Rejoinder

J. O. Ramsay

The discussants have extended the article in a number of useful directions, indicated some possible limits on its applicability, corrected some inaccuracies and asked some fascinating questions. Whatever the merits of the article, I take considerable satisfaction in providing the occasion for so much fine commentary. I will only add a few more remarks for further clarification.

Much attention is given to comparing regression splines and smoothing splines in the context of the linear smoothing of a bivariate relationship. Although this discussion is illuminating and useful, this problem is not the focus of the article, as was indicated at the beginning of Section 4. Rather, my concern was with the simultaneous nonlinear estimation of a number of transformations given a limited amount of data. In this context, simplifying assumptions or constraints such as monotonicity and knot placement with only a limited data dependency become important in order to reduce tendencies to over-fit as well as to give some protection against algorithmic instabilities such as that illustrated for ACE. As Wahba has indicated, simultaneous curve estimation is not as well understood, although much progress has been made in the additive model. Comparatively simple procedures such as those presented in the paper should offer useful benchmarks against which more sophisticated proposals can be assessed. Thus, the spirit of the article is closer to that of Box and Cox (1964) than to the large smoothing spline literature.

What is the essential difference between regression and smoothing (or interpolation) splines? In both cases data transformations are defined as linear combinations of certain basis functions. In either case the role of the data in defining the basis depends on how knots are chosen. The more knots and the more data-driven the procedure for positioning knots, the more dependency there is. Eubank and others are right to caution that adaptive knot positioning can be essential in some situations. In others simple procedures such as positioning at quantiles will suffice, and a variety of methods have already been explored to permit the user to detect when this is so.

The basis for natural smoothing splines in the context of the bivariate relation $\{x_i, y_i \mid i=1, \cdots, n\}$ is the set of n functions $k(x_i, x)$, where k is the bivariate reproducing kernel associated with the Hilbert function space underlying the nature of the smoothing process. Kimeldorf and Wahba (1971) and Besse and Ramsay (1986) discuss the relationship between L-splines and reproducing kernel Hilbert spaces. Thus, in all but exceptional cases the matrix of basis function values $k(x_i, x_j)$ for smoothing splines has rank equal to the number of observations.

For regression splines the matrix of basis function values will in general be of much smaller rank because the number of basis functions is as a rule rather less than the number of values x_i. Computational considerations thus tend to favor regression splines when the number of estimated curves is large and/or it is desirable to keep the number of estimated coefficients small.

On the other hand, Hastie and Tibshirani have provided a needed correction in pointing out the difference between rank and the number of effective degrees of freedom, and Buja, Hastie and Tibshirani (1988) is commended to the reader's attention. The situation is analogous to regression in the presence of a large number p of highly collinear predictors. Assuming that variation associated with very small singular values of the independent variable data matrix has little predictive value, the effective number of independent variables can be much smaller than p, and some rank reduction procedure such as regression on principal components can be very effective.

Is monotonicity really a useful feature? Although Breiman, Hastie and Tibshirani are not convinced, we can perhaps leave this to users to decide. Nevertheless, there are situations such as dose-response curve estimation where inversion of the transformation is essential, just as there are others where one should at least consider a comparison of nonmonotone and monotone transformations. Moreover, as noted above the imposition of monotonicity is itself a sort of smoothing procedure which can be helpful, especially when more than one transformation is being estimated.

I do agree that we know relatively little about interval estimation in this field, and I thank the discussants for some needed cautionary remarks. However, when interval estimates based on relatively restrictive assumptions such as conditioning on specific parametric families and using asymptotic arguments still show that a transformation is poorly estimated, as was the case for displacement in the automobile regression example, we can be reasonably confident that better techniques will not change the picture, and something useful has therefore been learned.

Breiman wonders why maximum likelihood estimation is favored rather than least squares, and in fact ACE and other techniques designed to deal with the generalized additive model tend toward the minimization of the ratio of error sum of squares to the variance of the transformation f_0 of the dependent variable y. This is a critical issue because transformation of the dependent variable affects the criterion itself in ways rather different from what is familiar in conventional linear modeling. This is reflected in the fact that MLE appends a Jacobian term to the error sum of squares in the log likelihood (7) for Gaussian error. Failure to do so can result in serious bias (Ramsay, 1977), and it is not clear to this author what the rationale for minimizing a variance ratio or maximizing squared correlation is when y is transformed.

To illustrate this problem, consider the results of the following modest simulation study. The three variables from the automobile data analyzed in Section 4.3 on the monotone spline regression were used to generate 100 simulated sets of data as follows. For each simulated sample, independent Gaussian errors with mean zero and standard deviation 0.174 were added to the 44 values of the sum of transformed values of displacement and weight. The transformations f_1 and f_2 used were those estimated for the original data and shown in Figure 5, and the error level was that estimated for these data. The perturbed values were then back-transformed to yield simulated city gas consumption values

$$y_r = f_0^{-1}[f_1(x_{r1}) + f_2(x_{r2}) + e_r], \quad r = 1, \cdots, 44,$$

with f_0 being that estimated for the original data.

Each of the 100 simulated samples was then analyzed using four procedures: (1) MLE as decribed in the article, (2) minimizing error sum of squares $S(y, x; a, \theta)$, (3) minimizing the variance ratio $S(y, x; a, \theta)/\mathrm{Var}[f_0(y)]$ and (4) ACE. The extent to which the four procedures were able to recover the dependent variable transformation f_0 is displayed in Figure 1, which shows for each value of y the

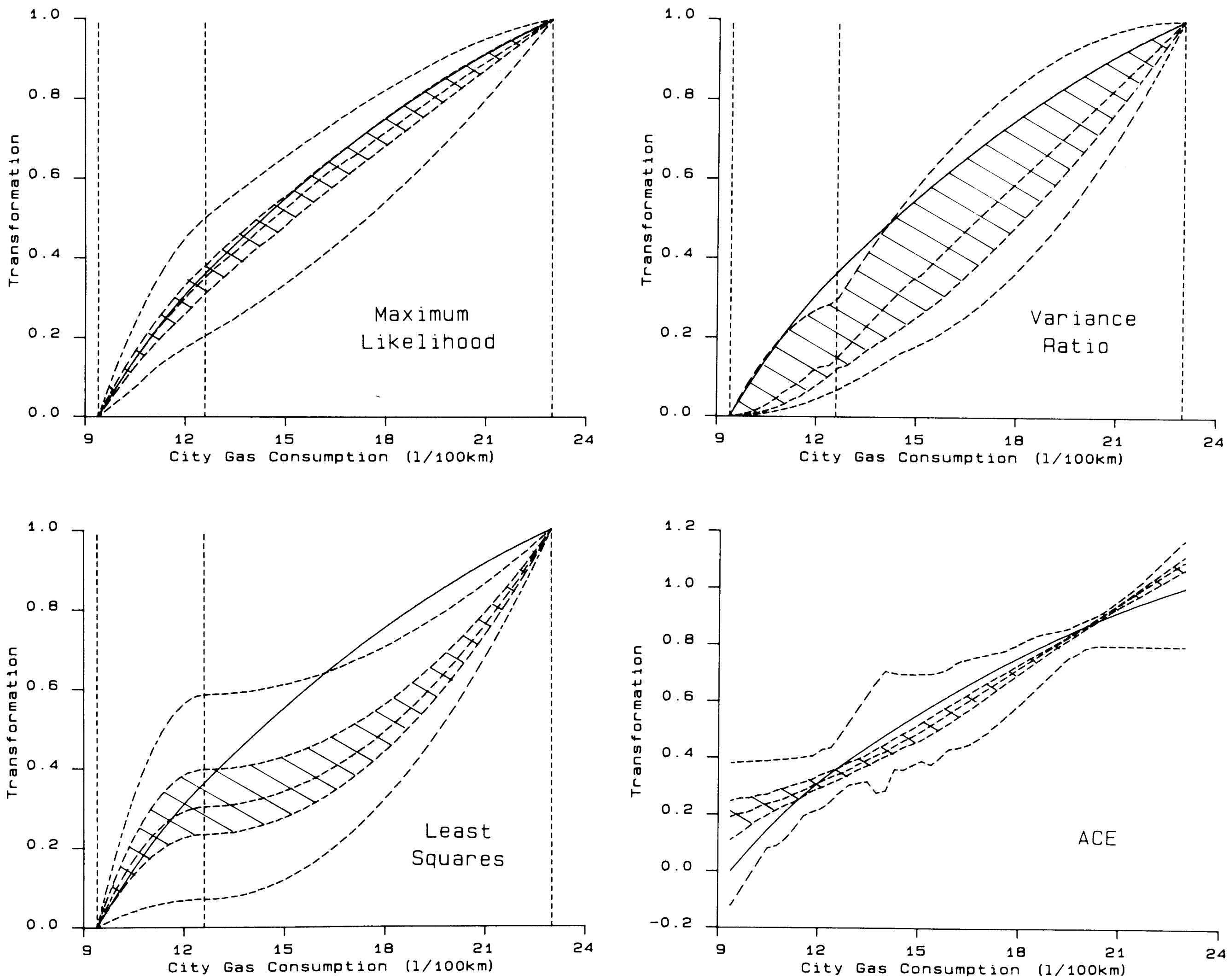

FIG. 1. *Results for the transformation of city gas consumption from analyzing* 100 *simulated sets of automobile data by four procedures:* (1) *monotone spline regression of simulated city gas consumption on displacement and weight using maximum likelihood estimation,* (2) *minimizing the ratio of error sum of squares to variance of transformed y,* (3) *minimizing error sum of squares and* (4) ACE. *For each value of y the minimum,* 25%, *median,* 75%, *and maximum estimated transformation values are indicated by dashed curves, and the true transformation by the solid curve. Vertical dashed lines for the monotone spline regressions indicate knot placement.*

minimum, 25%, median, 75%, and maximum values of $f_0(y)$ across the 100 analyses (dashed lines) as well as the actual transformation used (solid line). In the case of ACE, which uses a different normalization procedure, estimated transformation values were first matched to the true transformation values by least squares before assessment.

Maximum likelihood estimation works well in the sense that the true curve lies within the 25% and 75% values (the cross-hatched area) almost everywhere, and the variability seems reasonably small and quite consistent with the asymptotic estimates shown in Figure 5 of the article. Minimizing the variance ratio, on the other hand, produces large sampling variability and a suggestion of considerable bias. Worst of all is least squares estimation, and this is a consequence of that the fact that the error sum of squares can be partially reduced in the absence of the Jacobian term by compressing the variation in $f_0(y)$. ACE results are similar in terms of bias to those of variance ratio minimization because it, too, uses this criterion. I conjecture that this bias is at least partly due to the fact that the variance ratio can also be reduced somewhat by compressing variation in f_0 for the data in the central portion of their distribution while leaving $\mathrm{Var}[f_0(y)]$ constant by expanding the tails.

To conclude my response, the techniques in the paper, although possibly not competitive for the linear estimation of a single transformation with possible local high curvature, appear to have a useful role when many transformations are being estimated (transformation to multinormality, extensions of familiar multivariate procedures, estimation of item characteristic curves, etc.), where computational considerations are critical, and where the amount of data available gives concern about overfitting. But we have much to learn, and it is a pleasure to acknowledge the importance of the discussion.

ADDITIONAL REFERENCES

BESSE, P. and RAMSAY, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika* **51** 285–311.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1988). Linear smoothers and additive models (with discussion). *Ann. Statist.* To appear.

KIMELDORF, G. S. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–94.

RAMSAY, J. O. (1977). Monotonic weighted power transformations to additivity. *Psychometrika* **42** 83–109.