The Annals of Statistics
1997, Vol. 25, No. 4, 1371–1470

1994 WALD MEMORIAL LECTURE

POLYNOMIAL SPLINES AND THEIR TENSOR PRODUCTS IN EXTENDED LINEAR MODELING

By Charles J. Stone,[1] Mark H. Hansen, Charles Kooperberg[2] and Young K. Truong[3]

University of California, Berkeley, Bell Laboratories, University of Washington and University of North Carolina, Chapel Hill

Analysis of variance type models are considered for a regression function or for the logarithm of a probability function, conditional probability function, density function, conditional density function, hazard function, conditional hazard function or spectral density function. Polynomial splines are used to model the main effects, and their tensor products are used to model any interaction components that are included. In the special context of survival analysis, the baseline hazard function is modeled and nonproportionality is allowed. In general, the theory involves the L_2 rate of convergence for the fitted model and its components. The methodology involves least squares and maximum likelihood estimation, stepwise addition of basis functions using Rao statistics, stepwise deletion using Wald statistics and model selection using the Bayesian information criterion, cross-validation or an independent test set. Publicly available software, written in C and interfaced to S/S-PLUS, is used to apply this methodology to real data.

1. Introduction. The last two decades have witnessed an incredible change in the focus of statistical theory and methodology. Fueled in part by the explosion of available computer power, highly adaptive, functional procedures are now essential tools for modern data analysis. While freed from the rigid assumptions implicit in classical parametric models, the statistician is now expected to select not only the important variables in a model, but also the functional form of the dependence on these variables. To be practically successful, any new adaptive procedure must inevitably strike a balance between flexibility and the haunting "curse of dimensionality." It is in this capacity that statistical theory is critical to the success of emerging methodologies. Polynomial splines and their tensor products offer the flexibility required for modern data analysis, and when used in concert with low-dimensional analysis of variance (ANOVA) decompositions, effectively tame the curse of dimensionality.

Received August 1995; revised December 1996.

[1] Research supported in part by NSF Grants DMS-92-04247 and DMS-95-04463.

[2] Research supported in part by NSF Grant DMS-94-03371 and NIH Grant CA61937.

[3] Research supported in part by NSF Grant DMS-94-03800 and NIH Grant CA61937.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62J12.

Key words and phrases. ANOVA, density estimation, generalized additive models, generalized linear models, least squares, logistic regression, maximum likelihood, model selection, multiple classification, nonparametric regression, optimal rates of convergence, proportional hazards model, spectral estimation, survival analysis.

In the pages that follow, we will alternate between a discussion of the practical implementation of this methodology and a very broad theoretical investigation into the properties of this approach in the context of *extended linear models*. We have coined this term because our theoretical results apply to a group of estimation problems that subsumes the classical exponential family regression models [see McCullagh and Nelder (1989)]. While our initial motivation for introducing this family was to achieve a theoretical synthesis, we found that this framework also allows us to entertain a fairly general treatment of the associated methodology. Throughout our presentation, however, we maintain a distinction between the nonadaptive procedures that we can treat theoretically and the adaptive methodologies that we have implemented for density estimation, hazard regression, polychotomous regression and spectral density estimation. In this presentation, we concentrate on theoretical and methodological innovations developed through many collaborations involving various subsets of the authors of the present paper.

In Section 2, we define the notion of an extended linear model and use this framework simultaneously to discuss the L_2 rate of convergence for the nonadaptive version of our procedures in a variety of important statistical settings, while in Section 3, we translate these promising theoretical results into practically useful, adaptive methodology. Ultimately, however, the true measure of any statistical procedure is its performance on real data. In Sections 4–9 we focus on a number of specific modeling problems for which our approach has yielded successful data analysis tools. In each case, an S/S-PLUS implementation is (or will soon be made) publicly available so that the "true measure" of these procedures can be judged on the wealth of data that exist beyond the (necessarily narrow) confines of our examples. Logspline density estimation was our first attempt at an adaptive spline-based methodology, and in Section 4 we present the latest version of this procedure, LOGSPLINE. In Section 5 we describe our own version of MARS [Friedman (1991)] as a routine to handle regression problems involving many predictors. The motivation for reworking this routine stems from an application of linear splines to polychotomous regression, known as POLYCLASS, which is described in Section 6. In order to relax the proportionality and linearity assumptions in classical survival analysis, we have developed spline routines for hazard estimation with flexible tails (HEFT) and hazard regression (HARE). These are the subject of Section 7. Spectral density estimation is another area in which our adaptive methodology can easily capture all the relevant features of a given time series, and in Section 8 we discuss LSPEC, an implementation of this approach. We end the paper with a discussion of Triogram models, a methodology for bivariate function estimation through the use of splines defined over adaptively determined triangulations.

2. Extended linear models: theory.

Notation. Consider a $\mathscr{W}$-valued random variable $\mathbf{W}$, where $\mathscr{W}$ is an arbitrary set. Let $\mathscr{U} = \mathscr{U}_1 \times \cdots \times \mathscr{U}_M$ be a Cartesian product of compact intervals, each having positive length. Let K be a positive integer. Consider a vector-valued function $h = (h_1, \dots, h_K)$ on $\mathscr{U}$ whose *constituents* $h_1, \dots, h_K$ are real-valued functions on $\mathscr{U}$. Let $l(h, \mathbf{W})$ be a (not necessarily true) log-likelihood and let $\Lambda(h) = E[l(h, \mathbf{W})]$ be the corresponding expected log-likelihood. There may be some mild restrictions on h for the log-likelihood to be defined. We assume that, subject to such restrictions, there is an essentially unique function $\phi = (\phi_1, \dots, \phi_K)$ that maximizes the expected log-likelihood. (Here two functions on $\mathscr{U}$ are essentially equal if they differ only on a subset of $\mathscr{U}$ having Lebesgue measure zero.)

Let H be a linear space of real-valued functions on $\mathscr{U}$, let H^K denote the space of functions of the form $h = (h_1, \dots, h_K)$, where the constituents $h_1, \dots, h_K$ of h range over H, and consider the log-likelihood function $l(h, \mathbf{W})$, $h \in H^K$. We refer to any particular setup of this form as an *extended linear model*. The expected log-likelihood function is given by $\Lambda(h)$, $h \in H^K$. The model is said to be concave if $l(h, \mathbf{w})$ is a concave function of h for each $\mathbf{w} \in \mathscr{W}$ and $\Lambda(h)$ is a strictly concave function of h when restricted to those functions $h \in H^K$ such that $\Lambda(h) > -\infty$. Typically, when the model is concave, there is an essentially unique function $\phi^* = (\phi_1^*, \dots, \phi_K^*) \in H^K$ that maximizes the expected log-likelihood over H^K. It follows from the information inequality that if $\phi \in H^K$, then $\phi^* = \phi$.

In order to define ANOVA decompositions of the constituents of ϕ^*, we first need to define corresponding theoretical inner products and norms. To this end, let ψ be an absolutely continuous measure on $\mathscr{U}$ having a density function that is bounded away from zero and infinity on $\mathscr{U}$. Given square-integrable, real-valued functions h_1 and h_2 on $\mathscr{U}$, their theoretical inner product is defined by $\langle h_1, h_2 \rangle = \int_{\mathscr{U}} h_1 h_2 \, d\psi$. Given such a function h, its theoretical norm is defined by $\|h\|^2 = \langle h, h \rangle = \int_{\mathscr{U}} h^2 \, d\psi$. Conversely, if $\| \cdot \|$ is defined directly, then ψ is defined implicitly by the formula $\psi(A) = \|\mathrm{ind}_A\|^2$, where ind_A is the indicator function of A, which equals 1 on A and 0 on A^c.

Let $\mathbf{W}_1, \dots, \mathbf{W}_n$ be a random sample of size n from the distribution of $\mathbf{W}$. The log-likelihood function corresponding to this random sample is given by $l(h) = \sum_i l(h, \mathbf{W}_i)$. Let $G = G_n$ be a finite-dimensional subspace of H and let $G^K = G_n^K$ denote the corresponding subspace of H^K. (Note that if $K = 1$, then $H^K = H$ and $G^K = G$.) Under the assumptions of a concave extended linear model and reasonable additional conditions, except on an event whose probability tends to zero as $n \to \infty$, there is a unique maximum likelihood estimate $\widehat{\phi}$ in G^K of ϕ^*; that is, a unique function $\widehat{\phi} = (\widehat{\phi}_1, \dots, \widehat{\phi}_K)$ in G^K that maximizes the log-likelihood function over G^K.

In order to define ANOVA decompositions of the constituents of $\widehat{\phi}$, we need to define corresponding empirical inner products and norms. For $n \geq 1$, let ψ_n be an empirical product measure on $\mathscr{U}$ that is a transform (measurable function) of the random sample $\mathbf{W}_1, \dots, \mathbf{W}_n$. (Roughly speaking, ψ_n should approach ψ as $n \to \infty$.) Given real-valued functions h_1 and h_2 on $\mathscr{U}$, their empirical inner product is defined by $\langle h_1, h_2 \rangle_n = \int_{\mathscr{U}} h_1 h_2 \, d\psi_n$. Given such a function h, its empirical norm is defined by $\|h\|_n^2 = \int_{\mathscr{U}} h^2 \, d\psi_n$. The space G is said to be *identifiable* if the only function $g \in G$ such that $\|g\|_n = 0$ is given by $g = 0$. Under reasonable conditions, G is identifiable except on an event whose probability tends to zero as $n \to \infty$.

Many statistical problems of theoretical and practical importance can effectively be treated within the framework of concave extended linear models. Most of the investigations in this framework have involved a $\mathscr{U}$-valued random variable $\mathbf{U}$ that is a transform of $\mathbf{W}$. Let $\mathbf{U}_1, \dots, \mathbf{U}_n$ be the corresponding transforms of $\mathbf{W}_1, \dots, \mathbf{W}_n$, respectively. Here, we typically let ψ be the distribution of $\mathbf{U}$ and ψ_n the empirical distribution of $\mathbf{U}_1, \dots, \mathbf{U}_n$.

EXAMPLES.

Regression. Consider a random pair $(\mathbf{X}, Y)$, where $\mathbf{X}$ is $\mathscr{X}$-valued and Y is real-valued and has finite second moment. Set $l(h, \mathbf{X}, Y) = -[Y - h(\mathbf{X})]^2$. Then we get a concave extended linear model with $\mathbf{W} = (\mathbf{X}, Y)$, $\mathbf{U} = \mathbf{X}$ and $K = 1$. If H is the space of all functions h on $\mathscr{X}$ with $E[h^2(\mathbf{X})] < \infty$, then ϕ is the regression function of Y on $\mathbf{X}$. More generally, if H is a Hilbert space of such functions h, then ϕ^* is the best approximation in H to the regression function, where "best" means minimizing the mean squared error $E\{[Y - h(\mathbf{X})]^2\}$ in predicting Y by $h(\mathbf{X})$. Here maximum likelihood estimation in G coincides with least squares estimation.

Generalized regression. Suppose now that, for each $\mathbf{x} \in \mathscr{X}$, the conditional distribution of Y given that $\mathbf{X} = \mathbf{x}$ belongs to a fixed exponential family of distributions on $\mathbb{R}$ of the form $\exp[B(\theta)y - C(\theta)]\rho(dy)$, where the parameter θ ranges over $\mathbb{R}$. Here ρ is a nonzero measure on $\mathbb{R}$ that is not concentrated at a single point and $\int_{\mathbb{R}} \exp[B(\theta)y - C(\theta)]\rho(dy) = 1$ for $\theta \in \mathbb{R}$. The function $B(\cdot)$ is required to be twice continuously differentiable and its first derivative $B'(\cdot)$ is required to be strictly positive on $\mathbb{R}$. It is required that there be a subinterval S of $\mathbb{R}$ such that ρ is concentrated on S and $B''(\theta)y - C''(\theta) < 0$ for $\theta \in \mathbb{R}$ and $y \in S$. If S is bounded, it is required that it contain at least one of its endpoints. Let h be a candidate for the dependence of θ on $\mathbf{x}$. The corresponding (conditional) log-likelihood is given by $l(h, \mathbf{X}, Y) = B(h(\mathbf{X}))Y - C(h(\mathbf{X}))$. This has the form of a concave extended linear model with $\mathbf{W} = (\mathbf{X}, Y)$, $\mathbf{U} = \mathbf{X}$ and $K = 1$. As special cases, we get logistic regression, probit regression and Poisson regression models.

Polychotomous regression. Let Y be a qualitative random variable having $K + 1$ possible values. Without loss of generality, we can think of this random variable as ranging over $\mathscr{Y} = \{1, \dots, K+1\}$. Suppose that $P(Y = k|\mathbf{X} = \mathbf{x}) > 0$ for $\mathbf{x} \in \mathscr{X}$ and $k \in \mathscr{Y}$. For $1 \le k \le K$, let h_k be a candidate for the function

$$\log \frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y = K + 1|\mathbf{X} = \mathbf{x})}.$$

The corresponding log-likelihood is given by

$$\begin{aligned} l(h, \mathbf{X}, Y) = {} & h_1(\mathbf{X})I_1(Y) + \cdots + h_K(\mathbf{x})I_K(Y) \\ & - \log(1 + \exp h_1(\mathbf{X}) + \cdots + \exp h_K(\mathbf{X})), \end{aligned}$$

where $I_k(Y)$ equals 1 or 0 according as $Y = k$ or $Y \neq k$ and $h = (h_1, \ldots, h_K)$. This setup has the form of a concave extended linear model with $\mathbf{W} = (\mathbf{X}, Y)$ and $\mathbf{U} = \mathbf{X}$.

Density estimation. Let $\mathbf{Y}$ have an unknown positive density function on $\mathscr{Y}$. We can write its log-density function in the form $\phi - C(\phi)$, where $C(h) = \log \int \exp h(\mathbf{y})\, d\mathbf{y}$. The corresponding log-likelihood function is given by $l(h, \mathbf{Y}) = h(\mathbf{Y}) - C(h)$. This setup has the form of a concave extended linear model with $\mathbf{W} = \mathbf{U} = \mathbf{Y}$ and $K = 1$ provided that, for identifiability, we impose a restriction on the functions $h \in H$ such as $E[h(\mathbf{U}) = 0]$ and we impose a similar condition on the functions in G.

Hazard regression. Consider a positive survival time T, a positive censoring time C, the observed time $\min(T, C)$ and an $\mathscr{X}$-valued random vector $\mathbf{X}$ of covariates. Let $\delta = \operatorname{ind}(T \leq C)$ be the indicator random variable that equals 1 or 0 according as $T \leq C$ (T is uncensored) or $T > C$ (T is censored) and write $\min(T, C)$ as $T \wedge C$. Suppose T and C are conditionally independent given $\mathbf{X}$. For theoretical purposes, it is supposed that $P(C \leq \tau) = 1$, where τ is a known positive constant. Set $\mathbf{W} = (\mathbf{X}, T \wedge C, \delta)$ and $\mathbf{U} = (\mathbf{X}, T \wedge C)$. Let $\phi(\mathbf{x}, t) = \log f(t|\mathbf{x})/[1 - F(t|\mathbf{x})]$, $t > 0$, denote the logarithm of the conditional hazard function, where $f(t|\mathbf{x})$ and $F(t|\mathbf{x})$ are the conditional density and distribution functions, respectively, of T given that $\mathbf{X} = \mathbf{x}$. Since the likelihood equals $f(T \wedge C|\mathbf{X})$ for an uncensored case and $1 - F(T \wedge C|\mathbf{X})$ for a censored case, it can be written as

$$\begin{aligned}
&[f(T \wedge C|\mathbf{X})]^{\delta}[1 - F(T \wedge C|\mathbf{X})]^{1-\delta} \\
&\quad = \left(\frac{f(T \wedge C|\mathbf{X})}{1 - F(T \wedge C|\mathbf{X})}\right)^{\delta}[1 - F(T \wedge C|\mathbf{X})] \\
&\quad = [\exp\, \phi(\mathbf{X}, T \wedge C)]^{\delta} \exp\left(-\int_0^{T \wedge C} \exp\, \phi(\mathbf{X}, t)\, dt\right).
\end{aligned}$$

Thus the log-likelihood function is given by

$$l(h, \mathbf{W}) = \delta h(\mathbf{X}, T \wedge C) - \int_0^{T \wedge C} \exp\, h(\mathbf{X}, t)\, dt.$$

This setup has the form of a concave extended linear model with $K = 1$. Here the theoretical inner product is given by

$$\langle h_1, h_2 \rangle = E \int_0^{T \wedge C} h_1(t, \mathbf{X}) h_2(t, \mathbf{X})\, dt,$$

which defines ψ implicitly; the corresponding empirical inner product $\langle \cdot, \cdot \rangle_n$ and empirical measure ψ_n are defined in the obvious manner.

ANOVA decompositions and convergence rates. In the theoretical development of extended linear models, ANOVA decompositions of ϕ^*, $\widehat{\phi}$, and their constituents play important roles. For a simple illustration of such decompositions, consider a regression or generalized regression context with $M = 2$

and let H be the space of all square-integrable functions on $\mathscr{U}$. Then ϕ can be written as

$$\phi(x_1, x_2) = \phi_0 + \phi_1(x_1) + \phi_2(x_2) + \phi_{12}(x_1, x_2). \tag{2.1}$$

Here ϕ_0 is the constant component, ϕ_1 and ϕ_2 are the main effect components and ϕ_{12} is the two-factor interaction component. It is required that each component be theoretically orthogonal to all choices of the corresponding lower-order components; that is, ϕ_1, ϕ_2 and ϕ_{12} are each theoretically orthogonal to 1, and ϕ_{12} is orthogonal to all choices of ϕ_1 and ϕ_2. The maximum number d of factors in any component of the model is given by $d = 2$. Since $d = M$, the model is saturated.

Given a random sample, consider an estimate

$$\widehat{\phi}(x_1, x_2) = \widehat{\phi}_0 + \widehat{\phi}_1(x_1) + \widehat{\phi}_2(x_2) + \widehat{\phi}_{12}(x_1, x_2), \tag{2.2}$$

where each component is empirically orthogonal to all choices of the corresponding lower-order components. The right-hand sides of (2.1) and (2.2) are referred to as the ANOVA decompositions of ϕ and $\widehat{\phi}$, respectively.

Removing the interaction component, we get the additive ($d = 1$), unsaturated approximation

$$\phi^*(x_1, x_2) = \phi_0^* + \phi_1^*(x_1) + \phi_2^*(x_2)$$

to ϕ and the corresponding estimate

$$\widehat{\phi}(x_1, x_2) = \widehat{\phi}_0 + \widehat{\phi}_1(x_1) + \widehat{\phi}_2(x_2).$$

In general, given a subset s of $\{1, \dots, M\}$, let H_s denote the space of square-integrable, real-valued functions on $\mathscr{U}$ that depend only on the variables u_m, $m \in s$. (The space $H_\varnothing$ corresponding to the empty set $\varnothing$ is the space of constant functions.) Let $\mathscr{S}$ denote a hierarchical collection of subsets of $\{1, \dots, M\}$, where hierarchical means that if s is a member of $\mathscr{S}$ and r is a subset of s, then r is a member of $\mathscr{S}$. Let H now denote the space of functions on $\mathscr{U}$ of the form $\sum_{s\in\mathscr{S}} h_s$, where $h_s \in H_s$ for $s \in \mathscr{S}$. Let d denote the maximum cardinality of the sets $s \in \mathscr{S}$. We refer to this setup as being *saturated* if $d = M$ and *unsaturated* if $d < M$. If $d = 1$, then the functions in H are additive functions of the individual coordinates.

Let $h \perp H_r$ mean that $\langle h, h_r\rangle = 0$ for $h_r \in H_r$. Every function $h \in H$ can then be written in an essentially unique manner as $h = \sum_{s\in\mathscr{S}} h_s$, where, for $s \in \mathscr{S}$, $h_s \in H_s$ and $h_s \perp H_r$ for every proper subset r of s. We refer to h_s, $s \in \mathscr{S}$, as the *components* of the ANOVA decomposition of h. In particular, let ϕ_{ks}^*, $s \in \mathscr{S}$, denote the components of the ANOVA decomposition of ϕ_k^*. Also, set $\phi_s^* = (\phi_{1s}^*, \dots, \phi_{Ks}^*)$ for $s \in \mathscr{S}$.

For $1 \le m \le M$, let G_m denote a finite-dimensional space of functions on $\mathscr{U}_m$ containing the constant functions. Given a subset s of $\{1, \dots, M\}$, let G_s denote the tensor product of the spaces G_m, $m \in s$, which is the space spanned by functions on $\mathscr{U}$ of the form $\prod_{m\in s} g_m(u_m)$ as g_m ranges over G_m for $m \in s$. Observe that $G_r \subset G_s$ for $r \subset s$. Let G denote the space of functions on $\mathscr{U}$ of the form $\sum_{s\in\mathscr{S}} g_s$, where $g_s \in G_s$ for $s \in \mathscr{S}$.

Let $g \perp_n G_r$ mean that $\langle g, g_r \rangle_n = 0$ for $g_r \in G_r$. If G is identifiable, then every function $g \in G$ can be written uniquely as $g = \sum_{s \in \mathscr{S}} g_s$, where, for $s \in \mathscr{S}$, $g_s \in G_s$ and $g_s \perp_n G_r$ for every proper subset r of s. We refer to g_s, $s \in \mathscr{S}$, as the components of the ANOVA decomposition of g. In particular, let $\widehat{\phi}_{ks}$, $s \in \mathscr{S}$, denote the components of the ANOVA decomposition of $\widehat{\phi}_k$. Also, set $\widehat{\phi}_s = (\widehat{\phi}_{1s}, \dots, \widehat{\phi}_{Ks})$ for $s \in \mathscr{S}$.

We now restrict attention to spaces G_m of polynomial splines. For theoretical simplicity, for $1 \le m \le M$, let Δ_m be a partition of $\mathscr{U}_m$ into disjoint intervals having common length a. By a piecewise polynomial of degree q on $\mathscr{U}_m$, we mean a function g on $\mathscr{U}_m$ such that the restriction of g to each $\delta \in \Delta_m$ is a polynomial of degree q. Let G_m be a linear space of splines on $\mathscr{U}_m$—that is, piecewise polynomials of degree q on $\mathscr{U}_m$ subject to specified smoothness constraints, typically that of being $(q-1)$-times continuously differentiable on $\mathscr{U}_m$.

Given a real-valued function h on $\mathscr{U}$, let $\|h\|_\infty$ denote the supremum of $|h|$ on $\mathscr{U}$. Given a vector-valued function $h = (h_1, \dots, h_K)$ on $\mathscr{U}$, set $\|h\|_\infty = \max(\|h_1\|_\infty, \dots, \|h_K\|_\infty)$ and $\|h\|^2 = \|h_1\|^2 + \cdots + \|h_K\|^2$.

Next we consider the rates of convergence that can theoretically be established for the estimate $\widehat{\phi}$ of ϕ^* and for the corresponding estimates $\widehat{\phi}_s$ of the components ϕ_s^* of ϕ^*. Let $s \in \mathscr{S}$. Under various conditions on the spaces G_m, $m \in s$,

$$\inf_{g \in G_s} \|g - \phi_{ks}^*\|_\infty = O(a^p), \qquad 1 \le k \le K \text{ and } s \in \mathscr{S},$$

with p being a suitably defined measure of smoothness of the constituents of ϕ^*. Under various reasonable additional conditions,

$$\|\widehat{\phi}_s - \phi_s^*\|^2 = O_P\left(a^{2p} + \frac{1}{na^d}\right), \qquad s \in \mathscr{S},$$

and

$$\|\widehat{\phi} - \phi^*\|^2 = O_P\left(a^{2p} + \frac{1}{na^d}\right).$$

Thus, by optimally choosing $a \sim n^{-1/(2p+d)}$, we get the rate of convergence given by

$$\|\widehat{\phi}_s - \phi_s^*\| = O_P(n^{-p/(2p+d)}), \qquad s \in \mathscr{S}, \tag{2.3}$$

and

$$\|\widehat{\phi} - \phi^*\| = O_P(n^{-p/(2p+d)}). \tag{2.4}$$

In particular, by considering additive models ($d = 1$) or by allowing interactions involving only two factors ($d = 2$), we can get faster rates of convergence than by choosing $d = M$ and thereby ameliorate the "curse of dimensionality."

Hansen (1994) introduced the class of extended linear models and obtained the corresponding L_2 rates of convergence. The various cases of this theory that have previously been treated are as follows: regression in Stone (1985, 1994); generalized regression in Stone (1986, 1994), density estimation in Stone (1990, 1994); conditional density estimation in Stone (1991, 1994) and Hansen (1994); hazard regression in Kooperberg, Stone and Truong (1995b); and spectral density estimation in Kooperberg, Stone and Truong (1995d).

3. Extended linear models: adaptive methodology. In practice, it seems best to select G in an adaptive manner. Let J be the dimension of G, let $B_1, \dots, B_J$ be a basis of this space and write a candidate $g = (g_1, \dots, g_K)$ for the maximum likelihood estimate $\widehat{\phi}$ in G of ϕ^* as $g_k = \sum_j \beta_{jk} B_j$ for $1 \le k \le K$. Let $\boldsymbol{\beta}$ be the (suitably) ordered JK-tuple $(\beta_{jk})_{1 \le j \le J, 1 \le k \le K}$. Then the log-likelihood function based on the sample data can be written as $l(\boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathscr{B}$. Assume that this log-likelihood function is twice continuously differentiable, and let $\nabla l(\boldsymbol{\beta})$ and $\mathbf{H}(\boldsymbol{\beta})$ denote its gradient and Hessian matrix, respectively, at $\boldsymbol{\beta}$.

The quadratic approximation Q to the log-likelihood function about $\boldsymbol{\beta}_0 \in \mathscr{B}$ is given by

$$Q(\boldsymbol{\beta}) = l(\boldsymbol{\beta}_0) + [\nabla l(\boldsymbol{\beta}_0)]^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \tfrac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \mathbf{H}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0). \tag{3.1}$$

Suppose $\mathbf{H}(\boldsymbol{\beta}_0)$ is negative definite or, equivalently, that $\mathbf{I}(\boldsymbol{\beta}_0) = -\mathbf{H}(\boldsymbol{\beta}_0)$ is positive definite. Then Q is uniquely maximized at

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + [\mathbf{I}(\boldsymbol{\beta}_0)]^{-1} \nabla l(\boldsymbol{\beta}_0). \tag{3.2}$$

Using (3.2) in an iterative manner, we get the Newton–Raphson method for numerically determining the maximum likelihood estimate from any starting value $\boldsymbol{\beta}_0$. If the maximum likelihood estimate exists, the log-likelihood function is strictly concave, and we apply a suitable modification to the Newton–Raphson method (such as step-halving), then the method is guaranteed to converge to the maximum likelihood estimate from any starting value [see Kooperberg, Bose and Stone (1997) for details]. It follows from (3.1) and (3.2) that

$$2[Q(\boldsymbol{\beta}_1) - Q(\boldsymbol{\beta}_0)] = [\nabla l(\boldsymbol{\beta}_0)]^T [\mathbf{I}(\boldsymbol{\beta}_0)]^{-1} \nabla l(\boldsymbol{\beta}_0). \tag{3.3}$$

If $\boldsymbol{\beta}_0$ is the maximum likelihood estimate in a subspace of $\mathscr{B}$, then the right-hand side of (3.3) is the Rao (score) statistic for testing the hypothesis that the "true" value of $\boldsymbol{\beta}$ lies in this subspace.

Let Q now be the quadratic approximation to the log-likelihood function about the maximum likelihood estimate $\widehat{\boldsymbol{\beta}} \in \mathscr{B}$, and let $\mathscr{B}_0$ be the subspace of $\mathscr{B}$ consisting of all $\boldsymbol{\beta} \in \mathscr{B}$ such that $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{A}$ has full rank. Then the maximum of Q over $\mathscr{B}_0$ occurs uniquely at

$$\widehat{\boldsymbol{\beta}}_0 = \widehat{\boldsymbol{\beta}} - \mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{A}^T[\mathbf{A}\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{A}^T]^{-1}\mathbf{A}\widehat{\boldsymbol{\beta}}. \tag{3.4}$$

Moreover,

$$2[Q(\widehat{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}}_0)] = (\mathbf{A}\widehat{\boldsymbol{\beta}})^T[\mathbf{A}\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{A}^T]^{-1}\mathbf{A}\widehat{\boldsymbol{\beta}}. \tag{3.5}$$

The right-hand side of (3.5) is the Wald statistic for testing the hypothesis that $\boldsymbol{\beta} \in \mathscr{B}_0$ under the assumption that $\boldsymbol{\beta} \in \mathscr{B}$. Moreover, the right-hand side of (3.4) gives a good starting value for using the Newton–Raphson method to find the maximum likelihood estimate in $\mathscr{B}_0$ when the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ in $\mathscr{B}$ has already been determined.

An important aspect of the methodology for fitting extended linear models is the adaptive choice of the space G from a family $\mathscr{G}$ of allowable spaces that is typically assumed to satisfy the following properties:

1. For each $G \in \mathscr{G}$, the model has dimension $J \geq J_{\min}$.
2. There is only one $G \in \mathscr{G}$ with dimension $J_{\min}$, which we refer to as the minimum allowable space.
3. If $G_0 \in \mathscr{G}$ has dimension J, there is at least one space $G \in \mathscr{G}$ with dimension $J + 1$ that contains G_0 as a subspace.
4. If $G \in \mathscr{G}$ has dimension $J > J_{\min}$, there is at least one subspace $G_0 \in \mathscr{G}$ of G with dimension $J - 1$.

In our univariate methodologies (LOGSPLINE, LSPEC and HEFT) we use families of allowable spaces based on cubic splines. For each of these methodologies there are some extra restrictions on the allowable spaces, which are discussed in the relevant sections. Also, the HEFT and LSPEC methodologies involve some additional basis functions that are not cubic splines. Details are given in Sections 7 and 8.

For the multivariate methodologies POLYMARS (our version of MARS), POLYCLASS and HARE we make use of piecewise linear splines and selected tensor products. These spaces are discussed in detail in Section 5 about POLYMARS. In all of these applications we restrict attention to $d \leq 2$, so that main effects (polynomial splines in individual variables) and two-factor interactions (tensor products of polynomial splines in two different variables) may be allowed, but no three-factor or higher-order interactions are allowed in the model. The allowable spaces for the bivariate splines considered in Section 9 are discussed in that section.

Initially, we choose G as the minimum allowable space. Then we proceed with stepwise addition. Here we successively replace the $(J - 1)$-dimensional allowable space G_0 by a J-dimensional allowable space G containing G_0 as a subspace, choosing among the various candidates for a new basis function by a heuristic search that is designed approximately to maximize the corresponding Rao statistic. The reason for using Rao statistics here is to avoid the need for computing maximum likelihood estimates corresponding to the various candidate spaces G.

Upon stopping the stepwise addition process (for example, after we reach a default or user-specified maximum dimension), we carry out stepwise deletion. Here we successively replace the J-dimensional allowable space G by a $(J - 1)$-dimensional allowable subspace G_0 until we arrive at the minimal allowable space, at each step choosing the candidate space G_0 so that the Wald statistic for a basis function that is in G but not in G_0 is smallest in magnitude. The reason for using Wald statistics here is to avoid the need for computing maximum likelihood estimates corresponding to the various candidate subspaces G_0.

During the combination of stepwise addition and stepwise deletion, we get a sequence of models indexed by ν, with the νth model having $J_\nu K$ parameters. The (generalized) Akaike information criterion (AIC) can be used to select one model from this sequence. Let $\widehat{l}_\nu$ denote the fitted log-likelihood for the νth model and let

$$\mathrm{AIC}_{a,\nu} = -2\widehat{l}_\nu + aJ_\nu K \tag{3.6}$$

be the Akaike information criterion with penalty parameter a for this model. We select the model corresponding to the value $\widehat{\nu}$ of ν that minimizes $\mathrm{AIC}_{a,\nu}$. In light of practical experience, we generally recommend choosing $a = \log n$ as in the Bayesian information criterion (BIC) due to Schwarz (1978). (Choosing $a = 2$ as in classical AIC tends to yield models that are unnecessarily complex, have spurious features and do not predict well on test data.)

Alternatively, we can use an independent test set to obtain a more nearly unbiased estimate of the expected log-likelihood and select the model that maximizes this estimate. In the regression and classification contexts we could use the independent test set to obtain a nearly unbiased estimate of the mean squared error of prediction or the cost of misclassification and select the model that minimizes this estimate.

Finally, cross-validation can be used to select a so as approximately to maximize the expected log-likelihood or minimize the expected mean squared error of prediction or cost of misclassification. [For detailed discussions of the use of independent test sets or cross-validation in the related context of selecting classification and regression trees, see Breiman, Friedman, Olshen and Stone (1984).]

Regardless of the final criteria used to choose between competing estimates, it is likely that many of the models encountered during the stepwise addition and deletion processes will perform similarly. By examining which terms are present in these best fitting models, we can gain considerable insight into the underlying features of the data. Simulation can also be used to judge whether or not our procedures can reliably resolve important aspects of a given data set. In addition, simulation can be used to calibrate the choice of (the implicit smoothing parameter) a in the AIC criterion of (3.6). Illustrations of these procedures will be given in the context of the various adaptive methodologies presented in Sections 4–9.

As mentioned in Section 1, various adaptive methodologies and corresponding software products have already been developed. The current situation regarding software availability is as follows:

1. Versions of the HARE, HEFT, LOGSPLINE and LSPEC methodologies are available from statlib. (The publicly available version of the LOGSPLINE program is slightly older than the one discussed in Section 4; see that section for more discussion.) All these methodologies are written as C programs with an interface to the S/S-PLUS environment.

2. A commercial version of HARE is currently being implemented in S-PLUS.
3. Friedman's MARS program is available as a collection of Fortran subroutines from statlib.
4. The POLYMARS program discussed in Section 5 was not written as a standalone program.
5. The current version of POLYCLASS is available from Kooperberg. We are working on a modification to this methodology to make it computationally much less intensive when applied to huge data sets with many classes, features and cases. In this modification we plan to use a stochastic gradient method to obtain the maximum likelihood fit to the largest model selected by POLYMARS.
6. A library of S/S-PLUS routines for manipulating Triogram models is currently available from Hansen and will soon be available in version 4 of S.

Our eventual goal is to develop a comprehensive set of polynomial spline modeling routines.

4. Univariate density estimation (LOGSPLINE). In logspline density estimation a (univariate) log-density is modeled by a cubic spline. The LOGSPLINE project was the first methodology project employing model selection and polynomial splines on which we have worked. In this section we describe the fourth version of LOGSPLINE. Earlier versions are discussed in Stone and Koo (1986b) and Kooperberg and Stone (1991, 1992). The various versions of LOGSPLINE all employ cubic splines and maximum likelihood estimation. The way that the program positions knots, how it deals with the tails of the distribution and what types of data it can handle are among the things that have evolved over time. Before presenting any details about the LOGSPLINE methodology, we give a brief example.

In the left side of Figure 1 we show a density estimate based on a random sample of 7125 annual net incomes in the United Kingdom [Family Expenditure Survey (1968–1983)]. [The data have been rescaled to have mean 1 as in Wand, Marron and Ruppert (1991).] The spike near 0.24 is due to the UK national old age pension, which caused many people to have nearly identical incomes. The right side of Figure 1 zooms in on the neighborhood of this spike. In Kooperberg and Stone (1992) we concluded that the height and location of this spike are accurately estimated by LOGSPLINE.

The selection of knots in logspline density estimation is discussed in detail below. Here it suffices to note that the procedure involves stepwise addition and deletion of knots. The program starts with a fairly small number of knots. In Figure 1 these knots are indicated by the letter "s". It then adds knots in those regions where an added knot would have the most influence, using Rao statistics. The program continues adding until a prespecified maximum number of knots is reached. The knots for this largest model are indicated by the letter "m" in Figure 1. After the largest model has been fitted, knots are deleted one at a time, using Wald statistics to decide which one to delete next. The smallest model that is fitted has three knots. Out of the complete

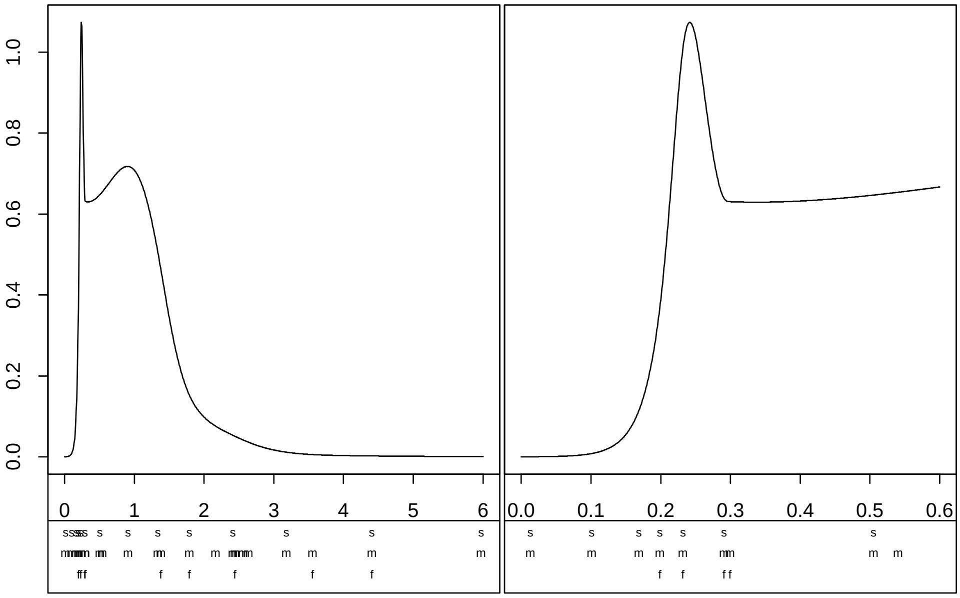


FIG. 1. *Left*: *Logspline density estimate for the income data. Right*: *Enlargement of the area near* $x = 0.24$. *The letters below the plots refer to the knot placement. See the text for details.*

sequence of models, LOGSPLINE selects the one having the smallest value for the AIC criterion. The knots for this "best" model are indicated by the letter "f" in Figure 1.

Usually, as is the case here, the final model based on the AIC criterion is fitted during the stepwise deletion stage of the procedure. The new LOGSPLINE procedure thus has the advantage that it adds knots in those parts of the density where they are most needed, for example, near the spike, while it deletes knots where they are not needed, for example, in the tails, thus creating an adaptivity that other density estimation procedures seem to lack. This is one of LOGSPLINE's main advantages.

LOGSPLINE has additional advantages over other density estimation methods:

1. While LOGSPLINE generally gives accurate estimates of the height and location of peaks, thanks to adaptivity, it avoids spurious bumps and gives smooth estimates in the tail of the distribution.
2. LOGSPLINE has a natural way to estimate densities with bounded support, which may be discontinuous at the end of their range.
3. LOGSPLINE can estimate the density even when some observations are censored.
4. A LOGSPLINE density is represented by a list of numbers of moderate length, making it convenient to use the density for further analysis.

The LOGSPLINE method is fairly fast: on our Sparc 10 workstation the estimate shown in Figure 1 was computed in about 9 s of CPU time.

In the following section we will discuss the LOGSPLINE methodology in some detail. In Section 4.2 we present an example of the application of the various LOGSPLINE algorithms to a much smaller data set.

4.1. *The LOGSPLINE methodology.*

LOGSPLINE models. As usual in our polynomial spline methodologies, there are two main issues to LOGSPLINE:

1. Given a linear space, how are the parameters estimated?
2. How is the linear space selected?

We now discuss the types of linear spaces that we consider in LOGSPLINE and the corresponding log-likelihood function. Then we discuss how to select a linear space in an adaptive manner.

Given the integer $K \geq 3$, the numbers L and U with $-\infty \leq L < U \leq \infty$ and the sequence $t_1, \dots, t_K$ with $L < t_1 < \cdots < t_K < U$, let G be the space of twice continuously differentiable functions s on (L, U) such that the restrictions of s to $[t_1, t_2], \dots, [t_{K-1}, t_K]$ are cubic polynomials and the restrictions of s to $(L, t_1]$ and $[t_K, U)$ are linear. The space G is K-dimensional. Set $J = K - 1$. Then G has a basis of the form $1, B_1, \dots, B_J$. We can choose $B_1, \dots, B_J$ such that B_1 is linear with negative slope on $(L, t_1]$, $B_2, \dots, B_J$ are constant on $(L, t_1]$, B_J is linear with positive slope on $[t_K, U)$ and $B_1, \dots, B_{J-1}$ are constant on $[t_K, U)$.

A column vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)^T \in \mathbb{R}^J$ is said to be *feasible* if

$$\int_L^U \exp(\beta_1 B_1(y) + \cdots + \beta_J B_J(y))\, dy < \infty$$

or, equivalently, if (i) either $L > -\infty$ or $\beta_1 < 0$ and (ii) either $U < \infty$ or $\beta_J < 0$. Let $\mathscr{B}$ denote the collection of such feasible column vectors. Given $\boldsymbol{\beta} \in \mathscr{B}$, set

$$f(y; \boldsymbol{\beta}) = \exp(\beta_1 B_1(y) + \cdots + \beta_J B_J(y) - C(\boldsymbol{\beta})), \qquad L < y < U,$$

where

$$C(\boldsymbol{\beta}) = \log\left(\int_L^U \exp(\beta_1 B_1(y) + \cdots + \beta_J B_J(y))\, dy\right).$$

Then $f(\cdot; \boldsymbol{\beta})$ is a positive density function on (L, U) for $\boldsymbol{\beta} \in \mathscr{B}$. If $U = \infty$, then the density function is exponential on $[t_K, \infty)$; if $L = -\infty$, then the density function is exponential on $(-\infty, t_1]$.

Let $Y_1, \dots, Y_n$ be a random sample of size n from a distribution on (L, U) having density function f. Let $A_1, \dots, A_n$ be subintervals of (L, U) such that it is known only that $Y_i \in A_i$ for $1 \leq i \leq n$. If Y_i is uncensored, then $A_i = \{Y_i\}$. If Y_i is right censored at $C_i < Y_i$, then $A_i = (C_i, U)$. If Y_i is left censored at $C_i > Y_i$, then $A_i = (L, C_i)$. In either case, we refer to C_i as the censoring value of Y_i. If Y_i is interval censored, then its censoring interval A_i is a subinterval of (L, U). Under the usual assumption that the random sample is independent of the censoring mechanism, the log-likelihood function corresponding to the LOGSPLINE model has the form given by

$$l(\boldsymbol{\beta}) = \sum_i \varphi(A_i; \boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathscr{B};$$

here

$$\varphi(y; \boldsymbol{\beta}) = \log f(y; \boldsymbol{\beta}) = \sum_j \beta_j B_j(y) - C(\boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathscr{B},$$

if A is the one-point set $\{y\}$ and

$$\varphi(A; \boldsymbol{\beta}) = \log\left(\int_A f(y; \boldsymbol{\beta})\, dy\right) = \log\left(\int_A \exp\varphi(y; \boldsymbol{\beta})\, dy\right), \qquad \boldsymbol{\beta} \in \mathscr{B},$$

if A has positive length. Formulas for the score function and Hessian can be found in Kooperberg and Stone [(1992), Section 2]. These formulas become rather complicated when A has positive length.

The maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ is given by $l(\widehat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta} \in \mathscr{B}} l(\boldsymbol{\beta})$, and the log-likelihood of the fitted model is given by $\widehat{l} = l(\widehat{\boldsymbol{\beta}})$. The corresponding maximum likelihood estimate of f is given by $\widehat{f}(y) = f(y; \widehat{\boldsymbol{\beta}})$ for $L < y < U$.

Model selection. The knot selection methodology involves initial knot placement, stepwise knot addition, stepwise knot deletion and final model selection based on AIC. In this subsection we assume that all the data are uncensored; that is, $A_i = \{Y_i\}$ for all i.

Initially we start with K knots, with $K = \min(2.5n^{1/5}, n/4, N, 25)$, where N is the number of distinct Y_i's. These K knots are positioned according to the rule described in Kooperberg and Stone (1992). This rule places knots at selected order statistics of the data. (The rule is suitably modified when some data are censored.) If $L = -\infty$ and $U = \infty$, the extreme knots are placed at the extreme observations and the interior knots are positioned such that the distances (on an order statistic scale) between knots near the extremes of the data are fairly small and almost independent of the sample size, while the knots in the interior are positioned approximately equidistantly. If $L > -\infty$ or $U < \infty$, the procedure is suitably modified.

The knot-addition/knot-deletion procedure that we employ is essentially the procedure described in Section 3. In particular, at each addition step of the algorithm we first find a good location for a new knot in each of the intervals (L, t_1), (t_1, t_2), $\dots$, (t_{K-1}, t_K), (t_K, U) determined by the existing knots $t_1, \dots, t_K$. To do this we maximize in each interval the Rao statistic for potential knots located at the quartiles of the data within each interval. The location is then further optimized, which may involve computing a few more Rao statistics [see Section 11.3 of Kooperberg, Stone and Truong (1995a) for our current implementation]. The search algorithm then selects among the best candidates within the various intervals. The default value for the maximum number of knots in a model is $K_{\max} = \min(4n^{1/5}, n/4, N, 30)$.

During knot deletion we successively remove the least significant knot, where Wald statistics are used to measure significance. We continue this procedure until only three knots are left. (Rarely, with extremely heavy tailed densities, there are numerical problems when the number of knots is too small. In such a situation we terminate the procedure as soon as these problems occur.)

Among all models that are fitted during the sequence of knot addition and knot deletion we choose the model that minimizes AIC with default penalty parameter $a = \log n$, as described in Section 3.

Innovations. As we mentioned in the introduction to this section, the present version of LOGPSLINE is the fourth version. In the first version [Stone and Koo (1986b)], a small fixed number of knots was placed equidistantly on an order-statistic logit scale. In Kooperberg and Stone (1991), stepwise knot deletion was employed, and the initial knot placement rule was very similar to the one we now employ. Both of these earlier papers used a preliminary transformation for densities on the positive half-line. In Kooperberg and Stone (1992) it was decided that such a transformation is not needed when the knot placement is sufficiently adaptive. In the 1992 paper we extended logspline density estimation to censored data and discussed a user interface based on S. The present version of LOGSPLINE is the only one that includes stepwise addition of knots. There are also several significant computational improvements, the two most important of which are as follows:

1. The starting values used during stepwise deletion are obtained by maximizing a quadratic approximation to the log-likelihood function, as described in Section 3. These starting values are significantly better than those proposed in Kooperberg and Stone (1992). Indeed, the number of Newton–Raphson iterations may be reduced by as much as 30%.
2. In the absence of censored data the log-likelihood function is strictly concave. Therefore, if a maximum of the log-likelihood function exists, it is unique. If some of the observations are censored, however, the log-likelihood function need not be concave. In Kooperberg and Stone (1992), this problem was circumvented by alternating between Newton–Raphson and steepest ascent. We now take the approach of adding a small negative constant times the identity matrix to the Hessian, if necessary, to ensure that this matrix is negative definite [see Kennedy and Gentle (1980), Section 10.2.2].

Note that the version of the program described in Kooperberg and Stone (1992) is available from statlib (statlib@stat.cmu.edu). The version described in this paper is not yet publicly available.

4.2. *An example.* The penalty parameter a in the AIC criterion (see Section 3) is the main parameter in the LOGSPLINE procedure that governs the complexity of the final density estimate. The default value for this parameter is $a = \log n$ as in BIC. Another commonly used value is $a = 2$ as in (traditional) AIC. One of the goals of this section is to study the influence of this penalty parameter by means of a small simulation study.

Besides the choice of the penalty parameter, it may matter whether we use the new LOGSPLINE procedure, as described in this paper, or the previous LOGSPLINE procedure, described in Kooperberg and Stone (1992). Since the new procedure positions some of the knots adaptively, so as approximately to maximize the log-likelihood, conceivably it may lead to a more flexible estimate.

We applied the new and previous LOGSPLINE procedures with both $a = 2$ and $a = \log n$ to the Buffalo snowfall data. This is a small data set ($n = 63$) that has been used extensively in the density estimation literature; see, for example, Parzen (1979) and Silverman (1986). The main issue here is the number of modes: is there one or are there three (or maybe two)? As can be seen from Figure 2, the different LOGSPLINE procedures provide different answers, as summarized in Table 1. From this table we see that the model that was selected using the new procedure with penalty parameter $a = 2$ would also have been selected for values of a between 0.45 and 3.01. From (3.6) we note that if a model with J basis functions is selected for some value of a, it will be selected for a range of values of a. Some models may not be

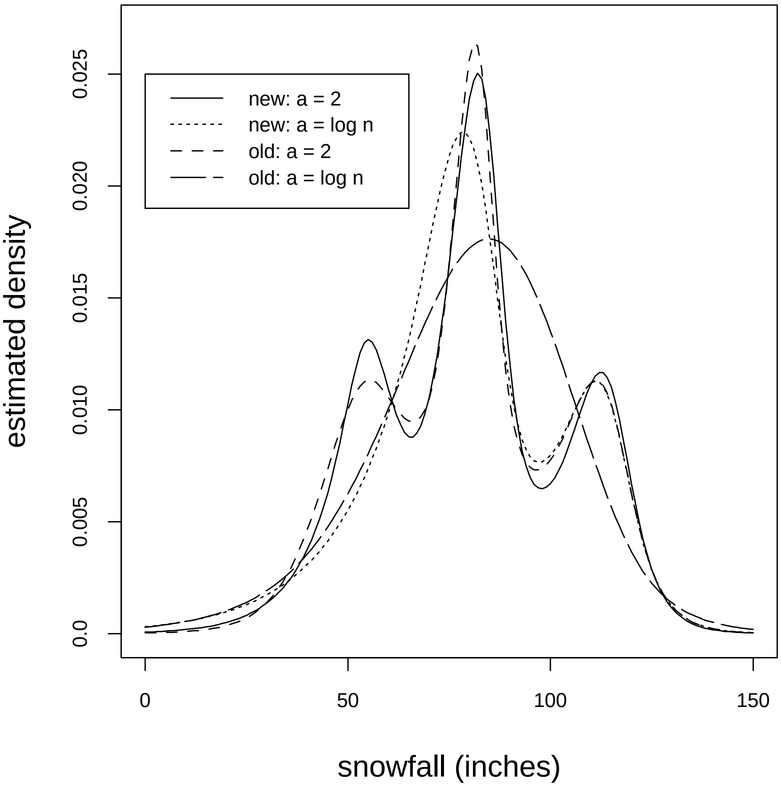


FIG. 2. *Logspline density estimates for the Buffalo snowfall data* ($n = 63$) *for the new and the previous LOGSPLINE procedure and two different values of the penalty parameter.*

TABLE 1
Knots and modes for LOGSPLINE estimates for the Buffalo snowfall data

Procedure	Optimal for a From	Optimal for a To	Number of Knots	Number of Modes
New procedure, $a = 2$	0.45	3.01	7	3
New procedure, $a = \log n \approx 4.14$	3.01	8.38	5	2
Previous procedure, $a = 2$	0.03	2.65	7	3
Previous procedure, $a = \log n \approx 4.14$	2.65	∞	3	1

optimal for any value of a [see Kooperberg, Stone and Truong (1995a), Table 6]. Note that for $n = 63$ the starting number of knots for the previous procedure is 10, while for the new procedure it is 6, with 4 knots being added by the algorithm.

To investigate the behavior of the LOGSPLINE estimation procedures in situations similar to the snowfall data, we generated 100 samples of size 63 from each of the densities shown in Figure 2, except for the estimate of the previous procedure with $a = 2$ since it is very similar to the estimate of the new procedure with $a = 2$. For each of the 300 samples that we obtained, we applied the same procedures with the same choices of a as in Figure 2, yielding four estimates for each sample. In Table 2 we summarize the number of modes in each of these estimates. Not unexpectedly, the procedures with $a = \log n$ tend to underestimate the number of modes, while the procedures with $a = 2$ tend to overestimate it. Although it would be possible to fine tune the penalty parameter to balance the number of times the procedure underestimates and overestimates the number of modes, we feel that it may be more useful to look at a few estimates with different values of the penalty parameter before deciding on the final estimate. From Table 2 we also see that the newer procedures are indeed a little more flexible than the old procedures, yielding even more overestimation of the number of modes for the $a = 2$ procedure, while the new procedure with $a = \log n$ falls in between the two old procedures. From this summary we thus see that with the present sample size it is virtually impossible to distinguish accurately between densities with one, two and three modes. However, when we generated samples from the

TABLE 2
*Number of modes in the simulation study with $n = 63$**

Data generated from:	**Previous $a = \log n$**				**New $a = \log n$**				**New $a = 2$**			
Correct number of modes:	**1**				**2**				**3**			
Estimated number of modes:	**1**	**2**	**3**	**≥ 4**	**1**	**2**	**3**	**≥ 4**	**1**	**2**	**3**	**≥ 4**
New $a = 2$	**39**	41	19	1	7	**74**	17	2	6	26	**64**	4
New $a = \log n$	**74**	23	3	0	34	**64**	2	0	29	40	**31**	0
Previous $a = 2$	**51**	37	11	1	16	**68**	16	0	12	22	**65**	1
Previous $a = \log n$	**84**	13	3	0	51	**46**	3	0	45	26	**29**	0

*The numbers of estimates having the correct number of modes are boldface.

TABLE 3
*Number of modes in the simulation study with $n = 250$**

Data generated from:	**Previous $a = \log n$**				**New $a = \log n$**				**New $a = 2$**			
Correct number of modes:	**1**				**2**				**3**			
Estimated number of modes:	**1**	**2**	**3**	**≥ 4**	**1**	**2**	**3**	**≥ 4**	**1**	**2**	**3**	**≥ 4**
New $a = 2$	**41**	26	25	8	0	**56**	32	12	0	3	**68**	29
New $a = \log n$	**88**	12	0	0	4	**90**	4	2	0	9	**89**	2
Previous $a = 2$	**74**	19	7	0	2	**79**	18	4	0	9	**90**	1
Previous $a = \log n$	**99**	1	0	0	16	**82**	2	0	5	17	**78**	0

*The numbers of estimates having the correct number of modes are boldface.

unimodal density (previous procedure, $a = \log n$) and estimated the density with one of the procedures with $a = 2$, we noticed that when we got two modes, the second mode was more often on the left side of the main mode than on the right side. This is not surprising since the density is slightly flatter on that side. Reversing this reasoning we are lead to believe that the existence of a side mode to the right of the main mode is more plausible than the existence of a side mode to the left of the main mode.

Although all procedures have trouble distinguishing between unimodal and multimodal densities when $n = 63$, most carry out this task well when the sample size gets larger. In Table 3 we summarize a similar simulation study as in Table 2, except that we generated samples of size 250 from the densities in Figure 2. For this sample size the starting number of knots for the previous procedure is 12, while the new procedure starts with eight knots and adds four more during the algorithm. Except for the new procedure with $a = 2$, all methods get the right number of modes at least 74% of the time. The new method with $a = \log n \approx 5.52$ gets it right at least 88% of the time for each of the three situations.

5. Regression (MARS). When viewing regression as a function estimation problem we recognize that the regression function may not be a linear additive function of the predictors and instead allow nonlinear and possibly also nonadditive functions. When there is only one predictor, nonparametric regression can be viewed as smoothing, for which there are numerous methods available. Some of the popular methods are kernel and local polynomial regression [Wand and Jones (1995); Fan and Gijbels (1996)], smoothing splines [Wahba (1990); Green and Silverman (1994)], and polynomial splines. Smith (1982) wrote the first paper to use polynomial splines with adaptively selected knots for regression problems. In her method, knots for cubic splines are positioned uniformly over the range of the data, after which a stepwise knot deletion algorithm is employed.

While many of the univariate nonparametric regression methods can be generalized to situations where there are a few predictors, the curse of dimensionality applies when there are many predictors. One attractive approach for ameliorating this curse is to model the regression function as an additive function of the predictors. This approach has been popularized by Hastie and Tibshirani (1990), who treat both linear regression and generalized regression, including logistic regression and Poisson regression, and emphasize the use of backfitting together with a one-dimensional smoother to fit additive models to data.

An early paper using polynomial splines for additive linear regression as well as additive logistic regression is Stone and Koo (1986a), in which knots were placed at nonadaptive (predetermined) quantiles. Stepwise knot selection, forward and backward, was used in the additive regression program TURBO by Friedman and Silverman (1989). A somewhat different approach to additive regression involving stepwise knot selection was developed by Breiman (1993). In the applications of cubic splines in these papers, linear constraints were placed on the tails of the splines mainly to control the variance of the corresponding estimates.

When nonadditive models are considered, the usual approach to nonparametric regression has been to restrict the model to additive main effects and selected low-order interactions. Gu and Wahba (1993) developed a smoothing spline approach to ANOVA modeling in function estimation. Friedman (1991) introduced multivariate adaptive regression splines (MARS), which is a polynomial spline methodology for estimating the regression function.

In this section we first give a brief description of Friedman's MARS program. When we were working on POLYCLASS [Kooperberg, Bose and Stone (1997)], we found it necessary to develop our own version of MARS to handle very large data sets with many predictors and basis functions. In Section 5.2 we describe this version of MARS and list some differences between our version and Friedman's version. In Section 5.3 we present a small example in which we compare the two programs.

From now on, when we mention "MARS" in this paper, we refer either to Friedman's version or to both versions simultaneously. We refer to our version of the MARS algorithm as "POLYMARS."

5.1. *MARS.* Let $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ denote a random sample from the distribution of $(\mathbf{X}, Y)$, where $\mathbf{X} \in \mathbb{R}^M$ and $Y \in \mathbb{R}$. We wish to estimate $f(\mathbf{X}) = E(Y|\mathbf{X})$. The MARS model [Friedman (1991)] can be written as

$$f(\mathbf{X}) = f(\mathbf{X}|\boldsymbol{\beta}) = \sum_{j=1}^{J} \beta_j B_j(\mathbf{X}). \tag{5.1}$$

For a given set of basis functions, the unknown parameters in MARS are estimated using least squares. The selection of the basis functions in MARS is not easily written in the allowable spaces framework of Section 3. Here we outline the main features of the MARS algorithm when piecewise linear splines are used. A refinement of this algorithm makes use of continuously differentiable functions that are similar, but not exactly identical to the cubic splines employed in various other sections of this paper. (Note that these cubic splines yield twice continuously differentiable functions.)

In the MARS program the one-dimensional model $f(\mathbf{x}) = \beta_1$ is initially fitted. Then, successively, models with J basis functions are replaced by models with $J+1$ or $J+2$ basis functions. This is done by considering the addition of all possible pairs of new basis functions $B_m(\mathbf{x})(x_i - t)_+$ and $B_m(\mathbf{x})(t - x_i)_+$, where x_i is one of the predictors, t is a new knot in that predictor and $B_m(\mathbf{x})$ is a basis function currently in the model that does not depend on x_i. [Some of these additions may involve adding only one genuinely new basis function since one new basis function would already be in the span of the existing basis functions and the other new basis function; see Friedman (1991).] In the MARS algorithm every data coordinate that is sufficiently far from existing knots for the corresponding variable is a candidate for a new knot for that variable. The best model of dimension $J+2$ or $J+1$ is chosen among such candidates for stepwise addition using a generalized cross-validation (GCV) criterion. The stepwise addition of basis functions continues until a user-specified maximum number of basis functions is reached. During the stepwise deletion stage of MARS, any of the nonconstant basis functions can be removed at any step. GCV is used to select the best overall model during the addition or deletion stage.

An option in MARS allows the user to restrict each basis function to depend on at most d predictors. The POLYMARS methodology described below corresponds to MARS with $d = 2$.

5.2. *POLYMARS.* The setup for POLYMARS is identical to that for MARS, except that with POLYCLASS (Section 6) in mind we allow the response Y to be in $\mathbb{R}^K$ with $K \geq 1$. For simplicity, however, we will assume here that $K = 1$ since all computations generalize trivially. As in the other methodologies, we model $f(\mathbf{X})$ in a linear space, so that (5.1) again holds.

For POLYMARS it is convenient to define an allowable space by listing its basis functions. For $1 \leq m \leq M$, let J_m be an integer with $J_m \geq -1$; if $J_m = -1$ there are no basis functions depending on x_m; if $J_m = 0$, consider the basis function $B_{m0}(x_m) = x_m$; if $J_m \geq 1$, consider the basis function $B_{m0}(x_m) = x_m$, let x_{mj} for $1 \leq j \leq J_m$ be distinct real numbers, and consider the additional basis functions $B_{mj}(x_m) = (x_m - x_{mj})_+$ for $1 \leq j \leq J_m$.

Let G be the linear space having basis functions 1, $B_{mj}(x_m)$ for $1 \leq m \leq M$ and $0 \leq j \leq J_m$, and perhaps certain tensor products of two such basis functions. It is required that if $B_{lj}(x_l)B_{mk}(x_m)$ be among the basis functions for some $j \geq 1$, then $B_{l0}(x_l)B_{mk}(x_m) = x_l B_{mk}(x_m)$ and hence (if $k > 0$) $x_l x_m$ be among the basis functions.

One reason for adding linear terms before knots and main effects before interactions is to yield models that are simpler and easier to interpret. A second reason is to reduce the variance associated with the overall modeling procedure, and a third is to reduce the likelihood of ending up with spurious terms in the final model. The requirement of adding main effects before interactions is also motivated by theoretical considerations regarding convergence rates (see Section 2).

It is easy to check that the collection $\mathscr{G}$ of such spaces satisfies the properties listed in Section 3. In particular, the minimal allowable space $G_{\min}$ for the POLYMARS model is the space of constant functions. Thus the minimal model for (5.1) has $J = 1$, $B_1 = 1$ and $f(\mathbf{X}) = \beta_1$ so that $f(\mathbf{X})$ does not depend on the vector $\mathbf{X}$ of predictors. Note that the highest order d of interactions allowed in a POLYMARS model is two.

Given the basis of an allowable space G as defined above, it is obvious whether any given basis function can be deleted in one step.

EXAMPLE. Let $M = 4$, $B_1 = 1$, $B_2 = x_1$, $B_3 = (x_1 - 1)_+$, $B_4 = x_2$, $B_5 = x_3$ and $B_6 = x_1 x_2$. Then $B_1, \dots, B_6$ span an allowable space G. In this example, B_3, B_5 or B_6 could be removed and the remaining space would still be allowable. If one of the basis functions B_2 or B_4 were removed, however, the remaining space would not be allowable since it would still contain $B_6 = B_2 B_4$ (as well as B_3 in the case of removing B_2). The constant basis function B_1 can never be removed.

Let G_0 be the allowable space having basis functions 1, $B_{mj}(x_m)$ for $1 \leq m \leq M$ and $1 \leq j \leq J_m$, and perhaps certain tensor products of two such basis functions. To decide which basis function to add to this model, we compute the Rao statistic as described in Section 3:

(i) For all spaces that can be obtained from G_0 by adding a basis function $B_{l0}(x_l) = x_l$ to G_0;

(ii) for all allowable spaces that can be obtained from G_0 by adding a basis function to G_0 that is a tensor product of two basis functions $B_{lj}(x_l)$ and $B_{mk}(x_m)$, $l \neq m$, that are in G_0;

(iii) for an allowable space that can be obtained from G_0 by adding a basis function corresponding to a potential new knot in predictor m for $1 \leq m \leq M$. For every predictor we consider a fixed number N_0 of potential new knots, which typically are preselected order statistics of the data.

As the new space G we choose the one corresponding to the largest absolute value of the Rao statistic among those candidates listed above that are nonvacuous.

EXAMPLE (Continued). Corresponding to (i), we can add the basis function x_4 to the space in the above example. Corresponding to (ii), we can add $B_2 B_5 = x_1 x_3$, $B_3 B_4 = (x_1 - 1)_+ x_2$ or $B_4 B_5 = x_2 x_3$ to the space. The basis function $B_3 B_5 = (x_1 - 1)_+ x_3$ cannot be added, since the resulting space would not contain $B_2 B_5 = x_1 x_3$ so it would not be allowable. Corresponding to (iii), a basis function $(x_1 - x_{1k})_+$ with $x_{1k} \neq 1$, $(x_2 - x_{2k})_+$ or $(x_3 - x_{3k})_+$ could be added. No basis function of the form $(x_4 - x_{4k})_+$ could be added before x_4 is added.

For a given allowable space, the parameters β_j in (5.1) can be estimated using least squares. The Rao and Wald statistics that are used to decide which basis function to add or delete now reduce to the difference in the residual sum of squares between two nested models. The AIC criterion to select the final model is replaced by a penalized residual sum of squares called GCV [Friedman, (1991)]. In particular, we select the model that minimizes

$$\frac{\mathrm{RSS}_J}{n} \Big/ \left[1 - \frac{a(J-1)}{n}\right]^2,$$

where RSS_J is the residual sum of squares for the model with J basis functions and a is a parameter that we typically set equal to 2.5.

Several computational tricks make it possible for the POLYMARS algorithm to be extremely fast, even for huge data sets and many basis functions. [See Kooperberg, Bose and Stone (1997) for more details.] In particular, since we limit the number of potential locations for new knots, inner products need to be computed at most once. If the maximum number of basis functions considered is $P_{\max}$, the complete POLYMARS program requires $O(N_0 n P_{\max}^2)$ floating point operations (flops), while MARS (which has to recompute inner products since there are too many candidate basis functions to store them all) requires $O(MnP_{\max}^3)$ flops. In particular, on an example with $n = 10,000$, $M = 63$, $N_0 = 20$ and $P_{\max} = 80$, the POLYMARS program required 474 s of CPU time, while MARS required 12,636 s on the same machine.

Besides these computational issues, there are other differences between MARS and POLYMARS:

1. The allowable spaces are different. This is most evident in the addition stage, during which we add first a linear term and perhaps later a knot, while in Friedman's program two basis functions, essentially corresponding to a linear function and a knot, are added at the same time.
2. During the deletion stage POLYMARS requires interaction basis functions to be removed before the corresponding main effects can be removed. Knots have to be removed before linear terms are removed. MARS has no such restrictions.
3. In MARS, but not in POLYMARS, a piecewise cubic approximation to the piecewise linear function is applied after a basis function is added.

5.3. *An example.* For a comparison of the two MARS programs on a small data set, we applied them to the well studied Boston housing data [see, e.g., Belsley, Kuh and Welsch (1980) and Breiman, Friedman, Olshen and Stone (1984)]. The response is the median value of homes in thousands of dollars and there are 13 predictors, many of which are highly collinear.

In our experiment we randomly divided the data into a training set of 304 cases and a test set of 202 cases. Both MARS programs were applied to the training set, using 30 as the maximum number of basis functions, GCV to select the final model and otherwise the default options in both programs. (In MARS we set the maximum number of terms in each basis function equal to 2, to make the program comparable to POLYMARS.) We then computed the mean squared error (MSE) on the test set. We repeated this experiment 10

TABLE 4
MARS fits for the Boston housing data

Method	MSE	CPU
MARS, linear fit	14.37	5.07
MARS, cubic approximation	15.91	5.07
POLYMARS	14.07	3.41

times. The results are summarized in Table 4, together with the average cpu time on our SGI workstation. Since MARS supplies both a piecewise linear fit and a piecewise cubic approximation to this fit, there are two MSE's for this program. The standard errors in the estimates of the mean squared error are all approximately 1.5, while the variation in the CPU times is negligible. Over these 10 repetitions, the correlation between the MSE of the POLYMARS fit and that of the piecewise linear MARS fit is 0.94, while the two other correlations are between 0.4 and 0.6. From this table we see that the difference between the two piecewise linear fits is negligible, while both are a little better than the piecewise cubic approximation.

We then applied both MARS procedures to the complete data, with 80 as the maximum number of basis functions. MARS used 78.6 s CPU time to select 53 basis functions, while POLYMARS used 33.7 s to select 41 basis functions. Both models were very complicated: for example, POLYMARS used 10 of the 13 covariates, and 12 pairs of covariates had at least one tensor-product basis function involving both covariates in the pair. MARS used 11 of the 13 covariates, and 22 pairs of covariates had at least one tensor-product basis function involving both covariates in the pair.

6. Polychotomous regression and multiple classification (POLYCLASS).

6.1. *The POLYCLASS model.* The multiple classification problem is well studied in statistics. Typically, there is a qualitative random variable Y that takes on a finite number $K+1$ of values, which we refer to as classes. Based on a vector of predictors $\mathbf{X} \in \mathbb{R}^M$, we want to predict Y.

In POLYCLASS we use piecewise linear splines and selected tensor products ($d \le 2$) to model the conditional class probabilities. Specifically, suppose $P(Y = k|\mathbf{X} = \mathbf{x}) > 0$ for $k \in \mathscr{K} = \{1, \dots, K+1\}$ and $\mathbf{x} \in \mathscr{X}$, where $\mathscr{X}$ is a subset of $\mathbb{R}^M$ over which $\mathbf{X}$ ranges. Set

$$\theta(k|\mathbf{x}) = \log \frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y = K+1|\mathbf{X} = \mathbf{x})}, \qquad \mathbf{x} \in \mathscr{X} \text{ and } k \in \mathscr{K}.$$

Then $\theta(K+1|\mathbf{x}) = 0$ for $\mathbf{x} \in \mathscr{X}$ and

$$P(Y = k|\mathbf{X} = \mathbf{x}) = \frac{\exp \theta(k|\mathbf{x})}{\exp \theta(1|\mathbf{x}) + \cdots + \exp \theta(K+1|\mathbf{x})}, \tag{6.1}$$
$$\mathbf{x} \in \mathscr{X} \text{ and } k \in \mathscr{K}.$$

We refer to (6.1) as the *polychotomous regression model*; when $K = 1$ it is referred to as the *logistic regression model*.

Let J be a positive integer and let G be a J-dimensional linear space of functions on $\mathscr{X}$ with basis $B_1, \dots, B_J$. Consider the model

$$\theta(k|\mathbf{x}) = \theta(k|\mathbf{x}; \boldsymbol{\beta}_k) = \sum_{j=1}^{J} \beta_{jk} B_j(\mathbf{x}), \qquad \mathbf{x} \in \mathscr{X} \text{ and } k \in \mathscr{K}; \tag{6.2}$$

here $\boldsymbol{\beta}_k = (\beta_{k1}, \dots, \beta_{kJ})^T$ for $1 \le k \le K$, $\boldsymbol{\beta}_{K+1} = 0$ and $\boldsymbol{\beta}$ is the JK-dimensional column vector consisting of the entries of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K$, which range over $\mathscr{B} = \mathbb{R}^{JK}$. Correspondingly, set

$$P(Y = k|\mathbf{X} = \mathbf{x}; \boldsymbol{\beta}) = \frac{\exp \theta(k|\mathbf{x}; \boldsymbol{\beta})}{\exp \theta(1|\mathbf{x}; \boldsymbol{\beta}) + \cdots + \exp \theta(K+1|\mathbf{x}; \boldsymbol{\beta})}$$

for $\boldsymbol{\beta} \in \mathscr{B}$, $\mathbf{x} \in \mathscr{X}$ and $k \in \mathscr{K}$.

In POLYCLASS the basis functions $B_j(\mathbf{x})$ that are used in (6.2) are piecewise linear splines and their selected tensor products. Based on sample data, the coefficients β_{jk} can be estimated by maximum likelihood, yielding a concave optimization problem; see Kooperberg, Bose and Stone (1997) for more details.

As in most of the procedures that we describe in this paper, we use stepwise addition based on Rao statistics and stepwise deletion based on Wald statistics to select the basis functions. Some details specific to POLYCLASS are discussed in Section 6.3. The model selection in POLYCLASS can be carried out using AIC, an independent test set or cross-validation [see Kooperberg, Bose and Stone (1997)].

6.2. *A phoneme recognition example.* In Kooperberg, Bose and Stone (1997), POLYCLASS is applied to a huge data set from the area of speech recognition. Here we present an abbreviated version of this analysis. The source of this data set is the Center for Spoken Language Understanding in Portland, Oregon [Cole, Roginski and Fanty (1992); Cole et al. (1994)]. It consists of 2165 utterances from telephone calls, which are numbers that typically are parts of addresses, zip codes and street numbers. Each utterance was processed by one or more listeners, who produced a time-aligned phonetic description of the utterance. For example, for one particular utterance, "3o3" (three-oh-three), it was determined that from 1 to 167 ms, the speaker produced phoneme `T`, followed by phoneme `r` from 167 to 193 ms and so on. It should be noted that the person who decided which phoneme was spoken was not aware of the text of the utterance. The phoneme transcription, which we obtained from the International Computer Science Institute (ICSI) in Berkeley, California, is based on the LIMSI phonetic alphabet [Gauvain, Lamel, Adda and Adda-Decker (1994)].

The utterances were also processed to produce perceptual linear predictive (PLP) features. Every 12.5 ms the audible spectrum, based on a concentric 25 ms piece of sound, is determined. Since we consider telephone data, which is sampled at the frequency of 8 kHz, there are 200 observations of the sound wave in such a 25 ms interval. A Hamming window is applied to these 200 observations before the spectrum is estimated using the discrete Fourier transform. The estimated spectrum is next transformed to yield a critical-band integrated power spectrum with an equal-loudness preemphasis and a cube root nonlinearity to simulate the auditory intensity–loudness relation. Then the eighth-order autoregressive all-pole model of the transformed spectrum is obtained. The coefficients of the Fourier transform representation of the log-magnitude of this model are known as its cepstral coefficients. The PLP features [Bourlard and Morgan (1994); Hermansky (1990); Rabiner and Juang (1993)] that we used are the log-gain of the model (similar to the variance) and the next eight cepstral coefficients (similar to the autoregressive coefficients).

The goal in our analysis is to estimate the probability distribution over all phonemes at intervals of 12.5 ms based on the (nine) features available at that time point as well as the features available at the c time points, each 12.5 ms apart, before and after the point at which we want to estimate the phoneme distribution.

Such a probability distribution (or, more precisely, a likelihood that is obtained by weighting the estimated probabilities by the empirically determined frequencies of the phonemes) can be used as input to train (estimate) a hidden Markov model, which in turn can be used for automatic speech recognition [Bourlard and Morgan (1994)]. In the hybrid approach described by Bourlard and Morgan, a multilayer perceptron network (a type of artificial neural network) is used to estimate these probabilities.

There were 45 different phonemes, yielding 247,039 cases (12.5 ms intervals). We randomly divided the data into a training set of approximately 112,000 cases and a test set of about 135,000 cases. We used the vector of features at seven different time points, so that $c = 3$ above. The eight cepstral coefficients were used exactly as we received them from ICSI. Since some speakers speak more loudly than others, the log-gain by itself is not an informative predictor of the phoneme that is being spoken. Differences in the log-gain may be more informative. If $e(i)$ is the log-gain at time instance i, we used

$$d(i) = e(i) - \tfrac{1}{7} \sum_{j=-3}^{3} e(i + j)$$

instead of $e(i)$.

The standard POLYCLASS methodology would be practically impossible to apply to the phoneme recognition data, for which $K = 44$, $M = 9 \cdot 7 = 63$ and the sample size is given by $n = 112,115$. In Kooperberg, Bose and Stone (1997) a number of modifications, which make it possible for POLYCLASS to deal with this data set, are discussed. The most important such modification is that instead of computing the regular Rao statistics during the stepwise addition stage, a related least squares problem is solved.

We fitted a POLYCLASS model with 350 basis functions to the data. This maximum number was constrained by the computing resources that were available to us on a network of workstations at the Maui High Performance Computing Center. We believe that a larger number of basis functions would give better results. Exhaustion of our computing resources also prevented us from applying the stepwise deletion algorithm to the largest model. However, intermediate results suggest that the deletion of some basis functions would not significantly improve our results.

Of the 350 basis functions that were selected by the POLYMARS algorithm, 1 is the constant function, 31 are of the form x_i, 45 are of the form $(x_i - x_{ik})_+$, 134 are of the form $x_i x_j$, 87 are of the form $(x_i - x_{ik})_+ x_j$ and 11 are of the form $(x_i - x_{ik})_+(x_j - x_{jl})_+$. Thus, of the 63 features, 32 are not used. Of the remaining 31, 10 are involved in all types of basis functions, 10 more are involved in all types of basis functions except for $(x_i - x_{ik})_+(x_j - x_{jl})_+$ and 8 are involved in basis functions of the types x_i, $(x_i - x_{ik})_+$, $x_i x_j$ and $x_i(x_j - x_{jk})_+$. Finally, two features have basis functions of the types x_i, $(x_i - x_{ik})_+$ and $x_i x_j$ only, and one feature appears only linearly in the model.

The 63 features can be organized in a 9 (cepstral coefficients) × 7 (time points) table. If we label the features from 1, for the feature that occurs only linearly, to 5, for the features that are involved in all types of basis functions, and we ignore the entries for the 32 features that are unused, we obtain Table 5. From this table we clearly see that the most important information is obtained from time points -3 (37.5 ms before the phoneme was spoken), 0 (when the phoneme is spoken) and 3 (37.5 ms after the phoneme was spoken). This table suggests that, in retrospect, it would have been better to use the cepstral coefficients at more than seven time points. (We also see that the log-gain and the shorter lags are more important than the longer lags.)

In Figure 3 we report the misclassification rate and the fitted log-likelihood

$$\frac{\sum_i \log P(Y = Y_i | \mathbf{X} = \mathbf{X}_i)}{n}$$

TABLE 5
The features in the POLYCLASS model

	Time						
Cepstral Coefficient	**−3**	**−2**	**−1**	**0**	**1**	**2**	**3**
Log-gain	5	4	3	5	3		
Lag 1	5		4	5			4
Lag 2	4		5	2			5
Lag 3	4			4			5
Lag 4	5			5	1		5
Lag 5	3			4			4
Lag 6				4			3
Lag 7	3		2				3
Lag 8	3		3				4

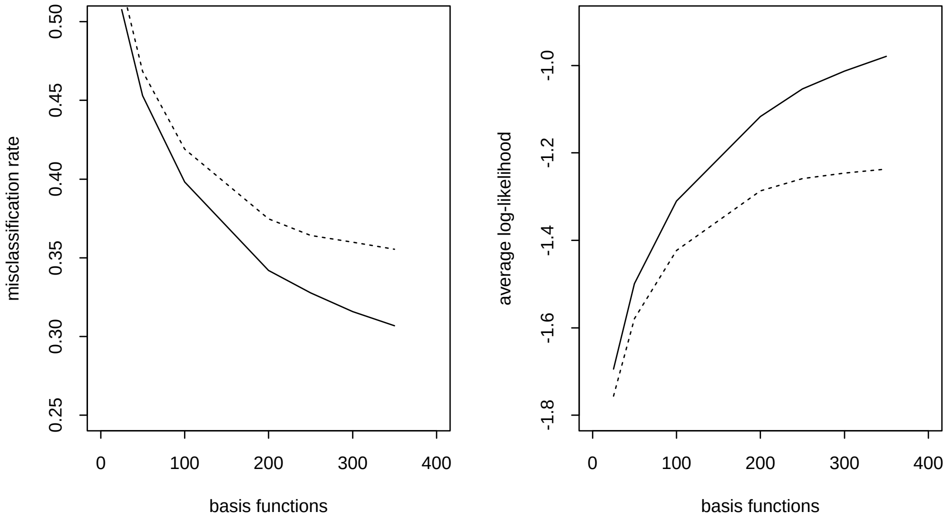


FIG. 3. *Misclassification rate (left) and fitted log-likelihood (right) versus the number of basis functions. Solid line, training set; dashed line, test set.*

for the training set and the test set combined. From these graphs it appears that the fit would continue to improve if we were to increase the number of basis functions.

As mentioned earlier, in this particular application the estimation of conditional class probabilities is more important than classification, since these probabilities can be used as inputs to the hidden Markov model for the approach to speech recognition described in Bourlard and Morgan (1994). POLYCLASS is particularly useful in this situation since, unlike most other classification methods, it provides viable estimates of the conditional class probabilities. In Figure 4 we plot the estimated probability that a case is a particular phoneme grouped in bins of size 0.01 on the horizontal axis and the fraction of cases with that probability that correspond to the correct phoneme on the vertical axis. Note that each case contributes 45 observations to this graph: one observation per candidate phoneme. These graphs are extremely close to the ideal straight line (fraction true class) = (estimated probability) for the test set (left side) and the training set (right side).

Clearly, not all phonemes are correctly estimated with the same probability. In Figure 5 we plot the average probability, over the test set, assigned to each phoneme. We see from Figure 5 that, not surprisingly, this probability is much larger for the frequently occurring phonemes than for the infrequently occurring ones.

Other aspects of the analysis that are discussed in Kooperberg, Bose and Stone (1997) are a comparison of POLYCLASS with other classification methods and an analysis of the patterns of misclassification by POLYCLASS. In

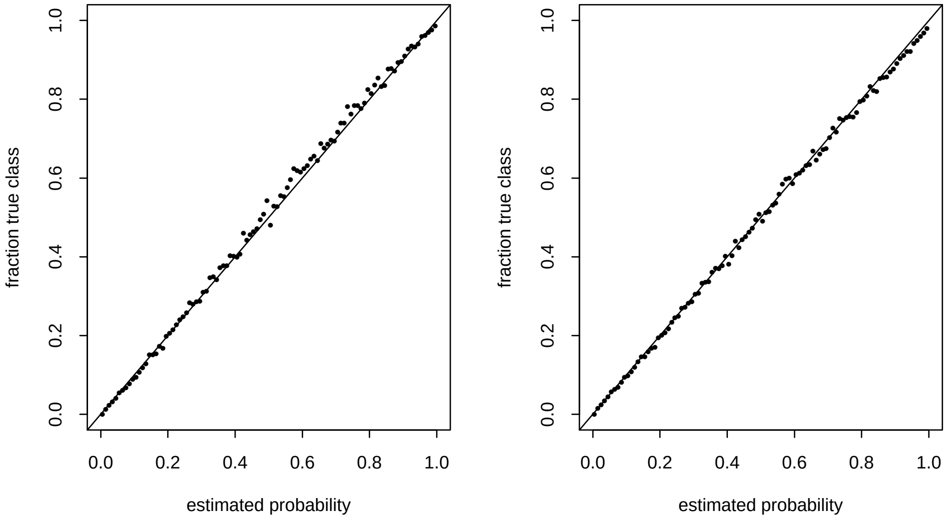


FIG. 4. *Fraction of phonemes that correspond to the true class versus the estimated probability. Data have been grouped in bins of size* 0.01. *Left*, *training set*; *right*, *test set.*

particular, it was found that most of the traditional classification methods either are not able to deal with such a large data set or are outperformed by POLYCLASS. Neural networks, however, do give better results on related, but not identical, data. It was hypothesized that for POLYCLASS to be competitive with neural networks it should be able to fit larger models faster, so

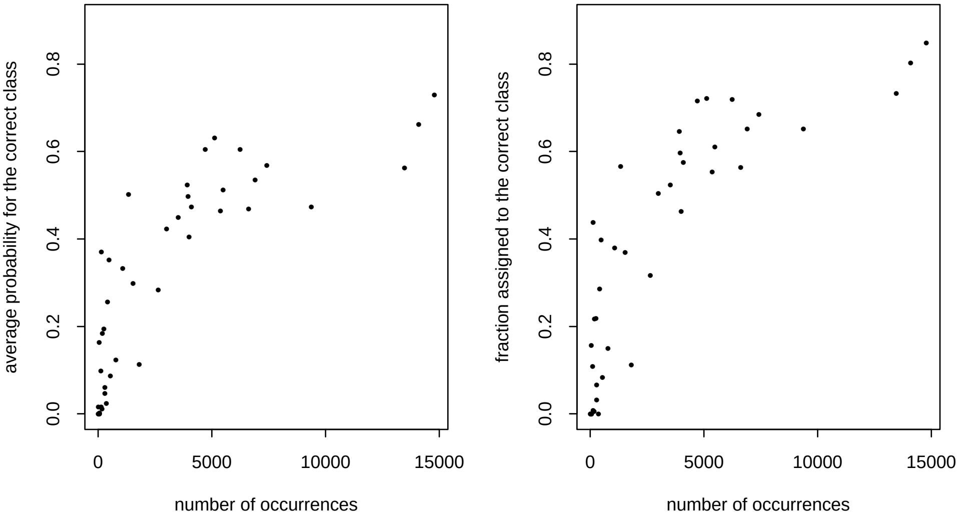


FIG. 5. *Average probability assigned to the correct class and fraction correctly classified versus the class frequency for the test set.*

that, for example, one could experiment with different sets of features. It may be that a stochastic gradient method (as in the backpropagation algorithm used in fitting neural networks) can give POLYCLASS the required computing power.

6.3. *Some more details of POLYCLASS.* The basis functions that are used in POLYCLASS are piecewise linear splines and their tensor products. We impose similar restrictions as in POLYMARS on which basis functions are allowed; that is, linear functions in one of the predictors are always allowed, while basis functions of the form $(x_i - x_{ik})_+$ are allowed in the model only when the corresponding linear function is already included in the model. Tensor products of basis functions involving two different predictors already in the model are allowed, except that if such a tensor product involves a knot in either or both of the predictors, the corresponding basis functions with linear terms must already be in the model. Thus, for $(x_i - x_{ik})_+(x_j - x_{jl})_+$ to be allowed in the model $x_i(x_j - x_{jl})_+$, $(x_i - x_{ik})_+ x_j$ and $x_i x_j$ need already be in the model.

The main difference between POLYCLASS and the other methodologies discussed in this paper is that in POLYCLASS there are K parameters for each basis function, while for the other methodologies there is only one parameter. This substantially increases the amount of computation needed for large data sets. For example, for the phoneme recognition problem discussed in the previous section the number of parameters for the largest model equals 15,400. Thus even storage of a (pseudo-) Hessian becomes prohibitively expensive, while the computation of one score function takes $O(JKn)$ floating point operations (flops) for a model with J basis functions and the computation of a Hessian takes $O(J^2K^2n)$ flops. The following modifications of the POLYCLASS algorithm, to make it feasible to deal with very large data sets, are discussed in Kooperberg, Bose and Stone (1997):

1. During the stepwise addition stage of the program we use a multiresponse least squares approximation to the POLYCLASS problem. That is, we regress $K + 1$ response vectors Z_k on the basis functions, where $Z_{ki} = \operatorname{ind}(Y_i = k)$, $i = 1, \dots, n$ and $k = 1, \dots, K + 1$, with $\operatorname{ind}(\cdot)$ being the usual indicator function. This least squares approximation can conveniently be carried out using a multiresponse version of the MARS algorithm described in Section 5. Selecting J basis functions now requires $O(50nJ(J + K))$ flops.
2. After the J basis functions have been selected using this least squares approximation, we immediately fit the largest model using maximum likelihood. To obtain good starting values we successively add basis functions to the model, using only a fraction of the cases, until all basis functions are in the model.
3. The code was modified to enable the maximum likelihood fitting to be carried out on a network of 64 workstations at the Maui High Performance Computing Center.

With these modifications, the time needed to fit the largest POLYCLASS model was reduced from an estimated several years to one day on the network of workstations.

7. Hazard regression.

Recall the discussion of hazard regression in Section 2. Let $F(t|\mathbf{X}) = P(T \leq t|\mathbf{X})$ denote the conditional distribution function of the survival time T given the random vector $\mathbf{X}$ of covariates and let $f(t|\mathbf{X})$ denote the corresponding conditional density function. Define the conditional hazard function by $\lambda(t|\mathbf{X}) = f(t|\mathbf{X})/[1 - F(t|\mathbf{X})]$ and set $\phi(t|\mathbf{X}) = \log \lambda(t|\mathbf{X})$. A proportional hazard model is specified by setting $\phi(t|\mathbf{X}) = \phi_0(t) + \mathbf{X}\boldsymbol{\beta}$; here $\phi_0(\cdot)$ is the baseline log-hazard function and $\boldsymbol{\beta} \in \mathbb{R}^M$ is a vector of parameters. Cox (1972) suggested a partial likelihood principle for estimating $\boldsymbol{\beta}$. Since then, analyses of censored outcome data have largely been confined to the estimation of linear covariate effects. See, for example, Andersen, Borgan, Gill and Keiding (1993), Cox and Oakes (1984), Fleming and Harrington (1991), Kalbfleisch and Prentice (1980) and Miller (1981).

The desire to relax the proportionality and linearity assumptions has led to many further developments in survival analysis. For example, Hastie and Tibshirani (1990), Sleeper and Harrington (1990) and Gray (1992) considered using splines to model nonlinear covariate effects in large clinical studies. In practice, it is even more desirable to estimate the conditional hazard, distribution and density functions. Based on proportional hazards models, Breslow (1972, 1974) suggested estimating the conditional distribution by combining Cox's partial likelihood principle for the covariate effects and the Kaplan and Meier (1958) method for estimating the baseline survival function. Following the extended linear modeling framework described in Sections 2 and 3, Kooperberg, Stone and Truong (1995a, b) developed a more general approach, which, without requiring the proportionality and linearity assumptions, yields estimates of the conditional hazard, density, survival and quantile functions in a unified manner using the relationships

$$F(t|\mathbf{x}) = 1-\exp\biggl(-\int_0^t \lambda(u|\mathbf{x})\,du\biggr) \quad \text{and} \quad f(t|\mathbf{x}) = [1-F(t|\mathbf{x})]\lambda(t|\mathbf{x}), \qquad t \geq 0.$$

In the remainder of this section, we describe the methodologies for hazard estimation with flexible tails (HEFT) and hazard regression (HARE), and we give an example to illustrate their practical application.

7.1. *The HEFT and HARE methodologies.*

HEFT. The HEFT methodology is designed to estimate the unconditional (or baseline) log-hazard function. Let f denote a positive density function on $(0, \infty)$ and let F, λ and ϕ be its distribution, hazard and log-hazard functions, respectively. Given the integer $J \geq 3$ and the sequence $t_1, \ldots, t_J$ with $0 < t_1 < \cdots < t_J < \infty$, let G_0 be the $(J-2)$-dimensional space of twice continuously differentiable, cubic spline functions s on $[0, \infty)$ with knots $t_1, t_2, \ldots, t_J$ such that s is constant on $[0, t_1]$ and on $[t_J, \infty)$. Let $B_1, \ldots, B_{J-2}$ be a basis of this space such that $B_{J-2} = 1$ on $[0, \infty)$ and $B_1, \ldots, B_{J-3} = 0$ on $[t_J, \infty)$.

To enhance its flexibility in estimating the hazard function, the space G_0 can be augmented by adding the basis functions

$$B_{-1}(t) = \log \frac{t}{t+c} \quad \text{and} \quad B_0(t) = \log(t+c), \qquad t > 0,$$

with $c > 0$ being a parameter. In fact, the linear space G spanned by $G_0 \cup \{B_{-1}, B_0\}$ includes Weibull and Pareto distributions as special cases [see Kooperberg, Stone and Truong (1995a)]. The collection $\mathscr{G}$ of such J-dimensional spaces G forms a family of allowable spaces.

Set

$$\boldsymbol{\beta} = (\beta_{-1}, \beta_0, \beta_1, \ldots, \beta_{J-2}) \in \mathbb{R}^J,$$

$$\phi(\cdot; \boldsymbol{\beta}) = \beta_{-1}B_{-1}(\cdot) + \beta_0 B_0(\cdot) + \beta_1 B_1(\cdot) + \cdots + \beta_{J-2}B_{J-2}(\cdot)$$

and

$$\mathscr{B} = \{(\beta_{-1}, \beta_0, \beta_1, \ldots, \beta_{J-2}) \in \mathbb{R}^J\colon \beta_{-1} > -1 \text{ and } \beta_0 \geq -1\}.$$

The above constraints ensure that $\int_0^t \exp\, \phi(u; \boldsymbol{\beta})\, du < \infty$ for $0 < t < \infty$ and $\int_0^\infty \exp\, \phi(t; \boldsymbol{\beta})\, dt = \infty$. We use $\phi(\cdot; \boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathscr{B}$, to model the log-hazard function.

Given a random sample, the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by using the Newton–Raphson method. (Note that the log-likelihood function here is easily obtained from that for hazard regression discussed in Section 2 by ignoring the covariates.) Estimates of the log-hazard, hazard, survival, distribution and density functions are given by $\widehat{\phi}(t) = \phi(\cdot; \widehat{\boldsymbol{\beta}})$, $\widehat{\lambda}(t) = \exp\, \widehat{\phi}(t)$, $\widehat{S}(t) = \exp\big(-\int_0^t \widehat{\lambda}(u)\, du\big)$, $\widehat{F}(t) = 1 - \widehat{S}(t)$ and $\widehat{f}(t) = \widehat{S}(t)\widehat{\lambda}(t)$, $t \geq 0$. The corresponding estimate of the pth quantile is given by $\widehat{Q}_p = \widehat{F}^{-1}(p)$.

Observe that the above log-hazard estimate depends on the choice of G. HEFT selects such a G adaptively from $\mathscr{G}$ by following the methodology for model selection described in Section 3. (In the current implementation of HEFT, the choice of which logarithmic terms to include in the model is made initially by the user and is not modified during the process of stepwise addition and deletion of knots.)

HARE. HARE is a routine for estimating covariate effects on a possibly censored response variable. Here the allowable spaces are similar to those used in POLYMARS, except that the conditional log-hazard function also depends on time. To this extent we also allow piecewise linear basis functions depending on time and tensor products of these with (piecewise linear) basis functions depending on a covariate. As with POLYMARS and POLYCLASS, the highest order of interactions allowed is two. Let $\mathscr{G}$ denote the collection of such allowable spaces.

For an allowable space in $\mathscr{G}$, we get estimates of the coefficients of basis functions by maximizing the log-likelihood function given in the discussion of hazard regression in Section 2. This procedure is carried out using the Newton–Raphson method. Estimates of the conditional log-hazard, conditional hazard, conditional survival, conditional distribution and conditional density functions are obtained in a manner similar to HEFT.

For model selection, the adaptive methodology is essentially the same as described in Section 3 with $d \leq 2$. In the current implementation of HARE, the fitted conditional log-hazard function has a constant tail. For details, see Kooperberg, Stone and Truong (1995a).

Besides providing a unified framework for estimating the conditional hazard, survival, density and quantile functions, HEFT and HARE also allow considerable flexibility in fitting survival data. If the fitted model contains an interaction involving time and a covariate, then the assumption of proportionality is questionable. On the other hand, HARE can be forced to fit a proportional hazards model or even an additive model ($d = 1$).

HEFT as preprocessor to HARE. Before applying HARE, it is useful to transform the time variable using HEFT. There are two advantages in doing this. First, because of the piecewise linear nature of HARE, the first derivative of the baseline hazard function can have big jumps at various knots in time. The HARE model for the transformed data, on the other hand, typically has fewer knots and the jumps in the first derivative of the hazard function at these knots tend to be smaller. Second, the fitted conditional hazard function beyond the last knot is necessarily constant when HARE is applied to the original data, but this is not the case when HARE is applied to the transformed values of time.

Let λ_0 denote the unconditional (baseline) hazard function of T and set $q_0 = -\log(1 - F_0)$ with F_0 being the distribution function corresponding to λ_0, so that q_0 is the baseline cumulative hazard function. Then $q_0(T)$ has constant hazard function [see Kooperberg, Stone and Truong (1995a)]. This motivates the use of HARE on the transformed responses.

We next describe relationships between the transformed and untransformed data. Let f_1, F_1 and λ_1 denote the conditional density, distribution and hazard functions of $q_0(T)$ given $\mathbf{X}$. Then the corresponding functions for T given $\mathbf{X}$ are given, respectively, by

$$f(t|\mathbf{X}) = \lambda_0(t) f_1(q_0(t)|\mathbf{X}), \qquad F(t|\mathbf{X}) = F_1(q_0(t)|\mathbf{X})$$

and

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\lambda_1(q_0(t)|\mathbf{X}).$$

Moreover, the pth conditional quantile function is given by

$$Q_p(\mathbf{x}) = F^{-1}(p|\mathbf{x}) = q_0^{-1}(F_1^{-1}(p|\mathbf{x})).$$

Given a random sample, our methodology starts by applying HEFT to the response variables (no covariates), yielding an estimate $\widehat{\lambda}_0$ of λ_0. Then $\widehat{q}_0$ is constructed based on the formula of the cumulative hazard function. Next the HARE methodology is applied to the transformed responses $\widehat{q}_0(T)$, yielding an estimate $\widehat{\lambda}_1$ of the conditional hazard function for the transformed data. Finally, we obtain estimates of the original conditional density, distribution, hazard and quantile functions using the relationships given above.

7.2. *An example.* In this section we use HEFT and HARE to analyze data from a clinical trial. The studies of left ventricular dysfunction [SOLVD (1990)] involves two double-blind, randomized clinical trials to test improved survival by treatment with enalapril, an inhibitor of angiotensin-converting enzyme, in patients with left ventricular dysfunction with or without congestive heart failure (CHF). The study started with a registry of 6273 patients involving 23 centers located in the United States, Canada and Belgium. Men and women aged 21–80 years with an ejection fraction (defined below) of at most 35% were eligible for the trials. In particular, patients with overt CHF were eligible for the treatment trial, whereas those with left ventricular dysfunction but no history of overt CHF were eligible for the prevention trail. Recruitment began in 1986, and the study terminated in 1991.

We will illustrate the use of HEFT and HARE on the treatment arm consisting of 2569 patients. Here the event is defined as death or hospitalization due to CHF. The response is time (in days). Among the 2569 observations, 1219 were censored. The censoring occurred when the patient was lost to follow-up or was still alive and never hospitalized due to CHF by the end of the study. We begin our analyses by applying HEFT to the possibly censored responses, yielding a model for the unconditional log-hazard function consisting of three knots and a log term (B_{-1}). Figure 6 shows estimates of the unconditional hazard and survival functions. As the right side of Figure 6 shows, our survival function estimate is remarkably close to the Kaplan–Meier estimate.

Next, HARE was applied to examine covariate effects on CHF. We used a set of 10 covariates: treatment (1=enalapril, 0=placebo); serum sodium level (serum); systolic blood pressure (SBP); dystolic blood pressure (DBP); smoking (1 = currently smoking, 0 = not currently smoking); sex (1 = female, 0 = male);

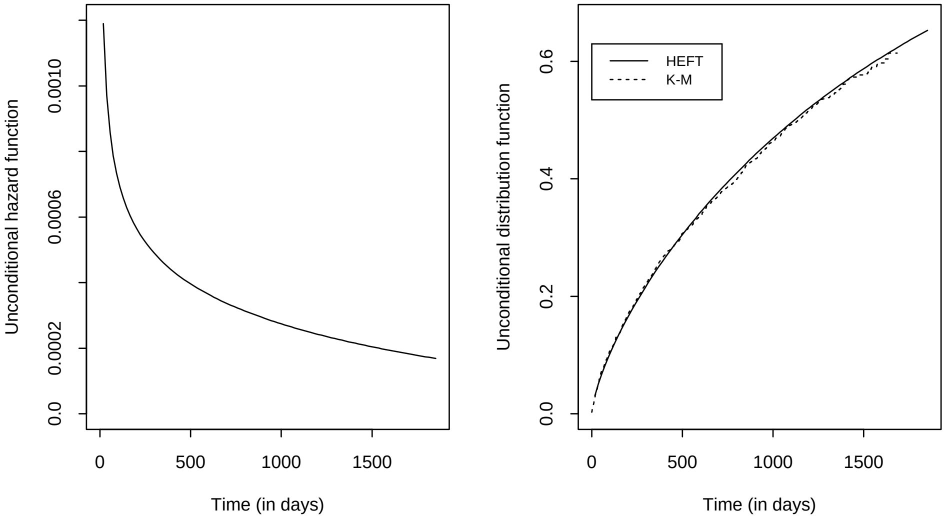


FIG. 6. *Estimated unconditional hazard and distribution functions using HEFT for the SOLVD data.*

age; adherence (a measure of treatment or placebo use in terms of numbers of pills taken and dispensed); New York Heart Association (NYHA) functional class I–IV (with I indicating the least severity of illness and IV indicating the greatest severity); and ejection fraction (EF).

The ejection fraction (EF) is the fraction (measured as a percentage) of the blood that is pumped from the left ventricle into the body's vascular system. After oxygenation in the lung, blood flows back to the left atrium of the heart and continues to the left ventricle. This is the chamber that "ejects" the blood from the heart into the body. Clearly, 100% of the blood cannot be ejected, but in normal hearts this fraction is at least 60%. In damaged hearts, where the muscle of the left ventricle is not working well (maybe from the effects of a previous heart attack), the fraction can be much lower, say 25–40%. Clinically, an EF of less than 35% is reason for concern. Below 15–20% the blood backs up into the atrium and lung, causing congestion and malfunctioning of the lung (CHF) and possibly death.

After removing the 69 cases with missing values on one or more covariates, we obtained a data set with 2500 observations and 1308 events. In our analyses we treated the covariate NYHA as an unordered categorical variable. Alternatively, we could have treated it as an ordinary variable having the four possible values 1, 2, 3 and 4.

Table 6 shows the results of applying HARE in various ways. Specifically, Model 1 summarizes the fit to the untransformed responses, which has 15

TABLE 6
*HARE analyses of the SOLVD data**

Basis Function	**Model 1**	**Model 2**	**Model 3**	**Model 4**
1	7.550	34.900	32.016	32.706
Age	0.013	0.010	0.009	0.011
Smoking	0.400			0.184
DBP		−0.424	−0.388	−0.400
EF	−0.567	−0.026	−0.026	−0.026
NYHA I			−0.294	−0.291
NYHA II	−0.462			
NYHA III	0.757	0.527	0.485	0.479
NYHA IV	1.210	0.980	18.577	19.004
Serum	−0.114	−0.248	−0.227	−0.233
Treatment	−0.124	−0.312	−0.302	−0.303
$(111 - t)_+$	0.006			
$(562 - t)_+$	0.002			
DBP × serum		0.003	0.003	0.003
EF × serum	0.004			
NYHA IV × serum			−0.127	−0.130
$(562 - t)_+ \times$ smoking	−0.001			
$(562 - t)_+ \times$ NYHA II	0.001			
$(562 - t)_+ \times$ treatment	−0.001			
BIC	21,620.17	21,562.30	21,561.83	21,562.32

*See text for the model descriptions.

basis functions and BIC = 21,620.17. As discussed in Section 7.1, the above analysis can further be refined by applying HARE to the transformed responses using $\widehat{q}_0(t) = -\log(1 - \widehat{F}_0(t))$, where $\widehat{F}_0(t)$ is shown on the right side of Figure 6. This yields a proportional hazards model having nine basis functions with no knots and BIC = 21,562.30. (Actually, BIC for the transformed data is 2480.49. We used the relationships described in Section 7.1 to retrieve BIC for the untransformed data.) The resulting fit is referred to as Model 2 in Table 6. Note that all of the interactions and the two nonlinear terms involving time have disappeared; this may be explained by the nature of the transformation $\widehat{q}_0(T)$. While HARE models allow for nonlinearity, this smaller model is linear and easier to interpret. In general, one of the strengths of HARE is that it chooses more complicated models only when simpler ones do not fit nearly as well [see the examples in Kooperberg, Stone and Truong (1995a)].

HARE facilitates the visual examination of covariate effects. For example, Figure 7 shows estimates of the conditional hazard and survival functions for a patient having the covariate values given by

$$\text{treatment} = 1, \quad \text{serum sodium} = 138.95, \quad \text{EF} = 24.85,$$
$$\text{DBP} = 76.81, \quad \text{NYHA} = \text{IV}, \quad \text{smoking} = 1, \quad \text{age} = 60.88.$$

These values were chosen to represent an average smoking, NYHA class IV, treated patient. Figure 7 also compares results from untransformed data (Model 1) and transformed data (Model 2). We remark that the estimated hazard function for the untransformed data exhibits a constant tail, as was

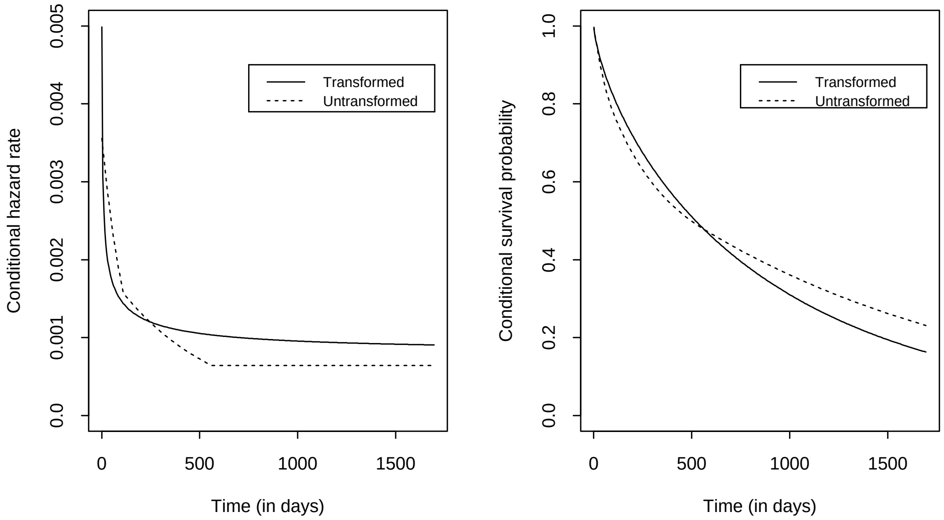


FIG. 7. *Estimated conditional hazard and survival functions for an average smoking, NYHA class IV, treated patient using HARE for the SOLVD data.*

discussed in Section 7.1. Estimates of the conditional density and quantile functions are also easily obtained using HARE.

We continue our analysis by using other options in HARE. Since Model 2 is a proportional hazards model, we decided to reapply HARE forcing it to fit such a model. Model 3 of Table 6 summarizes the resulting fit, indicating a slightly different proportional hazards model with 11 basis functions and BIC = 21,561.83. (BIC for the transformed data is 2480.01.) Comparing this model with Model 2, we note that HARE has reduced BIC slightly by including two more basis functions, NYHA I and NYHA IV × serum.

For a further comparison, we fitted the transformed values of time and the same covariates as above using `coxreg` from S-PLUS. In light of the analysis using HARE, we forced the two interaction terms of Model 3 into the Cox model (the default form of `coxreg` estimates main effects only). Table 7 provides a summary of the fit.

Observe that the interaction terms are highly significant and that the fit is similar to Model 3, except that the covariate smoking is significant and the constant term is not allowed in `coxreg`. Since there is no knot in Model 3, we felt that the default penalty value $\log(2500) \doteq 7.82$ of HARE might have been too high. (This is equivalent to using the chi-squared test with 1 degree of freedom and the significance level of $\alpha \doteq 0.005$ to test the model with 12 basis functions versus a submodel with 11 basis functions.) By using a smaller penalty value of 7.1 ($\alpha \doteq 0.007$) and refitting the data using HARE, we obtained Model 4 in Table 6, which has 12 basis functions. This model is in close agreement with the one obtained by using `coxreg` and shown in Table 7. Moreover, the standard errors of the coefficients in Model 4 (not shown) are remarkably close to the corresponding ones in Table 7. We conclude that Model 4 is our most reasonable HARE model for the data.

Note that the treatment effect is included in all five models discussed above. In fact, the treatment was so effective that, for ethical reasons, the trial was terminated early. Other important covariates are the ejection fraction

TABLE 7
Analyses of the SOLVD data using `coxreg` *from S-PLUS*

Variable	**Coefficient**	**SE**	***P*-value**
Age	0.011	0.003	0.000
Smoking	0.185	0.067	0.006
DBP	−0.401	0.106	0.000
EF	−0.027	0.004	0.000
NYHA I	−0.293	0.106	0.005
NYHA III	0.479	0.059	0.000
NYHA IV	19.480	6.040	0.001
Serum	−0.234	0.061	0.000
Treatment	−0.304	0.056	0.000
DBP × serum	0.003	0.001	0.000
NYHA IV × serum	−0.134	0.044	0.002

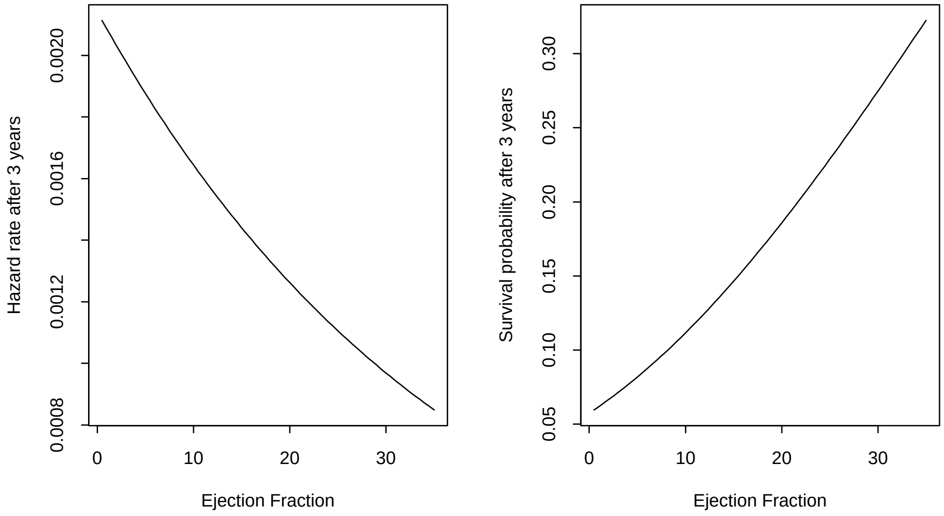


FIG. 8. *Left side*: *estimated conditional hazard rate after* 3 *years as a function of EF. Right side*: *estimated conditional survival probability after* 3 *years as a function of EF. Same covariates as in Fig.* 7.

(EF), age and the NYHA functional class. To demonstrate another strength of HARE, we use Model 4 to examine graphically some of the above covariate effects. Figure 8 illustrates estimates of the conditional hazard rate and survival probability after 3 years as a function of EF. We see that the hazard rate decreases and the survival probability increases with EF. Figure 9 shows

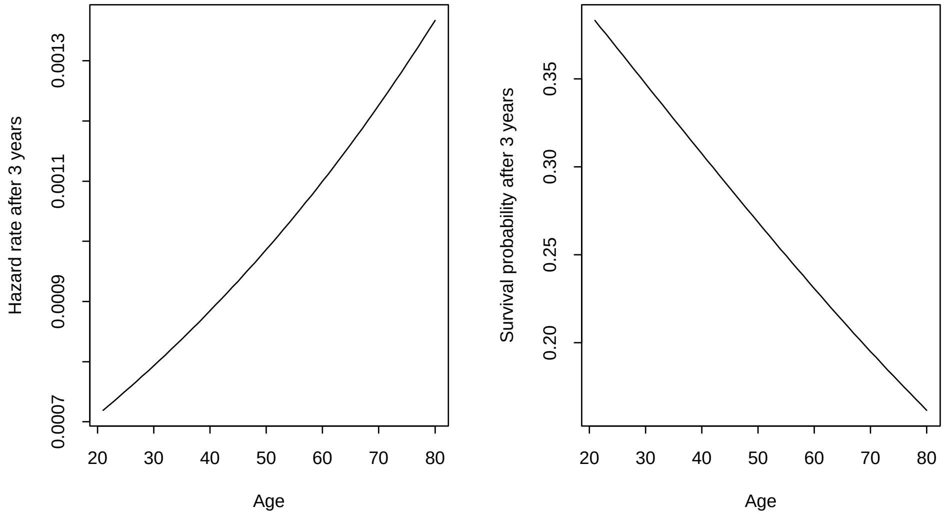


FIG. 9. *Left*: *Estimated conditional hazard rate after* 3 *years as a function of age. Right*: *Estimated conditional survival probability after* 3 *years as a function of age. Same covariates as in Figure* 7.

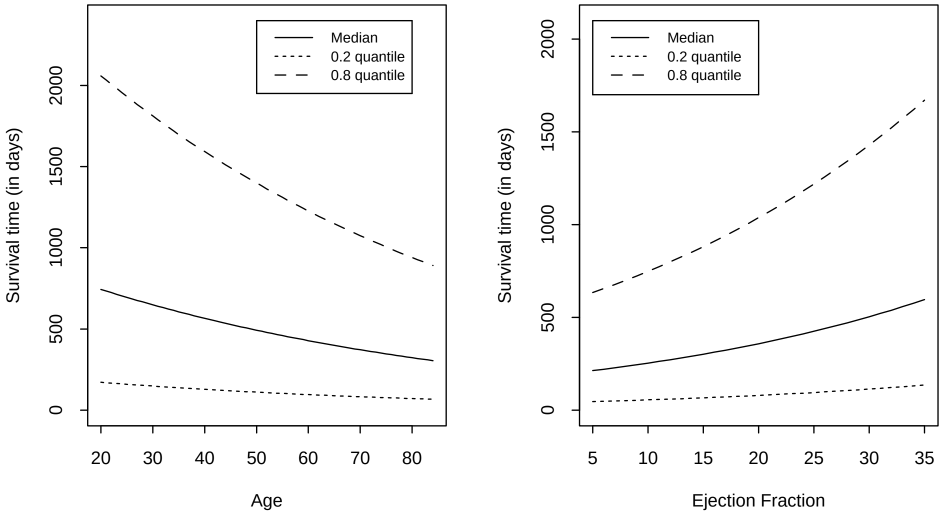


FIG. 10. *Estimated conditional quantile functions based on Model* 4 *as a function of age* (*left*) *and as a function of EF* (*right*). *Same covariates as in Figure* 7.

estimates of the hazard rate and survival probability after 3 years as functions of age. It is observed that older participants have a higher risk than younger ones.

As a final illustration of HARE, Figure 10 shows estimates of the 20th, 50th and 80th percentiles as functions of age and EF based on Model 4. Observe that the median survival time decreases with age, while it increases with EF.

In summary, in the above analyses the HEFT and HARE methodologies yielded estimates of the (conditional) hazard, survival, density and quantile functions in a consistent manner without requiring the proportionality assumption. Moreover, our highly adaptive methodology performs well in comparison with the traditional approach even when that approach is applicable. In light of this example and those given in Kooperberg, Stone and Truong (1995a), we find that HEFT and HARE are useful tools for survival analysis.

8. Spectral analysis. For stationary time series, it is known that the periodogram ordinates at the Fourier frequencies are approximately independent and have an exponential distribution with mean equal to the spectral density function. This implies that the periodogram is not a consistent estimate, but consistency can be achieved by smoothing the periodogram ordinates [see Brillinger (1981)]. In this section we present our version of the spectral estimate by treating it as a special case of the generalized regression problem discussed in Section 2. Specifically, we use the theory and methodology of extended linear models to estimate the logarithm of the mean of the exponential distribution function. Here the mean is the spectral density function.

To describe the possibly mixed spectral distribution, consider a real-valued, second-order stationary time series X_t with mean $E(X_t) = E(X_0)$ and covariance function $\gamma(u) = \operatorname{cov}(X_t, X_{t+u})$. Assume that the time series has the form

$$X_t = \sum_{j=1}^{p} R_j \cos(t\lambda_j + \varphi_j) + Y_t.$$

Here $0 < \lambda_j \le \pi$; φ_j are independent and uniformly distributed on $[-\pi, \pi]$; R_j are independent, nonnegative random variables such that R_j^2 has positive mean $4\rho_j$; and Y_t is a second-order stationary time series with $E(Y_t) = E(X_0)$ and autocovariance function $\gamma_c(u) = \operatorname{cov}(Y_t, Y_{t+u})$ satisfying $\sum_u |\gamma_c(u)| < \infty$.

The spectral distribution function of X_t is given by

$$F(\lambda) = \int_{-\pi}^{\lambda} f_c(\omega)\, d\omega + \sum_{\omega \le \lambda} f_d(\omega), \qquad |\lambda| \le \pi,$$

where

$$f_c(\lambda) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} \gamma_c(u) \exp(iu\lambda), \qquad |\lambda| \le \pi,$$

and

$$f_d(\lambda) = \begin{cases} \rho_j, & \text{if } \lambda = \pm\lambda_j, \\ 0, & \text{otherwise.} \end{cases}$$

The functions f_c and f_d are referred to as the *spectral density function* and *line spectrum* of the time series X_t.

Note that f_c and f_d are nonnegative and symmetric about zero and that they can be extended to periodic functions on $(-\infty, \infty)$ with period 2π. From now on we limit our attention to the interval $[0, \pi]$. Observe that if the indicated derivatives of f_c exist, then $f_c'(0)$, $f_c'''(0)$, $f_c'(\pi)$ and $f_c'''(\pi)$ all equal zero.

8.1. *The LSPEC methodology.* Let $\delta_a(\lambda)$ equal 1 or 0 according as $\lambda = a$ or $\lambda \ne a$. Given a time series $X_1, X_2, \dots, X_{T-1}$, set $f = f_c + (T/2\pi) f_d$, $\phi = \log f$ and $\phi_c = \log f_c$. Then $\phi = \phi_c + \phi_d$, where $\phi_d = \beta_1 \delta_{\lambda_1} + \cdots + \beta_p \delta_{\lambda_p}$ with $\beta_1, \dots, \beta_p > 0$. Moreover, $f_d = (2\pi/T)(\exp \phi_d - 1) f_c$. In the following discussion, we will use cubic splines to obtain a finite-dimensional approximation to ϕ_c and hence to ϕ.

First, we describe the space of splines that will be used to model the logarithm of the spectral density function. Given the positive integer J_c, let G_c be the J_c-dimensional space of twice continuously differentiable, cubic spline functions s with the knot sequence $0 \le t_1 < \cdots < t_{J_c} \le \pi$. We require that $s'(0) = s'(\pi) = 0$. Also, $s'''(0) = 0$ unless $t_1 = 0$, and $s'''(\pi) = 0$ unless $t_{J_c} = \pi$. Let $B_1, \dots, B_{J_c}$ be a basis of G_c. Then functions in G_c can be extended to splines on $(-\infty, \infty)$ that are symmetric about zero, periodic with period 2π, have a knot at zero if and only if $t_1 = 0$ and have a knot at π if and only if $t_{J_c} = \pi$.

Next, we describe the space that will be used indirectly to model the line spectrum. Given the nonnegative integer J_d and the increasing sequence $a_1, \dots, a_{J_d}$ of members of $\{2\pi j/T\colon 1 \le j \le T/2\}$, let G_d be the J_d-dimensional space of nonnegative functions s on $[0, \pi]$ such that $s = 0$ except at $a_1, \dots, a_{J_d}$. Set $B_{j+J_c}(\lambda) = \delta_{a_j}(\lambda)$ for $1 \le j \le J_d$. Then $B_{J_c+1}, \dots, B_J$ form a basis of G_d, where $J = J_c + J_d$.

Let G be the space spanned by $B_1, \dots, B_J$. The collection $\mathscr{G}$ of such J-dimensional spaces G forms a family of allowable spaces. Set

$$\phi_c(\cdot;\boldsymbol{\beta}_c) = \beta_1 B_1(\cdot) + \cdots + \beta_{J_c} B_{J_c}(\cdot), \qquad \boldsymbol{\beta}_c = (\beta_1, \dots, \beta_{J_c}) \in \mathbb{R}^{J_c},$$

$$\phi_d(\cdot;\boldsymbol{\beta}_d) = \beta_{J_c+1} B_{J_c+1}(\cdot) + \cdots + \beta_J B_J(\cdot),$$

$$\boldsymbol{\beta}_d = (\beta_{J_c+1}, \dots, \beta_J) \quad \text{with } \beta_{J_c+1}, \dots, \beta_J \ge 0,$$

and

$$\phi(\cdot;\boldsymbol{\beta}) = \phi_c(\cdot;\boldsymbol{\beta}_c) + \phi_d(\cdot;\boldsymbol{\beta}_d), \qquad \boldsymbol{\beta} = (\beta_1, \dots, \beta_J).$$

We use $\phi_c(\cdot;\boldsymbol{\beta}_c)$ to model the logarithm of the spectral density function and $\phi(\cdot;\boldsymbol{\beta})$ to model $\log f$. Thus, $f_c(\cdot;\boldsymbol{\beta}_c) = \exp\phi_c(\cdot;\boldsymbol{\beta}_c)$, $f(\cdot;\boldsymbol{\beta}) = \exp\phi(\cdot;\boldsymbol{\beta})$ and

$$f_d(\cdot;\boldsymbol{\beta}_c) = \frac{2\pi}{T}[\exp\,\phi_d(\cdot;\boldsymbol{\beta}_d) - 1] f_c(\cdot;\boldsymbol{\beta}_c).$$

Denote the Fourier frequencies by $\lambda_k = 2\pi k/T$ for $k = 0, 1, \dots, [T/2]$. Let I_k denote the kth ordinate of the periodogram, which is given by

$$I_k = I^{(T)}(\lambda_k) = (2\pi T)^{-1} \left| \sum_{t=0}^{T-1} \exp(-i\lambda_k t) X_t \right|^2.$$

For Gaussian time series, I_k, $1 \le k \le [T/2]$, are independent and have the exponential distribution with mean equal to $f(\lambda_k) = \exp\phi(\lambda_k)$. Hence, the log-likelihood function is given by

$$l(\boldsymbol{\beta}) = \frac{1}{[T/2]} \sum_{k=1}^{[T/2]} \left(\frac{\delta_\pi(\lambda_k)}{2} - 1 \right) [\phi(\lambda_k;\boldsymbol{\beta}) + I_k \exp(-\phi(\lambda_k;\boldsymbol{\beta}))], \qquad \boldsymbol{\beta} \in \mathbb{R}^J.$$

Observe that the log-likelihood is a concave function of $\boldsymbol{\beta}$.

Let $\widehat{\boldsymbol{\beta}}$ denote the maximum likelihood estimate of $\boldsymbol{\beta}$, which is obtained as usual by the Newton–Raphson method. The corresponding estimate of the function f is given by $\widehat{f}(\lambda) = f(\lambda;\widehat{\boldsymbol{\beta}})$. Similarly, estimates of the spectral density function and line spectrum are given by $\widehat{f}_c(\cdot) = f_c(\cdot;\widehat{\boldsymbol{\beta}}_c)$ and $\widehat{f}_d(\cdot) = f_d(\cdot, \widehat{\boldsymbol{\beta}}_d)$, where $\widehat{\boldsymbol{\beta}}_c = (\widehat{\beta}_1, \dots, \widehat{\beta}_{J_c})$ and $\widehat{\boldsymbol{\beta}}_d = (\widehat{\beta}_{J_c+1}, \dots, \widehat{\beta}_J)$.

As in other cases discussed in this paper, our spectral estimate depends on G. We follow the procedure described in Section 3 (with $d = 1$) to select G adaptively from $\mathscr{G}$. This methodology is referred to as LSPEC in Kooperberg, Stone and Truong (1995c). (In the current implementation of LSPEC, if an atom has a frequency that is not of the form $2\pi k/T$, then it is typically replaced by the two closest adjacent atoms with frequencies of this form. Also, LSPEC prevents atoms with small mass from entering the model.)

In the absence of atoms, the rate of convergence of the maximum likelihood estimate $\widehat{\phi}_c$ is given in Kooperberg, Stone and Truong (1995d). This result lends theoretical support to LSPEC.

8.2. *An example.* We will use LSPEC to analyze the result of a neurophysiological experiment consisting of 30 trials of electrical potential (EP) measurements [see Durka, Kelly and Blinowska (1995)]. It started with a 24 Hz (cycles/s), 500 μm peak-to-peak sinusoidal stimulus applied to the right fingertip. The responses are the EP measurements at the scalp and wrist. Each EP measurement lasted for 6 s, with the stimulus coming on at 2 s and staying on for the remainder of the trial. The channels were sampled at 256 times/s, giving a total of 1536 sampling points per channel.

Since the stimulus was not active for the first 2 s, our analyses were based on the last 4 s of recordings, so that $T = 1024$. Figure 11 shows the averages of 30 EP responses from the scalp and wrist, which appear to be stationary. The left side of Figure 12 shows the LSPEC estimate of the scalp EP spectrum. We observe two lines with frequencies of 9.25 and 9.75 Hz [the former frequency corresponds to $k = 4(9.25) = 37$ and $\lambda = 2\pi(37)/1024 \doteq 0.227$, and the latter frequency corresponds to $k = 39$ and $\lambda \doteq 0.239$]. These are approximately the alpha-rhythm frequencies. There is also a peak with a frequency of 48 Hz ($\lambda \doteq 1.178$), corresponding to the second harmonic of the stimulus frequency 24 Hz. In the right side of Figure 12, we observe that the wrist EP responded with a

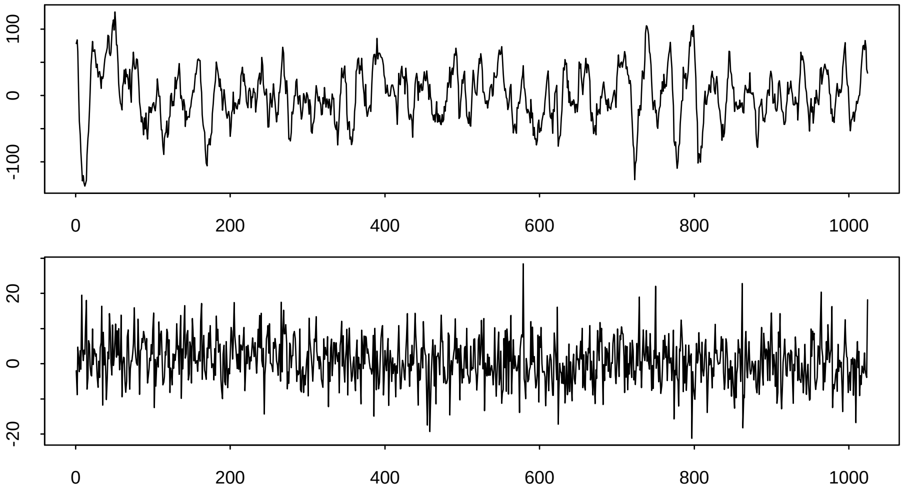


FIG. 11. *Averages of* 30 *series of electrical potential* (*EP*) *measurements from the scalp* (*top*) *and wrist* (*bottom*).

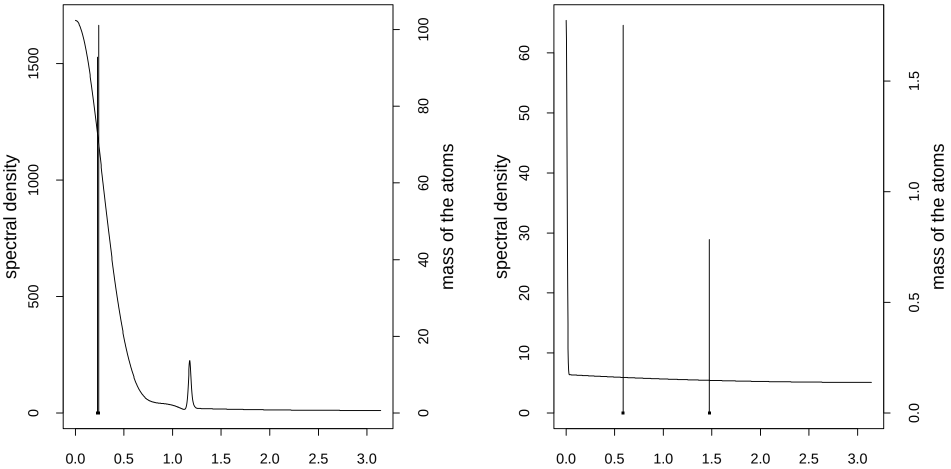


FIG. 12. *The scalp EP spectrum (left) has line frequencies equal to* 9.25 *and* 9.75 *Hz; the peak has a frequency equal to* 48 *Hz. The line frequencies of the wrist EP spectrum (right) are* 24 *and* 60 *Hz.*

frequency (the first line) at 24 Hz, while it also picked up the electrical power line frequency at 60 Hz. Note that the background noise level (the continuous spectrum) is much higher in the scalp EP than in the wrist EP.

The responses were then filtered to remove the unwanted (alpha-rhythm, electrical power line) signals and low frequency components of background noise, and sampled at 128 times/s, yielding a total of 512 sampling points. Applications of LSPEC to the filtered observations are illustrated in Figure 13.

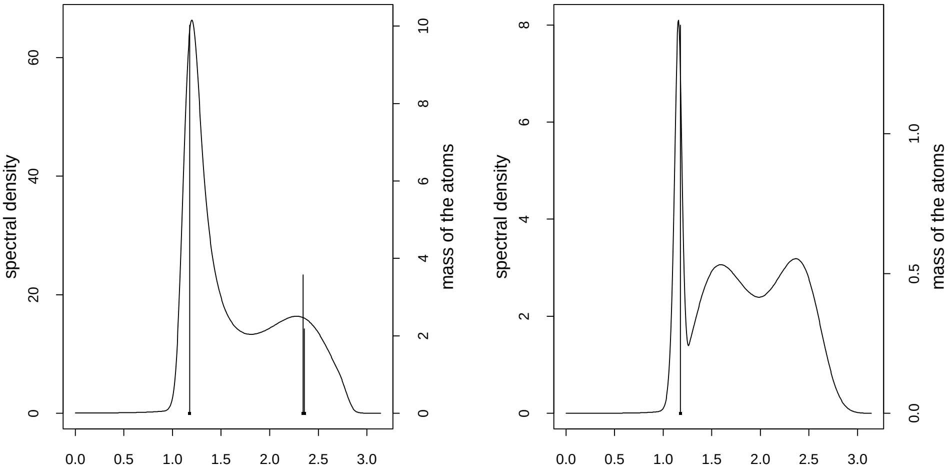


FIG. 13. *Spectra of the filtered EP data. The scalp (left) has line frequencies equal to* 24 *and* 48 *Hz. The wrist (right) has a line frequency equal to* 24 *Hz.*

For the scalp EP data, the resulting fit is a spline with seven knots and three lines in the model. The first line has a frequency of 24 Hz ($\lambda \doteq 1.178$), showing that LSPEC has located the desired signal. The other two lines correspond to the second harmonic. The fit for the wrist EP data shows a spline with eight knots and one line (at 24 Hz) in the model.

In summary, in this example the LSPEC methodology yielded a precise estimate of the stimulus frequency (24 Hz) and provided an informative description of the neurophysiological data. More generally, in the light of the present example and those given in Kooperberg, Stone and Truong (1995c), we find the LSPEC methodology to be both effective and of considerable practical value.

9. Models based on multivariate splines. In the last two decades, a considerable body of literature on multivariate spline spaces has been amassed by approximation theorists, numerical analysts and computer scientists. In this section, we demonstrate the practicality of these tools for statistical applications. We begin our survey on a theoretical note, developing rates of convergence for ANOVA decompositions based on multivariate splines and their tensor products. Then we shift our emphasis somewhat and consider techniques for adaptively constructing multivariate spline spaces, borrowing heavily from the ideas of knot addition and deletion presented in previous sections. Finally, we present a simple illustrative application of these ideas to bivariate logspline density estimation.

9.1. *The extended linear model revisited.* In Section 2, we introduced the notion of a concave extended linear model and discussed a variety of statistical problems that can be treated effectively within this framework. In each of these cases, our data consist of a sample from the distribution of a random vector **W**. In this section, we focus our attention on the derived variable **U**, which is typically a subvector of **W**. Broadly speaking, we are interested in estimating a (possibly) vector-valued function $\phi^* = (\phi_1^*, \dots, \phi_K^*)$, where the constituents ϕ_k^*, $1 \le k \le K$, are real-valued functions on a set $\mathscr{U} = \mathscr{U}_1 \times \cdots \times \mathscr{U}_M$, the range of **U**. So far, we have considered only the case in which each of the sets $\mathscr{U}_1, \dots, \mathscr{U}_M$ is (in theory) a compact interval with positive length. Under this restriction, we are naturally led to estimators of ϕ^* that are built up from univariate spline spaces defined on these intervals. From a methodological perspective, however, tensor products of univariate splines may not be flexible enough to capture all of the features exhibited by a particular data set. In addition, known structural relationships between the variables that constitute **U** might suggest that the domain of ϕ^* is something other than a hyperrectangle.

In the rest of our discussion, we allow $\mathscr{U}_1, \dots, \mathscr{U}_M$ to be compact subsets of $\mathbb{R}^{d_1}, \dots, \mathbb{R}^{d_M}$, respectively. In this case, the unknown function $\phi^* = \phi^*(u_1, \dots, u_M)$ is still defined on $\mathscr{U} = \mathscr{U}_1 \times \cdots \times \mathscr{U}_M$, with the distinction that now the individual variables u_m may be vectors. Recall that our approach to estimating $\phi^* \in H^K$ begins with an ANOVA decomposition $\phi^* = \sum_{s \in \mathscr{S}} \phi_s^*$ that decomposes ϕ^* into its components ϕ_s^*, $s \in \mathscr{S}$. A parallel construction is then used to define an ANOVA decomposition of the maximum likelihood estimate $\widehat{\phi} = \sum_{s\in\mathscr{S}} \widehat{\phi}_s$ in a space G^K consisting of smooth, piecewise polynomials. Not surprisingly, this approach can successfully be applied to derive the convergence properties of $\widehat{\phi}$ even when we allow the sets $\mathscr{U}_1, \dots, \mathscr{U}_M$ to be more complicated than compact intervals of the real line. Once we remove these restrictions, the components $\widehat{\phi}_s$, $s \in \mathscr{S}$, of the ANOVA decomposition of $\widehat{\phi}$ become *multivariate splines* and their tensor products.

To be more specific, for $1 \le m \le M$, let $\triangle_m$ be a partition of $\mathscr{U}_m \subset \mathbb{R}^{d_m}$ into disjoint (measurable) sets and for simplicity assume that each set has a common diameter a. By a piecewise polynomial of degree q over $\triangle_m$, we now mean a function g on $\mathscr{U}_m$ such that the restriction of g to each set $\delta \in \triangle_m$ is a polynomial of degree q in the d_m variables that constitute u_m. Let G_m be a linear space of *multivariate splines*; that is, piecewise polynomials of degree q on $\mathscr{U}_m$ that satisfy certain smoothness constraints. Following the development in Section 2, for each $s \in \mathscr{S}$, we let G_s denote the tensor product of the spaces G_m, $m \in s$.

The rate at which $\widehat{\phi}$ and its components approach ϕ^* and its components were derived in Hansen (1994). In the simple case described so far, if we assume that the spaces G_s are flexible enough to ensure that

$$\inf_{g\in G_s} \|g - \phi^*_{ks}\|_\infty = O(a^p), \qquad 1 \le k \le K \text{ and } s \in \mathscr{S},$$

where p is a measure of smoothness of the constituents of ϕ^*, we find that

$$\|\widehat{\phi}_s - \phi^*_s\|^2 = O_P\left(a^{2p} + \frac{1}{na^d}\right), \qquad s \in \mathscr{S},$$

and

$$\|\widehat{\phi} - \phi^*\|^2 = O_P\left(a^{2p} + \frac{1}{na^d}\right),$$

where $d = \max_{s\in\mathscr{S}} \sum_{m\in s} d_m$. As we collect more and more data, if the sets in our partition shrink so that $a \sim n^{-1/(2p+d)}$, then we obtain the rates in (2.3) and (2.4) with the indicated definition of d. Hansen (1994) extended these results and, in particular, derived L_2 rates of convergence for the case when the various constituents ϕ^*_s satisfy different smoothness conditions and the sets in the triangulations $\triangle_m$ do not share a common diameter.

9.2. *Bivariate splines and the extended linear model.* For simplicity, we now focus our discussion on saturated, bivariate models, where $\phi^* = \phi$. Assume that $\mathscr{U}$ is a compact region in the plane so that ϕ is a function of $\mathbf{u} \in \mathbb{R}^2$. In the context of our previous discussion, we now view $\mathbf{U}$ as a single variable and hence will not attempt to decompose ϕ into components based on individual spatial coordinates. In the remaining pages, we will discuss the use of bivariate splines to construct estimates of ϕ.

Triangulations and piecewise linear basis functions. Let $\triangle$ be a collection of closed subsets of $\mathscr{U}$ having disjoint interiors and satisfying $\mathscr{U} = \bigcup_{\delta\in\triangle} \delta$. In general, the set $\triangle$ is a tessellation of $\mathscr{U}$. If each element $\delta \in \triangle$ is a triangle, $\triangle$ is said to form a triangulation of $\mathscr{U}$. Furthermore, a triangulation $\triangle$ is said to be *conforming* if the nonempty intersection between pairs of triangles in $\triangle$ consists of either a single shared vertex or an entire common edge (see Figure 14). Throughout this section, we reserve the symbol $\triangle$ for this special type of tessellation.

Given such a conforming triangulation $\triangle$, we let G denote the space of continuous, piecewise linear functions over $\triangle$. There is a natural association between the vertices $\mathbf{v}_1, \dots, \mathbf{v}_J$ of the triangles in $\triangle$ and the basis functions $B_1(\mathbf{u}), \dots, B_J(\mathbf{u})$ of G. To be more precise, we define $B_j(\mathbf{u})$ to be the unique function that is linear on each of the triangles in $\triangle$ and takes on the value 1 at $\mathbf{v}_j$ and 0 at the remaining vertices in the partition. This collection of tent functions is frequently used in the finite element method and is often the starting point for defining multivariate splines of higher degrees [see Chui (1988), de Boor (1987) and Farin (1986)].

Many of the important properties of this basis can be obtained from a local representation of the tent functions. For the moment, consider a single triangle $\delta \in \triangle$ having vertices $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$. Relative to δ, the *barycentric coordinates* of any point $\mathbf{u} = (u_1, u_2) \in \mathbb{R}^2$ are defined as a triple $\varphi(\mathbf{u}) = (\varphi_1(\mathbf{u}), \varphi_2(\mathbf{u}), \varphi_3(\mathbf{u}))$ such that

$$\mathbf{u} = \varphi_1(\mathbf{u})\mathbf{v}_1 + \varphi_2(\mathbf{u})\mathbf{v}_2 + \varphi_3(\mathbf{u})\mathbf{v}_3 \quad \text{and} \quad \varphi_1(\mathbf{u}) + \varphi_2(\mathbf{u}) + \varphi_3(\mathbf{u}) = 1.$$

Casting these conditions into a simple set of linear equations we find that

$$\begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \varphi_1(\mathbf{u}) \\ \varphi_2(\mathbf{u}) \\ \varphi_3(\mathbf{u}) \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ 1 \end{pmatrix}. \tag{9.1}$$

Provided that δ has a nonempty interior, this system can be solved explicitly, and the solution is best written in terms of the function SignedArea$(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$,

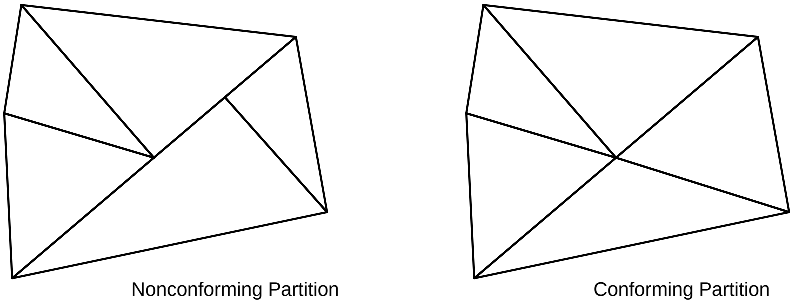


FIG. 14. *In a nonconforming partition, at least one vertex of a triangle in $\triangle$ falls along the interior of an edge of another triangle in the partition.*

which we define by

$$\mathrm{SignedArea}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = \tfrac{1}{2}\begin{vmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ 1 & 1 & 1 \end{vmatrix}.$$

As its name suggests, the absolute value of $\mathrm{SignedArea}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ is just the area of the triangle with vertices $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$. By applying Cramér's method to the set of equations (9.1) we find that $\varphi_1(\mathbf{u})$ is given by the ratio

$$\varphi_1(\mathbf{u}) = \varphi_1(u_1, u_2) = \frac{\mathrm{SignedArea}(\mathbf{u}, \mathbf{v}_2, \mathbf{v}_3)}{\mathrm{SignedArea}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)}. \tag{9.2}$$

Thus, the barycentric coordinates are linear functions of u_1 and u_2, where $\mathbf{u} = (u_1, u_2)$, and satisfy the interpolation conditions

$$\varphi_i(\mathbf{v}_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases} \qquad i, j = 1, 2, 3; \tag{9.3}$$

hence the vertices $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ have barycentric coordinates $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, respectively. Furthermore, from (9.2) we see that the points on the edge connecting $\mathbf{v}_2$ and $\mathbf{v}_3$ have barycentric coordinates of the form $(0, \alpha, 1-\alpha)$, $\alpha \in [0, 1]$.

Given the interpolation conditions (9.3) and the consequence of (9.2) that the barycentric coordinate functions are linear functions of $\mathbf{u}$, we now have an explicit representation of the basis functions of G that correspond to the vertices of δ; that is, for all $\mathbf{u} \in \delta$, $B_i(\mathbf{u}) = \varphi_i(\mathbf{u})$, $i = 1, 2, 3$. As an immediate consequence of this local (triangle by triangle) representation, we find that the basis functions $B_1, \dots, B_J$ associated with the triangulation $\triangle$ are bounded between 0 and 1 and satisfy

$$B_1(\mathbf{u}) + \cdots + B_J(\mathbf{u}) = 1, \qquad \mathbf{u} \in \mathscr{U}.$$

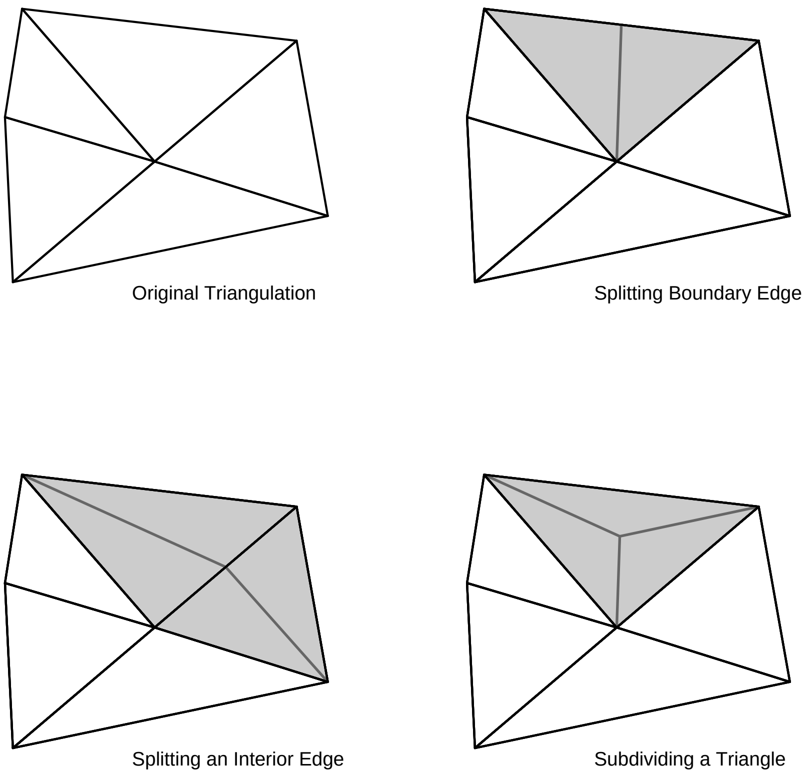
From (9.2) it is also possible to demonstrate that, for any nonsingular, 2-by-2 matrix A and any vector $\mathbf{b} \in \mathbb{R}^2$,

$$B_j(\mathbf{u}) = B_j^*(A\mathbf{u} + \mathbf{b}), \qquad \mathbf{u} \in \mathbb{R}^2,$$

where $B_1^*, \dots, B_J^*$ is the basis associated with vertices $A\mathbf{v}_1 + \mathbf{b}, \dots, A\mathbf{v}_J + \mathbf{b}$ of the transformed set $\mathscr{U}^* = \{A\mathbf{u} + \mathbf{b}, \mathbf{u} \in \mathscr{U}\}$. This means that models built from functions in G have a natural invariance under affine transformations. Using the barycentric coordinate functions, we will see in the next subsection that this invariance carries over to our adaptive methodology as well.

To summarize, we have derived some of the essential properties of a basis for the space of continuous, piecewise linear functions associated with a triangulation $\triangle$ of $\mathscr{U}$. An important observation here is that there is a simple correspondence between the structure of the partition $\triangle$ and the basis functions of G. As in the previous sections, this relationship will allow us to use simple model selection criteria to construct a functional form of our estimate $\widehat{\phi}$ of the unknown function ϕ. The only issue left to resolve is how we generalize the notion of stepwise addition and deletion of knots in this context.

Stepwise addition. The most natural way to proceed from one step to the next in the stepwise addition procedure is to introduce a new vertex into the existing triangulation, thereby adding one new basis function to the existing spline space. This operation requires a rule for connecting this point to the vertices in $\triangle$ so that the new mesh is also a conforming triangulation. In Figure 15, we illustrate three options for vertex addition: we can place a new vertex on either a boundary or an interior edge, splitting the edge, or we can add a point to the interior of one of the triangles in $\triangle$. Note that the space obtained by adding a vertex $\mathbf{v}$ to an interior edge of a triangle $\delta \in \triangle$ cannot be achieved as the limit of spaces constructed by adding $\mathbf{v}$ to the interior of δ. In this case, if $\mathbf{v}$ is very close to an edge of δ, the new triangulation is essentially nonconforming and the associated space of linear functions G contains elements that are discontinuous along that edge. Similar discontinuities arise when the new point $\mathbf{v}$ is positioned extremely close to an existing vertex. Degeneracies such as these are encountered in the context of univariate spline spaces when knots are allowed to coalesce [de Boor (1978)].



Original Triangulation

Splitting Boundary Edge

Splitting an Interior Edge

Subdividing a Triangle

FIG. 15. *Three ways to add a new vertex to an existing triangulation. Each addition represents the introduction of a single basis function, the support of which is colored gray.*

Given a triangulation $\triangle$, we construct a set of candidate vertices by considering the points with barycentric coordinates

$$\left(\frac{k_1}{K+1}, \frac{k_2}{K+1}, \frac{K+1-k_1-k_2}{K+1}\right)_\delta, \qquad \delta \in \triangle, \tag{9.4}$$

where k_1, k_2 and K are nonnegative integers satisfying $k_1 + k_2 \leq K + 1$ and no coordinate equals 1. We have introduced a subscript δ to make it clear that these points are calculated for each triangle in $\triangle$. At each step in the addition process, we select from this set of candidate vertices the point that maximizes the Rao statistic described in Section 3. Stability considerations may dictate that we do not consider for addition vertices in areas where there is little data. Moreover, we have found it useful to avoid creating triangles having one or two very small angles. Restrictions such as these are easily incorporated into the stepwise addition procedure.

Stepwise deletion. There are two possible strategies for reducing the dimension of an existing piecewise linear spline space. In each case, we enforce the condition that a function in the space be continuously differentiable across a given edge in the existing triangulation. Observe that a continuous, piecewise linear function has continuous partial derivatives across an edge if and only if the function is linear on the union of the two triangles that share the edge. Using the correspondence between vertices and basis functions described above, we can show that the subspace of spline functions satisfying this condition is characterized by a simple linear constraint of the type discussed in Section 3. In each of the examples in Figure 15, enforcing continuity of the first partial derivatives across any of the gray edges is equivalent to removing the added vertex, returning us to the original partition in the upper left corner of the figure. Thus, in light of the stepwise knot deletion strategy discussed in the previous sections, one procedure for stepwise deletion in the bivariate context involves using the Wald statistic to choose between continuity constraints across edges that fall into one of the three categories listed in Figure 15. An alternative deletion procedure is somewhat more aggressive and involves choosing from among all the continuity constraints, regardless of how the edge is positioned relative to the other edges in the partition. The important distinction between these two procedures is that only in the first case are we actually guaranteed that the structure of $\triangle$ is simplified at each step.

9.3. *Bivariate logspline density estimation.*

Maximum likelihood estimation. While the bivariate methodology introduced in the previous paragraphs has been implemented for a variety of extended linear models, we will focus mainly on logspline density estimation. In this context, we choose to model the logarithm of an unknown density ϕ of a random vector $\mathbf{U}$ as a bivariate spline. For ease of presentation, we restrict our attention to densities that are supported on a simply connected region $\mathcal{U} \in \mathbb{R}^2$ having a polygonal boundary. As usual, let $\triangle$ denote a conforming partition of

$\mathscr{U}$ and let $B_1(\mathbf{u}), \ldots, B_J(\mathbf{u})$ denote the basis functions of the corresponding space G of continuous, piecewise linear functions over $\triangle$.

Given a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J) \in \mathbb{R}^J$, we can define a density $f(\mathbf{u}; \boldsymbol{\beta})$ over $\mathscr{U}$ having the form

$$f(\mathbf{u}; \boldsymbol{\beta}) = \exp\bigl(\beta_1 B_1(\mathbf{u}) + \cdots + \beta_J B_J(\mathbf{u}) - C(\boldsymbol{\beta})\bigr),$$

where

$$C(\boldsymbol{\beta}) = \log \int_{\mathscr{U}} \exp\bigl(\beta_1 B_1(\mathbf{u}) + \cdots + \beta_J B_J(\mathbf{u})\bigr)\, d\mathbf{u}$$

is the normalizing constant. Based on a random sample $\mathbf{U}_1, \ldots, \mathbf{U}_n$ from the distribution of $\mathbf{U}$, we estimate ϕ by the function $\widehat{\phi} = f(\mathbf{u}; \widehat{\boldsymbol{\beta}})$, where $\widehat{\boldsymbol{\beta}}$ maximizes the "log-likelihood" $l(\boldsymbol{\beta}) = \log f(\mathbf{U}_1; \boldsymbol{\beta}) + \cdots + \log f(\mathbf{U}_n; \boldsymbol{\beta})$. While we do not believe that $l(\cdot)$ is the true log-likelihood function corresponding to our sample, we know from the discussion at the beginning of this section that as $n \to \infty$, $\widehat{\phi}$ tends to ϕ.

As in univariate logspline density estimation (see Section 4), the likelihood equations take on the very simple form

$$E_{\boldsymbol{\beta}} B_j(\mathbf{U}) = E_n B_j(\mathbf{U}), \qquad 1 \le j \le J, \tag{9.5}$$

where

$$E_{\boldsymbol{\beta}} B_j(\mathbf{U}) = \int_{\mathscr{U}} B_j(\mathbf{u}) f(\mathbf{u}; \boldsymbol{\beta})\, d\mathbf{u} \quad \text{and} \quad E_n B_j(\mathbf{U}) = \frac{1}{n} \sum_{i=1}^{n} B_j(\mathbf{U}_i).$$

Since the functions B_j are piecewise linear over $\mathscr{U}$, it is possible to evaluate the required integrals exactly. As in previous sections, the equations in (9.5) are solved using Newton–Raphson iterations. To obtain the Hessian matrix required for this procedure, we must also calculate expressions of the form $E_{\boldsymbol{\beta}}[B_{j_1}(\mathbf{U}) B_{j_2}(\mathbf{U})]$ for $1 \le j_1, j_2 \le J$. Since the basis functions are piecewise linear, however, we again do not require numerical quadrature to carry out these computations.

Implementing stepwise addition and deletion. Recall that we add basis functions to G by adding vertices to $\triangle$ and that our strategy for choosing between the competing basis functions is based on the heuristic maximization of Rao statistics. This process can be simplified considerably by making explicit use of the barycentric coordinate functions discussed above. For example, suppose that we want to add a node $\mathbf{v}$ inside δ, the right-hand triangle in Figure 16. Once again, suppose that δ has vertices $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ and let $\varphi_1(\mathbf{u})$, $\varphi_2(\mathbf{u})$ and $\varphi_3(\mathbf{u})$ denote the barycentric coordinates of a point $\mathbf{u} \in \mathbb{R}^2$ relative to δ. Now, if we let $B_1(\mathbf{u})$, $B_2(\mathbf{u})$ and $B(\mathbf{u})$ represent the piecewise linear basis functions associated with the points $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}$ in the updated triangulation, then it is straightforward to demonstrate that, for all points $\mathbf{u}$ in the shaded triangle on the right in Figure 16,

$$\varphi_1(\mathbf{u}) = B_1(\mathbf{u}) + \varphi_1(\mathbf{v}) B_3(\mathbf{u}), \qquad \varphi_2(\mathbf{u}) = B_2(\mathbf{u}) + \varphi_2(\mathbf{v}) B_3(\mathbf{u}) \quad \text{and}$$

$$\varphi_3(\mathbf{u}) = \varphi_3(\mathbf{v}) B_3(\mathbf{u}).$$

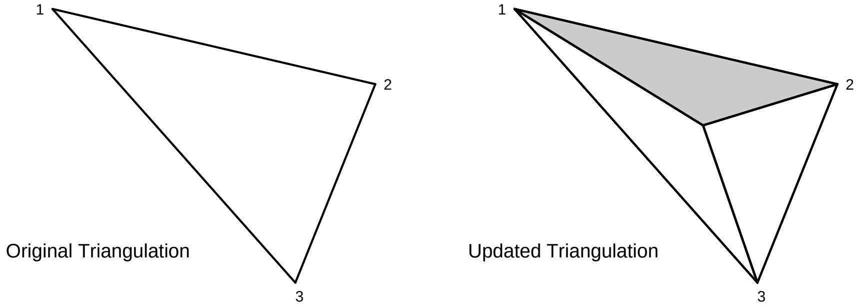


FIG. 16. *Adding a new vertex at the point* $\mathbf{v} = \varphi_1(\mathbf{v})\mathbf{v}_1 + \varphi_2(\mathbf{v})\mathbf{v}_2 + \varphi_3(\mathbf{v})\mathbf{v}_3$. *In this case, we are adding to G the continuous, piecewise linear function that takes on the value* 1 *at the point* $\mathbf{v}$ *and* 0 *at each of* $\mathbf{v}_1$, $\mathbf{v}_2$ *and* $\mathbf{v}_3$.

Combining these relationships with the fact that within δ, the piecewise linear basis functions associated with $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ are exactly the barycentric coordinate functions relative to δ, we arrive at simple formulas for calculating the necessary inner products and empirical moments that go into forming the Rao statistic for adding $\mathbf{v}$ to the partition $\triangle$. Similar expressions can be derived for evaluating the candidate function over the remaining two triangles in the right plot of Figure 16. In the numerical example discussed below, we introduce vertices at the points corresponding to $K = 5$ in expression (9.4).

Using these ideas, we can also derive a simple procedure for determining the constraint that a function in G be continuously differentiable across a given edge in $\triangle$. To make this more precise, consider the triangulation on the left in Figure 17 and let $\varphi_1(\mathbf{u})$, $\varphi_2(\mathbf{u})$ and $\varphi_3(\mathbf{u})$ denote the barycentric coordinates of a point $\mathbf{u} \in \mathbb{R}^2$ relative to the triangle with vertices $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$. Given a function $g \in G$, let β_1, β_2 and β_3 denote the coefficients of the basis functions associated with these vertices. Then for all points $\mathbf{u}$ in this triangle, $g(\mathbf{u})$ is the linear function given by $\beta_1\varphi_1(\mathbf{u}) + \beta_2\varphi_2(\mathbf{u}) + \beta_3\varphi_3(\mathbf{u})$. Now, if we let β_4 denote the coefficient of the basis function of G associated with the vertex $\mathbf{v}_4$, then $g(\mathbf{v}_4) = \beta_4$. Therefore, the function g is linear on the

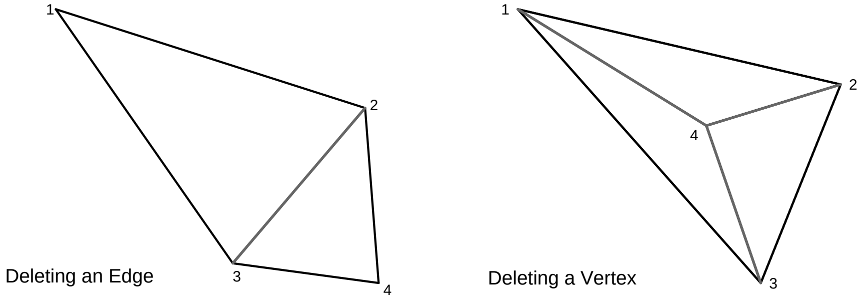


FIG. 17. *The effect of enforcing the constraint that functions in G be continuously differentiable across edges in two triangulations.*

union of the two triangles in the left portion of Figure 17 provided that

$$\beta_4 = g(\mathbf{v}_4) = \beta_1\varphi_1(\mathbf{v}_4) + \beta_2\varphi_2(\mathbf{v}_4) + \beta_3\varphi_3(\mathbf{v}_4).$$

By swapping the roles of $\mathbf{v}_1$ and $\mathbf{v}_4$ in this argument, we find that C^1 continuity of a function $g \in G$ can also be assured by the constraint

$$\beta_1 = g(\mathbf{v}_1) = \beta_2\tilde{\varphi}_2(\mathbf{v}_1) + \beta_3\tilde{\varphi}_3(\mathbf{v}_1) + \beta_4\tilde{\varphi}_4(\mathbf{v}_1),$$

where $\tilde{\varphi}_2(\mathbf{u})$, $\tilde{\varphi}_3(\mathbf{u})$ and $\tilde{\varphi}_4(\mathbf{u})$ denote the barycentric coordinates of a point $\mathbf{u}$ relative to the triangle with vertices $\mathbf{v}_2$, $\mathbf{v}_3$ and $\mathbf{v}_4$. It is not hard to demonstrate that these two constraints are equivalent up to a multiplicative constant. Observe, however, that when this condition is enforced, we are left with a single linear function over the pair of triangles that constitute $\triangle$, but we have not produced a simpler triangulation in the process.

Suppose instead that we want to remove the vertex $\mathbf{v}_4$ in the middle of the triangle in the right portion of Figure 17. Given $g \in G$ and $1 \leq i \leq 4$, we again let β_i correspond to the coefficient of the basis function associated with the vertex $\mathbf{v}_i$. It can be shown that each of the C^1 continuity constraints across the shaded interior edges shown in the figure is of the form

$$\beta_4 = \varphi_1(\mathbf{v}_4)\beta_1 + \varphi_2(\mathbf{v}_4)\beta_2 + \varphi_3(\mathbf{v}_4)\beta_3, \tag{9.6}$$

where $\varphi_1(\mathbf{u})$, $\varphi_2(\mathbf{u})$ and $\varphi_3(\mathbf{u})$ are the barycentric coordinates of a point $\mathbf{u}$ relative to the outer triangle in Figure 17. Observe that the expression on the left is the value at $\mathbf{v}_4$ of the unique linear function interpolating β_1, β_2 and β_3 at the points $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$, respectively. Recalling that $g(\mathbf{v}_4) = \beta_4$, we see that the constraint in (9.6) has considerable intuitive appeal.

9.4. *An example.* We end our discussion of bivariate logspline density estimation with an example suggested to us by Karl Broman. The points in the left panel of Figure 18 represent a collection of amino acids obtained from 100 protein structures taken from the Brookhaven Protein Data Bank [see Hobohm, Scharf, Schneider and Sander (1992)]. In order to characterize the *local environment* of each amino acid within a given protein structure, three pieces of information were recorded: the local structure of the protein at the given amino acid (whether the protein is twisting around a helix, for example), the fraction of the amino acid side-chain area that is buried in the protein structure and the fraction of the side-chain area that is covered by polar atoms. Because the unburied portion of the amino acid is exposed to a polar solvent, the final two quantities are restricted to the upper triangle of the unit square. In Figure 18, for example, we plot these two measurements for all of the occurrences of the amino acid lysine for which the local protein structure is a helix.

Bivariate density estimates computed for each amino acid and each local protein structure are the basis for an approach to solving the so-called inverse folding problem [see Bowie, Luthy and Eisenberg (1991) and Zhang and Eisenberg (1994)]. Evaluating the structure of a given protein is extremely difficult. Determining the sequence of amino acids that comprise the protein,

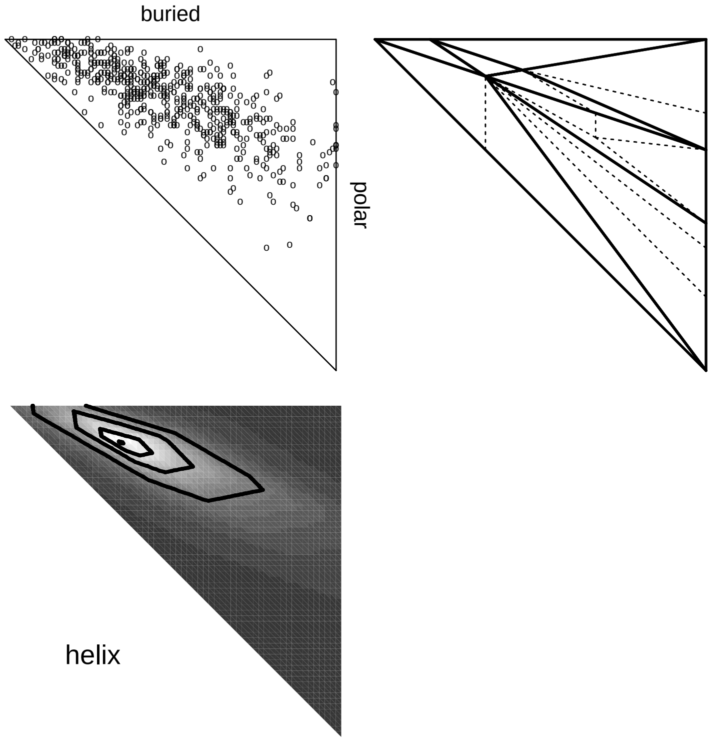


FIG. 18. *Applying the density estimation routine. In the top row we present the data and both the triangulation obtained from stepwise addition* (*thin*, *dashed line*) *and that obtained from stepwise deletion* (*thick*, *solid line*). *In the bottom row we present the data along with a contour plot of the final fit from the deletion process.*

however, is relatively simple. It would seem reasonable, therefore, to attempt to infer the protein's structure from its amino acid sequence. Unfortunately, many rather different sequences produce very similar structures, so the objective of the inverse folding problem is to determine which amino acid sequences might result in a given known structure. This can be accomplished by studying the propensity for certain amino acids to occur in certain local environments in a large collection of known protein structures. The procedure described by Zhang and Eisenberg involves a log-odds calculation, the main ingredient of which is a set of bivariate density estimates for the type of data given in Figure 18.

In the bottom panel of Figure 18, we present a contour plot of the density estimate obtained by stepwise addition followed by stepwise deletion. The model shown was encountered during stepwise deletion and attains the minimum BIC value among all the models obtained during both the stepwise addition and deletion processes. During this process, we selected candidate knots corresponding to $K = 5$ in (9.4), and did not consider any new vertices that would result in a triangle containing fewer than 25 points. In the panel on the upper right in the same figure, we present the final triangulation along with dashed edges to indicate the additional structure present when the stepwise deletion process began. The fits as well as the various plots in Figure 18 were produced using a library of S/S-PLUS routines that are available from Hansen.

In this section we have introduced a method for bivariate density estimation using piecewise linear, bivariate splines based on an adaptively constructed triangulation. We have also implemented this procedure for both regression and generalized regression. The resulting estimates, which we have named Triograms, have performed well on a variety of of bivariate data sets taken from a number of different estimation contexts. The interested reader is referred to Hansen, Kooperberg and Sardy (1996), where Triograms are compared to several existing function estimation routines. One advantage that Triograms have over these other methods is that the entire estimation procedure is invariant under affine transformations and is the most natural approach for modeling data when the domain of the predictor variables is a polygonal region in the plane. As anticipated by the convergence rate derived at the beginning of this section, if our underlying function ϕ is smooth, piecewise linear estimates are suboptimal. This problem can be corrected by using higher-order splines, and we are currently investigating how to extend the Triogram procedure to make use of the generalized vertex splines of Chui and He (1990).

REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

BELSEY, D. A., KUH, E. and WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

BOURLARD, H. A. and MORGAN, N. (1994). *Connectionist Speech Recognition*. Kluwer, Boston.

BOWIE, J. U., LUTHY, R. and EISENBERG, D. (1991). A method to identify protein sequences that fold into a known 3-dimensional structure. *Science* **253** 164–170.

BREIMAN, L. (1993). Fitting additive models to regression data. *Comput. Statist. Data Anal.* **15** 13–46.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

BRESLOW, N. E. (1972). Contribution to the discussion on the paper by D. R. Cox, Regression and life tables. *J. Roy. Statist. Soc. Ser. B* **34** 216–217.

BRESLOW, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30** 89–99.

BRILLINGER, D. R. (1981). *Time Series: Data Analysis and Theory.* Holden-Day, San Francisco.

CHUI, C. K. (1988). *Multivariate Splines*. SIAM, Philadelphia.

CHUI, C. K. and HE, T. (1990). Bivariate C^1 quadratic finite elements and vertex splines. *Math. Comp.* **54** 169–187.

COLE, R., NOEL, M., BURNETT, D. C., FANTY, M., LANDER, T., OSHIKA, B. and SUTTON, S. (1994). Corpus development activities at the Center for Spoken Language Understanding. Technical report, CSLU, Portland, OR.

COLE, R. A., ROGINSKI, K. and FANTY, M. (1992). A telephone speech database of spelled and spoken names. In *Proceedings of the International Conference on Spoken Language Processing* 891–893. Quality Color Press, Univ. Alberta, Edmonton.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.

DE BOOR, C. (1987). B-form basics. In *Geometric Modeling* (G. Farin, ed.) 131–148. SIAM, Philadelphia.

DURKA, P. J., KELLY, E. F. and Blinowska, K. J. (1995). Time-frequency analysis of stimulus-driven EEG activity by matching pursuit. In *Proceedings of the 18th Annual Conference of the IEEE EMBS, Amsterdam, October 31–November 3, 1996*. IEEE, New York.

FAMILY EXPENDITURE SURVEY (1968–1983). Annual base tapes and reports (1968–1983). Dept. Employment, Statistics Division, Her Majesty's Stationary Office, London.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

FARIN, G. (1986). Triangular Bernstein–Bézier patches. *Comput. Aided Geom. Design* **3** 83–127.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

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

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–39.

GAUVAIN, J. L., LAMEL, L. F., ADDA, G. and ADDA-DECKER, M. (1994). Speaker-independent continuous speech dictation. *Speech Communication* **15** 21–37.

GRAY, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Amer. Statist. Assoc.* **87** 942–951.

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.

GU, G. and WAHBA, G. (1993). Smoothing spline ANOVA with component-wise Bayesian "confidence intervals." *J. Comput. Graph. Statist.* **2** 97–117.

HANSEN, M. (1994). Extended linear models, multivariate splines and ANOVA. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

HANSEN, M., KOOPERBERG, C. and SARDY, S. (1996). Triograms models. *J. Amer. Statist. Assoc.* To appear.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

HERMANSKY, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America* **87** 1738–1752.

HOBOHM, U., SCHARF, M., SCHNEIDER, R. and SANDER, C. (1992). Selection of representative protein data sets. *Protein Science* **1** 409–417.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

KENNEDY, W. J. and GENTLE, J. E. (1980). *Statistical Computing*. Dekker, New York.

KOOPERBERG, C., BOSE, S. and STONE, C. J. (1997). Polychotomous regression. *J. Amer. Statist. Assoc.* **92** 117–127.

KOOPERBERG, C. and STONE, C. J. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.

KOOPERBERG, C. and STONE, C. J. (1992). Logspline density estimation for censored data. *J. Comput. Graph. Statist.* **1** 301–328.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995a). Hazard regression. *J. Amer. Statist. Assoc.* **90** 78–94.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995b). The L_2 rate of convergence for hazard regression. *Scand. J. Statist.* **22** 143–157.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995c). Logspline estimation of a possibly mixed spectral distribution. *J. Time Ser. Anal.* **16** 359–388.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995d). Rate of convergence for logspline spectral density estimation. *J. Time Ser. Anal.* **16** 389–401.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MILLER, R. G. (1981). *Survival Analysis*. Wiley, New York.

PARZEN, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* **74** 105–131.

RABINER, L. and JUANG, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

SLEEPER, L. A. and HARRINGTON, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *J. Amer. Statist. Assoc.* **85** 941–949.

SMITH, P. L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. Report NASA 166034, NASA, Langley Research Center, Hampton, VA.

SOLVD INVESTIGATORS (1990). Studies of left ventricular dysfunction (SOLVD)—rationale, design, and methods: two trials that evaluate the effect of enalapril in patients with reduced ejection fraction. *American Journal of Cardiology* **6** 315–322.

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. (1990). Large-sample inference for log-spline models. *Ann. Statist.* **18** 717–741.

STONE, C. J. (1991). Asymptotics for doubly flexible logspline response models. *Ann. Statist.* **19** 1832–1854.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

STONE, C. J. and KOO, C.-Y. (1986a). Additive splines in statistics. In *Proceedings of the Statistical Computing Section* 45–48. Amer. Statist. Assoc., Alexandria, VA.

STONE, C. J. and KOO, C.-Y. (1986b). Logspline density estimation. *Contemp. Math.* **59** 1–15.

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

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

WAND, M. P., MARRON, S. J. and RUPPERT, D. (1991). Transformations in density estimation (with discussion). *J. Amer. Statist. Assoc.* **86** 343–361.

ZHANG, K. and EISENBERG, D. (1994). The three-dimensional profile method using residue preference as a continuous function of residue environment. *Protein Science* **3** 687–695.

CHARLES J. STONE
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720-3860
E-MAIL: stone@stat.berkeley.edu

MARK H. HANSEN
BELL LABORATORIES
700 MOUNTAIN AVE., RM. 2C260
MURRAY HILL, NEW JERSEY 07030

CHARLES KOOPERBERG
FRED HUTCHSONSON CANCER RESEARCH CENTER
1100 FAIRVIEW AVE., MP 1002
SEATTLE, WASHINGTON 98109-1024

YOUNG K. TRUONG
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-7400
E-MAIL: truong@stat.unc.edu

DISCUSSION

JIANQING FAN

University of North Carolina, Chapel Hill
and Chinese University of Hong Kong

I would like to congratulate Stone, Hansen, Kooperberg and Truong for successfully outlining an ingenious principle on flexible statistical modeling. This principle is convincingly and successfully applied to a wide array of statistical problems. The availability of the software allows users to easily explore subtle nonlinear structure, which was unthinkable a decade ago. My perception is that the principle outlined in this seminal work will be widely used.

I like the name "extended linear modeling." In addition to the reasons given in the Introduction, the extended linear modeling emphasizes its continuity to parametric linear models and spells out clearly that the boundary between nonparametric and parametric modeling is moot. I heard a misperception that nonparametric modeling requires a very large amount of data. In Figure 2, the authors demonstrated that the approach can still be very useful when the number of observations is small, bearing in mind that there are many models that are indistinguishable for a small sample. The appeal of data-analytic (nonparametric) modeling is to reduce modeling biases via enhancing modeling flexibility. A satisfactory model should trade off the balance between the flexibility and estimability.

1. Basis mining. The basic idea of Stone's school stems from the variable selection of linear models. The concavity of likelihood modeling is stressed. The innovation is the use of the Rao statistics for adding variables and the Wald statistics for deleting variables. Using these techniques along with the celebrated concept of allowable spaces for interpretability, the intensive basis mining is avoided via some additional heuristics. The triumph of the algorithm is that it makes large and nearly ill-conditioned computing problems feasible.

I am puzzled why such a heuristic approach works out so well. I wonder whether the initial placement of equally spaced knots can be too reluctant to be deleted so that even better knots cannot be recruited. Why does the stepwise addition algorithm not start with zero initial knots in the program such as LOGSPLINE? There is always a risk that with poor choice in initial knots, poorly recruited new knots and additional errors of the knots deletion process, a suboptimal selection of bases is obtained. The traditional stepwise (addition and deletion) algorithms in the linear models can be used to improve the knot selection process. This algorithm can be expensive given the complexity of the current problems. However, computing cost can be reduced if a stepwise deletion algorithm is turned on when recruiting a few new variables.

2. Inference tools. One advantage of classical parametric models is that their parameters admit some clear interpretations. Standard errors of the estimated parameters can easily be computed. For the extended linear modeling, while the values of "standard errors" are available at the final model, they do not necessarily admit the conventional interpretation because of "data snooping" and possible modeling bias. The selected basis functions can vary from simulation to simulation. This makes confidence statements hard to construct. A less ambitious question is how large should the "t-statistic" be in order to have 95% confidence that a coefficient is significantly away from zero. A rough answer can be gained from the empirical experiences via extensive simulations. Despite the above technical difficulties, normalization of estimated coefficients gives us some vague ideas about the relative importance of each selected basis. The estimated coefficients in Table 6, for example, are not normalized. The variable $(111 - t)_+$ is on a smaller scale than that of $(562 - t)_+$, but they are hard to compare with the variables such as "age."

The above remarks have no intention of criticism. Most nonparametric methods face the same challenge. Constructing confidence bands (or pointwise confidence intervals) provides important inferential information, but is a challenging subject. In the univariate setting, Fan, Farmen and Gijbels (1997) outlined a general and simple principle for assessing biases and variances for the local polynomial modeling in likelihood-based models. See also Section 4.9 of Fan and Gijbels (1996). For nonparametric regression, confidence bands can be constructed along with the ideas in Eubank and Speckman (1993). For additive and interaction models, the developments remain to be done.

3. Model diagnostics. Does a model adequately fit the data? Traditional linear models rely on residual plots. The judgment can vary from person to person. These residuals are also available from the extended linear modeling. Here, I would like to describe a method which can be useful for any model fitting, including extended linear modeling. For simplicity, I use the regression setup to outline the idea.

Suppose that we have data $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ generated from the model

$$Y_i = m_0(\mathbf{X}_i) + \varepsilon_i.$$

Let $\hat{m}(\cdot)$ be the regression surface fitted by a method and let $\hat{\varepsilon}_i = Y_i - \hat{m}(\mathbf{X}_i)$ be the residuals. Let $m(\mathbf{x})$ be a function that the fitting method intends to estimate. In the linear regression case, $m(\mathbf{x})$ is simply the best linear approximation to the regression surface $m_0(\mathbf{x})$. Plotting residuals against a covariate variable or an index sequence amounts to visualizing whether the bias $m(\cdot) - m_0(\cdot)$ is negligible in a given direction in the presence of noise. This is not an accurate device because a bias of one-third (say) of the noise level of ε can hardly be detected.

Our idea is simple. If a fit is good, then the residuals should have nearly zero biases. Let $\{\hat{\varepsilon}_i^*\}$ be the Fourier transform of the residual vector $\{\hat{\varepsilon}_i\}$ ordered from low to high frequencies. This compresses useful signals into low frequencies and hence the dimensionality is reduced. Compute the adaptive Neyman test statistic

$$T_{\mathrm{AN}}^* = \max_{1 \le m \le n} \left\{ \left(\sqrt{\hat{\sigma}_2^2 m} \right)^{-1} \sum_{j=1}^{m} (\hat{\varepsilon}_j^{*2} - \hat{\sigma}_1^2) \right\},$$

which is the maximum of the normalized partial sum process, where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are, respectively, the sample standard deviations of $\{\hat{\varepsilon}_i^*,\ i = [n/2], \dots, n\}$ and $\{\hat{\varepsilon}_i^{*2},\ i = [n/2], \dots, n\}$. The reason for only using the high-frequency components to compute the sample variances is that their means are nearly zero. See Fan (1996) for a motivation of the adaptive Neyman test statistic. Reject the null hypothesis that the biases are negligible if T_{AN}^* is too large.

Let

$$T_{\mathrm{AN}} = \sqrt{2 \log \log n}\, T^{*}_{\mathrm{AN}} - \left\{2 \log \log n + 0.5 \log \log \log n - 0.5 \log(4\pi)\right\}$$

be the normalized form. Then, asymptotically,

$$P\{T_{\mathrm{AN}} > x\} \to \exp\{-\exp(-x)\}.$$

Because the class of alternative models is large, we would not reject the null hypothesis unless we had overwhelming evidence. This translates into choosing a small significance level α. If $\alpha = 1\%$, the asymptotic critical value is about 4.6, but our simulations show that for reasonable sample sizes this corresponds to $\alpha = 2.5\%$. In conclusion, T_{AN} larger than 4.6 is the evidence of lack of fit.

The above method depends on the ordering of the residual sequence $\{\hat{\varepsilon}_i\}$. What is a useful ordering scheme? Let us decompose

$$\hat{\varepsilon}_i = m_0(\mathbf{X}_i) - m(\mathbf{X}_i) + \varepsilon_i + \{m(\mathbf{X}_i) - \hat{m}(\mathbf{X}_i)\}.$$

Assume that the last summand is negligible. Then it is clear from Theorem 2.2 of Fan (1996) that the power depends on $\sum_{i=1}^{n}\{m_0(\mathbf{X}_i) - m(\mathbf{X}_i)\}^2$ (which is independent of ordering) and the smoothness of the sequence $\{m_0(\mathbf{X}_i) - m(\mathbf{X}_i)\}$ indexed by i. In other words, a powerful ordering is the one that makes the sequence $\{m_0(\mathbf{X}_i) - m(\mathbf{X}_i)\}$ smooth. Since $m(\cdot)$ is unknown, a good ordering scheme is to make two consecutive covariates have close distance. One possible ordering scheme is according to the diagonal projection of the standardized covariates $(\hat{\Sigma}^{-1/2}\mathbf{X}_i)^T\mathbf{1}$, where $\hat{\Sigma}$ is the sample standard deviation and $\mathbf{1}$ is a vector whose elements are all 1. Another possible scheme is to project $\mathbf{X}_i$ in a few important principal axes. Let λ_j and α_j be the jth largest eigenvalue and its associated eigenvector of the covariance matrix of $\hat{\Sigma}$. Let

$$s_i = \sum_{j=1}^{j_0} \lambda_j^{1/2} \mathbf{X}_i^T \alpha_j,$$

where j_0 is the value such that 80% (say) of variability is explained by the first j_0 principal axes. Then, order the residuals according to the scores s_i before using the adaptive Neyman test.

The last paragraph attempts to search for a good direction for ordering the residuals. Of course, we can choose any sensible direction to order the residuals or combine the test statistics in a few important directions.

4. Constrained models. The extended linear modeling has been successfully applied to a wide array of statistical problems. It is handy to use when one wants to model completely unknown functions. However, some extra thoughts are needed when it is applied to constrained models. Here I outline two problems.

Consider first the semiparametric model

$$Y = g(\mathbf{X}^T\beta_1, \ldots, \mathbf{X}^T\beta_p, \varepsilon)$$

in Li (1991). One can easily use a local modeling technique to solve this problem. The essence of sliced inverse regression (SIR) is to extract the directions by using the inverse regression $E(\mathbf{X}|Y) - E(\mathbf{X})$ via cutting variable Y into slices. See Li (1991) for details. The forward regression method is to use the idea of average directives. See for example Härdle and Stoker (1989) for the case $p = 1$. Let $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ be the regression surface. Then its gradient $\nabla m(\mathbf{x})$ can directly be estimated by local linear regression. This yields a crude estimator $\widehat{\nabla m}(\mathbf{x})$ with small biases and possibly large variances. Now average the derivative estimate to stabilize the variance. Let

$$\hat{\beta}_w = n^{-1} \sum_{i=1}^{n} \widehat{\nabla m}(\mathbf{X}_i) w(\mathbf{X}_i),$$

where w is a given function. Taking p independent functions w yields p independent directions whose linear span is a root-n consistent estimator of the space spanned by the directions $\beta_1, \dots, \beta_p$. An alternative method is to extract the directions via a principal component analysis of the weighted covariance matrix of $\{\widehat{\nabla m}(\mathbf{X}_i)\}$. See for example Wong and Shen (1996).

Direct expansion of g into polynomial spline space and maximizing the resulting likelihood can be difficult. Backfitting algorithms in Hastie and Tibshirani (1990) can be used to iteratively estimate the parametric component $\beta_1, \dots, \beta_p$ and the nonparametric component g via polynomial splines. However, the concavity structure of the likelihood will no longer be available and the success of this schematic implementation remains to be seen.

Next consider the constrained model

$$Y = f_1(X_1) + \cdots + f_p(X_p) + \mathbf{Z}^T\beta + \varepsilon, \quad f_j \text{ monotone},$$

where $X_1, \dots, X_p$ and $\mathbf{Z}$ are given covariates. It is conceptually simple to handle this problem via the smoothing spline approach. Find $f_1, \dots, f_p$ and β that minimize

$$n^{-1} \sum_{i=1}^{n} \{Y_i - f_1(X_{i1}) - \cdots - f_p(X_{ip}) - \mathbf{Z}_i^T\beta\}^2 + \sum_{j=1}^{p} \int \lambda_j \{f_j''(t)\}^2 \, dt$$

subject to constraints that f_j is monotone. See Green and Silverman (1994). Of course, the solution to this problem is not trivial. The local regression approach to this problem can also easily be formulated. Take the local constant modeling as an example. One minimizes

$$\int n^{-1} \sum_{i=1}^{n} \{Y_i - f_1(x_1) - \cdots - f_p(x_p) - \mathbf{Z}_i^T\beta\}^2 K_h(\mathbf{X}_i - \mathbf{x}) w(\mathbf{x}) \, d\mathbf{x}$$

subject to the constraints that f_j is monotone, where K is a given kernel function and w is a given weighting scheme. The solution to this can be found when w and K are in product form. See the last paragraph of our Section 5.

The polynomial spline approach can in principle be used to handle this problem. One needs to expand f_j into spline bases and optimize the parameters subject to appropriate constraints. The constraints are nontrivial and can be nonlinear. Additional difficulty arises whenever adding or deleting a knot because new constraints have to be set in force. New heuristics are needed. This includes setting simpler constraints at the expense that the resulting "allowable space" is only a subset of monotone function space.

5. Theoretical considerations and average regression surface. The polynomial spline estimators have been shown to possess the optimal rates of convergence in various statistical contexts by various subsets of the authors. Their implementations with knot addition and deletion yield appealing methodology. However, proving the sampling properties of the resulting method poses a challenge to the theoretical school. Carefully designed simulation studies can provide valuable insights into complex procedures and serve as a useful criterion. Another possible criterion, as the authors indicated, is that "the true measure of any statistical procedure is its performance on real data." While this criterion can be subjective, its emphasis on practical use is greatly appreciated.

Can theoretical studies provide useful practical relevance? Consider the additive model

$$Y = f_1(\mathbf{X}_1) + f_2(\mathbf{X}_2, \mathbf{X}_3) + \varepsilon,$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are continuous variables of dimension p and q, respectively, and $\mathbf{X}_3$ is a discrete random vector. While the dimensionality of $\mathbf{X}_2$ can be much larger than that of $\mathbf{X}_1$, the function f_1 can be estimated as well as in the case that f_2 is known in terms of asymptotic biases and variance [Fan, Härdle and Mammen (1995)]. This gives a theoretical endorsement to the additive and lower order interaction modeling in the sense that not knowing the components f_2 does not asymptotically cost us anything to estimate f_1. This valuable theoretical insight can hardly be understood without a foundational device.

Another important aspect of this theoretical study is that it yields a practical methodology. The basic idea is to directly estimate the nonparametric regression surface $m(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = E(Y|\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2, \mathbf{X}_3 = \mathbf{x}_3)$ via a local linear regression. Let $\hat{m}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ be the resulting estimator. Then, use averaging to stabilize the variance, resulting in

$$\hat{f}_1(\mathbf{x}) = \sum_{i=1}^{n} \hat{m}(\mathbf{x}_1, \mathbf{X}_{2i}, \mathbf{X}_{3i}) w(\mathbf{X}_{2i}, \mathbf{X}_{3i}).$$

This averaging surface method avoids an iterative estimation scheme and is asymptotically efficient with a suitable choice of weight w. It is also applicable to the additive partial linear model

$$Y = f_1(X_1) + \cdots + f_p(X_p) + \mathbf{Z}^T\beta + \varepsilon.$$

Each additive component f_j can efficiently and directly be estimated and a root-n consistent estimator of parameter β can be obtained via fitting the residuals $Y - \hat{f}_1(X_1) - \cdots - \hat{f}_p(X_p)$ on $\mathbf{Z}$. See Fan, Härdle and Mammen (1995) for details.

How robust is the above approach to the model misspecification? In Stone (1994), it is argued that the polynomial spline estimator will estimate the best additive approximation to the underlying regression curve. A similar result holds: the averages of the regression surface

$$f_j^*(x_j) = \int \{m(X_1, \ldots, x_j, \ldots X_p) - \mu^*\} \prod_{i \neq j} w_i(X_i)\, dX_i,$$

with $\mu^* = \int m(X_1, \ldots, X_p) \prod_i w_i(X_i)\, dX_i$, minimize

$$\int \{m(X_1, \ldots, X_p) - \mu - f_1(X_1) - \cdots - f_p(X_p)\}^2 \prod_i w_i(X_i)\, dX_i$$

subject to the usual identifiability constraints. A similar result holds for the best "interaction model" approximation:

$$\mu + \sum_i f_i(X_i) + \sum_{i<j} f_{i,j}(X_i, X_j).$$

Each term above can be represented as the average of the regression surface and can directly be estimated.

6. Local modeling versus global modeling. Extended linear modeling expands unknown functions into a spline basis, resulting in potentially large parametric models. This global modeling approach aims at capturing nonlinearity and reducing modeling bias. Similar objectives also can be achieved via local modeling: in a local neighborhood around a given point, a polynomial (usually linear or quadratic) function is fitted to the data. The size of the neighborhood or bandwidth is used to control biases and variances of the resulting estimators. This method is also applicable to most statistical problems in this paper. See Fan and Gibjels (1996) for details.

Like most tools, both methods have their own merits. Various discussions on this have already appeared in previous sections. First of all, both approaches include traditional linear models as their submodels. Computationally, the global modeling method solves one or many (depending on whether knots are adaptively chosen) large parametric likelihood problems, while the local modeling approach solves many small parametric (usually two or three parameters) problems. Depending on the implementations and efforts of exploring the data, both methods can be implemented at comparable computing cost. Typically, spline estimates give visually appealing estimated functions, while the local modeling method can be very flexible via varying bandwidths. Local data can often be homoscedastic. Hence, the effect of heteroscedasticity is automatically reduced via the local modeling approach. For boutique problems in Section 4, local modeling offers a natural solution, while for a problem as large and complex as in the phoneme recognition example the solution based on the local modeling approach remains to be seen.

The local polynomial regression as a convenient technical tool dates back at least to Stone (1977). One can use this device to gain some technical insights. A convincing example of this is given in the last section. Asymptotic pointwise minimaxity (rates and constants) can be obtained from the local polynomial fitting. Their sampling properties such as bias and variance and asymptotic distributions can be derived. Estimators of their biases and variances can easily be formulated.

The global modeling and the local modeling approach both have strengths in their own domain of applications. Together they provide invaluable tools for nonlinear data analyses and foundational insights.

REFERENCES

EUBANK, R. and SPECKMAN, P. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **88** 1287–1301.

FAN, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.* **91** 674–688.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

FAN, J., FARMEN, M. and GIJBELS, I. (1997). Local maximum likelihood estimation and inference. *J. Roy. Statist. Soc. Ser. B*. To appear.

FAN, J., HÄRDLE, W. and MAMMEN, E. (1995). Direct estimation of additive and linear components for high dimensional data. Mimeo 2339, Inst. Statistics, Univ. North Carolina, Chapel Hill.

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.* Chapman and Hall, London.

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

LI, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

WONG, W. H. and SHEN, X. (1996). Dimension reduction in regression. Unpublished manuscript.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260
E-MAIL: jfan@stat.unc.edu

DISCUSSION

CHONG GU

Purdue University

Stone, Hansen, Kooperberg and Truong are to be congratulated for their fine article summarizing the adaptive regression spline approach to nonparametric function estimation. With the unified asymptotic theory, the successful applications to a broad spectrum of problems and the availability of user-friendly software, the developments present very impressive achievements that leave many people envious.

Comprehensive and coherent as the authors' treatment is, there still exists an alternative approach that can achieve about as much. This other approach is the penalized likelihood method, pioneered by Good and Gaskins (1971) and extensively developed over the years by the Wisconsin spline school led by Wahba. My mandate here is to present in a nutshell what has been going on with this alternative line of research, and to provide some comparative comments where fit.

In Sections 2, 3 and 4, I will briefly describe what one can do with the penalized likelihood method. Before that, a bit more discussion of the analysis of variance (ANOVA) decomposition is presented in Section 1, which plays a pivotal role in many of the subsequent developments. Brief comparative comments appear here and there as we move along. Further thoughts on model selection are collected in Section 5.

1. ANOVA decomposition. Let us first look at a generic construction of ANOVA decomposition of functions on arbitrary product domains, one that does not involve the notion of inner product. Despite its extensive application in recent developments of the penalized likelihood method, that to some may seem to tie it with the specific method, the construction does have its independent conceptual identity.

Consider a function $\phi(x_1, \dots, x_M)$ on a product domain $\prod_{m=1}^{M} \mathscr{X}_m$. Let A_m be averaging operators acting on arguments x_m that satisfy $A_m^2 = A_m$. An ANOVA decomposition of the function can be defined as

$$\begin{aligned}\phi &= \left\{ \prod_{m=1}^{M} (I - A_m + A_m) \right\} \phi \\ &= \sum_{\mathscr{S} \subseteq \{1, \dots, M\}} \left\{ \prod_{m \in \mathscr{S}} (I - A_m) \prod_{m \in \mathscr{S}^c} A_m \right\} \phi \\ &= \sum_{\mathscr{S} \subseteq \{1, \dots, M\}} \phi_{\mathscr{S}},\end{aligned} \tag{1.1}$$

where $\mathscr{S}$ is the index set of active arguments in a component. $\phi_\varnothing = [\prod_{m=1}^{M} A_m]\phi$ is a constant, $\phi_m = \phi_{\{m\}} = \{(I - A_m)\prod_{l \neq m} A_l\}\phi$ are the x_m main effects, $\phi_{m,l} = \phi_{\{m,l\}} = \{(I - A_m)(I - A_l)\prod_{k \neq m,l} A_k\}\phi$ are the x_m-x_l interactions, and so on. The identifiability of such a decomposition is assured by the side conditions $A_m \phi_{\mathscr{S}} = 0$, $\forall \mathscr{S} \ni m$. The decomposition can also be obtained through recursive hierarchical construction.

For $\mathscr{X}_m = [a, b]$ a real interval, one may choose $A_m \phi = (b - a)^{-1} \int_a^b \phi \, dx_m$ or $A_m \phi = \phi(a)$, anything that satisfies $A_m^2 = A_m$.

For $\mathscr{X}_m = \{1, \dots, K\}$ a discrete domain, one may choose $A_m \phi = K^{-1} \sum_{x_m=1}^{K} \phi(x_m)$ or $A_m \phi = \phi(1)$ and so forth.

For $\mathscr{X}_m$ logically univariate but mathematically multivariate such as the geography, one does not need to decompose things further into say the longitude effect and the latitude effect that do not always make practical sense. A possible choice for the averaging operator is $A_m \phi = N^{-1} \sum_{j=1}^{N} \phi(x_{j,m})$, where $x_{j,m} \in \mathscr{X}_m$ provide a “normalizing mesh” on the domain.

Technically, the decomposition of (1.1) can be constructed explicitly using the tensor product spline technique based on the construction of tensor product reproducing kernel Hilbert spaces, with possibly a mixture of continuous, discrete, univariate or multivariate marginal domains. Technical details can be found in Aronszajn (1950), Wahba (1990) and other references to follow throughout this discussion. For the cursory exposition in this discussion, the reader only needs to know that the components can be independently attached or detached in the construction. For example, on $\mathscr{X}_1 \times \mathscr{X}_2$, one may well choose to consider only functions of the form $\phi = \phi_1 + \phi_{1,2}$, with the constant and the x_2 main effect eliminated. The decomposition obviously is dependent on the choices of A_m, which are usually based on the ease of interpretation or implementation.

Aside from the asymptotic theory, the ANOVA decomposition does not seem to play much of a role in the authors' treatment. For one thing, the authors do not seem to get an explicit ANOVA decomposition from their fit, which can be useful in the interpretation of the fit. Also, a mechanism to enforce selective exclusion of certain interaction terms would be very useful, if one is not already at work.

2. Penalized likelihood function estimation. The penalized likelihood estimate of a function ϕ can be defined by the minimizer of

$$L(\phi|\text{data}) + (\lambda/2)J(\phi), \tag{2.1}$$

where $L(\phi|\text{data})$ is usually the minus log-likelihood that measures the goodness-of-fit of ϕ to the data, $J(\phi)$ often is a quadratic functional that measures the roughness of ϕ and λ is a tunable smoothing parameter that balances the two conflicting goals of goodness-of-fit and smoothness. The minimizer of (2.1) is sought in a function space $\mathscr{H}$ in which $J(\phi) < \infty$. For ϕ on a product domain, the ANOVA decomposition of (1.1) can be built into the procedure via modular constructions of $\mathscr{H}$ and J using the tensor product spline technique. A penalized likelihood estimate is also called a smoothing spline.

Regression. Consider response data from exponential family distributions $Y|x \sim \exp\{(y\phi(x) - b(\phi(x)))/\sigma^2 + c(y, \sigma^2)\}$, where the dependence of the canonical parameter ϕ on the covariate x is to be estimated and the possibly unknown nuisance dispersion parameter σ^2 is assumed common to all observations. Based on observed pairs (x_i, Y_i), ϕ is estimated by minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\phi(x_i) - b(\phi(x_i))\} + \frac{\lambda}{2}J(\phi), \tag{2.2}$$

where σ^2 is absorbed into λ.

For ϕ the normal mean, (2.2) reduces to the classical penalized least squares procedure. Other common examples include ϕ the logit for binary data and ϕ the log intensity for Poisson data. The general formulation of (2.2) appeared in the literature no later than O'Sullivan, Yandell and Raynor (1986).

The covariate x resides on a generic domain $\mathscr{X}$, which, in particular, can be a product domain with a mixture of marginals. Unified numerical and theoretical treatments have been developed over the years; further discussion can be found in the next two sections.

Some of the recent developments in regression include diagnostics for aliasing or negligible terms in an ANOVA decomposition [Gu (1992a)], the incorporation of multivariate marginals in an ANOVA decomposition [Gu and Wahba (1993a)], interval estimates for the individual terms of an ANOVA decomposition [Gu and Wahba (1993b); Wahba, et al. (1995); Wang and Wahba (1995)] and the treatment of dependent observations and longitudinal data [Wang (1996a, b)].

Interval estimates seem to be lacking in the authors' treatment of regression even for the function ϕ itself.

Density estimation. Based on independent samples X_i from a probability density $f(x)$ on a domain $\mathscr{X}$, one may write $f = e^\phi / \int_{\mathscr{X}} e^\phi$, known as a logistic density transform [Leonard (1978)], and estimate ϕ by minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\phi(X_i) - \log\int_{\mathscr{X}} e^\phi\right\} + \frac{\lambda}{2}J(\phi). \tag{2.3}$$

To make the logistic density transform one-to-one, one may enforce a side condition $A\phi = 0$ with some averaging operator A on $\mathscr{X}$ [Gu and Qiu (1993)], as the authors also do. This can be done by the elimination of the constant term in an ANOVA decomposition, possibly one-way.

When $\mathscr{X}$ is a product domain, selective inclusion/exclusion of the ANOVA terms may be employed to incorporate (conditional) independence structures of the marginals, providing a means to the nonparametric fitting of certain graphical models [cf. Whittaker (1990)]. When $\mathscr{X}$ consists of only a portion of a product domain due to sampling truncation, such as in the protein data example in Section 9.4 of the paper under discussion, the ANOVA structure can be used to enforce pretruncation independence of the marginals, if desired.

Further details can be found in Gu and Qiu (1993) and Gu (1993, 1997). Earlier work on univariate density estimation can be found in Good and Gaskins (1971), Leonard (1978), Silverman (1982), O'Sullivan (1988a) and Cox and O'Sullivan (1990).

Conditional density estimation and polychotomous regression. Now consider a product domain $\mathscr{X} \times \mathscr{Y}$, with both marginals generic. Observing pairs (x_i, Y_i), the objective is to estimate the conditional probability density $f(y|x)$. Write the joint density as

$$f(x, y) = \frac{\exp(\phi_x + \phi_y + \phi_{x,y})}{\int_{\mathscr{X}\times\mathscr{Y}} \exp(\phi_x + \phi_y + \phi_{x,y})}, \tag{2.4}$$

where an ANOVA decomposition is explicitly spelled out and the constant term is trimmed for a one-to-one logistic density transform. The conditional density is easily seen to be $f(y|x) = \exp(\phi_y + \phi_{x,y}) / \int_{\mathscr{Y}} \exp(\phi_y + \phi_{x,y})$. This can be written as $f(y|x) = e^\phi / \int_{\mathscr{Y}} e^\phi$, with side conditions $A_y\phi = 0$, $\forall x$, where A_y is the averaging operator on domain $\mathscr{Y}$ that helps to define the ANOVA decomposition. The side conditions ensure a one-to-one logistic conditional density transform. The penalized likelihood estimation of $f(y|x)$ is then through the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\phi(x_i, Y_i) - \log\int_{\mathscr{Y}} \exp(\phi(x_i, y))\right\} + \frac{\lambda}{2}J(\phi) \tag{2.5}$$

in a function space with $A_y\phi = 0$.

While the conditional density can be derived from the joint density estimated from random pairs (X_i, Y_i) via (2.3) [with $\mathscr{X}$ in (2.3) replaced by $\mathscr{X} \times \mathscr{Y}$], the use of (2.5) is necessary when observations on the $\mathscr{X}$ domain are considered "fixed," as in a typical regression setting.

Unlike the regression procedure (2.2), which assumes a parametric model on the $\mathscr{Y}$ axis and estimates a parameter ϕ "univariate" in x, the present procedure estimates a "bivariate" function nonparametrically on both axes. The words "univariate" and "bivariate" are put in quotes for x (and y) can itself be multivariate in a hierarchical structure.

For $\mathscr{Y}$ a real interval, the procedure gets conditional mean and conditional quantiles all at once, without ever running into the quantile crossover problem that may trouble methods which target individual quantiles separately.

For $\mathscr{Y}$ discrete, (2.5) naturally reduces to a procedure for nonparametric polychotomous regression. When the class number is 2, the method reduces to exactly what one would get by applying (2.2) to Bernoulli data.

Further details can be found in Gu (1995a).

Density estimation under sampling bias. Distribution data may not always come from the generating density directly, and they may actually come from a variety of sources. The penalized likelihood method provides a convenient way to combine information in the estimation process.

Observing X_i on $\mathscr{X}$ from a density proportional to $f(x)w_i(x)$ with $w_i(x)$ known, the estimation of $f = e^{\phi}/\int_{\mathscr{X}} e^{\phi}$ is simply through the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\phi(X_i) - \log\int_{\mathscr{X}} w_i e^{\phi}\right\} + \frac{\lambda}{2}J(\phi). \tag{2.6}$$

Ordinary samples, length-biased samples, randomly truncated samples or a mixture of these are among those covered by (2.6). Further details can be found in Gu (1992b).

On a product domain $\mathscr{X} \times \mathscr{Y}$, one sometimes collects data from the "wrong" conditional density $f(x|y)$, but is interested in aspects of the other conditional density $f(y|x)$. This is the case with (unmatched) case-control studies in biostatistics and choice-based sampling in econometrics, collectively known as response-based sampling. When information comes only from $f(x|y)$, (2.5) can be used, with x and y interchanged, to estimate ϕ_x and $\phi_{x,y}$, where ϕ_x and $\phi_{x,y}$ are as in (2.4). The odds ratio that interests most is characterized by $\phi_{x,y}$. When supplemental information concerning the joint density is also available, such as in an enriched choice-based sample [cf. Cosslett (1981)], a simple modification of (2.5) combines all relevant information and all three terms ϕ_x, ϕ_y and $\phi_{x,y}$ are estimable. Further details can be found in Gu (1996a).

Hazard estimation. Let T be the lifetime of an item with a survival function $S(t,u) = P(T > t|u)$ and hazard function $e^{\phi(t,u)} = -\partial \log S(t,u)/\partial t$, where u is a covariate. Let Z be the left truncation time and let C be the right censoring time, independent of T and of each other. Observing $(Z_i, X_i, \delta_i, U_i)$, where $X = \min(T, C)$, $\delta = I_{[T \leq C]}$ and $Z < X$, one may estimate ϕ by minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i \phi(X_i, U_i) - \int_{Z_i}^{X_i} \exp(\phi(t, U_i))\, dt\right\} + \frac{\lambda}{2} J(\phi). \tag{2.7}$$

With an ANOVA decomposition $\phi = \phi_{\varnothing} + \phi_t + \phi_u + \phi_{t,u}$, the elimination of $\phi_{t,u}$ characterizes a proportional hazard model, and the inclusion of $\phi_{t,u}$ takes one beyond the proportional hazard model. The covariate domain $\mathscr{U}$ can be a product domain itself, on which hierarchical ANOVA structures can be recursively constructed. The procedure (2.7) estimates all components of ϕ simultaneously via penalized *full* likelihood.

When the covariate domain $\mathscr{U}$ degenerates to a singleton, (2.7) reduces to the log-hazard estimation procedure originally proposed by O'Sullivan (1988a).

Treating $\phi_{\varnothing} + \phi_t$ as nuisance parameters, penalized *partial* likelihood was used by O'Sullivan (1988b) to estimate ϕ_u in a proportional hazard model and by Zucker and Karr (1990) to estimate $\phi_u + \phi_{t,u}$ of the form $u\beta(t)$, a parametric model with time-varying parameter.

Further details concerning (2.7) can be found in Gu (1994, 1996b, 1997).

Spectral density estimation. Spectral density estimation was a major motivation for the early development of nonparametric function estimation, and the smoothing of a periodogram or log periodogram has been the main tool since day one. Cogburn and Davis (1974) appear to have been the first to use smoothing splines in spectral density estimation.

Based on the first two moments of the log periodogram, Wahba (1980) proposed a certain penalized least squares estimate for the log spectral density, and developed an optimal strategy for the selection of the smoothing parameter. As a refinement of Wahba's (1980) work, Pawitan and O'Sullivan (1994) replaced the least squares by the so-called Whittle log-likelihood of the log periodogram, and developed their version of an optimal smoothing parameter selector. The Whittle log-likelihood is virtually the same log-likelihood the authors use in their LSPEC procedure.

Wahba (1980) and Pawitan and O'Sullivan (1994) both reported extensive empirical studies to justify the optimality of their methods.

3. Asymptotics. Along with the recent methodological developments outlined in Section 2, a unified theme for the calculation of asymptotic convergence rates for penalized likelihood estimates has also emerged. Actually, the asymptotics has played an important role in bringing the method to practice.

Our asymptotic analysis is different from that of the authors. Instead of using the L_2 loss as the universal criterion, we use specific stochastic loss functions customized to specific problem settings. What is common is the analytical approach, together with the routine we follow to customize the loss functions and the regularity conditions.

Take density estimation of (2.3) for example. The loss we target is the symmetrized Kullback–Leibler,

$$\mathrm{SKL}(\phi, \phi_0) = \mu_{\phi}(\phi - \phi_0) - \mu_{\phi_0}(\phi - \phi_0), \tag{3.1}$$

where $\mu_g(h) = \int_{\mathscr{X}} he^g / \int_{\mathscr{X}} e^g$ and ϕ_0 is the "true" function. A related normed distance is

$$V(\phi - \phi_0) = \mu_{\phi_0}((\phi - \phi_0)^2) - \mu^2_{\phi_0}(\phi - \phi_0). \tag{3.2}$$

Under appropriate conditions, the minimizer $\hat{\phi}$ of (2.3) converges to ϕ_0 at a rate

$$\mathrm{SKL}(\hat{\phi}, \phi_0) \sim V(\hat{\phi} - \phi_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda), \tag{3.3}$$

where r is the decay rate of the eigenvalues of V with respect to J, which characterizes the smoothness of functions in space $\mathscr{H} \subseteq \{f\colon J(f) < \infty\}$ in which $\hat{\phi}$ is sought. In general, the space $\mathscr{H}$ is infinite dimensional and $\hat{\phi}$ is not computable. To bring the method to practice, an adaptive finite-dimensional subspace of $\mathscr{H}$, denoted by $\mathscr{H}_n$, is identified, and the minimizer $\hat{\phi}_n$ of (2.3) in $\mathscr{H}_n$ is shown to have the same convergence rate as given in (3.3). Technical details can be found in Gu and Qiu (1993). Customizations for conditional density estimation and for density estimation under sampling bias can be found in respective references cited in Section 2.

For regression, the loss functions are customized to be

$$\begin{aligned} \mathrm{SKL}(\phi, \phi_0) &= \int_{\mathscr{X}} (\phi - \phi_0)(\mu - \mu_0) f, \\ V(\phi - \phi_0) &= \int_{\mathscr{X}} (\phi - \phi_0)^2 v_0 f, \end{aligned} \tag{3.4}$$

where $\mu(x) = \dot{b}(\phi) = E(Y|x)$, $v(x) = \ddot{b}(\phi) \propto \mathrm{Var}(Y|x)$ and $f(x)$ is the limiting density of x_i. The rate given in (3.3) is established for the minimizer $\hat{\phi}$ of (2.2). Technical details are given in Gu and Qiu (1994). In regression problems, $\hat{\phi}$ is known to be in a finite-dimensional space, so no $\hat{\phi}_n$ is necessary.

For hazard estimation, the loss functions are customized to be

$$\begin{aligned} \mathrm{SKL}(\phi, \phi_0) &= \int_{\mathscr{U}} \int_{\mathscr{T}} (\exp(\phi) - \exp(\phi_0))(\phi - \phi_0)\tilde{S}m, \\ V(\phi - \phi_0) &= \int_{\mathscr{U}} \int_{\mathscr{T}} (\phi - \phi_0)^2 \exp(\phi_0)\tilde{S}m, \end{aligned} \tag{3.5}$$

where $\tilde{S}(t, u) = \mathrm{Prob}(Z < t \leq X|u)$ is the at-risk probability and $m(u)$ is the limiting density of U_i. The counting process and martingale structure of survival data are employed to obtain the convergence rate given in (3.3) for the corresponding $\hat{\phi}$ and $\hat{\phi}_n$. Gu (1996b) gives details.

When ϕ_0 resides outside of $\mathscr{H}$, say an additive model is fitted while ϕ_0 does contain interaction, as the authors discuss in Section 2 of the paper, minimal modification of the analysis yields the same rate for $\hat{\phi}$ and $\hat{\phi}_n$ converging toward the Kullback–Leibler projection of ϕ_0 in $\mathscr{H}$. For density estimation, the projection is the minimizer of the relative Kullback–Leibler, $\mathrm{RKL}(\phi|\phi_0) = \log\int_{\mathscr{X}} e^{\phi} - \mu_{\phi_0}(\phi)$, in $\mathscr{H}$. Further details and customizations in other settings are to be found in Gu (1995b).

References of influence include Silverman (1982), Cox and O'Sullivan (1990), Zucker and Karr (1990) and O'Sullivan (1993).

4. Model selection and computation. We shall now discuss smoothing parameter selection strategies, the single most important factor that determines the practical performance of the estimates. To facilitate the use of the methods in data analysis, software that implements the methods is made available to the public.

The most popular smoothing parameter selection method for penalized least squares regression is Craven and Wahba's (1979) generalized cross-validation (GCV), which was shown by Li (1986) to asymptotically minimize the mean square error, $n^{-1}\sum_{i=1}^{n}(\phi(x_i) - \phi_0(x_i))^2$. Generic algorithms have been developed by Gu, Bates, Chen and Wahba (1989) and Gu and Wahba (1991) for the calculation of automatic fits using GCV selected smoothing parameters. The algorithms are implemented in RKPACK [Gu (1989)], a collection of self-documented Fortran compatible routines (available at `http://www.stat.purdue.edu/~chong/software.html`). Information needed for the construction of interval estimates is also available from RKPACK routines.

Among earlier numerical work are GCVPACK by Bates, Lindstrom, Wahba and Yandell (1987) for models without an ANOVA decomposition (`ftp://ftp.stat.wisc.edu/pub/wahba/software`) and BART by O'Sullivan (1985) for the fast calculation of smoothing splines in one dimension (`http://www.netlib.org/gcv`).

For non-Gaussian regression, an iterative algorithm with a certain adaptation of GCV has been developed and justified in Gu (1990, 1992c). Through semitheoretical analysis and simulation, the GCV adaptation was shown to asymptotically minimize the symmetrized Kullback–Leibler of (3.4). The computation is conveniently conducted by direct calls to existing RKPACK routines in each step of the iteration. Portable code was put together by Wang (1995) in GRKPACK (available at `http://www.stat.purdue.edu/~chong/software.html`). An alternative GCV adaptation was developed by Xiang and Wahba (1996).

The computation of density estimates, including conditional densities and possibly with sampling bias, and that of hazard estimates have much in common numerically. A smoothing parameter selection strategy, designed to minimize the loss functions of (3.1), (3.2), (3.5) or others in their respective settings, was built into certain performance-oriented iteration algorithms by Gu (1993, 1994, 1997) and was shown to demonstrate favorable performance in simulation studies. Fortran compatible routines implementing the algorithms have been put together by the discussant in RKPACK-II (currently available in beta version at `http://www.stat.purdue.edu/~chong/software.html`).

What I like the most in the authors' treatment is their provision of user-friendly S functions, which we also hope to do, but probably not in the immediate future. By working with a selected few basis functions, the authors were able to confine the numerical task to a manageable magnitude even for large data sets. With execution speed $O(n^3)$ and memory requirement $O(n^2)$, however, the algorithms implemented in RKPACK and RKPACK-II are likely to hang S with large data sets. Progress is being made to improve the situation [Wahba and Luo (1995)], and with the chip capacity magnifying and the chip price declining at the current rate, larger and larger problems will soon come within reach.

5. Further thoughts on model selection. Being the single most important issue in function estimation, model selection is probably also the softest spot, because "ad hocness" is often the name of the game. The authors' strategy guided by the Wald statistic, the Rao statistic and AIC or BIC is certainly very appealing, and the examples presented indeed demonstrate adequate performance, yet more can be desired, especially in view of how first intuitions can be grossly misleading in this area [Gu (1992c, 1995c)].

What appears missing in the authors' treatment is a systematic assessment of the performance of the method. Rigorous theoretical justification such as Li's (1986) results on Craven and Wahba's (1979) GCV is probably too much to ask, but *systematic* empirical evidence ought to be supplied to present a real convincing case. With a systematic model indexing, such as that by λ for penalized likelihood estimates or that by bandwidth for kernel estimates, one can (and should) always assess the performance of a model selection strategy, at least in relatively simple settings, by gauging its choices against the best possible fits in simulation studies. Such an assessment is understandably less feasible with a recursive growing/pruning approach that the authors have adopted, for the best possible fits are almost impossible to identify. Until some assessment as convincing yet feasible is developed, however, one may not be fully confident that the method is likely to return a nearly optimal fit. Note that AIC, BIC or GCV are not loss functions themselves and do not define the notion of optimality.

In a promising recent development, Shen and Hu (1994) directly tracked some consistent estimate of the relative Kullback–Leibler during the addition/deletion of knots in an adaptive regression spline approach known as the universal sieve method. Backed by rigorous theoretical justifications, the method is somewhat more excusable of a systematic empirical performance assessment.

The lack of performance assessment may also translate into user's confusion in practice. Take the Buffalo snowfall data for example. Facing a rich selection of four possible recipes with neither a track record for each nor a house recommendation, and with markedly different results at least between three of them, I don't know whether a consumer will choose to roll a die or simply leave the house. With a sample size of $n = 63$, as the authors point out, it is virtually impossible to accurately estimate the number of modes, so the simulation presented in Table 2 of the paper offers little help.

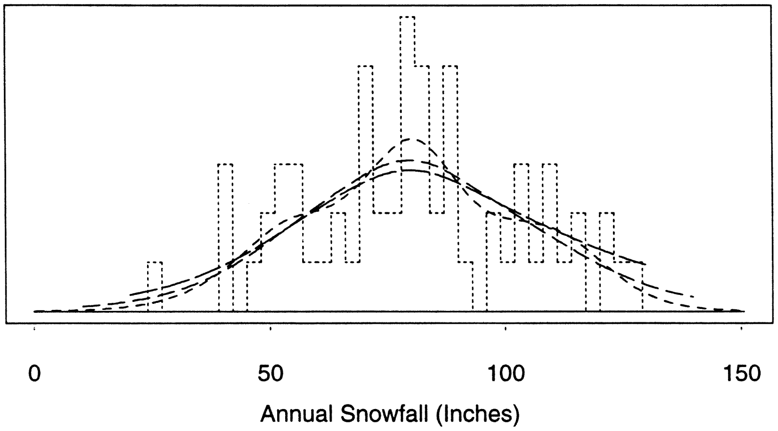


FIG. 1. *The distribution of Buffalo annual snowfalls. The three dashed lines are the automatic estimates under the three domain assumptions indicated by their running lengths. The dotted lines plot a finely binned histogram of the data.*

Reproduced in Figure 1 are three automatic fits to the Buffalo snowfall data using the penalized likelihood method, taken from Gu (1993). The user again has to make some choices, but the choice here is not for different model selection strategies, but for the domain $\mathscr{X}$ on which the log density is assumed to be smooth. The data range from 25.0 to 126.4, and the three fits are supported on $[20, 130]$, $[10, 140]$ and $[0, 150]$, respectively. As the support expands, the model selection strategy tries harder to take away the mass assigned to the empty space at the ends by smoothness, yielding rougher estimates. All three fits are unimodal, however, with the roughest barely showing two shoulders. Note that the relatively smoother fits are not due to a lack of flexibility in the estimation, as the space $\mathscr{H}_n$ (cf. Section 3) has a dimension of 64, but simply by the choice of the model selection procedure. To check out how well the model selection procedure tracks the optimal fits in terms of symmetrized Kullback–Leibler, the reader is referred to the simulation studies documented in Gu (1993).

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404.

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

COGBURN, R. and DAVIS, H. T. (1974). Periodic splines and spectral estimation. *Ann. Statist.* **2** 1108–1126.

COSSLETT, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica* **49** 1289–1316.

COX, D. D. and O'SULLIVAN, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18** 1676–1695.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403.

GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

GU, C. (1989). Rkpack and its applications: fitting smoothing spline models. *Proceedings of the Statistical Computing Section* 42–51. Amer. Statist. Assoc., Alexandria, VA.

GU, C. (1990). Adaptive spline smoothing in non Gaussian regression models. *J. Amer. Statist. Assoc.* **85** 801–807.

GU, C. (1992a). Diagnostics for nonparametric regression models with additive terms. *J. Amer. Statist. Assoc.* **87** 1051–1058.

GU, C. (1992b). Smoothing spline density estimation: biased sampling and random truncation. Technical Report 92-03, Dept. Statistics, Purdue Univ.

GU, C. (1992c). Cross validating non Gaussian data. *J. Comput. Graph. Statist.* **1** 169–179.

GU, C. (1993). Smoothing spline density estimation: a dimensionless automatic algorithm. *J. Amer. Statist. Assoc.* **88** 495–504.

GU, C. (1994). Penalized likelihood hazard estimation: algorithm and examples. In *Statistical Decision Theory and Related Topics 5* (S. S. Gupta and J. O. Berger, eds.) 61–72. Springer, New York.

GU, C. (1995a). Smoothing spline density estimation: conditional distribution. *Statist. Sinica.* **5** 709–726.

GU, C. (1995b). The destination and rates of convergence of penalized likelihood estimate when the model is wrong. Technical Report 255, Dept. Statistics, Univ. Michigan. (Available on-line at `http://www.stat.purdue.edu/~chong/manu.html`.)

GU, C. (1995c). Model indexing and smoothing parameter selection in nonparametric function estimation. Technical Report 93-55 (rev.), Dept. Statistics, Purdue Univ. (Available online at `http://www.stat.purdue.edu/~chong/manu.html`.)

GU, C. (1996a). Smoothing spline density estimation: response-based sampling. Technical Report 267, Dept. Statistics, Univ. Michigan. (Available on-line at `http://www.stat.purdue.edu/~chong/manu.html`.)

GU, C. (1996b). Penalized likelihood hazard estimation: a general procedure. *Statist. Sinica.* **6** 861–876.

GU, C. (1997). Structural multivariate function estimation: some automatic density and hazard estimates. *Statist. Sinica.* To appear. (Available on-line at `http://www.stat.purdue.edu/~chong/manu.html`.)

GU, C. and QIU, C. (1993). Smoothing spline density estimation: theory. *Ann. Statist.* **21** 217–234.

GU, C. and QIU, C. (1994). Penalized likelihood regression: a simple asymptotic analysis. *Statist. Sinica* **4** 297–304.

GU, C. and WAHBA, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Comput.* **12** 383–398.

GU, C. and WAHBA, G. (1993a). Semiparametric ANOVA with tensor product thin plate splines. *J. Roy. Statist. Soc. Ser. B* **55** 353–368.

GU, C. and WAHBA, G. (1993b). Smoothing spline ANOVA with component-wise Bayesian confidence intervals. *J. Comput. Graph. Statist.* **2** 97–117.

GU, C., BATES, D. M., CHEN, Z. and WAHBA, G. (1989). The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal. Appl.* **10** 457–480.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

LI, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in the ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112.

O'SULLIVAN, F. (1985). Discussion of "Some aspects of the spline smoothing approach to nonparametric regression curve fitting" by B. W. Silverman. *J. Roy. Statist. Soc. Ser. B* **47** 39–40.

O'SULLIVAN, F. (1988a). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Comput.* **9** 363–379.

O'SULLIVAN, F. (1988b). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Comput.* **9** 531–542.

O'SULLIVAN, F. (1993). Nonparametric estimation in the Cox model. *Ann. Statist.* **21** 124–145.

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

PAWITAN, Y. and O'SULLIVAN, F. (1994). Nonparametric spectral density estimation using penalized Whittle likelihood. *J. Amer. Statist. Assoc.* **89** 600–610.

SHEN, X. and HU, D. (1994). Universal sieve scheme and spline adaptation. Technical report, Dept. Statistics, Ohio State Univ.

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

WAHBA, G. (1980). Automatic smoothing of the log periodogram. *J. Amer. Statist. Assoc.* **75** 122–132.

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

WAHBA, G. and LUO, Z. (1995). Smoothing spline ANOVA fits for very large, nearly regular data sets, with application to historical global climate data. Technical Report 953, Dept. Statistics, Univ. Wisconsin.

WAHBA, G., WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23** 1865–1895.

WANG, Y. (1995). Grkpack: fitting smoothing spline ANOVA models for exponential families. Technical Report 942, Dept. Statistics, Univ. Wisconsin. (Available on-line at `http://www.sph.umich.edu/~yuedong`.)

WANG, Y. (1996a). Smoothing spline models with correlated random errors. Technical report, Dept. Biostatistics, Univ. Michigan. (Available on-line at `http://www.sph.umich.edu/~yuedong`.)

WANG, Y. (1996b). Mixed-effects smoothing spline ANOVA. *J. Roy. Statist. Soc. Ser. B*. To appear. (Available on-line at `http://www.sph.umich.edu/~yuedong`.)

WANG, Y. and WAHBA, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals. *J. Statist. Comput. Simulation* **51** 263–279.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

XIANG, D. and WAHBA, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian date. *Statist. Sinica* **6** 675–692.

ZUCKER, D. M. and KARR, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist.* **18** 329–353.

DEPARTMENT OF STATISTICS
1399 MATH-SCI BUILDING
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907-1399
E-MAIL: chong@stat.purdue.edu

DISCUSSION

W. HÄRDLE,[1] J. S. MARRON[2] AND L. YANG[1]

Humboldt Universität zu Berlin, University of North Carolina, Chapel Hill and Humboldt Universität zu Berlin

Stone, Hansen, Kooperberg and Truong have written an excellent review of the fine work they have done in making one type of spline modeling useful

[1] Supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" Deutsche Forschungsgemeinschaft.

[2] Supported by NSF Grant DMS-9504414.

in a wide variety of statistical problems. The unifying framework of *extended linear modeling* provides substantial insights about the essential ideas of this type of spline smoothing.

While the presentation is generally excellent, we question the chosen terminology "polynomial splines." The problem is that the rather popular "smoothing spline," which is a solution of a regularization problem and already the subject of two monographs [Wahba (1990) and Green and Silverman (1994)], is also a polynomial spline (although of a very different type). In our view, it would be more appropriate to call the type of spline in the present paper by their older names of "B-spline" or "regression spline."

This discussion consists of questions in four directions: interpretability, theory versus implementation, the effectiveness of knot deletion algorithms and how applicable the present methodology is to some problems in time series and model testing.

1. Interpretability. Many statisticians view simplicity and intuitive understanding of "what the smooth is doing to the data" as very important criteria in choosing a smoothing method. In this respect, we suggest moving average–kernel–local polynomial methods as being preferable, and we believe that they will continue to have an enduring attraction to many statisticians for this reason. We note that Kooperberg and Stone (1991) were not immune to this appeal, and used a very simple kernel method to show (quite convincingly) that the spline method at that time was doing a very good job of density estimation (in fact better than "higher tech" kernel methods). From the point of view of simplicity and interpretability, we ask: "if the kernel method is how one really understands what is going on in the data, why should one then construct the spline?"

2. Theory versus implementation. The gap between what is called "the nonadaptive procedures that we can treat analytically" and "the adaptive methodologies that we have implemented" is somewhat worrying. We are unsure about the suggestion that this is merely because the knot deletion/addition is not very tractable to mathematical analysis. Instead we wonder: "has nobody been able to show this adaptive method is statistically efficient because it is inefficient?"

An alternate, intuitively appealing approach to knot choice for B/regression splines, based on Bayes' ideas, has been developed in Smith and Kohn (1994, 1996a, b). See the Ph.D. dissertation by Smith (1996) for an excellent summary, and a compelling case made for the effectiveness of this approach. A direct comparison of this Bayesian approach with that of the present paper would be quite interesting in terms of statistical efficiency, flexibility and also computation time. We note that Smith and Kohn do much more extensive simulation, and we wonder if this is because their methods are faster to compute.

3. Knot deletion. In this section, we look carefully at some ideas which give us doubts concerning the issues raised in Section 2. We focus here on the one dimensional regression setting, which is probably the easiest to understand and interpret, but the issues we raise here likely exist as well in other settings.

In Section 5 of Stone, Hansen, Kooperberg and Truong, the minimal space is the space of constant functions. Here we show that more discussion of this issue is needed. In particular, we show that without this restriction, the knot deletion procedure can give poor performance both asymptotically and with a simulated example. We wonder if this is an anomaly of the minimal space or if this is what lies at the root of the fact that good asymptotic properties have not been demonstrated for knot deletion methods. In our asymptotics we show that in a "high noise case" a crucial term can be improperly eliminated by knot deletion, which leads to an inconsistent estimate. We then show that this effect is not just an artifact of our asymptotic model by considering a reasonable simulated example where this occurred.

For this, consider the simple regression model

$$Y = m(X) + \varepsilon,$$

where X is assumed to be uniformly distributed on $[0, 1]$, ε is independent of X and normally distributed with mean 0 and variance σ^2, and $m(\cdot)$ is a function defined on $[0, 1]$ with piecewise continuous derivative. Using equally spaced knots initially [as in Stone (1985, 1994)], we let the basis functions be

$$B_0(x) \equiv 1, \qquad B_1(x) = x, \qquad B_j(x) = \left(x - \frac{j-1}{J}\right)_+, \qquad j = 2, \dots, J,$$

where $J = O(n^{1/3})$ [because here $p = d = 1$ in the notation of Stone (1994), where p is the degree of smoothness and d is the highest degree of interaction allowed]. For simplicity, set $J = 2[n^{1/3}/2] + 1$. Given a random sample $(X_i, Y_i)_{i=1}^n$, let $\widehat{m}(x) = \sum_{j=0}^J \widehat{\beta}_j B_j(x)$ be the linear spline estimator of $m(\cdot)$, where the coefficients $\widehat{\beta}_j$ satisfy

$$(\widehat{\beta}_j) = \arg \min_{\beta \in \mathbb{R}^{J+1}} \frac{1}{n} \sum_{i=1}^n \left(Y_i - \sum_{j=0}^J \beta_j B_j(X_i)\right)^2.$$

The deletion rule in the regression setting is to use the residual sum of squares to decide which basis function to add or delete. According to the definition of allowable spaces, the function B_1 cannot be deleted unless all B_j, $j \geq 2$, had been deleted. Although new knots at preselected order statistics of the data could be added, we believe that the addition of these new knots would not significantly change the situation that we are addressing here. We let $\widehat{\beta}_j$, $j = 0, 1, 2, \dots, J$, denote the estimated coefficients for the allowable space that has the smallest GCV value, after the deletion process is done. An indication of difficulties in this context is given by:

PROPOSITION 1. *Under the above assumptions and without the restriction of constant functions being in the minimal space, if* $m(x) = 1 + bx$ *and* $\sigma = 2[n^{1/3}]$, *there exists a constant* $C_1 > 0$, *such that*

$$\lim_{n\to\infty} \inf P\big[\widehat{\beta}_0 = 0\big] \geq C_1.$$

The fact that this leads to inconsistency is summarized as:

COROLLARY 1. *Under the above assumptions and without the restriction of the constant function being in the minimal space, if* $m(x) = 1 + bx$ *and* $\sigma = 2[n^{1/3}]$, *there exist constants* $C_1 > 0$ *and* $C_2 > 0$, *such that*

$$\lim_{n\to\infty} \inf P\big[\big\|\widehat{m}(x) - m(x)\big\|_\infty \geq 1\big] \geq C_1.$$

The assumption $\sigma \approx n^{1/3}$ is a model for "high noise with respect to the sample size." Such noise levels often occur in econometrics. This version of the adaptive spline seems questionable in such applications, because it is inconsistent. This makes us wonder about possible similar inefficiencies in the knot deletion approach because of similar occurrences for other "important" basis functions. Note also that this estimation context is not impossibly difficult. For example, using the simple Nadaraya–Watson estimator, the L_∞ rate is

$$O\big[n^{1/3}(n^{-1}h^{-1}\log n)^{1/2} + h\big],$$

optimized when the bandwidth h is chosen at the rate $n^{-1/9}(\log n)^{1/3}$, which is also the optimal rate. See, for example, Györfi, Härdle, Sarda and Vieu [(1989), Theorem 3.3.0, page 23].

Figure 1 shows a simulated example which demonstrates that the problem of deleting important knots in this way is not an asymptotic oddity. The target function is linear, $m(x) = 0.6x + 0.2$. The data come from adding i.i.d. $N(0, 0.25)$ noise as shown. The estimate comes from doing knot deletion and then finding the AIC and BIC best choices of the number of knots (they were the same for this data set). The poor behavior on the right-hand side is "bad luck" because the data happen to be larger than usual in that area. However, the poor behavior on the left-hand side is caused by the fact that the intercept term of the model is deleted relatively early in the sequence. We are concerned about this because the intercept is actually part of the underlying model.

Our question here is: "are these simply artifacts of our ignoring the minimal space or are they indicators that in fact knot deletion is an inefficient method of adaptation?"

4. Time series and model testing. In nonlinear time series analysis, the conditional variance is often of interest, sometimes more than the conditional mean (for some econometrics data, for example). A review of some recent works in this area can be found in Härdle and Chen (1995) and the references therein. Simultaneous estimation of additive mean and multiplicative volatility functions in autoregressive time series has been done with the local polynomial method by Yang and Härdle (1996). Härdle, Tsybakov and Yang (1996) have also developed estimation procedures of the mean and covariance

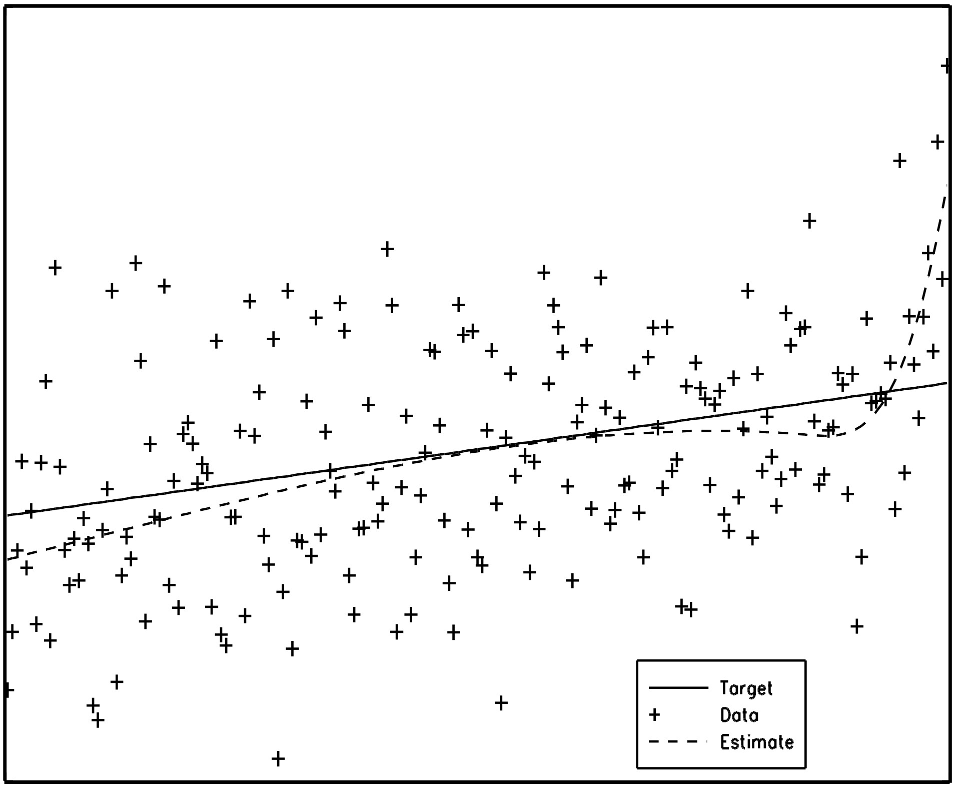


FIG. 1.

function in vector autoregression using local linear estimators. It would be interesting to see such results obtained with the spline approach.

Another area of interesting research where the local polynomial method has been successfully employed is the testing of models. In particular, Härdle, Mammen and Müller (1996) developed procedures for testing parametric versus semiparametric modeling in generalized regression, while Härdle and Yang (1996) developed procedures for testing linearity of main effects in generalized additive regression. Again, it would be interesting to see work in these areas using B/regression splines, which we suspect may be more difficult.

APPENDIX

PROOF OF PROPOSITION 1. We denote by $\mathbf{X}$ the vector $(X_1, X_2, \dots, X_n)^T$, by $\mathbf{Y}$ the vector $(Y_1, Y_2, \dots, Y_n)^T$ and by ε the vector $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$. The inner product on $\mathbb{R}^n$ is defined as $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle_n = (1/n) \sum_{i=1}^n x_i y_i$, while for functions as $\langle f, g \rangle = \int_0^1 f(x) g(x)\, dx$, the norms are defined accordingly. Also denote by $V_{\{j_1, j_2, \dots, j_k\}}$ the function space spanned by $B_{j_1}(x), B_{j_2}(x), \dots,$

$B_{j_k}(x)$ and by $\widehat{V}_{\{j_1,\, j_2, \dots,\, j_k\}}$ the space spanned by $B_{j_1}(\mathbf{X}), B_{j_2}(\mathbf{X}), \dots, B_{j_k}(\mathbf{X})$. For any vector v (or function f), we denote also by $v_{\{j_1,\, j_2, \dots,\, j_k\}}$ (or $f_{\{j_1,\, j_2, \dots,\, j_k\}}$) the projection of v (or f) and by $v^{\perp}_{\{j_1,\, j_2, \dots,\, j_k\}}$ (or $f^{\perp}_{\{j_1,\, j_2, \dots,\, j_k\}}$) $v - v_{\{j_1,\, j_2, \dots,\, j_k\}}$ (or $f - f_{\{j_1,\, j_2, \dots,\, j_k\}}$). The distance from v to $\widehat{V}_{\{j_1,\, j_2, \dots,\, j_k\}}$ is $\|v^{\perp}_{\{j_1,\, j_2, \dots,\, j_k\}}\|$, which we denote by $d(v)_{\{j_1,\, j_2, \dots,\, j_k\}}$ and so forth. For now, we fix $m(x) \equiv 1+bx$. Without loss of generality, we take $b = 0$ because the function $B_1(x) \equiv x$ cannot be deleted.

LEMMA 1. *For any* $j = 2, 3, \dots, J$, $1/4 \le (J/j)D_j \le 1/3$, *where*

$$D_j = d(m(x))^2_{\{1,\, j,\, j+1, \dots,\, J\}},$$

and consequently

$$\frac{1}{4}(1 + O_p(n^{-1/2})) \le \frac{J}{j}\mathbf{D}_j \le \frac{1}{3}(1 + O_p(n^{-1/2}))$$

where $\mathbf{D}_j = d(m(\mathbf{X}))^2_{\{1,\, j,\, j+1, \dots,\, J\}}$.

PROOF. It is easy to verify that $\|m(x) - (J/j)[B_1(x) - B_j(x)]\|^2 = j/(3J)$, which implies the right-hand side of the inequality. Note that among the functions $B_1(x), B_j(x), B_{j+1}(x), \dots, B_J(x)$, only $B_1(x)$ is nonzero on the interval $(0, j/J)$; thus,

$$d(m(x))^2_{\{1,\, j,\, j+1, \dots,\, J\}} \ge \min_t \int_0^{i/J} (1 - tx)^2\, dx = \frac{j}{4J}. \qquad \square$$

LEMMA 2. *For any* $j = 2, 3, \dots, J$,

$$\|\varepsilon_{\{1,\, j,\, j+1, \dots,\, J\}}\|^2 = \frac{4(J-j)}{n^{1/3}}(1 + o_p(1)).$$

PROOF. The two facts needed for the proof are ε is $N(\mathbf{0}_n, \sigma^2\mathbf{I}_{n\times n}) = N(\mathbf{0}_n, 4n^{2/3}\mathbf{I}_{n\times n}(1 + o(n^{-1/3}))$ and the projection subspace is of dimension $1 + (J - j + 1)$. $\square$

LEMMA 3. *For any* $j = 2, 3, \dots, J$,

$$\frac{n^{1/6}}{2\sqrt{\mathbf{D}_j}}\langle \varepsilon, m(\mathbf{X})^{\perp}{}_{\{1,\, j,\, j+1, \dots,\, J\}}\rangle \to N(0, 1).$$

The proof is similar to the previous lemmas.

To complete the proof of Proposition 1, we want to prove that the following event has a probability $\geq C_1 > 0$:

$$\begin{aligned}&\left\|\varepsilon_{\{0,1,\dots,j-1,j+1,\dots,J\}}\right\|^2 + \left\|m(\mathbf{X})_{\{0,1,\dots,j-1,j+1,\dots,J\}}\right\|^2\\&\quad + 2\left\langle\varepsilon_{\{0,1,\dots,j-1,j+1,\dots,J\}}, m(\mathbf{X})_{\{0,1,\dots,j-1,j+1,\dots,J\}}\right\rangle\\&\qquad < \left\|\varepsilon_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\|^2 + \left\|m(\mathbf{X})_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\|^2\\&\qquad\quad + 2\left\langle\varepsilon_{\{1,\dots,j-1,j,j+1,\dots,J\}}, m(\mathbf{X})_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\rangle\end{aligned}$$

for all $j = 2, 3, \dots, J$. This is proved by noting that in fact $m(\mathbf{X})_{\{0,1,\dots,j-1,j+1,\dots,J\}} = m(\mathbf{X})$,

$$\begin{aligned}&\left\|\varepsilon_{\{0,1,\dots,j-1,j+1,\dots,J\}}\right\|^2 - \left\|\varepsilon_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\|^2\\&\qquad \leq \left\|\varepsilon_{\{0,1,\dots,j-1,j,j+1,\dots,J\}}\right\|^2 - \left\|\varepsilon_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\|^2 \leq \frac{4}{n^{1/3}}(1+o_p(1))\end{aligned}$$

by Lemma 1 and

$$\begin{aligned}&\left\|m(\mathbf{X})_{\{0,1,\dots,j-1,j+1,\dots,J\}}\right\|^2 - \left\|m(\mathbf{X})_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\|^2\\&\qquad = \left\|m(\mathbf{X})\right\|^2 - \left\|m(\mathbf{X})_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\|^2 \leq \frac{1}{3J}\end{aligned}$$

by Lemma 2, while

$$\begin{aligned}&2\left\langle\varepsilon_{\{0,1,\dots,j-1,j+1,\dots,J\}}, m(\mathbf{X})_{\{0,1,\dots,j-1,j+1,\dots,J\}}\right\rangle\\&\quad - 2\left\langle\varepsilon_{\{1,\dots,j-1,j,j+1,\dots,J\}}, m(\mathbf{X})_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\rangle\\&\qquad = 2\left\langle\varepsilon, m(\mathbf{X})^{\perp}_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\rangle\end{aligned}$$

has variance of order $n^{-1/3}(1/J)$ or $n^{-2/3}$ by Lemma 3. Therefore the probability of

$$2\left\langle\varepsilon, m(\mathbf{X})^{\perp}_{\{1,\dots,j-1,j,j+1,\dots,J\}}\right\rangle + \frac{4}{n^{1/3}}(1+o_p(1)) + \frac{1}{3J} < 0$$

is greater than a positive constant, meaning that the constant basis B_0 would be the first one to be removed. It is easy to verify that the GCV for the space $\widehat{V}_{\{1,2,\dots,J\}}$ can also be made smaller than that of $\widehat{V}_{\{0,1,2,\dots,J\}}$ with positive probability. In other words, the event ($\alpha = 2.5$ according to the end of Section 5.2 of Stone, Hansen, Kooperberg and Truong)

$$\begin{aligned}&\left(\frac{n-\alpha J}{n-\alpha J-1}\right)^2 \left\|\varepsilon^{\perp}{}_{\{0,1,\dots,J\}}\right\|^2\\&\qquad > \left\|\varepsilon^{\perp}{}_{\{1,\dots,J\}}\right\|^2 + 2\left\langle\varepsilon^{\perp}, m(\mathbf{X})_{\{1,\dots,J\}}\right\rangle + \left\|m(\mathbf{X})^{\perp}_{\{1,\dots,J\}}\right\|^2\end{aligned}$$

can have a positive probability as well. Thus we have shown that the probability that the final model does not contain the constant term is positive. □

PROOF OF COROLLARY 1. Note again that among the functions $B_j(x)$, $1 \leq j \leq J$, only $B_1(x)$ is nonzero on the interval $(0, j/J)$, thus, when the constant term is not in the final model,

$$\left\|\widehat{m}(x) - m(x)\right\|_\infty \geq \inf_{t\in\mathbb{R}^1} \sup_{x\in(0,1/J)} |1 - tx| = 1. \qquad \square$$

REFERENCES

GREEN, P. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models, a Roughness Penalty Approach*. Chapman and Hall, London.

GYÖRFI, L., HÄRDLE, W., SARDA, P. and VIEU, P. (1989). *Nonparametric Curve Estimation from Time Series. Lecture Notes in Statist.* **60**. Springer, Berlin.

HÄRDLE, W. and CHEN, R. (1995). Nonparametric time series analysis, a selective review with examples. In *Proceedings of the 50th Session of the International Statistical Institute, Beijing.* International Statistical Institute, Voorburg.

HÄRDLE, W., MAMMEN, E. and MÜLLER, M. (1996). Testing parametric versus semiparametric modelling in generalized linear models. Discussion Paper 28, Sonderforschungsbereich 373.

HÄRDLE, W., TSYBAKOV, A. and YANG, L. (1997). Nonparametric vector autoregression. *J. Statist. Plann. Inf.* To appear.

HÄRDLE, W. and YANG, L. (1996). Generalized additive models: derivative estimation and hypotheses testing. Discussion paper, Sonderforschungsbereich 373.

SMITH, M. S. (1996). Nonparametric regression: a Markov chain Monte Carlo approach. Ph.D. dissertation, Univ. New South Wales.

SMITH, M. and KOHN, R. (1994). A Bayesian approach to additive nonparametric regression. In *Computationally Intensive Statistical Methods: Proceedings of the 26th Symposium on the Interface* (J. Sall and A. Lehman, eds.) **26** 96–104. Interface Foundation, Fairfax Station, VA.

SMITH, M. and KOHN, R. (1996). Nonparametric regression via Bayesian variable selection. *J. Econometrics* **75** 317–327.

SMITH, M. and KOHN, R. (1996b). Nonparametric bivariate regression. Unpublished manuscript.

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

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

TRUONG, Y. and STONE, C. J. (1992). Nonparametric function estimation involving time series. *Ann. Statist.* **20** 77–97.

TRUONG, Y. and STONE, C. J. (1994). Semiparametric time series regression. *J. Time Series Anal.* **15** 405–428.

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

YANG, L. and HÄRDLE, W. (1997). Nonparametric autoregression with additive mean and multiplicative volatility. *J. Time Series Anal.* To appear.

W. HÄRDLE
L. YANG
INSTITUT FÜR STATISTIK UND ÖKONOMETRIE
WIRTSCHAFTSWISSENSCHAFTLICHE FAKULTÄT
HUMBOLDT UNIVERSITÄT ZU BERLIN
SPANDAUER STRASSE 1
D-10178 BERLIN
GERMANY

J. S. MARRON
DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599

DISCUSSION

TREVOR HASTIE AND ROBERT TIBSHIRANI

Stanford University and University of Toronto

In this paper and their work in the past few years, the authors have done a terrific job in making spline fitting accessible in a wide variety of contexts. Our discussion focuses on the POLYMARS and POLYCLASS procedures, where we report on two discoveries we have made recently.

1. Masking. A popular approach to multiclass classification is via dummy variables and indicator matrices. Suppose we create a $K+1$ response random vector $\mathbf{Z}$ (using the notation of the present paper), such that $Z_k = \operatorname{ind}(Y = k)$. Thus $\mathbf{Z}$ is a vector of all zeros and a single 1 in the position corresponding to the class of Y. Then

$$\begin{aligned} E(\mathbf{Z}|\mathbf{X}=\mathbf{x}) &= \mathbf{P}(\mathbf{x}) \\ &= \big[P(Y=1|\mathbf{X}=\mathbf{x}), \dots, P(Y=K+1|\mathbf{X}=\mathbf{x})\big]. \end{aligned} \tag{1}$$

Since regression can be viewed as estimating a conditional expectation, (1) suggests that we can regress each of the $K+1$ elements Z_k on $\mathbf{x}$ to estimate the elements of $\mathbf{P}(\mathbf{x})$, the conditional probabilities needed for classification. Flexible regression procedures such as MARS, POLYMARS or neural networks are particularly appropriate because they tend to operate locally in $\mathbf{x}$.

A purist may complain because the estimates thus obtained are not guaranteed to be positive or to sum to 1 (although typically they will sum to 1 if the regression method includes an intercept). If we ignore the possible negativity of some of the (smaller) elements of $\mathbf{P}(\mathbf{x})$, we would typically classify to the class with the largest fitted value. This approach is consistent, in that as the regressions approach conditional expectations, this classifier approaches the Bayes-optimal classifier for 0/1 losses.

The authors suggest this approach using POLYMARS to select the basis functions to be used in POLYCLASS. We have used the same approach to select basis functions in the context of flexible discriminant analysis [Hastie, Tibshirani and Buja (1994)].

There are difficulties with this approach, which get worse as K/N gets large. Figure 1 shows some fabricated data with three classes and a single predictor variable x. The classes are perfectly separated, yet when we perform the indicator variable regressions (using linear regression) we see that the middle class never dominates. Of course, this problem can be easily solved by using quadratic regressions rather than linear, and since we anticipate adaptive regression procedures, why the concern?

1. Suppose there are 10 predictors and the 3 classes line up along a particular direction α in predictor space. In order to solve the problem via quadratic polynomials, we would need to fit a general quadratic surface with all the

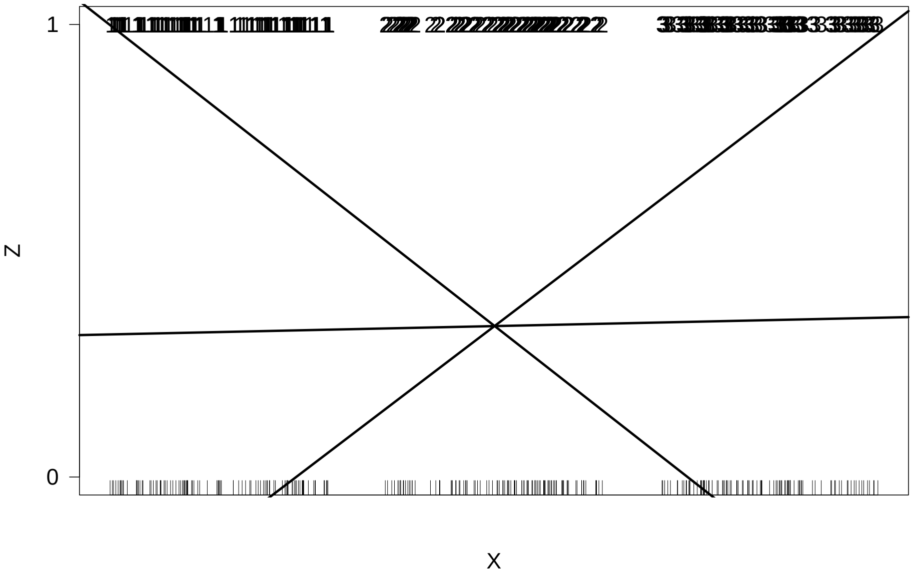


FIG. 1. *The three classes are perfectly separated by the single predictor X (the rug plot shows the distribution of the data). The three lines represent the linear regression fits of each of the three columns of the indicator response matrix Z on X. The* 1*'s,* 2*'s and* 3*'s at the top of the plot indicate the three response indicators Z_i, $i = 1, \dots, 3$, each to be matched with the zeros for each of the other two classes. The middle class is completely masked, in that its regression line (fitted probabilities) never dominate.*

bilinear terms included. Of course, if we knew about projection pursuit regression, we could be a bit smarter than that.

2. If four classes line up, the quadratic curves do not drop down sufficiently fast, and cubic curves are more appropriate. In general, if M classes line up, order M polynomials tend to be needed to completely untangle them.

When the number of classes is large relative to the number of predictors, masking or partial masking of this kind is relatively frequent. Procedures like MARS or POLYMARS will struggle in general to achieve the untangling, because they have difficulty creating the type of general interaction terms required here.

Now if the end result is POLYCLASS on the selected basis functions, then it seems the problem disappears. In these examples POLYCLASS with simple linear basis functions will achieve perfect separation, because the nonlinearity of the exponential is sufficient (actually, if the regions are really pure, then the optimal coefficients will be infinite). The point is we do not want POLYMARS or MARS, the basis-function selectors, to be spinning their wheels adding basis functions that are ultimately not needed in the POLYCLASS model. The solution is of course to perform the selection within the POLYCLASS model, perhaps via sequential use of the score tests, despite the heavy computational premium.

2. Naive Bayes. This (apparently unattractive) procedure is reported to be quite successful as a classifier [Michie, Spigelhalter and Taylor (1994)]; here we describe it in the context of the LOGSPLINE density estimation procedure. Again in the classification context, consider the following independence model for the class conditional densities $f_k(\mathbf{x})$:

$$\begin{aligned}\log f_k(\mathbf{x}) &= \log \prod_{m=1}^{M} f_{km}(x_m)\\ &= \sum_{m=1}^{M} \log f_{km}(x_m)\\ &= \sum_{m=1}^{M}\sum_{j=1}^{J_{km}} \beta_{kmj} B_{kmj}(x_m).\end{aligned} \tag{2}$$

This is a conditional independence model—in each class the densities are a product of marginal densities. We can fit each of these $(K+1)\times M$ marginal densities using the LOGSPLINE procedure.

Given these approximated class densities $\hat{f}_k$, we would classify to the class for which $\hat{f}_k(\mathbf{x})\pi_k$ is largest. The attractiveness of this approach is that we need only to estimate the marginal densities separately (and possibly in parallel.) The criticisms usually leveled are that (a) the conditional independence assumption seems rather stringent and (b) we might be optimizing to achieve subtle features of the separate class densities that ultimately *cancel* when we come to classify.

Suppose for simplicity we drop the k subscript from the basis functions; that is, we use the same J_m basis functions for the mth coordinate in each of the $K+1$ models. Using the $(K+1)$st class as the reference, the logit transform of this naive Bayes model is

$$\begin{aligned}\theta(k|\mathbf{x}) &= \log \frac{P(Y=k|\mathbf{X}=\mathbf{x})}{P(Y=K+1|\mathbf{X}=\mathbf{x})}\\ &= \log \frac{f_k(\mathbf{x})\pi_k}{f_{K+1}(\mathbf{x})\pi_{K+1}}\\ &= \alpha_k + \sum_{m=1}^{M}\sum_{j=1}^{J_m}(\beta_{kmj} - \beta_{(K+1)mj})B_{mj}(x_m)\\ &= \alpha_k + \sum_{m=1}^{M}\sum_{j=1}^{J_m}\alpha_{kmj}B_{mj}(x_m).\end{aligned} \tag{3}$$

This has the same form as the (additive) POLYCLASS model. What then is the distinction?

There is an analogy here to the distinction between linear discriminant analysis and logistic regression. The POLYCLASS model is more general and hence more robust, in effect allowing f_{K+1} to be arbitrary, and each of the f_k

to be an *exponentially additive tilt* of f_{K+1}:

$$f_k(\mathbf{x}) = f_{K+1}(\mathbf{x}) \exp\left[\alpha_k + \sum_{m=1}^{M} \sum_{j=1}^{J_m} \alpha_{kmj} B_{mj}(x_m)\right]. \tag{4}$$

The coefficients are fitted by maximizing the multinomial or *conditional likelihood* $[P(Y|\mathbf{X}=\mathbf{x})]$, where f_{K+1} does not play a role. The naive Bayes model assumes f_{K+1} is also log additive and fits all the parameters by maximizing the full likelihood [actually $P(\mathbf{X}|Y)$, but trivially including $P(Y=k)=\pi_k$ gives us $P(\mathbf{X}|Y)P(Y)=P(\mathbf{X},Y)$].

Given the naive Bayes model, the distinction appears to be between *discriminative* (multinomial) versus *nondiscriminative* training, with the latter offering a great deal of simplification.

We are currently exploring the trade-offs between these two approaches, which can also be extended to handle interactions (second order dependencies).

REFERENCES

HASTIE, T., TIBSHIRANI, R. and BUJA, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Amer. Statist. Assoc.* **89** 1255–1270.

MICHIE, D., SPIGELHALTER, D. and TAYLOR, C. (eds.) (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Chichester.

DEPARTMENT OF STATISTICS
SEQUOIA HALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-MAIL: trevor@stat.stanford.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO
CANADA

REJOINDER

CHARLES J. STONE, MARK H. HANSEN, CHARLES KOOPERBERG,
YOUNG K. TRUONG AND JIANHUA Z. HUANG

University of California, Berkeley, Bell Laboratories,
University of Washington, University of North Carolina, Chapel Hill
and University of California, Berkeley

We wish to thank our discussants for their interesting and thoughtful comments. With these comments in mind, we begin our rejoinder by clarifying some features of spline-based estimation in the context of an extended linear model.

1. Extended linear modeling revisited. Since our paper was written, the theoretical investigation of extended linear modeling has continued to expand. In particular, in Huang (1996a, b) the rate of convergence results for polynomial splines in Section 2 of our paper have been extended to general approximating spaces including polynomials and trigonometric polynomials, as well as polynomial splines. In Huang and Stone (1996) this more general framework is used to extend the rate of convergence results for hazard regression in Kooperberg, Stone and Truong (1995b) to event history analysis involving repeated events of multiple kinds and time-dependent covariates. In light of this work by Huang and his further work in progress on extended linear modeling, the authors of this paper have invited him to be a coauthor of our rejoinder.

In Section 2 of the paper, we presented a rather general asymptotic result for maximum likelihood estimates defined over spaces of (possibly smooth) piecewise polynomials. Suppose for simplicity that the separate intervals used to define the different (univariate) polynomial pieces have common length δ, corresponding to equally spaced knots. To achieve the optimal rate of convergence, δ must shrink to zero as the sample size tends to infinity, with the rate being chosen to balance the bias and variance of the resulting estimate. Stone and Koo (1986a, b) followed a similar nonadaptive recipe for knot placement in the context of generalized linear models and logspline density estimation. In practical applications, however, fixed knot splines are rarely adequate, so adaptive knot spline procedures have been developed that alternate between adding knots in regions where the unknown function being estimated exhibits significant features and deleting knots in regions where, subject to noise considerations, this function seems relatively unstructured. A long but still incomplete list of such developments includes Smith (1982); Breiman, Friedman, Olshen and Stone (1984); Friedman and Silverman (1989); Breiman (1991); Friedman (1991); Kooperberg and Stone (1991, 1992); Breiman (1993); Zhang (1994); Kooperberg, Stone and Truong (1995a,c); Hansen, Kooperberg and Sardy (1996); and Kooperberg, Bose and Stone (1997).

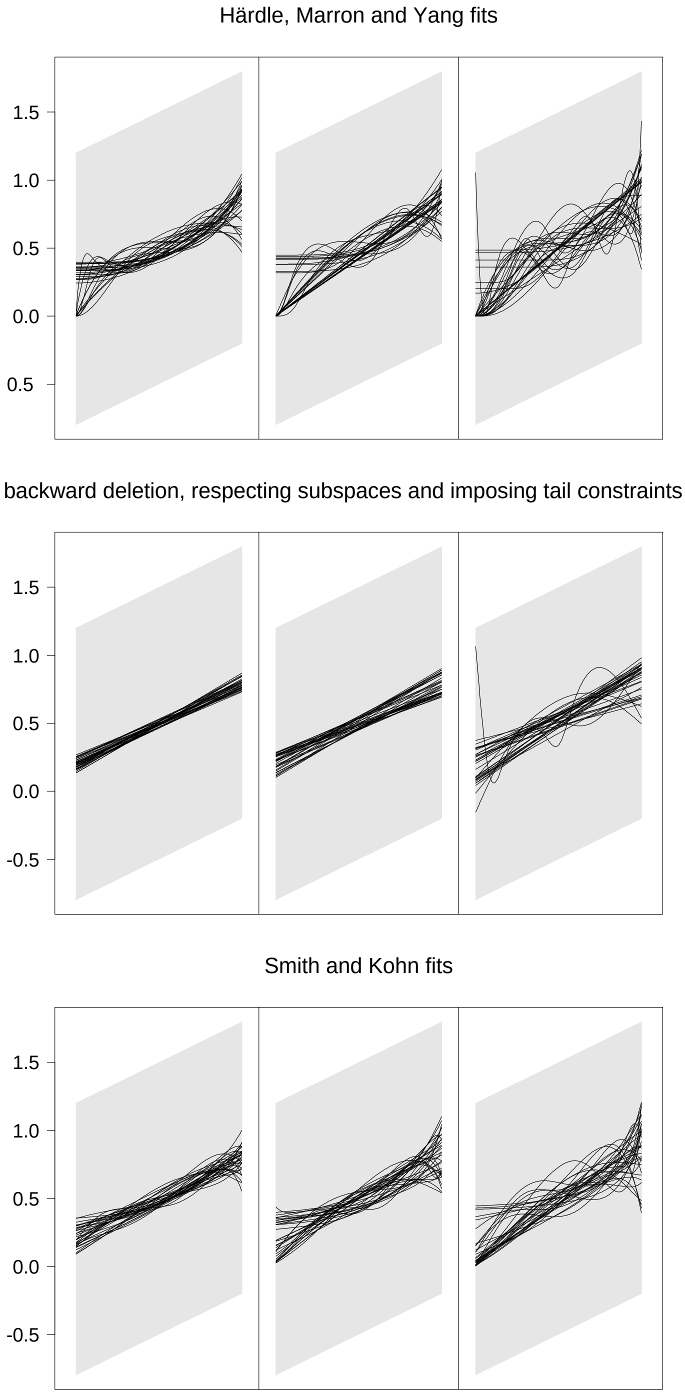
Knot placement is not variable selection. Being based on classical model building techniques, adaptive knot spline procedures are readily understood by practitioners. There is, however, a crucial distinction between knot placement and variable selection. Let G be the space of twice continuously differentiable cubic splines on a bounded interval $[a, b]$ having knots (jumps in the third derivative) at the points $x_1, \dots, x_M$ inside this interval. Aside from its poor numerical properties, the truncated power basis 1, x, x^2, x^3, $(x-x_1)^3_+, \dots, (x-x_M)^3_+$ is convenient for dealing with the addition and deletion of knots. Specifically, deleting a knot is equivalent to deleting one of the last M columns from the design matrix corresponding to the indicated basis, and adding a knot is equivalent to adding the appropriate column to this matrix. It is important, however, to restrict these column operations on the design matrix so that at each step we have made sensible alterations to the underlying function space. Thus we have been careful to treat adaptation in terms of spline spaces, stressing the notion of allowable spaces.

To illustrate what can go wrong, consider knot deletion. By treating this process as a problem in variable selection, we could remove any column of the design matrix corresponding to the truncated power basis. Unfortunately, however, if we attempt to remove the intercept or linear term x in the truncated power basis before the other basis functions have been removed, the resulting space will have potentially poor approximation power in the interval $[a, x_1]$. This is essentially the observation of Härdle, Marron and Yang in the arguments leading to their Proposition 1. While they are mainly concerned with the intercept being removed, the same effect will hold if the linear term is removed prematurely.

In the top row of Figure 1 we display the results of repeating the simulation described by Härdle, Marron and Yang 100 times. For each run, we created a data set of 200 observations: $y_i = 0.6(i - 0.5)/200 + 0.2 + \varepsilon_i$, $i = 1, \dots, 200$, where $\varepsilon_i \sim N(0, 0.25)$. We then performed backward deletion on the columns of the design matrix based on $M = 9$ equally spaced knots, ignoring the implications on the underlying space G. For each data set, we selected the best BIC fit from the sequence of models encountered during deletion, and in the three plots in the top row of Figure 1 we present the top third, middle third, and bottom third of the fits ordered by mean squared error. The gray regions denote the theoretical 95% (pointwise) prediction intervals for the simulated data. For comparison, we repeated this process on the same data sets, this time respecting the structure of the linear spline space; that is, we considered constant, linear, quadratic and cubic polynomial models together with cubic spline models corresponding to the original collection of knots and its nonempty, proper subsets. For each fit, we also imposed the "tail-linear constraints" $g''(0) = g''(1) = 0$ that are commonly employed in adaptive knot cubic spline procedures such as LOGSPLINE and Friedman's MARS. Because this constraint eliminated 2 degrees of freedom, we enlarged the space and considered $M = 11$ initial knots. The resulting fits are presented in the middle row of Figure 1, again ordered by mean squared error.

The welter of curves in the top row of Figure 1 exhibits the anticipated behavior when knot deletion is confused with variable selection. As can be readily verified from the subplots in this row, the intercept appears in only 39 of the 100 final fits. The only fit that actually included both the intercept and the linear term x also contained the other two polynomial terms x^2 and x^3 and two additional spline functions from the power basis. The essential problem is that in the first few deletion steps, the models being considered have very similar residual sums of squares, and a decision that compromises the approximation power of the underlying space can easily be made. This is the essential thrust of the arguments by Härdle, Marron and Yang. By contrast, consider the fits displayed in the second row of Figure 1, only four of which contain terms other than 1 and x. These fits underscore the need to perform stepwise deletion properly, respecting the underlying approximation space. When tail-linear constraints are not imposed, the second row changes only in the high MSE or third subplot, with the extra variation being seen near the boundaries of the interval.

The stepwise procedures suggested in our paper represent computationally efficient approaches to knot placement that have proved effective in practice, but, as the above simulation underscores, it is important to impose the correct



Härdle, Marron and Yang fits

backward deletion, respecting subspaces and imposing tail constraints

Smith and Kohn fits

FIG. 1. *The simulations of Härdle, Marron and Yang. Three separate approaches to model selection.*

hierarchical structure when adding or deleting columns from the design matrix. Not only did Härdle, Marron and Yang fail to appreciate the distinction between knot placement and variable selection, but this point was also missed in the original paper [Smith and Kohn (1996a)], on Bayesian nonparametric regression that they cite. However, this Bayesian approach can be corrected and it provides us with a nice framework to discuss more general schemes for allocating knots efficiently.

Other knot placement procedures. The idea behind the Smith–Kohn technique is to introduce a binary vector $\gamma = (\gamma_1, \dots, \gamma_{M+4})$ that indexes the columns of the design matrix X corresponding to the truncated power basis: γ_i equals 0 or 1 according as the coefficient β_i of the ith basis function does or does not equal 0. The components of γ are assumed to be a priori independent, with probability $\frac{1}{2}$ of equaling 0. This corresponds to giving all possible subsets of the set of $M+4$ variables the same prior probability. After also specifying prior distributions for $\beta = (\beta_1, \dots, \beta_{M+4})|(\gamma, \sigma^2)$ and $\sigma^2|\gamma$, Smith and Kohn derived the posterior distribution of γ given the vector of n observations $y = X\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I_n)$. Specifically, if we let β_γ and X_γ denote the coefficient vector and design matrix, respectively, corresponding to a model containing exactly those variables for which γ_i equals 1, then by setting

$$p(\beta_\gamma|\gamma, \sigma^2) = N(0, c(X_\gamma^T X_\gamma)^{-1}) \quad \text{and} \quad p(\sigma^2|\gamma) \sim 1/\sigma^2,$$

we find that the posterior probability function of γ is given by

$$p(\gamma|y) \sim (1+c)^{-q_\gamma/2}\left(y^T y - \frac{c}{c+1} y^T X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y\right)^{-n/2},$$

where $q_\gamma = \sum_i \gamma_i$ is the number of terms in the model and c is a user-specified constant. Smith and Kohn applied the Gibbs sampler to simulate from the posterior distribution of γ and either report the posterior mode of γ or the posterior mean of β. The model has been specified so that the sampling procedure steps through many models with high posterior probability in a computationally efficient manner. After K Gibbs iterations, two alternative approaches are applied to the samples $\gamma^{[k]}$, $k = 1, \dots, K$, to estimate β: (i) $\hat{\beta}$ is obtained by a least squares fit to those variables included in the model specified by the vector $\gamma^{[k]}$ that maximizes $p(\gamma^{[k]}|y)$ or (ii) the posterior mean $E(\beta|y)$ is estimated by the average value of $E(\beta|y, \gamma^{[k]})$, where the indicated conditional expectation is computed exactly using the fact that $\beta|(y, \gamma)$ has a multivariate t distribution.

Transforming the expression for the posterior probability function of γ into something more familiar to the smoothing community, Foster and George (1996) have found that under certain conditions the value of γ that maximizes this function also minimizes the quantity $\mathrm{RSS}(\gamma) + (1 + c^{-1})\log(c+1) q_\gamma \hat{\sigma}^2$, where $\mathrm{RSS}(\gamma)$ is the residual sum of squares for the model specified by γ, and $\hat{\sigma}^2$ is estimated from the full model with all $M+4$ variables. By selecting c properly, we can perform model selection with respect to Mallow's C_p, AIC or BIC. Making this connection, we see that the Gibbs sampler of Smith and Kohn is in fact an alternative to our approach of minimizing BIC in a stepwise fashion. Unfortunately, because of the way the vector γ treats all variables as candidates for inclusion or exclusion, we are left with the deficiencies described in the previous section.

As an illustration of the last remark, we applied the Smith–Kohn technique to the same data and basis as in our previous simulations. To this end we downloaded the S-PLUS function `br()` from statlib, which returns only the posterior mean estimate of β, and applied it with the option `density = 20`, thus specifying 20 data points per knot (or $M = 9$). The resulting fits, shown in the bottom row of Figure 1, exhibit behavior that is similar to (but somewhat less variable than) those in the top row. The problem is that the posterior model probability is roughly the same for the simple model involving 1 and x as it is for a large number of models that include a single polynomial term and one or more of the spline functions from the power basis. In short, because Smith and Kohn searched over models that would not be considered when the underlying approximation space G is respected, their final fits average together models with the same types of degeneracies as are found in the top row of Figure 1. Mike Smith (personal communication) has indicated several ways in which their technique could provide for the proper structure of spline spaces by imposing corresponding structure on the prior distribution of γ. Further results on the general model considered by Smith and Kohn and empirical Bayes procedures for estimating c and specifying different prior distributions for γ can be found in Foster and George (1996). Ultimately, we agree with Härdle, Marron and Yang that, properly implemented, the Smith–Kohn technique is a very attractive method for efficiently identifying good knot locations in the context of least squares regression.

The performance of this technique, however, is heavily dependent on the number of knots used to define the truncated power basis. An alternative approach followed by Denison, Mallick and Smith (1997) involves defining prior distributions for the number and location of knots as well as the coefficients in a spline expansion. The resulting "automatic Bayesian curve fitting" procedure makes use of reversible jump Markov chain Monte Carlo methods [Green (1995)] to compute the posterior distribution, this time over collections of models having different numbers and positions of knots. At each step in their sampling procedure, one of several possible transitions is chosen at random. These transitions include adding a new knot and either moving or deleting an existing knot, and they can in principle be made efficient through the use of Rao and Wald statistics as discussed in Section 3 of our paper.

Recently, Hansen and Kooperberg (1997) have applied Markov chain Monte Carlo to the problem of identifying promising triangulations for fitting the triogram models discussed in Section 9. In this investigation, the collection of addition and deletion steps described in Figures 15 and 16 in Section 9.2 has been augmented with two new moves, illustrated here in Figure 2: (i) swapping the diagonal of a convex quadrilateral; (ii) moving a vertex within the union of triangles that contain it. With these additional moves, a much wider vari-

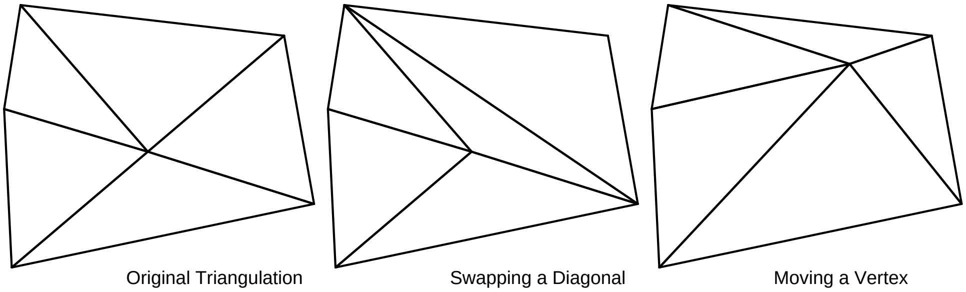


FIG. 2. *Additional moves to improve the search for promising triangulations.*

ety of triangulations can efficiently be examined than was possible by means of simple stepwise adaptation. For a similar technique applied to piecewise constant modeling, see Nicholls (1996).

2. Specific responses to our discussants.

Härdle, Marron and Yang. Härdle, Marron and Yang question the use of stepwise knot deletion, indicating that this may be "an inefficient method of adaptation." Their concerns stem primarily from their proof that if we are cavalier in our definition of a minimal space, then the resulting estimates can be inconsistent. As noted in Section 1 of this rejoinder, their proposal to remove the constant function from the minimal allowable space is not natural when this space is thought of as a linear space of functions rather than as the span of columns of the design matrix. Härdle, Marron and Yang also suggest that these inconsistencies could develop for other "important" basis functions. Like other locally adaptive procedures, the stepwise addition and deletion process can produce spurious effects. The arguments of Härdle, Marron and Yang, however, paint a needlessly pessimistic picture of the problems with our approach. We know, for example, that without adaptive knot selection, spline estimates yield an L_∞ rate of $O[n^{1/3}(n^{-1}h^{-1}\log n)^{1/2} + h]$; here $h = 1/J$ is the distance between uniformly spaced knots. As noted by Härdle, Marron and Yang, this expression is optimized when $h^{-1} = J = O[n^{1/9}(\log n)^{-1/3}]$, not when $J \sim n^{1/3}$ as they used in their theoretical work. Therefore, a fair comparison between the fixed bandwidth Nadaraya–Watson estimate and the nonadaptive spline estimate with equally spaced knots indicates that the two procedures achieve the same asymptotic performance when measured in the L_∞ norm.

As a practical matter, we depend on a mixture of simulated and real data to guide the development of each new application. For example, when testing the "deletion-only" version of the LOGSPLINE procedure, we experimented with various rules for selecting the number and location of the initial knots. If we apply stepwise deletion starting with too many knots, we encounter an unacceptable number of spurious fits to data drawn from various known densities.

If the initial number of knots is too small, on the other hand, the procedure is unable to adapt to the essential features of any but the most regular data sets. Ultimately, we decided that a sensible default for the LOGSPLINE software is to place $O(n^{1/5})$ knots at suitable order statistics. Interestingly enough, the theoretical inconsistency result derived by Härdle, Marron and Yang fails to hold unless the initial number of knots is much larger than $n^{1/5}$.

To summarize, this discussion reinforces the notion that we must be mindful that our adaptive procedure is defined on an underlying spline space and not on the columns of a particular parameterization of the space. This immediately defines a set of allowable operations on the resulting design matrix that we must not ignore. Their theoretical treatment also adds support to the prevailing notion that these procedures are sensitive to the size of the maximum allowable space and hence that this size must be tuned through extensive simulation.

We agree with these discussants that kernel and local polynomial methods will continue to have an enduring attraction to many statisticians. In particular, it is simple to understand what “vanilla” kernel smoothing does to data, which is one reason it was used as a *rough diagnostic* in Kooperberg and Stone (1991). On the other hand, once we include variable bandwidths [Silverman (1986)], bandwidth selection [Sheather and Jones (1991)] and transformations [Wand, Marron and Ruppert (1991)] in kernel density estimation, it loses its advantages in simplicity, intuitive understanding and mathematical elegance over logspline density estimation.

These three discussants also question the terminology “polynomial spline” and would prefer that we use the “older” names of “B-spline” or “regression spline.” Our use, however, is compatible with the bulk of the literature on splines, which is in the field of numerical approximation. For example, Schumaker (1981) starts out Chapter 9 (which along with the first eight chapters is about univariate splines) with the following:

> In the first eight chapters of this book we have dealt exclusively with polynomial splines. Here and in the following two chapters we develop the theory of similar spaces of generalized splines. We begin this development by studying Tchebycheffian splines, where, as we shall see, almost all of the results for polynomial splines can be carried over.

Tchebycheffian splines include, for example, smooth piecewise exponential functions. It should also be noted that “B-spline” and “polynomial spline” are not synonymous terms. While B-splines are a basis for certain spaces of polynomial splines in a single variable, in certain methodologies dealing with polynomial splines, a different choice of basis is often more convenient (see Section 5.2). Finally, it seems perverse to us to use “regression spline” to refer to a polynomial spline model for a log-density function.

Fan. The approximation power of adaptive knot splines makes them attractive tools for modeling functions with unknown smoothness properties

[Devore and Lorentz (1991)], but the usefulness of these tools in a statistical context depends on how well knot placement algorithms can be tuned in the presence of noise. As Fan points out, the general problem of knot placement is very intricate. The use of Rao and Wald statistics in a stepwise fashion is a practical attempt to deal with this intricacy while preserving the flexibility of adaptive knot splines. In order to control the variability of estimation based on this approach, however, we need to limit the maximum number of knots and the minimum spacing between knots.

As Fan suggests, in the context of fitting constrained models, additional ideas may be required to apply our approach and other techniques are sometimes preferable. In particular, Li's approach to dimensionality reduction is an attractive alternative to the use of low-order ANOVA models. Our concern about using the local polynomial method to estimate the effective directions in Li's model is that the resulting methodology lacks the simplicity of Li's original SIR method. The same issue would arise if we were to substitute our approach for the local polynomial method in the procedure described by Fan. However, our approach could be used after the directions have been selected by SIR.

Fan claims that the smoothing spline approach provides a conceptually simple solution to the monotone regression problem, but this is not obvious to us. Ramsay (1988) and Kelly and Rice (1990) provide alternative solutions involving B-splines. When the monotonicity assumption is valid, our approach should yield estimates that are monotone or nearly so; otherwise, the estimates should provide a warning that this assumption may be invalid.

Some constrained models are naturally handled by our approach. For example, in partly linear models without monotonicity constraints, the unknown function has the form $f_1(X_1)+\cdots+f_p(X_p)+Z^T\beta$. Here, to apply our approach we simply need to enforce the additive and linear constraints in choosing the allowable spaces.

The surface averaging technique developed by Fan, Härdle and Mammen (1995) provides a theoretically interesting alternative to our approach in estimating the components in ANOVA models, at least in the regression context. It is surprising that lack of knowledge of f_2 does not cost us anything asymptotically in estimating f_1. However, this result depends on smoothness assumptions on both f_2 and the joint density function of the covariates. Moreover, in estimating an additive component of the regression function at a single point, we need to obtain a local linear estimate of the multidimensional regression function at every "design" point. This necessity makes their procedure computationally intensive. Also, in light of the curse of dimensionality and their smoothness requirement on the joint density function, we are concerned that a very large sample size may be needed for their procedure to be competitive in statistical efficiency with ours. Fan mentions that a special case of their procedure will estimate the average of the regression surface defined with respect to a weight function of product form even if their additive model is misspecified, but we wonder about the motivation for such a weight function when the covariates may be dependent.

Gu. The smoothing spline approach to extended linear modeling developed by Wahba and her school, nicely summarized by Gu in his discussion, is an attractive alternative to our approach. While both approaches emphasize the use of low-order ANOVA models to overcome the curse of dimensionality, the actual implementations are quite different. In the smoothing spline approach, the penalization technique is used after the various main effect and interaction components in the ANOVA model are specified, with other techniques being employed to select these components [Wahba et al. (1995)]. This approach is competitive with ours in applications involving a few covariates, but it is computationally infeasible in applications such as SOLVD and phoneme recognition that involve many covariates.

In his discussion, Gu concentrates on a single smoothing parameter λ as a systematic model index. In the context of fitting low-order ANOVA models, however, it is more natural to use one such parameter per component. When there are many smoothing parameters, their automatic selection becomes computationally challenging. In the context of additive regression and additive generalized regression with polynomial splines, Burman (1990) treated the adaptive selection of a single smoothing parameter (common number of knots). Presumably, similar arguments would apply to other extended linear models.

An interesting modification to the adaptive knot spline approach is to perform a penalized fit using a basis built by stepwise addition. This alternative to stepwise deletion was suggested by Hastie [see Buja, Duffy, Hastie and Tibshirani (1991)] in his discussion of MARS, and it has been implemented by Hastie and Pregibon (1990) in the context of shrinking tree models. Luo and Wahba (1997), cited by Gu, narrow the gap between the smoothing spline and adaptive knot spline approaches both computationally and philosophically by considering the stepwise addition of cubic reproducing kernel functions. A penalized regression is then performed on the smaller basis, reducing the amount of computation considerably and also improving the local adaptability of the smoothing spline approach. In the simple univariate regression problems considered in their paper, for example, they consider adding up to a maximum of 150 such basis functions in problems with samples of size 2048. Interestingly, the authors state that they realized only small improvements in mean squared error when a penalized rather than ordinary least squares fit was performed on the basis built by stepwise addition.

We are pleased to see the many recent and interesting papers on the theoretical properties of the smoothing spline approach that were cited in Gu's discussion. In these papers, however, the underlying assumptions involve conditions on the eigenvalues and eigenfunctions of certain bilinear forms. In general, these conditions are collectively very difficult to verify from more primitive and statistically more natural conditions such as those in Stone (1994), Hansen (1994) and Kooperberg, Stone and Truong (1995b). This difficulty is heightened in the context of applying the smoothing spline approach to the fitting of additive and other unsaturated ANOVA models so as to ameliorate the curse of dimensionality (see Section 2 of our paper). Chen (1991) does deal successfully with such heightened difficulties, but only for regression. Moreover, for mathematical tractability, he was forced to replace the random points $\mathbf{X}_1, \dots, \mathbf{X}_n$ by deterministic points that form a suitably regular balanced complete factorial design.

Although it is not clear to us exactly what has been rigorously established in the paper by Shen and Hu that Gu cites, we agree that this paper is a promising recent development at least in the sense of suggesting a worthwhile direction for future research.

Hastie and Tibshirani. Hastie and Tibshirani's example illustrates that the use of multiple response linear regression is at best a cheap substitute for POLYCLASS and that it is much better either to use POLYCLASS directly or to fit a POLYCLASS model with basis functions selected by POLYMARS. Obviously, the situation in their toy example is not as bad when POLYMARS is used to select basis functions. In particular, when we applied POLYMARS to a similar data set, it immediately positioned two knots in each of the two gaps between classes, so that the fitted probabilities almost exactly matched the true probabilities. In this example a POLYMARS model with three basis functions can achieve perfect classification. While POLYCLASS and POLYMARS can achieve both "perfect classification" and "perfect probabilities" in this example, POLYCLASS can do so with fewer basis functions.

The extended linear modeling approach that we have discussed in the current paper always considers nonlinear functions of the predictors. This is precisely to prevent artifacts such as the one presented by Hastie and Tibshirani.

Ignoring computational efficiency, we agree with Hastie and Tibshirani in strongly prefering POLYCLASS to POLYMARS in the selection of basis functions for use in POLYCLASS because the latter approach would tend to add basis functions that are ultimately not needed at the POLYCLASS fitting stage and thus require more basis functions to ensure optimal performance. On the other hand, we doubt that this tendency is so great that the former approach should invariably be used whenever it is computationally feasible, despite its heavy computational premium. This issue deserves investigation.

In any case, we ended up using the combination procedure for the phoneme recognition example because POLYCLASS itself is not computationally feasible when applied to such huge data sets with so many classes. Currently, we are studying the application of the stochastic gradient method to the fitting of POLYCLASS models. With this approach, the selection of 400 basis functions using POLYMARS and the fitting of a POLYCLASS model together take less than 20 h of CPU time (on the Silicon Graphics workstation at our disposal), which would allow us to experiment with different sets of features and use more basis functions. (In fact, we have experimented with different sets of features and larger models, and we now get test-set error rates of approximately 30%, which is competitive with neural networks.)

Clearly, the density estimation approach proposed by Hastie and Tibshirani will be feasible for larger data sets than POLYCLASS by itself. For example, to apply it to the phoneme recognition data, one would have to estimate $45 \times 63 = 2835$ densities based on sample sizes of $112{,}115/45 \approx 2500$ on the average when no interactions are considered. However, this is actually very feasible using LOGSPLINE, since this procedure takes only about 10 s of CPU time when applied to the income data ($n = 7125$). If we do not increase the maximum number of basis functions beyond the 25 used for the income data (and this may be very reasonable since presumably not too much detail in the density estimates is needed), the CPU time for LOGSPLINE would be linear in the sample size. Thus the density estimation approach to polychotomous regression would take approximately $2835 \times (2500/7125) \times 10$ s ≈ 3 h of CPU time. Moreover, the computations could easily be divided over a number of workstations.

A related approach would be to fit separate logistic regressions for each of the $K+1$ classes; that is, we would obtain separate estimates of the functions

$$\phi(k|\mathbf{x}) = \log \frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y \neq k|\mathbf{X} = \mathbf{x})}, \qquad 1 \le k \le K+1.$$

We can then set

$$\theta(k|\mathbf{x}) = \phi(k|\mathbf{x}) - \phi(K+1|\mathbf{x}) = \log \frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y = K+1|\mathbf{X} = \mathbf{x})}, \qquad 1 \le k \le K.$$

As in the density estimation approach, this is a POLYCLASS model. The logistic regression approach could be considered as *partial discriminative* training.

For the logistic regression approach, we can use the POLYCLASS algorithm, which is quite feasible for very large data sets when there are only two classes. Kooperberg, Bose and Stone (1997) established that the POLYCLASS algorithm requires approximately $O(K^2 P_{\max}^3 n)$ flops independently of the number of predictors. Thus for the logistic regression approach, we would have to carry out $K+1$ times an algorithm that takes $O(P_{\max}^3 n)$ flops, so that we would gain a factor of K. This may not make the approach feasible for the phoneme recognition data, but it will definitely be applicable to much larger data sets than will a direct application of POLYCLASS.

3. Some publicly available software. Not surprisingly, the discussants introduced a number of viable alternative approaches to function estimation in various extended linear model settings. To practitioners, this diversity can represent new views of a given dataset and hence the potential for new insights. Unfortunately, current nonparametric procedures are sufficiently difficult to implement that without usable, publicly available software, it is unlikely that a new technique would ever be applied. During our own work to refine the various spline-based estimates presented in our paper, we have found that a language like S/S-PLUS allows us easily to share both code and results, speeding up the entire learning process.

With this in mind, we were naturally led to inquire about the availability of software for comparing our approach with those suggested by the discussants. Both Fan and Härdle, Marron and Yang mention local polynomial techniques, Härdle, Marron and Yang suggested that kernel procedures are dependable and understandable, and Gu summarized how the smoothing spline technology has been applied in many extended linear model settings. We decided to test the practicality of these various approaches by using each one to estimate the density associated with the income data discussed in Section 4 (see Figure 1) of our paper. Locating software that implements each of these approaches was fairly straightforward: a variety of local polynomial models can be fitted using LocFit, a collection of routines with a very polite S/S-PLUS interface being written by Clive Loader [Loader (1996)] and available from http://cm.bell-labs.com/stat/project/locfit; kernel density estimation comes with standard S/S-PLUS, and Simon Sheather kindly provided us with a Fortran subroutine for calculating the Sheather–Jones plug-in (SJPI) bandwidth estimate [Sheather and Jones (1991)]; Chong Gu's RKPK2 is available on the Web and is packaged with a number of density estimation examples [Gu (1993)].

Using an IRIS Challenger with twenty 150 MHz IP19 processors, we tried each procedure on the mean-scaled data used to generate Figure 1 in Section 4 of our paper. Based on the experience of Wand, Marron and Ruppert (1991), we attempted to fit the logarithm of the mean-scaled data as well. The CPU time for each procedure is recorded in Table 1. The nonadaptive LocFit time includes the time to select the global bandwidth via BIC. The available implementations of the adaptive LocFit and smoothing spline procedures required binning the data before the models could be fitted. Unfortunately, variable bandwidth selection is not yet implemented for density estimation in LocFit, but by applying a Poisson approximation we can make use of the adaptive features implemented for generalized linear models. The smoothing spline procedure is known to be computationally intensive, requiring $O(n^3)$ operations, so this procedure was not able to deal directly with the income data having $n = 7125$. After mapping the income data to the unit interval, we partitioned the data into 400 equally sized bins and passed these counts to the variable bandwidth LocFit and the smoothing spline programs. The CPU times in seconds listed in Table 1 do not include this preprocessing.

The results are compared in Figure 3:

(a) The LOGSPLINE fit from Section 4 of our paper, compared with a new version of the routine that uses simulated annealing to optimize knot locations [Hansen and Kooperberg (1997)]. SALSA (simulated annealing logspline

TABLE 1
CPU times in seconds

	Original Data ($n = 7125$)	**Binned (400 bins)**
LOGSPLINE	9.0	
LocFit	14.5	
Variable bandwidth LocFit		19.7
Smoothing splines		194.2
SJPI	341.0	

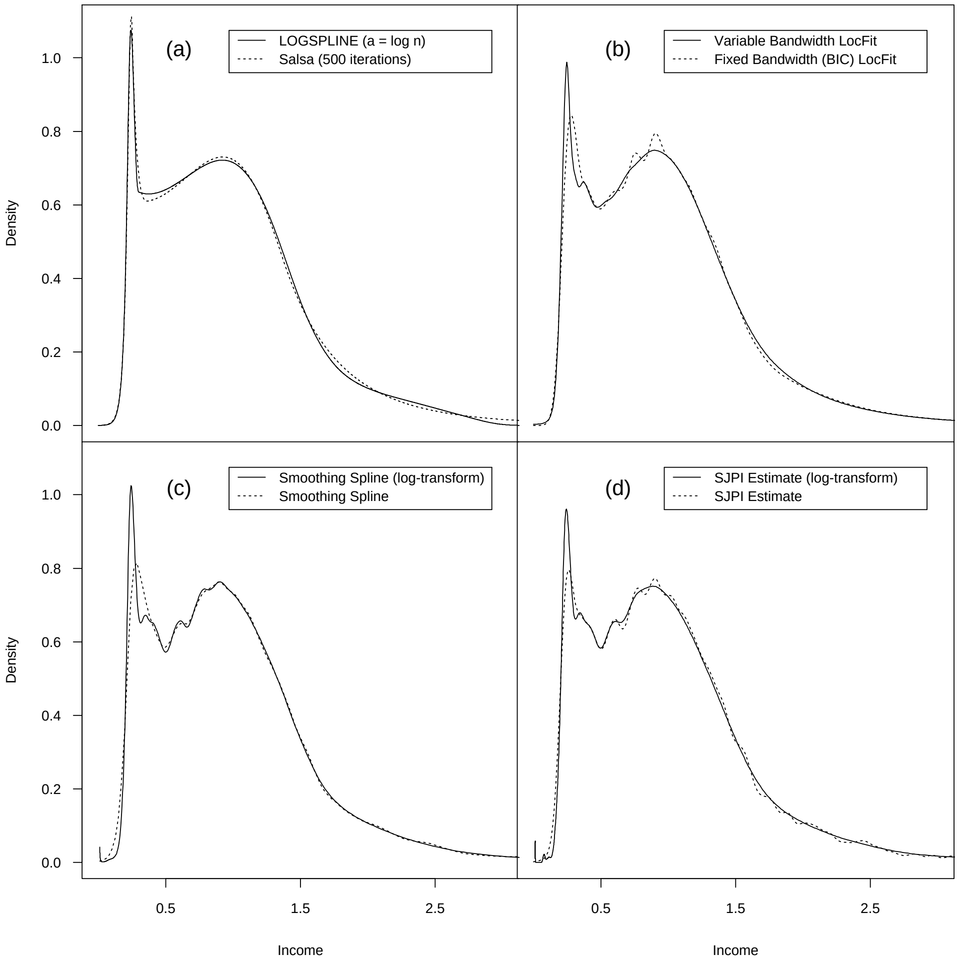


FIG. 3. *Comparison of various density estimation methods on the income data.*

approximation) requires a few minutes CPU time, so it is still competitive with the slower density estimation routines that are compared in the other examples.

(b) LocFit applied with both a global, fixed bandwidth (chosen by BIC) and a variable bandwidth using a Poisson approximation.

(c) The smoothing spline estimate using both log-transformed and untransformed mean-scaled data.

(d) The SJPI estimate using both log-transformed and untransformed mean-scaled data. (We also tried using the implementation `width.SJ` of SPJI mentioned by Venables and Ripley [(1994), page 139], which is available from statlib, and found that it required similar CPU time. The CPU time for SPJI could undoubtedly be reduced to some extent by binning.)

The smoothing spline estimate, the SJPI estimate and the fixed bandwith LocFit estimate clearly have problems capturing all the elements of the data, since each of these procedures selects only one (global) smoothing parameter. To truly capture the height of the spike, each estimate would have to be very rough near the larger mode and in the tail. In fact, the heights of the spikes of the smoothing spline and SJPI estimates based on the untransformed data are reduced by more than 20% from that suggested by the adaptive routines, but these estimates still produce ripples in other parts of the density. A logarithmic transformation considerably improves both nonadaptive estimates. We believe, however, that the efficacy here of such a simple transformation with one global smoothing parameter is the consequence of a fortuitous "interaction" between the spike and the tail. [Wand, Marron and Ruppert (1991) considered a two parameter family of power transformations for kernel density estimation that yields a very reasonable estimate for the income data.]

Currently, ignoring computing time, we prefer the SALSA estimate. It seems to have the correct height for the spike. Based on calculations not reported here, we also feel that the dip near 0.6 has approximately the right depth. Among the faster procedures, LOGSPLINE and the variable bandwidth LocFit (which required binning and using a Poisson approximation) produce very nice estimates.

REFERENCES

BREIMAN, L. (1991). The Π method for estimating multivariate functions from noisy data (with discussion). *Technometrics* **33** 125–160.

BREIMAN, L. (1993). Fitting additive models to regression data. *Comput. Statist. Data Anal.* **15** 13–46.

BREIMAN, L., FRIEDMAN, J. H. OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

BUJA, A., DUFFY, D., HASTIE, T. and TIBSHIRANI, R. (1991). Discussion of "Multivariate adaptive regression splines" by J. H. Friedman. *Ann. Statist.* **19** 93–99.

BURMAN, P. (1990). Estimation of generalized additive models. *J. Multivariate Anal.* **32** 230–255.

CHEN, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19** 1855–1868.

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1997). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc. Ser. B*. To appear.

DEVORE, R. and LORENTZ, G. (1991). *Constructive Approximation*. Springer, New York.

FAN, J., HÄRDLE, W. and MAMMEN, E. (1995). Direct estimation of additive and linear components for high dimensional data. Mimeo 2339, Inst. Statistics, Univ. North Carolina, Chapel Hill.

FOSTER, D. P. and GEORGE, E. I. (1996). Empirical Bayes variable selection. Technical report, Dept. MSIS, Univ. Texas, Austin.

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

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–39.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

GU, C. (1993). Smoothing spline density estimation: a dimensionless automatic algorithm. *J. Amer. Statist. Assoc.* **88** 495–504.

HANSEN, M. (1994). Extended linear models, multivariate splines and ANOVA. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

HANSEN, M. and KOOPERBERG, C. (1997). Strategies for spline adaptation. Unpublished manuscript.

HANSEN, M., KOOPERBERG, C. and SARDY, S. (1996). Triograms models. Technical Report 304, Dept. Statistics, Univ. Washington.

HASTIE, T. and PREGIBON, D. (1990). Shrinking trees. Technical report, AT&T Bell Laboratories.

HUANG, J. Z. (1996a). Projection estimation in multiple regression with application to functional ANOVA models. Technical Report 451, Dept. Statistics, Univ. California, Berkeley.

HUANG, J. Z. (1996b). Functional ANOVA models for generalized regression. Technical Report 458, Dept. Statistics, Univ. California, Berkeley.

HUANG, J. Z. and STONE, C. J. (1996). The L_2 rate of convergence for event history regression with time-dependent covariates. Technical Report 473, Dept. Statistics, Univ. California, Berkeley.

KELLY, C. and RICE, J. (1990). Monotone smoothing with application to dose–response curves and the assessment of synergism. *Biometrics* **46** 1071–1085.

KOOPERBERG, C., BOSE, S. and STONE, C. J. (1997). Polychotomous regression. *J. Amer. Statist. Assoc.* **92** 117–127.

KOOPERBERG, C. and STONE, C. J. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.

KOOPERBERG, C. and STONE, C. J. (1992). Logspline density estimation for censored data. *J. Comput. Graph. Statist.* **1** 301–328.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995a). Hazard regression. *J. Amer. Statist. Assoc.* **90** 78–94.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995b). The L_2 rate of convergence for hazard regression. *Scand. J. Statist.* **22** 143–157.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995c). Logspline estimation of a possibly mixed spectral distribution. *J. Time Ser. Anal.* **16** 359–388.

LOADER, C. (1996). Local likelihood density estimation. *Ann. Statist.* **24** 1602–1618.

LUO, Z. and WAHBA, G. (1997). Hybrid adaptive splines. *J. Amer. Statist. Assoc.* **92** 107–116.

NICHOLLS, G. (1996). Bayesian image analysis with Markov chain Monte Carlo and colored continuum triangulation models. *J. Roy. Statist. Soc. Ser. B*. To appear.

RAMSAY, J. O. (1988). Monotone regression splines in action (with discussion). *Statist. Sci.* **3** 425–461.

SCHUMAKER, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

SMITH, P. L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. Report NASA 166034, NASA, Langley Research Center, Hampton, VA.

SMITH, M. and KOHN, R. (1996a). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

SMITH, M. and KOHN, R. (1996b). Nonparametric bivariate regression. Unpublished manuscript.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

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. *Contemp. Math.* **59** 1–15.

VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern Applied Statistics with S-Plus*. Springer, New York.

WAHBA, G., WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23** 1865–1895.

WAND, M. P., MARRON, J. S. and RUPPERT, D. (1991). Transformations in density estimation (with discussion). *J. Amer. Statist. Assoc.* **86** 343–361.

ZHANG, H. (1994). Maximal correlation and adaptive splines. *Technometrics* **36** 196–201.

CHARLES J. STONE
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720-3860
E-MAIL: stone@stat.berkeley.edu

MARK H. HANSEN
BELL LABORATORIES
700 MOUNTAIN AVE., RM. 2C260
MURRAY HILL, NEW JERSEY 07030

CHARLES KOOPERBERG
FRED HUTCHSONSON CANCER RESEARCH CENTER
1100 FAIRVIEW AVE., MP 1002
SEATTLE, WASHINGTON 98109-1024

YOUNG K. TRUONG
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-7400
E-MAIL: truong@stat.unc.edu

JIANHUA Z. HUANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302