Statistical Science
2002, Vol. 17, No. 1, 2–51

Spline Adaptation in Extended Linear Models

Mark H. Hansen and Charles Kooperberg

Abstract. In many statistical applications, nonparametric modeling can provide insights into the features of a dataset that are not obtainable by other means. One successful approach involves the use of (univariate or multivariate) spline spaces. As a class, these methods have inherited much from classical tools for parametric modeling. For example, stepwise variable selection with spline basis terms is a simple scheme for locating knots (breakpoints) in regions where the data exhibit strong, local features. Similarly, candidate knot configurations (generated by this or some other search technique), are routinely evaluated with traditional selection criteria like AIC or BIC. In short, strategies typically applied in parametric model selection have proved useful in constructing flexible, low-dimensional models for nonparametric problems.

Until recently, greedy, stepwise procedures were most frequently suggested in the literature. Research into Bayesian variable selection, however, has given rise to a number of new spline-based methods that primarily rely on some form of Markov chain Monte Carlo to identify promising knot locations. In this paper, we consider various alternatives to greedy, deterministic schemes, and present a Bayesian framework for studying adaptation in the context of an extended linear model (ELM). Our major test cases are Logspline density estimation and (bivariate) Triogram regression models. We selected these because they illustrate a number of computational and methodological issues concerning model adaptation that arise in ELMs.

Key words and phrases: Adaptive triangulations, AIC, BIC, density estimation, extended linear models, finite elements, free knot splines, GCV, linear splines, multivariate splines, regression.

1. INTRODUCTION

Polynomial splines are at the heart of many popular techniques for nonparametric function estimation. For regression problems, TURBO (Friedman and Silverman, 1989), multivariate adaptive regression splines or MARS (Friedman, 1991) and Π (Breiman, 1991) have all met with considerable success. In the context of density estimation, the Logspline procedure of Kooperberg and Stone (1991, 1992) exhibits excellent spatial adaptation, capturing the full height of spikes without overfitting smoother regions. And finally, among classification procedures, classification and regression trees (CART) (Breiman, Friedman, Olshen and Stone, 1984) is a de facto standard, while the more recent PolyMARS models (Kooperberg, Bose and Stone, 1997) have been able to tackle even large problems in speech recognition. Stone et al. (1997) and a forthcoming monograph by Hansen, Huang, Kooperberg, Stone and Truong are the prime references for the application of polynomial splines to function estimation. In this paper, we review a general methodological framework common to procedures like MARS and Logspline, and contrast it with several Bayesian

Mark H. Hansen is a Member of the Technical Staff, Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974 (e-mail: cocteau@bell-labs.com). Charles Kooperberg is Member, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109-1024.

approaches to spline modeling. We begin with some background material on splines.

1.1 Splines

Univariate, polynomial splines are piecewise polynomials of some degree d. The breakpoints marking a transition from one polynomial to the next are referred to as *knots*. In this paper, we will let the vector $\mathbf{t} = (t_1, \ldots, t_K) \in \mathbb{R}^K$ denote a collection of K knots. Typically, a spline will also satisfy *smoothness constraints* describing how the different pieces are to be joined. These restrictions are specified in terms of the number of continuous derivatives, s, exhibited by the piecewise polynomials. Consider, for example, piecewise linear curves. Without any constraints, these functions can have discontinuities at the knots. By adding the condition that the functions be globally continuous, we force the separate linear pieces to meet at each knot. If we demand even greater smoothness (say, continuous first derivatives), we loose flexibility at the knots and the curves become simple linear functions. In the literature on approximation theory, the term "linear spline" is applied to a continuous, piecewise linear function. Similarly, the term "cubic spline" is reserved for piecewise cubic functions having two continuous derivatives, allowing jumps in the third derivative at the knots. In general, it is common to work with splines having *maximal smoothness* in the sense that any more continuity conditions would result in a global polynomial.

Given a degree d and a knot vector $\mathbf{t}$, the collection of polynomial splines having s continuous derivatives forms a linear space. For example, the collection of linear splines with knot sequence $\mathbf{t}$ is spanned by the functions

$$1, x, (x - t_1)_+, \ldots, (x - t_K)_+, \tag{1}$$

where $(\cdot)_+ = \max(\cdot, 0)$. We refer to this set as the *truncated power basis* of the space. In general, the basis for a spline space of degree d and smoothness s is made up of monomials up to degree d together with terms of the form $(x - t_k)_+^{s+j}$, where $1 \le j \le d - s$. Using this formula the classical cubic splines have $d = 3$ and $s = 2$ so that the basis has elements

$$1, x, x^2, x^3, (x - t_1)_+^3, \ldots, (x - t_k)_+^3. \tag{2}$$

From a modeling standpoint, the truncated power basis is convenient because the individual functions are tied to knot locations. In the expressions (1) and (2), there is exactly one function associated with each knot, and eliminating that function effectively removes the knot. This observation is at the heart of many statistical methods that involve splines and will be revisited shortly.

The truncated power functions (1) and (2) are known to have rather poor numerical properties. In linear regression problems, for example, the condition of the design matrix deteriorates rapidly as the number of knots increases. An important alternative representation is the so-called *B-spline basis* (de Boor, 1978). These functions are constructed to have support only on a few neighboring intervals defined by the knots. (For splines having maximal smoothness, this means $d + 1$ neighboring intervals.) A detailed description of this basis is beyond the scope of this paper, but the interested reader is referred to Schumaker (1993). For the moment, assume we can find a basis $B_1(x; \mathbf{t}), \ldots, B_J(x; \mathbf{t})$ for the space of splines of degree d with smoothness s and knot sequence $\mathbf{t}$ so that any function in the space can be written as

$$g(x; \boldsymbol{\beta}, \mathbf{t}) = \beta_1 B_1(x; \mathbf{t}) + \cdots + \beta_J B_J(x; \mathbf{t}), \tag{3}$$

for some coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^t$. If we are dealing with spline spaces of maximal smoothness, then $J = K + d + 1$, as we have seen in (1) and (2). Given this structure, we now briefly describe a broad collection of estimation problems that admit relatively natural techniques for identifying good fitting functions g.

1.2 Extended Linear Models

Extended linear models (ELMs) were originally defined as a theoretical tool for understanding the properties of spline-based procedures in a large class of estimation problems (Hansen, 1994; Stone et al., 1997; Huang, 1998, 2001). This class is extremely rich, containing all of the standard generalized linear models as well as density and conditional density estimation, hazard regression, censored regression, spectral density estimation and polychotomous regression. To describe an ELM, we begin with a probability model $p(W|h)$ for a (possibly vector-valued) random variable $W \in \mathcal{W}$ that depends on an unknown (also possibly vector-valued) function h. Typically, h represents some component of the probability model about which we hope to make inferences. For example, in a normal linear model, h is the regression function; while for density estimation, we take h to be the log-density.

Let $l(W|h) = \log p(W|h)$ denote the log-likelihood for an ELM, and assume that there exists a unique

function ϕ that maximizes the expected log-likelihood $El(W|h)$ over some linear space of real-valued functions H. The maximizer ϕ defines "truth," and is the target of our estimation procedures. [In Stone et al. (1997) a slightly more general notion of "truth" is developed to handle ANOVA-like functional decompositions.] We refer to this set-up as an *extended linear model* for ϕ. In this case, the term "linear" refers to our use of a linear model space H. The class H is chosen to capture our beliefs about ϕ, and is commonly defined through smoothness conditions (e.g., we might assume that the true regression function in a linear model has two continuous, bounded derivatives). These weak assumptions about ϕ tend to result in classes H that are infinite dimensional. Therefore, for estimation purposes we choose to work with flexible, finite-dimensional spaces G that have good approximation properties. That is, the elements $g \in G$ can capture the major features of functions $\phi \in H$, or $\min_{g\in G} \|g - \phi\|$ is small in some norm for all $\phi \in H$. Splines are one such approximation space.

Given a series of observations $W_1, \ldots, W_n$ from the distribution of W, we estimate ϕ by maximizing the log-likelihood

$$l(g) = \sum_i l(W_i|g) \qquad \text{where } g \in G. \tag{4}$$

Our appeal to maximum likelihood in this context does not imply that we believe $p(W|\phi)$ to be the true, data-generating distribution for W. Rather, p may be chosen for computational ease in the same way that ordinary least squares can be applied when the assumption of strict normality is violated. In theoretical studies, it is common to let the dimension of G depend on the sample size n. For example, if G is a spline space with K knots, we let $K = K(n) \to \infty$ as $n \to \infty$. As we collect more data, we are able to entertain more flexible descriptions of ϕ. Asymptotic results describing the number of knots $K(n)$ and their placement needed to achieve optimal mean squared error behavior are given in Stone (1985), Stone (1994), Hansen (1994), and Huang (1998, 2001), and Stone and Huang (2002).

An ELM is said to be concave if the log-likelihood $l(w|h)$ is concave in $h \in H$ for each value of $w \in \mathcal{W}$ and if $El(W|h)$ is strictly concave in h [when restricted to those h for which $El(W|h) > -\infty$]. Strict concavity holds for all of the estimation problems listed at the beginning of this section. Now, let G be a spline space with knot sequence $\mathbf{t}$ so that any $g \in G$ can be written in the form (3). Then since $g(\cdot) = g(\cdot; \boldsymbol{\beta}, \mathbf{t})$, the log-likelihood (4) can be written as $l(\boldsymbol{\beta}, \mathbf{t})$. Because of concavity, the maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\beta}}$ for the coefficients $\boldsymbol{\beta}$ and a fixed $\mathbf{t}$ can be found efficiently in reasonably-sized problems through simple Newton–Raphson iterations. Therefore, it is possible to compute

$$l(\mathbf{t}) = \max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \mathbf{t}). \tag{5}$$

After making the dependence on $\mathbf{t}$ explicit in this way, we can consider adjusting the knot locations $t_1 < \cdots < t_K$ to maximize the log-likelihood. It is intuitively clear that the knot sequence $\mathbf{t} = (t_1, \ldots, t_K)$ controls the flexibility of elements in g to track local features: tightly-spaced knots can capture peaks, while widely-separated knots produce smooth fits.

However, even in the simplest case, linear regression with a single univariate predictor, maximizing (5) over knot sequences is a difficult optimization problem. To see this, we first translate univariate regression into an ELM: let $W = (X, Y)$ and define $p(W|\phi)$ via the relationship

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

for an unknown regression function ϕ. The error ε is assumed independent of X with a normal distribution having mean zero and variance σ^2. For a spline space G, the negative log-likelihood for $\boldsymbol{\beta}$ is proportional to the regression sum of squares

$$\begin{aligned}\mathrm{RSS}(\boldsymbol{\beta}, \mathbf{t}) &= \sum_i \big(Y_i - g(X_i; \boldsymbol{\beta}, \mathbf{t})\big)^2 \\ &= \sum_i \big(Y_i - \beta_1 B_1(X_i; \mathbf{t}) - \cdots \\ &\qquad - \beta_J B_J(X_i; \mathbf{t})\big)^2.\end{aligned}$$

If we hold $\mathbf{t}$ fixed, then

$$\begin{aligned} l(\mathbf{t}) \propto -\mathrm{RSS}(\mathbf{t}) &= \max_{\boldsymbol{\beta}} \{-\mathrm{RSS}(\boldsymbol{\beta}, \mathbf{t})\} \\ &= -\mathrm{RSS}(\widehat{\boldsymbol{\beta}}, \mathbf{t}), \end{aligned} \tag{6}$$

where $\widehat{\boldsymbol{\beta}}$ is the ordinary least squares estimate of $\boldsymbol{\beta}$. Jupp (1978) demonstrated that $-\mathrm{RSS}(\mathbf{t})$ has local maxima along lines of the form $t_k = t_{k+1}$, making the solution to (6) difficult for standard optimization software. Not surprisingly, this problem persists in even more exotic ELMs.

Several authors have considered special transformations, penalties or ad hoc optimization schemes to maximize the log-likelihood with respect to $\mathbf{t}$ (Jupp, 1978; Lindstrom, 1999; Kooperberg and Stone, 2002). In this paper, we will instead consider an approximate solution that begins by connecting knot placement with model selection.

1.3 Model Selection

The concavity of an ELM together with the association of knot locations to terms in the truncated power basis suggests simple approximations to maximizing (5) based on fast, stepwise approaches to model selection. Consider splines of degree d having maximal smoothness and knot sequence $\mathbf{t}$. According to Section 1.1, this means that $s = d - 1$ and each knot point t_k in $\mathbf{t}$ is associated with only one of the truncated monomials $(x - t_k)_+^d$; the linear (1) and cubic (2) splines are two examples. Therefore, moving t_k effects only one basis element in G, and in fact removing t_k entirely is equivalent to deleting $(x - t_k)_+^d$ from the model. Many existing spline methods use this idea in some form. It was originally proposed by Smith (1982a, b) and it has been the workhorse of many procedures suggested since (TURBO, DKCV, MARS, PolyMARS).

Returning to the problem of maximizing (5), suppose we have a finite set of candidate knots $\mathcal{T} = \{t'_1, \ldots, t'_{K'}\}$, from which we want to select a subset of size K, $\mathbf{t} = (t_1, \ldots, t_K)$, $K \leq K'$. The connection between knots and basis functions suggests that finding a good sequence $\mathbf{t}$ is really a problem in model selection where we are choosing from among candidate basis functions of the form $(x - t)_+^d$, $t \in \mathcal{T}$. For linear regression and moderate numbers of candidate knots K', we can find the sequence of length K that minimizes (6) using traditional branch-and-bound techniques. However, when K' gets large, or when we have a more exotic ELM requiring Newton–Raphson iterations to evaluate (5), this approach quickly becomes infeasible.

For computational efficiency, the algorithms discussed by Stone et al. (1997) take a stepwise approach, introducing knots in regions where the unknown function ϕ exhibits significant features, as evaluated through the log-likelihood, and deleting knots in regions where ϕ appears relatively smooth. More formally, starting from a simple spline model, knots are added successively, at each step choosing the location that produces the greatest increase in the log-likelihood. This is followed by a pruning phase in which unnecessary knots are removed, at each stage eliminating the basis element that results in the smallest change in the log-likelihood. Because we are always taking the best single alteration to the current model, these schemes are often referred to as *greedy*. To prevent this process from tracking spurious patterns in the data, it is common to impose constraints on the initial model, the size M of the largest model fit during addition, and the minimal number of data points between each knot. These restrictions are defined in terms of *allowable spaces*, a topic we will discuss in more detail in the next section.

Several facts about ELMs make this approach attractive computationally. Consider placing a single knot in a linear regression model. Then, among all basis sets of the form $1, x, (x - t)_+$, we want to find the one that minimizes the criterion (6), which in this case is a function of t. It is not hard to show that $\mathrm{RSS}(t)$ is a piecewise smooth function of t, with breaks in the first derivative at each of the data points. This means we can derive fast heuristics to guide the search for new knots during the addition phase without having to evaluate all the candidates. Next, the concavity of the ELMs listed in Section 1.2 means that we can quickly approximate the change in log-likelihood from either adding or deleting a knot without actually fitting each candidate model. We now describe each alteration or "move" in more detail.

Knot addition. Let G be a J-dimensional spline space with a given knot sequence, $\mathbf{t}$, and let $\widehat{\boldsymbol{\beta}}$ denote the MLE of $\boldsymbol{\beta}$. When using the truncated power basis inserting a new knot is equivalent to adding a single basis function to G, taking us to a new $(J + 1)$-dimensional space G_1 with coefficient vector $\boldsymbol{\beta}_1$ and knot sequence $\mathbf{t}_1$ (where we let B_{J+1} be the basis function associated with the new knot). To evaluate the improvement, we employ a Taylor expansion of the log-likelihood $l(\boldsymbol{\beta}_1, \mathbf{t}_1)$ around $\boldsymbol{\beta}_1 = (\widehat{\boldsymbol{\beta}}, 0)$, which specifies a function in G_1. This approximation yields the well-known Rao (score) statistic and is convenient because it allows us to entertain a large number of candidate knot locations without having to compute the MLE $\widehat{\boldsymbol{\beta}}_1$ in each candidate space.

Knot deletion. Again, let G be a given spline space and $\widehat{\boldsymbol{\beta}}$ the associated MLE. Removing a knot from G reduces the dimension of G by one and takes us to a space G_0. To evaluate the impact of this alteration, we again employ a Taylor expansion, this time around $\widehat{\boldsymbol{\beta}}$. If $\mathbf{a} \in \mathbb{R}^J$ represents the linear constraint that effectively removes a given knot, this expansion yields the Wald statistic for testing the hypothesis that $\mathbf{a}^t \boldsymbol{\beta} = 0$. For the truncated power basis, $\mathbf{a}$ is a binary vector with a single nonzero entry. With this approach, we can compare the impact of removing each knot in G without having to compute the MLE in these reduced spaces.

Alternating phases of knot addition and deletion produces a sequence of models, from which we select

the single best according to some selection criterion like generalized cross validation (GCV)

$$\text{GCV}_a(\mathbf{t}) = \frac{\text{RSS}(\mathbf{t})}{n} \Big/ \left[1 - \frac{a(J(\mathbf{t})-1)}{n}\right]^2, \tag{7}$$

or a variant of the Akaike information criterion (AIC)

$$\text{AIC}_a(\mathbf{t}) = -2\hat{l}(\mathbf{t}) + aJ(\mathbf{t}) \tag{8}$$

(Akaike, 1974), where $J(\mathbf{t})$ is the dimension of the spline space. The parameter a in each of these expressions controls the penalty assigned to models with more knots and is introduced to offset the effects of selection bias (Friedman and Silverman, 1989; Friedman, 1991). In Stone et al. (1997) the default value of a in (8) is $\log n$, resulting in a criterion that is commonly referred to as BIC (Schwarz, 1978).

Notice that our search for good knot locations based on the log-likelihood (5) has led to a heuristic minimization of a selection criterion like (7) or (8). Several comments about this reduction are in order. First, greedy schemes are often criticized for not exploring a large enough set of candidate models. In the stepwise algorithms of Stone et al. (1997), for example, the simple two-pass scheme (knot addition to a model of size M followed by deletion) evaluates essentially $2M$ different knot sequences. These $2M$ candidates are also highly constrained, representing a potentially narrow path through the search space. As a result, when we identify the "best model" according to some selection criterion, we have visited at most a handful of its "good-fitting" neighbors, those spline spaces with about the same number of knots found during either addition or deletion. However, as is typical with variable selection problems, many spline models offer essentially equivalent fits (in terms of AIC or GCV).

Despite these caveats, examples in Stone et al. (1997) and other papers show that greedy algorithms for knot selection can work quite well. They lead to a surprising amount of spatial adaptivity, easily locating extra knots near sharp features, while removing knots in smooth areas. It is natural, however, to question whether or not alternative methods might prove more effective. In the discussion following Stone et al. (1997), for example, the Bayesian framework of Smith and Kohn (1996) is shown to approximately minimize the same objective function (8), but with a stochastic search algorithm. In general, the recent work on Bayesian model selection offers interesting solutions to the shortcomings of greedy methods.

1.4 A Bayesian Approach

The desire to compare alternative search schemes is half the motivation for this paper. As mentioned earlier, a major source of inspiration comes from the recent work on Bayesian model selection and the accompanying Markov chain Monte Carlo (MCMC) methods for identifying promising models. To date, several Bayesian spline methods have appeared that make the connections with model selection listed above. The first was Halpern (1973), who constructed a hierarchical model for regression with linear splines. This application necessarily focused on small problems with a limited number of potential knots, succumbing to the computational resources of the day. More modern research in this area has followed a similar approach in terms of prior assignment, but makes use of MCMC to sample from a (possibly very) large set of candidate knots. Perhaps the first such procedure was exhibited by Smith (1996) and Smith and Kohn (1996) for univariate and additive regression models. Similar in spirit are the Bayesian versions of TURBO and CART proposed by Denison et al. (1998a, b), which employ reversible jump MCMC (Green, 1995).

In a Bayesian setup, *model uncertainty* comes from both the structural aspects of the space G—knot placement—as well as from our selection of members $g \in G$—determining coefficients in expression (3). We now spell out a simple hierarchical formulation that we will revisit in the next section. At the first level of the hierarchy, we assign a prior distribution $p(G)$ to some set of candidate models G. In the setup for univariate regression using linear splines, for example, we would typically do that by first choosing a prior distribution on the number of knots $p(K)$, and then by choosing an additional prior on the collection of knots $\mathbf{t}$ given K, $p(\mathbf{t}|K)$. Through $p(\mathbf{t}|K)$ we can prevent knots from getting too close, reducing the chance that the fitted model will track spurious features in the data. Next, given a space G, we generate elements g according to the distribution $p(g|G)$. Consistent with our motivation for modeling with splines in the first place, our priors on $K, \mathbf{t}$ and g should somehow reflect our beliefs about the smoothness of the underlying function of interest in an ELM, ϕ. In the literature on smoothing splines we find a class of priors for g that given a basis for G and an expansion (3) involves the coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$. This amounts to a partially improper, normal distribution for $\boldsymbol{\beta}$ (Silverman, 1985; Wahba, 1990; and Green and Silverman, 1994), which we will return to in Section 2.

Given a prior for spline functions g, we can generate a sample from the posterior distribution of g using MCMC. In particular, in Sections 2 and 3 we will use the reversible jump algorithm of Green (1995) for Logspline density estimation and Triogram regression, respectively. Details about how to choose priors and how to tune algorithms are discussed in these sections. When properly tuned, these stochastic methods can identify many more good-fitting knot configurations than their greedy, deterministic competitors. By focusing their search in regions of model space that have high posterior probabilities, the MCMC schemes listed above visit many more "promising" configurations.

The second major motivation for this paper is the form of the final function estimate itself. Since deterministic searches identify a very small number of usable models, the unknown function is typically estimated by straight maximum likelihood applied to some basis for the identified spline space. Suppose for the moment, that the function being estimated is smooth in some region, perhaps requiring not more than a single knot to adequately describe the curve. From the point of view of mean squared error, there are many roughly equivalent ways to place this knot in the region. Therefore, if given a number of good knot configurations, it might be more reasonable to combine these estimates in some way. This is roughly a spline or knot-selection version of the classical motivation for Bayesian model averaging. In later versions of the Gibbs sampling approach of Smith and Kohn (1998) and the Bayesian versions of TURBO and MARS by Denison, Mallick and Smith (1998a, b), the final function estimate is a posterior mean.

In this paper, we compare greedy (stepwise) algorithms with nongreedy (stochastic, Bayesian) algorithms for model selection. We evaluate different approaches to adaptation by examining strategies for both knot placement and coefficient estimation. We focus on four classes of methods: greedy, stepwise procedures with maximum likelihood estimates in the final spline space; MCMC for selecting a single model; model averaging using maximum likelihood estimates of the coefficients; and finally a fully Bayesian approach with model and coefficient averaging. Our two main estimation problems will be Logspline density estimation and (bivariate) Triogram regression. We selected these because they illustrate a number of computational and methodological issues concerning model adaptation that arise in ELMs.

In Section 2 we discuss greedy and Bayesian model selection approaches in the context of Logspline density estimation. In Section 3 we turn to Triogram regression, contrasting it with Logspline. Finally, in Section 4 we identify areas of future research. Our goal in preparing this paper was not to advocate one scheme over another, but rather to investigate the performance of various approaches to model selection in the context of univariate and multivariate nonparametric estimation with splines.

2. LOGSPLINE DENSITY ESTIMATION

Recall that density estimation is an example of an ELM. In the notation of the previous section, the target of our analysis, ϕ, is a log-density, and $W = Y$, a random variable taking values in some interval (L, U). If the density of Y has infinite support, then L, U will be $\pm\infty$. In Stone and Koo (1986), Kooperberg and Stone (1991, 1992) and Stone et al. (1997), a technique known as Logspline is developed in which ϕ is modeled with a *natural cubic spline.* Like the ordinary cubic splines in (2), these functions are also twice continuously differentiable, piecewise polynomials defined relative to a knot sequence $\mathbf{t} = (t_1, \ldots, t_K)$. Within each interval $[t_1, t_2], \ldots, [t_{K-1}, t_K]$, natural cubic splines are cubic polynomials, but on $(L, t_1]$ and $[t_K, U)$ they are forced to be linear functions. It is not difficult to see that this *tail constraint* again yields a linear space, but with dimension K. Also, the space will contain spline terms providing we have at least $K \geq 3$ knots (otherwise we have only linear or constant functions). In this application, we use a basis of the form $1, B_1(y; \mathbf{t}), \ldots, B_J(y; \mathbf{t})$, where $J = K - 1$. We chose to make the constant term explicit in this way because it disappears from our model; recall that each density estimate is normalized to integrate to one. Therefore, let G denote the J-dimensional span of the functions $B_1, \ldots, B_J$. So that $g \in G$ is of the form $g(y; \boldsymbol{\beta}, \mathbf{t}) = \beta_1 B_1(y; \mathbf{t}) + \cdots + \beta_J B_J(y; \mathbf{t})$.

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

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

Let $\mathcal{B}$ denote the collection of such feasible column vectors. Given $\boldsymbol{\beta} \in \mathcal{B}$, we define a family of positive

density functions on (L, U) of the form

$$f(y;\boldsymbol{\beta},\mathbf{t}) = \exp\big(g(y;\boldsymbol{\beta},\mathbf{t}) - C(\boldsymbol{\beta},\mathbf{t})\big)$$
$$= \exp\big(\beta_1 B_1(y;\mathbf{t}) + \cdots + \beta_J B_J(y;\mathbf{t}) - C(\boldsymbol{\beta},\mathbf{t})\big), \qquad L < y < U. \tag{9}$$

Now, given a random sample $Y_1, \ldots, Y_n$ of size n from a distribution on (L, U) having an unknown density function $\exp(\phi)$, the log-likelihood function corresponding to the Logspline model (9) is given by

$$l(\boldsymbol{\beta},\mathbf{t}) = \sum_i \log f(Y_i;\boldsymbol{\beta},\mathbf{t}) = \sum_i \sum_j \beta_j B_j(Y_i;\mathbf{t}) - nC(\boldsymbol{\beta};\mathbf{t}), \quad \boldsymbol{\beta} \in \mathcal{B}.$$

The maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ is given by $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathcal{B}} l(\boldsymbol{\beta},\mathbf{t})$, corresponding to $\hat{g}(y) = g(y;\widehat{\boldsymbol{\beta}},\mathbf{t})$ for $L < y < U$.

Stepwise knot addition begins from an initial model with K_{init} knots, positioned according to the rule described in Kooperberg and Stone (1992). Given a knot sequence $t_1, \ldots, t_K$, the addition scheme finds a location for a candidate knot corresponding to the largest Rao statistic. For numerical stability, we do not allow the breakpoints $t_1, \ldots, t_K$ to be separated by fewer than n_{sep} data points. We say that in this context, a space G is *allowable*, providing the knot sequence satisfies this condition. Stepwise addition continues until a maximum number of knots $K_{\max}$ is reached. Knot deletion is then performed according to the outline in the previous section, and a final model is selected according to the generalized AIC criterion (8) with parameter $a = \log n$.

2.1 A Bayesian Framework

We set up the framework for a Bayesian approach to Logspline density estimation by selecting several priors: first a prior $p(G)$ on the structure of the model space G, and then a prior $p(g|G)$ on the splines g in a given space. In addition, we will need to specify how we sample from the posterior distributions.

Priors on model space. For Logspline we choose to specify $p(G)$ by creating a distribution on knot sequences $\mathbf{t}$ formed from some large collection of candidates $\mathcal{T} = \{t'_1, \ldots, t'_{K'}\}$. We construct $p(G)$ hierarchically, first choosing the number of knots $K < K'$ (in this case recall that the dimension J of G is $K - 1$) according to $p(K)$, and then given K, we generate $\mathbf{t}$ from the distribution $p(\mathbf{t}|K)$. Regularity conditions on the structural aspects of the associated spline space G can be imposed by restricting the placement of $t_1, \ldots, t_K$ through $p(\mathbf{t}|K)$. While other authors have also considered a discrete set of candidate knot sequences (Denison, Mallick and Smith, 1998a; Smith and Kohn, 1996), we could also specify a distribution that treats the elements of $\mathbf{t}$ as continuous variables (e.g., Green 1995). In our experiments we have found that for Logspline density estimation the discrete approach is sufficient, and we consider those spaces G for which all K knots are located at data points. This restriction is purely for convenience, but represents little loss of flexibility especially in the context of density estimation (where peaks in the underlying density naturally produce more candidate knots). For numerical stability, we require that there are at least n_{sep} data points in between any two knots.

This leaves us with the task of specifying $p(K)$. To the extent that the number of knots also acts as a smoothing parameter, this distribution can have a considerable effect on the look of the final curves produced. We explore several of the proposals that have appeared in the literature. The first is a simple Poisson distribution with mean γ suggested by Green (1995). Denison et al. (1998a) take the same distribution for more general spline spaces and argue that their results are somewhat insensitive to the value of γ. The next prior we will consider was suggested by Smith and Kohn (1996). Either by greatly reducing the number of candidate knots or by scaling the prior on the coefficients, these authors suggest that K be distributed uniformly on the set $K_{\min}, \ldots, K_{\max}$.

The final proposal for $p(K)$ is somewhat more aggressive in enforcing small models. To properly motivate this distribution, we think of the model selection procedure as two stages: in the first we find the posterior average of all models with k knots by integrating out $\mathbf{t}$ and g, to obtain, say $\bar{g}_k$ and its posterior probability $p(\bar{g}_k|Y_1, \ldots, Y_n, K = k)$. Suppose that we consider $\bar{g}_k$ to have k degrees of freedom (an admittedly questionable assumption). If we now were to use an AIC-like criterion to choose among the $\bar{g}_k$, we would select the model that minimized

$$-2\log p(\bar{g}_k|Y_1, \ldots, Y_n, K = k) + ak,$$

compare (8). On the other hand, using the posterior to evaluate the best model suggests maximizing

$$p(\bar{g}_k|Y_1, \ldots, Y_n, K = k)\, p(K = k).$$

If we take $p(K = k) \propto \exp(-ak/2)$ these two approaches agree. Thus, taking a geometric distribution

for $p(K)$ implies an AIC-like penalty on model dimension. In particular $a = \log n$ and $q = 1/\sqrt{n}$ imposes the same cost per knot as AIC with penalty $\log n$. For reasonable settings of $K_{\min}$ and $K_{\max}$, however, the expected prior number of knots under this prior will tend to zero with n. While it is certainly intuitive that the prior probability of K decreases monotonically with k, this drop may be at a faster rate than we would expect! If $a \geq 2$ then $p(K = k+1)/p(K = k) \leq 1/e$.

Priors on splines in a given space. We parameterize $p(g|G)$ through the coefficients $\boldsymbol{\beta}$ in the expansion (3), and consider priors on $\boldsymbol{\beta}$ that relate to our assumptions about the smoothness of g. Recall that as the solution to a penalized maximum likelihood fit, smoothing splines (Wahba, 1990) have a straightforward Bayesian interpretation (Silverman, 1985). In univariate smoothing, for example, G is a space of natural splines (given some knot sequence $\mathbf{t}$), and the "roughness" of any $g \in G$ is measured by the quantity $\int_L^U (g'')^2$. Expanding g in a basis, it is not hard to see that

$$\int_L^U (g'')^2 = \boldsymbol{\beta}' A \boldsymbol{\beta} \quad \text{where} \quad A_{ij} = \int_L^U B_i''(x) B_j''(x)\, dx \quad \text{for } 1 \leq i, j \leq J. \tag{10}$$

The traditional smoothing spline fit maximizes the penalized likelihood

$$\arg\max_{\boldsymbol{\beta}} \{l(\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}' A \boldsymbol{\beta}\},$$

for some parameter λ. Silverman (1985) observes that the solution to this problem can be viewed as a posterior mode, where $\boldsymbol{\beta}$ is assigned a partially improper, normal prior having mean $\mathbf{0}$ and variance-covariance matrix $(\lambda A)^{-1}$. This setup has the favorable property that it is invariant to our choice of basis. This is desirable, as the choice of the basis will often be made for computational reasons.

In our simulations we will compare this smoothing prior to the scheme of Denison et al. (1998a) in which no stochastic structure is assigned to the coefficients $\boldsymbol{\beta}$ once G is selected. Instead, these authors employ maximum likelihood to make a deterministic choice of $\boldsymbol{\beta}$.

Markov chain Monte Carlo. In order to treat a variety of estimation problems simultaneously, we have chosen the reversible jump MCMC scheme developed by Green (1995). Denison et al. (1998a) implement this technique in the context of general univariate and additive regression. We refer to these papers for the details of the scheme, and we instead focus on the type of moves that we need to implement the sampler. In general, we alternate (possibly at random) between the following moves.

- *Increase model dimension.* In this step, we introduce a new knot into an existing collection of breakpoints. Given the concavity properties of ELMs the change in the log-likelihood could either be computed exactly or approximated using the appropriate Rao statistic. In our experiments we have computed the change in the log-likelihood exactly. The new knot is selected uniformly from among the set that yields an allowable space.
- *Decrease model dimension.* As with the greedy scheme, knots are deleted by imposing a constraint on one or more coefficients in the spline expansion. We can either evaluate the drop in the log-likelihood exactly, or through the Wald statistics. Any knot can be removed at any time (assuming we have more than $K_{\min}$ breakpoints to chose from).
- *Make structural changes to G that do not change dimension.* Unlike our standard greedy scheme, nonnested steps like moving a knot are now possible. Moving a knot from t_k to t_k^* technically involves deleting t_k and then inserting a new breakpoint at t_k^*. With smart initial conditions on the Newton–Raphson steps, we can calculate the change in the log-likelihood exactly and still maintain an efficient algorithm.
- *Update (possibly) g.* In a nonlinear model like Logspline, we can either apply a suitable approximation to the posterior and integrate with respect to the coefficients $\boldsymbol{\beta}$, or we can fold sampling them into our Markov chain.

Following Green (1995) and Denison et al. (1998a), we cycle between proposals for adding, deleting and moving knots, assigning these moves probabilities b_J, d_J and $1 - b_J - d_J$ (see Denison et al., 1998a). New knots can be positioned at any data point that is at least n_{sep} data points removed from one of the current knots. Subject to this constraint, knot addition follows a simple two step procedure. First, we select one of the intervals $(L, t_1), (t_1, t_2), \ldots, (t_K, U)$ uniformly at random (where the t_k are the current breakpoints). Within this interval, the candidate knot is then selected uniformly at random from one of the allowable data points. When moving a knot, we either propose a large move (in which a knot is first deleted, and then added using the addition scheme just described) or a small

move (in which the knot is only moved within the interval between its two neighbors). Each of these two proposals have probability $(1 - d_J - b_J)/2$.

After each reversible jump step, we update the coefficients $\boldsymbol{\beta}$. To do this, we use the fact that for a given set of knots, we have a parametric model, and that the posterior distribution of $\boldsymbol{\beta}$ given G and the data is thus approximately multivariate normal with covariance matrix $\Sigma = (\lambda A + H)^{-1}$, and mean $\Sigma H\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ in G, and H is the Hessian of the log-likelihood function at $\widehat{\boldsymbol{\beta}}$. An observation from this distribution is used as a proposal in a Metropolis step. Because we are using (partially improper) smoothing priors, the acceptance ratio for this proposal is formally undetermined (recall that the prior covariance matrices are degenerate). We solve this problem by "canceling" the zero eigenvalue in the numerator and the denominator (see also Besag and Higdon, 1999).

2.2 A Simulation Study

To compare the performance of the various possible implementations of Logspline density model selection procedures, we carried out a simulation study. We generated data from three densities:

- *normal*—the standard normal density;
- *slight bimodal*—$f(y) = 0.5 f_Z(y; 1.25, 1) + 0.5 f_Z(y; -1.25, 1.1)$, where $f_Z(y; \mu, \sigma)$ is the normal density with mean μ and standard deviation σ;
- *sharp peak*—$f(y) = 0.8 g(y) + 0.2 f_Z(y; 2, 0.07)$, where $g(Y)$ is the density of the lognormal random variable $Y = \exp(Z/2)$ and Z has a standard normal distribution.

These three densities are displayed in Figure 1. From each we generated 100 independent samples of size $n = 50$, 200, 1,000 and 10,000. We applied a variety of Logspline methods, see Table 1. For all the Bayesian methods we estimated the posterior mean by a simple pointwise average of the MCMC samples. Otherwise, the Bayesian approaches differ in two aspects:

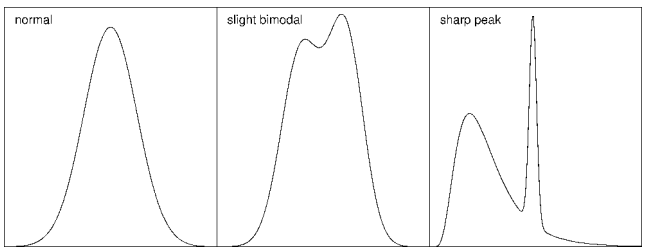

FIG. 1. *Densities used in the simulation study.*

TABLE 1
Versions of Logspline density estimation used in the simulation study

	Model size	**Parameters**
(i)	Greedy optimization of AIC proposed by Stone et al. (1997)	
(ii)	Simulated annealing optimization of AIC (SALSA)	
(iii)	Geometric	ML
(iv)	Poisson (5)	ML
(v)	Uniform	$\lambda = 1/n$
(vi)	Uniform	$\lambda = 1/\sqrt{n}$
(vii)	Uniform	$\lambda = 1$
(viii)	Geometric	$\lambda = 1/n$

- the prior on the model size—we used the geometric prior with parameter $p = 1 - 1/\sqrt{n}$, the Poisson prior with parameter 5, and a uniform prior;
- parameter estimates $\widehat{\boldsymbol{\beta}}$—we took either the maximum likelihood (ML) estimate, or we assigned a multivariate normal prior to $\boldsymbol{\beta}$ (for one of several choices for λ).

Table 1 summarizes the versions of Logspline which are reported here.

For simulated annealing (ii) (termed SALSA for "Simulated Annealing LogSpline Approximation") we ran the same MCMC iterations as for version (iii), but rather than selecting the mean of the sampled densities, we chose the density which minimizes AIC. As described above this is very similar to taking the density with the largest a posteriori probability (the mode), except that we ignore the prior on knot locations given the number of knots, K. This would have changed the penalty in the AIC criterion from $K \log n$ to $K \log n + \frac{1}{2} \log \binom{n}{K}$. Since version (ii) begins with the fit obtained by the greedy search (i), it is guaranteed to improve as far as AIC is concerned. Version (iii) uses the same penalty structure as version (ii), but averages over MCMC samples. Version (iv) is included since a Poisson (5) prior was proposed by Denison et al. (1998a). It applies a considerably smaller penalty on model size. Versions (v)–(viii) experiment with penalties on the coefficients. Generating the parameters using a multivariate normal prior distribution implies smoothing with a AIC-like penalty. As such, we would expect that using $\lambda = 1/n$ with a uniform prior [version (v)] may give reasonable results, but that using a geometric prior [version (ix)] would smooth too much. Choosing λ too large, as in versions (vi)–(vii), leads to oversmoothing, while choosing λ too small tends to produce overly wiggly fits.

TABLE 2
Mean integrated squared error (MISE) for the simulation study

		Version							
		(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
Distribution	***n***	**MISE**	**Ratio of MISE over MISE of the greedy version (i)**						
Normal	50	0.02790	0.73	1.52	1.84	0.66	0.40	0.26	0.67
Normal	200	0.01069	0.49	0.60	1.23	0.79	0.50	0.24	0.66
Normal	1,000	0.00209	0.59	0.58	1.33	0.87	0.90	0.42	0.73
Bormal	10,000	0.00020	0.33	0.49	1.45	1.35	1.10	0.80	0.87
Slight bimodal	50	0.02502	0.88	1.09	1.34	0.48	0.36	0.36	0.50
Slight bimodal	200	0.00770	0.80	0.61	1.14	0.70	0.38	0.46	0.61
Slight bimodal	1,000	0.00164	0.57	0.60	1.13	0.89	0.66	0.40	0.77
Slight bimodal	10,000	0.00020	0.77	0.61	0.88	0.71	0.82	0.51	0.84
Sharp peak	50	0.15226	0.97	0.78	0.81	0.68	0.90	1.12	0.72
Sharp peak	200	0.03704	0.89	0.75	0.94	0.93	2.02	3.62	1.13
Sharp peak	1,000	0.00973	0.81	0.67	0.81	0.67	2.01	8.90	0.74
Sharp peak	10,000	0.00150	0.72	0.57	0.57	0.64	0.58	21.43	0.76
	Average	1.00	0.71	0.74	1.12	0.78	0.89	3.21	0.75

For versions (iii) and (iv) we ran 600 MCMC iterations, of which we discarded the first 100 as burn-in. Some simple diagnostics (not reported) suggest that after 100 iterations the chain is properly mixed. For versions (v)–(viii) each structural change was followed by an update of the coefficients $\boldsymbol{\beta}$.

In Table 2, we report ratios of integrated squared errors between the greedy scheme and the other methods outlined above. In addition, we feel that it is at least as important for a density estimate to provide the correct general "shape" of a density as to have a low integrated squared error. To capture the shape of our estimates, we counted the number of times that a scheme produced densities having too few, too many and the correct number of modes. These results are summarized in Tables 3 and 4. Table 5 calculates the "total" lines of Tables 3 and 4. Note that for simulations of a normal distribution it is not possible for an estimate to have too few modes.

From Table 2 we note that most methods show a moderate overall improvement over the greedy version of Logspline, except for (vii). This scheme oversmoothes the data, so that the details (like the mode in the sharp-peaked distribution) are frequently missed. We note that version (iii), choosing the mode of a Bayesian approach, is the only version that outperforms the greedy version for all 12 simulation setups. Otherwise, the difference between versions (ii), (iii), and (viii) seems to be minimal. In particular, if we had chosen another set of results than those for (i) to normalize by, the order of the average MISE for these four methods was often changed.

TABLE 3
Number of times out of 100 *simulations that a Logspline density estimate had too few modes*

		Version							
Distribution	***n***	**(i)**	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**	**(vii)**	**(viii)**
Slight bimodal	50	45	52	4	0	21	74	99	31
Slight bimodal	200	6	22	13	0	1	18	96	19
Slight bimodal	1,000	5	17	19	0	7	6	45	16
Slight bimodal	10,000	4	12	4	1	3	4	2	10
Sharp peak	50	24	38	1	0	9	56	99	13
Sharp peak	200	0	1	0	0	0	0	89	1
Sharp peak	1,000	0	0	0	0	0	0	0	0
Sharp peak	10,000	0	0	0	0	0	0	0	0
	Total	84	142	41	1	41	158	430	90

TABLE 4
Number of times out of 100 *simulations that a Logspline density estimate had too many modes*

		Version							
Distribution	***n***	**(i)**	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**	**(vii)**	**(viii)**
Normal	50	18	11	94	100	49	5	0	28
Normal	200	34	9	38	100	81	21	0	24
Normal	1,000	26	4	15	91	68	54	32	32
Normal	10,000	4	1	7	61	31	29	1	17
Slight bimodal	50	4	1	84	99	6	0	0	4
Slight bimodal	200	16	1	19	99	55	4	0	5
Slight bimodal	1,000	15	1	13	93	51	31	1	17
Slight bimodal	10,000	6	1	8	68	33	39	0	6
Sharp peak	50	15	8	90	93	3	1	0	2
Sharp peak	200	36	19	46	94	43	5	0	5
Sharp peak	1,000	28	14	30	77	32	12	1	9
Sharp peak	10,000	25	12	15	31	20	30	11	7
	Total	227	82	459	1006	472	231	46	156

From Table 3 we note that version (vii), and to a lesser extent (ii) and (vi), have trouble with the slight bimodal density, preferring a model with just one peak. Versions (vi) and (vii) find too few modes, leading us to conclude that λ should be chosen smaller than $1/\sqrt{n}$ when using a uniform prior on model size. On the other hand, the Poisson prior leads to models exhibiting too many peaks, as do versions (iii) and (v).

Overall, it appears that the greedy, stepwise search is not too bad. It is several orders of magnitude faster than any of the other methods. The greedy approach, as well as SALSA have the advantage that the final model is again a Logspline density, which can be stored for later use. For the other methods, we must record the posterior mean at a number of points. This has the potential of complicating later uses of our estimate. Among the Bayesian versions that employ ML estimates, version (iii) seems to perform best overall, while among those that put a prior on the coefficient vector, versions (v) and (viii) (both of which set $\lambda = 1/n$) are best. It is somewhat surprising that version (viii) performs so well, since it effectively imposes twice the AIC penalty on model size: one coming from the geometric prior, and one from the normal prior on the parameters. Kooperberg and Stone (1992) argue that the Logspline method is not very sensitive to the exact value of the parameter, possibly explaining the behavior of version (viii). In Kooperberg and Stone (2002) a double penalty is also employed in the context of free knot Logspline density estimation.

2.3 Income Data

We applied the nine versions of Logspline used for the simulation study to the income data discussed in Stone et al. (1997), and the results are displayed in Figure 2. For the computations on the income data we ran the MCMC chain for 5000 iterations in which a new model was proposed, after discarding the first 500 iterations for burn-in. For the versions with priors on the parameters we alternated these iterations with updates of the parameters. The estimates for versions (ii), which was indistinguishable from version (iii), and versions (viii) which was indistinguishable from version (v) are not shown. In Kooperberg and Stone (1992) it was argued that the height of the peak should be at least about 1. Thus, it appears that versions

TABLE 5
Number of times that a Logspline density estimate had an incorrect number of modes

	(i)	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**	**(vii)**	**(viii)**
Too few modes	84	142	41	1	41	158	430	90
Too many modes	227	82	459	1,006	472	231	46	156
Total	311	224	500	1,007	513	389	476	246

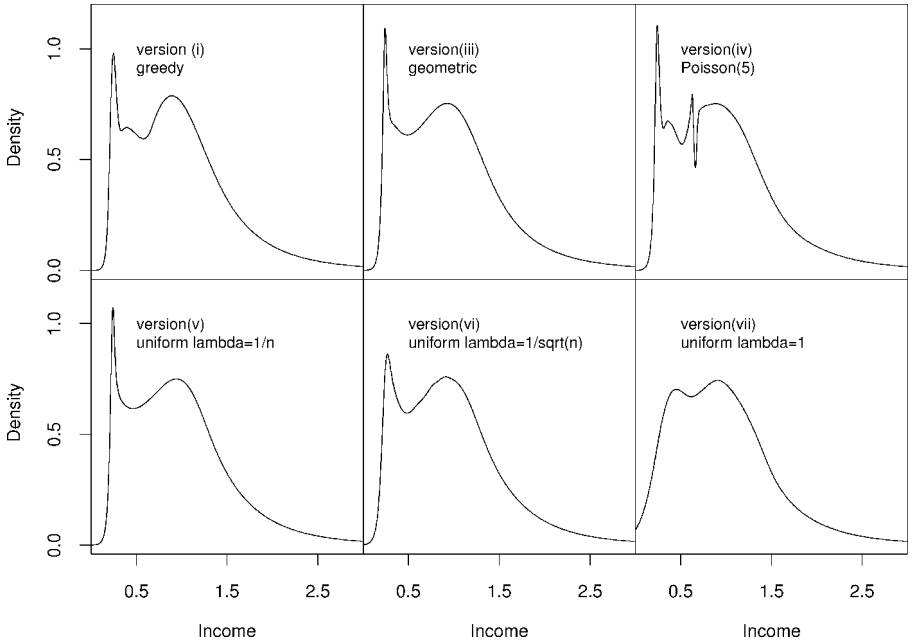

FIG. 2. *Logspline density estimates for the income data.*

(vi) and (vii) have oversmoothed the peak. On the other hand, version (iv) seems to have too many small peaks.

It is interesting to compare the number of knots for the various schemes. The greedy estimate (version i) has 8 knots, and the simulated annealing estimate (version ii) has 7 knots. The Bayesian versions (iii), (v) and (viii) have an average number of knots between 5 and 8, while the three versions that produced unsatisfactory results (iv, vi and vii) have an average number of knots between 14 and 17.

The MCMC iterations can also give us information about the uncertainty in the knot locations. To study this further, we ran a chain for version (iii) with 500,000 iterations. Since the knots are highly correlated from one iteration to the next (at most one knot moves at each step), we only considered every 250th iteration. The autocorrelation function of the fitted log-likelihood suggested that this was well beyond the time over which iterations are correlated. This yielded 2,000 sets of knot locations: 1,128 with five knots, 783 with six knots, 84 with seven knots, and 5 with eight knots. When there were five knots, the first three were always located close to the mode, the fourth one was virtually always between 0.5 and 1.25, and the last knot between 1 and 2. The locations of the first three knots overlap considerably. When there are six knots, the extra knot can either be a fourth knot in the peak, or it is beyond the fifth knot.

3. TRIOGRAM REGRESSION

When estimating a univariate function ϕ, our "pieces" in a piecewise polynomial model were intervals of the form (t_k, t_{k+1}). Through knot selection, we adjusted these intervals to capture the major features in ϕ. When ϕ is a function of two variables, we have more freedom in how we define a piecewise polynomial model. In this section we take our separate pieces to be triangles in the plane, and consider data-drive-techniques that adapt these pieces to best fit ϕ. Our starting point is the Triogram methodology of Hansen et al. (1998) which employs continuous, piecewise linear (planar) bivariate splines. Triograms are based on a greedy, stepwise algorithm that builds on the ideas in Section 1 and can be applied in the context of any ELM where ϕ is a function of two variables. After reviewing some notation, we present a Bayesian version of Triograms for ordinary regression. An alternative approach to piecewise linear modeling was proposed in Breiman (1993) and given a Bayesian extension in Holmes and Mallick (2001).

Let $\triangle$ be a collection of triangles δ (having disjoint interiors) that partition a bounded, polygonal region in the plane $\mathcal{X} = \bigcup_{\delta\in\triangle}\delta$. The set $\triangle$ is said to be a *triangulation* of $\mathcal{X}$. Furthermore, $\triangle$ is *conforming* if the nonempty intersection between pairs of triangles in the collection consists of either a single, shared vertex or an entire common edge. Let $\mathbf{v}_1, \ldots, \mathbf{v}_K$ represent the collection of (unique) vertices of the triangles in $\triangle$.

Over $\mathcal{X}$, we consider the collection G of continuous, piecewise-linear functions which are allowed to break (or hinge) along the edges in $\triangle$. It is not hard to show that G is a linear space having dimension equal to the number of vertices K. A simple basis composed of "tent functions" was derived in Courant (1943): for each $j = 1, \ldots, K$, we define $B_j(\mathbf{x}; \triangle)$ 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 triangulation. The set $B_1(\mathbf{x}; \triangle), \ldots, B_K(\mathbf{x}; \triangle)$ is a basis for G. Also notice that each function $B_j(\mathbf{x}; \triangle)$ is associated with a single vertex $\mathbf{v}_j$, and in fact each $g \in G$

$$g(\mathbf{x}; \boldsymbol{\beta}, \triangle) = \sum_{j=1}^{K} \beta_j B_j(\mathbf{x}; \triangle), \tag{11}$$

interpolates the coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ at the points $\mathbf{v}_1, \ldots, \mathbf{v}_K$.

We now apply the space of linear splines to estimate an unknown regression function. In the notation of an ELM, we let $W = (\mathbf{X}, Y)$, where $\mathbf{X} \in \mathcal{X}$ is a two-dimensional predictor and Y is a univariate response. We are interested in exploring the dependence of Y on $\mathbf{X}$ by estimating the regression function $\phi(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. Given a triangulation $\triangle$, we employ linear splines over $\triangle$ of the form (11). For a collection of (possibly random) design points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ taken from $\mathcal{X}$ and corresponding observations $Y_1, \ldots, Y_n$, we apply ordinary least squares to estimate $\boldsymbol{\beta}$. That is, we take $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_i [Y_i - g(\mathbf{X}_i; \boldsymbol{\beta}, \triangle)]^2$, and use $\widehat{g}(\mathbf{x}) = g(\mathbf{x}; \widehat{\boldsymbol{\beta}}, \triangle)$ as an estimate for ϕ.

As with the univariate spline models, we now consider stepwise alterations to the space G. Following Hansen, Kooperberg and Sardy (1998), the one-to-one correspondence between vertices and the "tent" basis functions suggests a direct implementation of the greedy schemes in Section 1. Stepwise addition involves introducing a new vertex into an existing triangulation, thereby adding one new basis function to the original spline space. This operation requires a rule for connecting the new point to the vertices in $\triangle$ so that the new mesh is again a conforming triangulation. In Figure 3, 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 an add a point to the interior of one of the triangles in $\triangle$. Given a triangulation $\triangle$, candidate vertices are selected from a regular triangular grid in each of the existing triangles, as well as a number of locations on each of the existing edges (for details see Hansen et al., 1998).

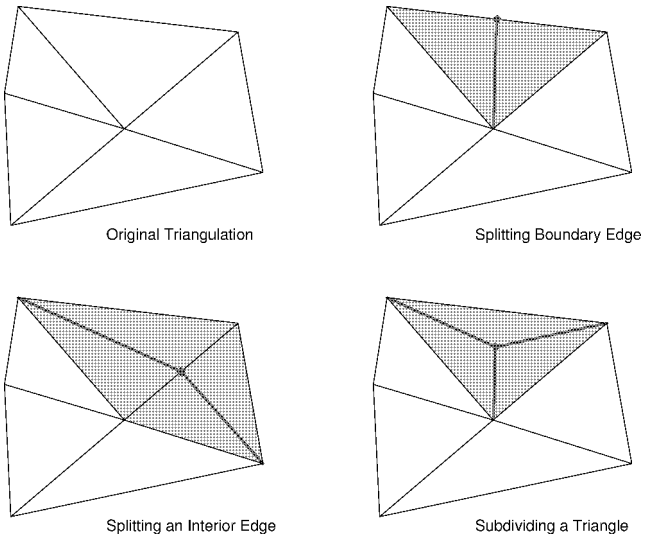

FIG. 3. *Three "moves" that add a new vertex to an existing triangulation. Each addition represents the introduction of a single basis function, the support of which is colored gray.*

We impose constraints on our search by limiting, say, the area of the triangles in a mesh, their aspect ratio, or perhaps the number of data points they contain. As with Logspline, spaces satisfying these restrictions are referred to as allowable. At each step in the addition process, we select from the set of candidate vertices (that result in an allowable space), the point that maximizes the decrease in residual sum of squares when the Triogram model (11) is fitted to sample data. (In regression, the Rao and Wald statistics are the same and reduce to the change in the residual sum of squares between two nested models.)

Deleting a knot from an existing triangulation can be accomplished most easily by simply reversing one of the steps in Figure 3. Observe that removing a vertex in one of these three settings is equivalent to enforcing continuity of the first partial derivatives across any of the "bold edges" in this figure. Such continuity conditions are simple linear constraint on the coefficients of the fitted model, allowing us to once again apply a Wald test to evaluate the rise in the residual sum of squares after the vertex is deleted.

3.1 A Bayesian Framework

Priors on model space. As with univariate spline models, a prior on the space of Triograms is most easily defined by first specifying the structure of the approximation space, which in this case is a triangulation $\triangle$. For any $\triangle$, we need to select the number of vertices K, their placement $\mathbf{v}$, and the triangles that connect them.

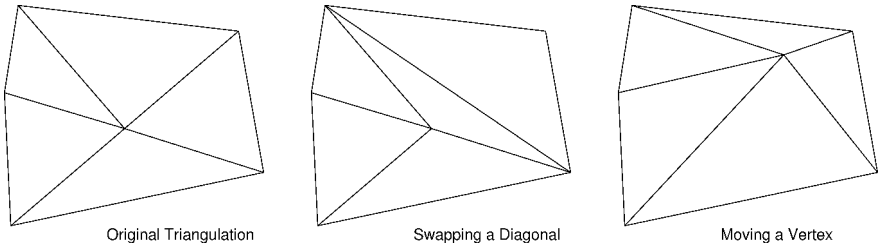

FIG. 4. *Additional structural moves for the reversible jump MCMC scheme. Note that these two proposals result in a nonnested sequence of spaces.*

Each set $\mathbf{v}$ can be joined by a number of different triangulations (assuming $\mathbf{v}$ has more than 3 points). Sibson (1978) shows that by starting from one triangulation of $\mathbf{v}$, we can generate any other by a sequence of "edge swaps." (This operation is given in Figure 4 and will come up later when we discuss MCMC for bivariate splines.) Unfortunately, a closed-form expression for the number of triangulations associated with a given set of vertices does not exist. Computing this number for even moderately sized configurations is difficult because two sets each with K vertices can have different numbers of triangulations.

To see how this complicates matters, suppose we follow the strategy for Logspline and propose a hierarchical prior of the form

$$p(\triangle|\mathbf{v}, K)p(\mathbf{v}|K)p(K), \tag{12}$$

where $\triangle$ is a triangulation of the vertices $\mathbf{v} = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$. Assigning any proper distribution to $\triangle$ given $\mathbf{v}$ introduces a normalizing constant in $p(\triangle|\mathbf{v}, K)$ that involves enumerating the different triangulations of $\mathbf{v}$. Therefore, when taking ratios of (12) for two different sets of vertices, we are usually left with a prohibitively expensive computational problem. MCMC methods for exploring the model space are not possible.

To avoid this problem, we will use a tractable prior on triangulations developed by Nicholls (1998). This distribution depends on a pair of Poisson point processes, one that generates vertices on the interior of $\mathcal{X}$ and one for the boundary. As constructed, there is one parameter β that controls the intensity of this process, where larger values of β produce triangulations with more vertices. Nicholls (1998) avoids counting triangulations by normalizing across all triangulations obtainable from all vertex sets generated by this point process, and produces a distribution $p(\triangle)$. Bounds on the number of triangulations obtainable from a given vertex set are used to show that this kind of normalization is possible. This construction also has the advantage that restrictions on the size and shape of triangles are easily enforced and only change the (global) normalization constant in $p(\triangle)$. In our experiments, we set β so that the expected number of vertices for this base process is 5. We then adapted Nicholls's approach, so that the underlying point process produces a geometric (with parameter $1 - 1/\sqrt{n}$) or a uniform (on $K_{\min}, \ldots, K_{\max}$) number of vertices, following the simulation setup in the previous section.

Priors on splines in a given space. Unlike the Logspline example, we do not have a single obvious choice for the smoothing prior for linear splines $g \in G$ defined relative to a triangulation $\triangle$. Dyn, Levin and Rippa (1990a, b) propose several criteria of the form

$$\sum_e s^2(g, e) \quad \text{for } g \in G,$$

where the summation is over all edges in $\triangle$. Their cost function $s(g, e)$ evaluates the behavior of g along an edge, assigning greater weight when the hinged linear pieces are farther from a single plane. Koenker and Mizera (2001) elegantly motivate a cost function $s(g, e) = \| \bigtriangledown g_e^+ - \bigtriangledown g_e^- \| \cdot \|e\|$, where $\bigtriangledown g_e^+$ and $\bigtriangledown g_e^-$ are the gradients of g computed over the triangles that share the common edge e having length $\|e\|$. This is similar to the approach taken by Nicholls (1998) who derived an edge-based smoothness penalty for piecewise constant functions defined over triangulations.

We choose to work with the cost function of Koenker and Mizera (2001). It is not hard to show that this gives rise to a quadratic penalty on the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ which can be written $\boldsymbol{\beta}^t A \boldsymbol{\beta}$ for a positive-semidefinite matrix A. Since constant and linear functions have zero roughness by this measure, A has two zero eigenvalues. As was done for Logspline, we use A to generate a partially improper normal prior on $\boldsymbol{\beta}$ (with prior variance σ^2/λ, where σ^2 is the error variance). Following Denison et al. (1998a), we assign a proper, inverse-gamma distribution to σ, and experiment with various fixed choices for λ that depend on sample size.

Markov chain Monte Carlo (*MCMC*). Our approach to MCMC for Triograms is similar to that with Logspline except that we need to augment our set of structural changes to $\triangle$ to include more moves than simple vertex addition and deletion. In Figure 4, we present two additional moves that maintain the dimension of the space G but change its structure. The middle panel illustrates swapping an edge, an operation that we have already noted is capable of generating all triangulations of a given vertex set $\mathbf{v}$. Quak and Schumaker (1991) use random swaps of this kind to come up with a good triangulation for a fixed set of vertices. In in the final panel of Figure 4, we demonstrate moving a vertex inside the union of triangles that contain it. These changes to $\triangle$ are non-nested in the sense that they produce spline spaces that do not differ by the presence or absence of a single basis function. For Triograms, the notion of an allowable space can appear through size or aspect ratio restrictions on the triangulations, and serves to limit the region in which we can place new vertices or to which we can move existing vertices. For example, given a triangle, the set into which we can insert a new vertex and still maintain a minimum area condition is a subtriangle, easily computable in terms of barycentric coordinates (see Hansen et al., 1998). As with Logspline, we alternate between these structural moves and updating the model parameters, following essentially the recipe in Denison et al. (1998a). Because we are working with regression, we can integrate out $\boldsymbol{\beta}$ and only have to update σ^2 at each pass. This approach allows us to focus on structural changes as was done by Smith and Kohn (1996) for univariate regression. [Of course, we can also integrate out σ^2, but to retain consistency with Denison et al. (1998a) we chose to sample.]

3.2 Simulations

In Figure 5, we present a series of three fits to a simulated surface plotted in the upper lefthand corner. A data set consisting of 100 observations was generated by first sampling 100 design points uniformly in the unit square. The actual surface is described by the function

$$f(\mathbf{x}) = 40\exp\{8[(x_1-0.5)^2+(x_2-0.5)^2]\} \cdot \big(\exp\{8[(x_1-0.2)^2+(x_2-0.7)^2]\} + \exp\{8[(x_1-0.7)^2+(x_2-0.2)^2]\}\big)^{-1},$$

to which we add standard Gaussian errors. This function first appeared in Gu et al. (1989), and it will be hereafter referred to as simply GBCW. The signal-to-noise ratio in this setup is about 3. In the lower lefthand panel in Figure 5, we present the result of applying the greedy, Triogram algorithm. As is typical, the

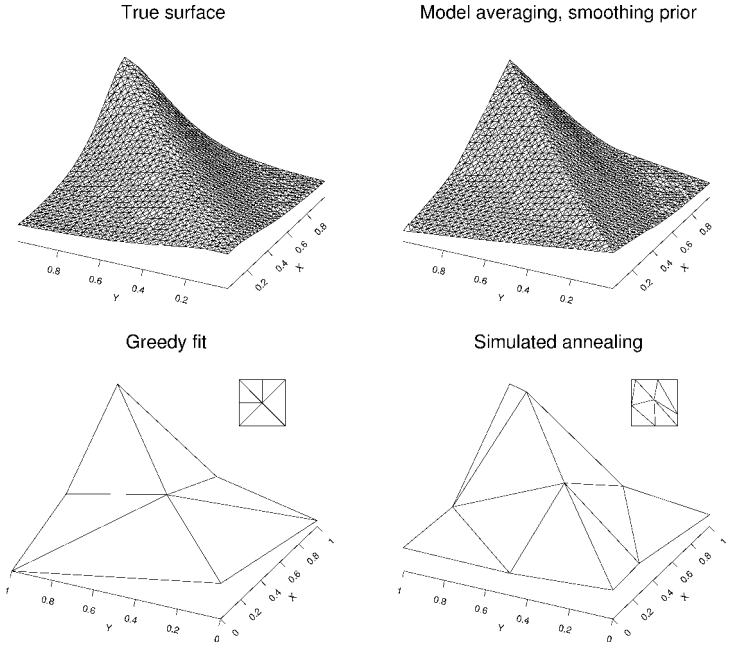

FIG. 5. *In the top row we have the true surface* (*left*) *and the fit resulting from model averaging* (*right*). *In the bottom row we have two isolated fits*, *each a "minimal" BIC model*, *the leftmost coming from a greedy search*, *and the rightmost produced by simulated annealing* (*the triangulations appear at the top of each panel*).

procedure has found a fairly regular, low-dimensional mesh describing the surface (the MISE is 0.31). For the fit plotted in the lower righthand panel, we employed a simulated annealing scheme similar to that described for Logspline. The geometric prior for $\triangle$ is used to guide the sampler through triangulations, and in each corresponding spline space G we consider $\hat{g}$, the MLE (or in this case the ordinary least squares fit). In this way, the objective function matches that of the greedy search, the generalized AIC criterion (8). The scheme alternates between (randomly selected) structural changes (edge swaps and vertex moves, additions and deletions) and updating the estimate $\widehat{\sigma}^2$ of the noise variance. After 6,000 iterations, the sampler has managed to find a less regular, and marginally poorer-fitting model (the MISE is 0.32). In the context of triangulations, the greedy search is subject to a certain regularity that prevents configurations like the one in Figure 5. We can recapture this in the MCMC simulations either by placing restrictions on the triangulations in each mesh (say, imposing a smallest allowable size or aspect ratio) or by increasing the penalty on dimension, specified through our geometric prior.

In the last panel, we present the result of model averaging using a uniform prior on model size and a smoothing prior on the coefficients ($\lambda = 1/n$). The sampler is run for a total of 6,000 iterations, of which 1,000 are discarded as burn-in. We then estimate the mean as a pointwise average of the sampled surfaces. The final fit is smoother in part because we are combining many piecewise-planar surfaces. We still see sharp effects, however, where features like the central ridge are present. The model in the lower righthand panel is not unlike the surfaces visited by this chain. As spaces G are generated, the central spine (along the line $y = x$) of this surface is always present. The same is true for the hinged portions of the surface along the lines $x = 0$ and $y = 0$. With these caveats in mind, the MISE of the averaged surface is about half of the other two estimates (0.15). We repeated these simulations for several sample sizes, taking $n =$ 100, 500 and 1000 (100 repetitions for each value of n). In Table 6, we present several variations in the prior specification and search procedure. In addition to GBCW, we also borrow a test function from Breiman (1991), which we will refer to as Exp. Here, points $\mathbf{X} = (X_1, X_2)$ are selected uniformly from the square $[-1, 1]^2$. The response is given by $\exp(x_1 \sin(\pi x_2))$ to which normal noise is added ($\sigma = 0.5$). The signal-to-noise ratio in this setup is much lower, 0.9. The results are presented in Table 7. It seems reasonably clear that the simulated annealing approach can go very wrong, especially when the sample size is small. Again, this argues for the use of greater constraints in terms of allowable spaces when n is moderate. It seems that model averaging with the smoothing prior ($\lambda = 1/n$) and the Poisson/ML prior of Denison et al. (1998a) perform the best. A closer examination of the fitted surfaces reveals the same kinds of secondary structure as we saw in Figure 5. To be sure, smoother basis functions would eliminate this behavior. It is not clear at present, however, if a different smoothing prior on the coefficients might serve to "unkink" these fits.

TABLE 6
Versions of Triogram used in the simulation study

	Model size	Parameters
(i)	Greedy optimization of AIC	
(ii)	Simulated annealing optimization of AIC	
(iii)	Poisson (5)	ML
(iv)	Geometric	ML
(v)	Uniform	$\lambda = 1/n$

TABLE 7
Mean integrated squared error (MISE) for two smooth test functions

		Version				
		(i)	(ii)	(iii)	(iv)	(v)
Distribution	n	MISE	Ratio of MISE over (i)			
GBCW (high snr)	100	0.31	1.35	0.85	0.78	0.77
GBCW (high snr)	500	0.10	1.0	0.64	0.76	0.80
GBCW (high snr)	1,000	0.08	0.91	0.82	0.94	0.79
Exp (low snr)	100	0.15	0.90	0.52	0.51	0.49
Exp (low snr)	500	0.04	0.85	0.46	0.50	0.47
Exp (low snr)	1,000	0.03	0.51	0.32	0.40	0.46

TABLE 8
Mean integrated squared error (MISE) for two piecewise-planar test functions

		Version				
		(i)	(ii)	(iii)	(iv)	(v)
Distribution	***n***	**MISE**	**Ratio of MISE over (i)**			
Model 1	50	0.16	0.97	0.70	0.35	0.80
Model 1	200	0.04	0.82	0.95	0.52	0.62
Model 1	1,000	0.01	0.63	0.72	0.76	0.40
Model 3	50	0.70	1.40	0.86	0.51	0.50
Model 3	200	0.17	0.85	0.63	0.27	0.30
Model 3	1,000	0.03	0.34	0.45	0.21	0.20

The performance of the Poisson (5) distribution is somewhat surprising. While for Logspline this choice led to undersmoothed densities, it would appear that the Triogram scheme benefits from slightly larger models. We believe that this is because of the bias involved in estimating a smooth function by a piecewise-linear surface. In general, these experiments indicate that tuning the Bayesian schemes in the context of a Triogram model is much more difficult than univariate set-ups. One comforting conclusion, however, is that essentially each of the schemes considered outperform the simple greedy search.

As a final test, we repeated the simulations from Hansen et al. (1998). We took as our trial functions two piecewise-planar surfaces, one that the greedy scheme can jump to in a single move (Model 1), and one that requires several moves (Model 3). In this case, the model averaged fits (iv) were better than both simulated annealing and the greedy procedure. The estimate built from the Poisson prior tends to spend too much time in larger models, leading to its slightly poorer MISE results, while the geometric prior extracts a heavy price for stepping off of the "true" model. (Unlike the smooth cases examined above, the extra degrees of freedom do not help the Poisson scheme.) The simulations are summarized in Table 8. One message from this suite of simulations, therefore, is that a posterior mean does not oversmooth edges, and in fact identifies them better than the greedy alternatives.

4. DISCUSSION

Early applications of splines were focused mainly on curve estimation. In recent years, these tools have proved effective for multivariate problems as well. By extending the concepts of "main effects" and "interactions" familiar in traditional d-way analysis of variance (ANOVA), techniques have been developed that produce so-called functional ANOVAs. Here, spline basis elements and their tensor products are used to construct the main effects and interactions, respectively. In these problems, one must determine which knot sequence to employ for each covariate, as well as what interactions are present.

In this paper we have discussed a general framework for adaptation in the context of an extended linear model. Traditionally, model-selection for these problems is accomplished through greedy, stepwise algorithms. While these approaches appear to perform reasonably well in practice, they visit a relatively small number of candidate configurations. By casting knot selection into a Bayesian framework, we have discussed an MCMC algorithms that sample many more promising models. We have examined various techniques for calibrating the prior specifications in this setup to more easily compare the greedy searches and the MCMC schemes. An effective penalty on model size can be imposed either explicitly (through a prior distribution on dimension), or through the smoothness prior assigned to the coefficient vector. In general, we have demonstrated a gain in final mean squared error when appealing to the more elaborate sampling schemes.

We have also gone to great lengths to map out connections between this Bayesian method and other approaches to the knot placement problem. For example, a geometric prior distribution on model size, has a natural link to (stepwise) model selection with BIC, while we can choose a multivariate normal prior on the coefficients to connect us with the penalized likelihood methods employed in classical smoothing splines. In addition, the Bayesian formalism allows us to account for the uncertainty in both the structural aspects of our estimates (knot configurations and triangulations) as

well as the coefficients in any given expansion. Model averaging in this context seems to provide improvement over simply selecting a single "optimal" model in terms of say BIC. The disadvantage of this approach is that we do not end up with a model based on one set of knots (or one triangulation).

While running our experiments, we quickly reached the conclusion that the priors play an important role: an inappropriate prior can easily lead to results that are much worse than the greedy algorithms. However, in our experiments we found out that, when the priors are in the right ballpark, Bayesian procedures do perform somewhat better than greedy schemes in a mean squared error sense. This improvement in performance is larger for a relatively "unstable" procedures such as Triogram, while the improvement for a "stable" procedure such as Logspline is smaller.

For the Triogram methodology there is an additional effect of model averaging: the average of many piecewise-planar surfaces will give the impression of being smoother. Whether this is an advantage or not probably depends on the individual user and her/his application: when we gave seminars about the original Triogram paper, there were people who saw the piecewise-planar approach as a major strength, while others saw it as a major weakness of the methodology.

ACKNOWLEDGMENTS

Charles Kooperberg was supported in part by NIH Grant R29 CA 74841. The authors wish to thank Merlise Clyde, David Denison, Ed George, Peter Green, Robert Kohn, Charles Stone and Bin Yu for many helpful discussions.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723.

BESAG, J. and HIGDON, D. (1999). Bayesian inference for agricultural field experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **61** 691–746.

BREIMAN, L. (1991). The Π-method for estimating multivariate functions from noisy data. *Technometrics* **33** 125–143.

BREIMAN, L. (1993). Hinging hyperplanes for regression, classification and function approximation. *IEEE Trans. Inform. Theory* **39** 999–1013.

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

COURANT, R. (1943). Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.* **49** 1–23.

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

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998a). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc. Ser. B* **60** 333–350.

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998b). A Bayesian CART algorithm. *Biometrika* **85** 363–377.

DYN, N., LEVIN, D. and RIPPA, S. (1990a). Data dependent triangulations for piecewise linear interpolation. *IMA J. Numer. Anal.* **10** 137–154.

DYN, N., LEVIN, D. and RIPPA, S. (1990b). Algorithms for the construction of data dependent triangulations. In *Algorithms for Approximation* **2** (J. C. Mason and M. G. Cox, eds.) 185–192. Chapman and Hall, 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.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

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

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

HALPERN, E. F. (1973). Bayesian spline regression when the number of knots is unknown. *J. Roy. Statist. Soc. Ser. B* **35** 347–360.

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

HANSEN, M., KOOPERBERG, C. and SARDY, S. (1998). Triogram models. *J. Amer. Statist. Assoc.* **93** 101–119.

HOLMES, C. C. and MALLICK, B. K. (2001). Bayesian regression with multivariate linear splines. *J. Roy. Statist. Soc. Ser. B* **63** 3–18.

HUANG, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.* **26** 242–272.

HUANG, J. Z. (2001). Concave extended linear modeling: A theoretical synthesis. *Statist. Sinica* **11** 173–197.

JUPP, D. L. B. (1978). Approximation to data by splines with free knots. *SIAM J. Numer. Anal.* **15** 328–343.

KOENKER, R. and MIZERA, I. (2001). Penalized Triograms: Total variation regularization for bivariate smoothing. Technical report. (Available at www.econ.uiuc.edu/roger/research/goniolatry/gon.html.)

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. and STONE, C. J. (2002). Comparison of parametric, bootstrap, and Bayesian approaches to obtaining confidence intervals for logspline density estimation. Unpublished manuscript.

KOOPERBERG, C. and STONE, C. J. (2002). Confidence intervals for logspline density estimation. Available at http://bear.fhcrc.org/∼clk/ref.html.

LINDSTROM, M. (1999). Penalized estimation of free-knot splines. *J. Comput. Graph. Statist.* **8** 333–352.

NICHOLLS, G. (1998). Bayesian image analysis with Markov chain Monte Carlo and colored continuum triangulation models. *J. Roy. Statist. Soc. Ser. B* **60** 643–659.

QUAK, E. and SCHUMAKER, L. L. (1991). Least squares fitting by linear splines on data dependent triangulations. In *Curves and Surfaces* (P. J. Laurent, A. Le Méhauté and L. L. Schumaker, eds.) 387–390. Academic Press, New York.

SCHUMAKER, L. L. (1993). *Spline Functions*: *Basic Theory*. Wiley, New York.

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

SIBSON, R. (1978). Locally equiangular triangulations. *Computer Journal* **21** 243–245.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

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

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–344.

SMITH, M. and KOHN, R. (1998). Nonparametric estimation of irregular functions with independent or autocorrelated errors. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Müller and D. Sinha, eds.) 133–150. Springer, New York.

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

SMITH, P. L. (1982b). Hypothesis testing in B-spline regression. *Comm. Statist. Part B—Simulation and Comput.* **11** 143–157.

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.

STONE, C. J., HANSEN M., KOOPERBERG, C. and TRUONG, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25** 1371–1470.

STONE, C. J. and HUANG, J. Z. (2002). Free knot splines in concave extended linear modeling. *J. Statist. Plann. Inference*. To appear.

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

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

Comment

Hugh A. Chipman, Edward I. George and Robert E. McCulloch

This paper uses ideas for stochastic search implementations of adaptive Bayesian models, such as those outlined in Denison, Mallick and Smith (1998a, b) and Chipman, George and McCulloch (1998a) and effectively applies these ideas to logspline density estimation and triogram regression. Interesting comparisons are made to assess the effect of greedy search, stochastic search and model averaging. Such comparisons are valuable, since readily available computing power enables the construction of many methods, and an understanding of what works is important in developing new methodology.

It is very important to note the role of the prior when adaptive models are used in conjunction with stochastic searches. Inevitably, priors guide and temper our wandering in a large space of models. This benefit comes with a price: the need to select a prior that is appropriate for the problem at hand. It is important to acknowledge the simple fact that a prior choice represents a bet on what kind of models we want to consider.

If we skip to the end of the paper and read the discussion, what lessons have been learned? We have that (i) "... we have demonstrated a gain... when appealing to the more elaborate sampling schemes" (relative to simple greedy search) and that (ii) "priors play an important role." These things we know to be

Hugh A. Chipman is Associate Professor, Department of Statistics and Acturial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1 Canada (e-mail: hachipman@uwaterloo.ca). Edward I. George is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302 (e-mail: edgeorge@wharton.upenn.edu). Robert E. McCulloch is Professor, Graduate School of Business, University of Chicago, Chicago, IL 60637 (e-mail: robert.mcculloch@gsb.uchicago.edu).

true in general from much experience. The question is: what should be done in practice?

In general, a practical approach usually involves first getting the prior specification down to a few hyperparameters (about which we hopefully have some understanding) and then developing a scheme for making reasonable choices. At one end of the spectrum we can use automatic methods such as cross-validation to choose hyperparameters that are appropriate for the problem at hand. At the other end of the spectrum we choose "reasonable values" based on our understanding and prior beliefs. Often, compromise strategies that combine a peek at the data with some judgment are effective and somewhat in the spirit of empirical Bayes. We believe Chipman, George and McCulloch (2002) is a good example of this middle ground approach.

We have some general Bayesian insights that help us understand the effects of these hyperparameters. Often we can think of prior in two stages: $p(M_k)$ a prior on "models;" and $p(\theta_k|M_k)$ a prior on the parameters of a given model. A set of hyperparameters would specify a choice for each of these components. In Section 2 of the paper, θ corresponds to the coefficients β, and M_k would be (K, t). Both choices can be important. Often we choose $p(M_k)$ to express the belief that the model is not too large. More subtle is the effect of a choice $p(\theta|M_k)$. If we make the prior too tight, we will miss parameter values that give good fit to the data, diminishing the posterior probability on model M_k. If we make the prior too spread out, the likelihood will be washed out and again we diminish the posterior probability. These are the basic facts of odds ratio calculations.

In Section 2 of the paper, the choices of A and λ are the hyperparameters that determine the spread of the prior given the model. We know from the general insight outlined above that these choices will be influential. The paper discusses these choices in terms of penalties and the AIC. We find the basic Bayesian intuition about odds ratio calculations is also helpful in understanding what is going on. It may be helpful to recall that the AIC is just a (very poor) approximation to the odds ratio calculation.

Table 2 compares the performance of algorithms for various values of λ. We see that the choice of λ matters. What choice is best? It depends. Based on Table 2, the authors state that choice (vii) is bad, yet it is best in several scenarios! The question remains: how do you choose λ?

While the authors consider the impact of different prior choices (e.g., for λ), methods for selection of the prior are not considered. Without such choices, the use of MCMC technology for stochastic search non-Bayesian is more limited.

One of the most important advantages of Bayesian methods in adaptive modeling problems is the effectiveness of stochastic search methods such as MCMC. In applications where the model space is complicated, constructing an effective chain can be challenging. For example, in the triogram regression problem, models are arranged somewhat hierarchically, with regions recursively subdivided into smaller and smaller triangles. Hierarchical structure makes the construction of an effective chain challenging because it constrains the possible set of proposals that can be made. Proposals making small local changes are easiest to make and most likely to be accepted, but a long succession of simple proposals may need to be accepted for the stochastic search to move on to a different posterior mode. With this dilemma in mind, we appreciate the importance of using good proposal steps in effective exploration of the model space. These transitions need to work within the model constraints (e.g., hierarchy in triograms) while not being so constrained as to have difficulty moving. Hansen and Kooperberg have effectively accomplished this by developing a set of proposal steps which move around the space in a natural way while respecting the nested nature of the models. In some problems, such as logspline density estimation, it may be easier to move around the space. In that case, the knots do not depend on the order in which they are added.

The authors use a single long chain to explore the model space, which can be an issue if the posterior on models has many sharp local peaks. In such situations, MCMC methods can tend to gravitate toward a single mode and have difficulty in moving to other regions of the model space. We expect such issues to arise in the triogram regression problem, for example. Denison, Mallick and Smith (1998b) use single chains as well and, by carefully controlling the early stages of the chain, achieve an algorithm which seems to explore a region of the model space around a single local maximum. We have found that another effective technique is to use multiple chains as a means of more fully exploring the space. Single and multiple chains were explored on a simulated dataset in Chipman, George and McCulloch (1998a) and the use of multiple chains resulted in a more complete exploration of the model space.

The authors examine the performance of Bayesian model averaging, which is an appealing and natural

means of improving predictive accuracy. We are not surprised that greedy methods can be improved upon by a better search and model averaging. What does surprise us is the omission of a trivial (and often effective) frequentist competitor: bootstrapping. The bootstrap has been used as a method of generating multiple models for model averaging (Breiman, 1996) and as an easy way to improve upon greedy search algorithms (Tibshirani and Knight, 1999). In this approach, multiple pseudodatasets are generated by resampling with replacement the rows of the data matrix, and a (often greedy) modeling algorithm is applied to each bootstrap dataset. Bootstrapping the data and averaging over models is an effective and easy way to model average. It enhances the search by perturbing the data and letting the greedy algorithm converge to different local maxima. Predictions are improved by averaging across all the different models. We have carried out some experiments with bootstrapping in the context of Bayesian CART (Chipman, George and McCulloch, 1998b). In the example we considered, we found that bootstrapping identified a wider variety of good models than a single greedy search, but the models identified by a bootstrap algorithm were still a subset of those identified by Bayesian stochastic search procedures.

Comment

C. C. Holmes

Mark Hansen and Charles Kooperberg have done an excellent job in tying (or should I say knotting) together the various methodological and philosophical approaches to random regression splines. My personal experience in this area derives from a fully probabilistic (Bayesian) standpoint and my comments will reflect this view. While I am aware that the authors may be familiar with much of what I am about to say I hope that the reader will benefit from the insights and subjective observations that follow.

1. INTRODUCTION

It is well known empirically that model averaging over flexible regression models tends to produce more accurate predictions. This is especially true when the individual model is "nonsmooth" such as for the triogram model described in Section 3. Flexible models typically show high variance to the data which the averaging tends to counter. The term "high variance" refers to the fact that small changes or perturbations to the data can lead to large changes in the form or output of the selected model. In this article Hansen and Kooperberg, hereafter H&K, consider some averaging methods using penalized likelihood. In the field of machine learning, algorithms such as Boosting, Bagging and Stacked Regression are popular methods to implement this strategy; see Hastie, Tibshirani and Friedman (2001), Chapters 8 and 10, for details and references.

In contrast to these mainly empirically motivated approaches, the probabilistic (Bayesian) modeller is forced into model averaging by the requirement to remain coherent. Bayesian inference is an axiomatic system; if you buy into the axioms you must follow the rules and the rules state that when fully quantifying measures of uncertainty in a response variable you should report the marginal distribution $p(y|x)$; this is subsequently combined with a loss function when reporting point estimates. The Bayesian approach requires a priori that we quantify, via probability distributions, measures of uncertainty in all aspects of the model.

Bayesian inference provides us with a rich modelling paradigm with probability as its central pillar. However, this richness comes with a price. In defining probability distributions over complex model spaces we must be careful about the implications on the joint marginal prior distribution $p(Y)$. I shall expand on this last point in Section 3 with reference to the precision parameter λ used in H&K. The Bayesian and non-Bayesian methods presented in H&K appear to be doing something similar, namely model averaging. However, the procedures behind them are very different. In the next section I will highlight what I believe to be the key difference.

C. C. Holmes is Lecturer, Department of Mathematics, Imperial College, London, SW7 2BZ, United Kingdom (e-mail: c.holmes@ic.ac.uk).

2. BAYESIAN INFERENCE AND PENALIZED LIKELIHOOD

In the averaging procedure of the non-Bayesian methods the weight given to each model is

$$w(\mathbf{t}) \propto \max_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}, \mathbf{t}) + a J(\mathbf{t}),$$

where a model is characterized by a knot configuration $\mathbf{t}$ that records the number and position of the splines, $\max_{\hat{\boldsymbol{\beta}}_\mathbf{t}} l(\hat{\boldsymbol{\beta}}_\mathbf{t}, \mathbf{t})$ is the *profile* log likelihood of model $\mathbf{t}$ and $J(\mathbf{t})$ denotes the dimension of $\mathbf{t}$. The corresponding weight given to $\mathbf{t}$ in the Bayesian procedure is

$$w(\mathbf{t}) \propto l(\mathbf{t}) + p(\mathbf{t}),$$

where $l(\mathbf{t})$ is the *marginal* likelihood and $p(\mathbf{t})$ the prior. Matching the user-defined quantities $p(\mathbf{t}) = a J(\mathbf{t})$ we see that the central difference between the procedures lies in the use of the marginal or profile likelihood. The marginal likelihood is special to Bayesian inference and is defined as

$$\begin{aligned} l(\mathbf{t}) &= p(Y = y|\mathbf{t}) \\ &= \int p(Y = y|\boldsymbol{\beta}, \mathbf{t}) p(\boldsymbol{\beta}|\mathbf{t})\, d\boldsymbol{\beta}, \end{aligned}$$

where Y denotes the observed data and the measure $p(Y = y|\mathbf{t})$ is also known as the *prior predictive* as it records the probability of observing the data before the data arrived. In contrast the profile likelihood conditions on the observed data through the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_\mathbf{t}$.

The prior predictive contains a natural penalty against overly complex models which is a direct consequence of using a fully probabilistic approach. Complex models spread their probability measure over a wide space of possible data generators (possible realizations $Y = y$). By definition this distribution is normalized (integrates to 1) and hence the actual quantity $p(Y = y|\mathbf{t})$ will be diluted as the space spanned by the model increases. Accordingly, simple models that are consistent with the observed data have greater marginal likelihood. The exact opposite is true of the profile likelihood.

It is worth stating now that I do not consider the "Bayesian ML" models considered by H&K as Bayesian. They do not relate to proper probabilistic models and do not use the marginal likelihood in their inference. Moreover, in Bayesian inference we are typically interested in reporting the full distribution $p(y|\mathbf{x})$ not just reporting the mean. By not taking into account uncertainty in $\boldsymbol{\beta}$ these will be artificially tight if ML is used.

The form of the marginal likelihood $l(\mathbf{t})$ highlights another important issue, namely the significant role played by the prior distribution $p(\boldsymbol{\beta}|\mathbf{t})$ within the posterior inference. This is noted in the results in H&K in their comparison of models (v)–(vii) given in Table 1. Throughout, the prior distribution $p(\boldsymbol{\beta}|\mathbf{t})$ is taken to be multivariate normal, $p(\boldsymbol{\beta}|\mathbf{t}) = N(0, (\lambda A)^{-1})$, for fixed positive-definite matrix A. The precision parameter λ is highly influential to the inference. In Section 2.2 the authors suggest that "choosing λ too large ... leads to oversmoothing, while choosing λ too small ... tends to produce overly wiggly fits." This statement is not strictly true and the reason for this is interesting.

Consider the smoothing spline introduced in Section 2.1 equation (10). The smoothing spline has a basis function representation with a knot point at every data value. The parameter λ controls the smoothness of the fit such that as $\lambda \to 0$ the model has n degrees of freedom and interpolates the data. In this case the estimates of $\boldsymbol{\beta}$ match the maximum likelihood estimates. For $\lambda = 0$ the prior $p(\boldsymbol{\beta}|\mathbf{t})$ is improper and is known as the reference prior which would be considered as noninformative. However, suppose we wish to entertain the prospect that some knots are not needed, as in the methods of this article. For the variable dimension case as $\lambda \to 0$ we find we select no knots at all regardless of the data and we are left with a smooth global polynomial! That is, all of the posterior mass lies on the simplest model. This effect, that the noninformative reference prior for the fixed dimension case is a maximumly informative prior in the the variable dimension case, is a consequence of the Lindley–Bartlett paradox.

3. LINDLEY–BARTLETT PARADOX AND COVARIANCE REPRESENTATION

Lindley (1957) and Bartlett (1957) discuss an apparent paradox in the hypothesis testing of the location of the mean of a normal distribution. The consequence of their result for the Bayes linear model, $y \sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2)$ with prior $p(\boldsymbol{\beta}) \sim N(0, \lambda^{-1}A)$ and random covariate set $\mathbf{x}$, is that for $\lambda \to 0$ the posterior distribution $p(\mathbf{x}|y)$ places ever greater mass on the simplest model. In this article, the basis set $\mathbf{x}$ would refer to the output from the set of splines with knot locations $\mathbf{t}$.

This result may seem surprising and even a little worrying to the non-Bayesian observer. However, it is

merely a consequence of the prior induced on $p(Y)$ by the model and is not really a paradox at all when considered in this way. To be specific, given a normally distributed prior on the coefficients $\boldsymbol{\beta} \sim N(0, (\lambda A)^{-1})$ we find that the marginal prior induced on Y is normal

$$p(Y = y|\mathbf{x}) = \int p(Y = y|\mathbf{x}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{x}) = N\big(0, X(\lambda A)^{-1} X'\big),$$

where X denotes the $n \times k$ design matrix, which in our situation records the k responses from the set of polynomial and spline bases for the n data points. We see that the normal prior on $\boldsymbol{\beta}$ induces a Gaussian process prior on Y with variance–covariance matrix $X(\lambda A)^{-1} X'$. Hence, setting $\lambda \to 0$ specifies huge variance along the column space spanned by X. Adding a column to X inflates the variance and hence typically reduces the probability of $Y = y$.

The covariance representation of the Bayesian regression spline is instructive and peculiar to the Bayesian model. The Bayesian methods discussed in H&K can be considered as modelling Y or $\log(Y)$ as a Gaussian process with covariance matrix defined by λ, A and the knot locations $\mathbf{t}$. The MCMC algorithm is then seen to perform a random walk on the state space of fixed dimensional $n \times n$ covariance matrices determined by the spline locations. Of course, the covariance form is rarely used in practice as it requires the inversion of an $n \times n$ matrix while the usual basis representation requires inversion of $k \times k$ matrices. Nevertheless, they are equivalent.

In this manner we can view the Bayesian method of free-knot splines as a flexible approach to automatically constructing appropriate covariance structures for a Gaussian process model for Y, or $\log(Y)$. The use of free knots readily allows the covariance matrix to be nonstationary in that different amounts of smoothing are achieved in different regions of $\mathbf{x}$. The covariance representation also highlights that fact that the setting of λ is critical to the inference. In the recent work of Holmes and Denison (2002) we advocate reducing the sensitivity to this parameter by adopting a further prior distribution on λ; see also the forthcoming monograph by Denison, Holmes, Mallick and Smith (2002) on Bayesian methods for nonlinear classification and regression.

To conclude, I greatly enjoyed reading the paper. You get a real sense that the authors have a great feel for the various approaches they discuss. To the committed Bayesian, phrases such as "an inappropriate prior" are inappropriate but these are minor quibbles.

Comment

Robert E. Kass and Garrick Wallstrom

Over the past several years Hansen and Kooperberg have contributed substantially to the development of spline-based nonparametric function estimation. As they show in this paper and in their 1997 paper with Stone and Truong, Logspline and its variants can be effective in diverse settings. Indeed, the unification of those settings with the rubric of extended linear models is itself an important contribution. Here, they emphasize the contrast between the deterministic, frequentist optimization methods and stochastic, Bayesian alternatives. We and our colleagues have been using Bayesian spline fitting in a variety of applications and we are pleased to be able to contribute to the discussion.

Hansen and Kooperberg are mainly interested in Bayesian methods for algorithmic, as opposed to inferential, reasons: reversible-jump MCMC can visit a large number of models, and the resulting posterior mode, and posterior mean based on model averaging, turn out to be very good estimates. In fact, as we discuss in Section 1, MCMC can be even better than Hansen and Kooperberg's simulations indicate. Hansen and Kooperberg provide interpretations for the methods they investigate in their simulation study, but this is a place where we disagree with them on technical grounds. In Section 2 we suggest somewhat different interpretations, which make it easier to understand the authors's numerical results. In Section 3 we briefly em-

Robert E. Kass is Professor and Department Head and Garrick Wallstrom is Visiting Assistant Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213 (e-mail: kass@stat.cmu.edu).

phasize the importance of maximization schemes such as Logspline; in Section 4 we discuss some related work; in Section 5 we summarize our view of the current state of the art.

1. WHEN IS A METHOD BAYESIAN?

Bayesian methods are optimal when the data are generated according to the assumed model (giving the likelihood function) and the prior correctly represents a priori knowledge. From a practical standpoint, when the model does a reasonably good job of describing the data, and the prior contributes relatively little erroneous information, Bayesian methods work well. With moderately large data sets we expect Bayesian methods to reflect the model, so that Bayes estimates are essentially maximum likelihood estimates.

Many researchers, including Hansen and Kooperberg, find it informative to try to recast frequentist procedures in Bayesian terms, so as to get a clearer understanding of the way those procedures work. Typically, this exercise shows a frequentist method to be approximately Bayesian in some sense, and one then hopes to have a relatively easy time seeing what the model and prior each contribute. For this to be justified, however, there must be a formal sense in which the method is "approximately Bayesian." That is, there must be a formal sense in which the procedure computes a probability (or an estimate) that is approximately equal to a probability (or estimate) coming from some posterior distribution. For example, maximum likelihood estimates may be shown to approximate posterior means as sample sizes become infinite.

Hansen and Kooperberg have made central to their study the algorithm of Denison, Mallick and Smith (1998a). We would like to take issue with their interpretation of it. We will call it DMS, and we will contrast it with a modification of DMS that DiMatteo, Genovese and Kass (2001) called BARS.

1.1 DMS Is neither Bayesian nor Approximately Bayesian in Large Samples

Green (1995) pointed out that reversible-jump MCMC was well-suited for finding breakpoints in piecewise constant functions. The paper by Denison, Mallick and Smith (1998a) was important because it extended this observation to cubic-spline curve-fitting.

From a Bayesian point of view, the likelihood function on the knot set $\mathbf{t}$ based on data y, model $p(y|\beta, \mathbf{t})$ and prior $\pi(\beta)$ is the marginal density

$$L(\mathbf{t}) = p(y|\mathbf{t}) = \int p(y|\beta, \mathbf{t})\pi(\beta)\, d\beta.$$

[Here, for simplicity, we are ignoring σ, which is handled by integration with respect to the prior $\pi(\sigma) = 1/\sigma$.] The DMS method replaces this marginal density $p(y|\mathbf{t})$ with the conditional density $p(y|\hat{\beta}, \mathbf{t})$, where, in the context of curve-fitting, $\hat{\beta}$ is the least-squares estimate. The advantage of doing so is that no prior distribution on β need be considered. Because there is no prior on β, DMS is not an exact Bayesian method. It might be thought that it is approximately Bayesian in the same sense that maximum likelihood is approximately Bayesian, but that is not the case either.

In many situations it is innocuous to ignore the prior or to assume a flat prior. For instance, if the knots $\mathbf{t}$ are fixed so that the problem is one of multiple linear regression, for moderately large samples the posterior of β based on a flat prior will become a good approximation to the posterior of β based on any relatively diffuse proper prior, and the least-squares estimate will become a good approximation to the posterior mean of β. In contrast, the conditional density $p(y|\hat{\beta}, \mathbf{t})$ will *not* approximate the marginal density $p(y|\mathbf{t})$ in large samples for *any* prior.

This point may seem pedantic but, as we spell out below, it has important implications both for practice and for interpretation.

1.2 For Gaussian Nonparametric Regression, BARS is Bayesian

For curve-fitting, a very simple alternative to DMS is to define an analytically tractable prior on β given $\mathbf{t}$. DiMatteo, Genovese and Kass (2001) chose the prior

$$\beta \sim N(0, n\sigma^2(X_{\mathbf{t}}^T X_{\mathbf{t}})^{-1}),$$

where $X_{\mathbf{t}}$ is the design matrix based on the spline basis. Because the Fisher information in the data is $X_{\mathbf{t}}^T X_{\mathbf{t}}/\sigma^2$, this prior may be understood as having the same amount of information as a single observation. Thus, Kass and Wasserman (1995) called it the unit-information prior. Many other authors have used this prior; see Pauler (1998). In the context of spline curve-fitting, Smith and Kohn (1996) used the same prior but with variance matrix $c\sigma^2(X_{\mathbf{t}}^T X_{\mathbf{t}})^{-1}$, where c was chosen to be between 10 and 1,000, values not terribly different than the sample size. Any

such choice yields an explicit form for the likelihood $L(\mathbf{t})$ required in the reversible-jump MCMC scheme. Thus, any such prior in conjunction with the MCMC scheme used in DMS yields an exactly Bayesian method for Gaussian nonparametric regression. The unit-information prior was chosen as a simple default but, aside from its connection with BIC described below, there is nothing special about the particular choice $c = n$.

An additional feature of the BARS implementation was the incorporation of a locality heuristic: continuous proposal distributions for addition of a knot are used; these attempt to place knots *close* to existing knots. This is based on the observation, stressed by Zhou and Shen (2001), that extra knots are needed where curves change sharply. Hansen and Kooperberg mention the general idea that more knots are needed in regions where the function changes rapidly, but their implementations apparently ignore this fact. The DMS implementation attempts to spread knots out, rather than clumping them as would be needed for rapidly varying curves. This produces a somewhat inefficient chain, that is, one that needs to be run a long time.

1.3 BARS Achieves Much Smaller MISE than Does DMS

When the conditional density $p(y|\hat{\beta}, \mathbf{t})$ replaces the correct marginal density $p(y|\mathbf{t})$ it is no longer possible to learn about the number of knots needed to model the data. The practical implication is that DMS severely overfits the data and suffers extreme sensitivity to the choice of prior. It is easy to see why: when using the conditional density, the likelihood increases as knots are added so that the posterior mode will be at the maximum allowable number of knots. Indeed, it is precisely to avoid this situation that dimensionality penalization methods such as AIC and BIC are introduced. The DMS likelihood on the number of knots (based on the conditional density) will level off as the number of knots increases and will become dominated by the prior.

As reported by Denison, Mallick and Smith (1998a), and also by Hansen and Kooperberg, fits obtained using DMS may sometimes appear reasonable, and it may also appear that the prior on the number of knots has relatively little effect. This happens for two reasons: first, overfitting will often produce a model with relatively small MISE and, second, short MCMC chains may traverse regions of reasonably good models. Long chains will allow the method to explore regions of higher probability, which in DMS are regions involving large numbers of knots leading to excessively rough (wiggly) models. One might then be tempted to advocate DMS with short chains—in fact, this is precisely what Hansen and Kooperberg used in their simulation study—but that approach is very worrisome: with MCMC we want to make sure we have adequate mixing, particularly in highly multimodal problems. A safer strategy is to use a Bayesian method.

In the three simulated examples reported by DiMatteo, Genovese and Kass (2001), which were similar in spirit to the three used by Hansen and Kooperberg except in the context of curve-fitting, BARS reduced the MISE compared to DMS by 70% or more (with chains of length 10,000). Gains for curves with very sharp jumps are most dramatic. Furthermore, when fits were examined, BARS was appropriately smooth while DMS was very rough, with very large numbers of local modes; and the number of DMS modes was highly dependent on the choice of prior.

1.4 BIC Is Approximately Bayesian with Respect to the Unit-Information Prior

Schwarz (1978) showed that the criterion now most commonly known as BIC provides consistent model selection, in the sense that BIC will choose the correct model (among alternative choices, one of which is assumed true) for sufficiently large samples. The argument was Bayesian: Schwarz used the log of the Bayes factor and then omitted constant-order terms (see Kass and Raftery, 1995, for additional references). Kass and Wasserman (1995) and Pauler (1998) showed that BIC approximates the log of the Bayes factor very well when the unit-information prior is used: analytically the error is of order $O_p(n^{-1/2})$ and numerically the results are remarkably close, even for moderate sample sizes. Thus, in very general terms, model selection using BIC is approximately the same thing as model selection via Bayes factors with the unit-information prior. We should caution that this asymptotic argument requires the sample size to be well defined (see Pauler, 1998) and grow to infinity while the number of parameters remains finite. In the spline applications we have faced, however, these seem to be reasonable assumptions and the BIC-based BARS procedure we describe below, in Section 1.5, performs well.

In the case of Gaussian curve-fitting it is particularly easy to see the BIC approximation in action. DiMatteo, Genovese and Kass (2001) noted that the likelihood ratio for adding a knot in the BARS reversible-jump MCMC is

$$\text{ratio} \approx \exp(-\text{BIC}/2), \tag{1}$$

where the approximation error is simply omission of the factor $\sqrt{n/(n+1)}$. DiMatteo, Genovese and Kass also pointed out that DMS multiplies the right-hand side by $\sqrt{n}$, which is another way to see why that method leads to severe overfitting.

Hansen and Kooperberg observe that BIC may be interpreted as using the conditional density $p(y|\hat{\beta}, \mathbf{t})$ together with a Geometric prior distribution on the number of knots, with parameter $p = 1 - n^{-1/2}$. We would prefer to state this differently, in line with Schwarz's analysis: for large samples, the marginal density $p(y|\mathbf{t})$ based on *any* prior on β will differ from the conditional density by a factor that behaves like this Geometric prior; that is, it falls at the exponential rate $n^{-k/2}$. [For consistency the factor multiplying $p(y|\hat{\beta}, \mathbf{t})$ must fall at the rate $e^{-kf(n)}$, where $f(n) \to \infty$ and $f(n)/n \to 0$; see Nishii, 1988.] Thus, DMS with a uniform prior could be interpreted as doing the same thing as would a Bayesian method that uses a proper prior on β together with an improper prior on the number of knots that increases roughly as $n^{k/2}$.

1.5 BARS May Be Generalized with Laplace's Method and BIC

We described BARS very briefly above in the case of Gaussian curve-fitting. There, with the unit-information prior or any other Gaussian prior on β, the integral required for the marginal density $p(y|\mathbf{t})$ may be evaluated analytically. For other ELMs the integral is no longer tractable. DiMatteo, Genovese and Kass (2001) approximated the integral by Laplace's method. With the unit-information prior this yields

$$\log p(y|\mathbf{t}) \approx \log p(y|\hat{\beta}, \mathbf{t}) - \frac{J}{2}\log n,$$

where J is the dimensionality of the basis. That is, for general models BARS uses BIC in its MCMC scheme. It would be equally easy to use Laplace's method for an alternative prior. It might also be more effective to use a different prior, but DiMatteo, Genovese and Kass (2001) were satisfied with the results obtained with BIC.

DiMatteo, Genovese and Kass (2001) approximated the integral defining $p(y|\mathbf{t})$ to improve simulation efficiency: marginal chains (here, a chain on $\mathbf{t}$) tend to be more efficient than joint chains (here, a chain on $(\beta, \mathbf{t})$; see Liu, Wong and Kong, 1994). In addition, chains on $(\beta, \mathbf{t})$ must be constructed with care to ensure detailed balance (see Genovese, 2000). From Hansen and Kooperberg's brief description of the procedure they used with their smoothing priors we could not understand whether they ran chains on $\mathbf{t}$ or $(\beta, \mathbf{t})$. (In addition, it appears that they used a Normal approximation to the conditional posterior distribution on β; note that this is less accurate than using Laplace's method, though it may be corrected by importance weighting as in DiMatteo, Genovese and Kass, 2001.)

2. REINTERPRETATION OF HANSEN AND KOOPERBERG'S RESULTS

We now return to the methods Hansen and Kooperberg examined in their simulation study. Keeping in mind that when Hansen and Kooperberg refer to the use of AIC in these methods they actually mean BIC (which, in contrast to the terminology in the literature, they choose to regard as a special case of AIC), we interpret the methods from our own Bayesian point of view as follows:

- method (i) is a quick-and-dirty way to find a model having high posterior probability (approximately, for moderate or large samples, using BIC), based on the unit-information prior on β and a flat prior on the number of knots;
- method (ii) is an attempt to find the posterior mode (again, approximately), based on the unit-information prior on β and a flat prior on the number of knots; (The authors call this simulated annealing but did not describe any annealing; if they actually did some annealing we would be very interested to find out the details: what did they learn about the appropriate cooling schedule?)
- method (iii) is an approximately Bayesian method, based on the unit-information prior on β and a flat prior on the number of knots. DiMatteo, Genovese and Kass (2001) called it modified-DMS in their simulation study and found that it worked reasonably well, though not as well as BARS because BARS also incorporated the locality heuristic, which used continuous proposal distributions and produced a more efficient MCMC scheme;
- method (iv) is the pseudo-Bayesian DMS method, which tends to overfit;

- method (v) is a Bayesian method, based on the smoothing prior on β with $\lambda = 1/n$ and a flat prior on the number of knots. The smoothing prior is very similar to the unit-information prior: both use the multiplier $1/n$ while the smoothing prior uses A in place of the information matrix. We would expect this method to behave fairly similarly to method (iii), except that the authors may have used different MCMC schemes;
- by increasing the multiplier to be of order greater than $1/n$, methods (vi) and (vii) make the prior on β much more informative than a single observation; in method (vii) it has the same order of information as does the full data set. In these cases, and more dramatically for (vii), the prior will substantially smooth the fit and the net result will be to destroy the local adaptation that is the chief benefit of knot location algorithms. We would expect to see these methods behave somewhat similarly to smoothing splines, and often poorly, sometimes overfitting and sometimes smoothing over sharp changes in the function;
- method (viii) is similar to method (v), using the same order of information on β as a single observation, except that it has the geometric prior on the number of knots. We would expect this method to work well on functions that require only small numbers of knots, and also when there is sufficiently much data to overcome the prior's very thin tails. We would expect this method to perform poorly with functions having many sharp peaks and moderate sample sizes.

To summarize, with our reinterpretation we would expect methods (ii), (iii), (v) and (viii) to perform well for the four examples used by Hansen and Kooperberg, and methods (iv), (v) and (vi) to be considerably worse. We would also expect the performance of (viii) to deteriorate on examples with moderate sample sizes and multiple peaks. Note that our interpretation of method (iii) should be contrasted with Hansen and Kooperberg's characterization of method (iii) as being ML with a geometric prior on the number of knots. We would expect (viii) to deteriorate on examples with moderate sample sizes and multiple peaks because of its geometric prior, but we would not expect such poor behavior from method (iii) because we think of it as having a flat prior rather than a geometric prior.

We also would be concerned that good performance of methods (ii), (iii), (v) and (viii) might sometimes require very long chains. It is worth repeating that short chains, such as those used by Hansen and Kooperberg in their simulation study, may fail to properly sample from the posterior distribution. This depends, of course, on the MCMC method and we mentioned that BARS can perform better due to its increased efficiency. Further improvements may well be possible. In addition, as we said above, for reasons of efficiency we prefer running the chain on $\mathbf{t}$ rather than on $(\beta, \mathbf{t})$.

As far as method (i) is concerned, we would expect it to do a reasonably good job, and its speed is very appealing. On the other hand, given the goal of finding the posterior mode (approximately, via BIC), alternative maximization methods may be even better.

We believe these remarks provide a straightforward understanding of Hansen and Kooperberg's numerical results, aside from a few perplexing anomalies: for the most part the methods perform just as we would expect. It is possible that the anomalies (the most egregious being method (iii) at 1.52 for normal and 1.09 for slightly bimodal with $n = 50$, method (v) at 1.35 for normal with $n = 10{,}000$ and method (viii) at 1.13 for sharp peak with $n = 200$) might go away with a more efficient MCMC method and/or substantially longer chains.

3. POSTERIOR MAXIMIZATION

In our applied work we always need not only fits for an unknown function but also assessments of uncertainty. We consider the availabilty of the posterior as an inference engine to be a big advantage of the Bayesian approach. However, in many of our applications we have sufficiently large sets of data (often hundreds of function-fitting problems, each having tens of thousands of binary observations) that computing time is a big issue. We desperately need fast methods. The speed of Logspline is appealing, but we cannot use it in practice unless we also have a way to assess uncertainty.

It seems to us that a reasonable compromise would be to apply a rapid fitting method, and then run a relatively short chain to get a rough idea of the posterior—enough of an idea to get at least some notion of uncertainty. The initial rapid fitting method should provide a good estimate of the function and, as a by-product, a sufficiently good starting value for the short chain that no burn-in is necessary. We could use Logspline for this purpose, as Hansen and Kooperberg

apparently did for method (iii), but our sense is that it may be possible to do better.

As a method of fitting, the big advantage of reversible-jump MCMC for finding good fits is its ability to "tunnel" across regions of low posterior density. Algorithms that try to maximize the posterior for a given number of knots can get stuck when knot movement requires traversing an interval where the posterior density is comparatively small; by instead moving to a model with an additional knot, and then subsequently deleting a knot, an algorithm can effectively move past the region of low posterior density. This makes us think that a maximization method should incorporate some of this "stepwise" alternation of models with different numbers of knots. Hansen and Kooperberg refer to Logspline as a stepwise method but, as we understand it, it actually uses a "forward selection" method to add knots and then a "backward elimination" method to delete them. While it apparently does a good job, we wonder whether alternatives that attempt to build some of the "tunneling" features of the MCMC method into a fast maximization scheme might be effective. We have tried simple variants on BARS that maximize rather than sampling, and they can work reasonably well, but we have not done systematic research and are unable to report any dramatic improvements in computation time. We would be interested in any further comments Hansen and Kooperberg may have on this point.

4. RELATED WORK

Spline-based approaches to function estimation have received much recent attention and there are variations that Hansen and Kooperberg did not touch on. Some relevant additional references may be found in Shively, Kohn and Wood (1999) and in DiMatteo, Genovese and Kass (2001).

Our own applied work has focused primarily on applications to neuroscience. In addition to using BARS for functional imaging and identification of neuronal firing patterns (as illustrated in DiMatteo, Genovese and Kass, 2001), this has included artifact removal from EEG's (Wallstrom et al., 2002). For some of our neuronal work we have also needed to fit two-dimensional surfaces and for this we have applied a mild modification of the method in Denison, Mallick and Smith (1998c), which (unlike their curve-fitting approach) is fully Bayesian. Here we would be very interested in any remarks Hansen and Kooperberg might be able to make regarding the performance of their triogram methods compared to those using the more familiar product spline bases as in Denison, Mallick and Smith (1998c). We have also implemented a generalization of BARS that fits multiple curves simultaneously, which is closely related to the method of Shi, Weiss and Taylor (1996). This provides a Bayesian approach to a basic problem in functional data analysis, that of describing the variation among many curves. We hope to report on it soon.

5. CONCLUSIONS

Here and in cited publications, Hansen and Kooperberg have taken advantage of important insights to produce useful methods and have enlightened us concerning their behavior.

Currently, BARS appears to be the most powerful available method for spline-based curve-fitting in ELMs. The essence of BARS is that it uses the following:

1. a proper prior on β that contains relatively little information;
2. a reversible-jump MCMC scheme on $\mathbf{t}$ with continuous proposals for knot addition that try to place new knots close to existing knots;
3. approximation of the marginal density $p(y|\mathbf{t})$ by Laplace's method when it is not available analytically.

Innovations that would improve behavior, or make the chain more efficient, are certainly possible. For example, one obvious idea is to replace the factor $1/n$ that multiplies the prior information in the unit-information prior (or the λ in Hansen and Kooperberg's smoothing priors) with a value $1/c$ (as in Smith and Kohn, 1996) and then put a prior on c (as in George and Foster, 2000). We have not tried this ourselves as yet.

In many applications, faster methods, such as Logspline, or alternative schemes designed to maximize the posterior, are highly desirable. However, evaluation of uncertainty is essential. Some hybrid approach involving maximization steps together with short simulation chains may provide a useful solution. We look forward to further work by Hansen, Kooperberg and others on this important problem.

ACKNOWLEDGMENTS

This work was supported in part through grants from the National Science Foundation and the National Institutes of Health.

Comment

Roger Koenker and Ivan Mizera

1. INTRODUCTION

Piecewise linear approximations on adaptively selected triangulations of planar domains provide an effective framework for many aspects of applied mathematics, from surface modeling in computer aided design to numerical methods for solving partial differential equations. Hansen and Kooperberg have convincingly demonstrated that these methods also deserve a prominent place in the statistical arsenal. On equivariance grounds alone a persuasive case can be made for their superiority to competing tensor product methods.

The most challenging aspect of the triogram approach lies in chosing a strategy for the adaptive triangulation. Hansen and Kooperberg have undertaken a wide ranging exploration of these strategies: initially in Hansen, Kooperberg and Sardy (1998) within the regression spline, model-selection paradigm and now from a more Bayesian viewpoint. The prior plays two important roles in the latter approach. It controls the number and position of the vertices of the triangulations, and it acts to shrink the parameters of the basis expansion toward a globally linear fit. Vertex selection and the attendant choice of the triangulation are the computationally difficult aspects of this process and seem to demand an MCMC implementation. Inherently, there is a trade-off between the flexibility allowed by the triangulation and the amount of shrinkage. Parsimonious triangulations, that is, those with few vertices, need little shrinkage; more profligate triangulations need more shrinkage. Hansen and Kooperberg opt for priors yielding rather parsimonious triangulations with only a handful of vertices. This is quite suitable for the test function of Section 3.2, but we are curious about how their MCMC methods would scale for problems that required many more vertices.

Roger Koenker is Professor, Departments of Economics and Statistics, University of Illinois, Champaign, Illinois 61820 (e-mail: rkoenker@uiuc.edu). Ivan Mizera is Associate Professor, Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1 (e-mail: mizera@stat.ualberta.ca).

2. A TOTAL VARIATION ROUGHNESS PENALTY

In recent work (Koenker and Mizera, 2001), we have been exploring total variation regularization methods for estimating triogram models. We take an extremely liberal attitude to the choice of the triangulation, allowing vertices at each of the distinct (x_i, y_i) points, and then adopting the resulting Delaunay triangulation. In the spirit of the smoothing spline literature we rely entirely on our "roughness penalty" to achieve the appropriate degree of smoothness. Our penalty may be interpreted as the total variation of the gradient of the fitted function defined as

$$J(g) = V(\nabla g) = \iint_\Omega \|\nabla^2 g\| \, dx \, dy,$$

where the integral may need to be interpreted in the sense of distributions as $\liminf$ of a sequence of smooth approximates. For general functions, g, the choice of the norm in the penalty may pose real problems; however, for triograms we are able to show that for any orthogonally invariant norm we obtain a scalar multiple of

$$J(g) = \sum_e \|\nabla g_e^+ - \nabla g_e^-\| \, \|e\|,$$

where e runs over all the interior edges of the triangulation, $\|e\|$ is the Euclidean length of the edge e and $\|\nabla g_e^+ - \nabla g_e^-\|$ is the Euclidean length of the difference between gradients of g on the two triangles adjacent to e.

This penalty is particularly convenient when paired with absolute error fidelity,

$$\sum_{i=1}^n |z_i - f(x_i, y_i)| + \lambda J(f),$$

where it can be reformulated as a data augmentation strategy and minimization can be accomplished by efficient linear programming methods. It is important to mention that the sparsity of the linear algebra in these problems enables us to solve large problems even

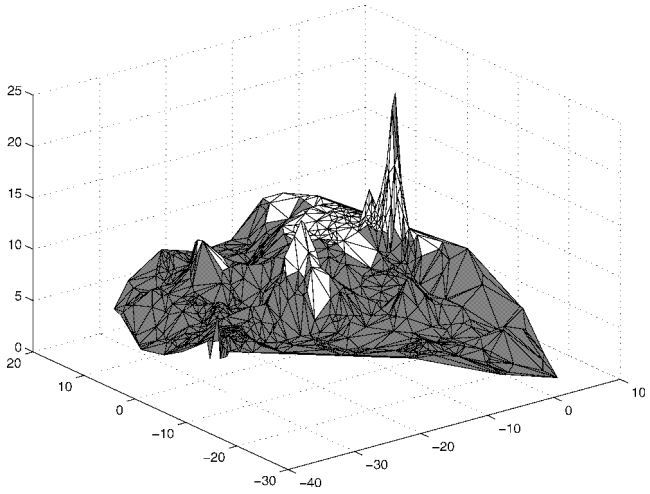

FIG. 1. *Median triogram fit of Chicago land values.*

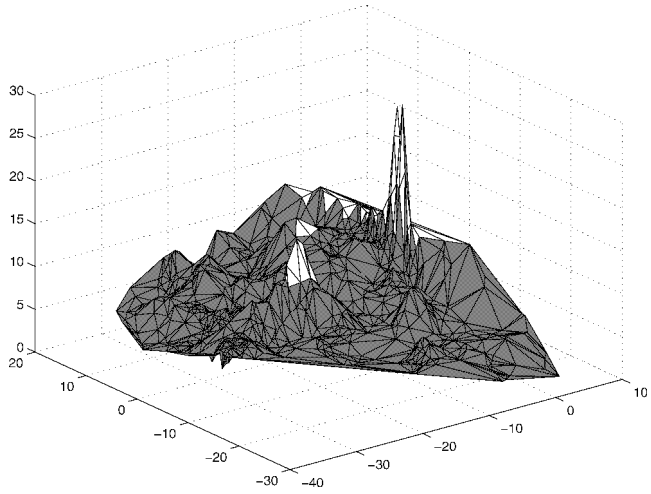

FIG. 2. *Mean fit of Chicago land values.*

though the nominal parametric dimension is several hundred.

Hansen and Kooperberg employ a modified version of our $J(\cdot)$ penalty that sums the *squares* of the contributions along the edges as a measure of roughness. This pairs nicely with Gaussian fidelity as a computational device, again yielding a data augmentation strategy. But the rationale for this quadratic form of the penalty remains, at least to us, a bit mysterious. The total variation form of the penalty has the advantage that it is less sensitive to sharp bends in g along the edges and thus may be better adapted to sharp edges and spiky behavior. This viewpoint has been emphasized in the image processing literature where penalties based on total variation of the function itself, rather than its gradient, have been widely employed and rationalized as an edge detection device. While there may be perhaps reason to believe that the quadratic form will retain similar features, we cannot offer any positive evidence for that—except perhaps for Figure 2.

3. AN EXAMPLE

To illustrate the fexibility of this approach we briefly describe an application. The data consist of 1,194 vacant land sales occuring at 761 distinct sites in the Chicago metropolitan area during the period 1995–97. Extending the model described in Koenker and Mizera (2001), we consider the partially linear model

$$\log(z_i) = g(x_i, y_i) + \beta_1 \log(s_i) + \beta_2 \log^2(s_i),$$

where z_i denotes the sale price of the land in dollars per square foot and depends on the geographic location (x_i, y_i), measured in miles from the intersection of State and Madison, and the effect of the size of the parcel, s_i, in square feet is modeled as quadratic in $\log(s_i)$.

In Figures 1 and 2 we illustrate two fits of this model: one representing the median fit using the absolute error fidelity and the total variation form of the triogram penalty, the other representing a mean fit using Gaussian fidelity and the quadratic version of the penalty. The smoothing parameter λ was chosen to achieve roughly the same (Gaussian) fidelity in both plots. In both plots the parcel size is set to its sample (geometric) mean. Evaluating the parcel size effect at the mean parcel size, we obtain for the median fit -0.5148, indicating that parcel prices are roughly proportional to frontage (square root of area) rather than to area itself.

ACKNOWLEDGMENTS

This research was supported in part by NSF Grant SES-99-11184 and by the National Scientific and Engineering Research Council of Canada. The authors thank Peter Colwell and Henry Munneke for providing the Chicago land value data.

Comment

Mary J. Lindstrom

The authors give an insightful overview of spline regression and extended linear models and compare a number of methods for model selection. Table 1 in this Comment divides the methods into the four classes described at the end of Section 1 of the paper. The methods grouped together in the third and fourth lines of Table 1 differ only in the specification of the prior on K (the number of knots) and (for the fourth line) the prior on the coefficient vector $\boldsymbol{\beta}$.

The methods are distinguished by the definition of the estimate, the handling of $\boldsymbol{\beta}$, the computational algorithm and the method for choosing the model size. The definitions of the estimates are relatively straightforward. The handling of $\boldsymbol{\beta}$ is not an issue for the first two methods since, for any fixed $\mathbf{t}$, $\mathrm{AIC}(\boldsymbol{\beta}, \mathbf{t})$ is minimized at $\mathrm{AIC}(\hat{\boldsymbol{\beta}}(\mathbf{t}), \mathbf{t})$, where $\hat{\boldsymbol{\beta}}(\mathbf{t})$ is the MLE for $\boldsymbol{\beta}$. In contrast, for the MCMC-based estimation, the choice of whether to remove $\boldsymbol{\beta}$ from the estimation problem [methods (iii) and (iv)] or to specify a prior for $\boldsymbol{\beta}$ [methods (v) through (viii)] will affect the resulting estimates. Complicating matters is the effect of the prior for $\boldsymbol{\beta}$ on the relative posterior probabilities of the various model sizes. A diffuse prior will tend to drive down the size of the model with highest posterior probability. Thus for methods (v) through (vii) the model size is controlled by the priors on both $\boldsymbol{\beta}$ and K. Model size is controlled by the penalty in the AIC for the first two methods and by the prior on K for methods (iii) and (iv).

The computational algorithms consist of the following: a simple stepwise method for optimizing the AIC; a simulated annealing algorithm to optimize the AIC (which the authors state is nearly equivalent to MCMC estimation where the posterior mode is chosen as the point estimate); and reversible jump MCMC (RJMCMC) algorithms which, while not usually thought of as optimization algorithms, can be viewed as maximizing an objective function (the posterior) by generating a sample from the domain of the function which is concentrated in places where the objective function is largest (assuming the chain mixes well). However, the estimate used for these methods is not simply the fit corresponding to the sampled parameter vector with largest posterior probability but is an average of the sampled fits.

It is clear that the eight methods are not simply alternate ways to optimize an objective function. Appropriately, the authors base their simulation comparisons on the ability to recover the truth. The comparison of the greedy algorithm to the stochastic methods is of interest since the greedy algorithm is in general use. However, it would also be of interest to fix some measure of overall computing time (perhaps number of model evaluations) and compare to the greedy algorithm with random restarts. That is, once no additional improvement is possible, select a new K and a new $\mathbf{t}$ at random and start the algorithm again, repeating until the time limit is reached. A more sophisticated scheme would be to allow knot addition, knot deletion or knot movement at each step. This sort of heuristic algorithm (often called steepest descent with random restarts) usually does quite well against other heuristic optimization schemes, including genetic algorithms, simulated annealing and taboo searches. The random restarts are crucial. No deterministic algorithm run from one starting value will do well against stochastic algorithms when there are numerous local optima.

The real advantage of the RJMCMC methods over heuristic algorithms for optimizing an information criterion is that the output from the algorithm can be used to assess the variability of the estimate—including the uncertainty due to estimating K. It would be of interest to compare the estimates of variability for method (iii) [$\boldsymbol{\beta}$ fixed at $\hat{\boldsymbol{\beta}}(\mathbf{t})$] and those for methods (v) and (viii) ($\boldsymbol{\beta}$ not fixed). The downside of the Bayesian formulation for models which have a variable number of parameters is that the results are more than usually dependent on the choice of prior. A noninformative prior cannot be used for K and, as mentioned above, the prior for $\boldsymbol{\beta}$ (or any parameter vector which changes dimension) cannot be too diffuse or the smallest model will have highest posterior probability.

The extension of the methods to multiple dimensions is an important problem. It would seem that the radial basis functions used in thin-plate smoothing

Mary J. Lindstrom is Associate Professor, Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, Wisconsin 53792-4675 (e-mail: lindstro@biostat.wisc.edu).

TABLE 1
Summary of the eight methods; here $\hat{\boldsymbol{\beta}}(\mathbf{t})$ *is the MLE for* $\boldsymbol{\beta}$ *given the knot sequence* $\mathbf{t}$, *and* K *is the number of knots; method* (iii) (*Geometric prior on* K) *did best in the third grouping, and methods* (v) ($\lambda = 1/n$, *Uniform prior on* K) *and* (viii) ($\lambda = 1/n$, *Geometric prior on* K) *did best in the fourth grouping*

Method(s)	Estimate	Handling of $\boldsymbol{\beta}$	Algorithm	Model size
(i)	Minimizer of AIC	Fixed at $\hat{\boldsymbol{\beta}}(\mathbf{t})$	Greedy	AIC
(ii)	Minimizer of AIC (approximately the posterior mode)	Fixed at $\hat{\boldsymbol{\beta}}(\mathbf{t})$	Simulated annealing	AIC
(iii), (iv)	Mean of predicted values for posterior sample	Fixed at $\hat{\boldsymbol{\beta}}(\mathbf{t})$	RJMCMC	Prior on K
(v)–(viii)	Mean of predicted values for posterior sample	Prior on $\boldsymbol{\beta}$ parameterized by λ	RJMCMC	Prior on K and prior on $\boldsymbol{\beta}$

splines would make a more natural generalization than the triangulation approach used by the authors. The thin-plate radial basis functions are smooth and easily calculated, and each basis function is defined by a single location, making addition and deletion straightforward.

Comment

Grace Wahba, Yi Lin and Chenlei Leng

1. INTRODUCTION AND THANKS

The authors present greedy and Bayesian model selection frameworks for studying adaptation in the context of an extended linear model, with application to logspline density estimation and bivariate triogram regression models. We will confine our remarks to the density estimation case. The authors define the setup of their "extended linear model" as finding $g \in G$ to maximize the log likelihood

$$l(g) = \sum_i l(g, W_i), \tag{1}$$

where G is a linear space, generally of much lower dimension than the sample size n. Generally the famous bias–variance tradeoff is controlled (most likely primarily) by the dimension of G, as well as other parameters involved in the choice of G or spline spaces in Hansen and Kooperberg (HK), the number of knots governs the dimension of the space and the number and location of the knots are to be chosen according to several Bayesian methods and compared with a greedy method. Knot selection in the context of (1) is a difficult but not impossible task, as the authors clearly show. The authors are to be thanked for an interesting study of Bayesian knot selection methods and their comparison with a greedy knot selection method.

To contrast with the ELM approach in the paper, we will examine a penalized likelihood method for the same (log) density estimation problem. It is based on solving a variational problem in an infinite dimensional (Hilbert) space, where the problem has a Bayesian flavor, and where the solution to the variational problem is (essentially) known to lie in a particular n-dimensional subspace. Then the smoothing parameter(s) are chosen by a predictive loss criteria. If the penalty functional is square integral second derivative, the n-dimensional subspace is spanned by a basis of cubic splines with knots at the observation points. At this point we can take one of several points of view. The three that are relevant to the discussion here are the following: (i) solve the variational problem exactly; (ii) find a good approximation to the solution of the variational problem, by using a representative or a random sample of the knots, instead of the complete set, when the sample size is large; and (iii) instead of using the solution of the variational problem as the "gold standard" as in (ii), use a greedy algorithm to choose a subset of

Grace Wahba is Professor, Yi Lin is Assistant Professor and Chenlei Leng is Research Assistant, Department of Statistics, University of Wisconsin–Madison, Madison, Wisconsin 53792 (e-mail: {wahba,yilin,chenlei}@stat.wisc.edu).

the knots, actually, a subset of the *representers* (Wahba, 1990), which reduce to the knots in the case of polynomial splines. This will have the effect of letting the "wiggliness" of the solution vary where there are more observations, and/or more variable responses. Then the variational problem is solved in the greedily chosen subspace. This so-called hybrid approach was taken in Luo and Wahba (1997) in a Gaussian regression problem, using a relatively simple greedy algorithm, and, as was also found in Stone, Hansen, Kooperberg and Truong (1997) more knots are located near sharp features, as well as where there are more observations.

We will focus on a density estimation version of (ii) in the rest of this discussion. To carry out this program we need a criterion for the choice of the smoothing parameters appropriate for density estimation and we will use randomized generalized approximate cross validation (GACV) for density estimation (to be defined), which is a proxy for the comparative Kullback–Leibler distance of the "truth" from the estimate. In this discussion we will first give some details for the univariate case and compare the results to Table 2 of HK. Loosely speaking, the results compare fairly favorably with all of the estimates whose MISE performance is given in Table 2 with the exception of the two largest sample sizes in the "sharp peak" example. After commenting on these results, we will then describe some work in progress, in which the penalized likelihood estimate is extended to several dimensions via a smoothing spline ANOVA (SS-ANOVA) model. We briefly demonstrate a three-dimensional result. The conceptual extension of the penalized likelihood method to higher dimensions is fairly straightforward, and the real thrust of the work is to be able to estimate densities in higher dimensions. One of the rationales behind the use of the SS-ANOVA model for density estimation in several dimensions is that the pattern of main effects and interactions has an interesting interpretation in terms of conditional dependencies and can thus be used to fit graphical models (Darroch, Lauritzen and Speed, 1980; Whittaker, 1990; Jordan, 1998) nonparametrically.

2. PENALIZED LOG LIKELIHOOD DENSITY ESTIMATION

Our density estimate is based on the penalized log likelihood estimate of Silverman (1982). When going to higher dimensions we will use the basic ANOVA decomposition idea in Gu (1993). Our density estimate will have compact support Ω, which will be scaled to the unit interval or the unit cube in E^d and then rescaled back after fitting. Let the density $p = e^g$ with g in some reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with square seminorm $J(g)$, where the null space of J contains the constant function and is low dimensional. Letting $x_i \in \Omega$, Silverman showed that the penalized log likelihood minimization problem min $g \in \mathcal{H}$

$$-\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \lambda J(g) \tag{2}$$

subject to the condition

$$\int_\Omega e^g = 1 \tag{3}$$

is the same as the minimizer of

$$\mathcal{l}_\lambda(g) = -\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \int_\Omega e^g + \lambda J(g). \tag{4}$$

We will describe the estimate in general form so that its extension from the univariate to the multivariate case is clear. Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is the null space of J, and let the reproducing kernel for $\mathcal{H}_1$ be $K(x, x')$. If the term $\int_\Omega e^g$ were not in (4), then (it is well known that) the minimizer of (4) would be in $\mathcal{H}^n \equiv \mathcal{H}_0 \oplus \text{span}\{\xi_i,\ i = 1, \ldots, n\}$, where $\xi_i(x) = K(x, x_i)$. (ξ_i is known as a representer.) We will therefore feel confident that the minimizer of (4) in $\mathcal{H}^n$ is a good approximation to the minimizer of (4) in $\mathcal{H}$. In fact, we will seek a minimizer in $\mathcal{H}^N = \mathcal{H}_0 \oplus \text{span}\{\xi_{i_r},\ r = 1, \ldots, N\}$, where the i_r is a representative subset chosen sufficiently large that the minimizer in $\mathcal{H}^N$ is a good approximation to the minimizer in $\mathcal{H}^n$.

In order to carry out penalized log likelihood estimation a method for choosing λ is required. We have obtained a randomized generalized approximate cross validation (ranGACV) estimate for λ, for density estimation. We briefly describe it here; details will be given elsewhere. Let f_λ be the estimate of the log density, and let $f_\lambda^{[-i]}(x_i)$ be the estimate with the ith observation left out. Define the ordinary leaving-out-one function as

$$V_0(\lambda) = \text{OBS}(\lambda) + D(\lambda), \tag{5}$$

where

$$\text{OBS}(\lambda) = -\frac{1}{n}\sum_{i=1}^{n} f_\lambda(x_i) \tag{6}$$

and

$$D(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[f_\lambda(x_i) - f_\lambda^{[-i]}(x_i)\right]. \tag{7}$$

TABLE 1

Distribution	Sample size	MISE (pen.log.lik)	HK [Table 2(i)]	Ratio (pen.log.lik/HK)
Normal	50	0.01859	0.02790	0.666
	200	0.00435	0.01069	0.407
	1,000	0.00071	0.00209	0.340
	10,000	0.00014	0.00020	0.700
Bimodal	50	0.01358	0.02502	0.543
	200	0.00372	0.00770	0.483
	1,000	0.00079	0.00164	0.482
	10,000	0.00011	0.00020	0.550
Peak	50	0.10011	0.15226	0.657
	200	0.03045	0.03704	0.822
	1,000	0.02152	0.00973	2.212
	10,000	0.01624	0.00150	10.83

Elsewhere (to appear) we show that $nD(\lambda)$ can be approximated by the trace of the inverse Hessian of $\mathcal{I}_\lambda$ with respect to $f_\lambda(x_i)$, $i = 1, \ldots, n$, and that it can be estimated by a randomization technique as follows. Let $\mathcal{I}_\lambda(g, y)$ be

$$\mathcal{I}_\lambda(g, y) = -\frac{1}{n}\sum_{i=1}^{n} y_i g(x_i) + \int_\Omega e^g + \lambda J(g). \tag{8}$$

When $y = (1, \ldots, 1)'$ then (8) becomes (4). Letting f_λ^y be the minimizer of (8), $D(\lambda)$ is estimated as

$$\hat{D}(\lambda) = \frac{1}{n\sigma_\varepsilon^2}\varepsilon'(f_\lambda^{y+\varepsilon} - f_\lambda^y), \tag{9}$$

where $y = (1, \ldots, 1)'$, ε is a random vector with mean 0 and covariance $\sigma_\varepsilon^2 I$ and, with some abuse of notation, $f_\lambda^z = (f_\lambda^z(x_1), \ldots, f_\lambda^z(x_n))'$. Several replicates in ε may be used for greater accuracy. Then

$$\text{ranGACV}(\lambda) = \text{OBS}(\lambda) + \hat{D}(\lambda). \tag{10}$$

Our numerical results (to appear) show that ranGACV is a good proxy for the comparative Kullback–Leibler distance between the density determined by f_λ and the true density.

3. THE UNIVARIATE ESTIMATE

The procedure is to start with N representers. In the one-dimensional case we choose roughly equally spaced order statistics. Fix λ large. Use a Newton–Raphson iteration to estimate the coefficients of f_λ in the basis functions spanning $\mathcal{H}^N$. Evaluate ranGACV(λ). Decrease λ and repeat, until the minimizer over λ is found. Double N and repeat. Compare the resulting estimates with N and $2N$; if they agree within a specified tolerance, stop, otherwise double N again. We tried this penalized log likelihood estimate on the examples in HK, using $\mathcal{H} = W_2^2 \equiv \{g : g, g' \text{ abs. cont.}, g'' \in \mathcal{L}_2\}$ and $J(g) = \int_0^1 (g''(x))^2$. In this case $\mathcal{H}_0$ is spanned by linear functions and $K(x, x') = k_2(x)k_2(x') - k_4([x - x'])$, $x \in [0, 1]$, where $[\tau]$ is the fractional part of τ and $k_m(x) = B_m(x)/x!$, where B_m is the mth Bernoulli polynomial. The estimate is a cubic spline (Wahba, 1990) with knots at the x_{i_r}. In the one-dimensional case this is not the most efficient way to compute this estimate, since a B-spline basis is available given the knots, and that will lead to a sparse linear system, whereas the present representation does not. However, this representation generalizes easily to higher dimensional estimates. In our experiment the maximum allowed N was 48. We made 100 replicates of each case in Table 2 of HK, and computed the MISE in the same way as HK did, by averaging the squared difference over 5,001 equally spaced quadrature points in the three intervals (for the normal, slight bimodal and sharp peak cases) of $[-5, 5]$, $[-7, 7]$ and $[0, 12]$. (See our Table 1.)

We note that the ratio column suggests that this estimate is among the better estimates in HK's Table 2 with the exception of the $n = 1{,}000$ and $n = 10{,}000$ cases for the peak example.

4. MULTIVARIATE SMOOTHING SPLINE ANOVA DENSITY ESTIMATION

The univariate penalized log likelihood density estimation procedure we have described can be generalized to the multivariate case in various ways. Here we describe the smoothing spline ANOVA (SS-ANOVA) model. The use of SS-ANOVA in a density estimate

was suggested by Gu (1993), who also gave a method for choosing the smoothing parameter(s). It can be shown that (for the same smoothing parameters) the estimates of Gu and Silverman are mathematically equivalent; however, we found the variational problem in Silverman easier to compute. The problem in d dimensions is transformed to the d-dimensional unit cube, and $x_i = (x_{i1}, \ldots, x_{id})$. $\mathcal{H}$ will be an RKHS on the d-dimensional cube which is formed as the direct sum of subspaces of the tensor product of d one-dimensional RKHSs. Details of SS-ANOVA models may be found in Wahba (1990), Wahba et al. (1995) and Lin et al. (2000). Letting $u = (u_1, \ldots, u_d) \in [0,1]^d$, we have

$$(11)\quad g(u) = \mu + \sum_{\alpha=1}^{d} g_\alpha(u_\alpha) + \sum_{\alpha\neq\beta} g_{\alpha\beta}(u_\alpha, u_\beta) + \cdots,$$

where the terms satisfy averaging conditions analogous to those in ordinary ANOVA that insure identifiability, and the series may be truncated somewhere. The interesting feature of this representation of a log density is the fact that the presence or absence of interaction terms determines the conditional dependencies, that is, a graphical model (see Whittaker, 1990). For example, the main effects model represents independent component random variables, and if, for example, $d = 3$ and the g_{23} and g_{123} terms are missing, then the second and third component random variables are conditionally independent, given the first.

Let $\tilde{\mathcal{H}}$ be the d-fold tensor product of W_2^2 and let $\mathcal{H}$ be the subspace of $\tilde{\mathcal{H}}$ consisting of the direct sum of subspaces containing the terms retained in the expansion. (They are orthogonal in $\tilde{\mathcal{H}}$). We have $\int_0^1 g_\alpha(u_\alpha)\,du_\alpha = 0$, and so forth. The penalty functional $J(g)$ of (4) becomes $J_\theta(g)$, where the θ represents a vector of (relative) weights on separate penalty terms for each of the components of (11). As before $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is the (low dimensional) null space of J_θ. Let $K_\theta(x, x')$, $x, x' \in [0,1]^d$, be the reproducing kernel for $\mathcal{H}_1$, where θ has been incorporated into the norm on $\mathcal{H}_1$. (See Wahba, 1990, Chapter 10.) Let $\xi_i(x) = \xi_{i\theta}(x) = K_\theta(x, x_i)$. The same arguments hold as in the one-dimensional case, and we seek a minimizer of (4) (with $J = J_\theta$) in $\mathcal{H}^N = \mathcal{H}_0 \oplus \operatorname{span}\{\xi_{i_r\theta},\ r = 1, \ldots, N\}$, and λ and θ are chosen using the ranGACV of (10).

We will give a three-dimensional example, essentially to demonstrate that the calculations are possible and the ranGACV reasonable in higher dimensions. The SS-ANOVA model for this example contained only the main effects and two factor interactions, and we had altogether six smoothing parameters, parameterized in a convenient manner (details to appear elsewhere). For fixed smoothing parameters λ, θ the coefficients in the expansion in $\mathcal{H}^N$ are obtained via a Newton–Raphson iteration. In this case integrations over $[0,1]^3$ are required, and we used quadrature formulae based on the hyperbolic cross points; see Novak and Ritter (1996) and Wahba (1978b). These quadrature formulae seem particularly appropriate for SS-ANOVA models and make high dimensional quadrature feasible. Then the ranGACV was minimized over smoothing parameters via a six-dimensional downhill simplex calculation.

The underlying true density used in the example is $p(x) = 0.5N(\mu_1, \Sigma) + 0.5N(\mu_2, \Sigma)$, where $\mu_1 = (0.25, 0.25, 0.25)$, $\mu_2 = (0.75, 0.75, 0.75)$,

$$\Sigma = \begin{pmatrix} 10 & 0 & 10 \\ 0 & 20 & 30 \\ 10 & 30 & 80 \end{pmatrix}^{-1} = \begin{pmatrix} 0.14 & 0.06 & -0.04 \\ 0.06 & 0.14 & -0.06 \\ -0.04 & -0.06 & 0.04 \end{pmatrix}.$$

(This density has a nonzero three-factor interaction which is not in our two-factor model.) In this example the sample size was $n = 1{,}000$. $N = 40$ and the 40 representers were randomly chosen from among the n possibilities. The $N = 80$ estimate was essentially indistinguishable from the $N = 40$ case. (Note that the smoothing parameters will not generally be

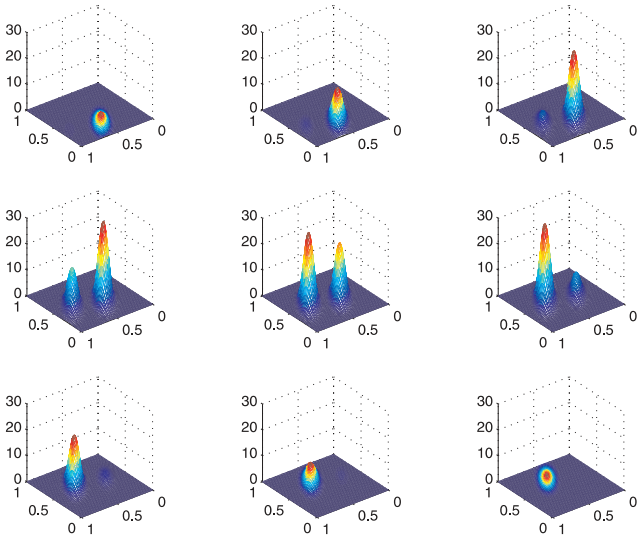

FIG. 1. *The true density: $x_1 = 0.1, \ldots, 0.9$ is fixed in the plots, left to right, then top to bottom.*

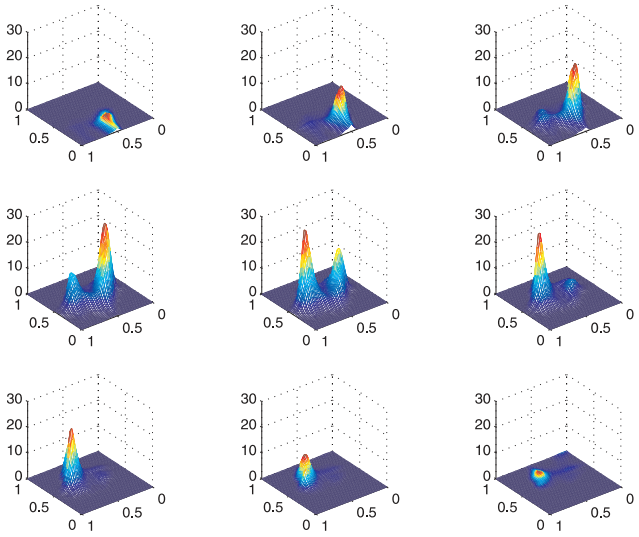

FIG. 2. *The estimated density*: $x_1 = 0.1, \ldots, 0.9$ *is fixed in the plots, left to right, then top to bottom.*

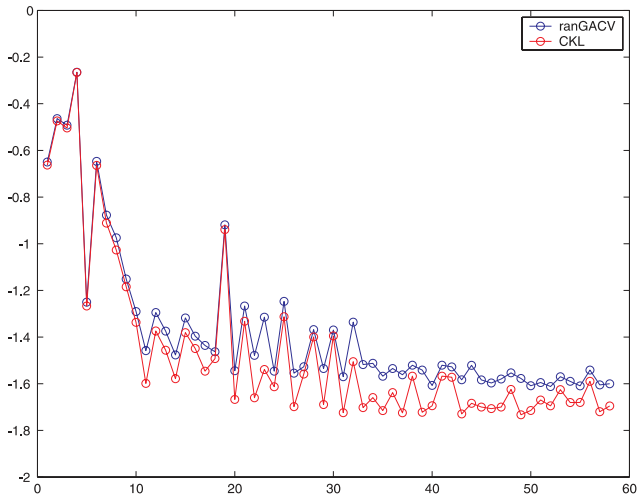

FIG. 3. *The ranGACV and the CKL compared*: *the horizontal axis is iteration number, using the downhill simplex method*; *the ranGACV is minimized and the ranGACV and CKL are computed at the minimizer at each step.*

the same in the two cases.) Figure 1 gives cross sections of the true density, and Figure 2 gives the SS-ANOVA penalized log likelihood estimate. Figure 3 compares the ranGACV and the CKL ($\mathrm{CKL}(\lambda) = -\int_\Omega f_{\lambda,\theta}(u)\,p(u)\,du$) as a function of iteration number in a downhill simplex minimization of the ranGACV.

5. CLOSING REMARKS

We have compared a penalized likelihood density estimate with ranGACV to choose the smoothing parameter(s) for the greedy density estimate and the Bayesian estimates in ELM models considered by HK. Fairly favorable results were obtained except in the highest n peak cases. We have shown that these penalized likelihood estimates can be extended to the multivariate case (work in progress). It remains to develop tests to allow the construction of graphical models from the SS-ANOVA estimates in higher dimensions.

We would be interested in knowing to what extent the Bayesian model selection methods can be incorporated in to ELM estimates for the multivariate case.

Splines of various flavors have been widely adopted in many statistical problems. It is interesting to compare the various flavors and we are pleased to compliment the authors and contribute to the discussion.

ACKNOWLEDGMENT

Research supported in part by NIH Grant RO1 EYO9946 and NSF Grant DMS-00-72292.

Comment

David Ruppert

I congratulate the authors for a very fine paper. There is nothing more satisfying than a broad theory, such as the extended linear models (ELM) methodology presented here, that solves a wide range of practical problems within a unified framework. The authors's clear introduction to ELM will be much appreciated by those seeking to understand and apply nonparametric models.

David Ruppert is Professor, Department of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853 (e-mail: davidr@orie.cornell.edu).

By nonparametric modeling I do not mean the absence of parameters but rather situations where the shape of the curve or surface fit to the data is determined by the data themselves, not by the dictates of a predetermined model. In fact, nonparametric modeling typically is implemented by the flexible use of *parametric* models. There are at least three general approaches to nonparametric modeling:

- *local fitting*—where a simple parametric model is fitted separately in neighborhoods defined by covariates;
- *model selection*—meaning selection of a model or models from within a rich class of parametric models; selection is followed by global maximum likelihood estimation either applied to a single model or averaged over all models according to posterior probabilities;
- *regularization*—meaning penalized global estimation using a large, overparametrized model.

In this discussion, I will attempt a brief overview of these three approaches to highlight their relative strengths and weaknesses. My hope is to show the place of ELM within the large toolkit of nonparametric estimation techniques. Hansen and Kooperberg's paper has already provided a very clear introduction to the model selection method of nonparametric modeling, so I will not discuss that approach much.

Local fitting includes the familiar local polynomial method of nonparametric regression where the regression function is estimated at each point x by weighted least-squares fitting of a low degree polynomial, with weights decreasing with distance from x. My experience with local regression has been quite positive overall, but local fitting methods are best suited to simpler situations such as univariate or bivariate regression.

For multivariate data, one usually needs to reduce the dimensionality by, for example, using an additive model or a semiparametric model. Then, "localness" varies across different aspects of the model. For example, when fitting an additive model, the local neighborhoods change with each covariate and an iterative backfitting algorithm is necessary. As another example, single index models assume that the expected response is $\eta(\boldsymbol{\alpha}^\top \mathbf{x})$, where $\mathbf{x}$ is a vector of covariates, $\boldsymbol{\alpha}$ is a parametric vector and η is an unknown univariate function. This is a semiparametric model where the *index* $\boldsymbol{\alpha}^\top \mathbf{x}$ is the parametric component and $\eta(\cdot)$ is the nonparametric component. Estimation of the parametric component requires an unweighted fit to all of the data, while estimation of $\eta(\cdot)$ uses local neighborhoods defined by the *estimated* indices $\widehat{\boldsymbol{\alpha}}^\top \mathbf{x}$. Carroll, Fan, Gijbels and Wand (1997) have developed an algorithm that iterates between global estimation of $\boldsymbol{\alpha}$ and local polynomial estimation of $\eta(\cdot)$ but the algorithm can be computationally unstable because the neighborhoods change as $\boldsymbol{\alpha}$ is updated (Wand, personal communication). The problem of differing local neighorhoods disappears if one uses model selection or regularization and for this reason I favor these techniques. For example, Yu and Ruppert (2002) use regularization to fit single index models. Their algorithm uses standard nonlinear least-squares software and is computationally stable. Also, for nonparametric regression with covariate measurement error, one would really like to base the local neighborhoods upon the true covariate values, but this is of course impossible. This problem is likely to be one main reason why regularization is the best current method for handling measurement error; see below.

The key idea of ELM and other model selection methods is that overfitting is avoided by careful choice of a model that is both parsimonious and suitable for the data. Regularization takes a different approach to the prevention of overfitting. Rather than seeking a parsimonious model, one uses a highly parametrized model and imposes a penalty on large fluctuations on the fitted curve. Suppose the nonparametric component of a model is a function f. We model $f(x)$ using a linear combination of the basis functions $\mathbf{B}(x) = \{B_1(x), \ldots, B_M(x)\}$, for example, the truncated power basis functions given by (1) or (2) of the paper. The dimension M is chosen to be large, so that some linear combination of the basis functions, say, $\mathbf{B}(x)^\top \boldsymbol{\beta}$, will be close to f. Let $\ell(\boldsymbol{\beta})$ be the log-likelihood and let $P(\boldsymbol{\beta})$ be a nonnegative penalty function. Then $\widehat{\boldsymbol{\beta}}$ minimizes

$$-\ell(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}),$$

where $\lambda \geq 0$ is a penalty parameter. For univariate regression, if $\mathbf{B}(x)$ is the natural cubic spline basis functions with knots at the unique values of x in the data and if $P(\boldsymbol{\beta}) = \int \{\mathbf{B}''(x)^\top \boldsymbol{\beta}\}^2 \, dx$, then one obtains a cubic smoothing spline.

As discussed in Ruppert (2002), penalized splines or P-splines, generalize smoothing splines by allowing any spline basis and any form of the penalty. For example, the knots might be any regularly spaced quantiles of the x values (e.g., every tenth unique x value) rather than a knot at every unique x. Although there are fast algorithms for univariate smoothing

splines, for more complex models such as nested curves in Brumback and Rice (1998), using an excess of knots can slow down computations significantly. However, there is no need for a knot at every data point and major computational speedups are possible with fewer knots.

As an alternative to the usual quadratic integral penalties of derivatives that yield smoothing splines as estimates, Ruppert and Carroll (2000) suggest penalizing the sum of the squared jumps at the knots of the pth derivative of a pth degree spline. One interesting feature of this penalty is that it has some useful variants that are easily implemented. If one uses the truncated power basis $\{1, x, \ldots, x^p, (x - t_1)_+^p, \ldots, (x - t_K)_+^p\}$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p+K})^{\mathsf{T}}$, then $P(\boldsymbol{\beta}) = \sum_{k=1}^{K} \beta_{p+k}^2$. An alternative penalty suggested by Ruppert and Carroll (2000) and studied in Yu and Ruppert (2001) is

$$P(\boldsymbol{\beta}) = \sum_{k=1}^{K} |\beta_{p+k}|^q, \tag{1}$$

where $q > 0$. For $q = 1$ we are using Tibshirani's (1996) lasso, originally developed for parametric linear regression. The lasso has the feature that many of the components of $\boldsymbol{\beta}$ are shrunk all the way to 0. In effect, these coefficients are deleted. Therefore, the lasso is similar to model selection. If $q < 1$, then this shrinkage of coefficients to 0 is even more pronounced.

A potentially serious problem with smoothing splines is a lack of spatial adaptivity. Spatial adaptivity is the ability to impose less smoothing where the regression function has sharp curvature or other features that require higher resolution and more smoothing where the regression function is relatively flat. Local fitting methods can achieve spatial adaptivity if they allow the bandwidth to vary spatially. Model selection methods are spatially adaptive because they can place knots at a higher density around features of the regression function. Unfortunately, smoothing splines, with their single global smoothing parameter, cannot adjust to spatial inhomogeneity of the regression function. There are at least three ways to make P-splines spatially adaptive. One is to use penalty (1) with $q \leq 1$ so that large jumps are not unduly penalized as with $q = 2$. Another is to let the penalty parameter λ vary spatially. Ruppert and Carroll (2000) developed a P-spline estimator with penalty

$$P(\boldsymbol{\beta}) = \sum_{k=1}^{K} \lambda(t_k) \beta_{p+k}^2.$$

Here $\log\{\lambda(\cdot)\}$ is a linear spline, with knots a small subset of the knots for f. The penalty $\lambda(\cdot)$ should be small where there are features in the regression function and large where that function is flat. This behavior can be achieved automatically by choosing the parameters in $\lambda(\cdot)$ by GCV; see Ruppert and Carroll (2001), who show that for inhomogeneous regression functions spatially adaptive smoothing splines are competitive with the model selection spline estimate of Smith and Kohn (1996).

A third method of achieving spatial adaptivity is to place the knots more closely together in regions containing features. Luo and Wahba's (1997) hybrid adaptive splines (HAS) do precisely that. However, the initial knot location step of their algorithm is very similar to ELM and it is not clear from Luo and Wahba's work whether the second step where a roughness penalty is imposed is really necessary. If not, then HAS could be considered primarily a model selection approach, not regularization.

There are interesting connections between regularization, Bayesian estimation and mixed models. Wahba (1978a) first noticed that a smoothing spline can be derived as the mode of the posterior density when a particular prior is used. The smoothing parameter is the ratio of the residual variance to the variance in the prior for the regression function; that is, it is a noise-to-signal ratio. This Bayesian formulation can also be viewed as a mixed model, with the regression function being a random effect and the smoothing parameter being the ratio of variance components. This fact allows one to estimate the smoothing parameter by residual maximum likelihood (REML), sometimes called generalized maximum likelihood (GML) or marginal likelihood in the spline literature; see Wahba (1985) and Kohn, Ansley and Tharm (1991). Therefore, nonparametric curve fitting including the choice of smoothing parameter can be implemented by standard mixed model software, for example, lme in S-PLUS and Proc Mixed in SAS. Within the mixed model framework other random effects can be introduced, for example, to handle longitudinal data as in Zhang, Lin, Raz and Sowers (1998). Moreover, modeling the regression function with random effects applies to non-Gaussian responses using generalized linear mixed models (GLMMs). See, for example, Lin and Zhang (1999) or Coull, Ruppert and Wand (2001), who have an application to pollen count data.

If the smoothing parameter is estimated by REML, for example, and then treated as fixed, which is a type

of empirical Bayes estimation, then confidence intervals do not fully reflect uncertainty in the smoothing parameter. However, a fully Bayesian analysis is possible using MCMC. Such an analysis can even be extended to data where the predictor variables are measured with error. See Berry, Carroll and Ruppert (2002), where in an empirical study a fully Bayesian analysis using P-splines was the best of several techniques for nonparametric regression with measurement error.

Local fitting will remain an attractive method for simple smoothing problems, but model selection such as ELM and regularization such as P-splines will grow in popularity as nonparametric and semiparametric models are applied to more complex types of data. Since model selection and regularization both work extremely well in practice, it is unlikely that either method will surplant the other. Which approach works best in a particular application will depend both on the nature of that application and, perhaps even more so, on which approach is most familiar to the researcher. Both model selection and regularization view semiparametric and nonparametric modeling as extreme cases of parametric modeling. This viewpoint is fortunate, since it brings about a unity in statistics. Rather than nonparametric modeling using its own special techniques, such as local kernel weighting, nonparametric modeling utilizes the same techniques as parametric modeling, for example, model selection, model averaging, mixed models, Bayesian analysis, REML and MCMC. This common framework is important for those of us teaching the next generation of statisticians. There is time to teach only so much, so the economies of a unified approach to statistics are essential.

Rejoinder

Mark H. Hansen and Charles Kooperberg

INTRODUCTION

First, we thank the Editor, George Casella, for inviting comments on our paper from such a distinguished group of researchers. The diversity of approaches represented here certainly enriches the practice of function estimation. To best address these different perspectives, we have organized our responses around each of their discussions separately. However, there are some common themes that we would like to briefly highlight here because we will return to them repeatedly during our Rejoinder.

- *Priors*—For function estimation, priors amount to statements about the smoothness of the underlying function ϕ. To the extent that subjective elicitation of prior distributions is possible here, formulations in terms of characteristics of ϕ will be most successful. Priors of convenience developed for generic model selection in regression or other settings may not be appropriate when working with finite-dimensional approximation spaces like splines.
- *Bayesian confidence intervals*—In principle, the Bayesian formalism provides an automatic mechanism for making inference about features of ϕ. Unfortunately, we find very few examples of confidence intervals for model-averaged spline estimates in the Bayesian literature on function estimation. Perhaps one difficulty here is that the object we are estimating is not really a cubic spline, and there is always bias present. In our own experience, it has been difficult to properly calibrate the prior distribution to produce both sensible point estimates and confidence intervals.
- *Bayesianism*—Many of the discussants comment on what is or what is not a Bayesian procedure. Our interest was mainly in evaluating computational methods, and each of the simulation setups we studied was motivated by our previous experience with greedy schemes. We will not offer opinions as to whether previously published approaches are legitimately Bayesian or not.

Finally, we want to remind readers of an early paper in this area that has gone somewhat unreferenced in the modern literature on Bayesian splines. Halpern (1973) presented a thoughtful treatment of Bayesian knot selection that is very close to the methods mentioned in our paper and in this discussion. Unfortunately, the state of Bayesian computation in 1973 made most applications intractable.

CHIPMAN, GEORGE AND McCULLOCH

We thank Chipman, George and McCulloch for sharing their experiences formulating highly adaptive Bayesian methods for function estimation. The work by this trio on model selection, stochastic search and Bayesian versions of CART has influenced much of our paper. In terms of their characterization of priors as "a bet on what kind of models we want to consider," we feel that this gamble is best worded in terms of the object being studied, a function. Chipman, George and McCulloch (2002) take an approach similar to ours when developing tree-based estimators. In their paper we find a nice division of prior specifications into a linear model class (representing separate tree structures; analogous to our linear spaces G) and then a member given the class (through choosing parameters; our selection of $g \in G$). Chipman et al. (2002) specify a prior on G to control the complexity of the trees that one expects, formulated in terms of the number of splits. This can be tuned to reflect one's beliefs that the data can best be described by small or large trees. As Chipman, George and McCulloch (CGM) point out, the prior $p(g|G)$ also plays a role in determining how the posterior weights models of different complexity (dimension). Given a knot sequence $\mathbf{t}$ and a basis set $B_1, \ldots, B_J$, we select this prior via a distribution on the coefficient vector $\boldsymbol{\beta}$ in the expansion

$$\text{(R1)} \qquad g(x; \boldsymbol{\beta}, \mathbf{t}) = \sum_j \beta_j B_j(x; \mathbf{t}).$$

In our paper, this was always taken to be normal with mean zero and covariance $(\lambda A)^{-1}$. With a special choice of A not related to smoothing, it is possible to calibrate the influence of $p(g|G)$ on posterior weights for each model class in terms of well-known selection criteria like AIC and BIC. We will have more to say about this in our discussions of Holmes and of Kass and Wallstrom.

When developing (greedy) Logspline and Triogram, we attempted to make the "smoothing parameters" understandable and controllable through our software implementation. In so doing, users can tune these methods to agree with their prior beliefs about the smoothness of the function ϕ they are estimating. This also allows users to experiment with strategies like CGMs cross-validation. In the early stages of an analysis, however, flexible smoothers are often used in an exploratory way, helping users formulate hypotheses about their data. In this setting, Logspline and Triogram function more like black boxes, and it is important to have reasonable automatic settings for hyperparameters. Fast computation is also important, given the iterative nature of data analysis. In crafting the simulations for this paper, we hoped to identify sensible default values for Bayesian versions of our ELM methods. By expressing the hyperparameters of our models in terms of the complexity of the underlying function, users can still easily incorporate their beliefs about ϕ in a natural way.

CGM are right about Logspline mixing more rapidly than the Triogram procedures. This difference also appears in the greedy implementation, where many of the Triogram models found during addition were visited again during deletion. At each deletion step, we are only able to remove certain vertices and still maintain a proper triangulation. In fact, at each step the number of candidates for deletion is limited by the structure imposed on the triangulations by the addition process itself; vertices added in the first few steps are typically too "connected" to be eliminated early in the deletion process. We hoped that a Bayesian Triogram procedure would allow us to visit more models. In thinking about how to implement the Triogram sampler, we borrowed from the experience of CGM with trees. In some sense, the structure imposed by our stepwise triangulation process is similar to the node splitting used for building trees. We also explored the tradeoffs between starting several chains versus a single, long chain. The simulations in our paper represent a technique that seemed to work best overall.

We agree with CGM on the potential gains of bootstrapping. Our present paper was inspired by an

TABLE R1
MISE for bagging logspline on the 12 *cases from Table* 2

	Distribution												
	Normal				**Slight bimodal**				**Sharp peak**				
n	50	200	1,000	10,000	50	200	1,000	10,000	50	200	1,000	10,000	**Average**
Ratio over (i)	0.88	0.96	1.05	0.80	0.76	0.97	0.95	0.65	0.72	0.84	0.77	0.69	0.84

attempt to assess the usefulness of (then) developing Bayesian computational methods. Over the last five years, techniques like stacking, bagging and boosting have emerged from, and in response to, ideas in computer science and machine learning. At this point, they have developed into extremely powerful, general-purpose methods for combining estimators. To see what kinds of gains we can expect in our applications, we implemented a bagged version of Logspline based on 25 bootstrap samples. The column for Table 2 in our original paper corresponding to this new estimate is given in Table R1. We are not sure why bagging seems to have the hardest time with moderate sample sizes (200 and 1,000) for the two smoothest distributions. Overall bagging improves over the greedy algorithm, but not as much as some of the MCMC approaches.

HOLMES

We are pleased to have Holmes as a discussant. In the past few years, Holmes and his collaborators have applied the latest sampling techniques to problems in function estimation. From Bayesian versions of MARS that employ reversible jump Markov chain Monte Carlo to wavelet methods via perfect sampling, this group has contributed substantially to the practice of Bayesian curve and surface fitting.

Holmes highlights a rather general problem encountered when specifying priors for Bayesian model selection. For simplicity, consider the normal linear regression model in a J-dimensional spline space G defined by a given knot sequence $\mathbf{t}$ of length K, and a basis (R1). Let $\boldsymbol{\beta}, \sigma^2$ have a conjugate normal-inverse gamma prior distribution

$$w(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{(-d+k+2)/2} \cdot \exp\left[\frac{-(\boldsymbol{\beta}-\mathbf{b})^t V^{-1}(\boldsymbol{\beta}-\mathbf{b})+a}{2\sigma^2}\right] \tag{R2}$$

that depends on several hyperparameters: $a, d \in \mathbb{R}$, the vector $\mathbf{b} \in \mathbb{R}^J$ and a $J \times J$ symmetric, positive definite matrix V. Valid ranges for these parameters include all values that make (R2) a proper density. We considered covariance matrices of the form $V = (\lambda A)^{-1}$, where A is chosen to reflect our beliefs about the smoothness of an unknown function ϕ. For a fixed set of knots, the prior precision parameter λ affects the look of the fitted curve in a fairly predictable way. Given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, the posterior distribution of $\boldsymbol{\beta}$ is multivariate t with mean

$$(B^t B + \lambda A)^{-1} B^t Y, \tag{R3}$$

where B is the $n \times J$ design matrix based on the basis (R1), $[B]_{ij} = B_j(X_i)$ and $Y = (Y_1, \ldots, Y_n)^t$. Therefore, selecting a small value of λ can eliminate the effect of smoothing, often producing a wiggly fit; while large values tend to shrink $\boldsymbol{\beta}$ toward the null-space of A (for the prior associated with cubic smoothing splines, this is simply the space of linear functions). In Bayesian terms, a small value of λ indicates little information about $\boldsymbol{\beta}$, while a large value of λ indicates a strong belief that ϕ is very smooth. As Holmes points out, however, things become more complicated when comparing spline spaces of different dimensions (having different numbers of knots). In this case λ partially specifies the weight assigned to each space and we are subject to the so-called Lindley–Bartlett paradox; a noninformative prior for a fixed model class becomes maximally informative when comparing models of different dimensions. In short, the posterior piles mass on the spline space with the fewest knots.

Holmes presents a well-known covariance-representation that alleviates the paradox and is meant to calm the non-Bayesian. Unfortunately, it does not eliminate the difficulty in specifying λ. Many researchers, including Holmes, have dealt with the problem of prior choice by calibrating Bayesian procedures with frequentist measures. In Holmes and Denison (1999), for example, we find a prior on the precision parameter λ that is tuned to the degrees of freedom of the fit, DF, or explicitly

$$p(\lambda|\sigma^2) \propto \exp(-c\mathrm{DF}),$$

where DF is computed as the trace of the smoothing matrix (and the domain of the distribution is truncated to obtain a proper prior). The posterior dependence on c is then compared in terms of an equivalence with classical model selection criteria that is similar to our range of AIC statistics.

Holmes correctly points out that one needs to be careful in blindly matching terms to establish this kind of correspondence, noting the difference between the maximized or profile likelihood

$$p(y|\widehat{\boldsymbol{\beta}}, \mathbf{t}) = \max_{\boldsymbol{\beta}} p(y|\boldsymbol{\beta}, \mathbf{t}) \tag{R4}$$

and the marginal likelihood

$$p(y|\mathbf{t}) = \int p(y|\boldsymbol{\beta}, \mathbf{t})\, p(\boldsymbol{\beta}|\mathbf{t})\, d\boldsymbol{\beta}. \tag{R5}$$

Naturally, we agree that the two are very different, and it was a comparison with model selection criteria similar to Holmes's that led us to the geometric prior on

model size. In the discussion from Kass and Wallstrom we are reminded that

$$\log p(y|\mathbf{t}) \approx \log p(y|\widehat{\boldsymbol{\beta}}, \mathbf{t}) - \frac{J(\mathbf{t})}{2} \log n.$$

Therefore, by adjusting the prior on model size, we could recover the general form of BIC and restore some of the protection against overfitting Holmes claims for the marginal likelihood. Inevitably, this kind of benchmarking against known criteria seems useful even for the purest Bayesian, and we will have more to say about it in our response to Kass and Wallstrom. Before we leave this point, we should mention that the latest minimum description length criteria for model selection developed by Barron, Rissanen and Yu (1998) and Hansen and Yu (2001) attempt to make up for the shortcomings of the maximized likelihood by a certain (re)normalization procedure.

In evaluating procedures like Logspline and Triogram, we have looked to theoretical results for guidance. While the results in Stone (1994), Hansen (1994) and Huang (1998, 2001) are asymptotic, they do provide an indication that a spline-based approach to ELMs is sensible. A similar theory for Bayesian versions of these methods is still emerging. For recent results, we refer the interested reader to articles by Zhao (1993, 1998), Barron, Schervish and Wasserman (1999) and Shen and Wasserman (2001). In the last paper we find rates of convergence associated with so-called sieve-priors, a setup that most closely mimics our own use of finite-dimensional approximating spaces. Shen and Wasserman point out that, while early work in this area produced a rich set of priors that led to consistent estimators, achieving the optimal rate is much more delicate.

KASS AND WALLSTROM

We first want to thank Kass and Wallstrom for a their contribution to the discussion of Denison, Mallick and Smith (1998a).

It seems that the work described by Kass and Wallstrom and our own motivation for exploring Bayesian procedures for splines share a common source. In our Rejoinder to Stone, Hansen, Kooperberg and Truong (1997), we presented a connection between the Bayesian method of Smith and Kohn (1996) and our own greedy schemes that attempted an approximate minimization of a generalized AIC statistic. We began with the straightforward observation that, by using Zellner's g-prior (Zellner, 1986) for the coefficients, the resulting posterior is mathematically equivalent (or nearly so) to AIC, BIC or anything "in between." In terms of (R2), this amounts to setting $V = c(B^t B)^{-1}$ and $a = d = 0$, and letting c vary. Ed George pointed us to this posterior calibration in a paper he crafted in 1997, which eventually appeared as George and Foster (2000). Of course the desire to relate Bayesian and frequentist methods is an old one, and explicitly expressing a posterior in terms of known model selection criteria goes back at least to Smith and Spiegelhalter (1980) (these authors speak of a "generalized AIC" criterion, of which BIC is just one example).

In their 1996 paper, Smith and Kohn were selecting the "best" knot configuration having the highest posterior probability. Given the equivalence between this posterior and our own generalized AIC statistics for ELM's, both procedures were attempting to optimize the same quantity. Having identified a good knot sequence, Smith and Kohn used the posterior mean of $\boldsymbol{\beta}$ to form a curve estimate. Under the g-prior version of (R2) this is simply $\frac{c}{c+1}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the OLS estimate [see (R3)]. Since the values of c used were at least 100, the posterior mean was essentially $\widehat{\boldsymbol{\beta}}$, and now the only difference between our greedy methods and the Bayesian approach of Smith and Kohn was the search scheme. This was our starting point.

To reduce computation, Smith and Kohn restricted their attention to at most 30 knots, placed at order statistics of the data. With such a small problem, it was possible to even use branch and bound techniques to search all possible combinations of candidate knots from within S-PLUS. Since that point, many other Bayesian spline-based estimators have appeared. Building on examples of Green (1995), Denison, Mallick and Smith (1998a) treated the knot sequence more like our own greedy schemes, in that each of the data points was a candidate knot. BARS (Bayesian regression splines), proposed by DiMatteo, Genovese and Kass (2001), is a hybrid that borrows the prior structure of Smith and Kohn, but adapts Green's sampler. In writing our paper, we were naturally aware of the shortcomings of DMS, although our insight came not from formal Bayesian thinking but by a simple comparison of the "objective function" represented by their posterior. Without a prior on the coefficients $\boldsymbol{\beta}$, the posterior did not provide enough of a penalty for large models. We attempted to correct this with the geometric prior on model sizes, yielding a posterior that agrees with BIC. In so doing, our penalty on dimension and that for BARS are the same.

Based on our experience with "free-knot splines," we also considered simulations that paired the geometric

prior along with a proper prior on $\boldsymbol{\beta}$, imposing an extra penalty on large models. Recall when working with free-knot splines, we treat the sequence $\mathbf{t}$ not as a discrete index for separate models, but rather as a continuous vector of parameters. Kooperberg and Stone (2002) used free-knot splines for density estimation and argued that the most appropriate form of AIC was

$$\log p(y|\widehat{\boldsymbol{\beta}}, \mathbf{t}) - \frac{J+K}{2}a,$$

where again J is the dimension of the spline space and K is the number of knots in $\mathbf{t}$. In short, an extra penalty on dimension was necessary to produce reasonable point estimates. No matter what principle is applied to determine the penalty, it is clear that the tradeoffs between sample size, model dimension and assumed smoothness of ϕ are complex. Since any theoretical guidance is likely to be asymptotic, new procedures should be tested on a wide variety of simulations. However, returning to the BARS paper, we find only three simulations, each of which are based on modestly sized problems with very similar characteristics. That is, one expects the "optimal" number of knots needed for approximating the three test functions is similar and the signal-to-noise ratios were all close (basically 3:1). Curiously, the simulations for BARS only involve 10 runs per test case, making the very strong claims by Kass somewhat premature. It would be good to see BARS run under many more conditions.

Aside from the posterior calibration to BIC, BARS differs from DMS in two other ways. First, in designing the moves for BARS, DiMatteo et al. (2001) borrow a "key idea" of Zhou and Shen (2001), that more knots are needed in regions where a curve changes rapidly. This is nothing more than the definition of a spline space with repeated knots (Schumaker, 1993); that is, as one knot approaches another, the splines lose a derivative. In a cubic spline space, two coalescent knots allow the fit to break in the first derivative, and three coalescent knots produce a break in the function itself. This behavior has guided the design of our own ELM procedures. In regression, if we have reason to believe the underlying curve has sharp features, we might consider relaxing the separation condition in the definition of an allowable space. However, in most cases, for reasonable sample sizes, this condition is not restrictive and the final curves can track strong features when they exist. We should add that, in BARS, there is no encouragement in the prior itself for nearby (or coalescent) knots. The "locality property" is purely a function of the proposal.

Aside from the locality property, there is another big difference between BARS and DMS. DiMatteo et al. (2001) use a natural spline basis for BARS, while DMS uses ordinary B-splines. While both are smooth, piecewise polynomials, the natural splines are forced to blend into linear functions outside the interval $[t_1, t_K]$. It is well known that this reduced space improves the variance at the ends of the data dramatically (we know that one of the referees of DMS encouraged the authors of that paper to use a natural spline basis). Kass and Wallstrom state that BARS performs better than a version of DMS that has been "corrected" to agree with BIC. Unfortunately, it is not clear whether the knot proposal distribution or the natural spline basis is responsible for this improvement. DiMatteo et al. (2001) have only one panel of plots comparing the variants of DMS and BARS, and this consists of a single estimated curve per method taken from one simulation setup (the test function having a very sharp break in the middle of the domain).

We should add that working with natural splines is a bit trickier than regular B-splines because of the boundary conditions. Breiman (1990) proposes a constrained least squares fit to enforce linearity in the tails during knot deletion. Luo and Wahba (1997) describe a stepwise algorithm for natural splines that would allow BARS to take advantage of one-degree-of-freedom alterations as we have done for Logspline and Triogram. For natural splines that are linear outside the interval [0, 1], Luo and Wahba use a basis of the form

(R6) $\quad \phi_1(x),\ \phi_2(x)$ and $R(x, t_k), \quad k = 1, \ldots, K,$

where $\phi_1(x) = 1$, $\phi_2(x) = k_1(x)$ and

(R7) $\quad R(x, x') = k_2(x)k_2(x') - k_4(|x - x'|).$

The functions k_1, k_2 and k_4 are constant multiples of Bernoulli polynomials and are given by

$$k_1(x) = x - 1/2, \quad k_2(x) = \big(k_1^2(x) - 1/12\big)/2$$

and

$$k_4(x) = \big(k_1^4(x) - k_1^2(x)/2 + 7/240\big)/24.$$

Unlike the truncated power basis, the so-called kernel functions $R(\cdot, \cdot)$ are not "one-sided" but have global support. They do, however, share the property that each candidate knot t_l corresponds to a single function $R(\cdot, t_l)$. As it stands, it is not clear how BARS is

implemented and whether or not this kind of shortcut has been used.

Given (R3), taking $V = c(X^tX)^{-1}$ in the regression context makes things much easier computationally because the posterior mean for $\boldsymbol{\beta}$ given a single space G is a scale version of the OLS estimate $\frac{c}{1+c}\widehat{\boldsymbol{\beta}}$. What is troubling, however, is that the implied shrinkage here is toward zero and depends on the basis (R1) used. For moderate values of c, we expect to see odd smoothing behavior which will become even more problematic when we start averaging fits. This does not seem appropriate to us, so for triogram regression we chose a smoothing prior with a null space consisting of planar functions. (The classical cubic smoothing splines also reduce to linear functions.) Unfortunately, our experiments with triograms are somewhat inconclusive because any kind of reasonable averaging can improve the piecewise linear fits if the underlying function is smooth (there are bigger effects to overcome before the behavior at the peaks is a concern). Kass and Wallstrom also mention the idea of asigning a prior to λ or setting it in an empirical-Bayes way. The latter approach is followed in George and Foster (2000) for $V = c(X^tX)^{-1}$. For a general prior covariance matrix, a simple iteration is required to select c in this way. Details can be found in Hansen and Yu (2001).

When considering normal priors on $\boldsymbol{\beta}$ for density estimation, the story is a bit clearer. First, taking a normal prior with covariance proportional to $(X^tX)^{-1}$ does not represent a reduction in computation. Shrinking toward zero in this context produces a uniform density. Despite this seemingly bad property, we investigated the use of this prior and found that it smooths away peaks (even when $\lambda = 1/n$, as proposed by Kass and Wallstrom), and it makes methods (v) and (viii) look very much like (vii). Unlike triograms which generically benefit from averaging, the proper smoothing can have a big effect on the performance in Logspline.

In attempting to recast our results for Logspline, Kass and Wallstrom fail to appreciate the fact that density estimation is a very different problem from (generalized) regression. As mentioned above, the closer you position knots, the greater the (potential) discontinuity in the fitted curve. In density estimation, this extra flexibility can drive the likelihood to infinity if knots are placed too close to each other relative to the scale of the data, an artifact that does not reflect the suitability of a model with nearby knots. When this happens, the search procedures underlying both the stepwise and our MCMC versions of Logspline can get stuck in local solutions. Therefore, in each case we require knots to be at least a few (usually three) data points apart. Keep in mind that, for density estimation, regions with sharp features (peaks) contain many more data points, and hence many more candidate knots; there is a natural coupling between structures in the function we are estimating and the spread of the data. In short, the restriction on knot placement does not hinder Logspline's ability to track strong features. In fact, under this restriction, we can ensure that knots proposed fairly close together are much more likely to be retained when they are near a peak, as the likelihood will increase much more with the addition of such knots (see Stone et al., 1997, Figure 1). We have experimented with many different knot location schemes, but for Logspline proposing too many close knots actually reduces the ability of the MCMC algorithm to explore the complete distribution, as the chain gets stuck more easily. On the other hand, with careful knot location, the convergence of MCMC chains in Logspline is very good. We carried out a substantial amount of additional chains, suggesting that the Logspline chains mix well. See also the discussion of Lindstrom, suggesting that a chain of length 500 is almost as good as a chain of length of 5,000. (For single calculations, rather than a large simulation study, we would advise somewhat longer chains too.) Finally, we should add that the situation for (univariate) regression is different because the likelihood is well behaved as knots coalesce providing we use a sensible basis (and do not add so many knots that we lose identifiability of the model). Still, it is not clear if the locality heuristic used in DiMatteo et al. (2001) actually performs better or if the potential improvements claimed by Kass and Wallstrom are due to their implementation of the reversible jump algorithm (as compared to that of DMS), their use of the natural spline basis or their selection test cases (small problems with common noise levels).

We mostly agree with Kass and Wallstrom's interpretation of our results; the label attached to (iii), whether it be Bayesian or approximately Bayesian or quasi-Bayesian, is of little concern for us. As suggested method (vii) looks like smoothing splines; judging by the MISE results, the method proposed by Wahba, Lin and Leng in their discussion behaves most like this approach. We believe that the "perplexing anomalies" cited by Kass and Wallstrom are primarily caused by a lack of appreciation for Logspline and the stepwise

algorithm, which is really quite good. When we compare approaches (iii) and (viii) we note that for small sample sizes (iii) tends to overfit, so that the prior on $\boldsymbol{\beta}$ [which makes it (viii)] helps, and without that prior the overfitting for (iii) hurts relative to the greedy algorithm, which overfits much less, but for large sample sizes there is enough variability in (iii) to compensate for the overfitting. Also, we actually do carry out annealing for method (ii). Details were omitted when we were asked to reduce the length of the paper during revisions.

We agree with Kass and Wallstrom that one of the interesting features of a Bayesian analysis is the assessment of uncertainty. While these Bayesian schemes seem to give sensible point estimates, we do not hold out much hope for confidence intervals. The BARS paper, for example, does not present any results with simulated data on coverage, but reserves plots with intervals for real-data problems (assessing their quality subjectively). In Kooperberg and Stone (2002) we examine Bayesian confidence intervals for the income data for versions (v) and (vi) based on an MCMC run of length 100,000. Based on comparisons with bootstrapping procedures we concluded there that these intervals were considerably narrower than frequentist confidence intervals; that is, calibrating the posterior to achieve a reasonable point estimate yielded optimistic intervals. This really emphasizes the points made by Lindstrom in her discussion that to obtain "credible" Bayesian confidence intervals the priors need to be selected much more carefully. We would challenge the Bayesian function estimators to produce a few confidence intervals in their papers. It is very hard to find any that come from simulations where we can judge their usefulness. Wang and Wahba (1995) compare Bayesian intervals for smoothing splines with those for bootstrapping. A similar study for the model-averaged fits would be useful.

The goal of our paper was to compare Bayesian and stepwise methods, and to compare how well they are doing. Our goal was not to advocate one approach or another. From Kass and Wallstrom's discussion of BARS we get the impression that it is a useful procedure. We would like to see a comparison of their procedure with the one by Smith and Kohn (1996), which sounds very similar. However, we feel that the conclusion that "BARS appears to be the most powerful available method for spline-based curve-fitting in ELMs" is somewhat premature. In particular, we would like to see the performance of BARS:

- illustrated on a suite of very different test functions with different sample sizes and different signal-to-noise ratios;
- extended to demanding nonlinear applications, such as density estimation and survival analysis;
- tested in high dimensional applications with thousands of cases;
- compared to greedy (stepwise) algorithms in terms of computing time.

KOENKER AND MIZERA

The L-1 methods made popular by Koenker and his co-authors have filtered into extremely effective spline methods for estimating median fits (Koenker, Ng and Portnoy, 1994; He, Ng and Portnoy, 1998). Koenker and Mizera base their version of triograms on a Delaunay triangulation, producing an underlying linear space G from which a single model is selected via a roughness-penalized L-1 error criterion. In Rippa (1990) we find an interesting characterization of the Delaunay triangulation based on roughness that we think is worth mentioning. Suppose we have data points $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$. We want to interpolate the values Y_i at $\mathbf{x}_i$ and will do so by creating a triangulation with $\mathbf{x}_i$ as the vertices. There are many many ways to do this. For each one, we will measure its roughness via the (Sobolev seminorm)

$$\sum \int_\delta \left[\left(\frac{\partial g}{\partial x_1} \right)^2 + \left(\frac{\partial g}{\partial x_2} \right)^2 \right] dx_1\, dx_2,$$

where the sum is over all triangles $\delta \in \triangle$ and $\mathbf{x} = (x_1, x_2)$. Then, Rippa (1990) shows that the Delaunay triangulation of a set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is a minimal roughness triangulation. Therefore, with at least one sensible smoothness constraint, the Delaunay triangulation is the appropriate thing to do. However, it is likely that, for other smoothness measures, the Delaunay is not optimal and that other structures are more appropriate. Rippa (1992) explores a similar line of reasoning and uses an edge-swap algorithm to find promising triangulations. It is known that we can generate all the triangulations of a given set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ by the first move type in Figure 4 of our original paper. In developing the greedy version of triogram and its Bayesian counterpart, however, we felt that there was much to be gained by taking smaller triangulations that were better adapted to the underlying function being estimated. Hence, it was important to develop a suite of moves that allowed us to step through models of different sizes.

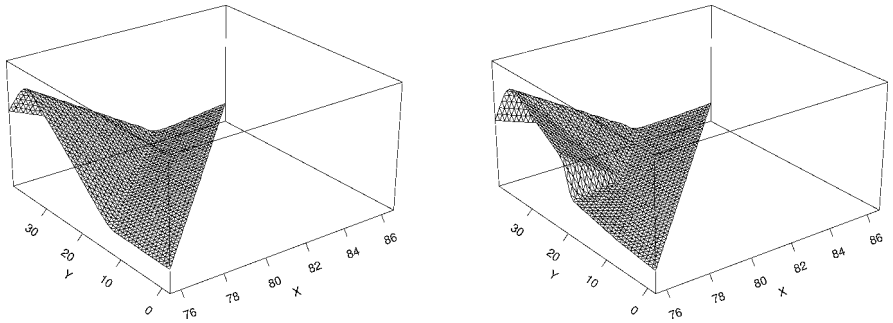

FIG. R1. *The volt data from Cleveland and Fuentes* (1996): (*left*) *the fit from prior specification* (v); (*right*) *the fit for* (iii).

For our prior on $\boldsymbol{\beta}$, we borrowed the penalty from Koenker and Mizera, first communicated to us by Koenker. We were pleased with the form because it built both on their work and on the approach of Nicholls (1998), who considered piecewise constant surfaces over adaptively chosen triangulations. We tried a variety of other penalties, including those suggested in Dyn, Levin and Rippa (1990a, b). Each have a null space of planar functions. In simulations with this form using both smooth functions and simple examples exhibiting sharp edges, it seems to perform well. As a further test, we consider a dataset known to exhibit a simple hinge, not aligned with either coordinate axis (Cleveland and Fuentes, 1996). While we have seen that a certain amount of smoothing is possible with the Bayesian estimator when the underlying target function is smooth, in this case, we hope that the sampler will spend time in very simple, "nearby" ridge models. This would allow the nongreedy schemes to still capture ridges effectively. In Figure R1 we present two surfaces, one from simulation setup (v) and one from (iii). A careful analysis by Cleveland and Fuentes suggested that the best fit was a hinged pair of planes, very similar to the fit in the left-hand panel corresponding to (v). It is clear from this figure that Poisson prior (iii) yields a chain that spends too much time in overly complex models and the fit is badly degraded. The surface obtained by prior specification (v), on the other hand, is an improvement over the greedy scheme. While it is difficult to tell from the perspective plot, the ridge or central hinge more closely follows the line found by Cleveland and Fuentes (1996).

LINDSTROM

Lindstrom has a considerable background in working with free-knot splines. She has developed an extremely attractive penalized appraoch to fitting such models (Lindstrom, 1999) and has a long history of working with functional data in medical applications (see, e.g., Lindstrom, 1995). We appreciate her extensive comments on the computational aspects of our Bayesian schemes. In our implementation of the samplers in this paper, we used the same number of iterations for each of the model selection schemes. The greedy scheme naturally takes far fewer "iterations" or moves. For the income data, on our current machine (which is not the machine on which the original calculations were carried out), the greedy Logspline algorithm takes about 0.5 second of CPU time, while a simulated annealing or MCMC run of 5,500 iterations takes about 200 seconds. The differences between the various versions of the sampling methods do not have a major influence on the CPU time.

Lindstrom asks us to compare this to stepwise algorithms with random restarts. This comparison appears most relevant to the greedy version (i) and the simulated annealing version (ii). For the income data, version (i) has a BIC value of 161,936.4, while the simulated annealing version, using a chain length of 5,500 and the best greedy solution as initial knots, has a BIC value of 161,918.3. Table R2 contains some alternative approaches using (several) shorter chains, or the stepwise algorithm with several random restarts. (For the random restarts, we positioned 7 knots at randomly selected datapoints, followed by the stepwise addition and deletion algorithm of Stone et al., 1997, allowing a maximum of 15 knots.) The bottom line of this seems to be that, at least for this example, random restarts are just as good as (but not better than) simulated annealing using the same amount of CPU time. It remains open what works better if we want to get close to the "best" solution. However, for all purposes the differences in BIC are not very large.

We fully agree with Lindstrom's remark that when we want to use Bayesian methods to carry out inference, the choice of priors becomes much more critical

TABLE R2
CPU and BIC results for various alternative stochastic optimization methods: for the last three methods the BIC value is the average over five repetitions of the complete procedure, and the SD is the corresponding standard deviation

Method	Number of steps	Number of seeds or restarts	BIC	SD	Seconds CPU time
Stepwise (version (i))	1	1	161,936.4	—	0.4
Simulated annealing (ii)	5,500	1	161,918.3	—	200
Shorter annealing chains	500	1	161,922.1	2.5	20
Combining real short chains	100	5	161,921.7	1.9	20
Stepwise with restart	1	50	161,921.8	0.8	20

than when we primarily want to get a point estimate. See our remarks about the contributions by Kass and Wallstrom and by Chipman, George and McCulloch and also the example in Kooperberg and Stone (2002). Finally, we agree with Lindstrom that radial basis functions are a sensible approach to multivariate modeling. In this case, we might consider an expansion of the form

$$g(x) = \sum \beta_i R(\mathbf{x}, \mathbf{v}_i) \tag{R8}$$

for a fixed "kernel function" R and a set of "knots" $\mathbf{v}_1, \ldots, \mathbf{v}_K$. Holmes and Mallick (1998) describe a Bayesian setup that also uses reversible jump Markov chain Monte Carlo, but selects how many knots and where they should be placed. The basis (R7) is an example of this kind of expansion for univariate functions. Having said that, we still feel that the Triogram basis has its role in multivariate function estimation. It is better able to capture sharp features in the data and has the added advantage that the entire procedure is invariant to affine transformations. This property makes it ideal for many spatial applications.

RUPPERT

Ruppert is a pioneer in the area of function estimation, contributing kernel, local polynomial and now spline methods, and we are pleased to have such an informative contribution to the discussion. Ruppert presents a penalized formulation that mimics smoothing splines. The so-called P-spline approach originally put forward by Eilers and Marx (1996) was designed as a shortcut to smoothing splines and illustrated for generalized linear models. Ruppert's version of this technique is really nothing more than ridge regression applied to a special truncated power basis. As such, it inherits a certain degree of familiarity and represents an accessible smoothing method for people with a basic introductory regression course. In comparing Ruppert's method and that of Eilers and Marx we do find one important difference: the basis. Ruppert makes use of the truncated power basis and (for cubic splines) he penalizes jumps in the second derivative at each knot. That is, penalties of the form

$$\sum_{k=1}^{K} |\beta_{p+k}|^q, \tag{R9}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{K+p})$ is associated with the truncated power basis $\{1, x, \ldots, x^p, (x - t_1)_+^p, \ldots, (x - t_K)_+^p\}$. Eilers and Marx choose the numerically stable B-spline basis and derive a quadratic penalty by taking differences of coefficients associated with adjacent basis elements (this shortcut is a rough approximation to the derivative-based penalties we applied in our paper, although we are not convinced of its reasonableness). Ruppert's use of values of q in (R9) other than 2 is interesting. As he mentions, this amounts to the lasso of Tibshirani. In Hastie, Tibshirani and Friedman (2001), we find a close connection between forward stagewise fitting like boosting and the lasso. These authors suggest that boosting is like combining all possible models (R8) with a lasso penalty.

Finally, we comment on a couple of the P-spline improvements mentioned by Ruppert. First, reducing the number of knots to ease the computational burden of traditional cubic smoothing splines was examined in detail by O'Sullivan (1988) and is implemented in S-PLUS. To help improve the spatial adaptivity, Ruppert suggests two approaches; the first involves a variable penalty that is again modeled as a spline. A form of this was also suggested in Wahba (1995).

TABLE R3
Geometric average for the MISE ratios for the Logspline simulation study reported in Table 2

	Version									
	(i)	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**	**(vii)**	**(viii)**	**Wahba et al.**	**Bagging**
Geometric average	1.00	0.69	0.70	1.07	0.76	0.75	0.95	0.74	0.79	0.83

WAHBA, LIN AND LENG

Over the years Wahba and her co-workers have pioneered the use of smoothing splines in function estimation. Wahba's influence has been mentioned several times throughout this Rejoinder already. As for smoothing spline density estimation, we would have liked to see some of the plots of the estimates corresponding to their Table 1, as the numbers reported suggest that the proposed method best compares to a less extreme version of our approach (vii), which, as was demonstrated in our Table 3 and Figure 2 smoothes away details too much.

When we read Wahba, Lin and Leng's discussion, we realized that the average in Table 2 is not a good summary of the table, and that for a particular method the geometric average would be a fairer comparison. Those are given in Table R3. This new summary table, even more than the one in the main paper, shows that we cannot just judge performance by MISE: except for (iv) and (vii) we have no way to choose. Summaries like the number of peaks need to play a role in deciding which version to use. See also the first example in Kooperberg and Stone (1991).

The smoothing spline ANOVA models for multivariate density estimation look quite promising. We are looking forward to seeing actual estimates. In particular we wonder how efficient the algorithm is if many λ parameters have to be estimated simultaneously. Also, we wonder how one would choose the terms in the ANOVA decomposition. Methods like MARS do this automatically (during passes of addition and deletion), but it does not appear to be an easy calculation in the smoothing spline world. Perhaps it would be best to combine the greedy search with the penalized fit as was done in Luo and Wahba (1997) for saturated models.

ADDITIONAL REFERENCES

BARRON, A., RISSANEN, J. and YU, B. (1998). The minimum description length principle in coding and modeling. Information theory: 1948–1998. *IEEE Trans. Inform. Theory* **44** 2743–2760.

BARRON, A., SCHERVISH, M. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561.

BARTLETT, M. S. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika* **44** 533–534.

BERRY, S., CARROLL, R. and RUPPERT, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *J. Amer. Statist. Assoc.* **97** 160–169.

BREIMAN, L. (1993). Fitting additive models to regression data: Diagnostics and alternative views, computational statistics and data analysis. *Comput. Statist. Data Anal.* **15** 13–46.

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

BRUMBACK, B. and RICE, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Amer. Statist. Assoc.* **93** 961–994.

BRUMBACK, B., RUPPERT, D. and WAND, M. (1999). Comments on "Variable selection and function estimation in additive nonparametric regression using a data-based prior," by T. Shively, R. Kohn and S. Wood. *J. Amer. Statist. Assoc.* **94** 794–797.

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998a). Bayesian CART model search (with discussion). *J. Amer. Statist. Assoc.* **93** 935–960.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998b). Making sense of a forest of trees. In *Proceedings of the 30th Symposium on the Interface* (S. Weisberg, ed.) 84–92. Interface Foundation of North America, Fairfax Station, VA.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2002). Bayesian treed models. *Machine Learning* **48** 299–320.

CLEVELAND, W. S. and FUENTES, M. (1996). Multipanel conditioning: modeling data from designed experiments. Technical Memorandum 96-1, Bell Laboratories, Lucent Technologies.

COULL, B., RUPPERT, D. and WAND, M. (2001). Simple incorporation of interactions into additive models. *Biometrics* **57** 539–545.

DARROCH, J., LAURITZEN, S. and SPEED, T. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8** 522–539.

DENISON, D. G. T., HOLMES, C. C., MALLICK, B. K. and SMITH, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regressing*. Wiley, New York.

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998c). Bayesian MARS. *Statist. Comput.* **8** 337–346.

DIMATTEO, I., GENOVESE, C. R. and KASS, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88** 1055–1071.

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–102.

GENOVESE, C. R. (2000). A Bayesian time-course model for functional magnetic resonance imaging data (with discussion). *J. Amer. Statist. Assoc.* **95** 691–719.

GEORGE, E. and FOSTER, D. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747.

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

HANSEN, M. H. and YU, B. (2001). Model selection and the principle of minimum description length. *J. Amer. Statist. Assoc.* **96** 746–774.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

HE, X., NG, P. and PORTNOY, S. (1998). Bivariate quantile smoothing splines. *J. Roy. Statist. Soc. Ser. B* **60** 537–550.

HOLMES, C. C. and DENISON, D. G. T. (1999). Bayesian wavelet analysis with a model complexity prior. In *Bayesian Statistics* **6** (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 769–776. Oxford Univ. Press.

HOLMES, C. C. and DENISON, D. G. T. (2002). Classification with Bayesian MARS. *Machine Learning*. To appear.

HOLMES, C. C. and MALLICK, B. K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation* **10** 1217–1233.

JORDAN, M. (1998). *Learning in Graphical Models*. Kluwer, Dordrecht.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90** 928–934.

KOENKER, R., NG, P. and PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680.

KOHN, R., ANSLEY, C.F. and THARM, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Amer. Statist. Assoc.* **86** 1042–1050.

LIN, X., WAHBA, G., XIANG, D., GAO, F., KLEIN, R. and KLEIN, B. (2000). Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.* **28** 1570–1600.

LIN, X. and ZHANG, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *J. Roy. Statist. Soc. Ser. B* **61** 381–400.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika* **44** 187–192.

LINDSTROM, M. J. (1995). Self modeling with random scale and shift parameters and a free-knot spline shape function. *Statistics in Medicine* **14** 2009–2021.

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40.

LUO, Z. and WAHBA, G. (1997). Hybrid adaptive splines. *J. Amer. Statist. Assoc.* **92** 107–114.

NISHII, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.* **27** 392–403.

NOVAK, E. and RITTER, K. (1996). High dimensional integration of smooth functions over cubes. *Numer. Math.* **75** 79–97.

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

PAULER, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85** 13–27.

RIPPA, S. (1990). Minimal roughness property of the Delaunay triangulation. *Comput. Aided Geom. Design* **7** 489–497.

RIPPA, S. (1992). Adaptive approximation by piecewise linear polynomials on triangulations of subsets of scattered data. *SIAM J. Sci. Statist. Comput.* **13** 1123–1141.

RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* To appear.

RUPPERT, D. and CARROLL, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Aust. N. Z. J. Statist.* **42** 205–223.

SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714.

SHI, M. S., WEISS, R. E. and TAYLOR, J. M. G. (1996). An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *J. Roy. Statist. Soc. Ser. C* **45** 151–163.

SHIVELY, T. S., KOHN, R. and WOOD, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior (with discussion). *J. Amer. Statist. Assoc.* **94** 777–806.

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

SMITH, A. F. M. and SPIEGELHALTER, D. J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. Ser. B* **42** 213–220.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

TIBSHIRANI, R. and KNIGHT, K. (1999). Model search by bootstrap "bumping." *J. Comput. Graph. Statist.* **8** 671–686.

WAHBA, G. (1978a). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

WAHBA, G. (1978b). Interpolating surfaces: high order convergence rates and their associated designs, with applications to x-ray image reconstruction. Technical Report 523, Dept. Statistics, Univ. Wisconsin, Madison.

WAHBA, G. (1985). A comparison of GCV and GML for choosing the smoothing parameters in the generalized spline smoothing problem. *Ann. Statist.* **13** 1378–1402.

WAHBA, G. (1995). Discussion of "Wavelet shrinkage or asymptopia," by D. Donoho, I. Johnstone and G. Kerkyacharian. *J. Roy. Statist. Soc. Ser. B* **57** 360–361.

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.

WALLSTROM, G. L., KASS, R. E., MILLER, A., COHN, J. F. and FOX, N. A. (2002). Correction of ocular artifacts in the EEG using Bayesian adaptive regression splines. In *Case Studies in Bayesian Statistics* **6** (C. Gatsonis, R. E. Kass, A. Carriquiry,

A. Gelman, D. Higdon, D. Pauler and I. Verdinelli, eds.) Springer, New York. To appear.

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 Mathematical Multivariate Statistics*. Wiley, New York.

YU, Y. and RUPPERT, D. (2002). Penalized spline estimation for partially linear single index models. *J. Amer. Statist. Assoc.* **97**. To appear.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.) 233–243. North-Holland, Amsterdam.

ZHANG, D., LIN, X., RAZ, J. and SOWERS, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *J. Amer. Statist. Assoc.* **93** 710–719.

ZHAO, L. (1993). Frequentist and Bayesian aspects of some nonparametric estimation problems. Ph.D. dissertation, Cornell Univ.

ZHAO, L. (1998). A hierarchical Bayesian approach in nonparametric function estimation. Technical report, Dept. Statistics, Univ. Pennsylvania.

ZHOU, S. and SHEN, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *J. Amer. Statist. Assoc.* **96** 247–259.