Comment

Randy Eubank

The author is to be congratulated on his insightful and far-reaching article. This article represents a substantial contribution to the field of nonparametric function estimation due in particular to the many new application areas it will introduce to statisticians who are concerned with developing such methodology. The collection of examples Professor Ramsay presents pose challenging problems, that are also of practical importance, for study by nonparametric data modelers. It will be interesting to see the alternative solutions that are sure to be developed as a result of this article. I am delighted to get one of the first shots at this and will suggest some other possible approaches in Section 2 below.

I found myself in agreement with much of what Professor Ramsay has said. My few points of disagreement are methodological, rather than philosophical, and therefore minor. My primary concerns are related to the problems of knot selection and inference in spline fitting. Here my experiences appear to be almost the opposite of the author's. I will elaborate further on this in the next section.

It seems to me that nonparametric and parametric estimation methods are too often viewed as competitors to one another. There is no reason that these methods cannot or should not be used in tandem. Indeed, it is foolish to do otherwise when conducting exploratory data analyses. I would therefore like to expand on Professor Ramsay's point concerning his example in Section 4.1. He notes that although his nonparametric approach did not lead to new results, it provided us with reassurance concerning a parametric fit. This is an illustration of how nonparametric procedures can provide diagnostic checks concerning the validity of parametric models. Whereas no difficulties were uncovered with the parametric model for the data in question, this need not always be the case. Serious parametric modeling deficiencies can be uncovered by comparison with nonparametric fits. An excellent illustration of this is the growth curve analysis conducted by Gasser, Müller, Köhler, Molinari and Prader (1984).

1. CHOOSING KNOTS, INFERENCE AND RELATED ISSUES

The transformations Professor Ramsay employs are splines of some specified order k with n knots having locations contained in a set t_n. In practice k, n and t_n must all be selected in some fashion. I agree with the author that the choice of k is generally not crucial; with $k = 4$ (i.e., cubic splines) being statisfactory for most purposes (provided the number and locations of the knots have been chosen correctly). However, he also indicates that good choices of n and t_n can be easily made and that the shape of a spline function is robust with respect to these choices. This view is inconsistent with my experiences and those of many others.

To illustrate my point, consider the data in Figure 1 which represent a property, Y, of titanium as a function of heat, X. (See de Boor (1978), page 222 for the actual data.) Three cubic splines have been fitted to the data via least squares, the first uses five uniformly spaced knots whereas the second has five knots that have been selected more carefully. Notice that although both fits have the same order and number of knots, they are not even remotely similar in shape. The spline based on uniformly spaced knots is also a woeful fit being very little (if at all) better than fits obtained using polynomials.

This example illustrates that knot placement can be crucial for both the shape and quality of a spline fit. Some ad hoc rules for good knot placement can be found in Wold (1974). They are motivated by the same considerations which led the author to propose his two guidelines for this purpose in his Section 3.

A more objective knot selection can be accomplished by optimizing the estimation criterion of interest with respect to both the knot set t_n and the vector a of spline basis coefficients. For example, both t_n and a could have been chosen to maximize the sample likelihood in the examples discussed in Sections 4.1 and 4.2. This idea is by no means new and has even been suggested by the author (Winsberg and Ramsay, 1980). The second fit in Figure 1 was obtained by optimizing the knot locations and therefore illustrates the gains to be realized from optimal knot selection.

Once the knot locations have been optimized, n can be chosen using various model selection techniques. A criterion such as that of Akaike (1974) would be easy to use with likelihood-based fitting methods.

Unfortunately, I am not optimistic about the practicality of methods for optimal knot selection. In the context of least squares estimation this is a very poorly behaved nonlinear optimization problem (Jupp, 1978). It is unlikely that matters will improve for more general likelihood-based methods.

An alternative approach to selecting knots is to

Randy Eubank is Professor, Department of Statistics, Texas A&M University, College Station, Texas 77843-3143.

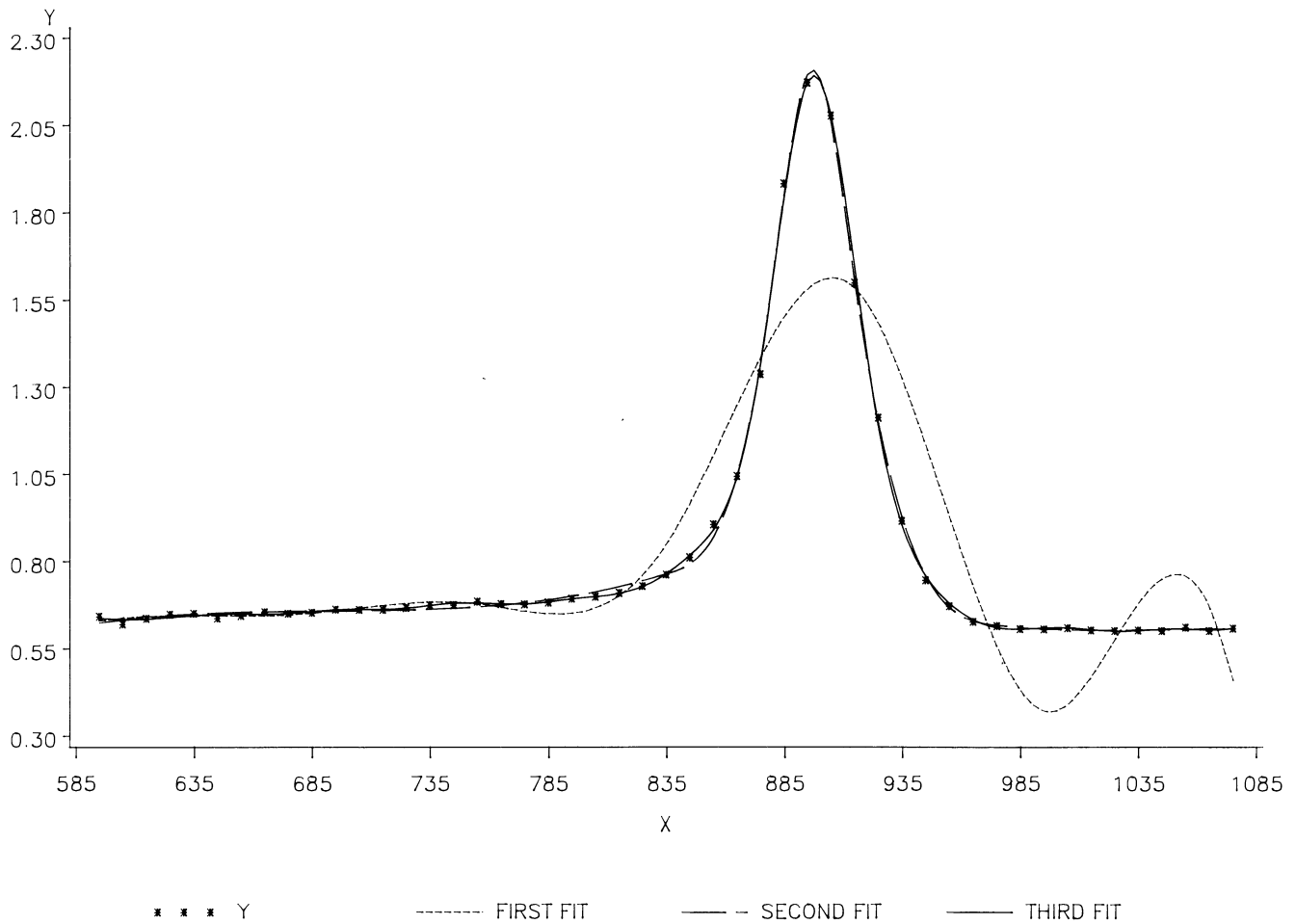

FIG. 1. *Fits to the titanium heat data.*

choose the knots to be the n-tiles of some density, such as a uniform, and then optimally select n via model selection methodology. Knots which do not contribute to the fit can be eliminated using appropriate statistical tests for their corresponding basis function coefficients. This is in the spirit of proposals by Smith (1983) and Stone (1985). An illustration of this procedure is provided by the third fit in Figure 1. Here the knots were uniformly spaced with the number of knots selected by generalized cross-validation (GCV). (See, e.g., Golub, Heath and Wahba, 1979.) The fit selected by GCV includes 19 knots (only five of which really contribute to the fit) and is quite similar to the fit obtained using five optimal knots.

The above discussion is intended to indicate that the issue of knot selection is not, in general, as elementary as the author seems to suggest. However, Professor Ramsay is primarily concerned with monotone transformations which exclude the possibility of functions with peaks and valleys and, in particular, my example in Figure 1. I wonder if monotonicity simplifies the knot selection problem and would be interested in the author's experiences and insights concerning this case. I am especially curious as to why good results should be expected from the placement he suggests of locating knots at the n-tiles of the variable to be transformed.

Another point on which I would like to comment concerns the author's approach to interval estimation of the transformed predicted values. The intervals he employs are constructed using variances that are computed as if the actual transformation which makes the data lognormal, or whatever, is a spline. In fact, there is a true transformation $f(\cdot)$ which accomplishes this that is being approximated in some sense by a spline function $f(\cdot; a)$. Thus there are two sources of estimation error: the difference between $f(\cdot)$ and $f(\cdot; a)$ and the difference between $f(\cdot; a)$ and its estimator $f(\cdot; \hat{a})$. Typically, the method of inference utilized by Professor Ramsay can be expected to adequately account for the random error resulting from the difference between $f(\cdot; a)$ and $f(\cdot; \hat{a})$ but will not include appreciable information on the deterministic error $f(\cdot) - f(\cdot; a)$. If the latter component is substantial, this can create problems with interval estimates for fitted or predicted values because they will not be correctly centered. Such problems will not improve by optimal selection of n and t_n. I also feel it is unlikely

that jackknifing or bootstrapping will prove to be a panacea for these difficulties, although I would like to be proven wrong. It therefore seems prudent to at least utilize root mean square error estimates in this type of interval estimation rather than standard errors that are based on an admittedly wrong transformation model. A possible advantage of the alternate methods proposed in the next section is that they have a Bayesian interpretation that has been shown in some cases to lead to "confidence intervals" that provide some compensation for bias.

2. PENALTY FUNCTION METHODS

In this section I would like to outline an alternative approach to estimating transformations of a variable Y. The class of procedures I will describe might be termed penalty function methods and are formulated as follows. Given a vector y of R observations on a variable Y that is to be transformed, a transformation is estimated by minimizing a criterion of the form

$$Q(y; f) = G(y; f) + \theta S(f), \quad \theta > 0,$$

over an appropriate collection of transformations or functions f. The criterion Q is a sum of two components. The first of these, $G(y; f)$, is assumed to assess the goodness-of-fit of the transformation to the data, whereas the second, $S(f)$, will measure the smoothness of f. The parameter θ then governs how much weight is given to the two competing aspects of the criterion. A typical choice for S is $S(f) = \int (f''(t))^2 dt$ which measures the curvature of the transformation. In the context of regression analysis $G(y; f)$ is usually taken to be the sum of squared errors. However, for general applications G can be chosen to be many things including the negative of the log of the sample likelihood.

General results which can be used to answer questions concerning the existence and uniqueness of estimators derived from Q are given in Tapia and Thompson (1978) and O'Sullivan (1983). Techniques for computing and analyzing the properties of such estimators are provided by O'Sullivan (1983, 1986) and O'Sullivan, Yandell and Raynor (1986). Code which might be adapted for computational purposes is discussed in Bates, Lindstrom, Wahba and Yandell (1987).

If the transformation is required to be monotone, convex, positive or whatever, this can be accomplished by restricting the functions over which Q is minimized accordingly. Methodology that may prove useful in the analysis of the resulting constrained minimization problem is described in Tapia and Thompson (1978), Wright and Wegman (1980), Utreras (1985) and Villalobos and Wahba (1987).

Objective data-driven choices for the parameter θ in Q can be made using natural parallels of the generalized cross-validation criterion. The basic idea is described in O'Sullivan, Yandell and Raynor (1986).

To illustrate the idea, consider the problem of transforming $\log(Y)$ to normality. Then one might attempt to find a transformation f which minimizes the penalized likelihood criterion

$$Q(y; f) = \sum_{r=1}^{R} \left(\frac{(f(\log(y_r))^2}{2\sigma^2} - \log(f'(\log(y_r))) + \log(y_r) \right) + \theta \int_{-\infty}^{\infty} (f''(t))^2 \, dt$$

subject to $f' \geq 0$. One can show that if Q is minimized over the set of all absolutely continuous functions with absolutely continuous, nonnegative derivatives and square integrable second derivatives, there will be a unique minimizer that is a cubic spline function with one continuous derivative. Variants of this particular suggestion could be employed to give alternate analyses for the problems discussed in Sections 4.1 and 4.2. Specific details will be treated elsewhere. However, one of the interesting and desirable aspects of this proposal is that when $\theta = \infty$ the estimator of f reverts back to the one obtained by using the parametric power transformation bY^{λ}, because then f must be linear in $\log(Y)$. To quote Silverman (1982), "Since one of the objects of nonparametric methods is to investigate the effect of relaxing parametric assumptions, it seems sensible that the limiting case of a nonparametric ... estimate should be a natural parametric estimate."

Some specific penalty function methods for the analysis of the other examples studied by Professor Ramsay are also easy to formulate and have already been studied in some cases. For example, an analysis similar to the one conducted on the data in Section 4.3 using an additive approximation could be carried out using the interaction splines developed by Wahba (1986). For situations such as those of Sections 4.6 and 4.7 where dichotomous response variables are to be analyzed, penalty function techniques have been used by O'Sullivan, Yandell and Raynor (1986) and Villalobos and Wahba (1987). Concerning the case of the canonical correlation example in Section 4.4 and the principal components example of Section 4.5, G might be chosen to be the canonical correlation coefficient and Ramsay's trace criterion which could then be paired with suitable smoothness penalties to obtain transformation estimates.

When G is chosen to be the log of the sample likelihood, penalty function methods will be intimately related to Bayesian procedures with $S(f)$ representing a prior for the transformation. (See Wahba

(1978, 1983) and references in Eubank (1988).) This connection can be exploited to derive Bayesian prediction intervals that can be used for interval estimation concerning values of f. Results in Wahba (1983) and Nychka (1986), for the special case of regression analysis, indicate that these Bayesian intervals use what is very nearly a mean squared error estimate in their assessment of estimation error. One might hope this would carry over to the more general setting discussed above; although, at present, this remains an open issue.

3. SPLINE SMOOTHING VERSUS SPLINE REGRESSION

The two approaches to estimating transformations set out in Professor Ramsay's article and the previous section are fundamentally different in nature. These differences are easiest to see in the setting of regression analysis. Therefore, let us assume a simple regression model with a uniform design wherein

$$y_r = f(r/R) + \varepsilon_r, \quad r = 1, \cdots, R,$$

with f representing an unknown regression function and the ε_r assumed to be zero mean, uncorrelated random errors having some common variance. A smoothing spline estimator of f can be obtained by minimizing $\sum_r (y_r - f(r/R))^2 + \theta \int (f''(t))^2 dt$. When formulated appropriately, the solution to this problem is a cubic natural spline with knots at the data values.

Another possible estimator, that is more in line with the article under discussion, is a cubic spline regression estimator of f. For simplicity assume that the knot set t_n has been chosen *a priori* and let $f(\cdot; a)$ denote a cubic spline with knots at t_n. Then a least squares or regression spline estimator of f is obtained by minimizing $\sum_r (y_r - f(r/R; a))^2$ with respect to the vector of basis coefficients a.

Professor Ramsay suggests that the principal difference between these two methods for estimating f is that the cubic smoothing spline has R knots, one at each data value, whereas the least squares spline estimator has some specified set of knots t_n. He then equates the number of knots to relative degrees of freedom for inference and flexibility of the two methods. Such statements are, in my view, both misleading and incorrect.

First note that the knot locations play no role whatsoever in controlling the actual shape of a smoothing spline. This is governed by the smoothing parameter θ. In contrast, the number and the locations of the knots are entirely responsible for the shape or smoothness of a least squares regression spline. It is therefore clear that relative location and number of knots represents only a superficial difference between the two methods. For an accurate comparison one must look a little deeper than this.

Both smoothing splines and least squares splines are linear estimators. Thus, in each case there is a function $W(\cdot, \cdot)$ such that the fitted or predicted value at a point x can be written as $\sum_r y_r W(x, r/R)$. In the case of smoothing splines, work by Silverman (1984) has the consequence that $W(x, y) \doteq W(|x - y|/\theta^{1/4})/R\theta^{1/4}$ for a specific symmetric kernel function $W(\cdot)$. Thus smoothing spline estimators behave essentially like kernel estimators in the way they use the data. The corresponding form of $W(x, y)$ for least squares splines is not so simple. In particular, it cannot be expressed as, or even well approximated by, a function of only $|x - y|$. Its shape is governed by where x and y lie relative to the knot set t_n. Thus, in my opinion, the principal difference between smoothing and least squares splines lies in how they smooth the data. Smoothing splines can be expected to perform like kernel estimators whereas least squares splines will not.

Because least squares splines are obtained by a projection, it makes sense to tie the number of knots (or equivalently the number of basis functions) to degrees of freedom. Smoothing splines do not stem from a linear projection and, hence, what is meant by degrees of freedom in this case is less clearcut. Some suggestions for what should be used for degrees of freedom for inference with smoothing splines have been made by Wahba (1983) and Green, Jennison and Seheult (1985).

4. SUMMARY

The article by Professor Ramsay describes a number of exciting application areas and analysis techniques for those interested in nonparametric data modeling. In the above discussion I have indicated some possible alternative methodology and some points of disagreement with the inference procedures that were employed. However, such details should not be confused with what I feel is the main issue here: namely, the benefits that can be realized from using flexible, nonparametric estimation procedures in lieu of, or in conjunction with, parametric methods. It is my suspicion that many people choose to use parametric methods, even when they know they are not appropriate, because they feel the results will be more interpretable. Professor Ramsay has aptly demonstrated in his article that important and interpretable conclusions can also be drawn from data using nonparametric procedures. For this he has my vote of thanks.

ACKNOWLEDGMENT

I am grateful to Paul Speckman for a helpful discussion concerning the material in Section 3.

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.