STONE, C. J. and KOO, C.-Y. (1985). Additive splines in statistics. *Proc. Statist. Comp. Sec.* 45–48. Amer. Statist. Assoc., Washington.
TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559–568.
TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass.
UTRERAS, F. D. (1979). Cross-validation techniques for smoothing spline functions in one or two dimensions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 196–232. Springer, Berlin.
VAN DER BURG, E. and DE LEEUW, J. (1983). Non-linear canonical correlation. *British J. Math. Statist. Psych.* **36** 54–80.
WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.
WAHBA, G. (1986). Partial and interaction splines for the semiparametric estimation of functions of several variables. Technical Report 784, Dept. Statistics, Univ. Wisconsin, Madison.
WATSON, G. S. (1964). Smooth regression analysis. *Sankyā Ser. A* **26** 359–372.
WHITTAKER, E. (1923). On a new method of graduation. *Proc. Edinburgh Math. Soc.* **41** 63–75.

ANDREAS BUJA
BELL COMMUNICATIONS RESEARCH
MORRISTOWN, NEW JERSEY 07960-1910

TREVOR HASTIE
AT&T BELL LABORATORIES
600 MOUNTAIN AVENUE
MURRAY HILL, NEW JERSEY 07974-2070

ROBERT TIBSHIRANI
DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO
CANADA M5S 1A8

DISCUSSION

LEO BREIMAN

University of California, Berkeley

After finishing the ACE paper [Breiman and Friedman (1985)] I hoped that others would tie up some of the significant loose ends. The work under discussion does a good part of that admirably.

But is it interesting that since that time both Friedman and myself have veered off in the direction of using splines for additive and more general models, thus circumventing the problem of convergence of iterated smooths which occupies much of the present paper.

I think it would be useful, in the context of the present paper, to give the itinerary of my journey from smoothers to splines. In addition, another problem that has occupied me is the incorporation of bivariate interaction into the model and I will also comment on that below.

Bivariate smoothers, in and of themselves are not of undying statistical interest. The interest in them developed because of realization, in the ACE paper, that additive models could be fitted through an iterated sequence of bivariate smooths. Now additive models are very interesting, since they form a useful and often revealing extension to linear models.

Thus, most of my thinking about where to go with smoothers has taken place within the context of additive models.

With additive models and numerous predictor variables it is imperative to keep control over the number of degrees of freedom being used in the fitting. Suppose there are 10 predictor variables and each one is fit using a smoother with five degrees of freedom. Assuming additivity in the number of degrees of freedom, then 50 degrees of freedom are being used in the fit.

This will usually result in "overfitting—too many parameters being estimated, and an inflated variance. Control over this was exercised in the original ACE algorithm through the use of supersmoother. This is an elaboration on running linear smoothers deviced by Friedman and Stuetzle (1982) which uses a locally adaptive window size. The idea is to use a large window when the underlying function is slowly changing and a small window when it is rapidly changing. The local window size is data determined.

Supersmoother is intrinsically nonlinear and unsymmetric. While ACE functioned extremely well on most problems, it because clear early on that these supersmoother properties were the source of some undesirable artifacts. For instance, the transformations for weak variables were somewhat dependent on the order of entry of the variables. These artifacts were especially apparent for small sample sizes and we advised users to use a fixed window size when the number of cases was appreciably below 100.

This started me thinking about what would be done to improve the performance. It was clear that convergence could, at best, be established for symmetric smoothers. It was not at all clear how to symmetrize running linear smoothers. The only natural symmetric smoother around was smoothing splines.

Up to this point, my thinking seems to have paralleled that of the authors. At a point several years ago, I would have been in agreement with their statement, "We find fixed knot cubic splines less appealing than their immediate competitors, smoothing splines."

But while it was clear that smoothing splines would do the trick, the issue of control over the degrees of freedom seemed unsurmountable. One cannot assume the same number of degrees of freedom for each variable. The appropriate number of degrees of freedom may vary considerably from variable to variable. Some transformations may be almost linear and some may be quite complex.

Smoothing splines use one parameter per variable to govern the degree of smoothness, and determine the value of this parameter by a search using cross-validation. To conduct such a search over, say, a 10-dimensional space, seemed computationally unfeasible.

In fact, I would challenge to the authors to construct an additive model using smoothing splines for, say 10 predictor variables and 300 cases computing a nearly "optimal" degree of smoothing for each variable *individually* that can run in a reasonably length of time on anything short of a Cray.

There are other problems about using iterated smoothers to fit an additive model. As the authors point out, it is difficult to compute standard inferential statistics, that is, confidence intervals, or standard regression diagnostics, such as influence.

On the other hand, it is very simple to set up a spline basis for each variable and then to treat the whole problem of fitting an additive model to an untransformed y-response as a classical regression. This too has problems associated with it, but the problems are more surmountable.

If one is doing a bivariate fit, and using splines with knots as the x-points $t_1, t_2, \ldots, t_J$, then a basis for the cubic splines is given by

$$1, x, x^2, x^3, \left[(x - t_1)^+\right]^3, \ldots, \left[(x - t_J)^+\right]^3.$$

Usually, the knots are put at equally spaced order statistics and in my experimentation, this has generally worked better than anything else. Here are some problems in the bivariate case:

1. The self-influence near the endpoints is large. The fitted curves have high variance near the endpoints and tend to wagglè around there.
2. The fitted function is dependent on the knot location. This was succinctly pointed out by Hastie and Tibshirani (1988).

There are good fixes to these problems. To decrease the self-influence near the endpoints, impose the condition that the fitted spline function be linear at the lowest and highest x-value. For equally spaced x-values this approximately halves the self-influence at the endpoints. A salutary side effect is that this end condition knocks the functions x^2, x^3 out of the spline basis, leaving only the linear functions and the functions corresponding to knots.

There are two aspects to problem 2. The first is that the influence of x-values will depend on their location relative to the knots. This can be seen by looking at a graph of the self-influence curve. For an ordinary spline fit, this curve ripples up and down depending on where the point is in relation to the adjacent knots.

One way to decrease this dependence of influence on the knot locations is to *interlace the knots*. Here is a simple example of interlacing. Suppose one is doing a spline fit with knots at the order statistics $x(0.25)$, $x(0.50)$, $x(0.75)$. Do the regression on the spline basis using these knots getting a predictor function $y_1(x)$. Do another regression using splines at the interlaced points $x(0.125)$, $x(0.375)$, $x(0.625)$, $x(0.875)$, and denote the resulting predictor function by $y_2(x)$. Let the final prediction function $y(x)$ be the average of $y_1(x)$ and $y_2(x)$.

If the corresponding self-influence is now graphed, it will appear almost constant except near the endpoints. It is interesting that the smoother corresponding to $y(x)$ is linear and symmetric, but is not a projection matrix.

However, the most important element in reducing the dependence on knot placement is a procedure suggested by Smith (1982). Her idea was this—put down many knots along the x-axis, say, as closely as possible to equispaced order statistics. Do the regression into this space. Now delete knots from the fit, at each stage deleting that knot whose deletion causes the least rise in residual sum of squares. Continue the deletion under the "best" fit is found.

There are many advantages to this procedure. First, in the basis given above, each knot corresponds to a coefficient. Thus, deleting knots is nothing else than the classical regression procedure of deleting variables. Second, deleting a knot at any knot point has the effect of broadening the window size in the neighborhood

of that point. Thus, knot deletion has the effect of using a locally adaptive window size—the feature that made supersmoother so attractive.

The result, when finished with the deletion, is that the remaining knots will be located in the vicinity of rapid change of the function, with no knots in those intervals where it can be adequately fit by a cubic.

Peters and I recently completed a fairly extensive simulation [Breiman and Peters (1988)] that compared four automatic smoothers. Here the word automatic indicates that the smoother had a mechanism to select its own window size in a data-dependent way. We compared smoothing splines, supersmoother, a kernel-type smoother, and our version of the above procedure which we called the DKS smoother (delete knot splines).

We used sample sizes 25, 75 and 225 with a variety of functions of X-designs. In DKS, we used 9 knots for sample size 25, 12 for 75 and 16 at 225. The decision on the number of knots to retain by was based on minimum C_p. The version of DKS used in the simulation has since been refined and made more accurate. Even so, the simulation showed that the early version was quite competitive at almost every X-design, sample size and function. In terms of computing time, it is second only to supersmoother.

As a result of this approach, we have a simple method for fitting a multivariate additive model that also permits good control over the number of degrees of freedom used in the fit. Here is the procedure—put down many knots on each predictor variable; fit the full model; do knot deletion where we delete at any stage that knot on any variable giving the least rise to RSS; and use some method to decide how many knots to retain in the model. For instance, some variant of cross-validation could be used in making this latter decision.

I am currently working on developing and testing this method, using cross-validation to decide how many knots to retain. Early results are quite promising. For those interested in computing details, I note that to prevent ill-conditioning, we convert to a B-spline basis for the matrix inversion and then convert back to the power basis for the deletion. The algorithm is computationally rapid—in any instances even faster than the original ACE code.

By using this approach, we have gotten around all of the issues of convergence of iterated sequences of smoothers. At the same time, we keep tight control on the number of degrees used in the model and tailor the number of degrees of freedom to individual variables.

Note that regression diagnostics can be easily computed. Confidence intervals can also be computed. Once the final model is arrived at, forget that it has been gotten by data-directed variable selection, and compute the intervals in the classical manner based on the final model. Of course this is cheating, but the issue is whether these confidence intervals are reasonable approximations. That is under investigation.

Another issue that I (and the authors) consider important in the context of nonparametric modeling is that of modeling bivariate interaction. Suppose, for example, that we have two predictor variables and we want to fit a model of the form $f(x_1, x_2)$. For example, we may have already fit an additive model and want to explore any signs of bivariate interactions in the residuals.

Early on, it was clear that one way to do this was to construct a surface gotten from a smooth of y on x_1, x_2. But is was also clear that this path had some serious difficulties. The authors refer to some of these in Section 3.1. I add one more—if the two-dimensional smoother has p degrees of freedom in each dimension, then, approximately, it has overall about p^2 degrees of freedom. Thus, it results in a quite nonparsimonious fit with little possibility of adjusting the fit to each variable.

I have stewed over this problem for a while. Recently, I developed the following promising approach: To explain, it is easier to go into random variable space. Consider the following problem: given random variables Y, X_1, X_2, find functions f and g that minimize

$$E(Y - f(X_1)g(X_2))^2.$$

Of course, f and g are nonunique up to at least a multiplicative constant. We can fix this up by requiring $E(f(X_1))^2 = 1$.

Here is a tentative algorithm for finding the minimizing f, g. Hold f fixed. Then the minimizing g is given by

$$g(X_2) = E(Yf(X_1)|X_2)/E(f^2(X_1)|X_2).$$

Now hold g fixed. Then the minimizing f is given by

$$f(X_1) = E(Yg(X_2)|X_1)/E(g^2(X_2)|X_1).$$

Iterate this process—hold f fixed and find the minimizing g. Then hold that g fixed and find the minimizing f. Continue until convergence.

If X_1 and X_2 are independent, then this is a classical problem in approximation theory, the optimal f, g are solutions of linear eigenvalue equations, there is one global minimum and no local minima; and the iteration can be easily be shown to converge to the global minimum.

Ignoring, for the nonce, what happens in the case of nonindependence, note that again, as in the ACE algorithm, the iteration only depends on bivariate conditional expectations. This gets us out of the nasty problem that any surface smoother depends on an arbitrary two-dimensional metric.

The algorithm can be easily extended to the problem: Find $\{f_j, g_j\}$, $j = 1, \ldots, J$, to minimize

$$E(Y - f_1(X_1)g_1(X_2) - \cdots - f_J(X_1)g_J(X_2))^2.$$

Again, the trick is to keep the $\{f_j\}$ fixed and minimize on the $\{g_j\}$, then keep the $\{g_j\}$ fixed and minimize on the $\{f_j\}$, and so forth.

The algorithm can also be done stepwise. That is, find the minimizing f, g. Now apply the same procedure to $Y - f(X_1)g(X_2)$. In the independence case, stepwise gives the same results as the simultaneous minimization. But in either case it can be shown that the sequence of approximations converges to $E(Y|X_1, X_2)$.

This is easy to translate into finite data form. Write f and g in terms of a finite spline basis. Hold the coefficients in f fixed and minimize the RSS over the coefficients in g. Now hold those in g constant and minimize over those in f.

This algorithm is a fast and parsimonious way for representing interaction. For example, if, in their spline bases, f and g have p degrees of freedom, then the minimizing product fg has about p degrees of freedom in it. One adds more multiplicative terms until there is no significant decrease in RSS. Furthermore, the multiplicative terms are easy to interpret.

Unfortunately, numerical results indicate that in the nonindependence case, there are a number of local minima in addition to the global minimum. The algorithm always converges, but it may not converge to the global minimum. This makes the selection of a good starting point important. Our experimental results have been that if we use the starting point given by assuming independence, then the iterates have always converged toward the global minimum.

I am currently working on straightening up the details of this representation of bivariate interaction and hope to go public soon.

REFERENCES

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* **80** 580–619.

BREIMAN, L. and PETERS, S. (1988). Comparing automatic bivariate smoothers. Technical Report, Dept. Statistics, Univ. California, Berkeley.

FRIEDMAN, J. H. and STUETZLE, W. (1982). Smoothing of scatterplots. Technical Report, Orion 3, Dept. Statistics, Stanford Univ.

HASTIE, T. and TIBSHIRANI, R. (1988). Comment on "Monotone regression splines in action" by J. O. Ramsay. *Statist. Sci.* **3** 450–456.

SMITH, P. (1982). Curve fitting and modelling with splines using statistical variable selection techniques. NASA Contractor Report 66034.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

ZEHUA CHEN, CHONG GU AND GRACE WAHBA[1]

University of Wisconsin-Madison

We must begin by thanking the authors for a thought-provoking work. As is well known [Kimeldorf and Wahba (1971) and Wahba (1978)], quadratic penalized likelihood estimates (with nonnegative definite penalty functionals) are Bayes estimates. Let $\mathbf{y} = \mathbf{g} + \boldsymbol{\varepsilon}$ with $\mathbf{g} \sim N(0, \Sigma)$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$, then

$$\hat{\mathbf{g}} = \Sigma(\Sigma + \sigma^2 I)^{-1}\mathbf{y} = A\mathbf{y}, \quad \text{say},$$

which also minimizes $(1/\sigma^2)(\mathbf{y} - \hat{\mathbf{g}})'(\mathbf{y} - \hat{\mathbf{g}}) + \mathbf{g}'\Sigma^{+}\mathbf{g}$, the resulting smoother matrices are all symmetric nonnegative definite with their eigenvalues in $[0, 1)$. This generalizes to the case where Σ is improper, which gives eigenvalues $+1$.

[1]Research supported by AFOSR Grant AFOSR 87-0171.