Statistical Science
2000, Vol. 15, No. 3, 196–223

Bayesian Backfitting

Trevor Hastie and Robert Tibshirani

Abstract. We propose general procedures for posterior sampling from additive and generalized additive models. The procedure is a stochastic generalization of the well-known backfitting algorithm for fitting additive models. One chooses a linear operator ("smoother") for each predictor, and the algorithm requires only the application of the operator and its square root. The procedure is general and modular, and we describe its application to nonparametric, semiparametric and mixed models.

Key words and phrases: Additive models, backfitting, Bayes, Gibbs sampling, random effects, Metropolis–Hastings procedure.

1. INTRODUCTION

In this paper we propose general procedures for posterior sampling from additive and generalized additive models. The main idea evolves from the close relationship between the backfitting algorithm for fitting additive models, and the Gibbs sampler for drawing realizations from a posterior distribution.

As an example, Figure 1 shows the results of an additive model fit to four air pollution variables, in a dataset with 330 observations. The response variable is log ozone concentration. The fit is represented by the solid curves in each of the panels and was obtained using cubic smoothing splines within the popular "backfitting" algorithm. Also shown are posterior realizations from a Bayesian version of the additive model. The posterior realizations were produced from a stochastic version of backfitting, which we call "Bayesian backfitting." That is the central topic of this paper.

An additive model is a popular tool for modelling regression data. It expresses the response variable as a sum of (typically nonlinear) functions of the predictor variables. The backfitting procedure is a modular way of fitting an additive model. It cycles through the predictors, replacing each current function estimate by a curve derived from smoothing a partial residual on each predictor. The Bayesian backfitting procedure, introduced here, smooths the same partial residual and then adds appropriate noise to obtain a new realization of the current function. This is equivalent to Gibbs sampling for an appropriately defined Bayesian model.

In the important special case of an additive cubic smoothing spline model with n observations, we obtain an $O(n)$ algorithm for sampling from the posterior. This is not the first such procedure: Wong and Kohn (1985) derive an $O(n)$ algorithm using the state-space representation of splines; see also Carter and Kohn (1994). Denison, Mallick and Smith (1998) employ polynomial splines and backfitting in a Bayesian additive model. Our proposal has the advantage of being conceptually simple, modular and general; it can be used with a wide range of operators representing nonparametric smoothers, as well as linear fixed and random effects models.

We begin with an exposition of posterior sampling for cubic smoothing splines in Section 2 and then discuss our general proposal for additive models (Section 3) and give an example involving growth curves. In Section 4 we discuss approaches for estimation of the variance components (including Bayes, empirical Bayes, REML and ML), and how to choose appropriate priors. The relationship with bootstrap sampling is briefly explored in Section 5. Generalized additive models and the Metropolis–Hastings procedure are discussed in Section 6, and we end with a discussion, including a description of a new public domain S-plus function for Bayesian backfitting.

Trevor Hastie is Professor, Department of Statistics and Division of Biostatistics, Stanford University, Stanford, California 94305 (*e-mail:* trevor@stat.stanford.edu). *Robert Tibshirani is Professor, Departments of Health Research and Policy and Statistics, Stanford University, Stanford, California 94305* (*e-mail:* tibs@stat.stanford.edu).

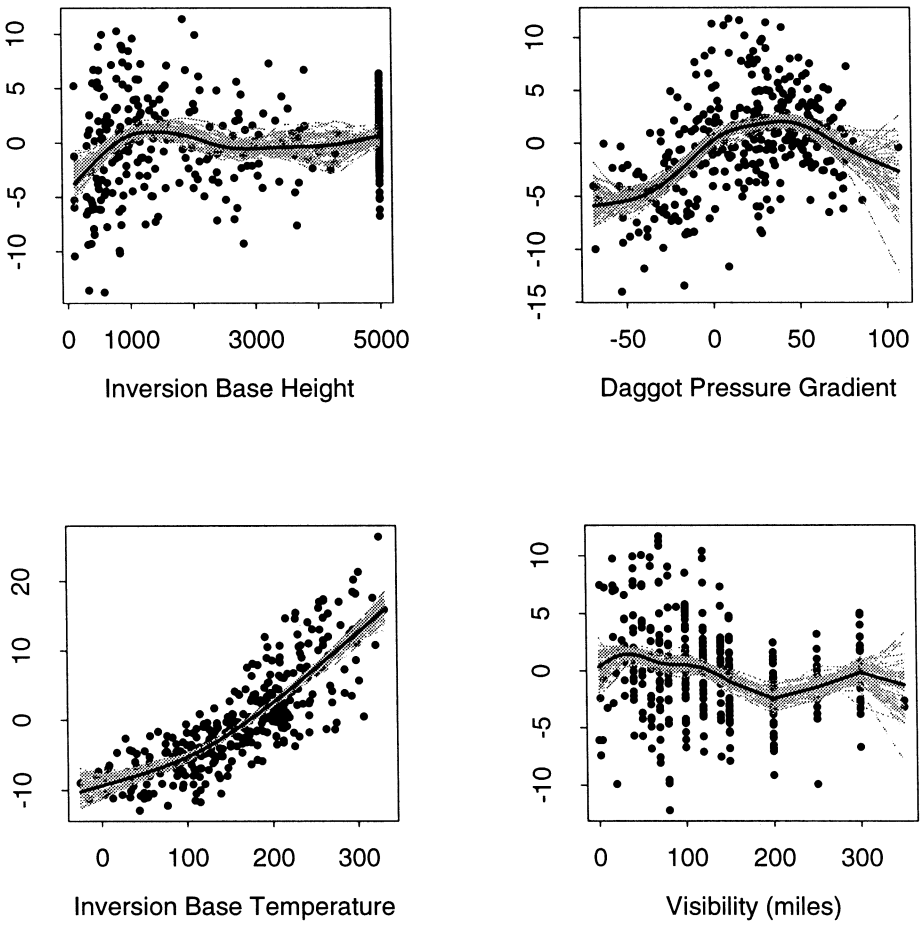

FIG. 1. *Fifty posterior realizations (grey curves) for an additive model fit to four air-pollution variables. The additive model fitted functions are shown with thick, dark curves. The points are partial residuals from the posterior means and give an idea of the spread of the data available for each posterior sample.*

2. POSTERIOR SAMPLING FOR A CUBIC SMOOTHING SPLINE

Consider a scatterplot smoothing problem with data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Here y_i are the response values and x_i are the inputs (predictors). We postulate a model

$$y_i = f(x_i) + \varepsilon_i; \qquad \varepsilon_i \sim N(0, \sigma^2). \tag{1}$$

The smoothing spline is a popular model for representing $f(x)$, and is usually derived as the minimizer of the penalized sum of squares criterion

$$J(f) = \sum_i (y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 \, dx \tag{2}$$

over all functions $f(x)$ such that the integral exists. The constant $\lambda \geq 0$ is a tuning parameter, with larger values resulting in smoother curves. The solution function $\hat{f}$ is a natural cubic spline, with knots at each of the unique values of x_i. This implies a representation

$$f(x) = \sum_{j=1}^{M} b_j(x)\theta_j, \tag{3}$$

where the $M \leq n$ basis functions b_j represent the linear space of such functions.

Letting $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))^T$, the fitted values at the n input values x_i can be written as

$$\hat{\mathbf{f}} = S(\lambda)\mathbf{y}. \tag{4}$$

Here $S(\lambda)$ is a symmetric $n \times n$ operator matrix, called the *smoother matrix*. It depends on the values x_i and the tuning parameter λ, but not on $\mathbf{y}$. We will sometimes write it simply as S. It is also possible to estimate λ in an adaptive (nonlinear) manner, depending on the response $\mathbf{y}$, but we do not consider that here. Most smoothers, and in particular smoothing splines, give a prediction at any values of x, not just the ones in the dataset, since they produce a function $\hat{f}$.

It is often convenient to parametrize the smoothing spline in terms of this fitted vector $\mathbf{f}$ rather than θ in (3). This is an equivalent representation (Green and Silverman, 1994), since $\mathbf{f} = B\theta$, where B is the full-rank $n \times M$ basis matrix evaluated at the n values of x_i. An advantage is that one obtains expressions that immediately suggest generalizations to other smoothing methods.

There is a Bayesian characterization of $\hat{\mathbf{f}}$. By choosing a particular partially improper Gaussian prior for $\mathbf{f}$,

$$\mathbf{f} \sim N(0, K^{-}\tau^2), \tag{5}$$

the resulting posterior distribution of $\mathbf{f}$ has the form

$$\mathbf{f}|\mathbf{y} \sim N(S(\lambda)\mathbf{y}, S(\lambda)\sigma^2) \tag{6}$$

with $\lambda = \sigma^2/\tau^2$. Hence the smoothing spline is the mean of the posterior distribution. Often it is convenient to parametrize $S(\lambda)$ using $df(\lambda) = \operatorname{tr} S(\lambda)$, the *effective degrees of freedom*, which is monotone in λ. We have assumed that σ^2, τ^2 and hence λ, are fixed.

Here and elsewhere, the notation K^{-} refers to a generalized inverse of a matrix K, with the understanding that an eigenvalue of zero for K gives an eigenvalue of $+\infty$ for K^{-}. In the case of smoothing splines and the parametrization $\mathbf{f}$, K computes the penalty in (2): $\int [f''(x)]^2\,dx = \mathbf{f}^T K \mathbf{f}$, and the zero eigenvectors correspond to linear functions of x. The prior therefore gives infinite variance to linear functions (is vague), and hence they are unrestricted. More details on K^{-} for splines are given in Appendix A, as well as Hastie and Tibshirani (1990). More generally, for symmetric smoother operators $S(\lambda)$ we can identify a prior covariance

$$K^{-} = \lambda[S(\lambda)^{-} - I]^{-}, \tag{7}$$

where S^{-} indicates a generalized matrix inverse. See Buja, Hastie and Tibshirani (1989) for more details.

In this paper our interest is not just the mean but the entire posterior distribution of $\mathbf{f}$ given in (6). Throughout the paper we use the notation $\mathbf{z} = (z_1, z_2, \ldots, z_n)^T$ to represent a vector of independent $N(0, 1)$ variates. Notice that (6) can be written as

$$\mathbf{f} = S\mathbf{y} + \sigma S^{1/2}\mathbf{z}, \tag{8}$$

where we have dropped the dependence on λ. Therefore we can generate a posterior realization of $\mathbf{f}$ by adding the noise $S^{1/2}\mathbf{z}$ to the fitted smoothing spline. The quantity $S^{1/2}\mathbf{z}$ can be generated efficiently with the same order of computations as $S\mathbf{y}$, typically $O(n)$. In Appendix A, we give two algorithms for this, one exclusively for smoothing splines and the other for general smoothing operators. Although (8) is derived for cubic smoothing splines, by analogy we can use it for any smoothing operator, even nonlinear smoothers. Once again, expression (8) can be used to generate posterior realizations at any arbitrary input values $t_1, t_2, \ldots, t_m$, including ones not in the dataset.

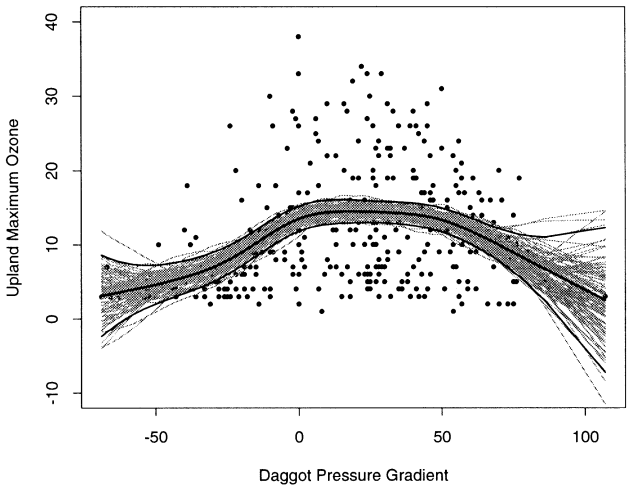

FIG. 2. *Los Angeles air pollution data: Upland Maximum Ozone vs Daggot Pressure Gradient. Shown* 100 *realizations from the posterior distribution* $f|\mathbf{y}$, *with the smoothing parameter fixed at* $df = 5$. *The smoothing-spline* (*posterior mean*) *is shown with a thick*, *dark curve*. *Included are the pointwise* 95% *posterior intervals, computed exactly*.

Figure 2 shows an example. The response variable represents ozone measurements on 330 days in the Los Angeles basin, and the predictor variable is the pressure gradient measured at the Daggot Airport. The figure shows 100 realizations of the posterior distribution, using a cubic smoothing spline with a fixed number ($df = 5$) degrees of freedom, and σ fixed at the unbiased estimate $\sigma^2 = \sum (y_i - \hat{y}_i)^2/(n - df)$. We used a "burn-in" period of 500 iterations. Convergence issues for Markov chain Monte Carlo methods are important, but there is insufficient space to address here. See, for example, Gelman, Stern and Rubin (1995) for a general discussion and references. The figure suggests that the variance of log ozone is not constant as a function of Daggot pressure gradient. An appropriate transformation of the response might help alleviate this, but we do not pursue that here.

The figure includes pointwise 95% posterior bands, which can in fact be computed exactly from the diagonal of $S\sigma^2$ [also in $O(n)$ operations]. While they show the shadow of the posterior distribution, they do not show the individual realizations. In Section 3.1 we make use of the individual realizations, and display the posterior distributions of

interesting functionals of them. Here the smoothing parameter λ is fixed at $df(\lambda) = 5$; in Section 4 we show how to incorporate priors for the smoothing parameters and σ^2.

Notice that adding noise $S\mathbf{z}\sigma$ in (8) would give posterior covariance $S^2\sigma$, which is not the same as $S\sigma^2$ since S is not idempotent. In fact, $S^2\sigma^2$ is the frequentist covariance of $S\mathbf{y}$, and $S\sigma^2 \geq S^2\sigma^2$: the posterior covariance exceeds the frequentist covariance because it incorporates prior uncertainty.

There is a version of result (8) for simple linear and multiple regression. Suppose

$$y_i = \beta x_i + \varepsilon_i; \qquad \varepsilon_i \sim N(0, \sigma^2), \quad \beta \sim N(0, \tau^2). \tag{9}$$

Letting $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, the posterior distribution of $\mathbf{x}^T\beta$ is

$$\mathbf{x}^T\beta|\mathbf{y} \sim N(H(\tau^2)\mathbf{y}, H(\tau^2)\sigma^2), \tag{10}$$

where

$$H(\tau^2) = \mathbf{x}(\mathbf{x}^T\mathbf{x} + I/\tau^2)^{-1}\mathbf{x}^T \tag{11}$$

with I being the $n \times n$ identity matrix. As for general smoothers we can write the posterior realizations as

$$\mathbf{x}^T\beta = H(\tau^2)\mathbf{y} + H(\tau^2)^{1/2}\mathbf{z}. \tag{12}$$

Taking $\tau \to \infty$ to represent prior ignorance, then $H(\tau^2) \to H(\infty) = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T = H$, the *hat matrix*. The operator H is an idempotent projection matrix and $H^{1/2} = H$, so that the posterior realizations can be written in the simpler form,

$$\mathbf{x}^T\beta = H\mathbf{y} + \sigma H\mathbf{z}. \tag{13}$$

3. ADDITIVE MODELS AND BAYESIAN BACKFITTING

We now consider the main topic of this paper, Bayesian posterior sampling for additive models. Our data consists of n observations of an outcome variable and p inputs: we write this as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ with $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Our model is

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i; \qquad \varepsilon_i \sim N(0, \sigma^2). \tag{14}$$

For identifiability between α and the f_j, $j > 0$ we assume

$$\sum_i f_j(x_{ij}) = 0 \qquad \forall j. \tag{15}$$

Suppose we define a cubic smoothing spline operator $S_j(\lambda_j)$ for each input variable j. Then the backfitting procedure for estimating the f_js uses iterations of the form

$$\mathbf{f}_j \leftarrow S_j\Big(\mathbf{y} - \bar{y}\mathbf{1} - \sum_{k \neq j} \mathbf{f}_k\Big) \tag{16}$$

for $j = 1, 2, \ldots, p, 1, 2, \ldots$. At each stage, the most current values of the functions $\mathbf{f}_k$ are used on the right-hand side, forming a partial residual that is smoothed as a function of x_j.

Rather than obtain estimates of the $\mathbf{f}_j$, which in Bayesian language means to compute the MAP estimates (here, the posterior means) we want to generate from their joint distribution. To achieve this we simply add the appropriate noise to the estimate at each backfitting step. For ease of notation define $\mathbf{f}_0 = \mathbf{1}\alpha$ and the associated operator $S_0 = \mathbf{1}\mathbf{1}^T/n$. Recall that the variance σ^2 is considered to be fixed. We define the *Bayesian backfitting algorithm* as follows.

ALGORITHM 3.1. *Bayesian backfitting.*

- Take initial values for $\mathbf{f}_j^0$, $j \geq 0$.
- Do for $t = 1, 2, 3, 4, \ldots$:
  - Do for $j = 0, 1, \ldots, p$:
    - Define the partial residual $\mathbf{r}_j^t = \mathbf{y} - \sum_{k<j} \mathbf{f}_k^t - \sum_{k>j} \mathbf{f}_k^{t-1}$.
    - Generate $\mathbf{z}_j^t \sim N(0, 1)$ and update $\mathbf{f}_j^t \leftarrow S_j\mathbf{r}_j^t + \sigma S_j^{1/2}\mathbf{z}_j^t$
- Until the joint distribution of $(\mathbf{f}_0^t, \mathbf{f}_1^t, \mathbf{f}_2^t, \ldots, \mathbf{f}_p^t)$ doesn't change.

The most appropriate starting values in the first step are the fitted curves from a standard additive model fit to the data. At the end of the procedure, the phrase "doesn't change" means that the procedure has converged to an appropriate stationary distribution. Convergence may be checked in practice in a number of ways; see, for example, the discussion in Gelman et al. (1995).

Bayesian backfitting is the Gibbs sampling procedure applied to additive models. Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) for general random variables $A_1, A_2, \ldots, A_p$ operates by successive sampling of each A_j conditional on the other A_k. At steady state a complete cycle delivers a sample from their joint distribution. The connection between Bayesian backfitting and Gibbs sampling is established by examining the conditional distribution of each $\mathbf{f}_j$. For cubic smoothing splines this connection is clear from the above development, but we give a more general result for a larger class of operators S_j.

Let S_j be any symmetric matrices with eigenvalues in $[0, 1]$. Define priors on the $\mathbf{f}_j$ by $\mathbf{f}_j \sim N(0, (S_j^- - I)^-\sigma^2)$. The matrices $(S_j^- - I)^-$ are symmetric with eigenvalues in $[0, +\infty]$ and hence are nonnegative definite. Consider σ^2 to be fixed (as well

as the *smoothing* parameter implicit in S_j). Then

$$\begin{aligned}\mathscr{L}(\mathbf{f}_j|\mathbf{y}, \mathbf{f}_k, k \neq j) &= \mathscr{L}\left(\mathbf{f}_j|\mathbf{y} - \sum_{k\neq j}\mathbf{f}_k, \{\mathbf{f}_k, k \neq j\}\right) \\ &= N\left(S_j\left(\mathbf{y} - \sum_{k\neq j}\mathbf{f}_k\right), S_j\sigma^2\right).\end{aligned} \tag{17}$$

Hence Bayesian backfitting corresponds to sampling from the conditional distribution of each $\mathbf{f}_j$. By the results in Tierney (1994), the joint distribution of the iterates $(\mathbf{f}_0, \mathbf{f}_1^t, \mathbf{f}_2^t, \ldots, \mathbf{f}_p^t)$ convergences to that of the true distribution of $(\mathbf{f}_0, \mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_p)|\mathbf{y}$. Furthermore, sample averages of functions of these quantities converge to their true values. This holds since the conditional densities are everywhere positive and hence the Markov chain is ergodic.

Figure 1 shows 50 realizations for the additive model fit to four air pollution variables. We fixed the degrees of freedom of the smoothers at $4.6, 4.8, 2.8, 6.0$, which are the values obtained from generalized cross-validation using an adaptive backfitting procedure (Hastie and Tibshirani, 1990, Chapter 9: the BRUTO procedure). From this information we can form posterior bands for the functions or carry out Bayesian inference for any other quantity of interest.

Recall the additive model constraints (15). These are necessary to ensure that the posterior distribution of α and the f_j is not singular. Practically speaking, it means that in the Bayesian backfitting algorithm, we have to center the fits after smoothing and generation. We discuss this and more sophisticated decorrelation procedures in Appendix A.

The standard backfitting algorithm is a general, modular method for fitting a wide variety of additive models. One chooses the smoother operator S_j for each input, and then backfits to estimate the joint model. The operator S_j can fit a flexible smooth, a linear regression (including dummy variable fits), an adaptive regression (e.g., wavelet smoother), and, more generally, any regression operator. Convergence has only been proved for a certain class of fixed, nonadaptive operators (Buja, Hastie and Tibshirani, 1989), such as smoothing splines, but the algorithm seems well behaved in general.

In the same way, we can choose an operator S_j for each input, and then paste them together as conditional sampling steps of the form

$$\mathbf{f}_j \leftarrow S_j\mathbf{r}_j + \sigma S_j^{1/2}\mathbf{z}_j \tag{18}$$

in the Bayesian backfitting algorithm (see Appendix A for a general procedure for computing $S^{1/2}\mathbf{z}$.) Given a single input $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$, we summarize some of the possibilities for choice of S_j:

1. *Smoothing Splines.* S_j computes a cubic smoothing spline. The conditional sampling step corresponds to the Gaussian process prior $\mathbf{f}_j \sim N(0, \tau^2 K_j^-)$. Both S_j and $S_j^{1/2}$ can be computed in $O(n)$ operations, the latter discussed in Appendix A.
2. *General nonparametric smoother.* S_j defines the smoothing operation, with implicit prior $\mathbf{f}_j \sim N(0, (S_j^- - I)^-\sigma^2)$. The operator $S_j^{1/2}$ is applied using Algorithm A.1 in the Appendix.
3. *Fixed linear effects.* $S_j = X_j(X_j^TX_j)^{-1}X_j^T$, where X_j is a matrix consisting of one or more predictors. This results from the model $\mathbf{f}_j = X_j\beta_j$ with $\beta_j \sim N(0, \tau^2 D)$ and D diagonal, and $\tau \to \infty$. Then $S_j^{1/2} = S_j$ and is easily applied. For the intercept term, for example, we simply obtain $\alpha \sim N(\text{ave}[y - \sum_1^p \mathbf{f}_j], \sigma^2/n)$.
4. *Random linear effects.* $S_j = X_j(X_j^TX_j + \sigma^2\Sigma^{-1})^{-1}X_j^T$. This results from $\mathbf{f}_j = X_j^T\beta_j$ with $\beta_j \sim N(0, \Sigma)$. Algorithms for implementing these random effects smoothers are very similar to those used in smoothing splines, which we discuss in Appendix A. We look more closely at a special case in the mixed effects example below.

3.1 Example: Growth Curves

The data in the top left panel of Figure 3 are measurements of spinal bone mineral density for a sample of 153 girls, as a function of age. There are between two and four measurements per girl, 471 in all. The consecutive data for each girl are connected in the plot. We see a great deal of between-girl variation, and a clear indication of the growth spurt around age 12. There is also a strong ethnic effect that is hidden in the variation of the growth fragments. A goal is to characterize the growth behavior and establish whether ethnic differences exist.

We consider the *mixed effects* model:

$$y_{ij} = f(t_{ij}) + \mathbf{x}_i^T\beta_E + V_i + \varepsilon_{ij}, \tag{19}$$

where:

- y_{ij} is the bone mineral density for girl i measured on occasion j, for $i = 1, \ldots, 153$, and $j = 1, \ldots, n_i$ with $n_i \in \{2, 3, 4\}$.
- $f(t_{ij})$ is the population growth curve as a function of the age measurement t_{ij} made on girl i on occasion j.
- β_E is an effect due to ethnic class; the data consist of white, black, Asian and Hispanic North American girls. $\mathbf{x}_i$ is any appropriate coding of contrasts to represent the 4-level factor.
- V_i is a random girl effect that allows a separate vertical shift in f.

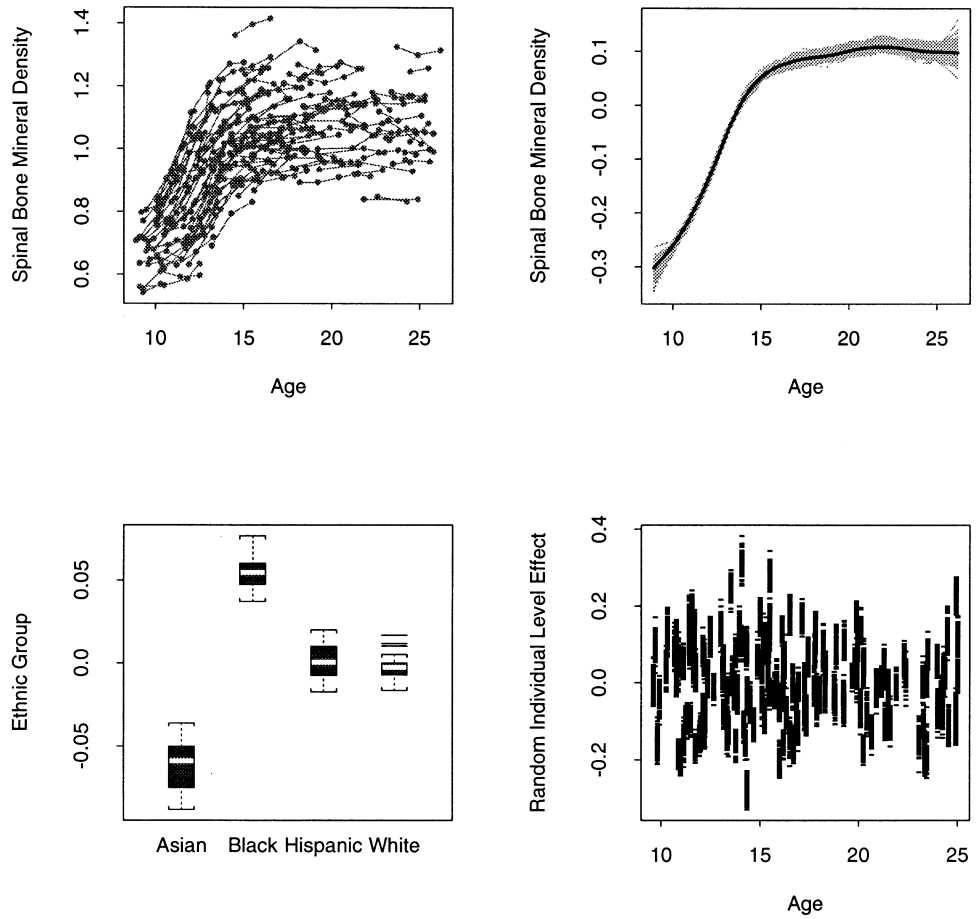

FIG. 3. *The top left panel contains* 471 *measurements of bone mineral density against age for* 153 *girls of different ethnic origin. Repeated measurements are connected. The remaining three panels show* 100 *posterior realizations from model* (19). *In the final panel, each random effect distribution is plotted against the mean age for that girl.*

- ε_{ij} is measurement and other variation, which we assume to be i.i.d.

A standard frequentist approach for fitting such data would be to treat f and the parameters in β_E as parametric fixed effects, and V_i as a random effect. One could model f by polynomials or more flexibly by splines with selected knots in age. Typically one assumes the $V_i \sim N(0, \sigma_V^2)$ independently across girls, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ independently across all measurements. Estimation of these mixed effects models is typically done by maximum likelihood (Laird and Ware, 1982), and focuses on:

- The parameters of the fixed effects and their standard errors,
- The variance components σ_V^2 and σ^2,
- The posterior mean or BLUP estimates of the V_i.

Here we take a Bayesian approach, and treat everything as random. We treat f as random, and use the smoothing spline process prior. We also treat the coefficients β_E as random, but with a flat prior. For the moment we assume the variance components are fixed ($df \approx 9$, $\sigma_V \approx 0.11$ and $\sigma \approx 0.03$), and focus on generating realizations from the joint posterior. In Section 4 we describe a variety of methods for estimating the variance components as well, including the REML method and the fully Bayesian procedure that was used here.

The set-up is tailor-made for the Bayesian backfitting procedure. The random effects have conditional posterior distributions

$$(20)\quad V_i|\mathbf{y}, f, \beta_E \sim N\left(\sum_{j=1}^{n_i} \frac{r_{ij}}{n_i + \lambda_V}, \frac{\sigma^2}{n_i + \lambda_V}\right),$$

where $r_{ij} = y_{ij} - f(t_{ij}) - \mathbf{x}_i^T \beta_E$ and $\lambda_V = \sigma^2/\sigma_V^2$.

The remaining three panels in Figure 3 show 100 realizations from the model. The 153 posterior realizations for the random effects are shown vertically in the last panel, centered at the average age for that girl. Figure 4 focuses on the posterior realiza-

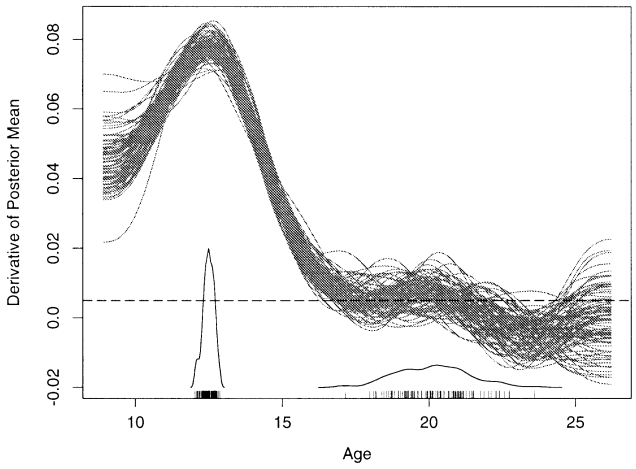

FIG. 4. *One hundred posterior realizations of the derivative of f. Included in the plot are the distributions of two functionals: the location of the maximum and the location of the point at which growth is less than 0.5% per year.*

tions for f. Since each realization is a natural cubic spline, we are easily able to produce the derivatives for each curve (in which the *natural* boundary conditions are evident). These are displayed, along with the posterior distributions of two functionals:

- The location of the maximum, which is the age at which the growth velocity is fastest. This distribution is fairly tightly concentrated at 12.5 years old.
- The location of the point at which bone growth increase levels off. We have used a threshold of 0.005, which corresponds to 0.5% per annum. This distribution is rather spread out; indeed, the derivative posterior is rather wiggly in this region.

4. ESTIMATING THE VARIANCE COMPONENTS

In the preceding development, the smoothing parameters or variances σ^2 and τ_j^2, $j = 0, 1, \ldots, p$, were considered fixed. In practice they have to be determined as well. There are several approaches:

- The full Bayesian approach, where we put priors on the variance components and estimate their posterior along with the functions;
- The empirical Bayes approach, that treats the variance components as parameters, which are estimated by maximum (marginal) likelihood.
- The frequentist approach, that treats everything as a smoother or regression fitting method, including the random effects operators. All the parameters are then estimated by cross-validation, GCV (Wahba, 1990), or related methods aimed at minimizing prediction error on future observations.

We give more details on the first two of these (in reverse order).

4.1 REML, ML and Empirical Bayes

Model (19) can be regarded as an hierarchical mixed effects model. The function f is random at level "0" (a single coefficient vector), while the V_i are random at level "1" (a coefficient per cluster). Mixed effects models are typically fit by maximum likelihood or REML (Laird and Ware, 1982), and the popular packages such as SAS and Splus have routines for fitting them. Maximum likelihood provides estimates of the variance components, τ^2, σ_V^2 and σ^2 in this case, the parameters of the fixed effects, and the BLUP or posterior mean estimates $E(f|\{y_{ij}\})$ and $E(V_i|\{y_{ij}\})$ of the random effects. Restricted maximum likelihood (REML) is a slight modification which takes into account the degrees of freedom used in estimating the fixed effects, when estimating the variance components (like the $n-1$ versus n correction in the sample variance.)

The empirical Bayes approach is to form the marginal likelihood by integrating out everything random and then estimating the remaining hyperparameters by maximum likelihood. It turns out, that if the "fixed effects" are given a flat prior, then empirical Bayes is equivalent to REML (Laird and Ware, 1982)

Treating smoothing splines as random effects and estimating τ^2 by REML is not a new idea (Speed, 1991), also known as GML in the spline literature (Wahba, 1990). Lin and Zhang (1997) in fact use REML in this way to estimate the smoothing parameters for additive spline models. Their approach is to represent the functions as $\mathbf{f}_j = P_j\theta_j$ with $\theta_j \sim N(0, \tau_j^2 I)$, and treat the P_j as a block of regression variables with random coefficients θ_j. The number of columns in P_j is $m_j - 2$, where m_j is the number of *unique* elements of x_j. In general their algorithm is $O(\min(\sum_j m_j, n)^3)$ computations, and so defeats our purposes here of efficiency. A promising alternative is to approximate $P_j\theta_j$ by $P_j^*\theta_j^*$, where P_j^* has a fixed number (10 or 15) columns, for the purpose of estimating the variance components efficiently. Approximations of this kind, based on the leading eigenvectors of K^-, are developed in Hastie (1995) (but are put to different uses there).

4.2 Priors for the Variance Components

A more mainstream Bayesian approach would be to provide priors for the variance components and integrate. Wong and Kohn (1996) suggest the fol-

lowing priors:

$$p(\tau_j^2) \sim \frac{1}{\tau_j^2}\exp(-\rho_j/\tau_j^2) \text{ with } \rho_j = 10^{-10}, \tag{21}$$

$$p(\sigma^2) \sim \frac{1}{\sigma^2}. \tag{22}$$

These priors are almost indistinguishable. The prior for σ^2 makes the prior for $\log(\sigma^2)$ flat. The prior for τ_j^2 is almost flat and still improper, and we give some insight into the additional term involving ρ_j later in this section. Both these priors are conjugate for the Gaussian distribution and lead to inverse gamma posterior distributions. More generally, one can use proper inverse gamma priors

$$p(\sigma^2) \sim \left(\frac{1}{\sigma^2}\right)^{r+1}\exp(-\rho/\sigma^2), \qquad r > 0 \tag{23}$$

for which both the above are degenerate special cases.

Since τ_j^2 enters the model only through $\mathbf{f}_j$, the corresponding conditional distributions are

$$\begin{aligned} &p(\tau_j^2|\mathbf{y}, \sigma^2, \{\mathbf{f}_j, j = 0, 1, \ldots, p\}) \\ &\quad = p(\tau_j^2|\mathbf{f}_j) \sim (\tau_j^2)^{-(\frac{n}{2}+r+1)} \\ &\qquad \cdot \exp\left(-\frac{\frac{1}{2}\mathbf{f}_j^T K_j \mathbf{f}_j + \rho_j}{\tau_j^2}\right). \end{aligned} \tag{24}$$

This is an inverse gamma distribution $IG(n/2 + r, \frac{1}{2}\mathbf{f}_j^T K_j \mathbf{f}_j + \rho_j)$.

Similarly the conditional distribution of σ^2 is $IG(n/2 + r, \frac{1}{2}||\mathbf{e}||^2 + \rho)$ where $\mathbf{e} = \mathbf{y} - \sum_j \mathbf{f}_j$. To generate from the full posterior distribution, we include conditional sampling steps for the τ_j^2 and σ^2 in the Bayesian backfitting algorithm. Theoretical convergence of the procedure to stationarity is unaffected. Note that calculation in (24) of $\mathbf{f}_j^T K \mathbf{f}_j \propto \mathbf{f}_j^T(S_j^- - I)\mathbf{f}_j = \mathbf{f}_j^T S_j^- \mathbf{f}_j - \mathbf{f}_j^T \mathbf{f}_j$ requires no new computation besides inner products, since $\mathbf{f}_j^T S_j^- \mathbf{f}_j = \mathbf{r}_j^T S_j \mathbf{r}_j + 2\sigma \mathbf{r}_j^T S_j^{1/2}\mathbf{z} + \sigma^2 \mathbf{z}^T\mathbf{z}$, and $S_j\mathbf{r}_j$ and $\sigma S_j^{1/2}\mathbf{z}$ are already available.

The posterior realizations of the air-pollution functions do not look any different from those in Figure 1, so we do not repeat them here. The realizations produced for the bone data in Figures 3 and 4 were obtained in the manner just described.

Figure 5 shows the posterior distributions of the degrees of freedom df_j from Bayesian backfitting, both for the air pollution data and the bone-growth data. The degrees of freedom are a one-to-one function of $\lambda_j = \sigma^2/\tau_j^2$: $df(\lambda) = \operatorname{tr} S(\lambda)$. Estimated optimal values from generalized cross-validation (GCV) are indicated by horizontal broken lines for the air pollution data. Compared to GCV, the fully Bayesian procedure applies slightly less smoothing (more degrees of freedom) to inversion base temperature, but the fit does not change much.

For the lower two panels, the horizontal lines indicate the df chosen by REML, which appear to match these posterior realizations more closely. Since the between-girl variation, σ_V^2, is very large compared to the within-girl σ^2, not much shrinking is done from the maximum of 153 df for the V effect.

For the remainder of this section, we investigate the priors (21) and (22) in terms of the prior degrees of freedom and develop a more general framework for any smoother. One might ask what the implicit prior is for df, given particular priors on σ^2 and τ^2. Holmes and Mallick (1997) and Hodges and Sargent (1998) similarly investigate priors based on degrees of freedom. Figure 6 shows the implied prior distributions for df for two commonly used priors on the variance components, obtained by simulation. The X-values are taken to be 50 uniformly spaced observations on [0, 1]. For any given pair σ^2 and τ^2, we compute $\lambda = \sigma^2/\tau^2$, and $df(\lambda) = \operatorname{tr} S(\lambda)$, where $S(\lambda)$ is the smoothing spline operator applied to the 50 values of X. Notice that if $K = UDU^T$ is the eigen-decomposition of K in $S(\lambda) = (I+\lambda K)^{-1}$, then $df(\lambda) = \sum_{j=1}^n 1/(1 + d_j\lambda)$ and can be computed efficiently for different values of λ.

- In the left panel, we have used *flat improper* priors (22) for both $\log\sigma^2 \sim 1$, $\log\tau^2 \sim 1$. The prior for df puts mass $\frac{1}{2}$ on 2 and 50, the linear and interpolating fits! This is easily proved (see appendix), and has some negative consequences on the Gibbs sampler. It implies that these two states are absorbing, and hence the real posterior would end up in one of these states as well.
- Using the prior (21) for both $p(\sigma_j^2) \sim (1/\sigma_j^2)\exp(-\rho_j/\sigma_j^2)$ and likewise for τ^2, one sees exactly the same behavior. This prior is still improper, but the presence of $\rho = 10^{-10}$ appears to prevent the absorptions at the two extreme states.
- In the right panel, we use $IG(0.01, 0.01)$, considered to be reasonably flat proper priors in the MCMC literature (Spiegelhalter, Best, Gilks and Inskip, 1996). The histogram was obtained by simulating 10,000 values from these priors. It exhibits very similar behavior to the first, although it appears there is support everywhere. A Gibbs sampler, starting at some value of df in this flat interior immediately concentrates the posterior away from the boundaries and appears not to run into trouble.

In all cases the strong U-shape is troublesome and does not seem very sensible as a prior for df. After some experimentation, we found that priors $\sigma^2 \sim$

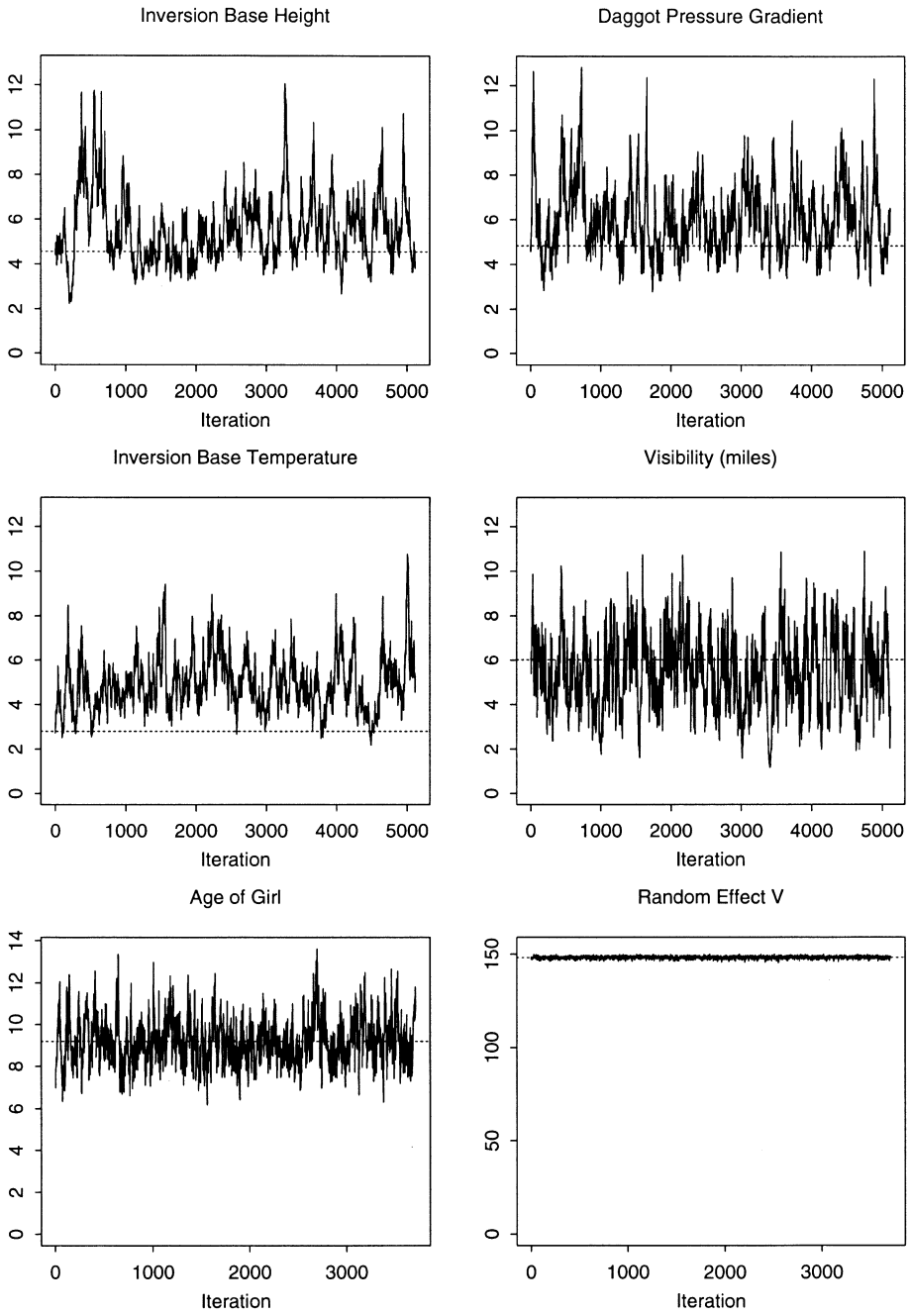

FIG. 5. *Top four panels*: 5000 *posterior realizations of the df for each predictor for the air pollution data. Estimated optimal values from generalized cross-validation are indicated by horizontal broken lines. Lower two panels*: 3700 *posterior realizations of df for the age curve and the random effect V for the bone growth data.*

$IG(2, 0.01)$ and $\tau^2 \sim IG(0.5, 0.01)$ gave a reasonable prior for df, without the right spike (Figure 7).

Here is an alternative strategy that one might use for prior selection. One could use the usual prior for σ^2, but then pose a prior $p(df)$ for df itself, rather than indirectly through τ^2, and avoid the rather strange right tail behavior. This prior might put more mass on smoother models than rough (as in Figure 7), or might itself be flat over the entire range of df. Since df is monotone with λ, a measure of noise-to-signal ratio, it is quite reasonable to generate these independently of each other.

The posterior (24) is expressed in terms of K, the penalty matrix for a smoothing spline or similar smoother, which is based on a prior covariance $\tau^2 K^-$ for $\mathbf{f}$. Since $K(\lambda) = (S(\lambda)^- - I) = \lambda K$, and parametrizing the smoother through df rather than λ, we get an equivalent representation for the prior

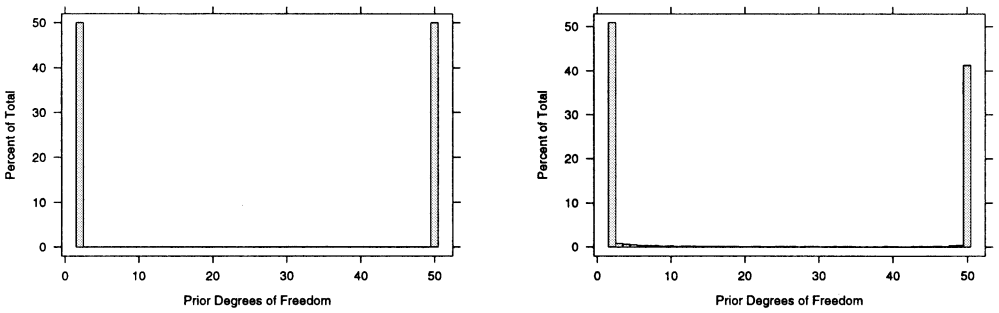

FIG. 6. *Left panel*: *The prior distribution of df based on a flat improper prior for* $\log \tau^2$ *and* $\log \sigma^2$. *Right panel*: *the prior for df based on fairly noninformative proper priors* $IG(0.01, 0.01)$ *prior for* σ^2 *and* τ^2.

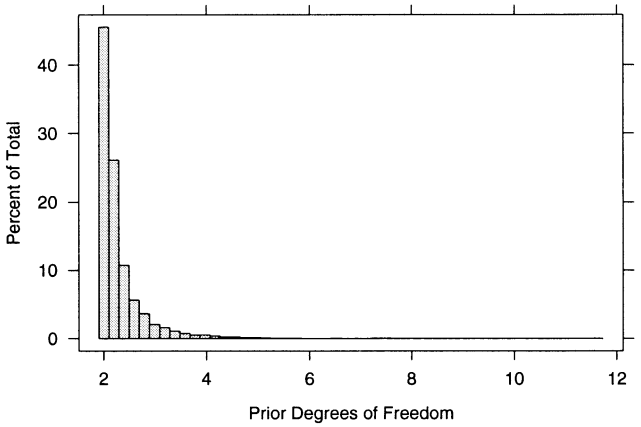

FIG. 7. *The implied prior for df based on* $\sigma^2 \sim IG(2, 0.01)$ *and* $\tau^2 \sim IG(0.5, 0.01)$.

covariance: $\tau^2 K^- = \sigma^2 K^-(df) = \sigma^2(S(\lambda)^- - I)^-$. The posterior distribution for df_j is then

$$p(df_j|\mathbf{y}, \sigma^2, \{\mathbf{f}_j, j = 0, 1, \ldots, p\}) \sim \frac{|K_j(df_j)|^{1/2}}{\sigma^2} p(df) \exp\left(-\frac{\frac{1}{2}\mathbf{f}_j^T K_j(df)\mathbf{f}_j}{\sigma_j^2}\right). \tag{25}$$

4.3 Example: Growth Curves Continued

The bone growth model (19) in Section 3.1 assumes that each girl has her growth spurt at the same age. There is some evidence of the deficiency of this model in the lower right panel of Figure 3, where the random effects distributions seem to have larger variance σ_V^2 around age 13. Here we consider a richer model, that attempts to correct for this deficiency:

$$y_{ij} = f(t_{ij} - \theta_i) + \mathbf{x}_i^T \beta_E + V_i + \varepsilon_{ij}. \tag{26}$$

The parametrization is the same as before, except we introduced an additional random effect, the *age shift* θ_i, which we assume has distribution $N(0, \sigma_\theta^2)$. The joint posterior distribution is now

$$p(f, \theta_i, V_i, \beta_E|\mathbf{y}) \propto \sum_{i=1}^{K}\left\{\sum_{j=1}^{n_i} \frac{(y_{ij} - f(t_{ij} - \theta_i) - \mathbf{x}_i^T\beta_E - V_i)^2}{\sigma^2} + \frac{\theta_i^2}{\sigma_\theta^2} + \frac{V_i^2}{\sigma_V} + \frac{J(f)}{\sigma_f^2}\right\} \tag{27}$$

up to a constant and the components of variance $\sigma_\theta, \sigma_V, \sigma_f$ and σ. The prior of β_E is flat. We first produce the MAP estimates for all the random and fixed effects. This is a large penalized nonlinear least squares problem.

- We have introduced an additional variance component σ_θ. In practice this needs to be estimated as well, either via empirical or full Bayes methods. For expediency, we selected $\sigma_\theta = 1.5$ based on a crude grid search and the BIC statistic and base our subsequent analysis on this value.
- We alternate between:

1. Fitting all the parameters holding the θ_i fixed. This includes the variance components, as well as the MAP estimates of V_i and f and β.
2. Fixing all the parameters in step (i), and computing the MAP estimates of the θ_i.

The first step (i) requires exactly the same technology as in Section 3.1.

- The alternating procedure requires initial values for the θ_i. Ignoring the V_i and fixed effects, we can produce an approximate collapsed version of the model,

$$\tilde{y}_i = f(\lambda_i) + \tilde{\varepsilon}_i,$$
$$\tilde{t}_i = \lambda_i + \theta_i,$$

where $\tilde{y}_i$ represents an average of all the y values for subject i, and so on. This has the form of a nonlinear errors-in-variables model, and can be

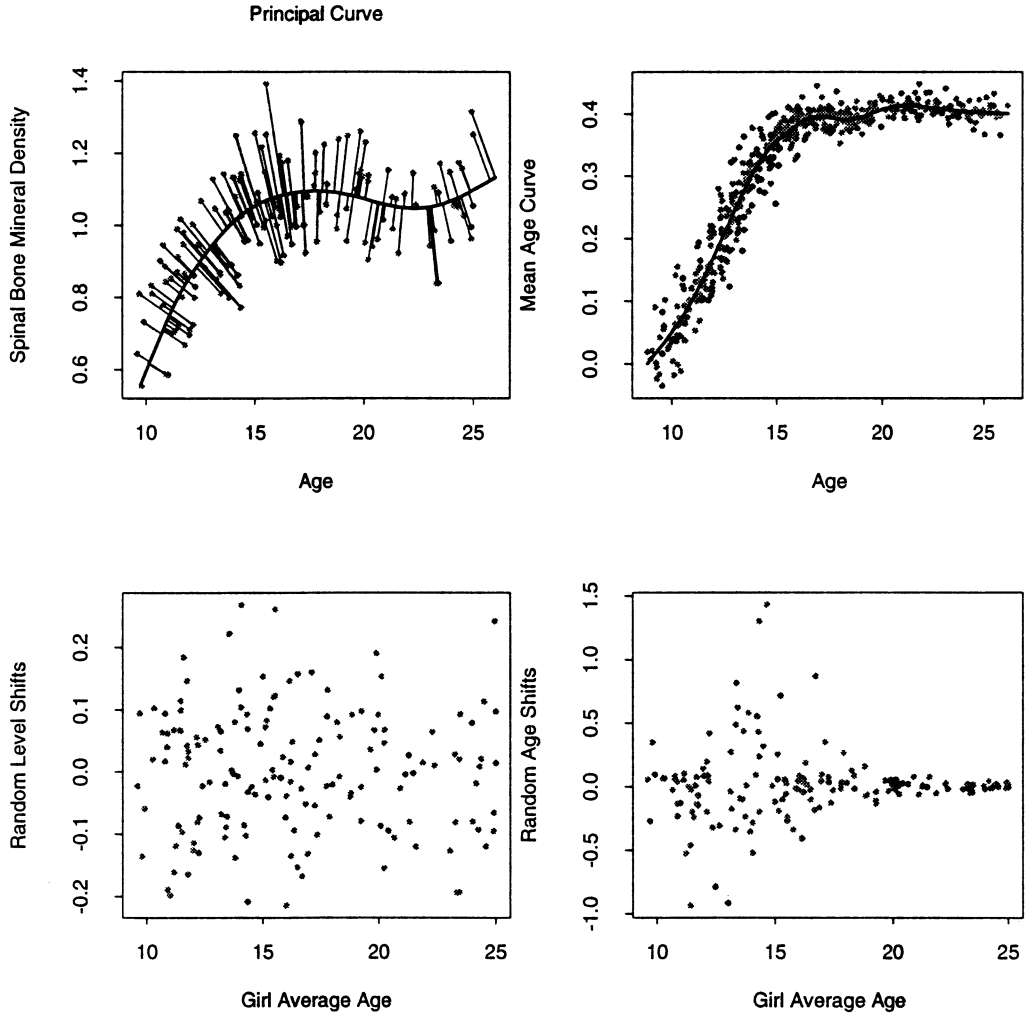

FIG. 8. *The MAP estimates of the nonlinear random effects model. The top left panel shows a principal curve fit to the reduced data, from which initial estimates of θ_i were obtained. The top right panel shows the estimate of f, along with the overall residuals ε_{ij}. The lower left panel shows the estimated random effects V_i and the lower right the estimated θ_i. The latter are far more variable around the growth spurt.*

estimated by the *principal curves* algorithm (Hastie and Stuetzle, 1989).

Figures 8 and 9 show the MAP estimates and illustrate the effect of the inclusion of the random age shift effects θ_i on the fitted random BMD effects V_i.

This sample modification to the model has made it quite nonlinear, and in particular the joint posterior of (θ_i, V_i) will depend on where the observations lie. We do not expect to learn much about θ_i if the growth spurt is over and expect the posterior distributions to look much like the prior. The area where we can learn something is at the earlier ages, where large deviations are attributable to both horizontal and vertical shifts.

The Gibbs simpler for fixed values of θ_i proceeds exactly as before. The only difficult part is sampling from the posterior for θ_i given the rest. The posterior

$$p(\theta_i|\text{rest}) \sim \sum_{j=1}^{n_i} \frac{(r_{ij} - f(t_{ij} - \theta_i))^2}{\sigma^2} + \frac{\theta_i^2}{\sigma_\theta^2}, \tag{28}$$

where $r_{ij} = y_{ij} - \mathbf{x}^T\beta_E - V_i$. This is a univariate simulation problem, and we resort to a simple Taylor approximation to f in (28),

$$f(t_{ij} - \theta_i) \approx f(t_{ij} - \tilde{\theta}_i) - f'(t_{ij} - \tilde{\theta}_i)(\theta_i - \tilde{\theta}_i), \tag{29}$$

where $\tilde{\theta}_i$ is the previous realization of θ_i. We let $a_{ij} = f(t_{ij} - \tilde{\theta}_i)$, $b_{ij} = f'(t_{ij} - \tilde{\theta}_i)$ and $u_{ij} = b_{ij}\tilde{\theta}_i + a_{ij} - r_{ij}$, and after some simple calculations we find that

$$p(\theta_i|\text{rest}) \approx N\left(\frac{\sum_{j=1}^{n_i} b_{ij}u_{ij}}{\sum_{j=1}^{n_i} b_{ij}^2 + \sigma^2/\sigma_\theta^2}, \frac{\sigma^2}{\sum_{j=1}^{n_i} b_{ij}^2 + \sigma^2/\sigma_\theta^2}\right). \tag{30}$$

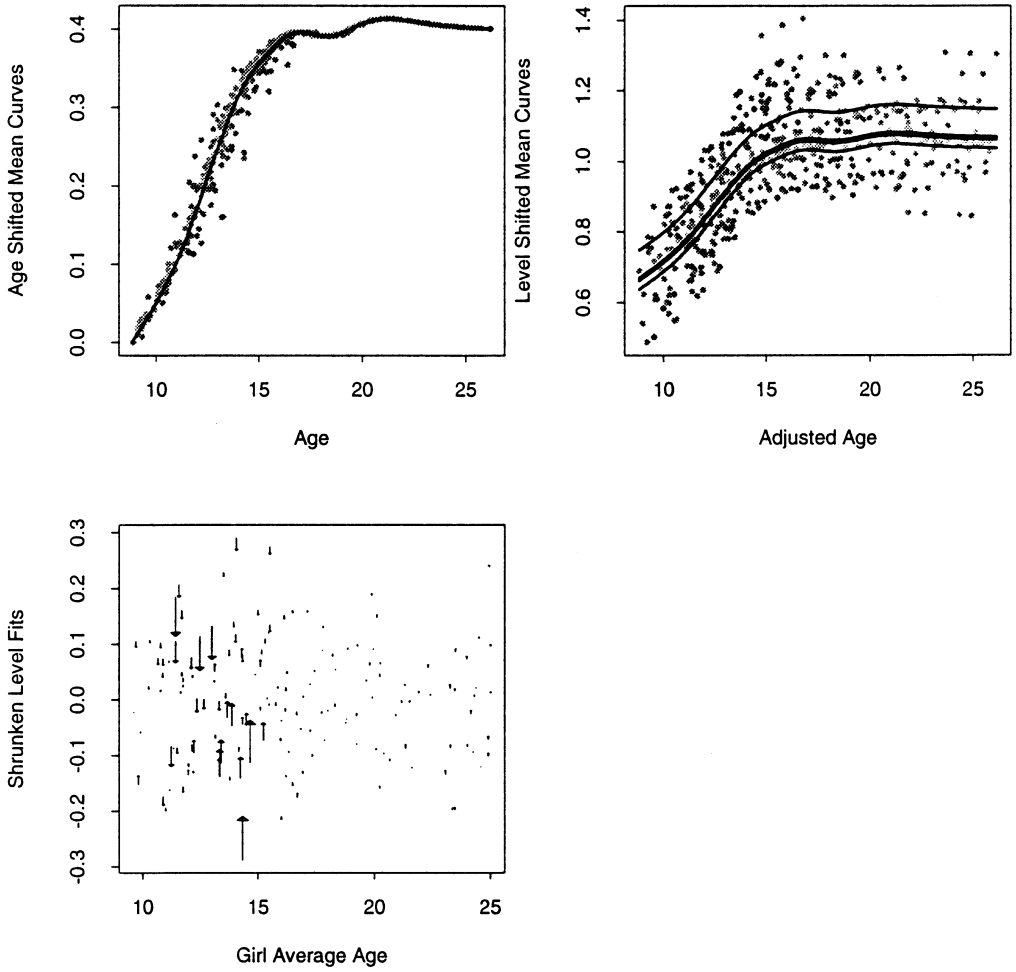

FIG. 9. *The top left panel shows a single average age curve, along with the various shifted versions obtained by adjusting for individual values of* θ_i. *The top right panel is similar, except the adjustments are now vertical shifts caused by the estimated random effects* V_i. *The lower left panel shows the movement of the* V_i *when the* θ_i *are included in the model.*

Figure 10 (left figure) shows the joint distribution of (θ_i, V_i) for nine particular values of i. The right figure show the original data, with the four MAP curves for each ethnic class and with the data for the nine values of i indicated.

5. RELATIONSHIP TO BOOTSTRAP SAMPLING

There is a close relation between Bayesian backfitting for additive models and the bootstrap applied to standard backfitting procedure.

Assume for simplicity that σ^2 is known. In the standard backfitting algorithm with smoothers S_j, the fitting values $\hat{\mathbf{y}}$ and functions $\hat{\mathbf{f}}_j$ satisfy $\hat{\mathbf{y}} = A\mathbf{y}$, $\hat{\mathbf{f}}_j = A_j\mathbf{y}$ where the matrices A and A_j are functions of S_j, $j = 1, 2, \ldots, p$.

It can be shown that the Bayes posterior functions have marginal distributions,

$$\mathbf{f}_j \sim N(A_j\mathbf{y}, (I - A_j)S_j(I - S_j)^{-}\sigma^2). \tag{31}$$

On the other hand, suppose we carry out parametric bootstrap sampling by adding residuals $\mathbf{r}^* \sim N(0, I\sigma^2)$ to the fit $\hat{\mathbf{y}} = A\mathbf{y}$ giving responses $\mathbf{y}^* = A\mathbf{y} + \mathbf{r}^*$. We then apply standard backfitting to the data $\mathbf{y}^*$, giving

$$\mathbf{f}_j^* = A_j\mathbf{y}^* = A_j(A\mathbf{y} + \mathbf{r}^*) \sim N(A_jA\mathbf{y}, A_j^2\sigma^2). \tag{32}$$

In the simple case of only one function ($p = 1$), we have $A_j = A = S_j$, and the Bayesian and bootstrap distributions are $N(S_j\mathbf{y}, S_j\sigma^2)$ and $N(S_j^2\mathbf{y}, S_j^2\sigma^2)$. Now $S_j^2 < S_j$ for cubic spline smoothers and many other smoothers, but typically the two are not very different.

In the general case with p functions, the bootstrap mean $A_jA\mathbf{y}$ is what we obtain if we apply backfitting twice: once to $\mathbf{y}$ to obtain $A\mathbf{y}$ and then again to the response $A\mathbf{y}$. Hence it will tend to be smoother than (but similar to) the Bayesian mean $A\mathbf{y}$. The Bayesian covariance matrice reduces to $S_j\sigma^2$ in the orthogonal case, that is, the inputs are

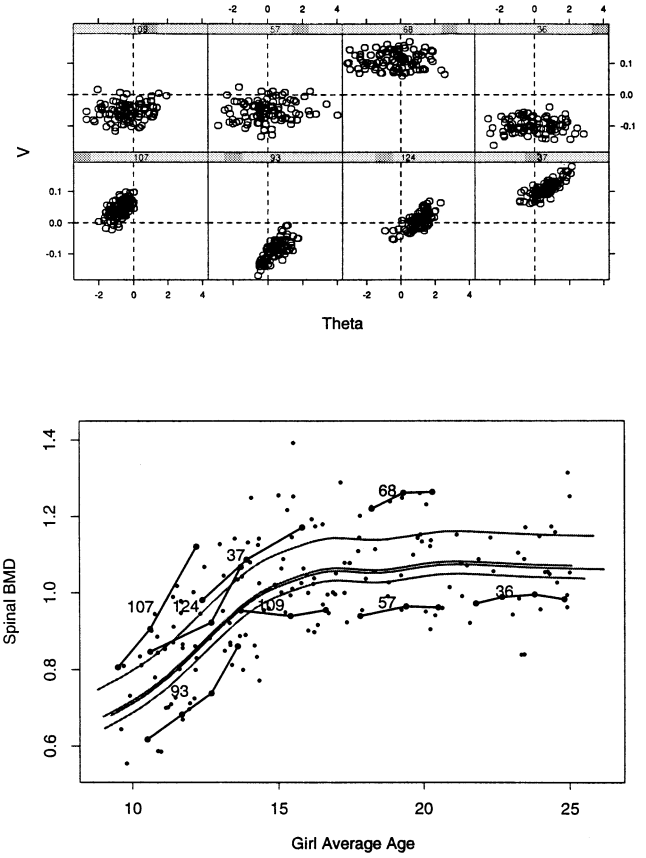

FIG. 10. *Upper figure: the joint posterior scatterplot for* 100 *realizations of* (θ_i, V_i) *for eight different values of* i, *ordered from left to right* (*and bottom to top*) *in age. For large values of age, where* f *is flat, the posterior distribution of* θ_i *has spread similar to the prior* (1.5 *units*). *For values of* i *corresponding to age at the growth spurt, the posterior has smaller spread and is correlated with* V_i. *Lower figure: the numbered data fragments show the data and values of* i *corresponding to the eight panels in the left figure.*

arranged on a lattice. In general however, it is not clear how the Bayesian covariance compares to the bootstrap covariance.

We did a small simulation with $n = 20$ observations and two bivariate standard Gaussian predictors having correlation ρ, for $\rho = 0, 0.5, 0.9$. We computed the matrices S_1 and A_1, and the resulting square root of the average diagonal of the covariance matrices. Table 1 shows the results. We see that the standard deviations are roughly equal, except when the correlation between the inputs is very high. In that case, the bootstrap standard deviation is nearly twice the Bayes posterior standard deviation, and hence would lead to confidence bands that are nearly twice as wide. This may be due to the assumption of proper independence in the Bayes model, reducing the effect of collinearity in the posterior. This interesting issue deserves further study.

TABLE 1
Average standard deviation for Bayes posterior f_1 *and bootstrap realization* f_1^*

ρ	Bayes	Bootstrap
0.0	0.41	0.45
0.5	0.43	0.47
0.9	0.51	0.64

6. GENERALIZED ADDITIVE MODELS

Hastie and Tibshirani (1986) introduced the generalized additive model for modeling non-Gaussian data. This includes members of the exponential family of distributions and other models such as the proportional hazards model for survival data. For a Bayesian analysis of this model, the conditional distributions do not have a simple form in general, as they do in the Gaussian case. Hence Gibbs sampling is no longer convenient. However the basic Gibbs step $\mathbf{f}_j = S_j\mathbf{r}_j + \sigma S_j^{1/2}\mathbf{z}$ can instead be used as a proposal distribution in a Metropolis–Hastings algorithm, as we outline below.

We first give some background on generalized additive models. In the exponential family, the mean μ_i of the response variable Y_i is assumed to be related to the inputs via

$$\eta_i \equiv g(\mu_i) = \sum_j f_j(x_{ij}), \tag{33}$$

where $g(\cdot)$ is a specified function, known as the *link function*. The functions $f_j(\cdot)$ are estimated by maximizing a penalized log-likelihood analogous to the penalized least squares criterion used for Gaussian additive models. Using vector notation, this criterion has the form

$$J(\boldsymbol{\eta}, \boldsymbol{\theta}) = \log L(\boldsymbol{\eta}, \boldsymbol{\theta}) - \sum_j \lambda_j \mathbf{f}^T K_j \mathbf{f}. \tag{34}$$

The function $L(\boldsymbol{\eta}, \boldsymbol{\theta})$ is the likelihood of the data, $\boldsymbol{\eta} = \sum_j \mathbf{f}_j$, $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_p, \sigma^2)$, the tuning parameters and K_j is the penalty matrix, as defined previously. The *local scoring* algorithm for maximization of $J(\boldsymbol{\eta}, \boldsymbol{\theta})$, proposed in Hastie and Tibshirani (1986), is equivalent to a Newton–Raphson procedure. It works by approximating the log likelihood by a quadratic, resulting in a working response variate v_i. A weighted backfitting algorithm is applied to the response v_i, and then v_i is recomputed and the process is repeated, until convergence. The actual

forms for v_i and w_i are

$$v_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right), \qquad w_i^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right) v_i, \tag{35}$$

where v_i is the variance of y_i at μ_i.

How can we simulate from the posterior density $\exp(J(\boldsymbol{\eta}, \boldsymbol{\theta}))$? The Metropolis–Hastings procedure (Hastings, 1970) is a convenient approach here. In general this method works as follows. Given a posterior density $\pi(u)$ from which we wish to generate realizations u, we define a proposal distribution $q(u, v)$ that specifies the probability of moving from state u to v. If we are currently at state u, we generate a random state v according to $q(u, v)$, and then move to v with probability

$$\alpha(u, v) = \begin{cases} \min\left\{\dfrac{\pi(v)q(v,u)}{\pi(u)q(u,v)}, 1\right\}, & \text{if } \pi(u)q(u,v) > 0, \\ 1, & \text{if } \pi(u)q(u,v) = 0. \end{cases} \tag{36}$$

In our application, we consider a move for a single function $\mathbf{f}_j \to \mathbf{f}_j'$, with all other parameters held fixed. Let $\boldsymbol{\eta} = \sum \mathbf{f}_j$, $\boldsymbol{\eta}' = \sum_{k \neq j} \mathbf{f}_k + \mathbf{f}_j'$ and let the corresponding working responses and diagonal weight matrices be $\mathbf{v}, \mathbf{v}'$ and W, W'. We choose as the proposal distribution the normal approximation $q(\mathbf{f}_j, \mathbf{f}_j') = N(S_j\mathbf{v}, S_j W^{-1}\sigma^2)$ which results from expanding the log-likelihood in a second-order Taylor series. The move from $\mathbf{f}$ to $\mathbf{f}'$ has the form

$$\mathbf{f}' = S_j\mathbf{v} + \sigma S_j^{1/2} W^{-1/2}\mathbf{z}, \tag{37}$$

which is just a weighted version of the basic operation in the Bayesian backfitting procedure given in Algorithm 3.1. The acceptance probability works out to be

$$\frac{\pi(\mathbf{f}')\mathrm{q}(\mathbf{f}', \mathbf{f})}{\pi(\mathbf{f})\mathrm{q}(\mathbf{f}, \mathbf{f}')} = \frac{\mathrm{e}^{\mathrm{J}(\boldsymbol{\eta}, \boldsymbol{\theta})}}{\mathrm{e}^{\mathrm{J}(\boldsymbol{\eta}', \boldsymbol{\theta})}} \frac{|\mathrm{S}_j'\mathrm{W}'^{-1}|}{|\mathrm{S}_j'\mathrm{W}^{-1}|} \frac{\mathrm{e}^{-(1/2)(\mathbf{f}_j' - \mathrm{S}_j\mathbf{v})^{\mathrm{T}}[\mathrm{S}_j\mathrm{W}^{-1}]^{-1}(\mathbf{f}_j' - \mathrm{S}_j\mathbf{v})}}{\mathrm{e}^{-(1/2)(\mathbf{f}_j - \mathrm{S}_j'\mathbf{v}')^{\mathrm{T}}[\mathrm{S}_j'\mathrm{W}'^{-1}]^{-1}(\mathbf{f}_j - \mathrm{S}_j'\mathbf{v}')}}. \tag{38}$$

For smoothing splines the operator that uses observation weights W has the form $S_j = (W + \lambda_j K_j)^{-1} W$, and the acceptance probability is easily computed in $O(n)$ operations.

The tuning parameters $\tau_j^2 = \sigma^2/\lambda_j$ and σ^2 can be sampled in a similar way. In expression (24) for the conditional distribution of τ_j^2, we again need to compute the penalty $\mathbf{f}_j^T K_j \mathbf{f}$. $\mathbf{f}^t(S_j^- - I)\mathbf{f}$ is replaced by $\mathbf{f}^t W(S_j^- - I)\mathbf{f}$ and we have $\mathbf{f}^t W S_j^- \mathbf{f} = \mathbf{r}^T S W \mathbf{r} + 2\mathbf{r}^T S^{1/2} W^{1/2}\mathbf{z} + \mathbf{z}^T\mathbf{z}$, all quantities that are already available.

The determinant ratio is not as easy to compute, but is typically very close to 1 (all that changes are the weights), and we can ignore it in the calculations. Details of this procedure, including an SPlus software implementation and a comparison to the related approach of (Zeger and Karim, 1991), will appear elsewhere.

7. DISCUSSION

The additive model used here is a special case of the Gaussian process model for flexible regression. In this class of models, a Gaussian process prior is assumed for the regression function, and inference is carried out from the posterior. Different choices for the prior covariance function lead to particular models: the additive smoothing spline model results from the prior discussed in this paper. The general Gaussian process model was proposed by O'Hagan (1978). More recently Neal (1996) and Williams and Rasmussen (1996) have explored the computational aspects in depth. They use MCMC for the tuning parameters, and an $O(n^3)$ procedure to obtain the the mean and covariance of the Gaussian posterior. This $O(n^3)$ operation can make the analysis infeasible for large n. Because of the banded nature of the matrices arising in the cubic smoothing spline model, we are able to reduce this computation to $O(nM)$ where M is the number of Gibbs sampling steps.

As mentioned in the introduction, Wong and Kohn (1996) provide an $O(n)$ algorithm for the additive spline model using the state-space representation of splines, introduced in Ansley and Kohn (1996). This framework is formally equivalent to that of Wahba (1980). We make no claims that our procedure is more efficient that theirs in the additive spline model; rather we believe that our proposal has the advantages of conceptual simplicity and generality.

As suggested by a referee, extensions to scale mixture and auto-correlated errors are possible using the methods proposed in Smith, Wong and Kohn (1998). We have also restricted ourselves to the use of proper priors. In general, choosing the prior to ensure that the posterior is proper can be a tricky exercise, especially in random effects models. See Hobert and Casella (1996) for detailed discussion of this issue.

We have written several functions in the S-plus language for implementing the ideas in this paper. In particular, a function `gibbs.gam()` takes as input a fitted `gam` object (Chambers and Hastie, 1991) and samples from the posterior distribution. The follow-

ing lines produced the essential ingredients for the figures in Section 3.1:

```
bonefit <- gam(spnbmd ~ s(age,9) + ethnic
  + random(factor(idnum)), data=Bonef)
bone.samples <-gibbs.gam(bonefit, nwarm=3600,
  nkeep=100, var.comp=T)
```

Even though `bonefit` requested $9df$ in `s(age,9)`, this acts simply as a starting value in the call to `gibbs.gam()`. The `gam()` object can specify any number of smooth terms and random effects, and they all get accommodated automatically. Although `random()` is a rather simple random intercept "smoother," it is not difficult for users to provide their own random effects methods.

The `gibbs.gam` collection will be made available from the public archive at Carnegie-Mellon University: `http://www.lib.stat.cmu.edu`.

From a mixed effects or empirical Bayes point of view, we have provided an $O(n)$ algorithm for sampling from the posterior distributions, given the variance components, even when the random effects (smoothing splines) have dimension n. The usual backfitting procedure delivers the posterior means or BLUPs in $O(n)$ computations. We are currently exploring approximations that allow the estimation of the variance components as well in $O(n)$ computations. We gave one such approximation in Section 4.1.

APPENDIX A

ALGORITHMIC DETAILS

Algorithms for Generating $S^{1/2}\mathbf{z}$

We present two algorithms for generating an n-vector $S^{1/2}\mathbf{z}$, where as before $\mathbf{z}$ is a vector of $N(0,1)$ variates. The first algorithm is iterative, and uses repeated applications of the smoothing operator S. It has the same order of complexity as the smoother and hence is $O(n)$ if the smoother can be applied with only $O(n)$ calculations. This is the case for many popular smoothers including cubic smoothing splines, kernels and wavelet smoothers. The second procedure is specifically designed for cubic smoothing splines, and uses the banded nature of the covariance kernel to generate $S^{1/2}\mathbf{z}$. It is more efficient than the first algorithm but applicable only to cubic smoothing splines.

GENERAL ALGORITHM. Consider the Taylor series

$$S^{-1/2} = I - \tfrac{1}{2}(S-I) + \tfrac{3}{8}(S-I)^2 - \tfrac{5}{16}(S-I)^3\cdots \tag{39}$$

and premultiply by S, giving

$$\begin{aligned} S^{1/2} &= S\cdot S^{-1/2} \\ &= S - \tfrac{1}{2}S(S-I) + \tfrac{3}{8}S(S-I)^2 \\ &\qquad - \tfrac{5}{16}S(S-I)^3\cdots. \end{aligned} \tag{40}$$

Hence we can apply $S^{1/2}$ by repeated applications of S representing the right-hand side of (40). This leads to the following algorithm.

ALGORITHM A.1. *General procedure for generating* $S^{1/2}\mathbf{z}$.

1. *Take* $\mathbf{z}\sim N(0,I)$. *Set* $\mathbf{z}' = S\mathbf{z}$, $\mathbf{z}'' = \mathbf{z}$.
2. *Do for* $b = 2, 3, \ldots,$
   - $\mathbf{z}'' \leftarrow \frac{3/2-b}{(b-1)}\cdot(S\mathbf{z}'' - \mathbf{z}'')$;
   - $\mathbf{z}' \leftarrow \mathbf{z}' + S\mathbf{z}''$.
3. *Until* $||S\mathbf{z}''||$ *is small.*

In step 2, the strange-looking multiplier $(\frac{3}{2} - b)/(b-1)$ generates the coefficients in the Taylor series (40). It is easy to show that $\mathbf{z}' \to S^{1/2}\mathbf{z}$, as long as $S(S-I)^b \to 0$. This is true for any symmetric smoother having eigenvalues in $[0,1]$: this includes cubic smoothing splines and some symmetrized kernel smoothers. Note that for projections, $S(S-I) = 0$ and so convergence is immediate (no iterations of step 2).

Table 2 shows the results of a simulation experiment to examine the convergence of this procedure. With a sample size $n = 100$, we generated a random normal vector $\mathbf{z}$ and applied the above algorithm with a cubic smoothing spline operator with degrees of freedom randomly chosen between 2 and 20. The convergence criterion was $\max|S\mathbf{z}''| < 0.01$. The number of iterations until convergence for 1000 simulations is shown in Table 2. The convergence is quite fast, never requiring more than 12 iterations and usually no more than six or seven.

TABLE 2
Number of iterations until convergence in 1000 *experiments, for general* $S^{1/2}\mathbf{z}$ *algorithm*

3	4	5	6	7	8	9	10	11	12
25	162	252	237	157	104	41	14	5	3

Algorithm for smoothing splines. When the smoother S represents a smoothing spline, we can implement a more precise and efficient algorithm for generating $S^{1/2}z$. Our implementation of smoothing splines follows de Boor (1978), where we represent the fitted functions in a basis of cubic

B-splines,

$$f(x) = \sum_{j=1}^{M} b_j(x)\theta_j. \tag{41}$$

The number of basis functions M depends on the number of unique values of x among the n input values x_i, as well as the particular representation used. In our case n_u unique values of x define $n_u - 2$ *interior* knots and a corresponding basis of $M = n_u + 2$ cubic B-splines. If all the n values of x are unique, then $M = n + 2$. The smoothing spline solution is given by

$$\begin{aligned} \hat{\mathbf{f}} &= S\mathbf{y} \\ &= B(B^TB + \lambda\Omega)^{-1}B^T\mathbf{y} \\ &= B\hat{\boldsymbol{\theta}}, \end{aligned} \tag{42}$$

where the n rows of the $n \times M$ basis matrix B consist of the vector of M basis functions $\mathbf{b}(x)$ evaluated at the n sample values x_i. The $M \times M$ penalty matrix Ω has elements

$$\Omega_{ij} = \int b_i''(t)b_j''(t)\,dt.$$

Likewise, the fitted function at an arbitrary input value x is given by

$$\begin{aligned} \hat{f}(x) &= \mathbf{b}^T(x)(B^TB + \lambda\Omega)^{-1}B^T\mathbf{y} \\ &= \mathbf{b}^T(x)\hat{\boldsymbol{\theta}}. \end{aligned} \tag{43}$$

The coefficient estimates $\hat{\boldsymbol{\theta}}$ are the posterior mean for $\boldsymbol{\theta}$ based on a model $y = \mathbf{b}^T(x)\boldsymbol{\theta} + \varepsilon$, where:

- $\varepsilon \sim N(0, \sigma^2)$.
- $\boldsymbol{\theta}$ has a (degenerate) prior normal distribution $N(0, \tau^2\Omega^-)$ with $\lambda = \sigma^2/\tau^2$. Here Ω has a two-dimensional null space corresponding to parameters leading to linear functions of x. It also gives effectively infinite penalty to nonzero second derivatives at the boundary knots, and hence enforces the *natural* boundary conditions.
- The prior covariance matrix K^- for $\mathbf{f}$ in (5) is $B\Omega^-B^T$ evaluated at the data.
- The above expressions generalize easily to the case where each observation has a weight. This happens naturally in nonlinear likelihood settings as in the next section, and also when the x values are tied. In the latter case the observations are collapsed onto the unique values of x_i, the y_i are replaced by the average at the tied values of x_i, and the observations receive weights proportional to the counts at each unique x.

Thus the posterior distribution for $\boldsymbol{\theta}$ is $\boldsymbol{\theta}|\mathbf{y} \sim N(\hat{\boldsymbol{\theta}}, \sigma^2(B^TB + \lambda\Omega)^{-1})$. Likewise, the posterior distribution of $\mathbf{f}$ is

$$\begin{aligned} \mathbf{f}|\mathbf{y} &\sim N(B\hat{\boldsymbol{\theta}}, \sigma^2B(B^TB + \lambda\Omega)^{-1}B^T) \\ &= N(S\mathbf{y}, \sigma^2S). \end{aligned} \tag{44}$$

Hence to simulate from this posterior, it is sufficient to simulate a parameter $\boldsymbol{\gamma} \sim N(0, \sigma^2(B^TB+\lambda\Omega)^{-1})$, and hence we can produce a posterior realization of the entire function.

It turns out that there is no additional computational burden over and above the usual smoothing spline $O(n)$ computations. The matrix B has four nonzero bands, and both B^TB and Ω are 4-banded. This means that $B^TB + \Omega = L^TL$ has a 4-banded cholesky factorization L (Silverman, 1984). This L is computed as part of the smoothing-spline calculations, and is available as part of the fit. Hence $\boldsymbol{\gamma} = L^{-1}\mathbf{z} \sim N(0, (B^TB + \lambda\Omega)^{-1})$ if $\mathbf{z} \sim N(0, I)$. We obtain γ by solving $L\boldsymbol{\gamma} = \mathbf{z}$, which takes $O(n)$ computations, because of the banded nature of L.

Modified Backfitting and Efficiency

In Section 3 we mentioned that the output of the smoothers have to be centered, to avoid identifiability problems. Here we describe a more general centering that speeds up convergence of the Bayesian backfitting algorithm.

In the standard backfitting algorithm, strong correlations among the inputs can cause slow convergence, because the procedure slowly seesaws towards the solution. In Buja, Hastie and Tibshirani (1989) a modified backfitting algorithm was proposed, in which all of the (linear) projections for all of the inputs were fit together, while the iterative one-at-a-time smoothing was applied just to the nonlinear parts of each function. This can noticeably speed up the convergence of backfitting, because it immediately captures the linear correlations. We let $X\beta$ denote the linear part of the model (including intercept) with projection operator H, and H_j the operator that projects onto the two-dimensional linear subspace of eigenvalue 1 of S_j. Then the modified backfitting algorithm uses the smoothers H and $\tilde{S}_j = S_j - H_j$ (for symmetric smoothers, such as smoothing splines). $\tilde{S}_j$ produces the nonlinear part of the fit for variable x_j.

An analogous strategy can be used to speed up the Bayesian backfitting procedure (Liu, Wong and Kong, 1994). We can separate each function $\mathbf{f} = X_j\beta_j + \tilde{\mathbf{f}}$, where $X_j\beta_j$ includes the constant and linear part of $\mathbf{f}_j$. The prior and posterior distributions factor accordingly, $X_j\beta_j|\mathbf{y} \sim N(H_j\mathbf{y}, \sigma^2H_j)$ and $\tilde{\mathbf{f}}|\mathbf{y} \sim N(\tilde{S}_j\mathbf{y}, \sigma^2\tilde{S}_j)$, and they are independent. Notice as well that $\tilde{S}_j^{1/2} = S_j^{1/2} - S_j$. Then we alternately generate realizations of the linear component

$X\beta$ all grouped together, separately from the nonlinear functions $\tilde{\mathbf{f}}_j$. The latter is achieved by first generating the usual realization $\mathbf{f}_j$, and then removing the linear trend $\tilde{\mathbf{f}}_j = \mathbf{f}_j - H_j\mathbf{f}_j$.

APPENDIX B

EXACT PRIOR FOR *df* BASED ON FLAT PRIORS FOR VARIANCE COMPONENTS

THEOREM. *Consider the Bayesian smoothing spline model, on N unique values of x. Let the prior for τ^2 and σ^2 both be improper and flat on the* log*-scale*; $p(\tau^2) \sim 1/\tau^2$ *and* $p(\sigma^2) \sim 1/\sigma^2$. *Then the implicit prior on df is discrete, and puts mass* $\frac{1}{2}$ *on* 2 *and* N.

LEMMA. *Consider the random variable* $V_D = 1/(1 + \kappa/Z_D)$, *where* $\log(Z_D) \sim U[-D, D]$. *Then* $V = \lim_{D\to\infty} V_D$ *is* 0 *or* 1 *with probability* $\frac{1}{2}$.

PROOF OF LEMMA.

$$P(V_D > v_0) = P(\log(Z_D) < \log\left(\frac{\kappa v_o}{1 - v_0}\right)$$

$$= \frac{1}{2}\left[1 + \frac{\log\left(\kappa v_o/1 - v_0\right)}{D}\right] \tag{45}$$

$$\text{for } v_0 \in \left[\frac{1}{1 + \kappa/e^D}, \frac{1}{1 + \kappa/e^{-D}}\right].$$

Now for any $v_0 \in (0, 1)$,

$$\lim_{D\to\infty} P(V_D > v_0) = P(V > v_0) \tag{46}$$

$$= \tfrac{1}{2}.$$

PROOF OF THEOREM. For a smoothing spline,

$$df = \sum_{j=1}^{N} \frac{1}{1 + \lambda d_j}. \tag{47}$$

Here $\lambda = \sigma^2/\tau^2$, the d_i are the eigenvalues of the $N \times N$ penalty matrix K, and $d_1 = d_2 = 0$ and $d_j > 0$ for $j = 3, \ldots, N$. So for any fixed value of σ^2, each of the contributions for $j > 2$ is the same and either 0 or 1, and the first two contributions are always 1. Thus for fixed σ^2 df is 2 or N with probability $\frac{1}{2}$. Since this does not depend on σ^2, this is also unconditionally true. Finally, since $p(\log(\tau^2)) \sim 1$, we see that $\lim_{D\to\infty} \mathscr{L}(\log Z_D) = \mathscr{L}(\log \tau^2)$. □

We have not proved this for the nearly-flat prior $p(\tau^2) \sim (1/\tau^2) \times \exp(-\rho/\tau^2)$ (which is in fact not integrable and so also improper). Empirical evidence suggests that this has the same distribution, obtained by simulating from an $IG(\varepsilon, \rho)$ density, and studying the behavior of the quantiles of V_ε as ε gets small.

One puzzling aspect of this prior distribution is that it has no support except on these two extreme points. This implies the posterior should be the same. One explanation for this somewhat contradictory behavior is that the Gibbs samplers are never run long enough! $df = N$ is an absorbing state, since this implies that $\|\mathbf{e}\| = 0$ and hence the posterior for σ^2, $IG(N/2, \|\mathbf{e}\|^2)$ will produce a 0 with probability 1, leading to $\lambda = 0$ and another exact fit. Likewise $df = 2$ is an absorbing state if $1/\tau^2$ is used for the prior. This is not the case for $(1/\tau^2)\exp(-\rho/\tau^2)$, whose posterior is $IG(N/2, \frac{1}{2}\mathbf{f}^T K\mathbf{f} + \rho)$. Even though the penalty may be zero (for exact linear fits), the presence of $\rho > 0$ protects!

ACKNOWLEDGMENTS

We thank Radford Neal for suggesting the use of the Metropolis–Hastings procedure in Section 6, Bernard Silverman for help with the smoothing spline representation, Larry Wassermen, the Editor and two referees for helpful comments. Xihong Lin was especially helpful in providing (personal communication) an up-to-date survey of the mixed effects field and making her preprints available. Trevor Hastie was supported in part by NSF Grant DMS-95-04495 NIH Grant ROI-CA-72028-01. Robert Tibshirani was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

ANSLEY, C. and KOHN, R. (1985). Estimation, filtering and smoothing in state space models with diffuse initial conditions. *Ann. Statist.* **13** 1286–1316.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555.

CARTER, C. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553.

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

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

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

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D. (1995). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* **6** 721–741.

GREEN, P. and SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.
HASTIE, T. (1995). Pseudosplines. *J. Roy. Statist. Soc. Ser. B* **58** 379–396.
HASTIE, T. and STUETZLE, W. (1989). Principle curves. *J. Amer. Statist. Assoc.* **84** 502–516.
HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 295–318.
HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.
HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.
HOBERT, J. and CASELLA, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91** 1461–1473.
HODGES, J. and SARGENT, D. (1998). Counting degrees of freedom in hierarchical and other richly parametrized models. Technical report, Div., Biostatistics, Univ. Minnesota.
HOLMES, C. and MALLICK, B. (1997). Bayesian wavelet networks for nonparametric regression. *IEEE. Trans. Neural Networks.* To appear.
LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.
LIN, X. and ZHANG, D. (1997). Inference in generalized additive mixed models. Technical report, Biostatistics, Dept., Univ. Michigan.
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.
NEAL, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer, New York.
O'HAGAN, A. (1978). Curve fitting and optimal design for regression (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 1–42.
SILVERMAN, B. (1984). Spline smoothing: the equivalent kernel method. *Ann. Statist.* **12** 898–9164.
SMITH, M., WONG, C. and KOHN, R. (1998). Additive nonparametric regression with autocorrelated errors. *J. Roy. Statist. Soc. Ser. B* **60** 311–332.
SPEED, T. (1991). Comment on "That BLUP is a good thing: the estimation of random effects." *Statist. Sci.* **6** 42–44.
SPIEGELHALTER, D., BEST, N., GILKS, W. and INSKIP, H. (1996). Hepatitis B: a case study in mcmc methods. In *Markov Chain Monte Carlo in Practice* (W. Gilks, S. Richardson and D. Spegelhalter, eds.) Chapman and Hall, London.
TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762.
WAHBA, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. In *Proceedings of the International Conference on Approximation Theory in Honour of George Lorenz*. Academic Press, Austin, TX.
WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.
WILLIAMS, C. and RASMUSSEN, C. (1996). Gaussian processes for regression. In *Neural Information Processing Systems* **8** (D. S. Touretzky, M. C. Mozer and M. E. Hasselmo, eds.) MIT Press.
WONG, C. and KOHN, R. (1996). A Bayesian approach to estimating and forecasting additive nonparametric autoregressive models. *J. Time Ser. Anal.* **17** 203–220.
ZEGER, S. and KARIM, M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

Comment

R. Dennis Cook and Iain Pardoe

1. INTRODUCTION

Hastie and Tibshirani propose an intriguing idea, neatly linking Bayesian modeling of the functions in a generalized additive model with Gibbs sampling to obtain posterior realizations of these functions. Since their procedure utilizes only smoother matrices for individual predictors, S_j, partial residuals, $\mathbf{r}_j$, and normal random vectors, $\mathbf{z}_j$, the method would appear to be applicable to any models with additive components that can be expressed in the form $S_j\mathbf{y}$.

A natural question to ask of any proposed methodology is "to what use can it be put?" Hastie and Tibshirani's examples, while interesting in themselves, left us questioning what information could be gleaned from plots such as Figures 1 and 2 for the ozone data and Figure 3 for the growth curves data. For example, do the individual realizations in Figure 2 add anything to the information already provided by the pointwise posterior intervals? Figure 4 goes some way to addressing these thoughts with a graphical display of two functionals of the posterior realizations. We decided to pursue these thoughts in a different direction, that of model checking, and we outline our findings in Section 2. We discuss other potential applications in Section 3 and make some more general comments in Section 4.

R. Dennis Cook is Professor and Iain Pardoe is Graduate Student at the School of Statistics, University of Minnesota, St. Paul, Minnesota 55108 (*e-mail:* dennis@stat.umn.edu).

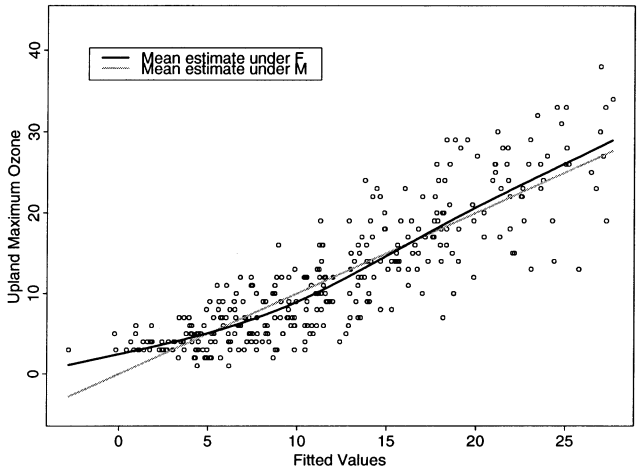

FIG. 1. *Marginal model plot for the fitted values for the additive model fit to four air pollution variables.*

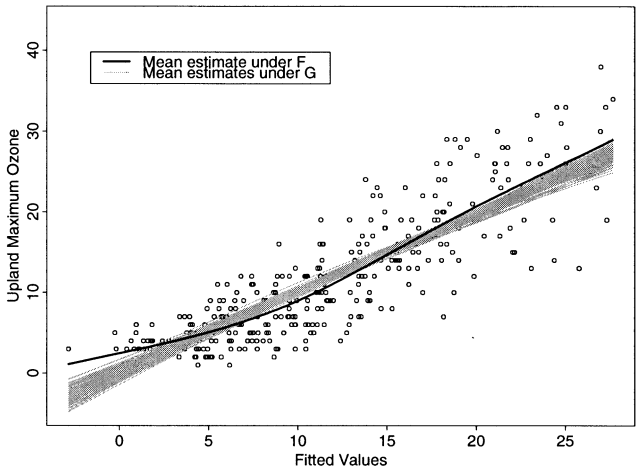

FIG. 2. *Gibbs marginal model plot for the fitted values for the additive model fit to four air pollution variables.*

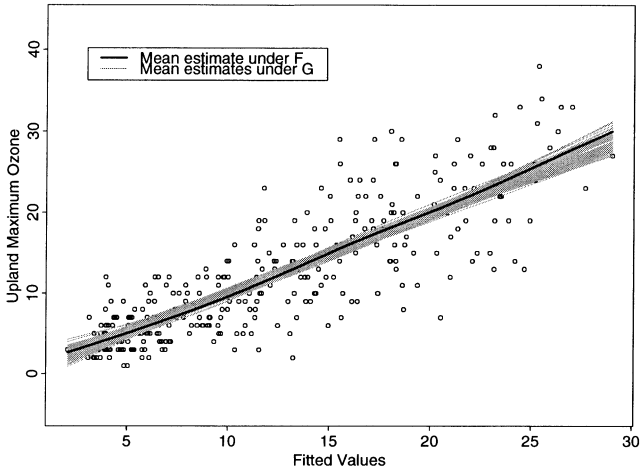

FIG. 3. *Gibbs marginal model plot for the fitted values for the additive model fit to* (w_1, w_2, w_{12}).

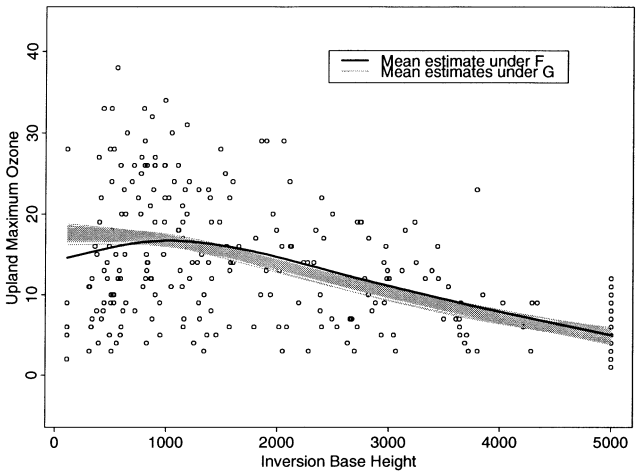

FIG. 4. *Gibbs marginal model plot for inversion base height for the additive model fit to* (w_1, w_2, w_{12}).

2. MARGINAL MODEL PLOTS

The goal of a regression analysis can be expressed as inference about the dependence of an unknown cdf F of the conditional random variable $y \mid \mathbf{x}$ on the value of $\mathbf{x}$. Consider a generic regression model for $y \mid \mathbf{x}$ represented by the cdf M; estimating this model gives rise to an estimated cdf $\widehat{M}$. We now consider graphics for comparing selected characteristics of F to the corresponding characteristics of $\widehat{M}$. We use the fact that $F(y \mid \mathbf{x}) = M(y \mid \mathbf{x})$ for all values of $\mathbf{x}$ in its sample space if and only if $F(y \mid h) = M(y \mid h)$ for all functions $h = h(\mathbf{x})$. This is a more general version of the approach proposed by Cook and Weisberg (1997) which sets $h = \mathbf{a}^T\mathbf{x}$, where $\mathbf{a} \in \mathbb{R}^p$. In particular, we focus on comparing a nonparametric estimate of the mean of $y \mid h$ to the corresponding mean computed from $\widehat{M}$, for various functions h.

For some fixed h, plot y versus h. Add a nonparametric mean estimate, say a cubic smoothing spline with fixed degrees of freedom, to the plot; denote this $\widehat{E}_F(y \mid h)$, where E_F denotes expectation under F. We wish to compare this mean estimate with a mean estimate under $\widehat{M}$, $\widehat{E}_{\widehat{M}}(y \mid h)$, where $E_{\widehat{M}}$ denotes expectation under $\widehat{M}$. Since $E_{\widehat{M}}(y \mid h) = E[E_{\widehat{M}}(y \mid \mathbf{x}) \mid h]$, we can obtain $\widehat{E}_{\widehat{M}}(y \mid h)$ from a nonparametric mean estimate for the regression of the fitted values under M, $E_{\widehat{M}}(y \mid \mathbf{x})$, on h. We can then add this to the plot to obtain a marginal model plot (MMP) for h; this can be thought of as a plot for checking the model in the (marginal) direction h. Using the same method (and smoothing parameter) to obtain this estimate as that used to obtain the mean estimate under F allows point-wise comparison of the two estimates, since any estimation bias should cancel. See Bowman and Young (1996) for further discussion of this point. If the model is

a close representation of F, we can expect that for any quantity h the marginal mean estimates should agree, $\widehat{E}_{\widehat{M}}(y \mid h) \approx \widehat{E}_F(y \mid h)$.

Ideas for selecting which MMPs (i.e., which functions h) to consider in practice are given in Cook and Weisberg (1997), with additional discussion provided in Cook (1998) and Cook and Weisberg (1999). Some examples of useful MMPs include those for fitted values, individual predictors and linear combinations of the predictors. Any indication that the estimated marginal means do not agree for one particular MMP suggests that the model could perhaps be improved; if they agree for a variety of plots, we have support for the model. The ideas above can be extended to variance estimates to provide further ways for checking models.

Consider, for example, a MMP for the fitted values for Hastie and Tibshirani's ozone data example with four predictor variables. The plot in Figure 1 shows a systematic discrepancy between the (black) mean estimate under F and the (gray) mean estimate under $\widehat{M}$; the mean estimate under $\widehat{M}$ is too low on the left, too high in the middle and too low again on the right. Both mean estimates were calculated using the S-plus function `smooth.spline` with (the default) four degrees of freedom.

On the other hand, relative to the variation in the data, the mean estimate under $\widehat{M}$ does not appear to be too far from the mean estimate under F. So, are the discrepancies enough to indicate any potential for model improvement? Porzio and Weisberg (1999) provide some frequentist methodology to address this issue: pointwise reference bands to aid visualization and statistics to calibrate discrepancies. Hastie and Tibshirani's procedure also provides methodology to address this issue. They make the well-taken points that we can make use of the individual realizations of the posterior distributions of the functions in an additive model, and display the posterior distributions of *interesting functionals* of them. They also note that we can carry out Bayesian inference for any quantity of interest. This would appear to offer a Bayesian way to aid visualization in a MMP, with potential possibilities for calibrating discrepancies.

For any particular MMP, it would be useful to display mean estimates for the individual realizations from the posterior distribution of the fitted values, where the fitted-value realizations are just $\sum_{j=0}^{p} \mathbf{f}_j^t$, $t = 1, 2, 3, \ldots$ and $\mathbf{f}_j^t = S_j \mathbf{r}_j^t + \sigma S_j^{1/2} \mathbf{z}_j^t$. So, instead of adding the mean estimate under $\widehat{M}$ to the plot of the mean estimate under F, we can instead add a mean estimate for each Gibbs sample, G^t, and obtain what we call a *Gibbs marginal model plot* (GMMP). If enough samples are taken, say 50 or 100, the Gibbs mean estimates will form an approximate mean estimate *band* under $\widehat{M}$. This plot may provide a visual way of determining whether there is any evidence to contradict the possibility that $F(y \mid h) = M(y \mid h)$. Intuitively, if, for a particular h, the mean estimate under F lies substantially outside the mean estimate band under $\widehat{M}$ (formed from the mean estimates under G^t), then perhaps the model can be improved. If, no matter what the function h is, the mean estimate under F lies broadly inside the mean estimate band under $\widehat{M}$, then perhaps the model provides a reasonable description of the conditional distribution of $y \mid \mathbf{x}$. It would appear to be possible to supplement this purely graphical methodology with more formal Bayesian inference.

Consider a GMMP for the fitted values for the ozone data. Hastie and Tibshirani kindly provided us with the S plus functions for implementing the ideas in their paper, as well as with help in using their code. This enabled us to construct the GMMP in Figure 2. The Gibbs sampling was carried out using the fully Bayesian procedure described in Hastie and Tibshirani's Section 4, with a warm-up period of 300 iterations. The plot shows the (black) mean estimate under F lying mostly outside the mean estimate band under $\widehat{M}$ [formed from 100 (gray) mean estimates under G^t]. This appears to offer clear evidence that the fitted model can be improved.

As curious applied statisticians, we couldn't resist trying to see if we could come up with a better model for these data. One particular technique we applied was *sliced average variance estimation* (SAVE), introduced by Cook and Weisberg (1991) and developed by Cook and Lee (1999). SAVE is a model-free method for estimating the smallest subspace $\mathcal{S}$ of $\mathbb{R}^p$ so that y and $\mathbf{x}$ are independent given the projection of $\mathbf{x}$ onto $\mathcal{S}$, $P_{\mathcal{S}}\mathbf{x}$. In words, all the information about y that is available from $\mathbf{x}$ is contained in $P_{\mathcal{S}}\mathbf{x}$. Following Li (1991), $\mathcal{S}$ is a *dimension reduction subspace* for the regression of y on $\mathbf{x}$. The smallest such $\mathcal{S}$ is called the *central* subspace, $\mathcal{S}_{y|\mathbf{x}}$ (Cook, 1994; Cook, 1998); SAVE yields a subspace estimate, $\mathcal{S}_{\mathrm{SAVE}} \subset \mathcal{S}_{y|\mathbf{x}}$. This estimate can then be used to postulate a model, as described by example below.

Since the additive model fit above appears unable to account for the curvature in the MMP for the fitted values, we felt that SAVE might be able to provide us with a better model. We used the SAVE methodology to infer the dimension of $\mathcal{S}_{y|\mathbf{x}}$ to be two, and obtained two linear combinations of predictors, w_1 and w_2, as an estimate of a basis for $\mathcal{S}_{y|\mathbf{x}}$. A three-dimensional plot of y versus w_1 and w_2 indicated that an interaction term, w_{12}, might also be

important. So, we decided to fit an additive model: $E(y \mid \mathbf{x}) = \alpha + f_1(w_1) + f_2(w_2) + f_{12}(w_{12})$. Smoothing splines with (the S plus default) four degrees of freedom were used to estimate the f functions. A GMMP for the fitted values for this model is shown in Figure 3. The plot shows the (black) mean estimate under F lying inside the mean estimate band under $\widehat{M}$ (formed from the (gray) mean estimates under G^t). There is little evidence in *this* plot to suggest that the fitted model can be improved.

However, there is evidence from a MMP for one of the original predictors, inversion base height, that this model too could be improved. Again, the discrepancy between the marginal mean estimates in this plot (not shown) is difficult to assess relative to the variability in the data. The corresponding GMMP in Figure 4 allows this discrepancy to be evaluated visually, and the plot reinforces the supposition that the model could possibly be improved (at least for low values of inversion base height).

Having applied Hastie and Tibshirani's methodology to these data, GMMPs appear to offer a quick and easy way to graphically check models. The Gibbs sampling only needs to be done once for each model; with Hastie and Tibshirani's S plus code this is straightforward. The analyst can then cycle through a variety of GMMPs to get some guidance on whether (and how) an alternative model might provide an improvement. For example, in the above analysis, a next step might be to develop a model that deals with low values of inversion base height more satisfactorily, say by increasing the degrees of freedom for the smoothers in the additive model, or by trying different smoothers such as *loess*.

Does a GMMP suffer the same shortcoming as Hastie and Tibshirani's Figure 2; namely, would we be able to obtain equivalent information by plotting pointwise posterior intervals instead of individual posterior realizations? The answer to this question would surely be yes, were it not for the fact that it is not clear how such intervals might be defined in practice. For example, posterior intervals could be calculated for the fitted values in an additive model by summing the posterior intervals for the individual functions in the model. It would then be straightforward to plot the pointwise intervals on a MMP for the fitted values. But, for MMPs for any other function h, it is unclear what pointwise posterior intervals should be defined to be. One possibility would be to smooth the pointwise upper and lower limits for the fitted values using the same method as used to obtain the mean estimates under F and $\widehat{M}$, but it is not clear that this will give us pointwise posterior intervals for $E_{\widehat{M}}(y \mid h)$.

3. OTHER POTENTIAL APPLICATIONS

Returning to Hastie and Tibshirani's Figure 1, can *these* plots (of partial residuals versus individual predictors) be used for model checking? The answer to this question would appear to be no. The black curves are smooths of the partial residuals, $\mathbf{f}_j = S_j \mathbf{r}_j$, while the gray curves are the Gibbs posterior realizations, $\mathbf{f}_j^t = S_j \mathbf{r}_j^t + \sigma S_j^{1/2} \mathbf{z}_j^t$. These plots would appear to offer visualization only of the variability in the fitted functions. Appropriate plots for model checking in this context are GMMPs for the individual predictors, as shown for example in Figure 4.

There are other plots used in model checking and regression diagnostics that can be difficult to assess relative to the variation in the data. Some examples include: residual plots; CERES plots, which are a generalization of partial residual plots and were introduced by Cook (1993); net effect plots, which aid in assessing the contribution of a selected predictor to a regression and were introduced by Cook (1995). The ideas discussed above would appear to have a rôle to play in the analysis of such plots. Work is in progress on these issues, as well as on developing supplementary Bayesian inference methodology.

4. MISCELLANEA

Hastie and Tibshirani's procedure appears to live up to its claim of modularity and generality. Although the procedure derives from the backfitting algorithm for fitting additive models, it could probably be applied fairly easily to other families of models such as generalized linear models. Whether the procedure could also be described as conceptually simple is perhaps more open to debate. For example, choosing priors for the variance components is far from trivial, and MCMC convergence should always be checked in practice. That said, there is clearly a wealth of potential applications for the posterior samples generated with this technique.

SAVE techniques can be applied using *Arc* (Cook and Weisberg, 1999), a comprehensive regression program. Information about the program is available at the Internet site `www.stat.umn.edu/arc`.

ACKNOWLEDGMENT

Research Supported in part by National Science Foundation.

Comment

Alan E. Gelfand

This generous manuscript offers much food for thought. I admire its generality, the ability to handle both Gaussian and non-Gaussian likelihoods, to accommodate both nonparametric and semiparametric forms, to handle a broad range of smoothers. I applaud its pragmatic stance. Much energy is invested on details of the model fitting, worrying about efficiency of algorithms and introducing useful approximations. Finally, I appreciate its effort to be comparative, frequently linking Bayesian, empirical Bayes and classical perspectives, attempting a bridging of ideologies within a rich regression setting.

But then, what is there that is worth commenting upon? I will focus on two main issues. First, I believe the authors are a bit too casual in their Bayesian formulation. There is confusion throughout with regard to singular versus improper priors, with regard to proper versus improper posteriors. If the contribution is viewed as primarily algorithmic, so be it. But if the claim is that legitimate Bayesian inference is being implemented, that credible posterior analysis is being provided, then I have reservations.

Second, at least in the Gaussian case, within the authors' general objectives, there seems to be no need to introduce iterative simulation. A direct simulation formulation is straightforward, avoiding all concerns with MCMC model fitting.

To my first point, at the outset (Section 2), we begin with a prior on $\boldsymbol{\theta}$ which induces a prior on $\mathbf{f} = B\boldsymbol{\theta}$. If the dimension of $\boldsymbol{\theta}$ is less than that of $\mathbf{f}$, a proper Gaussian prior on $\boldsymbol{\theta}$ induces a singular but proper distribution on $\mathbf{f}$. An improper distribution on $\boldsymbol{\theta}$ necessarily yields an improper distribution on $\mathbf{f}$. For smoothers, it is apparently more typical that the dimension of $\boldsymbol{\theta}$ is greater than that of $\mathbf{f}$ and that an improper (*not* degenerate) prior is implicit for $\boldsymbol{\theta}$, hence an induced improper prior for $\mathbf{f}$. Thus, these priors are *not* normal. Expression (5) should be written as

$$(1) \qquad \mathbf{f} \mid \tau^2 \propto \frac{1}{(\tau^2)^a} \exp\bigl(-\mathbf{f}^T K \mathbf{f}/2\tau^2\bigr),$$

where the power a is arbitrary since the distribution is improper. Similarly,

$$(2) \qquad \boldsymbol{\theta} \mid \tau^2 \propto \frac{1}{(\tau^2)^a} \exp\bigl(-\boldsymbol{\theta}^T \Omega \boldsymbol{\theta}/2\tau^2\bigr).$$

If we start with (2), adopting a specific generalized inverse B^- and operate formally [since (2) is improper] we obtain

$$(3) \qquad \mathbf{f} \mid \tau^2 \propto \frac{1}{(\tau^2)^a} \exp\bigl(-\mathbf{f}^T B^{-T} \Omega B^- \mathbf{f}/2\tau^2\bigr),$$

thus determining K.

Introduction of K^-, Ω^-, and $S(\lambda)^-$ [e.g., expression (7)] serves to cloud matters. Ultimately, in (6), if $\mathbf{f} \mid \mathbf{y}$ is proper, $I + \lambda K$ is full rank and $S(\lambda) = (I + \lambda K)^{-1}$ (which doesn't appear until halfway through Section 4.2) is all we need. If $\mathbf{f} \mid \mathbf{y}$ is not proper, how can we speak of its posterior? In this case, when the dim of $\boldsymbol{\theta}$ exceeds the dimension of $\mathbf{f}$, an improper posterior distribution for $\boldsymbol{\theta}$ may induce a unique proper posterior for $\mathbf{f}$. See, for example, Gelfand and Sahu (1999) in this regard.

When we move to Section 3, the situation becomes even more disturbing. Now, p functions are introduced additively in the mean structure, along with an intercept. A centering constraint is introduced on each function "for identifiability." In fact, from a Bayesian point of view, with a proper posterior, there is no identifiability issue (as in, e.g., Lindley, 1971). With improper priors, such constraints are customarily introduced to achieve proper posteriors as well as to provide well-behaved posteriors, yielding well-behaved MCMC algorithms. The recommended "centering on-the-fly," that is, after each iteration, has become standard in these situations. See, for example, Besag, Green, Higdon and Mengersen (1995).

In any event, while the Bayesian backfitting method is presented as an algorithm, it does not appear to correspond to a Gibbs sampler for a well-defined Bayes model. The existence of proper full conditionals for all model unknowns says nothing about the propriety of the joint posterior (see, e.g., Casella and George, 1992). I cannot see how a proper posterior can be associated with the model in (14), using the various improper priors proposed for the $\mathbf{f}_j$. Furthermore, what is the prior on α? Why isn't it updated in the sampler? What is the distributional justification for inserting $\bar{y}$ into (16)? In

Alan E. Gelfand is Professor, Department of Statistics, University of Connecticut, Storrs, Connecticut 06269-3120 (*e-mail:* alan@stat.uconn.edu).

this spirit, the updating using the approximation in (29) can be implemented, perhaps as a Metropolis step, but it is not a draw from the full conditional distribution for θ_i.

As to my second point, the authors concede that fixing σ^2 in (1), or, more generally in (14), is clearly inappropriate, as is fixing the various τ_j^2. The improper prior for σ^2 in (22) along with the improper choices discussed for the τ^2's will surely produce an improper posterior, following Hobert and Casella (1996). Such priors, along with the improper priors for $\mathbf{f}_j$, appear to run counter to the authors' claim in Section 7 that, "We have restricted ourselves to the use of proper priors," citing Hobert and Casella in this regard! The essentially improper prior in (21) in place of (22) cannot be expected to yield a convergent MCMC algorithm if the posterior is improper using (22).

Moreover, even if priors are introduced for σ^2 and the τ_j^2 such that a proper posterior is ensured, the Gibbs sampler in this setting will be very slow. Apart from the usual convergence concerns, at each iteration each S_j must be updated since λ_j changes with each iteration.

An alternative specification enables direct simulation of the entire posterior and permits S_j to be constant for each simulation. Convergence and updating problems vanish. Suppose we retain the same first-stage specification as in (14) but take the prior on $\{\mathbf{f}_j\}$ to be of the form

$$\{\mathbf{f}_j\} \mid \sigma^2 \propto \exp\left\{-\left(\Sigma \mathbf{f}_j^T \widetilde{K}_j \mathbf{f}_j\right)/2\sigma^2\right\}, \tag{4}$$

adding a proper $N(\alpha_0, \sigma^2/\lambda_0)$ prior for α. Here $\widetilde{K}_j = \lambda_j K_j$ with λ_j specified but not σ^2. With the λ_j fixed, $\widetilde{K}_j$ need only be computed once at the outset. We may view λ_j as a "relative precision." Again, if (4) is improper, there is no unique choice for the power of σ^2 in the proportionality constant, as in (1)–(3).

However, once this power is provided, if σ^2 follows an inverse gamma prior, we may factor the joint posterior density for α, $\{\mathbf{f}_j\}$ and σ^2 as

$$p(\alpha, \{\mathbf{f}_j\}, \sigma^2 \mid \mathbf{y}) = p(\alpha, \{\mathbf{f}_j\} \mid \sigma^2, \mathbf{y}) \cdot p(\sigma^2 \mid \mathbf{y}). \tag{5}$$

If (5) is proper, both densities on the right-hand side are. In fact, the first is an updated normal, the second an updated inverse gamma. Sampling σ^{2*} from $p(\sigma^2 \mid \mathbf{y})$ and then, $\alpha^*, \{\mathbf{f}_j^*\}$ from $p(\alpha, \{\mathbf{f}_j\} \mid \sigma^{2*}, \mathbf{y})$ directly provides a posterior realization. This approach appears to fit well with the authors' goals of pragmatism and efficiency.

In summary, while the ideas in this paper are attractive, the Bayesian modeling and resultant simulation-based inference, which are at its heart, are somewhat uncomfortable.

Comment

Peter J. Green

I warmly congratulate the authors on this paper. I am sure they will succeed in broadening acceptance of the Bayesian paradigm in inference in regression by providing this well-written and accessible treatment of the use of Markov chain Monte Carlo (MCMC) in fitting the important class of (generalized) additive models.

The paper promotes several ideas. It can be interesting and revealing to examine Bayesian analogues of familiar frequentist models and procedures; MCMC is important in Bayesian inference; Gibbs sampling is a convenient general recipe for MCMC; if that isn't available, try Metropolis–Hastings; Gibbs sampling is a close analogue of backfitting. None of these points are individually very original, of course! But it is very appealing to see their combination applied to additive models, with a number of practical details worked out to produce an efficient methodology, especially since Splus software to implement the resulting methodology is provided.

My comments focus on some of these practical details, and some other relations and connections to the proposed methods.

BAYESIAN FUNCTION ESTIMATION

The reader who comes to this work from a background of backfitting in (generalized) additive models rather than experience of inference about functions and surfaces in the Bayesian paradigm might get an impression that this paper is close to the state-of-the-art in Bayesian function estimation. In fact, the authors do not claim this; researchers have

Peter J. Green is Professor, Department of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom (*e-mail:* P.J.Green@bristol.ac.uk).

been investigating and using Bayesian models and MCMC calculations for more complicated situations than this, almost from the earliest days of MCMC in statistics. One could even claim that a driving force in the broader acceptance of practical Bayesian methodology for inference in complex data structures has been that using MCMC methods there was a comparatively trivial computational penalty to be paid in moving on from simple models to more complicated ones (in the case of inference about functions, for example, replacing Gaussian priors on functions by non-Gaussian ones).

MCMC in statistics is generally accepted to have begun in statistical image analysis, and of course nonlinear smoothers, which do not destroy boundaries between objects in the scene, are routinely needed there. In discrete spatial settings, such as arise in pixellized images and region-based geographical and ecological problems, particular use has been made of pairwise-difference priors that are not Gaussian, but have heavier tails; for example

$$p(f) \propto \exp\left\{-\beta \sum_{i\sim j} \phi(f_i - f_j)\right\},$$

where f_i is the true image intensity in pixel i, and $i \sim j$ means the summation is over pairs of neighboring pixels (see Geman and McClure, 1985; Green, 1990; Besag, Green, Higdon and Mengersen, 1995, among many others). Instead of using $\phi(u) = u^2$, taking it to be $|u|$, or something more complicated such as $\log\cosh u$ or $-1/(1+u^2)$ has proved successful in many image restoration contexts. (The apparent connection here with M-estimation and robustness is not entirely superficial.)

These methods are all for *discrete* spatial problems, whether on lattices or irregular graphs. Nonparametric Bayesian surface fitting methods for *continuous* space include that of Heikkinen and Arjas (1998) for inference on a Poisson intensity, using ideas of model averaging over appropriately defined step functions to yield smooth posterior mean surfaces, and the variogram-based methods of Diggle, Tawn and Moyeed (1998). Turning to regression on more general, nonspatial, covariates, apart from the work of Denison, Mallick and Smith (1998) and Holmes and Mallick (1997) that is mentioned in this paper, and other methods investigated by these researchers and colleagues, there is the interesting approach of Müller, Erkanli and West (1996), based on Dirichlet process priors.

BAYES FROM CONVICTION OR CONVENIENCE?

The conceptual connection between smoothing methods, especially those based on penalized likelihood, and Bayesian formulations of inference about functions is often made, but there always seems to be an implicit or explicit warning: "Don't take this too literally." Formal use of such connections is rarely made, a notable exception being Wahba's important 1983 paper, exploiting the Bayesian connection to construct confidence intervals about spline estimates. However, the frequentist properties of such intervals are well known to be problematical.

This is all well understood by the authors, but I do think that if they seek to use the Bayesian paradigm—and presenting credible intervals for functions is certainly doing that—they should try to take it a bit more seriously! There are several issues here.

The first concerns the use of the roughness penalty as a negative-log-prior on the regression function. Does the usual penalty $\lambda \int [f''(x)]^2\,dx$ have particular merit in this context, over other (say) quadratic forms in f that are zero for constant or first-order f, or is the choice driven, as in straightforward smoothing, by the computational advantages? Specifically, do autocorrelations decline with lag in a reasonable way, and what about homogeneity of variance? (Answers to these questions are complicated by the partially improper nature of the prior in this case.) Are Gaussian prior assumptions reasonable, or should we consider heavy-tailed modifications? Choice of functional form of the penalty has far greater consequences for the Bayesian procedure than for the simple smoother, as we are going to use it to generate much more subtle inferences, invested, presumably, with real probabilistic interpretations. To be fair, the authors are in good company in not raising these questions; they are almost never considered by anybody else either!

The second issue is the treatment of smoothing (tuning) parameters (or, equivalently, variance components). This is explicitly discussed, in Section 4 of the paper. I must say that I find the full Bayesian version much more compelling than either of the alternatives.

A third concern is about sensitivity to prior assumptions, always problematic in hierarchical models. It would be good to see at least an empirical study of the effect on posterior inference of variations in the authors' assumptions. I would anticipate that everything about the inference except the posterior means is actually rather sensitive.

THE ROOT-S APPROXIMATION

I cannot be the only reader to worry about the quality of the approximation to $S^{1/2}$, proposed in Algorithm A.1. The approach looks simplistic: why should the size of the first neglected term of the series be a reliable guide to the size of the sum of *all* neglected terms? and indeed a small numerical experiment bears out this concern.

Although an alternative, superior, approach using Cholesky decomposition is available for the cubic spline smoother, this case provides a convenient choice for a numerical check. Taking $x_i = i$, $i = 1, 2, \ldots, n = 100$ and drawing a single vector $\boldsymbol{z} \sim N(0, I)$, I compared $\|S^{1/2}\boldsymbol{z} - S_{\mathrm{HT}}^{1/2}\boldsymbol{z}\|$ with the tolerance on $\|S\boldsymbol{z}''\|$ used in the authors' algorithm, where $S_{\mathrm{HT}}^{1/2}$ represents their approximation. Table 3 shows that, especially if higher precision is sought, the method becomes both expensive and much less successful.

Are better approximations available for an amount of work comparable, say, to Algorithm A.1 with a 10^{-2} tolerance? If not, I wonder if there is merit in turning the approach around, and taking $S^{1/2}$ to be some convenient (symmetric) linear smoother, and defining S to be its square, that is, the smoother obtained by applying $S^{1/2}$ twice? This obviously changes the prior covariance structure so we would have to revisit that question. However, there is a clear potential for saving computational effort.

To a rough degree of approximation, there can be surprisingly little change; for example, with $x_i = i$, $i = 1, 2, \ldots, n = 100$ again, S_{10}^2 is close to $S_{8.07}$ where S_{df} is the cubic spline smoother with df degrees of freedom; the eigenvalues differ by a maximum of about 0.05, and their outputs are visually nearly indistinguishable.

TABLE 3
Relationship between tolerance on first neglected term, number of terms included and overall precision for Algorithm A.1 ($\|\cdot\|$ *is the sup norm*)

$df = 6$, $\lambda = 2493.2$			$df = 12$, $\lambda = 105.55$		
$\|S\boldsymbol{z}''\|$	final b	$\|S^{1/2}\boldsymbol{z} - \mathbf{S}_{\mathbf{HT}}^{\mathbf{1/2}}\boldsymbol{z}\|$	$\|S\boldsymbol{z}''\|$	final b	$\|S^{1/2}\boldsymbol{z} - S_{\mathrm{HT}}^{1/2}\boldsymbol{z}\|$
10^{-1}	2	0.100	10^{-1}	2	0.139
10^{-2}	5	0.073	10^{-2}	6	0.105
10^{-3}	25	0.032	10^{-3}	25	0.069
10^{-4}	95	0.024	10^{-4}	175	0.042
10^{-5}	620	0.015	10^{-5}	847	0.042
10^{-6}	2875	0.010	10^{-6}	2702	0.043

SOME MCMC DETAILS

There are several ideas for improving MCMC performance in the literature that could be beneficial in the present context. For example, the modifications to backfitting introduced at the end of Appendix A appear to be related to the ideas of "hierarchical centering" in normal linear mixed models, of Gelfand, Sahu and Carlin (1995).

Organizing variables in a Gibbs sampler into blocks to be simultaneously updated is a common strategy, and often pays off if the blocked variables are highly correlated and the multivariable update is not expensive to implement. The authors' backfitting strategy is precisely an example of this idea. The variables $(f_j(x_{ij}))$ in the jth block will indeed be strongly correlated. Many questions of MCMC strategy about blocking, updating schedules and reparameterization, precisely for the present case of multivariate Gaussian models, are discussed by Roberts and Sahu (1997), which is strongly recommended reading.

A posteriori, there are of course also strong correlations between some variables in different blocks, especially if the predictors are highly correlated themselves, and the anticipated impact of this on MCMC convergence provides another explanation for poor performance in this case.

In other contexts, the correlation between variance parameters and their associated sums-of-squares has proved damaging for MCMC performance; a commonly successful work-around has been to integrate out the variance parameter (assuming we are in the usual conjugate setting with normal random effects and inverse gamma hyperpriors) and then update the random effects by Metropolis–Hastings targetted at a t density. This approach may be useful if the sampler based on equation (24), for example, should mix slowly in a particular case.

Regarding the approach to generalized additive models in Section 6, I have also been a great enthusiast (in other contexts) for Metropolis–Hastings based on a Gaussian approximation to the full conditional. People tell me this is not a free lunch, however. Whether the resulting chain is even geometrically ergodic is, I think, not fully understood. In similar problems, this depends on the relative size of the tails of the proposal and target densities (see, e.g., Roberts and Tweedie, 1996), and so geometric ergodicity may be problematical in skew cases such as GAMs. Can the authors reassure me? Is the sampler provably good?

Finally, brief mention should be made of the required length of MCMC runs. The authors are

rather silent on this; a passing reference in the Discussion seems to suggest that only 100 sweeps were made, after burn-in, in one example. This seems very few for such a large-dimensional problem, especially if full posterior distributions are being estimated, not just their means. In all their emphasis on order-n computation, the authors do not allow for any increase in MCMC convergence times with n.

Rejoinder

Trevor Hastie and Robert Tibshirani

We knew we were taking a chance writing a paper with "Bayesian" in the title, since neither of us work in this area, and we have not paid our dues by attending a Valencia meeting. Visions of the Bayesian–Frequentist battles that have raged over the years in the Royal Statistical Society journals left many a sleepless night. We breathed a sigh of relief when Professor Green's discussion arrived; we appear to have got off lightly with a few small raps on the knuckles from a well-respected applied Bayesian. The discussion of Professor Cook and Mr Pardoe did no damage either. Just as we began to relax, the computer screen started to quiver, and after a quick degauss, there it was! The discussion of Professor Gelfand had arrived in all its fury.

We thank all the discussants for their contributions. All three were complimentary in their opening paragraphs about our pragmatic approach to the problem, and we thus feel that our main mission in writing this paper was accomplished. We will address their comments and concerns separately.

Professor Green gives a very useful history of MCMC and its use in Bayesian function estimation and image smoothing.

He is quite right; we have not spent much time investigating other priors, and realize, especially in spatial statistics and signal processing, that there are many other considerations besides smoothness. In (1) below we express the prior in a slightly different format, which allows perhaps for relatively easy tailoring for function approximation.

The prior assumptions on the individual effects in model (27) clearly have an important impact on the model. Without priors to pin them down, θ_i and V_i are strongly aliased. Consider an individual with bone measurements above the curve near the growth spurt (e.g., girl 124 in Figure 10). She could either have a positive value for θ_i (and $V_i = 0$) or a negative value for V_i (and $\theta_i = 0$) or values in between for both. Priors can save the day, since we have some idea from many different sources of the distribution of the onset of puberty in girls.

It seems our root-S approximation is not too precise; truth be told, our Splus implementation, `gibbs.gam()`, handles smoothing splines only, where this approximation is not needed.

We are reassured by Professor Green's endorsement of our blocking and efficiency strategy outlined in the Appendix, and grateful for his providing more details and references. Unfortunately, we cannot vouch at this time for the geometric ergodicity of our Metropolis–Hastings sampler for GAMs, but expect such results to be forthcoming.

Professor Cook and Mr. Pardoe provide some interesting graphical techniques for model assessment. Almost surely an additive model is an approximation to the truth, so we are not surprised by the small discrepancy in their Figure 1 between the MMP and the additive fit. One has to decide whether the gains obtained by fitting a more complex model (based on projections) is worth the sacrifice in simplicity. One small concern: since typically smoothers are not projection operators ($\|SS\mathbf{y}\| \leq \|S\mathbf{y}\|$), we wonder whether the bands in the GMMP are artificially narrow due to double smoothing?

Professor Gelfand reproaches us for our *vague* description of the prior distribution in Section 2, motivated by smoothing splines. Our goal in the section was to avoid details and inspire generalities; in fact, we purposely postponed details for smoothing splines till the Appendix. Professor Gelfand then attempts a more precise statement of the prior distribution (apparently to correct our treatment for smoothing splines in the Appendix), but does not get it quite right.

For simplicity we assume the n values x_i are distinct. When a smoothing spline is represented in an M-dimensional B-spline basis, with $M = n + 2$, the coefficients $\boldsymbol{\theta}$ have an M-dimensional prior distribution which is both:

- Improper in a two-dimensional subspace, corresponding to constant and linear functions of x;

• Degenerate or singular in a two-dimensional subspace, corresponding to the two natural boundary conditions.

So our statement in the the second bulleted item below (43) in the Appendix does have errors, but *not* the errors claimed by Professor Gelfand. $\boldsymbol{\theta}$ has both a degenerate and improper prior distribution.

It appears that Professor Gelfand's (1) is simply a more precise way of stating our (5). That does not make our (5) incorrect; we admit it is sloppy, but it appears to be the style used by many Bayesian authors.

There is a better way of expressing the prior distribution for smoothing splines or similar methods. Let $K = UDU^T$ be an eigen-decomposition of the $n \times n$ penalty matrix K [see (5) and below in our article; also Green and Silverman (1994) for a detailed description of K and an algorithm for computing it]. The null-space of K is two-dimensional, and spans the column space of $(\mathbf{1}, \mathbf{x})$ (linear functions of x). Suppose we partition $U = (U_1 : U_2)$ such that U_1 spans this null space, and the diagonal matrix $D = \text{diag}(D_1, D_2)$ with the 2×2 matrix $D_1 = 0$. U_2 is $n \times (n-2)$ and represents nonlinear functions in x. U_2 can be represented as a linear transformation of the $n \times M$ B-spline matrix B, which imposes (a) the natural boundary conditions, (b) orthogonality to U_1 and (c) orthogonal columns. The same transformation applied to the M B-spline basis functions $b_j(x)$ yields the $n-2$ *Demmler–Reinch* basis functions $h_\ell(x)$, $\ell = 1, \ldots, n$ [and U_2 is the matrix of sample realizations of these $h_\ell(x)$].

This leads to a representation for the smoothing spline model:

$$f(x) = \alpha_0 + \alpha_1 x + \sum_{\ell=1}^{n-2} h_\ell(x)\beta_\ell. \tag{1}$$

The parameters are divided into $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n-2})$. The prior on $\boldsymbol{\alpha}$ is noninformative, and $\boldsymbol{\beta} \sim N(\mathbf{0}, \tau^2 D_2)$ is proper. This is commonly referred to as a *mixed effects* model, with $\boldsymbol{\alpha}$ regarded as a fixed effect. This representation makes explicit the proper and improper parts of the prior for smoothing splines and similar models and avoids the degeneracy due to overparametrization. The roughness (as computed by the second-derivative penalty) of the $h_\ell(x)$ increases with ℓ, and the prior variances on the diagonal of $\tau^2 D_2$ decrease toward zero accordingly.

We are not quite sure what is bothering Professor Gelfand in our Section 3. An additive model $\mathbf{f} = \mathbf{1}\alpha + \mathbf{f}1 + \cdots + \mathbf{f}_p$ in which each $\mathbf{f}_j$ includes the constant term has a (p-fold) degeneracy which can be removed by assuming each $\mathbf{f}_j^T \mathbf{1} = 0$. This is all we are doing. The priors are easily modified to accommodate this centering [see (2) below]. Professor Gelfand is concerned about the existence of a proper posterior for the additive model (14). This is most easily demonstrated by extending (1) to the additive case (Lin and Zhang, 1999):

$$f() = \alpha_0 + \boldsymbol{\alpha}_1^T \mathbf{x} + \sum_{j=1}^{p} \sum_{\ell=1}^{n-2} h_{\ell j}(x_j)\beta_{\ell j}, \tag{2}$$

where the $h_{\ell j}$ represent the j different series of Demmler–Reinch basis functions defined separately for each predictor. There is an improper prior on the $p+1$ ("fixed effects") $\boldsymbol{\alpha}$, and proper (and independent) normal priors on all the $\beta_{\ell j}$,

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) \sim N\big(\mathbf{0}, \text{diag}(\tau_1^2 D_1, \ldots, \tau_p^2 D_p)\big).$$

This has the same structure as the one-dimensional smoothing spline and has a proper Gaussian posterior for the same reasons. The composed functions are simple linear combinations of the parameters and have proper posteriors as well.

We do sample the constant in Algorithm 3.1; the index j runs from 0, and step 0 samples the constant from a $N(0, \sigma^2/n)$. Possibly Professor Gelfand was misled by (16), which is simply an iterative algorithm (backfitting) for computing the posterior mean.

Although the additive model can be sampled without using Gibbs sampling, there is an overhead of $O(2(p+2)n^3)$ computations up front to compute the relevant posterior covariances and diagonalize them. Each realization from the Gibbs sampler takes $O(n)$ computations and can represent a dramatic savings for large problems.

ACKNOWLEDGMENTS

The authors thank Jun Liu for reassuring discussions while preparing this rejoinder. They also thank the editors for arranging this forum.

REFERENCES

BESAG, J., GREEN, P., HIGDON, D. and MENGERSEN, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10** 3–66.

BOWMAN, A. and YOUNG, S. (1996). Graphical comparison of nonparametric curves. *Appl. Statist.* **45** 83–98.

CASELLA, G. and GEORGE, E. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46** 167–174.

COOK, R. D. (1993). Exploring partial residual plots. *Technometrics* **35** 351–362.

COOK, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Physical and Engineering Sciences* Amer. Statist. Assoc., 18–25. Alexandria, VA.

COOK, R. D. (1995). Graphics for studying net effects of regression predictors. *Statist. Sinica* **5** 689–708.

COOK, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York.

COOK, R. D. and LEE, H. (2000). Dimension reduction in binary response regression. *J. Amer. Statist. Assoc.* To appear.

COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction." *J. Amer. Statist. Assoc.* **86** 316–342.

COOK, R. D. and WEISBERG, S. (1997). Graphics for assessing the adequacy of regression models. *J. Amer. Statist. Assoc.* **92** 490–499.

COOK, R. D. and WEISBERG, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.

DIGGLE, P. J., TAWN, J. A. and MOYEED, R. A. (1998). Model-based geostatistics (with discussion). *J. Roy. Statist. Soc. Serv. C* **47** 299–350.

GELFAND, A. E. and SAHU, S. K. (1999). Identifiability, improper priors and Gibbs sampling for generalized linear models. *J. Amer. Statist. Assoc.* **94** 247–253.

GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika* **82** 479–488.

GEMAN, S. and MCCLURE, D. E. (1985). Bayesian image analysis: an application to single photon emission tomography. In *Proceedings of the Statistical Computing Section* 12–18. Amer. Statist. Assoc., Alexandria, VA.

GREEN, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Medical Imaging* **9** 84–93.

HEIKKINEN, J. and ARJAS, E. (1998). Nonparametric Bayesian estimation of a spatial Poisson intensity. *Scand. J. Statist.* **25** 435–450.

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

LIN, X. and ZHANG, D. (1999). Mixed inference in generalized additive models. *J. Roy. Statist. Soc. Ser. B* **61** 381–400.

LINDLEY, D. V. (1971). The estimation of many parameters (with discussion). In *Foundations of Statistical Inference*. (V. P. Godambe and D. A. Sprott, eds.) 435–452. Holt, Rinehart and Winston, Toronto.

MÜLLER, P., ERKANLI, A. and WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79.

PORZIO, G. C. and WEISBERG, S. (1999). Tests for lack-of-fit of regression models. Technical report 634, School Statistics, Univ. Minnesota.

ROBERTS, G. O and SAHU, S. K. (1997). Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B* **59** 291–317.

ROBERTS, G. O. and TWEEDIE, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110.

WAHBA, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.