Biller, Fahrmeir:

Bayesian spline-type smoothing in generalized regression models

Sonderforschungsbereich 386, Paper 31 (1996)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Bayesian spline-type smoothing in generalized regression models

C. Biller and L. Fahrmeir

Institut für Statistik, Universität München,
Ludwigstr. 33, D-80539 München

Summary

Spline smoothing in non- or semiparametric regression models is usually based on the roughness penalty approach. For regression with normal errors, the spline smoother also has a Bayesian justification: Placing a smoothness prior over the regression function, it is the mean of the posterior given the data. For non-normal regression this equivalence is lost, but the spline smoother can still be viewed as the posterior mode. In this paper, we provide a full Bayesian approach to spline-type smoothing. The focus is on generalized additive models, however the models can be extended to other non-normal regression models. Our approach uses Markov Chain Monte Carlo methods to simulate samples from the posterior. Thus it is possible to estimate characteristics like the mean, median, moments, and quantiles of the posterior, or interesting functionals of the regression function. Also, this provides an alternative for the choice of smoothing parameters. For comparison, our approach is applied to real-data examples analyzed previously by the roughness penalty approach.

1 Introduction

Let us first consider classical curve estimation for metrical bivariate data (y_i, x_i), $i = 1, \ldots, n$, with strictly ordered covariate values $x_1 < \ldots < x_n$. It

is assumed that responses y_i depend on x_i by

$$y_i = f(x_i) + \epsilon_i \quad , \tag{1}$$

with i.i.d. errors $\epsilon_i \sim N(0, \sigma^2)$ and a regression function f to be estimated from the data.

The roughness penalty approach makes the compromise between faith with the data and smoothness explicit: Find f as a twice-differentiable function that minimizes the penalized sum of squares

$$PS(f) = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 + \lambda \int \left(f''(u)\right)^2 du \quad , \tag{2}$$

with the smoothing parameter $\lambda > 0$ controlling the trade-off between fit and smoothness. The minimizing function $\hat{f}$ is a cubic smoothing spline, see Reinsch (1967) or e.g., Green and Silverman (1994).

Wahba (1978) showed that (2) has a Bayesian justification by placing a smoothness prior over f as the solution of the stochastic differential equation

$$\frac{d^2 f(x)}{dx^2} = \lambda^{-1/2} \sigma \frac{dW(x)}{dx} \quad , \quad x \geq x_1 \quad . \tag{3}$$

In (3), $W(x)$ is a standard Wiener process with $W(x_1) = 0$, independent of the errors ϵ_i.

Initial conditions for $x = x_1$ are diffuse, i.e.

$$\left(f(x_1), f^{(1)}(x_1)\right)' \sim N(0, kI) \tag{4}$$

with $k \to \infty$. Then, the cubic smoothing spline $\hat{f}$ is the posterior mean of f given the data, i.e.

$$\hat{f} = E(f|y) \tag{5}$$

for $k \to \infty$. This equivalence can also be established for more general classes of spline functions, see e.g. Kohn and Ansley (1987). These authors also derive a stochastic difference equation from (3) and apply Kalman filtering and smoothing for efficient computation of the smoothing spline $\hat{f}$.

For non-normal responses, the observation equation (1) will be defined by a non-normal regression model, for example a logit model in the case of binary responses. Accordingly, the sum of squares in (2), which is essentially the log-likelihood for normal responses, is replaced by a sum of log-likelihoods $l_i(y_i|f(x_i))$ for non-normal responses. This leads to the penalized log-likelihood criterion

$$PL(f) = \sum_{i=1}^{n} l_i(y_i|f(x_i)) - \frac{1}{2}\lambda \int \left(f''(u)\right)^2 du \to \max_f \quad . \tag{6}$$

The solution $\hat{f}$ is again a cubic spline smoother, see Green and Silverman (1994).

However, the Bayesian justification as the posterior mean (5) is lost, since the posterior is no longer Gaussian: The cubic spline smoother $\hat{f}$ can now be seen as the posterior mode given the data. From this point of view, a Bayesian analysis that allows for wider inference is obviously desirable and provides motivation for the full Bayesian approach developed and discussed in this paper.

A direct approach to evaluate posteriors via Bayes' theorem would involve computationally intractable high-dimensional integrations. Therefore, we use Markov Chain Monte Carlo (MCMC) simulations to draw samples from the posterior. Based on these samples, estimates of means, medians, quantiles and other characteristics can be computed, without assuming any normality approximation for the posterior. In addition, Bayesian data-driven choice of smoothing parameters is carried out simultaneously. As a further advantage, the Bayesian formulation also allows to estimate posterior distributions of any functionals, e.g. maxima or minima, of regression functions.

Our focus is on non- and semiparametric analysis of generalized additive models. Section 2 gives a Bayesian framework for these models, including the more general case of polynomial splines of order $2m-1$ instead of cubic splines. In Section 3 we describe MCMC algorithms for simulation-based inference. The proposed sampling schemes are close to those in Knorr-Held (1995) and Fahrmeir and Knorr-Held (1996), developed in the related context of dynamic generalized linear models. For the case of a single covariate a somewhat different suggestion has recently been made by Shephard and Pitt (1995). Carter and Kohn (1995) discuss MCMC sampling for robustified models of the form (1), with mixtures of normals as error distributions. However, their sampling schemes are not applicable in our context. Section 4 contains applications to data sets analyzed previously by the roughness penalty approach. Extensions to other kinds of regression, e.g. varying coefficient models, are outlined in the concluding remarks in Section 5.

2 Bayesian models for generalized additive regression

Consider now the regression situation where observations $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, on a response y and on a vector $x = (x_1, \ldots, x_p)$ of covariates are given. The response y may be non-normal, e.g. nonnegative or discrete. To simplify presentation, we first assume that all covariates are metrical and that covariate values are strictly ordered, i.e. $x_{1j} < \ldots < x_{nj}$ for $j = 1, \ldots, p$.

Generalized additive models assume that, given $x_i = (x_{i1}, \ldots, x_{ip})$, the distribution of y_i belongs to an exponential family with mean $\mu_i = E(y_i|x_i)$

linked to an additive predictor η_i by an appropriate response function h, i.e.

$$\mu_i = h(\eta_i) \quad , \quad \eta_i = \gamma + f_1(x_{i1}) + \ldots + f_p(x_{ip}) \quad . \tag{7}$$

To estimate the regression functions $f_1, \ldots, f_p$, the roughness penalty approach (6) is generalized to the penalized log-likelihood criterion

$$PL(f_1, \ldots, f_p) = \sum_{i=1}^{n} l_i(y_i|\eta_i) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \int \left(f_j^{(m)}(u)\right)^2 du \to \max_{f_1,\ldots,f_p} \quad , \tag{8}$$

with likelihood contributions l_i from $y_i|x_i$, and with separate penalty terms and smoothing parameters. The maximizing functions $\hat{f}_1, \ldots, \hat{f}_p$ are polynomial splines of order $2m-1$, e.g. cubic splines for $m=2$. The constant γ is added to guarantee uniqueness of the smoothers in the backfitting algorithm, see Hastie and Tibshirani (1990) or Fahrmeir and Tutz (1994, ch.5).

For a Bayesian formulation, (7) together with a specific choice of the exponential family defines the observation model. It is supplemented by placing smoothness priors over the regression functions similarly as in (3). We make the following assumptions:

For $j = 1, \ldots, p$, the regression function f_j obeys the stochastic differential equation

$$L^m f_j(x) = \sigma_j \frac{dW_j(x)}{dx} \quad , \quad x \geq x_{1j} \tag{9}$$

with $L^m = d^m/dx^m$ as the m-th order differential operator. The standard Wiener processes $W_j(x)$, with $W_j(x_{1j}) = 0$, are mutually independent.

Initial conditions are

$$(f_j(x_{1j}), f_j^{(1)}(x_{1j}), \ldots, f_j^{(m-1)}(x_{1j}))' \sim N(0, k_j I) \quad , \tag{10}$$

becoming diffuse for $k_j \to \infty$. We also assume a normal or diffuse prior, independent from (9), for the constant γ.

The spline smoothers $\hat{f}_j$ obtained from (8) with $\lambda_j = 0.5/\sigma_j^2$ can then be viewed as posterior mode estimators. In an empirical Bayes approach, σ_j^2 will be regarded as an unknown constant that can be estimated from the data, e.g. by cross-validation or by maximum likelihood. Here we adopt a full Bayesian model and impose independent inverse gamma priors

$$\sigma_j^2 \sim IG(a_j, b_j) \quad , \quad j = 1, \ldots, p \quad , \tag{11}$$

on the variances. By appropriate specification of hyperparameters a_j, b_j, these priors can be made more or less informative. Posterior estimation of σ_j^2 then provides an alternative data-driven choice of smoothing parameters.

In order to compute Bayesian spline-type smoothers based on MCMC simulations from the posterior, we reformulate (9) as a stochastic difference equation for the vector $f_j(x_{1j}), \ldots, f_j(x_{nj})$ of evaluations of f_j. For the case

of a normal regression model (1) with a single covariate, such a derivation is already given in Kohn and Ansley (1987) and previous work of these authors. Since this derivation has nothing to do with the observation model, we can exploit their results for the present purpose. For $j = 1, \ldots, p$, we define

$$\alpha_{ij} = (f_j(x_{ij}), f_j^{(1)}(x_{ij}), \ldots, f_j^{(m-1)}(x_{ij}))' \quad , \quad i = 1, \ldots, n \quad .$$

Then the sequence $\alpha_{1j}, \ldots, \alpha_{nj}$ obeys the stochastic difference equation

$$\alpha_{ij} = F_{ij}\alpha_{i-1,j} + \sigma_j u_{ij} \quad , \quad i = 2, \ldots, n \quad . \tag{12}$$

The $(m \times m)$- transition matrices F_{ij} are given by

$$F_{ij} = \begin{pmatrix} 1 & \delta_{ij} & \cdots & \cdots & \frac{\delta_{ij}^{m-1}}{(m-1)!} \\ & 1 & \delta_{ij} & \cdots & \frac{\delta_{ij}^{m-2}}{(m-2)!} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \delta_{ij} \\ & & & & 1 \end{pmatrix}$$

with $\delta_{ij} = x_{ij} - x_{i-1,j}$. The errors u_{ij} are independent and normal

$$u_{ij} \sim N(0, Q_{ij}) \quad ,$$

with elements $Q_{ij}(k,l)$ of Q_{ij} given by

$$Q_{ij}(k,l) = \frac{\delta_{ij}^{2m-k-l+1}}{(2m-k-l+1)(m-k)!(m-l)!} \quad , \quad k,l = 1, \ldots, m \quad .$$

According to (10), initial values α_{1j} have a diffuse prior. For $m = 2$, corresponding to cubic smoothing splines, we have $\alpha_{ij} = (f_j(x_{ij}), f^{(1)}(x_{ij}))'$, and transition and covariance matrices are

$$F_{ij} = \begin{pmatrix} 1 & \delta_{ij} \\ 0 & 1 \end{pmatrix} \quad , \quad Q_{ij} = \begin{pmatrix} \delta_{ij}^3/3 & \delta_{ij}^2/2 \\ \delta_{ij}^2/2 & \delta_{ij} \end{pmatrix} \quad .$$

The stochastic difference equation (12) then defines a smoothness prior on a sequence of "parameters" $\alpha_{1j}, \ldots, \alpha_{nj}$ that is equivalent to the one obtained from (9). The generalized additive observation model (7) together with (12) is now similar in form to dynamic generalized linear models (see, e.g., Fahrmeir and Tutz, 1994, ch.8 for a survey). Therefore we may conceive MCMC sampling schemes based on suggestions and experience made in this related area.

The restriction to strictly ordered covariate values made at the beginning of this section can be easily dropped. First, for each covariate $x_j, j = 1, \ldots, p$,

the observed covariate values are ordered. Then $x_{1j} \leq x_{2j} \leq \ldots \leq x_{ij} \leq \ldots \leq x_{nj}$ for each component, where x_{ij} now denotes the i-th covariate value in the ordered sequence. The stochastic process prior (9) and its representation by the stochastic difference equation (12) remain formally unchanged, even in the presence of tied covariate values. If, for example, $x_{i-1,j} = x_{i,j}$, then $\delta_{ij} = 0$, $F_{ij} = I$ and $Q_{ij} = 0$, so that also $\alpha_{i-1,j} = \alpha_{ij}$. For computational purposes as in the next section, it is more convenient to group observations with same covariate value x_{ij}, say. Here grouping is realized separately for each covariate. Therefore, after relabeling, one gets ordered values $x_{1j} < \ldots < x_{rj} < \ldots < x_{n_j,j}$, $n_j < n$. With the number w_{rj} of repetitions of covariate value x_{rj} we define the grouped response

$$y_{rj} = \frac{1}{w_{rj}} \sum_{s=1}^{w_{rj}} y_i^s$$

as the mean of the responses y_i^s, $s = 1, \ldots, w_{rj}$, with same covariate value x_{rj}. Doing so, we remain within the exponential family framework. To connect the original covariate values x_{ij}, $i = 1, \ldots, n$, with the ordered and grouped values x_{rj}, $r = 1, \ldots, n_j$, we use the $n \times n_j$ incidence matrix N_j, with entries $N_j^{i,r} = 1$ if $x_{ij} = x_{rj}$, and 0 otherwise (see Green and Silverman, 1994, p. 65).

In Section 3, we always assume that data are ordered and grouped in the described way. The definitions of the quantities α_{ij}, F_{ij}, ... made above remain unchanged, if we use the index i instead of r also for the grouped data, but with i now running from 1 to n_j.

Another extension concerns partial splines, where the predictor has semiparametric additive form

$$\eta_i = \alpha + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \beta' z_i \quad .$$

For example, z_i may contain binary or categorical covariates, or the effect of some metrical covariate is supposed to be linear. To deal with such models, we add normal or diffuse priors for the components of β.

3 Posterior analysis by MCMC sampling

Bayesian inference for the unknown parameters $\alpha_j = (\alpha_{1j}, \ldots, \alpha_{ij}, \ldots)'$, σ_j^2, $j = 1, \ldots, p$, is based on joint or marginal posterior distributions like $p(\alpha_j|y)$, $p(\alpha_{ij}|y)$ or $p(\sigma_j^2|y)$. Direct evaluation of posteriors generally becomes computationally intractable due to high-dimensional integrations. Markov Chain Monte Carlo (MCMC) methods circumvent this problem by drawing samples indirectly. Estimates of the posteriors and functionals like moments and quantiles are available from these samples. Recent expositions of MCMC are given, e.g., in Besag, Green, Higdon and Mengersen (1995), Tierney (1995) and in the first chapter of Gilks et. al. (1996).

The key tool for the design of MCMC techniques is the definition of so-called full conditionals, i.e. conditional distributions for a part of the parameters given the rest and the data. For example $p(\alpha_j|\alpha_k, k \neq j, \sigma_1^2, \ldots, \sigma_p^2, y)$, $p(\alpha_{ij}|\alpha_{lj}, l \neq i, \alpha_k, k \neq j, \sigma_1^2, \ldots, \sigma_p^2, y)$, $p(\sigma_j^2|\alpha, \sigma_k^2, k \neq j, y)$ are such conditional distributions. The full set of such conditionals defines an ergodic Markov Chain on the state space of parameters with marginal posteriors as limiting distributions. Starting from some initial values, a sequence of samples drawn from the full conditional will then converge in distribution to the marginal posteriors, for example to $p(\alpha_j|y)$, $p(\alpha_{ij}|y)$, or $p(\sigma_j^2|y)$. Obviously, two properties are essential for designing efficient MCMC schemes: Firstly, samples from the conditionals should be available in a computationally efficient way. Secondly, the constructed Markov Chain should possess good convergence properties. Both goals are possibly conflicting, and some compromise will often be useful.

Our proposals for MCMC samplings for generalized additive models are based on suggestions and experience made in the related field of state space models and dynamic generalized linear models. Carlin, Polson and Stoffer (1992) first suggested the use of Gibbs sampling for state space models, and Fahrmeir, Hennevogl and Klemme (1992) adopted their method to dynamic generalized linear models. More general MCMC schemes for this class of models are proposed in Knorr-Held (1995, 1996), Gamerman (1995) and Shephard and Pitt (1995). These authors also discuss the important issue of single versus block moves.

In the following, we describe a single move sampling scheme and outline a generalization to block moves. For the derivation of full conditionals it is useful to note that the joint priors for $\alpha_j, j = 1, \ldots, p$ are multivariate normal with

$$p(\alpha_j|\sigma_j^2) \propto \exp(-\frac{1}{2\sigma_j^2}\alpha_j^{'} K_j \alpha_j) \quad . \tag{13}$$

The penalty matrix K_j in (13) is symmetric and block-tridiagonal :

$$K_j = \begin{pmatrix} K_{11} & K_{12} & & & \\ K_{21} & K_{22} & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & K_{n_j-1,n_j} \\ & & & K_{n_j,n_j-1} & K_{n_j,n_j} \end{pmatrix}$$

with blocks

$$\begin{aligned} K_{11} &= F_{2j}^{'} Q_{2j}^{-1} F_{2j} \\ K_{ii} &= Q_{ij}^{-1} + F_{i+1,j}^{'} Q_{i+1,j}^{-1} F_{i+1,j} \quad , i = 2, \ldots, n_j - 1 \\ K_{i,i-1} &= K_{i-1,i}^{'} = -Q_{ij}^{-1} F_{ij} \quad , i = 2, \ldots, n_j \\ K_{n_j,n_j} &= Q_{n_j,j}^{-1} \end{aligned}$$

compare Fahrmeir and Tutz (1994, p.265).

Single moves update the components α_{ij} of α_j one at a time. For a more compact notation, we suppress the conditioning variables $\sigma_1^2, \ldots, \sigma_p^2$ and the data y. Then

$$p(\alpha_{ij}|\alpha_{rj}, r \neq i, \alpha_k, k \neq j) \propto l(\alpha_{ij}) p(\alpha_{ij}|\alpha_{rj}, r \neq i) \quad . \tag{14}$$

In (14), the first factor $l(\alpha_{ij})$ denotes the joint likelihood of all responses that are observed at the (same) covariate value x_{ij}. For untied covariate values, $l(\alpha_{ij})$ reduces to the exponential family density $p(y_i|\alpha_{ij})$ of y_i. The second factor is the conditional prior for α_{ij} given the remaining components. Since the joint prior (13) is normal, it is multivariate normal,

$$p(\alpha_{ij}|\alpha_{rj}, r \neq i) \propto \phi(\alpha_{ij}|\mu_{ij}, \sigma_j^2 \Sigma_{ij})$$

with mean μ_{ij} and covariance matrix $\sigma_j^2 \Sigma_{ij}$, say. Following Carlin, Polson and Stoffer (1992) or Fahrmeir and Tutz (1994, Section 8.3.2), we have

$$\mu_{ij} = \begin{cases} (F_{2j}' Q_{2j}^{-1} F_{2j})^{-1} F_{2j}' Q_{2j}^{-1} \alpha_{2j} & , \quad i = 1 \\ (Q_{ij}^{-1} + F_{i+1,j}' Q_{i+1,j}^{-1} F_{i+1,j})^{-1} (Q_{ij}^{-1} F_{ij} \alpha_{i-1,j} + F_{i+1,j}' Q_{i+1,j}^{-1} \alpha_{i+1,j}) & \\ & , \quad i = 2, \ldots, n_j - 1 \\ F_{ij} \alpha_{i-1,j} & , \quad i = n_j \end{cases}$$

and

$$\Sigma_{ij} = \begin{cases} (F_{2j}' Q_{2j}^{-1} F_{2j})^{-1} & , \quad i = 1 \\ (Q_{ij}^{-1} + F_{i+1,j}' Q_{i+1,j}^{-1} F_{i+1,j})^{-1} & , \quad i = 2, \ldots, n_j - 1 \\ Q_{ij} & , \quad i = n_j \end{cases} \quad .$$

(Note that α_{1j} has a diffuse normal prior with $Q_{1j}^{-1} = 0$.)

For updating a current value α_{ij}^c, we use Metropolis-Hastings steps with a conditional independence proposal $\alpha_{ij}^* \sim \phi(\alpha_{ij}^c | \mu_{ij}^c, (\sigma_j^c)^2 \Sigma_{ij}^c)$, with "$c$" denoting values available at the current iteration. Then the acceptance probability of α_{ij}^* is

$$a(\alpha_{ij}^*, \alpha_{ij}^c) = \min \left\{ \frac{l(\alpha_{ij}^*)}{l(\alpha_{ij}^c)}, 1 \right\} \quad .$$

This updating procedure was introduced by Knorr-Held (1995) in the context of dynamic generalized linear models. Other updating schemes like those in Gamerman (1995) and Shephard and Pitt (1995) may also be used instead. However, they require more CPU time, since score functions and information matrices have to be evaluated in every update step.

A drawback of single move schemes is that convergence can be slow if neighboring parameters are highly correlated. This is likely to happen if

the likelihood $l(\alpha_{ij})$ contains little information, as for example for binary responses. Then block move schemes as the one outlined below are generally preferable.

The other parameters are updated more conventionally. For the constant γ, we have

$$p(\gamma|\alpha_1,\ldots,\alpha_p,y) \propto p(y|\alpha_1,\ldots,\alpha_p,\gamma)p(\gamma) \quad ,$$

where $p(y|\cdot)$ is the likelihood of all observations and $p(\gamma)$ a normal diffuse prior. MH steps with a simple random walk proposal are used for updating.

With inverse Gamma priors (11) for $\sigma_1^2,\ldots,\sigma_p^2$, posteriors are again inverse Gamma

$$\sigma_j^2|\alpha_j,y \sim IG(a_j^{'},b_j^{'}) \quad ,$$

with hyperparameters

$$\begin{aligned} a_j^{'} &= a_j + \frac{m}{2}(n_j-1) \\ b_j^{'} &= \left(\frac{1}{b_j} + \frac{1}{2}\sum_{i=2}^{n_j} u_{ij}^{'} Q_{ij}^{-1} u_{ij}\right)^{-1} \end{aligned}$$

and $u_{ij} = (\alpha_{ij} - F_{ij}\alpha_{i-1,j})/\sigma_j$. Therefore, $\sigma_1^2,\ldots,\sigma_p^2$ can be directly updated by Gibbs sampling.

For block moves, the vector α_j is divided into several blocks $\alpha_j = (\alpha_{(1)j}, \ldots,\alpha_{(r)j},\ldots,\alpha_{(s)j})$, say, each block $\alpha_{(r)j}$ consisting of several components α_{ij} of α_j. Then, instead of updating single components α_{ij} one at a time, blocks $\alpha_{(r)j}$, $r = 1,\ldots,s$, are updated. The idea for block moves is that corresponding likelihoods $l(\alpha_{(r)j})$ will contain more information, leading to less autocorrelation and better convergence. The conditionals $p(\alpha_{(r)j}|\cdot)$, given the rest of parameters, have a similar structure as in the case of single moves:

$$p(\alpha_{(r)j}|\cdot) \propto l(\alpha_{(r)j})\phi(\alpha_{(r)j}|\mu_{(r)j},\sigma_j^2\Sigma_{(r)j}) \quad .$$

Here, $l(\alpha_{(r)j})$ denotes the joint likelihood of all responses with densities depending on (components of) $\alpha_{(r)j}$, and ϕ is multivariate normal with $\mu_{(r)j}$ and $\Sigma_{(r)j}$ depending on neighboring blocks. Explicit formulae can be derived from the joint multivariate prior (13) with similar arguments as for single moves. Updating of blocks is done by MH steps with proposal densities $\phi(\alpha_{(r)j}|\cdot)$, in analogy to single moves, see Knorr-Held (1996). Updating of γ and $\sigma_1^2,\ldots,\sigma_p^2$ remains unchanged.

A different proposal, based on multivariate normals, centered near posterior modes, is suggested in Shephard and Pitt (1995), but computational efforts are distinctly higher.

4 Applications

For comparison, we apply the suggested Bayesian smoothing techniques to data sets already analyzed by other approaches.

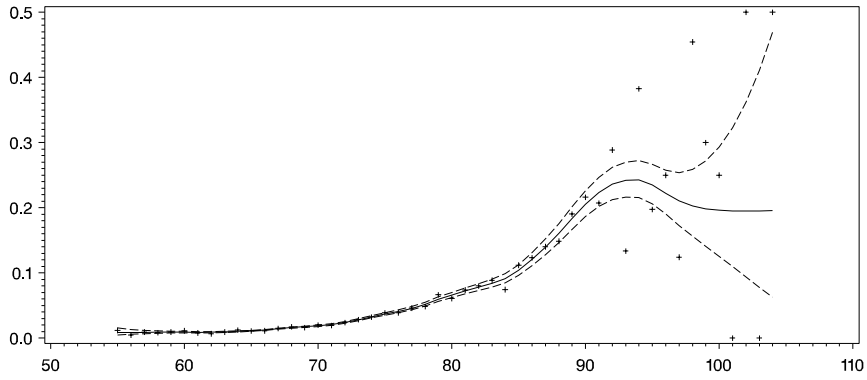

Figure 1: Posterior mean estimates (solid line) and pointwise two standard deviation confidence bands together with crude death rates (plus)

Application 1: Smoothing mortality tables

Green and Silverman (1994, pp. 101–104) smooth crude death rates $y_x = d_x/n_x$ for a population of retired American white females with cubic splines, using the roughness penalty approach. Here n_x is the size of the population at risk at age x, and d_x is the number of corresponding recorded deaths. Figure 1 shows crude death rates, from age 55 to 104, together with a Bayesian smoother. Since n_x becomes rather small for higher age, varying between 2 and 11 for $x \geq 98$, a binomial logit model

$$Ey_x = \mu_x = \frac{\exp(f(x))}{1 + \exp(f(x))}$$

is used, with smoothness prior $d^2 f(x)/dx^2 = dW(x)/dx$, corresponding to cubic spline type smoothing. Figure 1 shows the posterior mean smoother with ± 2 standard deviation pointwise confidence bands, based on single move MCMC sampling with 100.000 iterations and a burn-in period of 2500 iterations. After the burn-in period every 10 th sample is used to estimate posterior means and variances. The average acceptance rate in the MH steps was 0.91. Hyperparameters for the inverse Gamma prior for σ^2 were set to $a = 1$, $b = 200$, corresponding to a very flat prior. The posterior mean was estimated by $\hat{\sigma}^2 = 0.0037$. Shephard and Pitt (1995) analyzed the same data set with their computationally more demanding block move scheme. Comparing results, the methods yield more or less identical smoothers. This shows that the simple single move MCMC updating scheme described in Section 3 may give satisfactory results.

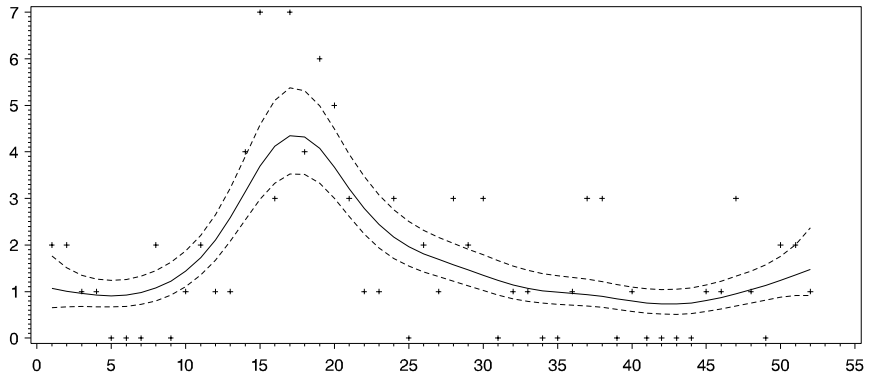

Figure 2: Posterior mean estimates (solid line) and pointwise one standard deviation confidence bands from single move together with observations (plus)

Application 2: Lymph node syndrome incidence

Figure 2 contains observations y_t of the weekly incidence of muco-cutaneous lymph node syndrome (MCLS) in Tottori-prefecture in Japan during 1982.

This data set is analyzed by Kashiwagi and Yanagimoto (1992), assuming a dynamic loglinear Poisson model

$$\lambda_t = \exp(f(t))$$

and a first order random walk $f(t) = f(t-1)+u(t)$ as smoothness prior. They obtain a posterior mean estimate based on numerical integrations similarly as in Kitagawa (1987).

We take a cubic spline type prior (9) and an inverse Gamma priori $\sigma^2 \sim IG(a,b)$ with $a = 1$, $b = 200$. For comparison, single and block move schemes are used to estimate posterior moments and quantiles. Figure 2 displays the estimated posterior mean and ± 1 standard deviation confidence band obtained from taking every 10 th sample out of 100.000 iterations after a burn-in period of 2500 iterations. Samples and autocorrelations for selected values α_i are seen in Figure 3.

Figure 4 gives corresponding results for block moves of block size 3, displaying posterior means and medians that are almost identical and very close to the posterior means for single move sampling. This seems to indicate that single moves may do their job quite well. Figure 5 contains block move samples and autocorrelations for the same selected values α_i. Here a distinct improvement can be seen in comparison to Figure 3. The graphs indicate that block moves might yield better mixing and convergence behaviour.

The following application is to be considered as a benchmark example: The response is (almost) purely binary, and there are three metrical covariates

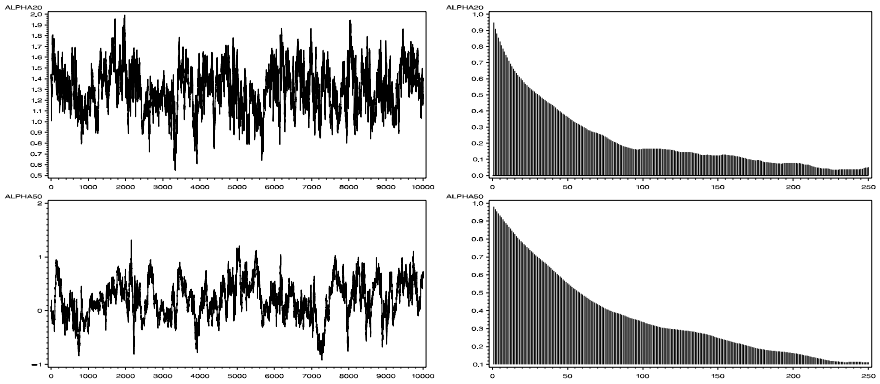

Figure 3: Single move samples (left) and autocorrelations (right) for α_{20} (top) and α_{50} (bottom)

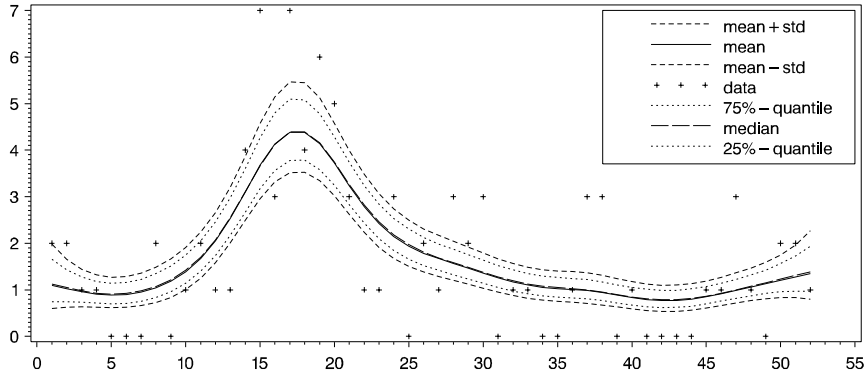

Figure 4: Posterior mean estimates and median together with pointwise one standard deviation confidence bands and 50% credible regions from block move.

with unknown functional forms $f_1(x_1), f_2(x_2)$ and $f_3(x_3)$ to be estimated from sparse data.

Application 3: Kyphosis in laminectomy patients

To illustrate the use of generalized additive modelling, Hastie and Tibshirani (1990, Section 10.2) analyze data on 83 patients undergoing corrective spinal surgery. The response y of interest is the presence or absence of kyphosis, defined to be the forward flexion of the spine of at least 40 degrees from vertical, following surgery. Risk factors are age in months (x_1), the starting

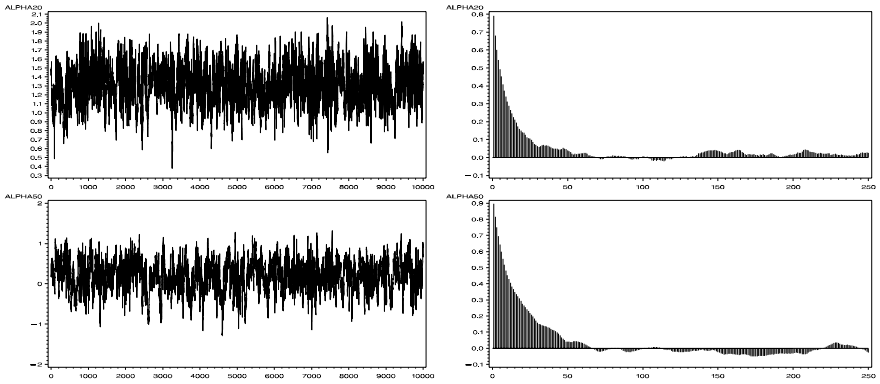

Figure 5: Block move samples (left) and autocorrelations (right) for α_{20} (top) and α_{50} (bottom)

vertebrae level of the surgery (x_2) and the number of vertebrae level involved (x_3). They fit an additive logit model for $P(y = 1|x)$, with predictor $\eta(x) = \gamma + f_1(x_1) + f_2(x_2) + f_3(x_3)$, using the roughness penalty approach with cubic splines. We analyze the data with Bayesian cubic spline-type priors (9) for the functions $f_j(x_j)$.

In contrast to the previous two examples, mixing and convergence behaviour detoriates distinctly, regardless whether single or block move schemes are used. Figure 6 shows estimated posterior means $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ together with confidence bands based on a burn-in period of 2500 iterations and a sampling period of 100.000 iterations taking every 10 th sample.

Hyperparameters for the variance priors were chosen as in the foregoing examples. Comparing the estimates with those in Hastie and Tibshirani (1990) and looking at confidence bands, it is seen that the overall pattern of estimated regression curves is similar and will lead to analogous interpretations given there. However, details of curves differ more distinctly than for the previous examples. In view of our experience with this sparse data set, it is surely worthwile to develop and investigate alternative sampling schemes to improve mixing and convergence, but this is beyond the scope of this paper.

5 Concluding remarks

Semiparametric Bayesian smoothing as discussed here has some attractive features compared to the roughness penalty approach: It provides a natural framework for Bayesian analysis beyond posterior modes, and MCMC techniques allow to estimate posterior means, medians, quantiles and other functionals of regression functions. No asymptotic normality approximations have to be assumed. Bayesian data-driven choice of smoothing parameters is

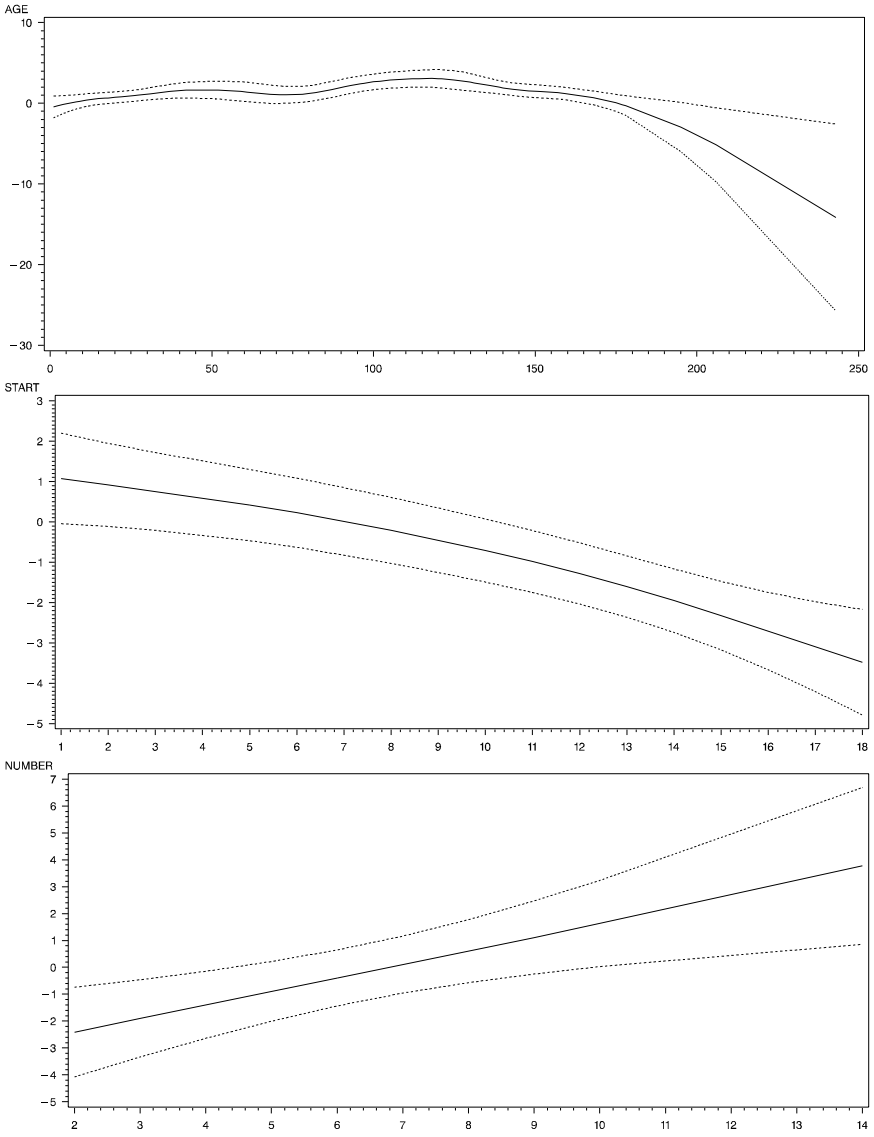

Figure 6: Posterior mean estimates (solid line) and pointwise one standard deviation confidence bands of the risk factors age, start and number.

automatically incorporated. A further advantage of MCMC techniques, not considered in this paper, is the convenient handling of missing values.

To some extent, the flexibility of MCMC techniques is also a certain weakness: The general Metropolis-Hastings algorithm allows for a wide variety of proposals for updating steps, at least in theory. As our application 3 shows, further experience is needed for developing and investigating sampling schemes that are efficient in sparse data situations.

We focussed on generalized additive models and polynomial spline-type smoothing, but modifications and extensions to other models are possible by other choices of observation models and smoothness priors. In particular, the approach can be extended to varying coefficient models and regression models for survival and event history data. More general splines, e.g. log-splines, can be considered as in Kohn and Ansley (1987). Introduction of mixtures of normals as in Carter and Kohn (1995) instead of normal errors in smoothness priors, is an appropriate device to detect jumps or discontinuities in regression functions. We intend to consider some of these topics in future work.

Acknowledgement: This work was supported by a grant from the German National Science Foundation, Sonderforschungsbereich 386. Special thanks go to Leo Knorr-Held for providing a prelimenary version of his doctoral thesis.

References

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995). Bayesian Computation and Stochastic Systems, *Statistical Science* **10**(1), 3–66.

Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992). A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling, *J. A. Statist. Assoc.* **87**(418), 493–500.

Carter, C. K. and Kohn, R. (1995). Robust Bayesian nonparametric regression, Preprint, Australian Graduate School of Management, Univ. of New South Wales.

Fahrmeir, L. and Knorr-Held, L. (1996). Dynamic discrete-time duration models, *Discussion Paper 14*, Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer–Verlag, New York.

Fahrmeir, L., Hennevogl, W. and Klemme, K. (1992). Smoothing in dynamic generalized models by Gibbs sampling, *Advances in GLIM and Statistical Modelling*, Springer Verlag, New York.

GAMERMAN, D. (1995). Monte Carlo Markov Chains for Dynamic Generalized Linear Models, *Discussion paper*, Instituto de Matemática, Universidade Federal do Rio de Janeiro.

GILKS, W. R., RICHARDSON, S. AND SPIEGELHALTER, D. J. (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

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

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

KASHIWAGI, N. AND YANAGIMOTO, T. (1992). Smoothing Serial Count Data Through a State–Space Model, *Biometrics* **48**, 1187–1194.

KITAGAWA, G. (1987). Non-Gaussian state-space modelling of nonstationary time series, *J. A. Statist. Assoc.* **82**(400), 1032–1063.

KNORR-HELD, L. (1995). Markov Chain Monte Carlo Simulation in Dynamic Generalized Linear Mixed Models, *Discussion Paper 8*, Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München.

KNORR-HELD, L. (1996). *Bayesian Hierachical Modelling of Discrete Longitudinal Data*, Dissertation, Universität München. forthcoming.

KOHN, R. AND ANSLEY, C. F. (1987). A New Algorithm For Spline Smoothing Based On Smoothing A Stochastic Process, *SIAM J. Sci. Stat. Comput.* **8**(1), 33–48.

REINSCH, C. (1967). Smoothing by spline functions, *Numerische Mathematik* **10**, 177–183.

SHEPHARD, N. AND PITT, M. K. (1995). Parameter–Driven Exponential Family Models, Preprint, Nuffield College, Oxford

TIERNEY, L. (1994). Markov Chains for exploring Posterior Distributions, *Ann. Statist.* **22**(4), 1701–1762.

WAHBA, G. (1978). Improper Priors, Spline Smoothing and the Problem of Guarding against Model Errors in Regression, *J. R. Statist. Soc. B* **40**(3), 364–372.