Biller, Fahrmeir:

Bayesian Varying-coefficient Models using Adaptive Regression Splines

Sonderforschungsbereich 386, Paper 206 (2000)

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

Bayesian Varying–coefficient Models using Adaptive Regression Splines

Clemens Biller and Ludwig Fahrmeir

Sonderforschungsbereich 386, Institute of Statistics

Ludwig Maximilians University Munich

Ludwigstr. 33, 80539 Munich, Germany

email: biller@stat.uni-muenchen.de, fahrmeir@stat.uni-muenchen.de

June 14, 2000

Abstract

Varying–coefficient models provide a flexible framework for semi– and nonparametric generalized regression analysis. We present a fully Bayesian B–spline basis function approach with adaptive knot selection. For each of the unknown regression functions or varying coefficients, the number and location of knots and the B–spline coefficients are estimated simultaneously using reversible jump Markov chain Monte Carlo sampling. The overall procedure can therefore be viewed as a kind of Bayesian model averaging. Although Gaussian responses are covered by the general framework, the method is particularly useful for fundamentally non–Gaussian responses, where less alternatives are available. We illustrate the approach with a thorough application to two data sets analyzed previously in the literature: the kyphosis data set with a binary response and survival data from the Veteran's Administration lung cancer trial.

Keywords: B–spline basis; knot selection; non–Gaussian response; non– and semiparametric regression; reversible jump Markov chain Monte Carlo.

1 Introduction

Generalized linear models (GLM, see McCullagh and Nelder, 1989) and extensions provide a unified framework for exploring the relation between a response variate y_i and a vector $x_i = (x_{i1}, \ldots, x_{ip})$ of covariates observed for $i = 1, \ldots, n$ individuals. They relate the expectation $\mu_i = \mathrm{E}(y_i|x_i)$ to a predictor η_i through the relation $\mu_i = h(\eta_i)$, where h is a response function. Classical parametric GLM's assume a linear predictor $\eta_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p$. Various non– and semiparametric extensions have been proposed to generalize parametric GLM's. Varying–coefficient models (VCM, Hastie and Tibshirani, 1993) comprise many other models as special cases. They are defined by a predictor of the form

$$\eta_i = x_{i1}f_1(r_{i1}) + \ldots + x_{ip}f_p(r_{ip}), \tag{1}$$

where $r_{i1}, \ldots, r_{ip}$ are metrical covariates and $f_1, \ldots, f_p$ are unspecified functions to be estimated nonparametrically. The covariates $r_{i1}, \ldots, r_{ip}$ can be interpreted as effect modifiers, since the effects of $x_{i1}, \ldots, x_{ip}$ vary through the functions $f_1, \ldots, f_p$. Semiparametric or partially linear models

$$\eta_i = f_1(r_{i1}) + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p \tag{2}$$

and generalized additive models (Hastie and Tibshirani, 1990)

$$\eta_i = f_1(r_{i1}) + \ldots + f_p(r_{ip}). \tag{3}$$

are obtained as special cases.

For the modeling and estimation of the functions f_j there exist some alternatives. Hastie and Tibshirani (1993) consider a penalized log–likelihood approach, where the smoothness of the f_j is controlled by a penalty term using a separate smoothness parameter for each f_j. The simultaneous data–driven selection of the smoothing parameters is so time consuming for more than one or two functions f_j, that usually it is not done. Instead, the smoothness is determined by the degree of freedom of the smoothing matrices (see Hastie and Tibshirani, 1990 and 1993). The estimates $\hat{f}_j$ of this approach are given as weighted cubic smoothing splines. An alternative mentioned in Hastie and Tibshirani (1993) as good choice for modeling the varying effects f_j are regression splines, which are defined as a linear combination of a vector of unknown basis coefficients and a vector of known basis functions. These basis functions are defined by a vector of knots, that lie

within the support of the respective effect modifier r_{ij}. The shape and smoothness of f_j here only is determined by the number and the location of these knots. There are some advantages of regression splines when compared to smoothing splines. First, regression splines need only few knots and also only few unknown basis coefficients, for example 5 to 10, while smoothing splines are defined with one knot for each distinct value of the effect modifier r_{ij}, resulting in a large number of unknown parameters. Another advantage is the fact that regression splines define an ordinary linear predictor, so that all standard inferential tools for GLM's can be used. However, one obstacle with regression splines has been the choice of the number and the location of the knots. Only minor changes in these parameters may cause major differences in the fitted functions $\hat{f}_j$. Eubank (1988, Section 7.2) pointed out that finding the right number and location of knots by visual inspection of the data is impossible in most cases. Therefore data–driven methods for adaptive knot placement are needed for (in some sense) nearly optimal estimators $\hat{f}_j$. Frequentist approaches (see e.g. Friedman and Silverman, 1989, or Stone, Hansen, Kooperberg and Truong, 1997) use first forward steps to add knots which are optimal with respect to some chosen criterion (for example Rao statistics) and afterwards delete knots in backward steps using another criterion (for example the AIC criterion). The results of these approaches are only optimal with respect to the chosen selection criterions. Bayesian approaches using Markov chain Monte Carlo (MCMC) techniques, however, are more general. During all iterations both the number and the location of the knots may vary. Hence, the uncertainty in the knot placement is taken into consideration and the estimation of the regression splines in each iteration of the algorithm is based on different knot settings. The final estimator then is built as the mean of the estimators in each iteration resulting in a great flexibility of the estimated spline function. Smith and Kohn (1996) proposed a Bayesian approach for univariate curve fitting and additive models with normal response using Gibbs sampling. In each iteration of their algorithm significant knots are chosen from a set of candidate knots by Bayesian variable selection. A Bayesian approach for univariate curve fitting with normal response using reversible jump Markov chain Monte Carlo (RJMCMC, see Green, 1995) is presented by Denison, Mallick and Smith (1998). They suggest a kind of hybrid algorithm: In each iteration they choose the set of knots by RJMCMC methods, but given these knots the unknown function is estimated by the usual

least squares approach. They also extend this approach to additive models, but due to the use of the least squares method they need backfitting in each iteration. Mallick, Denison and Smith (2000) proposed Bayesian multivariate adaptive regression splines (BMARS) for the GLM. They emphasize that "the Bayesian MARS method is just an extension in many dimensions of the Bayesian curve fitting methodology given in Denison et al. (1998)." For the extension to the GLM they use a simple Metropolis–Hastings proposal. No example of the convergence properties of the method is given, but they state that the sampler has slow convergence. A fully Bayesian approach for the semiparametric generalized linear model (2) also using RJMCMC for knot selection was presented in Biller (2000). In contrast to Denison et al. (1998), this approach was generally defined for responses from the exponential family, and for estimation of the regression spline given the knots MCMC techniques are used instead of least squares. The approach showed good convergence properties of the reversible jump technique in choosing both the number and location of the knots. A conceptually different type of Bayesian nonparametric inference is based on smoothness priors for unknown functions as a stochastic generalization of penalized likelihood approaches, see Hastie and Tibshirani (2000) or Fahrmeir and Lang (1999) for recent work.

In this paper we present an extension of the adaptive Bayesian regression spline approach for semiparametric GLM's in Biller (2000) to a Bayesian version of the VCM (1). Due to the use of MCMC techniques to estimate the spline given the knot placement, such an extension is possible without the need for backfitting in each iteration as in Denison et al. (1998). This fully Bayesian approach via MCMC has several advantages: First, RJMCMC is tailor–made for adaptive regression splines. Second, different building blocks are easier to combine or to extend, as the extension of the approach of Biller (2000) in this paper shows. Other possibilities are mentioned in the conclusions, e.g., the incorporation of random effects. Further, we do not need to rely on asymptotics. For example to define confidence regions of the splines, we only have to compute the 0.05 and the 0.95 quantiles of the generated sample of a function f_j to get a 90% confidence region for that f_j. In a similar way, any functionals of the model may be estimated. These functionals are simply computed in each iteration of the algorithm to create samples of the functionals. From these samples any probability statements of interest can be estimated. Also, the complete

MCMC output is available for model diagnostics.

The paper is organized as follows: The Bayesian varying–coefficient model is defined in Section 2. Section 3 describes, together with a brief introduction of MCMC techniques, the algorithm to estimate this model. In Section 4 the model is applied to known data sets from the literature. Some concluding remarks follow in Section 5.

2 The Bayesian varying–coefficient model

For the definition of the Bayesian varying–coefficient model (BVCM) we use a formulation, that directly combines the special cases (2) and (3) of the VCM (1). Additionally to the covariates $x_{i1}, \ldots, x_{ip}$ with effect modifiers $r_{i1}, \ldots, r_{ip}$ we consider covariates $z_i = (z_{i1}, \ldots, z_{iq})$ with fixed effects $\beta = (\beta_1, \ldots, \beta_q)'$ on the response. Then the BVCM is defined as

$$\eta_i = z_i\beta + x_{i1}f_1(r_{i1}) + \ldots + x_{ip}f_p(r_{ip}). \tag{4}$$

Here we get the classical parametric GLM for $p = 0$, the semiparametric GLM (2) for $p = 1$, and the GAM (3) for $q = 0$ and $x_{ij} \equiv 1$ for all i, j.

Each of the varying coefficients f_j for $j = 1, \ldots, p$ is defined to lie in the k_j–dimensional space of natural cubic splines. That is, with a vector $c_j = (c_{j1}, \ldots, c_{jk_j})'$ of unknown basis coefficients and a vector $B_j = (B_{j1}, \ldots, B_{jk_j})$ of basis functions for the space of natural splines, each f_j can be represented as spline

$$f_j(r_{ij}) = \sum_{l=1}^{k_j} c_{jl}B_{jl}(r_{ij}) = B_j(r_{ij})c_j. \tag{5}$$

The known basis functions $B_{j1}, \ldots, B_{jk_j}$ are computed with a k_j–vector of knots $t_j = (t_{j1}, \ldots, t_{jk_j})$ from the support of each effect modifier r_{ij}. An appropriate choice is the widely used B–spline basis with local support. For details and efficient algorithms for computing this basis see De Boor (1978), Eubank (1988), Schumaker (1993) or Dierckx (1993), and especially for natural splines Lyche and Schumaker (1973) or Lyche and Strøm (1996).

With the basis functions approach for each f_j the predictor (4) of the BVCM is given as GLM

$$\eta_i = z_i\beta + x_{i1}B_1(r_{i1})c_1 + \ldots + x_{ip}B_p(r_{ip})c_p, \tag{6}$$

with constant effects β, $c_1, \ldots, c_p$. As mentioned, the shape and the smoothness of the splines (5) is determined by the number k_j and the location of the knots t_j. An important distinction is that we assume that both k_j and t_j are unknown and have to be estimated together with the constant effects of model (6).

For the joint estimation of the knots t_j and the basis coefficients c_j defining the spline f_j, for each $j = 1, \ldots, p$ we define the following hierarchical model: The number k_j of knots is from some countable set $\mathcal{K}_j$ (which is defined below) and serves as model indicator. Each value of k_j defines a model for the spline f_j, that is determined by the parameters t_j and c_j. In such a hierarchical model we define the model parameter $\theta_{k_j} = (t_j, c_j) \in \mathbb{R}^{2k_j}$, which is combined with the model indicator k_j to build the parameter $\theta_j = (k_j, \theta_{k_j})$ of the spline f_j.

For the Bayesian approach we need a prior specification for each of the unknown parameters. Each of the model indicators k_j for $j = 1, \ldots, p$ is constrained to lie in a set $\mathcal{K}_j = \{k_{j,\min}, k_{j,\min+1}, \ldots, k_{j,\max}\} \subset \mathbb{N}$. Due to the definition of f_j as natural spline, $k_{j,\min}$ is restricted to $k_{j,\min} \geq 4$. As in Biller (2000) we propose three different priors for k_j: A Poisson distribution with parameter λ, but restricted to the set $\mathcal{K}_j$, is a usual and widely used prior in the reversible jump literature, see for example Green (1995) or Denison et al. (1998). Alternatives are a discrete uniform distribution on $\mathcal{K}_j$ and a negative binomial prior with parameters $m = 1$ and $p \in (0, 1)$. The probabilities of the last prior are globally monotonically decreasing in k_j, which avoids too complex models resulting from a prior that favours larger k_j. When compared to the Poisson prior, the two latter priors lead to models with small average numbers k_j of knots. As demonstrated by the examples in Biller (2000) resulting curves are often too smooth. However, in the examples in Section 4 also these two latter priors lead to convincing results.

Given k_j for $j = 1, \ldots, p$ we assume the elements t_j and c_j of the model parameter θ_{k_j} to be independent and treat them separately. The knots t_j are supposed to lie in a discrete set of candidate knots $\mathcal{T}_{j0} = \{t_{j0,1}, t_{j0,2}, \ldots, t_{j0,k_{j,\max}}\}$, which may consist of the sorted distinct values of the effect modifier r_{ij}. An alternative, that is used in the applications in Section 4, is to distribute $t_{j0,1}, \ldots, t_{j0,k_{j,\max}}$ equidistantly over the interval $[r_{\min,j}, r_{\max,j}]$. The prior of t_j is defined by assuming that all possible samples $t_j = (t_{j1}, \ldots, t_{jk_j})$ out of

$\mathcal{T}_{j0}$ have equal probability

$$p(t_j|k_j) = \binom{k_{j,\max}}{k_j}^{-1} = \frac{k_j!(k_{j,\max} - k_j)!}{k_{j,\max}!} \; . \tag{7}$$

Hence, this prior depends only on k_j and $k_{j,\max}$. For the basis coefficients c_j we use a multivariate normal prior distribution $c_j|k_j \sim N_{k_j}(0, \Sigma_{c_j})$, where the covariance matrix is defined as $\Sigma_{c_j} = \sigma^2_{c_j} I_{k_j}$ with a scalar $\sigma^2_{c_j}$.

The fixed effects β are also assumed to be multivariate normal, that is $\beta \sim N_q(0, \Sigma_\beta)$. Here possible correlations between the coefficients $\beta = (\beta_1, \ldots, \beta_q)'$ are modeled by defining $\Sigma_\beta = \sigma^2_\beta R_\beta$ with a scalar σ^2_β and a q–dimensional correlation matrix R_β.

All the parameters $\theta_1, \ldots, \theta_p$ and β are assumed to be pairwise independent and are combined to the joint unknown parameter $\theta = (\beta, \theta_1, \ldots, \theta_p)$. For the estimation of θ we consider the joint posterior distribution

$$p(\theta|y) \propto p(y|\theta)\, p(\beta) \prod_{j=1}^{p} p(\theta_{k_j}|k_j) p(k_j) \tag{8}$$

neglecting the covariates for ease of presentation. The factor $p(y|\theta)$ denotes the likelihood of the response $y = (y_1, \ldots, y_n)$.

3 MCMC estimation techniques

The estimation of the joint unknown parameter θ is done by simulating the posterior (8) with MCMC techniques. They are based on samples from a Markov chain with the distribution of interest as its stationary limiting distribution. Thus, these stochastic simulation methods avoid the necessity of a complete knowledge of the interesting distribution. This allows to simulate from very complex distributions in hierarchical Bayesian models as the posterior (8). The Metropolis–Hastings algorithm (the most general MCMC technique, see for example Gilks, Richardson and Spiegelhalter, 1996) ensures the convergence of the Markov chain against the considered distribution. Here one has to choose an appropriate proposal density $q(\theta, \theta')$, from which a new value θ' can be drawn given the current state θ of the Markov chain. Since this proposal density usually does not agree with the distribution of interest (8), the proposal value θ' is only accepted with a certain probability $\alpha(\theta, \theta')$ as new state of the Markov chain. For more informations about MCMC techniques

see Tierney (1994), Besag, Green, Higdon and Mengersen (1995), Gilks, Richardson and Spiegelhalter (1996), or Gamerman (1997b).

The Metropolis–Hastings algorithm is defined for models with known and fixed dimension of the parameter. However, such an algorithm is not suitable, when the dimension of the interesting parameters is also unknown. This is the case for the posterior (8), where for each spline f_j the model indicator k_j is not known (for $j = 1, \ldots, p$). The reversible jump MCMC algorithm of Green (1995) extends the Metropolis–Hastings technique to such problems with unknown and varying dimensions. Here the model indicators k_j are defined to vary during the iterations leading to different state spaces of the Markov chain with different dimensions, since with k_j the dimension of the model parameter θ_{k_j} varies. For state transitions without a change in dimension, i.e., when k_j does not vary and the transitions take place within one state space, the ordinary Metropolis–Hastings algorithm mentioned above is applicable. For transitions between different state spaces, the method of Green (1995) proposes steps for increasing and reducing k_j. These "birth" and "death" steps have to be defined as related pair of steps, where birth is the reversal of death and vice versa (this feature is called "dimension matching"). For a birth step, that is a transition from $\theta_j = (k_j, \theta_{k_j})$ to $\theta'_j = (k_j + 1, \theta'_{k_j+1})$ with an increase of k_j by 1, we have to create both one new knot and one new basis coefficient. This is done by drawing a two–dimensional random vector u_B independent of θ_j and setting the new proposal θ'_j by an appropriately chosen invertible deterministic function $\theta'_j(\theta_j, u_B)$. The reverse death step from θ'_j to θ_j is accomplished by using the inverse transformation leading to a deterministic proposal.

For the simulation of the joint posterior given in (8) it follows that we have to design different reversible jump steps for the different parts of θ both with and without a change in the dimension of the state space of the Markov chain, leading to a hybrid MCMC algorithm.

For each spline f_j both the number k_j and the location of the k_j knots t_j have to be chosen, what can be done separately for $j = 1, \ldots, p$ by the move types birth and death of a knot and the movement of a knot to another position as proposed by Biller (2000) for the semiparametric model (2) with only one spline. Given the placement of the knots, the estimation of the remaining parameters β, $c_1, \ldots, c_p$ can be done by a standard MCMC

technology for Bayesian GLM's due to the representation (6) of the model. Each iteration of the reversible jump algorithm then consists in the following steps:

(a) Update the fixed effects β by the method of Gamerman (1997a) for GLM's adapted to blocks of fixed effects.

(b) Update the splines f_j separately for $j = 1, \ldots, p$:

1. *Position change:* Move a given knot $t_{j,l}$ to another position (without change in k_j).
2. *Dimension change:* Birth or death of one knot $t_{j,l+1}$, that is, adding or deleting a $t_{j,l+1}$ with changing k_j by 1 and corresponding changes in c_j; the choice between birth and death is done randomly.
3. *Update of basis coefficients:* Update the basis coefficients c_j by the method of Gamerman (1997a) for GLM's adapted to blocks of fixed effects (without change in k_j).

Details of the update of the fixed effects β and the basis coefficients $c_1, \ldots, c_p$ are given in Appendix A. For details of the reversible jump moves *position change* and *dimension change* we refer to Biller (2000), Sections 3.3 and 3.4, which are applied separately to each f_j for $j = 1, \ldots, p$.

4 Applications

This section illustrates the BVCM with two data sets from the literature: the kyphosis data set presented in Hastie and Tibshirani (1990), and the data of the Veteran's Administration lung cancer trial, given in Kalbfleisch and Prentice (1980).

For each data set we used the three alternative prior distributions for the model indicators k_j mentioned in Section 2. The results in Biller (2000) indicate that the prior of k_j has minor influence on the smoothness of f_j provided that there is enough information in the data. However, the prior had influence on the estimation of k_j, where a reasonable convergence and mixing of the chain only was achievable with a Poisson prior (with parameter λ between about 20 and 35), whereas the discrete uniform and the negative binomial (or geometric) prior led to inadequate convergence with small acceptance rates.

In contrast to these results the discrete uniform and the geometric prior with $p \in (0,1)$ lead to a reasonable convergence and mixing of the Markov chains in the applications below. Since with the Poisson and the geometric prior the chosen average number of knots k_j depends on the choice of the hyperparameters λ and p, respectively, we use the discrete uniform prior for k_j for the applications, where no hyperparameter has to be specified. Only for the first example with the kyphosis data we compare the results for the three alternative prior distributions. The results for the second example are similar.

In both examples we compare several models with the deviance information criterion (DIC) defined by Spiegelhalter, Best and Carlin (1998) measuring the fit and the complexity of each model. For the Bernoulli distributed response of the following examples the saturated deviance

$$D(\phi) = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\mu_i}\right)\right]$$

is used (see McCullagh and Nelder, 1989, page 34). With the parameter $\phi = (\beta, c_1, \ldots, c_p)$ only the GLM (6) given the knot placement can be considered. The fit of the respective model is measured by the posterior expectation $\overline{D} = \mathrm{E}_{\phi|y}(D)$ of the deviance. The complexity is given by the effective number of parameters p_D that is defined by the difference of the expected posterior deviance $\overline{D}$ and the deviance computed at the posterior expectation $\overline{\phi} = \mathrm{E}_{\phi|y}(\phi)$ of the parameter, i.e., $p_D = \overline{D} - D(\overline{\phi})$. Hence, p_D is a penalty term that penalizes a better fit by greater complexity. The DIC then is defined as

$$\mathrm{DIC} = \overline{D} + p_D. \tag{9}$$

The algorithm is implemented and performed in C++ on a Windows NT 4.0 personal computer with a 333 MHz Intel Pentium II processor. Based each on 10000 iterations after a burn–in of 5000 iterations the algorithm ran for about 7 and 40 minutes in the applications in Sections 4.1 and 4.2, respectively. The plotted graphs show the median of each sample together with pointwise 90% Bayesian credible regions.

4.1 Kyphosis data

The binary response of the kyphosis data is given as presence (1) or absence (0) of kyphosis, a postoperative deformation that follows a corrective spinal surgery commonly performed in children for tumor and congenital or developmental abnormalities. Kyphosis is

	Model	$\overline{D}$	$D(\overline{\phi})$	p_D	DIC
(1)	$\eta_i = \beta_0 + f_A(A_i) + f_N(N_i) + f_S(S_i)$	56.08	45.99	10.08	66.16
(2)	$\eta_i = \beta_0 + A_i\beta_A + f_N(N_i) + f_S(S_i)$	60.84	52.65	8.19	69.03
(3)	$\eta_i = \beta_0 + f_N(N_i) + f_S(S_i)$	64.53	56.97	7.56	72.08
(4)	$\eta_i = \beta_0 + f_A(A_i) + N_i\beta_N + f_S(S_i)$	56.69	49.14	7.54	64.23
(5)	$\eta_i = \beta_0 + f_A(A_i) + f_S(S_i)$	58.38	50.79	7.59	65.96
(6)	$\eta_i = \beta_0 + f_A(A_i) + f_N(N_i) + S_i\beta_S$	59.88	51.86	8.02	67.89
(7)	$\eta_i = \beta_0 + f_A(A_i) + f_N(N_i)$	69.22	62.22	7.00	76.22

Table 1: Models for analyzing the kyphosis data together with DIC.

defined as forward flexion of the spine of at least 40 degrees from vertical. The data set contains 81 patients of which 17 had kyphosis after the surgery. The available predictors are age in month at time of operation (A), the starting range of vertebrae levels involved in the operation (S), and the number of levels involved (N). A frequentist analysis of the data based on splines is described in Hastie and Tibshirani (1990, Section 10.2).

To analyze the influence of the covariates on the response we fit the seven generalized additive logistic models shown in Table 1. Model 1 uses a regression spline f_j for each of the three predictors, together with an intercept term β_0. In the models 2 to 7 separately for each of the three covariates we either replace the respective nonparametric covariate effect by a linear parametric term or we completely leave it out.

Figure 1 shows the estimates of the nonparametric functions f_A, f_S and f_N for model 1. The plots for the predictors age A and start S have striking nonlinear features, while the effect of number N perhaps also could be modeled by a parametric term with fixed covariate effect.

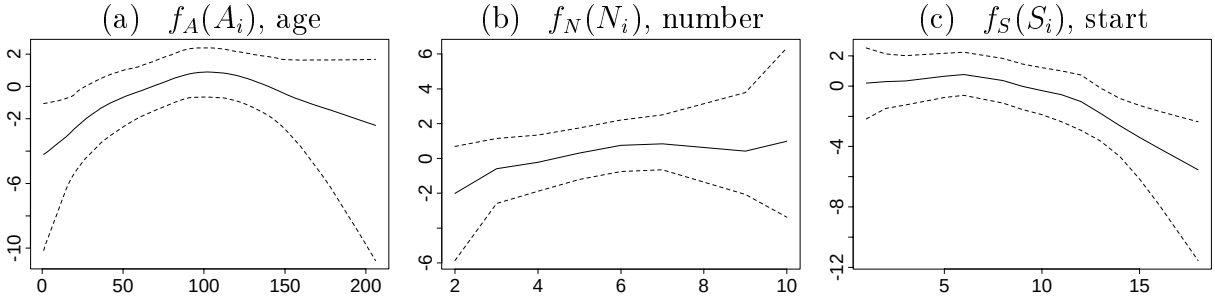

Figure 1: Estimates of splines with 90% Bayesian credible intervals for model 1.

To compare the seven models, Table 1 additionally shows the value of DIC for each model. Similar to the results of Hastie and Tibshirani (1990, Section 10.2, Table 10.1), linearizing or leaving out the covariates age A or start S (models 2, 3, 6 and 7) leads to a worse fit by increasing the DIC when compared to model 1. Hastie and Tibshirani (1990) state that only age and start seem to be important, for which reason they leave number N completely out. Inspection of the deviances given in their analysis leads to the conclusion that a linear effect of covariate N yields the best model. Due to the smallest value of DIC for model 4, this result can also be seen in Table 1 of our analysis, while model 5 with leaving out the covariate N shows the second best fit.

The plots of f_A and f_S for model 4 are very similar to the respective plots of model 1 in Figure 1 and therefore are not shown, while the linear effect of number N in model 4 has median 0.3547 with the 90% Bayesian credible region $(0.0600, 0.7078)$.

As example for estimating the model indicators k_j, Figure 2 gives some details of the sample of k_S for estimating the spline f_S of covariate start S in model 1. The left part of Figure 2 shows the sample k_S with values between 4 and 19. With an acceptance rate of 0.34 for the steps birth and death the mixing over k_S is good. In the middle of Figure 2 there is the frequency of the accepted values of k_S. The mode is at $k_S = 4$, and we see, that in more than one third of the iterations we use a spline f_S with four knots. The right part of Figure 2 depicts the cumulative occupancy fractions $p(k_S < j|y)$ for the different values of k_S against the number of iterations, what is a useful check on the stationarity of k_S. After the burn-in phase these cumulative occupancy fractions stay on a stable level speaking for an adequate length of the burn-in. The samples of the model indicators k_A and k_N for the splines f_A and f_N behave similar and hence are not shown.

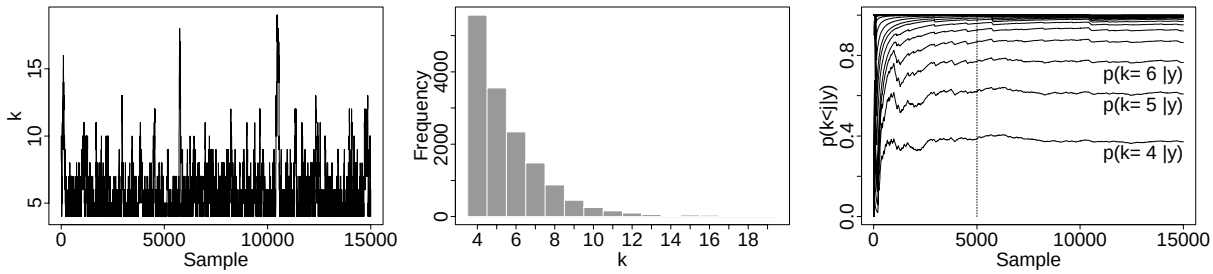

Figure 2: Sample path (left), frequencies (middle) and cumulative occupancy fractions (right) for the samples of the model indicator k_S for estimating the spline f_S of covariate start in model 1.

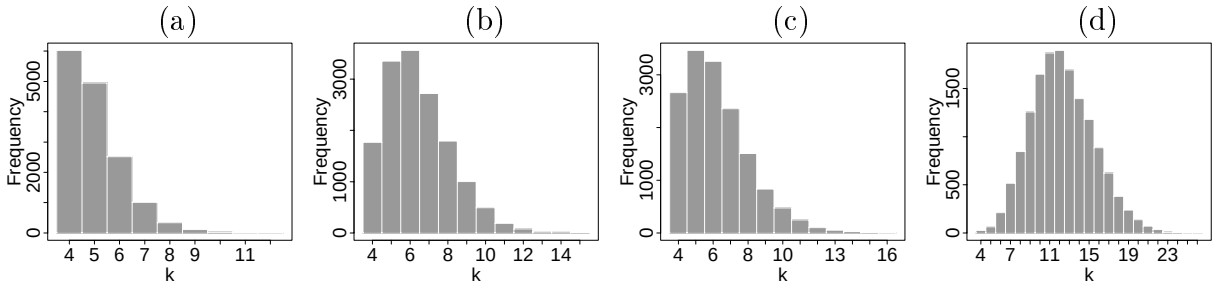

Figure 3: Frequencies of the model indicator k_S in model 1, when using the negative binomial prior with $p = 0.7$ (a) and $p = 0.3$ (b), and the Poisson prior with $\lambda = 10$ (c) and $\lambda = 30$ (d).

To compare the results of different priors for the model indicators, Figure 3 again exemplarily shows the frequencies of the samples of k_S in model 1, when using the negative binomial prior with $p = 0.7$ (a) and $p = 0.3$ (b), and the Poisson prior with $\lambda = 10$ (c) and $\lambda = 30$ (d). Prior (a) has a mode at $k_S = 4$, while reducing the parameter p, for example to $p = 0.3$ as in prior (b), leads to an increase in the mode of k_S to 6. A greater dependence from choosing the hyperparameter is given with the Poisson prior. With $\lambda = 10$ in prior (c) we have a mode of $k_S = 5$, while $\lambda = 30$ yields a mode $k_S = 12$ (d). In the estimation of the splines f_j the first three priors (a) to (c) lead to curves that are very similar to the estimates in Figure 1 for the discrete uniform prior, while the Poisson prior (d) with large hyperparameter $\lambda = 30$ leads to more rough and unsmooth estimates, see also the comments in Biller (2000) regarding the choice of the prior of the model indicators. Regarding the sample paths of k_S and the cumulative occupancy fractions, the four priors (a) to (d) yield very similar behaviours as given in Figure 2 for the discrete uniform prior. Due to the dependence on the hyperparameters of the negative binomial and the Poisson prior, we propose to use the discrete uniform prior for the model indicators k_j, where no hyperparameter has to be chosen by the user and hence is the most objective choice.

4.2 Veteran's Administration lung cancer trial

The Veteran's Administration lung cancer data are from a clinical trial to compare a standard and a test chemotherapy (see Kalbfleisch and Prentice, 1980, Appendix 1). The data set consists of the censored survival times of $n = 137$ male patients. The observed

G	Treatment group (randomized):
	1 = standard chemotherapy, 0 = new test chemotherapy.
K	Performance status of patient (Karnofsky scale), dummy coded in three categories:
	K_1 scale 10–30, completely hospitalized,
	K_2 scale 40–60, partial confinement,
	K_3 scale 70–90, able to care for self.
A	Age in years: 34 to 81 years.
M	Time in months from diagnosis to randomization: 1 to 87 months.
P	Prior therapy: 1 = yes, 0 = no.
H	Histological type of tumor, dummy coded in four categories:
	H_1 squamous,
	H_2 small cell,
	H_3 adeno,
	H_4 large cell.

Table 2: Covariates of the Veteran's Administration lung cancer data.

event is the death of a patient and only 9 of the 137 times are censored. To consider the possibility of existing heterogeneity between patients, a number of covariates was measured, see Table 2. With progressing observation time the number of patients at risk decreases strongly. For example, after about 8 months only 10 patients are at risk in each therapy group, while beyond month 20 no patient with standard chemotherapy is under observation. Therefore we group the survival time (originally given in days) into months. Hence, for each patient $i = 1, \ldots, n$ the survival time is measured at discrete time points t_i with maximal time $T_{\max} = 34$ months. Since splines are very sensitive to data situations with very sparse data at the end of the observation period, we additionally reduce the influence of the sparse data following a proposal of Grambsch and Therneau (1994) by using the monotonous transformation $L_t = \log(t)$ of the original time scale t.

To analyze the survival of patients in dependence of the covariates x_t given in Table 2 at survival time $t = 1, \ldots, T_{\max}$, we consider the discrete hazard rate $\lambda(t|x_t) = P(T = t|T \geq t, x_t)$. This is the conditional probability for the death of a patient at time t given the patient has survived up to that time. To analyze the hazard rate

within the framework of the GLM and especially the BVCM presented in this paper, the discrete survival data have to be transformed in the following way: For each patient $i = 1, \ldots, n$ and each time point $t = 1, \ldots, t_i$ we define binary event indicators by $y_{it} = 1$, if patient i dies at the discrete time point t, otherwise $y_{it} = 0$. With the covariates $x_{it} = (L_t, G_i, K_{1i}, K_{2i}, A_i, M_i, P_i, H_{1i}, H_{2i}, H_{3i})$ of patient i at time t and the histories y^*_{t-1} and x^*_t of event indicators and covariates of all patients up to time $t-1$ and t, respectively, the distributional assumption $y_{it}|y^*_{t-1}, x^*_t \sim B(1, \mu_{it})$ holds, and the discrete hazard rate of patient i at time t,

$$\lambda(t|x_{it}) = P(y_{it} = 1|y^*_{t-1}, x^*_t, y_{i1} = \ldots = y_{i,t-1} = 0) = \mu_{it},$$

is modeled within the framework of the GLM as $\lambda(t|x_{it}) = h(\eta_{it})$ with the logistic distribution function h. For details on discrete survival models see Fahrmeir and Tutz (1997).

Model 1 considers all covariates and is defined by the predictor

$$\begin{aligned}(1) \quad \eta_{it} \;=\; & \beta_0 + f_0(L_t) + G_i\, f_G(L_t) + K_{1i}\, f_{K_1}(L_t) + K_{2i}\, f_{K_2}(L_t) + P_i\, f_P(L_t) \\ & + H_{1i}\, f_{H_1}(L_t) + H_{2i}\, f_{H_2}(L_t) + H_{3i}\, f_{H_3}(L_t) + f_A(A_i) + f_M(M_i).\end{aligned}$$

The effects of the binary covariates G_i, K_{1i}, K_{2i}, H_{1i}, H_{2i}, H_{3i} and P_i are modeled by coefficients that vary over the transformed time L_t, while the functions f_A and f_M vary over the metrical variables A_i and M_i.

Figure 4 shows the estimates of the varying coefficients together with 90% Bayesian credible intervals. The effect of therapy in graph (a) is negative at the beginning, after 5 months the zero line is intersected, and then it stays positive. This implies that at the beginning the classical therapy is better for surviving, while from month 5 on the new test therapy is better. As in Kalbfleisch and Prentice (1980) with a pure parametric approach, or Mau (1986) with time varying coefficients, the effect of therapy may be considered as non–significant, since the zero line is included in the credible region for almost the whole observation period. The effect of Karnofsky scale 10–30 in graph (b) starts at value 4.4 and then decreases monotonously. Near month 20 it is approximately zero. A similar behaviour is seen for Karnofsky scale 40–60 in graph (c) which starts at value 2.2 and intersects the zero line after 4 months. This implies that the patients with Karnofsky scale 10–30 have the greatest risk of death in the first 8 months of treatment when compared to patients with Karnofsky scale 70–90 (the reference category). After month 8, the effect is

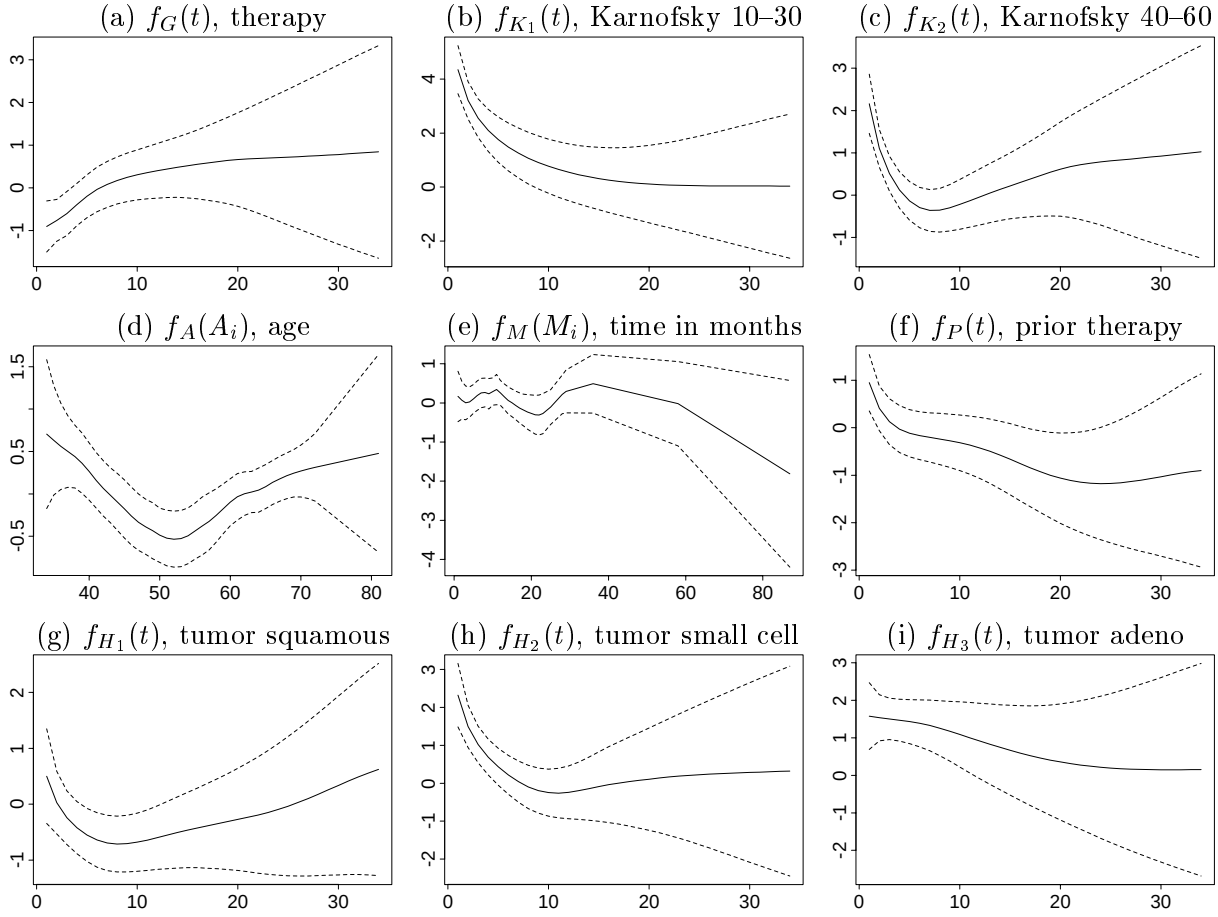

Figure 4: Estimates of varying coefficients with 90% Bayesian credible intervals (model 1).

non–significant, since the credible region includes the zero line. Patients with Karnofsky scale 40–60 have a greater risk of death in the first 4 months when compared to patients with Karnofsky scale 70–90. Notice however that this risk is below the one of patients with Karnofsky scale 10-30. From month 4 forward this effect is non–significant, since the credible region includes the zero line. The effect of age in graph (d) shows different risks for different age groups. The lowest risk can be seen for patients with age about 50 years. The effect of time in months from diagnosis to randomization in graph (e) varies about a horizontal line, and hence could be considered as being non–significant. Graph (f) depicts the effect of a prior therapy which monotonously falls until about month 24. Here the credible region includes the zero line over almost the whole observation period. Hence, this effect may be considered to be non–significant. The graphs (g) to (i) give the effects of the dummy variates of tumor type with reference category "large cell". The effect of tumor type "squamous" may be considered as non–significant, since again the credible region includes the zero line for almost the whole observation time. However, the

tumor type "small cell" is significantly positive up to month 4, after that time there is no effect. The effect of tumor type "adeno" is positive over the whole observation period, and approximates at the end to zero. We see that the effect could also be modeled by a straight line. When compared to the reference category tumor type "large cell", this results indicate that patients with type "small cell" and "adeno" have (at least in the first 8 to 10 months) a greater risk of death, whereby the risk of type "adeno" is above the risk of type "small cell".

To discover the covariates that are relevant for the survival of patients, we fit the following reduced models, which are compared by the deviance information criterion (9):

$$
\begin{aligned}
(2) \quad \eta_{it} &= \beta_0 + f_0(L_t) + G_i\, f_G(L_t) + K_{1i}\, f_{K_1}(L_t) + K_{2i}\, f_{K_2}(L_t) + P_i\, f_P(L_t) \\
&\quad + H_{1i}\, f_{H_1}(L_t) + H_{2i}\, f_{H_2}(L_t) + H_{3i}\, f_{H_3}(L_t) + A_i\beta_A + M_i\beta_M, \\
(3) \quad \eta_{it} &= \beta_0 + f_0(L_t) + G_i\, f_G(L_t) + K_{1i}\, f_{K_1}(L_t) + K_{2i}\, f_{K_2}(L_t) + P_i\, f_P(L_t) \\
&\quad + H_{1i}\, f_{H_1}(L_t) + H_{2i}\, f_{H_2}(L_t) + H_{3i}\, f_{H_3}(L_t), \\
(4) \quad \eta_{it} &= \beta_0 + f_0(L_t) + G_i\, f_G(L_t) + K_{1i}\, f_{K_1}(L_t) + K_{2i}\, f_{K_2}(L_t) \\
&\quad + H_{1i}\, f_{H_1}(L_t) + H_{2i}\, f_{H_2}(L_t) + H_{3i}\, f_{H_3}(L_t), \\
(5) \quad \eta_{it} &= \beta_0 + f_0(L_t) + G_i\, f_G(L_t) + K_{1i}\, f_{K_1}(L_t) + K_{2i}\, f_{K_2}(L_t) \\
&\quad + H_{1i}\, f_{H_1}(L_t) + H_{2i}\, f_{H_2}(L_t) + H_{3i}\beta_{H_3}, \\
(6) \quad \eta_{it} &= \beta_0 + f_0(L_t) + K_{1i}\, f_{K_1}(L_t) + K_{2i}\, f_{K_2}(L_t) \\
&\quad + H_{1i}\, f_{H_1}(L_t) + H_{2i}\, f_{H_2}(L_t) + H_{3i}\beta_{H_3}.
\end{aligned}
$$

When compared to model 1 in model 2 only the effects of age A and time in months from diagnosis to randomization M are modeled as fixed, while these two covariates are completely left out in model 3. Model 4 results from model 3 by leaving out the covariate prior therapy P. In model 5 additionally the effect of tumor type "adeno" H_3 is considered to be constant over time. Model 6 finally results from model 5 by leaving out the effect of the covariate treatment group G (that was considered as being non–significant).

Table 3 shows the model fit of models 1 to 6 computed with the deviance information criterion (DIC) (9). With the greatest value of DIC model 1 has the worst model fit resulting from the greatest complexity p_D. Modeling the effects of A and M as constant in model 2 yields a greater $\overline{D}$ but a much smaller p_D, resulting altogether in a clearly

Model	$\overline{D}$	$D(\overline{\theta})$	p_D	DIC
(1)	545.93	504.74	41.19	587.12
(2)	546.39	518.24	28.14	574.53
(3)	544.54	518.22	26.32	570.86
(4)	545.79	522.04	23.75	569.54
(5)	545.24	522.66	22.58	567.81
(6)	547.40	526.89	20.52	567.92

Table 3: Model fit of models 1 to 6 computed with the deviance information criterion (DIC).

better fit by DIC. The estimates of these constant effects, $\hat{\beta}_A = -0.0028$ and $\hat{\beta}_M = -0.0052$, are almost zero. With the 90% Bayesian credible regions $(-0.0212, 0.0150)$ and $(-0.0340, 0.0246)$ these two effects are non–significant. The omission of the covariates A and M in model 3 yields a further clear improvement of the fit by a smaller DIC. Also the omission of covariate prior therapy P in model 4 results in a somewhat better fit. We mentioned above, that the effect of tumor type "squamous" could be considered as non–significant. But both leaving out this covariate and modeling the effect as constant yields a greater value of DIC and hence a worse fit (this result is not shown in Table 3). However, we yield a better fit, if in model 5 the effect of tumor type "adeno" is modeled as constant over time with estimate $\hat{\beta}_{H_3} = 1.3073$ and 90% credible region $(0.5950, 2.0824)$. This results both in a smaller deviance $\overline{D}$ and in a smaller complexity p_D when compared to model 4. If the covariate tumor type with its dummies H_1, H_2 and H_3 would be left completely out, corresponding to the results of Mau (1986) where only the covariate Karnofsky scale is considered as significant, we would yield a very bad model fit, which would be worse than that of model 1 (also this result is not shown in Table 3). A similar model fit as with model 5 results if we additionally leave out the covariate treatment group G (model 6). This corresponds to the results of Kalbfleisch and Prentice (1980) and Mau (1986), where the treatment group is not significant for the survival of patients.

The presented results indicate, that the Veteran's Administration lung cancer data are best described by model 5 with the covariates treatment group, Karnofsky scale and histological type of tumor.

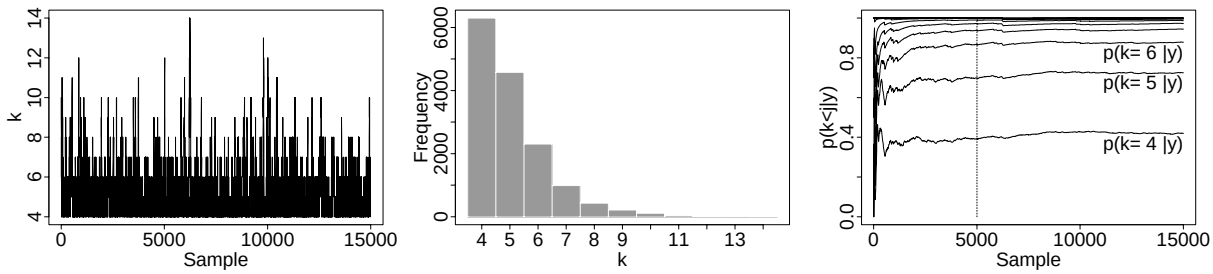

Figure 5: Sample path (left), frequencies (middle) and cumulative occupancy fractions (right) for the samples of the model indicator k_G for estimating the varying effect f_G of covariate treatment group in model 5.

As an example of the samples of the model indicators k_j, Figure 5 gives details for the sample of k_G for estimating the varying effect f_G of the covariate treatment group in model 5. The left part of Figure 5 shows the sample of k_G with values between 4 and 14. With an acceptance rate of 0.31 for the steps birth and death, the mixing over k_G is good. In the middle of Figure 5 there is the frequency of the accepted values of k_G with mode at $k_G = 4$. The right part of Figure 5 depicts the cumulative occupancy fractions $p(k_G < j|y)$ for the different values of k_G against the number of iterations. After the burn–in phase these cumulative occupancy fractions stay on a stable level speaking for an adequate length of the burn–in. The samples of the other model indicators k_j for $j \in \{0, K_1, K_2, H_1, H_2\}$ of model 5 behave similar and hence are not shown.

5 Conclusions

As we demonstrated in the last section, Bayesian non– and semiparametric regression is a valuable tool for practical data analysis. MCMC techniques provide a rich output for inference, prediction and model comparison. No approximations based on asymptotic arguments have to be made, and data–driven choice of smoothing or tuning parameters is incorporated as part of the model.

The main advantage of Bayesian modeling and inference with modern Monte Carlo techniques is the modular structure. This allows to generalize and to modify the existing approach in a conceptually straightforward way. Some future extensions are: inclusion of basis functions which admit edges or jumps, two–dimensional basis functions such as tensor products of B–splines, and incorporation of random effects for longitudinal or spatial data.

Appendix A: Update of fixed effects

With this move type both the fixed effects β and the basis coefficients $c_1, \ldots, c_p$ are updated by a method for GLM's with fixed effects. Since the dimensions k_j of the coefficients c_j are varying from iteration to iteration, here we need a more sophisticated MCMC technology that avoids tuning of the parameters of the proposal distribution. Suitable approaches are the adaptive rejection Metropolis sampler of Gilks, Best and Tan (1995) or the approach of Gamerman (1997a), the so–called weighted least squares proposal. As mentioned in Biller (2000), the latter approach has some advantages regarding computing time and provides the incorporation of correlations between the fixed effects β. Another advantage is the possibility to adapt this method in a straightforward way to GLM's where the vector of fixed effects is split up in several blocks, which have to be simulated separately as in the BVCM (6).

For this adaptation of the approach of Gamerman (1997a) we consider a GLM with fixed effects $\alpha = (\alpha'_{(1)}, \ldots, \alpha'_{(p+1)})'$ split up in $p+1$ blocks $\alpha_{(j)}$ yielding the predictor $\eta_i = z_{i(1)}\alpha_{(1)} + \ldots + z_{i(p+1)}\alpha_{(p+1)}$. The BVCM (6) then is given by $z_{i(1)} = z_i$, $\alpha_{(1)} = \beta$ and $z_{i(j+1)} = x_{ij}B_j(r_{ij})$, $\alpha_{(j+1)} = c_j$ for $j = 1, \ldots, p$. The blocks $\alpha_{(j)}$ are assumed to be a priori independent and multivariate normal $N(\alpha_{(j0)}, \Sigma_{\alpha_{(j)}})$. For each $j = 1, \ldots, p+1$ we consider the full conditional $p(\alpha_{(j)}|\alpha_{(-j)}, y)$ of block $\alpha_{(j)}$, where $\alpha_{(-j)}$ denotes the vector α without $\alpha_{(j)}$. In a single Fisher scoring step this full conditional now is maximized with regard to $\alpha_{(j)}$, resulting in a MAP (maximum a posteriori) estimate $\hat{m}_{(j)}$ of $\alpha_{(j)}$ and the inverse of the expected Fisher information $\hat{C}_{(j)} = \hat{F}^{-1}_{(j)}$. Details are given in Gamerman (1997a).

For the separate simulation of each block $\alpha_{(j)}$, the two estimates $\hat{m}_{(j)}$ and $\hat{C}_{(j)}$ are computed in each iteration of the algorithm by a single Fisher scoring step, given the estimate of $\alpha_{(j)}$ of the preceding iteration. The new proposal for $\alpha_{(j)}$ then is drawn from the multivariate normal proposal distribution $N(\hat{m}_{(j)}, \hat{C}_{(j)})$. This procedure incorporates the structure of the observational model in the proposal distribution, leading to a very efficient algorithm with good convergence and mixing properties.

Acknowledgement: This work was supported by a grant from the German National Science Foundation, Sonderforschungsbereich 386.

References

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

Biller, C. (2000). Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models, *Journal of Computational and Graphical Statistics* **9**: 122–140.

De Boor, C. (1978). *A Practical Guide to Splines*, Springer–Verlag, New York.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic Bayesian curve fitting, *J. R. Statist. Soc. B* **60**: 333–350.

Dierckx, P. (1993). *Curve and surface fitting with splines*, Oxford University Press, Oxford.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*, Marcel Dekker, New York.

Fahrmeir, L. and Tutz, G. (1997). *Multivariate Statistical Modelling Based on Generalized Linear Models*, corrected third printing edn, Springer–Verlag, New York.

Fahrmeir, L. and Lang, S. (1999). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors, *Discussion Paper 169*, Sonderforschungsbereich 386, Ludwig–Maximilians–Universität München. Revised for Applied Statistics.

Friedman, J. H. and Silverman, B. W. (1989). Flexible Parsimonious Smoothing and Additive Modeling (with discussion), *Technometrics* **31**: 3–39.

Gamerman, D. (1997a). Efficient sampling from the posterior distribution in generalized linear mixed models, *Statistics and Computing* **7**: 57–68.

Gamerman, D. (1997b). *Markov Chain Monte Carlo—Stochastic simulation for Bayesian inference*, Champman and Hall, London.

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995). Adaptive Rejection Metropolis Sampling within Gibbs Sampling, *Applied Statistics* **44**: 455–472.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazard tests and diagnostics based on weighted residuals, *Biometrika* **81**: 515–526.

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

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (1993). Varying–coefficient Models, *J. R. Statist. Soc. B* **55**: 757–796.

Hastie, T. and Tibshirani, R. (2000). Bayesian Backfitting, *Statistical Science* . to appear.

Kalbfleisch, J. and Prentice, R. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Lyche, T. and Schumaker, L. L. (1973). Computation of smoothing and interpolating natural splines via local bases, *SIAM J. Numer. Anal.* **10**: 1027–1038.

Lyche, T. and Strøm, K. (1996). Knot Insertion for Natural Splines, *Annals of Numerical Mathematics* **3**: 221–246.

Mallick, B. K., Denison, D. G. T. and Smith, A. F. M. (2000). Semiparametric generalized linear models: Bayesian approaches, *in* D. K. Dey, S. K. Ghosh and B. K. Mallick (eds), *Generalized linear models: A Bayesian perspective*, Marcel–Dekker, New York.

Mau, J. (1986). On a Graphical Method for the Detection of Time–dependent Effects of Covariates in Survival Data, *Applied Statistics* **35**: 245–255.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd. edn, Chapman and Hall, New York.

Schumaker, L. L. (1993). *Spline functions: basic theory*, reprinted with corrections edn, Krieger Publishing Company, Malabar, Florida.

Smith, M. and Kohn, R. (1996). Nonparametric Regression using Bayesian Variable Selection, *Journal of Econometrics* **75**: 317–343.

Spiegelhalter, D. J., Best, N. G. and Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models, *Research Report 98–009*, Division of Biostatistics, University of Minnesota.

Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997). The use of polynomial splines and their tensor products in extended linear modeling (with discussion), *Ann. Statist.* **25**: 1371–1470.

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