REFERENCES

DE LEEUW, J. (1982). Nonlinear principal component analysis. In *COMPSTAT 1982: Proceedings in Computational Statistics* (H. Caussinus, P. Ettinger and R. Tomassone, eds.) 77–86. Physica, Vienna.

DONNELL, D., BUJA, A. and STUETZLE, W. (1994). Analysis of additive dependencies and concurvities using smallest additive principal components. Unpublished manuscript.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141.

KETTENRING, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika* **58** 433–451.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

BELLCORE
445 SOUTH STREET
MORRISTOWN, NEW JERSEY 07962–1910

TREVOR HASTIE

AT&T Bell Laboratories

Professor Stone has done an admirable job in leading us through the difficult mathematics needed to build a firmer theoretical framework around high-dimensional nonparametric regression and density estimation techniques. ANOVA decompositions of regression surfaces are no longer confined to the case when the predictors are categorical; we can now play the same games in function spaces. Gu and Wahba (1991) describe similar decompositions in reproducing-kernel Hilbert spaces using tensor-product smoothing splines.

This comment moves us to the opposite boundary of the field and describes some computational tools for expressing and fitting tensor-product spline models of this kind in the S language [Becker, Chambers and Wilks (1988)].

In S there is a *formula language* for expressing models, primarily aimed at traditional ANOVA and linear models. For example, the formula $\sim a * (b + c)$ expands to $\sim a + b + c + a{:}b + a{:}c$ and expresses a model with main effects and interactions. Typically the variables `a`, `b` and `c` are factors. The formula is converted into a *model matrix* where the factors are coded via contrast matrices, and their interactions as matrix tensor products of these. The contrast matrix for a factor is a basis for representing the piecewise constant effect as a function of its levels; this is the default behavior for factors, and in fact a default contrast coding is used. This notion is extended by allowing the following in formulas: (i) variables representing matrices and (ii) expressions that are calls to functions, which evaluate to matrices.

We now elaborate in the context of regression splines.

There are some primitive functions in S, for example, `poly(x,...)`, `bs(x,...)` and `ns(x,...)`, for producing polynomial, B-spline and natural B-spline bases, respectively. The function `bs( )` (which we focus on here) has additional arguments relating to *knot placement* and *degree*, and returns a matrix corresponding to the specified B-spline basis evaluated at the values of `x`. For example,

`bs(x, knots = c(1.5, 3))` defaults to a cubic-spline basis with two interior knots. More simply, `bs(x, df = 6)` will return a B-spline basis matrix with three interior knots selected automatically at the appropriate interior quantiles of the supplied `x`. The argument `df = 6` refers to the *degrees of freedom* of the basis, or number of linearly independent columns (a column corresponding to the intercept is excluded).

Using these basis functions in the formula language, we can express polynomial tensor-product spline models in a natural way. For example,

$$\sim \texttt{bs(w,5)} + \texttt{bs(x,5)} + \texttt{bs(z,5)} + \texttt{bs(x,5):bs(z,5)},$$

or, more simply,

$$\sim \texttt{bs(w,5)} + \texttt{bs(x,5)} \star \texttt{bs(z,5)}$$

represents a model with main effects plus a selected tensor-product interaction term. Formulas such as these are used in the modelling software, as in the following example:

$$\texttt{glm(y} \sim \texttt{bs(w,5)} + \texttt{bs(x,5)} \star \texttt{bs(z,5)},\ \texttt{family = binomial)}.$$

This fits a logistic regression model to the response variable `y`. The right-hand side of the model formula is used to construct a model matrix built up from the specified main effect and tensor-product B-spline basis. The `family = binomial` argument implies a logistic link function by default, but other links are possible, for example, `family = binomial(link = probit)`. Among other families are `poisson` for log-linear models, `gamma`, as well as specialized families such as `robust` for fitting models resistant to outliers.

Although no explicit care is taken to orthogonalize the appropriate collections of terms in the model, this happens automatically. The columns of the model matrix corresponding to the modelling formula are arranged in an hierarchical order, and the successive orthogonalization is achieved when the model is fit by iteratively reweighted least squares via the Gramm–Schmidt method.

Models can have terms of mixed types, such as factor-by-spline interactions. Facilities are available for plotting the fitted terms in a variety of different ways. Stepwise model selection procedures are available at a high level. One can specify the highest-order model, such as

$$\texttt{full} \leftarrow \texttt{glm(y} \sim \texttt{bs(w,5)} \star \texttt{bs(x,5)} \star \texttt{bs(z,5)},\ \texttt{family = binomial)}$$

and then `step(full)` will search for the best-fitting submodel in a hierarchical fashion. Note that in this case the number and placement of knots is fixed per variable, and one is looking for the best ANOVA subspace.

The software and modelling tools are described in detail in Chambers and Hastie (1991). As yet no software is provided for density estimation or conditional density estimation, but these could easily be built on the current facilities.

REFERENCES

BECKER, R. A., CHAMBERS, J. M. and WILKS, A. R. (1988). *The New S Language*. Wadsworth/Brooks Cole, Pacific Grove, CA.

CHAMBERS, J. M. and HASTIE, T. J. (1991). *Statistical Models in S*. Wadsworth/Brooks Cole, Pacific Grove, CA.

GU, C. and WAHBA, G. (1991). Smoothing spline ANOVA with componentwise Bayesian "confidence intervals." Technical Report 881, Dept. Statistics, Univ. Wisconsin–Madison.

AT&T BELL LABORATORIES
MURRAY HILL, NEW JERSEY 07974

REJOINDER

CHARLES J. STONE

University of California, Berkeley

I wish to thank Buja and Hastie for their interesting and stimulating remarks. In particular, Buja's improvement over my Lemmas 3.3 and 3.4 is very elegant and may be useful in other contexts. His joint work with Donnell and Stuetzle on the analysis of additive dependencies in data sounds intriguing, and I look forward to reading about it soon.

Hastie gives a brief but excellent description of the formula language in S and the ease with which it can be used in the context of linear and generalized linear models to specify main effects as polynomial splines and selected interactions in terms of the corresponding tensor products. He points out that stepwise model selection procedures are also available in S for determining which main effects and interactions to include; that is, in the notation of the present paper, for adaptively choosing $\mathcal{S}$. As he also notes, however, these facilities are not convenient for selecting the number and placement of knots. The high-level stepwise model selection facilities that are currently available in S are compatible with the spirit of the theory developed in the present paper, but not with that of methodologies such as MARS that are adaptive at the level of the individual basis functions, that is, that adaptively select the individual knots and tensor product basis functions.

Recently, in Kooperberg, Stone and Truong (1993b), the theory developed in the present paper has been modified to handle hazard regression, which can be nonproportional and which includes a smooth model for the baseline hazard function. The corresponding MARS-like adaptive methodology is described in Kooperberg, Stone and Truong (1993a). Kooperberg has written a program in C that implements this methodology and an interface based on S. The combined software is available from statlib by sending an email with the body `send hare from S` to statlib@stat.cmu.edu. Concurrently, Kooperberg and Stone (1993) described similar methodology and software for hazard estimation without covariates. Kooperberg, Truong and I are now working on the theory and methodology for logspline spectral density estimation, while Bose, Kooperberg