MO, M. (1990a). Robust additive regression I: Population aspect. Unpublished manuscript.
MO, M. (1990b). Robust additive regression II: Finite sample approximations. Unpublished manuscript.
MO, M. (1991). Nonparametric estimation by parametric linear regression (I): global rate of convergence. Unpublished manuscript.
MORGAN, J. N. and SONQUIST, J. A. (1963). Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.* **58** 415–434.
NEWEY, W. N. (1991). Consistency and asymptotic normality of nonparametric projection estimators. Unpublished manuscript.
RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.
SCHUMAKER, L. L. (1981). *Spline Functions*: *Basic Theory*. Wiley, New York.
SMITH, P. L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. Report NASA 166034, NASA, Langley Research Center, Hampton, VA.
STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.
STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.
STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.
STONE, C. J. (1989). Uniform error bounds involving logspline models. In *Probability*, *Statistics and Mathematics*: *Papers in Honor of Samuel Karlin* (T. W. Anderson, K. B. Athreya and D. L. Iglehart, eds.) 335–355. Academic, New York.
STONE, C. J. (1990a). Large-sample inference for log-spline models. *Ann. Statist.* **18** 717–741.
STONE, C. J. (1990b). L_2 rate of convergence for interaction spline regression. Technical Report 268, Dept. Statistics, Univ. California, Berkeley.
STONE, C. J. (1991a). Asymptotics for doubly flexible logspline response models. *Ann. Statist.* **19** 1832–1854.
STONE, C. J. (1991b). Multivariate logspline conditional models. Technical Report 320, Dept. Statistics, Univ. California, Berkeley.
STONE, C. J. and KOO, C.-Y. (1986a). Additive splines in statistics. In *Proceedings of the Statistical Computing Section* 45–48. Amer. Statist. Assoc., Washington, DC.
STONE, C. J. and KOO, C.-Y. (1986b). Logspline density estimation. In *Automated Theorem Proving*: *After 25 Years* (W. W. Bledsoe and D. W. Loveland, eds.). *Contemp. Math.* **29** 1–15. Amer. Math Soc., Providence, R.I.
TAKEMURA, A. (1983). Tensor analysis of ANOVA decomposition. *J. Amer. Statist. Assoc.* **78** 894–900.

DEPARTMENT OF STATISTICS
STATISTICAL LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DISCUSSION

ANDREAS BUJA
Bellcore

Previous work by Stone has been impressive, and the present paper commands even more respect. In one grand sweep, he develops convergence rates for B-spline interaction models in LS regression, in ML generalized regression, in log-density estimation and in conditional log-density estimation. In some ways, Stone's work provides a unifying theory unmatched by existing methodology—a relatively uncommon situation. As he states at the end of Section 2, the usefulness of extensions of MARS [Friedman (1991)] to generalized regression, density estimation and conditional density estimation is suggested by this work.

Again with MARS in mind, Stone is very modest in stating the limitations of his theory: adaptive selection of effects and knots is not theoretically tractable within his framework. In this regard, in the regression context, his modeling approach is more similar to the penalized interaction splines of the Wisconsin school [e.g., Wahba (1990), Chapter 10] than to MARS. Penalized interaction splines are also based on a careful a priori selection of a functional ANOVA model. In MARS, the ANOVA decomposition is just a postprocessing step after a voraciously greedy search among component terms that generally do not belong to any orthogonalized main effect or interaction space.

With a piece covering as much ground as Stone's paper, it is impossible to comment on all or even only the major aspects. I therefore confine myself to two points that are somewhat arbitrary but close to my own interests: (1) the problem of identifiability in nonadaptively chosen models and (2) a suggestion with respect to the aesthetics of some proof details.

Identifiability as a theoretical and practical problem. As a rule, to prove strong results, strong assumptions must be made, and Stone's work is no exception. In the regression context, for example, he models the predictors as realizations of a random vector taking on values in an M-dimensional cube (w.l.o.g. $[0, 1]^M$) and having a probability density that is bounded away from zero and infinity. This smacks more of a somewhat unbalanced factorial experiment than the type of messy observational data to which we often apply nonparametric models. It would be premature, though, to criticize Stone's results on grounds of his choice of technically motivated conveniences. It turns out that the lemmas derived from these assumptions express the foundations of his results much better than the assumptions themselves. A case in point is Section 3, where he tackles identifiability problems among other things. A striking statement is Lemma 3.1: Its assertion,

$$E\left[\left(\sum_s h_s(\mathbf{X})\right)^2\right] \geq C\sum_s E[h_s^2(\mathbf{X})], \tag{1}$$

for some $C > 0$, is one way of limiting confounding (unidentifiability) in the ANOVA decomposition for the population model. Complete confounding describes a situation where one main effect or interaction can be expressed as the sum of other types of main effects or interactions. This is equivalent to the existence of a nontrivial additive relation, $\sum_s h_s(\mathbf{X}) = 0$, that would clearly violate (1). Now, confounding is, of course, a real problem in real data. Curiously, the technical assertion (1) is a pointer to a practical method for the detection of confounding. To this end, consider the following constrained optimization problem:

$$(2)\qquad E\left[\left(\sum_s h_s(\mathbf{X})\right)^2\right] = \min \quad \text{subject to} \quad \sum_s E\left[h_s^2(\mathbf{X})\right] = 1.$$

If, for given data, a solution to this problem could be estimated, one would obtain a fair idea of the nature of confounding and the extent of its danger. As it turns out, this minimization problem can be tackled and solutions can be practically estimated. We [Donnell, Buja and Stuetzle (1994)] hope to have a paper on this topic soon ("Analysis of additive dependencies and concurvities using smallest additive principal components"). Our work concerns the analysis of confounding in additive models, but it is trivial to extend the idea to the analysis of main effects and interactions in an ANOVA-type decomposition of a space of fits. The underlying principle is always the same: Given a decomposition of a Hilbert space $\mathbf{H} = \mathbf{H}_1 + \cdots + \mathbf{H}_p$, minimize $\|\Sigma h_i\|^2$ under $\Sigma\|h_i\|^2 = 1$, $h_i \in \mathbf{H}_i$. In Stone's context, the component spaces $\mathbf{H}_i$ are the orthogonalized main effects and interaction spaces H_s^0.

In one guise or another, this idea of "generalized canonical analysis" has been around for a long time, usually formulated in terms of data rather than populations, and most frequently in the form of the corresponding maximization problem, as, for example, in multiple correspondence analysis [e.g., de Leeuw (1982)]. Apparently, it is Kettenring (1971) who first introduced the minimization problem. The current preoccupation with nonparametric estimation was missing in those days, so the ideas were presented as generalizations of canonical analysis to more than two blocks of variables. Consequently, the interpretation in terms of the detection of confounding in the decomposition of a space of fits is missing as well.

It would be interesting to hear in more detail how the author thinks about the problem of confounding, both theoretically and in terms of guidance for practitioners.

Some technical remarks. The second point I would like to discuss concerns the proofs of Lemmas 3.3 and 3.4. There, Stone shows that, on spaces of fits, the empirical inner product is uniformly close to the population inner product. The lemmas are crucial as it is here that the rates emerge at which the number K of spline intervals can be increased as a function of the sample size n. (The main results of the paper are formulated in terms of J, the dimension of the space of splines obtained when subdividing a variable into K intervals. Rates in terms of J are the same as those in terms of K due to an obvious linear relationship. See Stone's remarks after his Condition 2.)

Stone's strategy is to prove closeness of empirical and population inner products for spaces of multivariate polynomials on a single cube (Lemma 3.3) and then extend this to spaces of functions that are piecewise polynomials on small cubes (Lemma 3.4). His technique for stitching the cubes together is through conditioning on the lattice of cubes and applying Lemma 3.3 to the polynomials in each cube. This may sound straightforward, but the details are hairy, involving estimates for the convergence of the numerators and denominators of the conditional expectations. Having used Hoeffding's inequality for Lemma 3.3, Stone needs Bernstein's inequality to bound denominators in his proof of Lemma 3.4.

This approach violated my sense of aesthetics to such a degree that I tried an alternative method. It turns out that a mild modification of the proof of Lemma 3.3 yields both a simplified proof and a slight strengthening of Lemma 3.4. To rid the proofs of conditioning, we calculate unconditional bounds on small cubes of edge length $1/K$ within the unit cube $[0, 1]^M$, the main difference being that we look at small cubes with small mass, while Stone's conditioning normalizes the mass of small cubes to 1, thus artificially introducing unpleasant denominators.

We use Stone's assumptions and notations: the unit cube $[0,1]^M$ supports the data distribution given by a density $f(\mathbf{x})$ that is bounded away from zero and infinity, $1/M_1 \le f(\mathbf{x}) \le M_2$. Let $\mathbf{X}$ denote a random vector with density f. Deviating from Stone for now, we denote by $I_{(1/K)}$ a cube of edge length $1/K$ within $[0, 1]^M$, and by $I_{(1/K)}(\mathbf{x})$ we denote its indicator function. By $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ we denote M-variable polynomials of degree less than or equal to m_1 in each variable. To shorten notation, we write $|E_n - E|(Y)$ for $|E_nY - EY|$, where E_n is the mean operator based on an i.i.d. sample of size n.

LEMMA 3.3′. *For $t > 0$, the inequalities*

$$\begin{aligned}&|E_n - E|\,[p_1(\mathbf{X})p_2(\mathbf{X})I_{(1/K)}(\mathbf{X})]\\ &\qquad \le c_{m_1}M_1M_2\left(E\,[p_1^2(\mathbf{X})I_{(1/K)}(\mathbf{X})]\,E\,[p_2^2(\mathbf{X})I_{(1/K)}(\mathbf{X})]\right)^{1/2}t \quad \forall\, p_1,p_2,\end{aligned}$$

hold, except on an event having probability at most

$$2(1+2m_1)^M \exp\left(-\frac{n}{K^M}\frac{M_2t^2}{2(1+\frac{1}{3}t)}\right).$$

PROOF. We need Bernstein's inequality in the following form: If $|Y| \le b$ and $\sigma^2 = \operatorname{Var}[Y]$, then, for $s > 0$,

$$|E_n - E|\;[Y] \le s$$

holds, except on an event having probability at most

$$2\,\exp\left(-n\frac{s^2/\sigma^2}{2\left(1+\frac{1}{3}bs/\sigma^2\right)}\right).$$

(See Stone's reference to Hoeffding. We use s rather than t in anticipation of a reparametrization.) We apply the inequality to

$$Y = (K\mathbf{X})^{j_1+j_2}I_{(1/K)}(\mathbf{X}),$$

where w.l.o.g. $I_{(1/K)} = [0, 1/K]^M$ (otherwise apply a simple shift to the cube and the monomials Y). Thus, a simple bound for all multiexponents $\mathbf{j}_1, \mathbf{j}_2$ is $b = 1$. We observe that the exception probability of the inequality is increasing in σ^2, hence it is conservative to replace σ^2 by a rough upper bound:

$$\sigma^2 \le E[Y^2] \le M_2/K^M.$$

Since Lemma 3.3′ is about polynomials of partial degrees less than or equal to m_1, we constrain $\mathbf{j}_1$ and $\mathbf{j}_2$ to $\{0, 1, \dots, m_1\}^M$. Note that $\mathbf{j}_1+\mathbf{j}_2 \in \{0, 1, \dots, 2m_1\}^M$. We want Bernstein's inequality to hold simultaneously for all $\mathbf{j}_1$ and $\mathbf{j}_2$, whence we inflate the exception probability conservatively:

$$(3) \qquad |E_n - E| \left[(K\mathbf{X})^{\mathbf{j}_1+\mathbf{j}_2} I_{(1/K)}(\mathbf{X})\right] \le s \quad \forall\, \mathbf{j}_1, \mathbf{j}_2 \in \{0, 1, \dots, m_1\}^M,$$

except on an event having probability at most

$$2(1 + 2m_1)^M \exp\left(-n \frac{s^2 K^M/M_2}{2\left(1 + \frac{1}{3} s K^M/M_2\right)}\right).$$

The extension from monomials to polynomials follows Stone's proof, with the exception of a scaling step to adjust for the size of the cube $I_{(1/K)}$: Let $q_1(\mathbf{x}) = \sum_{\mathbf{j}_1} a^{(1)}_{\mathbf{j}_1} \mathbf{x}^{\mathbf{j}_1}$ and $q_2(\mathbf{x}) = \sum_{\mathbf{j}_2} a^{(2)}_{\mathbf{j}_2} \mathbf{x}^{\mathbf{j}_2}$ be two polynomials of partial degrees less than or equal to m_1, that is, the sum is over $\mathbf{j}_1, \mathbf{j}_2 \in \{0,1,\dots, m_1\}^M$. On event (3) it follows that

$$|E_n - E| \left[q_1(K\mathbf{X})\, q_2(K\mathbf{X}) I_{(1/K)}(\mathbf{X})\right] \le \sum_{\mathbf{j}_1} \left|a^{(1)}_{\mathbf{j}_1}\right| \sum_{\mathbf{j}_2} \left|a^{(2)}_{\mathbf{j}_2}\right| s, \quad \forall\, q_1, q_2.$$

The desired bounds for the right-hand side are obtained as follows:

$$\begin{aligned} E\left[q_1^2(K\mathbf{X}) I_{(1/K)}(\mathbf{X})\right] &\ge \frac{1}{M_1} \int q_1^2(K\mathbf{x}) I_{(1/K)}(\mathbf{x})\, d\mathbf{x} \\ &= \frac{1}{M_1 K^M} \int q_1^2(\mathbf{x}) I_{(1)}(\mathbf{x})\, d\mathbf{x} \\ &\ge \frac{1}{M_1 K^M c_{m_1}} \left(\sum_{\mathbf{j}_1} \left|a^{(1)}_{\mathbf{j}_1}\right|\right)^2, \end{aligned}$$

where the middle equality is a change of scale from $I_{(1/K)}$ to the unit cube $I_{(1)} = [0, 1]^M$, and the second inequality makes use of the fact that any two norms on a finite-dimensional linear space are equivalent [see Stone's remark (3.5)]. The same holds of course for q_2, so on event (3) we get

$$\begin{aligned} &|E_n - E| \left[q_1(K\mathbf{X})\, q_2(K\mathbf{X})\, I_{(1/K)}(\mathbf{X})\right] \\ &\quad \le M_1 K^M c_{m_1} \left(E\left[q_1^2(K\mathbf{X})\, I_{(1/K)}(\mathbf{X})\right] E\left[q_2^2(K\mathbf{X})\, I_{(1/K)}(\mathbf{X})\right]\right)^{1/2} s \quad \forall\, q_1, q_2. \end{aligned}$$

By renaming $p_1(\mathbf{x}) = q_1(K\mathbf{x})$, $p_2(\mathbf{x}) = q_2(K\mathbf{x})$ and reparametrizing $t = sK^M/M_2$, Lemma 3.3′ follows. □

We turn to Lemma 3.4: Following Stone's notation, let $I_{\mathbf{k}}$ be the cube of edge length $1/K$ rooted at $((k_1 - 1)/K, (k_2 - 1)/K, \dots, (k_M - 1)/K)$ [where $\mathbf{k} = (k_1, k_2, \dots, k_M) \in \{1, 2, \dots, K\}^M$]. Denote by $g_1(\mathbf{x}) = \sum_{\mathbf{k}} p_{1\mathbf{k}}(\mathbf{x}) I_{\mathbf{k}}(\mathbf{x})$ and $g_2(\mathbf{x}) = \sum_{\mathbf{k}} p_{2\mathbf{k}}(\mathbf{x}) I_{\mathbf{k}}(\mathbf{x})$ functions that are polynomials of partial degrees less than or equal to m_1 on the cubes $I_{\mathbf{k}}$. The following is a slightly sharper version of Lemma 3.4, in the form referred to in Stone's proof.

LEMMA 3.4′. *For $t > 0$ the inequalities*

$$|E_n - E|[g_1(\mathbf{X}) g_2(\mathbf{X})] \\ \le c_{m_1} M_1 M_2 \left(E\left[g_1^2(\mathbf{X})\right] E\left[g_2^2(\mathbf{X})\right]\right)^{1/2} t \quad \forall g_1, g_2$$

hold, except on an event having probability at most

$$2(1 + 2m_1)^M K^M \exp\left(-\frac{n}{K^M} \frac{M_2 t^2}{2(1 + \frac{1}{3}t)}\right).$$

PROOF. The inequality of Lemma 3.3′ holds simultaneously on all cubes $I_{\mathbf{k}}$, except on an event having probability at most K^M times the exception probability of Lemma 3.3′. We then get the following bounds for all g_1 and g_2:

$$\begin{aligned} &|E_n - E|[g_1(\mathbf{X}) g_2(\mathbf{X})] \\ &\quad \le \sum_{\mathbf{k}} |E_n - E|[p_{1\mathbf{k}}(\mathbf{X}) p_{2\mathbf{k}}(\mathbf{X}) I_{\mathbf{k}}(\mathbf{X})] \\ &\quad \le c_{m_1} M_1 M_2 \sum_{\mathbf{k}} \left(E\left[p_{1\mathbf{k}}^2(\mathbf{X}) I_{\mathbf{k}}(\mathbf{X})\right] E\left[p_{2\mathbf{k}}^2(\mathbf{X}) I_{\mathbf{k}}(\mathbf{X})\right]\right)^{1/2} t \\ &\quad \le c_{m_1} M_1 M_2 \left(\sum_{\mathbf{k}} E\left[p_{1\mathbf{k}}^2(\mathbf{X}) I_{\mathbf{k}}(\mathbf{X})\right] \sum_{\mathbf{k}} E\left[p_{2\mathbf{k}}^2(\mathbf{X}) I_{\mathbf{k}}(\mathbf{X})\right]\right)^{1/2} t \\ &\quad = c_{m_1} M_1 M_2 \left(E\left[g_1^2(\mathbf{X})\right] E\left[g_2^2(\mathbf{X})\right]\right)^{1/2} t. \end{aligned}$$

The second inequality uses Lemma 3.3′, and the third inequality is an application of Cauchy–Schwarz in the form $\sum y^{1/2} z^{1/2} \le (\sum y \sum z)^{1/2}$. □

From Lemma 3.4′, we see that in order to let the exception probability go to zero, the growth of $K = K_n$ must be limited by the condition

$$C \frac{n}{K^M} - \log K^M \longrightarrow \infty, \qquad (n \longrightarrow \infty) \quad \forall C > 0. \tag{4}$$

Stone's assumption $K^M = o(n^{1-\delta})$ is certainly sufficient for (4).

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 ~ a * (b + c) expands to ~ 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,