ADDITIONAL REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723.

BATES, D. M., LINDSTROM, M. J., WAHBA, G. and YANDELL, B. S. (1987). GCVPACK—Routines for generalized cross-validation. *Comm. Statist. B—Simulation Comput.* **16** 263–297.

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

GASSER, TH., MÜLLER, H. G., KÖHLER, W., MOLINARI, L. and PRADER, A. (1984). Nonparametric regression analysis of growth curves. *Ann. Statist.* **12** 210–229.

GOLUB, G., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223.

GREEN, P., JENNISON, C. and SEHEULT, A. (1985). Analysis of field experiments by least-squares smoothing. *J. Roy. Statist. Soc. Ser. B* **47** 299–315.

JUPP, D. L. B. (1978). Approximation to data by splines with free knots. *SIAM J. Numer. Anal.* **15** 328–343.

NYCHKA, D. (1986). The mean posterior variance of a smoothing spline and a consistent estimate of the mean squared error. Unpublished.

O'SULLIVAN, F. (1983). The analysis of some penalized likelihood schemes. Technical Report 726, Dept. Statistics, Univ. Wisconsin-Madison.

O'SULLIVAN, F. (1986). Estimation of densities and hazards by the method of penalized likelihood. Technical Report 58, Dept. Statistics, Univ. California, Berkeley.

O'SULLIVAN, F., YANDELL, B. S. and RAYNOR, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–103.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

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

SMITH, P. (1983). Curve fitting and modeling with splines using statistical variable selection techniques. Unpublished.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation.* Johns Hopkins Univ. Press, Baltimore, Md.

UTRERAS, F. (1985). Smoothing noisy data under monotonicity constraints: Existence, characterization, and convergence rates. *Numer. Math.* **47** 611–625.

VILLALOBOS, M. and WAHBA, G. (1987). Inequality constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Amer. Statist. Assoc.* **82** 239–248.

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 semiparametric estimation of functions of several variables. In *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface* (T. E. Boardman, ed.) 75–80. Amer. Statist. Assoc., Washington.

Comment

Trevor Hastie and Robert Tibshirani

Professor Ramsay has written an informative paper about a topic that is new (at least to us) and deserves exposure. The techniques that he describes and his software implementations are potentially useful in a number of different areas. However, we found that after careful reading of the paper and experimenting with monotone splines, we are in substantial disagreement with him over a number of important points. In particular:

- The monotonicity assumption inherent in monotone splines will sometimes (often?) be unwarranted. A more useful modeling technique allows a choice of smoother for each variable, perhaps between linear, monotone and nonmonotone, together with a strategy for selecting the appropriate form. A general estimation procedure called *backfitting* can be used to estimate models of this kind.
- The number and position of knots *can* make a difference and we can see no clear way to make these choices. Other smoothing techniques such as smoothing splines have the significant advantage that a single smoothing parameter controls the smoothness of the output.
- The number of parameters inherent in a monotone spline is *not* "far fewer" than the number in a cubic smoothing spline or other common smoothers, given a comparable amount of smoothness.
- The data analysis in the paper are somewhat weak and potentially misleading.

Trevor Hastie is a member of the Statistics and Data Analysis Research Department, AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974. Robert Tibshirani is Assistant Professor and NSERC University Research Fellow, Department of Preventive Medicine and Biostatistics, and Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A8.

Regarding the last point, we note that we ourselves have been rightly accused of limp data analyses in this journal; see Brillinger (1986), Hodges (1987) and Draper (1987).

We now elaborate on these and other points.

WHY MONOTONE?

Nonparametric tools such as smoothers are meant to be exploratory. Thus it doesn't seem wise to restrict a function to be monotone *a priori* unless there is a very good reason for doing so. For example, a monotone restriction makes sense for a response transformation because it is necessary to allow predictions of the response from the estimated model. Similarly, in Ramsay's factor analysis model, the monotone transformations can be thought of as a different metameter for the variables. On the other hand, why restrict predictor transformations (such as for displacement and weight in the city gas consumption problem) to be monotone? Instead, why not leave them unrestricted and let the data suggest the shape of the relevant transformation? In some situations, the issue is murky. For example in a dose-response experiment, we might have reason to expect a monotone relationship: then it would be informative to try both a monotone and unrestricted fit.

The following strategy would seem to be useful in general: fit an unrestricted model and then see if a monotone model (with some or all of its components monotone) fits as well. If this were the case, the monotonicity would allow easier interpretations. Of course one should also check whether a simple linear fit is adequate for any of the components.

A BETTER WAY

Through use of the "backfitting" procedure (Friedman and Stuetzle, 1981; Breiman and Friedman, 1985; Hastie and Tibshirani, 1986) one can mix monotone splines with other smoothers to estimate more general models. Consider for example the Down's Syndrome data (Section 4.7). For illustration, suppose there were additional predictors available besides mothers age. Then one can choose a smoother for each variable: perhaps a monotone spline for one, a cubic spline for another, linear least squares fit for another, etc. The choice would be based on the nature of the variable and *a priori* scientific considerations. Then the overall model is estimated in an iterative fashion, using each smoother in turn. Hastie and Tibshirani (1986) give details. This hybrid approach may be used in most of the settings that Ramsay considers (for example ACE works in this way) and to us provides important flexibility.

NUMBER OF PARAMETERS IN THE FITTED FUNCTION

Ramsay says in Section 2 that "far fewer coefficients" are estimated in regression splines than in smoothing splines. We doubt that this is the case.

To investigate this, we first need a suitable definition of the number of parameters of a smoother. Cleveland (1979), Wahba (1983) and Hastie and Tibshirani (1986) develop various notions of "degrees of freedom" or "effective number of parameters" in a fitted smooth. Consider a set of n fitted values $\hat{\mathbf{y}} = (\hat{y}_1, \cdots, \hat{y}_n)^t$ based on a response $\mathbf{y} = (y_1, \cdots, y_n)^t$. Then the simplest definition is $\mathrm{df}(\hat{\mathbf{y}}) = \sum \mathrm{var}(\hat{y}_i)/\sigma^2$ where $\sigma^2 = \mathrm{var}(y_i)$. If a smoother is linear so that $\hat{\mathbf{y}}$ can be written as $S\mathbf{y}$, where S is an $n \times n$ "smoother" matrix not involving $\mathbf{y}$, then $\mathrm{df}(\hat{\mathbf{y}}) = \mathrm{trace}(SS^t)$ and thus can be computed without knowledge of $\mathbf{y}$. (Note that smoothing splines with a fixed smoothing parameter are linear whereas monotone splines, even considering the knots fixed, are not.)

Now Ramsay seems to imply that for a smoothing spline, $\mathrm{df}(\hat{\mathbf{y}})$ is close to the number of data points n. However, when we apply a cubic smoothing spline to data (or some other smoother like a running line or kernel), the value of $\mathrm{df}(\hat{\mathbf{y}})$ ranges from about 2 to 6. According to Ramsay a monotone spline of order 3 (piecewise cubic polynomials) with m interior knots uses $m + 3$ parameters. Thus one or two interior knots would result in about the same complexity as a cubic smoothing spline, at least for smoothing parameters in the range that we use in practice. The actual number of parameters in a monotone (cubic) spline is probably less than $m + 3$ because of the restrictions necessary to enforce monotonicity. A more exact calculation would require estimating $\sum \mathrm{var}(\hat{y}_i)$ by simulation. More details on degrees of freedom can be found in Cleveland and Devlin (1988) and Buja, Hastie and Tibshirani (1988).

A COMPARISON OF SMOOTHING AND REGRESSION SPLINES

One method for comparing the behavior of two smoothers is to examine the equivalent kernel over the domain x. For a linear smoother, we can plot the rows of the smoother matrix S against x. Note that if s_i is the ith row, then $\hat{y}_i = \sum_{j=1}^{n} s_{ij} y_j$. In this demonstration we compare cubic smoothing splines to (nonmonotone) regression splines, because our focus is on the effect of the knot placement. As data (x values) we use the 128 unique values of a pressure gradient variable described in a later section. We used a cubic regression spline with one interior knot at the median. The cubic smoothing spline was calibrated so that $\mathrm{trace}(SS^t) = 5$, which is the number of parameters in

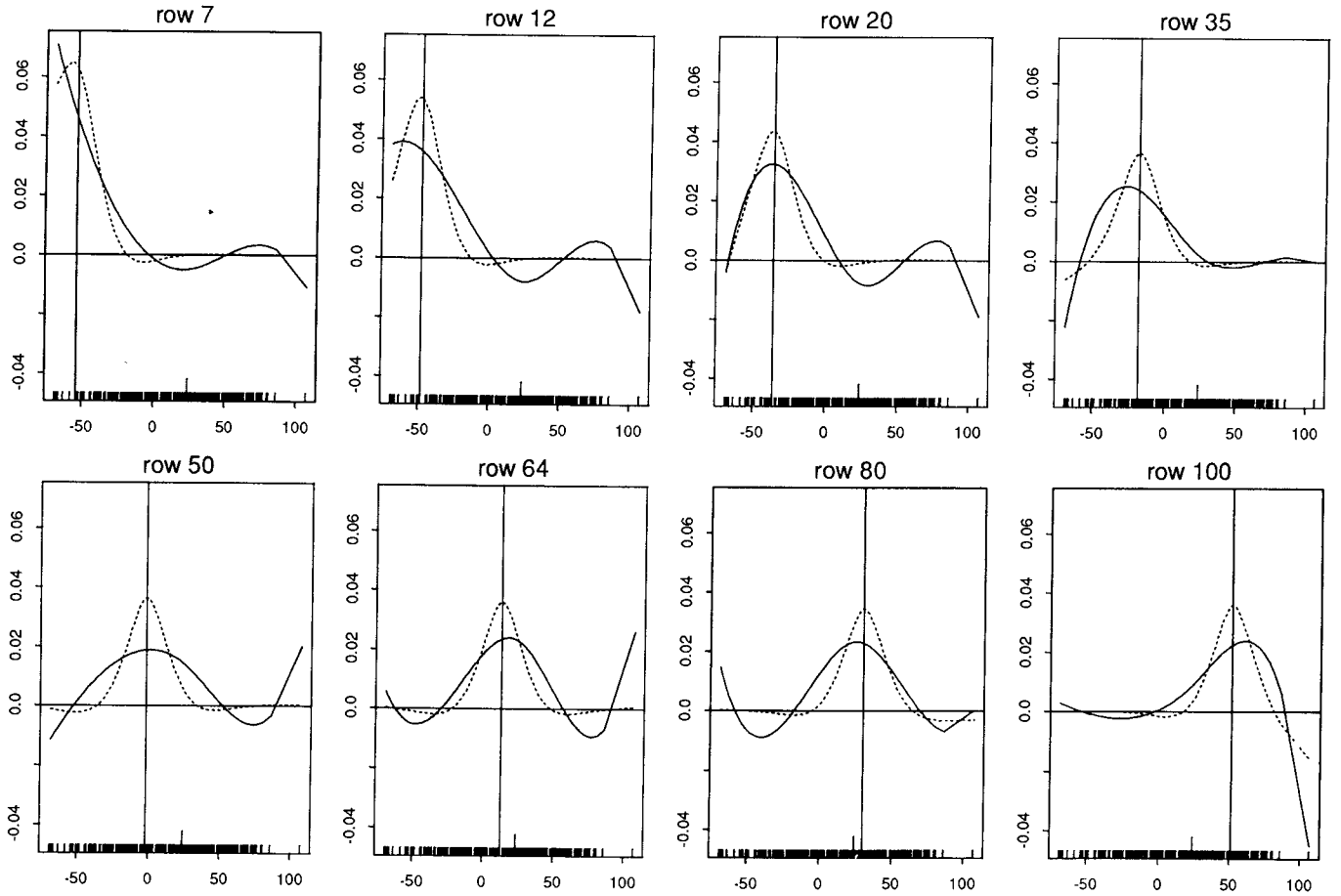

FIG. 1. *A comparison of the equivalent kernels of a cubic regression spline with one interior knot and a cubic smoothing spline with* 5 *"degrees of freedom." The x values are the* 128 *unique values of* DAGGOT PRESSURE GRADIENT *used in Figure* 4, *and are represented by the "whiskers" at the base of the plot. The one interior knot at* 24 *is indicated, and the vertical line is the target point of the kernels. The broken curves represent smoothing splines, the solid curves the regression splines.*

a (nonmonotone) regression spline. Figure 1 compares arbitrarily selected rows of the smoother matrices (dashed curves are for smoothing splines, solid curves for regression splines). The whiskers at the base of the plot indicate values of x, the longer tick indicating the one interior knot. The vertical line indicates the target point at which the fitted value is desired. The most alarming feature of these plots is that the regression spline shows some very nonlocal behavior in contrast to the smoothing spline. It is also influenced more by the extreme right "outlier."

Another question of interest is: how big is the class of functions that the smoother can recover or detect? The eigendecomposition of the respective smoother matrices is illuminating in this regard, because the eigenvalues tell us by how much the smoother "shrinks" the corresponding eigenvector. Figure 2 compares the eigenvalues of the smoothing spline to those of the regression spline used above. Because the regression spline is a projection, we know that it has 5 eigenvalues equal to 1 and the rest zero. The smoothing spline has two eigenvalues equal to 1 (corresponding to the space of linear functions), and then they shrink toward zero. Because the sum of squared eigenvalues is 5 in both cases, we see that the cubic spline sacrifices exact fits of lower order functions for some detection of higher order ones.

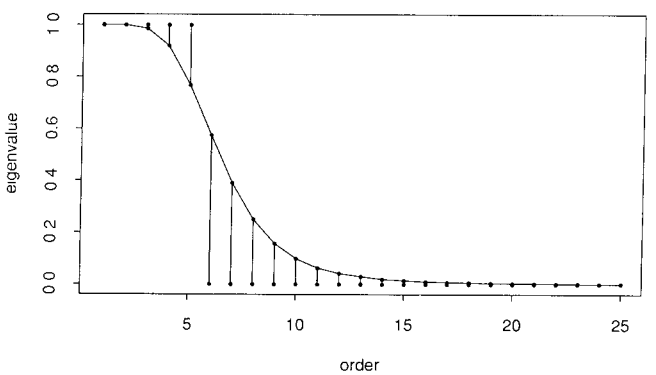

FIG. 2. *A comparison of the eigenvalues of the regression spline and smoothing spline used in Figure* 1, *truncated at number* 25. *The regression spline is a projection, and has* 5 *eigenvalues of* 1, *the rest are zero.*

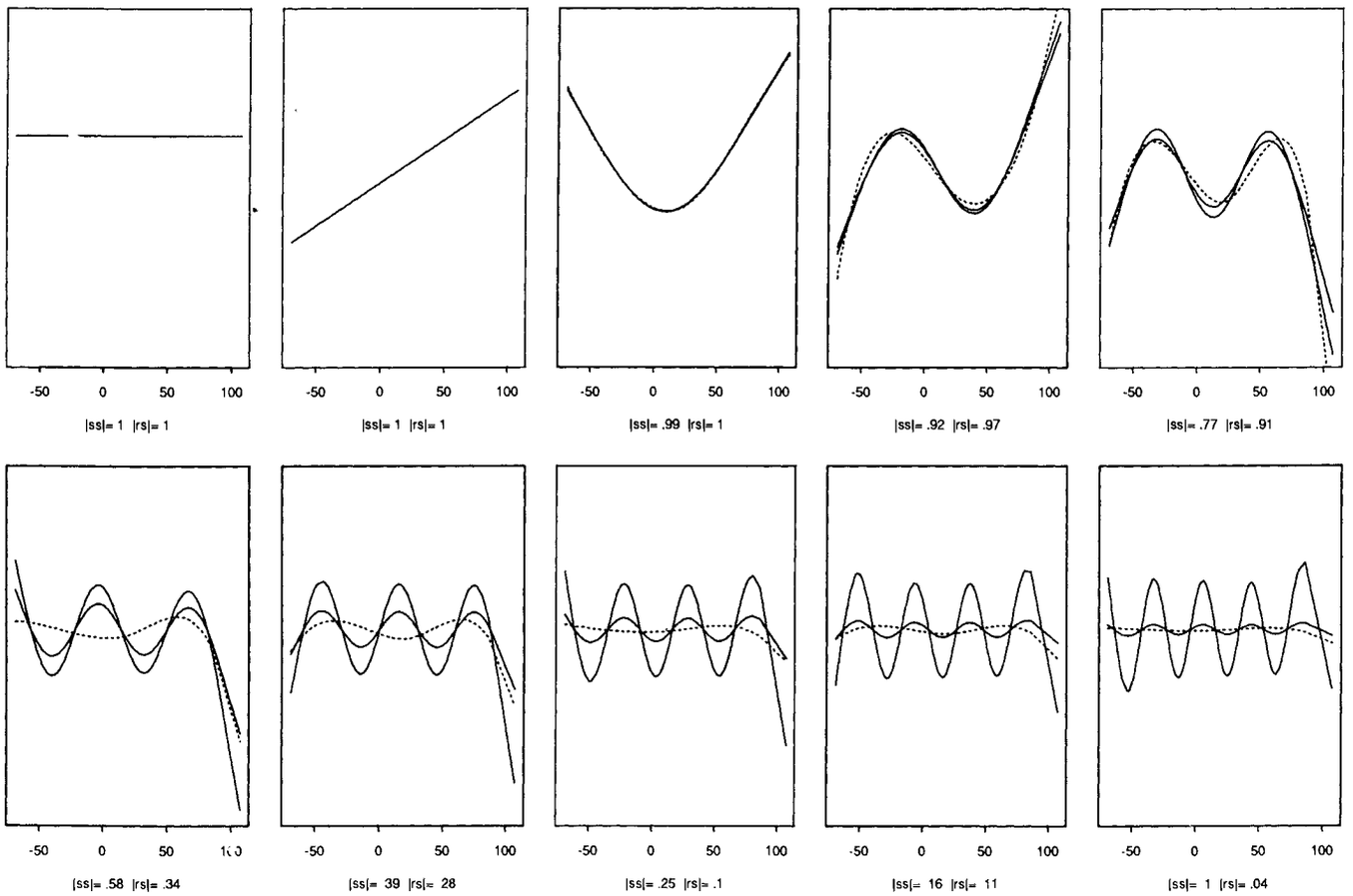

FIG. 3. *The effect of smoothing the first* 10 *eigenvectors of the smoothing spline by itself (shrunken broken curve) and by the regression spline (solid curve). The numbers at the base of the plots give the norms of the fitted functions, and hence an indication of how much shrinking was done (ss = smoothing spline; rs = regression spline).*

The eigenvectors for the smoothing spline resemble orthogonal polynomials, in that the number of zero-crossings increases with the order of the eigenvalue. It is instructive to see what happens when we smooth these eigenvectors with either smoother. Figure 3 shows the first 10 eigenvectors. The solid curve is the smoothing spline eigenvector, the shrunken solid curve is the result of the smoothing spline applied to this eigenvector, and the dashed line is the result of the regression spline. The space of eigenvectors for the regression spline corresponding to eigenvalue 1 is spanned by the four-dimensional space of cubic polynomials plus a piecewise cubic that resembles a quartic polynomial. The regression spline therefore has no trouble with the first five eigenvectors of the smoothing spline, and does better than the smoothing spline itself! After that it essentially annihilates all higher order functions, whereas the smoothing spline produces increasingly shrunken versions of these. Again the smoothing spline has sacrificed a little bit on the higher order functions, in order to partially recover higher order functions. This comparison may be a little unfair, though, because we are playing ball in the smoothing spline's home court.

AN ACTION REPLAY

In Section 4.4 Ramsay uses monotone splines to enhance a canonical correlation analysis. He reports $\rho^2 = .806$, and displays all the monotone transformations. We ran a standard linear canonical correlation analysis on the same data, and the resulting ρ^2 was .803!

In Section 4.3, monotone additive regression was applied to the car data. The conclusions are an itemized list of interpretations, separating the roles of the two regressors over their ranges. Yet the correlation between these two variables is 0.90! The "confidence curves" certainly do not support these separate interpretations either. Even the need for transformations is unclear in this case. We fit a linear regression model to these data using the same transformed response and untransformed covariates. A crude test of the effect of the transformations, using approximate degrees of freedom described earlier, does not reject the linear model.

One of the more challenging problems in additive regression modeling is to provide appropriate guidelines for when separate interpretations are valid, and

to gauge the sensitivity of the models to different functional forms for the covariates. If simpler models (such as the linear model) fit as well, we would certainly want to know about it.

AN EXAMPLE

We tried fitting a monotone spline regression (Ramsay's Section 4.3) using the ozone concentration data analyzed by Breiman and Friedman (1985). For this purpose Dr. Ramsay kindly supplied us with his FORTRAN code for monotone spline regression. There are 330 observations, the response being OZONE CONCENTRATION and the two predictors, INVERSION BASE HEIGHT and DAGGOT PRESSURE GRADIENT chosen, for simplicity, from a set of 10 predictors. (ATTENTION DRAPER AND HODGES: We are using this for illustration only; a more complete analysis can be found in Breiman and Friedman, 1985!) Figure 4 a–c shows the effect of increasing the number of knots on the estimated monotone splines. The estimate for DAGGOT PRESSURE GRADIENT varies quite a bit, especially when one remembers that the monotonicity assumption reduces its freedom quite substantially. Figure 4, d, e and f, shows the estimated transformations from ACE. The ACE results suggest that the effect of DAGGOT PRESSURE GRADIENT and possibly INVERSION BASE HEIGHT are not monotone. A crude F-test rejected monotonicity in both cases.

A MISCONCEPTION ABOUT ACE

Ramsay makes some incorrect statements about ACE at the end of Section 4.3: "There is also a theoretical explanation for this phenomenon ...." This is an attempt to explain the fact that for the city gas consumption data, the estimates depend on the order that the variables are entered into the algorithm. Although it is true that the population version of ACE solves an eigenvalue problem, so does the data version, if linear smoothers are used in the algorithm. Details can be found in Breiman and Friedman (1985). Buja,

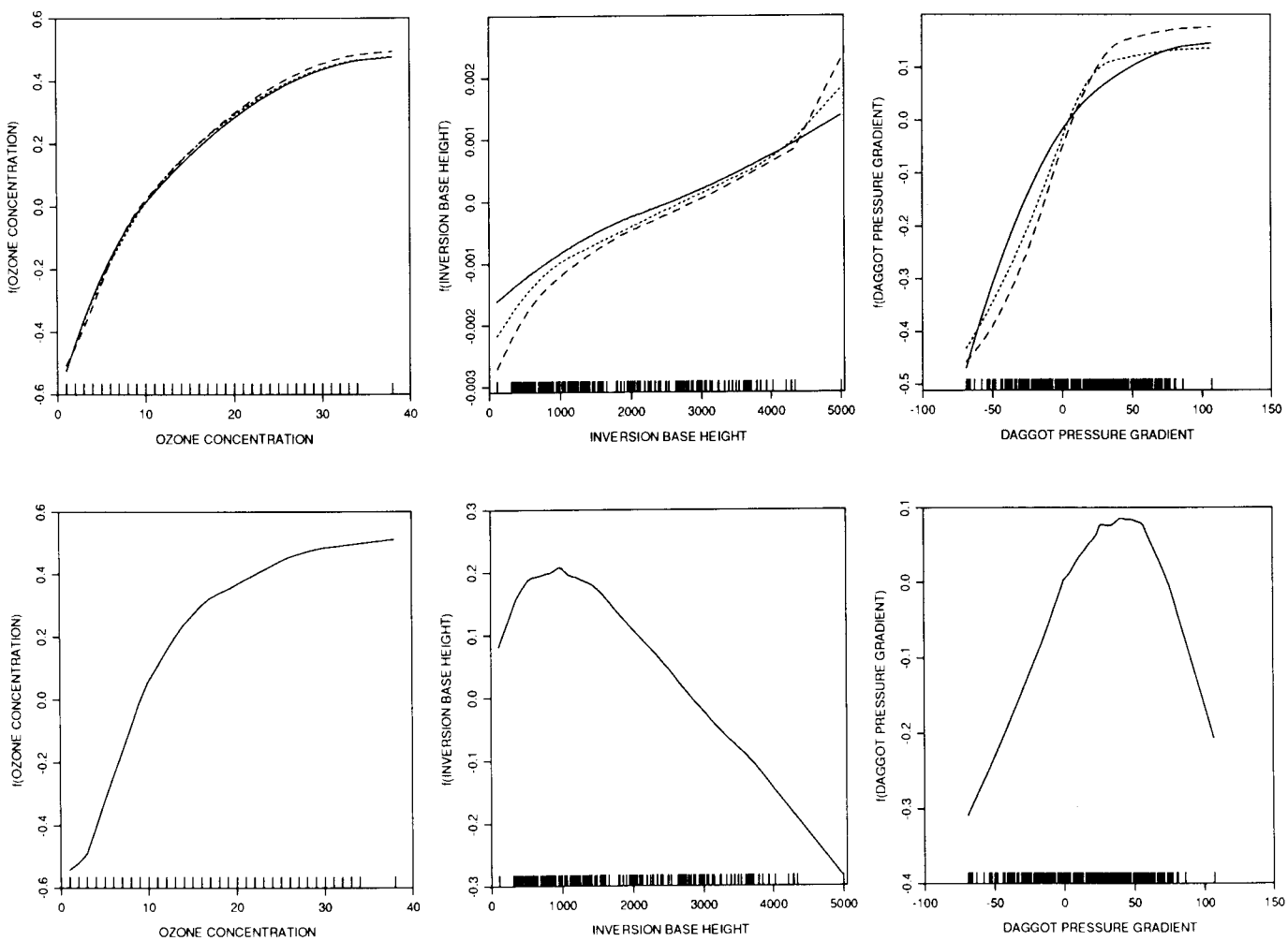

FIG. 4. a–c, *the fitted functions using monotone splines on the meteorological data. The different curves show the effect of adding knots. Solid line, 1 knot at median; dotted line, 2 knots at tertiles; dashed line, 3 knots at quartiles.* d–f, *the fitted functions produced by ACE for the same data.*

Hastie and Tibshirani (1988) extend the results of Breiman and Friedman and show that if cubic spline smoothers are used in ACE, the algorithm converges to the desired solution, independent of the order or values of the starting functions, unless there is an exact collinearity among the predictors.

The implementation of ACE used by Ramsay utilizes "supersmoother," a highly nonlinear smoother. Thus one can't make definitive statements about convergence. However, we found similar convergence difficulties using cubic smoothing splines in this problem. The source of the trouble is the high collinearity between the predictors. This causes the algorithm to converge quite slowly and the residual sum of squares can change very little in the iterations despite the fact that the estimated functions are still changing considerably. The ACE implementation, used by Ramsay, decides to terminate based on the change in residual sum of squares. Reducing the default threshold value doesn't seem to help in this example. The solution is to use the change in the functions as the termination criterion: this alleviates the problem. Perversely, we might say that ACE is warning us about the strong correlation between the regressors!

Actually, there is a strong similarity between ACE and the monotone spline procedure. To facilitate the comparison, let's use monotone splines as the smoother in ACE. Then one can show that ACE minimizes the residual sum of squares between the transformed variables, subject to $\text{var}(f_0(Y)) = 1$ and the monotonicity constraints. We assume Ramsay is using the Box-Cox type likelihood criterion. In this case one can show that the criterion he minimizes is residual sum of squares, subject to $(\prod_{i=1}^{n}[f_0'(Y_i)]^2)^{1/n} = 1$ and the monotonicity constraints. Both criteria are the same up to the scale functional used to penalize the transform of Y. In particular, this version of the ACE problem can be solved without the iterations needed for more general smoothers. It also seems clear that if the ACE criterion has multiple minima, then so does the likelihood criterion.

An alternative method for this problem uses the notion of a variance stabilizing transformation (the "AVAS" procedure, Tibshirani, 1988).

CORRELATION AMONG FITTED VALUES

Ramsay states in Section 4.1 that the "lack of coupling among distinct regions where the curve is most flexible is one of the great virtues of splines." We don't understand this point. A binning smoother (that is, taking means in disjoint partitions) results in zero correlation in fitted values between different partitions, but this is not a very useful smoother. Put another way, the process of smoothing uses local averaging to "borrow strength" across predictor values, so we should expect correlations among fitted values. This is highlighted in the equivalent kernels in Figure 1. The smoothing spline borrows strength from observations close by. The regression spline borrows strength in a very nonlocal fashion.

MODELS FOR BINOMIAL DATA

Ramsay suggests smoothing binomial data with a monotone spline constrained between 0 and 1. Apart from our objection to monotonicity, we note that this procedure works only for one covariate. For more than one, Hastie and Tibshirani (1986) suggest the generalized additive model: $\log\{P(x_j)/[1 - P(x_j)]\} = \beta_0 + \sum_{k=1}^{p} f_k(x_{jk})$ which forces the fitted probabilities to lie in (0, 1). The f_k can be monotone or arbitrary smooth functions. Bachetti (1987) describes a closely related procedure for binary data, using isotonic regression in a backfitting algorithm. We also note that Friedman and Tibshirani (1984) proposed an ad hoc method for monotone smoothing that entails smoothing the data with an arbitrary smoother, then applying isotonic regression to the smoothed values. This can also be used in the backfitting procedure.

CONCLUSIONS

Our discussion has emphasized points of disagreement (naturally) and hence has been critical. Overall, we feel that monotone splines are potentially useful, especially in settings for which a monotonicity requirement is natural or in conjunction with other smoothers in a backfitting algorithm. We are concerned about the difficulty of choosing the number and position of knots.

With the increasing flexibility of modern regression tools, there is, more than ever, a danger of over-interpreting results. We feel that for any of these tools to be useful, we need a strategy for selecting an appropriate model that will protect us from over-interpretation, and we must be guided by considerations of the scientific (data) problem at hand. Thus we endorse wholeheartedly the crusade of Draper and Hodges, namely that "a pint of technique combined with a quart of numbers" does not yield a data analysis. We hope to pursue some of these questions and hope that Dr. Ramsay will consider them in future research as well.

ADDITIONAL REFERENCES

BACHETTI, P. (1987). Binary additive isotonic regression. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

BRILLINGER, D. R. (1986). Comment on "Generalized additive models" by T. Hastie and R. Tibshirani. *Statist. Sci.* **1** 310–312.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1988). Linear smoothers and additive models (with discussion). *Ann. Statist.* To appear.
CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.
CLEVELAND, W. S. and DEVLIN, S. (1988). Locally -weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 596–610.
DRAPER, D. (1987). Comment on "Prediction in growth curves" by C. R. Rao. *Statist. Sci.* **2** 454–461.
FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.
FRIEDMAN, J. and TIBSHIRANI, R. (1984). The monotone smoothing of scatterplots. *Technometrics* **26** 243–250.
HODGES, J. S. (1987). Uncertainty, policy analysis and statistics (with discussion). *Statist. Sci.* **2** 259–291.
TIBSHIRANI, R. J. (1988). Estimating transformations for regression via additivity and variance stabilization. *J. Amer. Statist. Assoc.* **83** 394–405.
WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.

Comment

Grace Wahba

Professor Ramsay is to be congratulated for writing a lively and interesting paper and giving us a handy descriptive tool.

Without at all intending to criticize "eyeball" methods, which play an important role in data analysis, it should be clear that the success of the method depends on the ability of the user to select the number and location of the knots to give a pleasing picture. As the author observes, in the examples given, the results are fairly insensitive to knot location. This is, of course why it is difficult to select knots in the computer by an objective numerical criterion—numerically, that is an ill-posed problem. I would expect that the picture would be different if the number of knots is changed drastically.

Subjective notations of what the answer "ought to" look like appear to play an important role in the proposed method.

Having said this, I would like to raise the issue of "subjective" versus "objective" inference, both of which clearly play a role in statistics. Of course, the dividing line between these types of inference are blurred, every "objective" method has some subjectivity behind it, namely, the statistician had some preconceived framework about the truth when selecting a technique (no matter how "objective" the technique is). Conversely, any good "subjective" method, ideally will display the data in such a way that the "facts" about the truth are helped to come out.

One way of classifying subjective versus objective techniques is the following. A technique may be viewed to be on the objective end of the spectrum, if at least in principle, one could discover its properties on at least some useful class of "truths," by simulating data from various "truths," applying the technique, and studying how well the inference matched the simulated "truth." If I had an "objective" method for constructing confidence intervals or estimating variances, then I could run a big Monte Carlo study and see whether in fact the confidence intervals or estimated variances had an appropriate relation to the simulated "truth."

I am somewhat concerned here with the use of, for example, the "estimated sampling variance" of a. It appears that these estimates are conditioned on certain subjective choices made by the statistician. If I really wanted to claim that these estimates had some objective properties if used in the future, I should do a simulation study, sampling from a population of users who are going to use the eyeball method for choosing the location and number of knots.

On a different tack, I would like to thank Professor Ramsay for his kind reference to my work on smoothing splines and to take this opportunity to compare and contrast smoothing and regression splines. Positivity and monotonicity can also be imposed on smoothing splines (see Villalobos and Wahba (1987) and references cited there), and there is quite a bit of activity in the development of efficient alogrithms for doing this, but, in the absence of user-oriented software, it is work to start from scratch to implement a relatively objective constrained smoothing spline as described in Villalobos and Wahba (1987).

The monotone regression splines, as proposed by Professor Ramsay, appear to be quite accessible to relatively unsophisticated users who know how to call a quadratic programming algorithm.

In examples with larger data sets, smoothing splines do have the ability to resolve finer structure than

Grace Wahba is Bascom Professor of Statistics, Department of Statistics, University of Wisconsin, 1210 West Dayton Street, Madison, Wisconsin 53706. These comments were written while the author was Clare Boothe Luce Visiting Professor of Statistics at Yale University.