis of Numerical Methods*. Wiley, New York.

[5] KELLER, H. B. (1965). On the solution of singular and semidefinite linear systems by iteration. *SIAM J. Numer. Anal.* **B2** 281–290.

[6] VARGA, R. S. (1962). *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, N.J.

INSTITUT FÜR WISSENSCHAFTLICHES RECHNEN
EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE
CH-8092 ZÜRICH
SWITZERLAND

COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

THEO GASSER AND ALOIS KNEIP

Zentralinstitut für Seelische Gesundheit and Universität Heidelberg

We would like to congratulate the authors for a stimulating paper. Additive models for approximating high-dimensional regression problems have been around for quite some time, but a number of important problems have remained

unsolved. The present paper fills a large gap by studying in detail and with mathematical rigor numerical and computational problems within a statistical framework.

The first part of the paper is devoted to a discussion of various smoothing methods and some of their properties. A rather bewildering "fauna" of smoothing techniques to be found in the literature is presented one after the other. It then is shown that the important methods can be formalized as penalized least-squares methods, which give access to a powerful machinery both in statistical and computational terms. It should be noted that an equivalent approach is to consider smoothing techniques within the framework of the so-called "method of sieves" as presented by Geman and Hwang (1982). As it stands, the penalized least-squares approach does not, however, answer the question whether some smoothing method should be preferred uniformly or for some situation. There is, therefore, no strict rationale for choosing one of them.

Except for regression splines with variable knots, a common feature of the more sophisticated methods is that all are some linear weighted scheme of the data with the amount of smoothing tuned by one parameter. As a matter of fact, methods like kernel estimators, smoothing splines, k-nearest-neighbor estimators and locally weighted running lines are quite similar in an asymptotic sense. It can be proved that asymptotically all can be put into the framework of kernel estimation theory [for smoothing splines this was shown by Silverman (1984) and for locally weighted running lines by Müller (1987)]. Trying to understand how these methods are related theoretically, we have studied kernel estimators with bandwidth $\lambda(t)$ varying with the density d of the design as follows [see Jennen-Steinmetz and Gasser (1988)]:

$$\lambda(t) = \lambda_0 d(t)^{-\alpha}, \qquad 0 \le \alpha \le 1.$$

A value $\alpha = 0$ corresponds to ordinary kernel estimation, $\alpha = 1/4$ to smoothing splines and $\alpha = 1$ to k-nearest-neighbor estimation. Choosing the optimal method thus means choosing the optimal value of α. Interestingly, there is no uniformly optimal α, whereas $\alpha = 0$ is minimax optimal. In the examples inspected, $\alpha = 1/4$ (i.e., smoothing splines) was never much more inferior compared to $\alpha = 0$, but $\alpha = 1$ often was.

There are some further comments we would like to make with respect to the smoothing part of the paper.

(i) Kernel smoothers do not need $O(n^2)$ operators, but can be computed in $O(n)$ operations for polynomial kernels W [which include optimal and minimum variance kernels, see Gasser, Müller and Mammitzsch (1985)]. Following the definition of Gasser and Müller (1984), the smoothed value $\hat{y}_i$ at design point x_i is given by

$$\hat{y}_i = \sum_{j=1}^{n} w_j(x_i; \mathbf{x}; \lambda) y_j,$$

where λ is the bandwidth and w_j is a weight derived from the kernel W. If W is polynomial of some order p, for some $a_1, \ldots, a_{p+1}$ (given by W) w_j is obtained

by

$$w_j = \sum_{k=1}^{p+1} a_k \left\{ \left(\frac{s_j - x_i}{\lambda} \right)^k - \left(\frac{s_{j-1} - x_i}{\lambda} \right)^k \right\},$$

where $s_j = (x_{j-1} + x_j)/2$. Straightforward calculations yield

$$\hat{y}_i = \sum_{j=1}^{n} \sum_{k=1}^{p+1} \sum_{l=0}^{k} \frac{a_k}{\lambda^k} \binom{k}{l} x_i^{k-l} \left\{ s_j^l - s_{j-1}^l \right\} y_j.$$

By a rearrangement of sums, one notes that an $O(n)$ algorithm is possible since the sum over j needs only to be computed once,

$$\hat{y}_i = \sum_{k=1}^{p+1} \sum_{l=0}^{k} \frac{a_k}{\lambda^k} \binom{k}{l} x_i^{k-l} \sum_{j=1}^{n} \left(s_j^l - s_{j-1}^l \right) y_j.$$

The basic idea is simple, but unfortunately a naive implementation is numerically unstable (note the awkward difference). A less simple but stable algorithm has been found in the meantime.

(ii) Bias at the boundaries for kernel estimators can (and should) be reduced by using appropriate boundary kernels [Gasser, Müller and Mammitzsch et al. (1985)]. These allow the same rates of convergence as for interior points.

(iii) Estimating σ^2 by the residual sum of squares renders a heavily biased estimate as the authors note (Section 2.7.2). There is, however, a simple nonparametric estimator with a negligible bias [Gasser, Sroka and Jennen-Steinmetz (1986)], which performed well in subsequent work.

(iv) The present paper avoids the problem of the choice of the smoothing parameter, which is acceptable. It should nevertheless be stressed that Section 2.7 only leads to a comparison of bandwidths for different estimators and not to an objective choice of the optimal smoothing parameter (see also Section 2.2).

For anyone who plans to apply additive modeling to his or her data, the second part of the paper is a major step forward. It becomes clear how practical computations can be done and what the main properties of the algorithm are. It seems to us that studying backfitting algorithms could be of even more general interest: In our work on self-modeling nonlinear regression we found it essential to break down a large (nonlinear) least-squares problem into smaller ones by an iterative procedure similar to a backfitting algorithm [Kneip and Gasser (1988)]. This type of procedure might be of general interest in sophisticated model building.

After seeing quite a lot of theory and algorithmic work, interesting applications could now provide feedback for further statistical work. Nevertheless, we find two areas to be most interesting right now.

1. It would be important to understand which classes of multivariate problems can be well analyzed by additive modeling, and also when it fails by leading to wrong conclusions. The analogy to ANOVA tells us that often interaction effects are more interesting than main effects.

2. Statistical problems like appropriate goodness-of-fit tests, confidence bounds and selection of important or elimination of unimportant covariates should be dealt with.

REFERENCES

GASSER, TH. and MÜLLER, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171–185.

GASSER, TH., MÜLLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

GASSER, TH., SROKA, L. and JENNEN-STEINMETZ, CH. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.

GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

JENNEN-STEINMETZ, CH. and GASSER, TH. (1988). A unifying approach to nonparametric regression estimation. *J. Amer. Statist. Assoc.* **83** 1084–1089.

KNEIP, A. and GASSER, TH. (1988). Convergence and consistency results for self modeling nonlinear regression. *Ann. Statist.* **16** 82–112.

MÜLLER, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231–239.

SILVERMAN, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.* **12** 898–916.

ZENTRALINSTITUT FÜR SEELISCHE GESUNDHEIT
ABTEILUNG BIOSTATISTIK
POSTFACH 5970
D-6800 MANNHEIM 1
WEST GERMANY

INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
HEIDELBERG
WEST GERMANY

ROBERT KOHN AND CRAIG F. ANSLEY

University of New South Wales and University of Auckland

We would like to congratulate the authors for presenting us with such a broad overview of this important topic, and in particular on their proof of the convergence of the backfitting method. The linear smoother to which they have paid the most attention is the cubic spline smoother. Now smoothing splines can be represented as signal extraction estimates in a model where the unknown regression function is generated by a stochastic process. This allows us to take a model-based approach to smoothing and estimating the components of an additive model using smoothing splines, and in this comment we wish to contrast this approach with that of the authors. A model-based approach for estimating the additive components has much to commend it because: (i) All assumptions are stated explicitly. (ii) It is a comprehensive approach which is able to deal with a variety of problems including polynomial smoothing splines. (iii) Unlike ad hoc approaches such as running means and medians, the model-based approach can deal with unequally spaced data. (iv) It suggests reasonable ways of estimating unknown parameters either by maximum likelihood or Bayesian methods. (v) It provides a framework for doing statistical inference, that is, for setting confidence intervals for the unobserved components and the unknown