Biller:

Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models

Sonderforschungsbereich 386, Paper 133 (1998)

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

Projektpartner

Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models

Clemens Biller

SFB 386, Institute of Statistics

Ludwig Maximilians University

Ludwigstr. 33, 80539 Munich, Germany

email: biller@stat.uni-muenchen.de

Abstract

This paper presents a fully Bayesian approach to regression splines with automatic knot selection in generalized semiparametric models for fundamentally non–Gaussian responses. In a basis function representation of the regression spline we use a B–spline basis. The reversible jump Markov chain Monte Carlo method allows for simultaneous estimation both of the number of knots and the knot placement, together with the unknown basis coefficients determining the shape of the spline. Since the spline can be represented as design matrix times unknown (basis) coefficients, it is straightforward to include additionally a vector of covariates with fixed effects, yielding a semiparametric model. The method is illustrated with data sets from the literature for curve estimation in generalized linear models, the Tokyo rainfall data and the coal mining disaster data, and by a credit–scoring problem for generalized semiparametric models.

Keywords: B–spline basis; knot selection; nonnormal response; nonparametric regression; reversible jump Markov chain Monte Carlo.

1 Introduction

Let us consider observations (y_i, x_i, z_i), $i = 1, \ldots, n$, on $p+1$ explanatory variables, with a metrical covariate x and a p–vector $z = (z_1, \ldots, z_p)$, and on a univariate nonnormal response y, e.g. a discrete or nonnegative response. The aim is to analyze the dependence of the response on the covariates in a semiparametric generalized linear model. That means, the distribution of y_i given the covariates x_i, z_i is assumed to belong to an exponential family where the mean $\mu_i = \mathrm{E}(y_i|x_i, z_i)$ is linked to the predictor η_i by an appropriate response function h, i.e., $\mu_i = h(\eta_i)$. Since we consider a semiparametric approach, the predictor is of the form

$$\eta_i = f(x_i) + z_i\beta, \tag{1}$$

with an unknown real valued regression function f and a p-dimensional parameter vector β.

A common approach to estimate the unknowns f and β from the data is maximizing the penalized log–likelihood

$$\mathrm{pl}(f, \beta) = \sum_{i=1}^{n} l_i(y_i|\eta_i) - \frac{1}{2}\lambda \int \left(f''(x)\right)^2 dx \tag{2}$$

over f and β, where the log–likelihood contribution $l_i(y_i|\eta_i)$ is defined as the logarithm of the probability density function $p(y_i|\eta_i)$ of the response y_i. Here the parameter λ controls the compromise between faith with the data and smoothness of the maximizing function $\hat{f}$, which is a natural cubic smoothing spline with knots at each distinct x_i, see Green and Yandell (1985) or Green and Silverman (1994). For rather small λ, $\hat{f}$ will nearly interpolate the data, while a large λ yields a very smooth and nearly linear $\hat{f}$. Data driven methods for choosing the smoothing parameter λ exist, see e.g. Fahrmeir and Tutz (1997), Chapter 5.

Another approach starts by defining the unspecified function f as a spline. With a certain number k of knots $t_1 < \ldots < t_k$, with $t_j \in [x_{\min}, x_{\max}]$, we consider the space $\mathcal{S}_q(t_1, \ldots, t_k)$ of splines of order q. In this space of dimension $K = k + q$ we can represent f by

$$f(x) = \sum_{j=1}^{K} c_j B_j(x) = B(x)\, c, \tag{3}$$

with (known) basis functions $B(x) = (B_1(x), \ldots, B_K(x))$ of $\mathcal{S}_q(t_1, \ldots, t_k)$, and unknown basis coefficients $c = (c_1, \ldots, c_K)'$. As alternative one may define f to lie in a subspace of $\mathcal{S}_q(t_1, \ldots, t_k)$,

the k–dimensional space $\mathcal{NS}_q(t_1, \ldots, t_k)$ of natural splines of order $q = 2m$, $m \in \mathbb{N}$. For $q = 4$, $\mathcal{NS}_q(t_1, \ldots, t_k)$ is the space of natural cubic splines, where the function $\hat{f}$, resulting from the maximization of the penalized log–likelihood (2), lies in. Defining $K = k$, $f \in \mathcal{NS}_q(t_1, \ldots, t_k)$ has the representation (3), too. An appropriate basis $B_1(x), \ldots, B_K(x)$ for both spaces is the widely used B–spline basis with local support. For details and efficient algorithms for computing this basis see Eubank (1988) or Schumaker (1993), and especially for natural splines Lyche and Schumaker (1973) or Lyche and Strøm (1996).

Both with $f \in \mathcal{S}_q(t_1, \ldots, t_k)$ and with $f \in \mathcal{NS}_q(t_1, \ldots, t_k)$ the predictor (1) now has the form

$$\eta_i = B(x_i)\, c + z_i \beta = \Big(B(x_i), z_i \Big) \begin{pmatrix} c \\ \beta \end{pmatrix}, \tag{4}$$

i.e., we have a generalized linear model with fixed coefficients c and β, that can be estimated by standard methods. The shape and smoothness of the estimator $\hat{f}$ is given by the number k and the location of the knots $t_1, \ldots, t_k$. With only a few knots the function $\hat{f}$ is very smooth, while increasing the number k of knots allows a more wiggly and flexible $\hat{f}$. Placement of a knot in a certain aera yields more flexibility of $\hat{f}$ in that aera. Since finding the right number and location of knots by visual inspection of the data is impossible in most cases (see Eubank, 1988, Section 7.2), we need data driven methods for knot placement to get (in some sense) nearly optimal estimators $\hat{f}$.

For normal response y, such data driven methods exist. Friedman and Silverman (1989) present an adaptible knot placement algorithm with forward and backward steps. In the forward steps they add knots which are optimal with respect to the average squared residual criterion, while in the backward steps they delete knots yielding the model being optimal for the generalized cross-validation score. A Bayesian approach using reversible jump Markov chain Monte Carlo (RJMCMC, see Green, 1995) is presented by Denison, Mallick and Smith (1998). In each iteration they choose the set of knots by RJMCMC methods, and given these knots the spline is estimated by the usual least squares approach. The estimator $\hat{f}$ then is built by averaging over all iterations. For generalized linear models, Stone, Hansen, Kooperberg and Troung (1997) also propose forward and backward steps. They add basis functions using Rao statistics until a maximum number of knots

is reached. Subsequently, basis functions are deleted using Wald statistics. Out of the complete sequence of models, the one optimal with respect to the AIC criterion is chosen.

This paper presents a Bayesian approach to automatic knot selection in generalized semiparametric models (1) with nonnormal response using RJMCMC methods for knot selection and (ordinary) MCMC methods for estimating the resulting generalized linear model (4). So, in contrast to Denison et al. (1998), where the estimation of the basis coefficients given the knots is done by ordinary least squares methods for normal response, we use a fully Bayesian approach in nonnormal cases. And contrary to Friedman and Silverman (1989) and Stone et al. (1997), where the result is one somehow 'optimal' knot placement, the RJMCMC method can neither find one optimal number k of knots nor an optimal placement of these k knots. But in each iteration of the RJMCMC algorithm both the number of knots and the knot placement may vary. So in each iteration the estimation of the coefficients c and consequently the estimation of f is based on different knot settings. The final estimator of f is built as the mean of the estimators in each iteration, and hence a great flexibility of $\hat{f}$ is provided. Hence, this procedure is a sort of Bayesian model averaging.

The paper is organized as follows: Section 2 gives a brief overview over ordinary and reversible jump MCMC methods, which are applied in Section 3 to the Bayesian approach to adaptive regression splines. Applications of the proposed method to curve estimation in the generalized linear model and to the generalized semiparametric model follow in Section 4. Some concluding remarks and possible extensions of the model are given in Section 5.

2 Markov chain Monte Carlo methods

Here we give a brief summary of Markov chain Monte Carlo (MCMC) theory. The emphasis, in Section 2.2, is on reversible jump MCMC, proposed by Green (1995), which allows Bayesian estimation in hierarchical models where the parameter dimension varies. As introduction, Section 2.1 contains the 'ordinary' MCMC methods, but restricted to the Metropolis–Hastings algorithm, one of the most popular methods in MCMC computation. For more details including other methods, e.g. the Gibbs sampler, see Tierney (1994), Besag, Green, Higdon and Mengersen (1995) or Gilks, Richardson and Spiegelhalter (1996). In contrast to the remainder of the paper, where x is the

metrical covariate, in this section the parameters of interest are denoted by x.

2.1 Metropolis–Hastings MCMC

Consider a distribution $\pi(x)$, which in Bayesian inference is the posterior of the parameters of interest $x = (x_1, \ldots, x_n)$ given the data. Since in most cases $\pi(x)$ is not completely known, e.g. due to analytically intractable normalizing constants, direct sampling from the distribution of interest is not possible. To overcome this problem, in MCMC computation we create a Markov chain $x^{(1)}, \ldots, x^{(N)}$, whose transition kernel $P(x, x')$ has the limiting distribution $\pi(x)$. This sample is used for estimating parameters of interest, e.g. $\mathrm{E}(x)$ is estimated by the sample mean. In the Metropolis–Hastings method, we have to construct the kernel $P(x, x')$ in a way that it satisfies aperiodicity and irreducibility, and also the detailed balance

$$\pi(x)P(x, x') = \pi(x')P(x', x) \quad \forall x, x'. \tag{5}$$

Often the vector $x = (x_1, \ldots, x_n)$ is split up into components x_T, with T a subset of $\{1, \ldots, n\}$, and we consider the so–called full conditionals $\pi(x_T|x_{-T})$ instead of $\pi(x)$, where $x_{-T} = \{x_i, i \notin T\}$. Given the current state x_T of the Markov chain, a new value x'_T is drawn from an arbitrarily chosen proposal density $q_T(x_T, x'_T)$. With probability

$$\alpha_T(x_T, x'_T) = \min\left\{1, \frac{\pi(x'_T|x_{-T})q_T(x'_T, x_T)}{\pi(x_T|x_{-T})q_T(x_T, x'_T)}\right\} \tag{6}$$

the proposed value is accepted as new state of the chain, otherwise we leave x_T unchanged. Definition (6) ensures, that the necessary properties of the transition kernel $P(x, x')$ hold.

2.2 Reversible jump MCMC

Consider the following hierarchical model: let $k \in \mathcal{K}$ be an indicator from a countable set $\mathcal{K}$; given k we have a parameter vector $\theta^{(k)} \in \Theta^{(k)}$; finally the data y. Here each k determines a model $\mathcal{M}_k$ defined by the parameter $\theta^{(k)}$, with dimension of parameter space $\Theta^{(k)}$ possibly varying with k. The hierarchy is also reflected in the joint probability

$$p(k, \theta^{(k)}, y) = p(k)p(\theta^{(k)}|k)p(y|k, \theta^{(k)}),$$

the product of model probability, prior and likelihood, and in the joint posterior $p(k, \theta^{(k)}|y) \propto p(k, \theta^{(k)}, y)$, on which Bayesian inference is based on. For convenience we abbreviate $(k, \theta^{(k)})$ as x, and $p(k, \theta^{(k)}|y) = p(x|y)$ as $\pi(x)$. Given k, x lies in $\mathcal{C}_k = \{k\} \times \Theta^{(k)}$, while generally $x \in \mathcal{C} = \bigcup_{k \in \mathcal{K}} \mathcal{C}_k$.

For a fully Bayesian analysis with joint estimation of both model parameter $\theta^{(k)}$ and model indicator k, we need a method, that switches between parameter subspaces $A, B \subset \mathcal{C}$ of possibly different dimension. For that reason we use different move types m, and for each of these move types we construct a transition kernel P_m, which satisfies the detailed balance

$$\int_A \int_B \pi(dx) P_m(x, dx') = \int_B \int_A \pi(dx') P_m(x', dx), \quad \forall A, B \subset \mathcal{C}.$$

That means, the equilibrium probability of moves from A to B equals that from B to A. Given the current state x, a move of type m is proposed to state dx' with probability $q_m(x, dx')$, and is accepted with probability

$$\alpha_m(x, x') = \min\left\{1, \frac{\pi(dx') q_m(x', dx)}{\pi(dx) q_m(x, dx')}\right\}. \tag{7}$$

The ratio (7) makes sense for moves between subspaces A and B from $\mathcal{C}$, if we consider that proposal degrees of freedom are matched. This 'dimension matching' assumption is attained as follows. For a move from state x to state x' we generate random numbers u and set x' to be a deterministic function $x'(x, u)$. The reverse move from x' to x has to be defined symmetrically by generating random numbers u' and setting $x = x(x', u')$. For dimension matching there must be a bijection between (x, u) and (x', u'), i.e., the dimensions n_1 and n_2 of u and u' must satisfy $n_1 + m_1 = n_2 + m_2$, with m_1, m_2 the dimensions of x and x', respectively. The acceptance probability (7) then results in

$$\alpha_m(x, x') = \min\left\{1, \frac{p(x'|y) j_2(x') q_2(u')}{p(x|y) j_1(x) q_1(u)} \left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right\}. \tag{8}$$

Here j_1, j_2 are probabilities for the move types given x and x', and q_1, q_2 are the distributions of u and u'. The Jacobian results from deriving the proposal of move $B \to A$ $(q_m(x', dx))$ from the proposal of the reverse move $A \to B$ $(q_m(x, dx'))$ using the distributions q_1, q_2.

In practice often $n_1 + m_1 = m_2$ holds, i.e., only for the birth step a random number u is

necessary, while the reverse death step is deterministic, given the element to be removed. In this case, in (8) the terms $q_2(u')$ and u' in the Jacobian have to be omitted.

For move types without switching between subspaces, i.e., moves within one subspace, the ordinary MCMC theory of Section 2.1 holds, and the acceptance probabilities (6) and (8) are equal, since then the Jacobian is 1 and all ratios depending on k cancel out. For further details on the reversible jump MCMC method see Green (1995).

3 A Bayesian approach to adaptive regression splines

Consider the semiparametric generalized linear model of Section 1 with predictor (4) and the function f being defined as spline (3) with B–spline basis functions $B_1, \ldots, B_K$ and basis coefficients $c = (c_1, \ldots, c_K)'$. With $K = k$ and $q = 4$ here we consider $f \in \mathcal{NS}_4(t_1, \ldots, t_k)$, i.e., f is defined as cubic natural spline. By means of comparisons, in Section 3.5 we give adjustments to the case $f \in \mathcal{S}_4(t_1, \ldots, t_k)$ (with $K = k + 4$).

Now assume, that both the number of knots k and the placement of these knots are not given and so have to be estimated jointly with the model parameters. For a Bayesian approach let us formulate the following hierarchical model: the number k of knots is from some countable set $\mathcal{K}$ (which is specified in Section 3.1) and serves as model indicator. Each value of k defines a model, that is determined by the parameters $t = (t_1, \ldots, t_k)$ and $c = (c_1, \ldots, c_K)'$. For given k, the model for the data (y_i, x_i, z_i), $i = 1, \ldots, n$, with $y = (y_1, \ldots, y_n)$, is defined by the choice of the exponential family and the semiparametric predictor (4). Including additionally the fixed effects β we define the model parameter

$$\theta_k = (t, b) \in \mathbb{R}^{k+K+p}$$

with $b = (c', \beta')'$. Model indicator and model parameter are combined to $\theta = (k, \theta_k)$. The hierarchical model is also expressed by the joint posterior

$$p(k, \theta_k|y) \propto p(k)p(\theta_k|k)p(y|k, \theta_k),$$

the product of model probability, prior and likelihood, where for notational convenience we neglect the covariates x and z.

For the joint estimation of (k, θ_k) with variable model indicator k using the reversible jump MCMC method, we have to develope appropriate reversible jump moves as mentioned in Section 2.2. The scanning of these moves may be done randomly, as in Green (1995), or in systematical order, see Richardson and Green (1997). Following the latter approach we define three move types:

(1) birth or death of one knot t_{j+1}, i.e., adding or deleting a t_{j+1} with changing k by 1 and corresponding changes in c; the choice between birth and death is done randomly;

(2) move a given knot t_j to another position (without change in k);

(3) update the coefficients $b = (c', \beta')'$ (without change in k).

The set of move types is denoted by $S = \{B, D, P, C\}$, where B means a birth, D a death, P a position change, and C an update of coefficients. One iteration or sweep of our algorithm is made of the move types (1) to (3). Working out moves (1) to (3), each sweep may result in a different knot setting, i.e., a different number and location of knots. Defining the final estimator $\hat{f}$ as mean of the estimators of f in each sweep, which are based on these different settings, a great flexibility of $\hat{f}$ and a good fit to the data is provided.

The following section gives details to the specification of the prior distributions, while Sections 3.2 to 3.4 present the reversible jump move types (1) to (3) defined above (but in reverse order).

3.1 Prior specifications

The model indicator k is supposed to lie in a set $\mathcal{K} = \{k_{\min}, k_{\min+1}, \ldots, k_{\max}\} \subset \mathbb{N}$. For $f \in \mathcal{NS}_4(t_1, \ldots, t_k)$ $k_{\min}$ is restricted to $k_{\min} \geq 4$ due to the definition of natural splines, otherwise $k_{\min} \geq 1$ holds. As prior for k there are two plausible alternatives: a Poisson distribution with parameter λ restricted to the set $\mathcal{K}$, or a discrete uniform distribution on $\mathcal{K}$. Both priors will be considered.

Given k we assume the model parameters t and b to be independent, i.e., $p(\theta_k|k) = p(t|k)p(b|k)$, and we treat t and b separately.

The knots t are supposed to lie in a discret set of candidate knots $\mathcal{T}_0 = \{t_{01}, t_{02}, \ldots, t_{0k_{\max}}\}$, which may consist of the sorted distinct values of covariate x. An alternative is to distribute

$t_{01}, \ldots, t_{0k_{\max}}$ equidistantly over the intervall $[x_{\min}, x_{\max}]$. To define the prior for t we assume, that all possible samples $t = (t_1, \ldots, t_k)$ out of $\mathcal{T}_0$ have equal probability

$$p(t|k) = \binom{k_{\max}}{k}^{-1} = \frac{k!(k_{\max} - k)!}{k_{\max}!}, \tag{9}$$

hence, the prior $p(t|k)$ depends only on k and $k_{\max}$.

A widely used prior for the coefficients $b = (c', \beta')'$ of a generalized linear model is the multivariate normal distribution $b|k \sim N_{K+p}(0, \Sigma_0)$ (see e.g. Gamerman, 1997). While the basis coefficients c are assumed to be uncorrelated, possible correlations between the coefficients $\beta = (\beta_1, \ldots, \beta_p)'$ are modelled by defining $\Sigma_0 = \sigma_0^2 \text{diag}(I_K, R_p)$ with the K–dimensional identity matrix I_K and a p–dimensional correlation matrix R_p.

3.2 Update of coefficients

Move type (3), the update of the coefficients, is a conventional update of the fixed parameters of a generalized linear model, where ordinary MCMC methods of Section 2.1 are applicable. The simplest choice is a Metropolis random walk proposal with tuning of the scale parameter. Since the dimension of the parameter b may change with k from one sweep to another, tuning is not possible here. So we have to use more sophisticated methods which avoid tuning, but normally need more computing time. Dellaportas and Smith (1993) use adaptive rejection sampling for Gibbs sampling proposed by Gilks and Wild (1995) for univariate log–concave densities. To deal with non–log–concave distributions, Gilks, Best and Tan (1995) propose a generalization, the adaptive rejection Metropolis sampling. A completely different approach for multivariate distributions, the so–called weighted least squares proposal, is proposed by Gamerman (1997). In a single Fisher scoring step the posterior distribution of the parameter b given the data is maximized, resulting in a MAP estimate of b and the expected Fisher information. These two values are used as mean and covariance of a Gaussian proposal of the Metropolis–Hastings algorithm, and so the structure of the observation model is incorporated in the proposal distribution. Here we use the approach of Gamerman (1997), since in contrast to adaptive rejection (Metropolis) sampling it provides the incorporation of correlations between the fixed effects β, and has some advantage regarding computing time.

3.3 Position change

In the position change, i.e., the move of a knot t_j to another position, only this knot changes, while the remaining parameters k, $t \backslash t_j$ and b stay unchanged. So here we have a transition from state θ to state $\tilde{\theta}$, where θ and $\tilde{\theta}$ only differ in t_j. With a new value for t_j the B–spline basis functions defining the spline f change, but due to the local support of the basis only the five functions B_i, $i = j-2, \ldots, j+2$, have to be recomputed.

As a first step we have to determine the set of moveable knots. With $t_0 = -\infty$ and $t_{k+1} = +\infty$, a knot $t_j \in \{t_1, \ldots, t_k\}$ is called moveable, if the number m_j of vacant candidate knots $t_{0i} \in \mathcal{T}_0$ with $t_{j-1} < t_{0i} < t_{j+1}$ is at least 1. The number $n(t)$ of moveable knots then is defined as

$$n(t) = \#\{\, t_j \text{ with } m_j > 0,\ j = 1, \ldots, k \,\}.$$

The next step is to draw the knot t_j uniformly from the $n(t)$ moveable knots with probability $p(t_j) = 1/n(t)$. Given t_j, again uniformly we draw the proposal $\tilde{t}_j$ for the new position of t_j from the set of m_j vacant candidate knots $t_{0i} \in \mathcal{T}_0$, with $p(\tilde{t}_j | t_j) = 1/m_j$. The proposal distribution for the position change results as joint distribution of $(t_j, \tilde{t}_j)$:

$$q_P(\theta, \tilde{\theta}) = p(\tilde{t}_j | t_j) p(t_j) = \frac{1}{n(t) m_j}\,.$$

For defining the acceptance probability of the position change, we need the proposal for the reverse step from $\tilde{\theta}$ to θ, i.e., the move from $\tilde{t}_j$ to t_j. The number of moveable knots here is $n(\tilde{t})$, while $\tilde{m}_j = m_j$ holds, since all candidate knots again lie in the interval (t_{j-1}, t_{j+1}). The proposal for the reverse move is then

$$q_P(\tilde{\theta}, \theta) = p(t_j | \tilde{t}_j) p(\tilde{t}_j) = \frac{1}{n(\tilde{t}) m_j}\,.$$

Since the prior (9) of $t|k$ only depends on k and $k_{\max}$, which are unchanged here, the ratio of priors is 1, and the acceptance probability is given as

$$\alpha_P(\theta, \tilde{\theta}) = \min\left\{1, \frac{p(y|\tilde{\theta})}{p(y|\theta)} \frac{n(t)}{n(\tilde{t})}\right\}.$$

3.4 Dimension change

In each sweep of the algorithm we have to choose by random, if a birth move or a death move is performed. For this choice we define the probabilities b_k for birth and d_k for death by $b_{k_{\min}} = d_{k_{\max}} = 1$, $b_{k_{\max}} = d_{k_{\min}} = 0$, and otherwise $b_k = d_k = 0.5$.

Given the k knots $t = (t_1, \ldots, t_k)$, in the birth move we add a new knot t^* lying within some interval (t_j, t_{j+1}), with $j \in \{0, 1, \ldots, k\}$ and $t_0 = -\infty$, $t_{k+1} = +\infty$. The resulting model now is defined by the new model indikator $k+1$, the new knots $\tilde{t} = (\tilde{t}_1, \ldots, \tilde{t}_{k+1})$ (with $\tilde{t}_i = t_i$ for $i \leq j$, $\tilde{t}_{j+1} = t^*$ and $\tilde{t}_i = t_{i-1}$ for $i \geq j+2$), and the new basis coefficients $\tilde{c} = (\tilde{c}_1, \ldots, \tilde{c}_{K+1})$, which have to be adjusted appropriately. Hence, the function f now is from $\mathcal{NS}_4(t_1, \ldots, t_{k+1})$, where in contrast to (3) the summation in the basis function approach is over 1 to $K+1$. Formally the birth move can be defined as a transition from state $\theta = (k, \theta_k)$ to state $\tilde{\theta} = (k+1, \tilde{\theta}_{k+1})$. With $\theta_k = (t, c, \beta)$ and $\tilde{\theta}_{k+1} = (\tilde{t}, \tilde{c}, \beta)$ there is a change in dimension from $\dim(\theta_k) = k + K + p$ to $\dim(\tilde{\theta}_{k+1}) = k + K + 2 + p$.

The death move is symmetrically defined as the reverse move from $\tilde{\theta}$ to θ.

Following Section 2.2, for the birth move we have to compute $\tilde{\theta}_{k+1}$ as function of θ_k and two random numbers t^* and u, with $u_B = (t^*, u)$. The proposal knot t^* is drawn uniformly with probability $p(t^*) = 1/(k_{\max} - k)$ from the set of the $k_{\max} - k$ vacant candidate knots $t_{0i} \in \mathcal{T}_0$.

In the literature about B–splines there exist deterministic rules for deriving $\tilde{c}$ from c when inserting a knot $t^* \in (t_j, t_{j+1})$. According to Lyche and Strøm (1996), we only have to compute the coefficients

$$\tilde{c}_i = r_i c_i + (1 - r_i) c_{i-1}, \quad i = j, j+1, j+2,$$

with $r_i \in (0, 1)$ (determined by a ratio of the knots), while for the remaining coefficients $\tilde{c}_i = c_i$, $i \leq j-1$, and $\tilde{c}_i = c_{i-1}$, $i \geq j+3$, hold. For $j \leq 2$ and $j \geq k-2$ some boundary conditions in the definition of the natural splines have to be considered. Here we only want to emphasize, that for each $j \in \{0, 1, \ldots, k\}$ $\tilde{c}_1 = c_1$ and $\tilde{c}_{k+1} = c_k$ hold.

Using these deterministic rules, the required symmetry between the birth and the death move is destroyed and the reversible jump method does not work. Hence, we only use these rules as a basic idea for deriving $\tilde{c}$ as function of c and a uniform random variate $u \in (0, 1)$. So for $2 \leq j \leq k-2$

we define $\tilde{c}_i = c_i$, $i = 1, \ldots, j-1$, $\tilde{c}_i = c_{i-1}$, $i = j+3, \ldots, k+1$, and

$$\begin{aligned} \tilde{c}_{j+1} &= uc_j + (1-u)c_{j+1} \\ \tilde{c}_j &= c_j - r_j\tilde{c}_{j+1} \\ \tilde{c}_{j+2} &= c_{j+1} - (1-r_j)\tilde{c}_{j+1}, \end{aligned} \tag{10}$$

with

$$r_j = (t^* - t_j)/(t_{j+1} - t_j). \tag{11}$$

Here $\tilde{c}_{j+1}$, the new coefficient corresponding to the new knot $t_{j+1} = t^*$, is the weighted mean of the old coefficients c_j and c_{j+1}, where the weight u is drawn randomly. The new coefficients $\tilde{c}_j$ and $\tilde{c}_{j+2}$, corresponding to the knots $\tilde{t}_j = t_j$ and $\tilde{t}_{j+2} = t_{j+1}$, are determined by the old values c_j and c_{j+1}, respectively, adjusted by the new coefficient $\tilde{c}_{j+1}$ weighted by r_j in (11). Definition (10) ensures, that in the reverse death move, given the knot $\tilde{t}_{j+1}$ to be deleted, the computation of c from $\tilde{c}$ is deterministic and the required dimension matching holds:

$$\begin{aligned} c_j &= \tilde{c}_j + r_j\tilde{c}_{j+1} \\ c_{j+1} &= \tilde{c}_{j+2} + (1-r_j)\tilde{c}_{j+1} \end{aligned}$$

and $c_i = \tilde{c}_i$, $i = 1, \ldots, j-1$, $c_i = \tilde{c}_{i+1}$, $i = j+2, \ldots, k$. To consider both the dimension matching and the boundary conditions of natural splines, we define for $j \leq 1$ and $j \geq k-1$:

$$\begin{aligned} j \in \{0,1\}: \qquad \tilde{c}_1 &= c_1, \quad \tilde{c}_i = c_{i-1},\ i = 4, \ldots, k+1 \\ \tilde{c}_2 &= uc_2 \\ \tilde{c}_3 &= c_2 - \tilde{c}_2 \end{aligned} \tag{12}$$

$$\begin{aligned} j \in \{k-1,k\}: \qquad \tilde{c}_i &= c_i,\ i = 1, \ldots, k-2, \quad \tilde{c}_{k+1} = c_k, \\ \tilde{c}_k &= uc_{k-1} \\ \tilde{c}_{k-1} &= c_{k-1} - \tilde{c}_k. \end{aligned} \tag{13}$$

Due to the definitions (12) and (13), the reverse death move again is deterministic, given the knot $\tilde{t}_{j+1}$ is to be deleted.

For calculating the acceptance probability for the birth move, consider the transformation

$$\alpha_B(\theta, \tilde{\theta}) = \min\{\, 1,\ \mathcal{L} \cdot \mathcal{A} \cdot \mathcal{P} \cdot \mathcal{J}\} \tag{14}$$

of (8), with the ratio of likelihoods $\mathcal{L} = p(y|\tilde{\theta})/p(y|\theta)$, the ratio of priors $\mathcal{A}$, the ratio of proposals $\mathcal{P}$, and the Jacobian $\mathcal{J}$.

The ratio of priors results in

$$\mathcal{A} = \frac{p(k+1)}{p(k)} \frac{p(\tilde{\theta}_{k+1}|k+1)}{p(\theta_k|k)} = s(k)\,(2\pi\sigma_0^2)^{-1/2} \exp\left[\frac{1}{2\sigma_0^2}(c'c - \tilde{c}'\tilde{c})\right],$$

where the factor $s(k)$ depends on the alternative priors of k, specified in Section 3.1. With the Poisson prior $k \sim Po(\lambda)$ we get $s(k) = \lambda/(k_{\max} - k)$, while the discrete uniform prior on $\mathcal{K}$ yields $s(k) = (k+1)/(k_{\max} - k)$.

In deriving the ratio of proposals $\mathcal{P}$, we have to consider the remarks in Section 2.2 for the case $n_1 + m_1 = m_2$, where the reverse death move is deterministic given the element to be removed (here the knot $\tilde{t}_{j+1}$). Following these remarks, $\mathcal{P}$ is given as

$$\mathcal{P} = \frac{d_{k+1}}{b_k\, p(u_B)} = \frac{d_{k+1}}{b_k\, p(t^*) p(u)}.$$

However, the knot $\tilde{t}_{j+1}$ to be removed is only known in defining the death step as reversal of the birth step with new knot $\tilde{t}_{j+1} = t^*$. In fact, in the death step the knot $\tilde{t}_{j+1}$ to be removed is not known and has to be drawn with probability $p(\tilde{t}_{j+1}) = 1/(k+1)$ from the current knots $\tilde{t}_1, \ldots, \tilde{t}_{k+1}$. Hence, the ratio of proposals results in

$$\mathcal{P} = \frac{d_{k+1}(k_{\max} - k)}{b_k(k+1)}.$$

Considering $\tilde{\theta}_{k+1}$ as function of θ_k and u_B, the Jacobian is

$$\mathcal{J} \;=\; \left|\frac{\partial \tilde{\theta}_{k+1}}{\partial(\theta_k, u_B)}\right| \;=\; \begin{cases} |c_2|, & j \in \{0, 1\} \\ |c_j - c_{j+1}|, & j \in \{2, \ldots, k-2\} \\ |c_{k-1}|, & j \in \{k-1, k\}. \end{cases}$$

Due to the symmetric definition of birth and death move, the acceptance probability of death is just the inverse of the acceptance probability (14) of birth, i.e.,

$$\alpha_D(\tilde{\theta}, \theta) = \min\left\{\, 1,\ (\mathcal{L} \cdot \mathcal{A} \cdot \mathcal{P} \cdot \mathcal{J})^{-1}\right\}.$$

3.5 Adjustments for the case of ordinary splines

With $q = 4$ and $K = k + q$, now we consider the case $f \in \mathcal{S}_q(t_1, \ldots, t_k)$, which requires only little modifications in the dimension change steps of Section 3.4, namely in the transitions between c and $\tilde{c}$, and hence in the definition of the Jacobian $\mathcal{J}$.

For the definition of the B–spline basis of the space $\mathcal{S}_q(t_1, \ldots, t_k)$, we have to consider the so–called "extended partition" $s_1 < s_2 < \ldots < s_{k+2q}$, with $s_{q+j} = t_j$ for $j = 1, \ldots, k$. The additional $2q$ knots can arbitrarily be chosen as $s_1 < \ldots < s_q < t_1$ and $t_k < s_{k+q+1} < \ldots < s_{k+2q}$.

In the birth step we insert a new knot t^* with $s_{q+j} < t^* < s_{q+j+1}$ for $j \in \{0, 1, \ldots, k\}$. Since in $\mathcal{S}_q(t_1, \ldots, t_k)$ we do not need the boundary conditions of the space $\mathcal{NS}_q(t_1, \ldots, t_k)$ of natural splines, now for all $j \in \{0, 1, \ldots, k\}$ the definition (10) holds, but we have to adjust the indices to the knots of the extended partition:

$$\begin{aligned} \tilde{c}_{j+3} &= u c_{j+2} + (1-u) c_{j+3} \\ \tilde{c}_{j+2} &= c_{j+2} - r_j \tilde{c}_{j+3} \\ \tilde{c}_{j+4} &= c_{j+3} - (1-r_j) \tilde{c}_{j+3}, \end{aligned}$$

with $r_j = (t^* - s_{j+4})/(s_{j+5} - s_{j+4})$. As a consequence, also the Jacobian simplifies to $\mathcal{J} = |c_{j+2} - c_{j+3}|$, $j \in \{0, 1, \ldots, k\}$. All other definitions made above in Section 3 remain the same.

4 Applications

In this section we illustrate the Bayesian approach to adaptive regression splines with three examples. The first two are data from the literature for curve estimation with discrete response, the Tokyo rainfall data (see e.g. Kitagawa, 1987, or Fahrmeir and Tutz, 1997) and the coal mining disaster data (see e.g. Eilers and Marx, 1996). The third example is an application of the semiparametric model to credit–scoring data described in Fahrmeir and Tutz (1997).

4.1 Rainfall data

The response is given by the number of occurrences of rainfall in Tokyo for each calender day during the years 1983 and 1984: $y_i = 0$, if there is no rainfall over 1 mm on day i in both years, $y_i = 1$, if

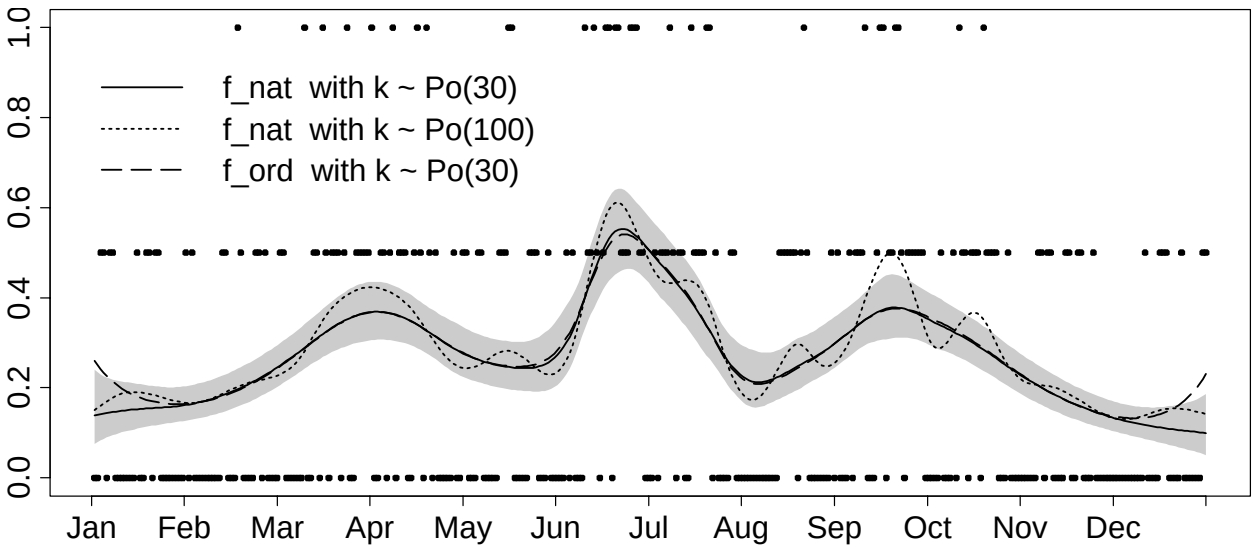

Figure 1: Rainfall data, estimates of f with different priors for k.

rain over 1 mm occurs on day i in only one of the two years, and $y_i = 2$, if in both years on day i there is rainfall over 1 mm. The big dots in Figure 1 show the data, but with response rescaled to $\{0, 0.5, 1\}$. The aim is to detect some seasonal yearly pattern for the probability π_i of rainfall, which is modelled with the logistic response function as $\pi_i = 1/(1 + \exp(-f(x_i)))$.

Figure 1 illustrates the differences between defining f as natural spline ($f_{\text{nat}} \in \mathcal{NS}_4(t_1, \ldots, t_k)$) or as ordinary spline ($f_{\text{ord}} \in \mathcal{S}_4(t_1, \ldots, t_k)$). The estimates of the natural spline (solid line) and the ordinary spline (dashed line) are rather similar, they differ only in the boundaries. While f_{ord} shows clear boundary effects, which are not explained by the data, there are none for f_{nat}. These effects result from the definition of natural and ordinary splines. For the latter there are no boundary conditions, and hence they are cubic over their whole support, while the natural cubic splines are linear at the boundaries.

The rainfall data show clear sensivity to the choice of the prior of k. A discrete uniform prior over $\mathcal{K}$ and also a Poisson prior with parameter λ about less or equal 10 yield very smooth estimates with the shape of a parabola, where no details are recognizable. This result can be explained by the fact that these priors lead in most iterations to a number k of knots between 4 and 5, which is too less for this data. A Poisson prior with λ between about 15 and 40 shows good results, i.e.,

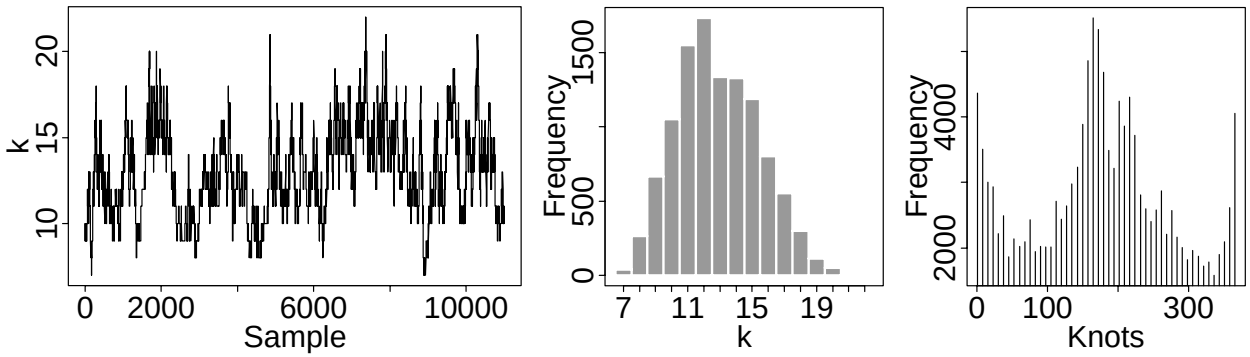

Figure 2: Rainfall data, sample of k and frequency of knots.

smooth estimates with clear details, as the estimate of f_{nat} with $k \sim Po(30)$ in Figure 1 (solid line). The shaded region is the pointwise one standard deviation confidence region. The estimate is very smooth, but reflects the character of weather in Tokyo, as described by Kitagawa (1987): dry winter, unsettled spring, clear sky in May, rainy season in late June to mid–July, stable hot summer in late July through August, generally fine but with a occasional typhoon in September and October. The dotted line in Figure 1 shows the estimate with prior $k \sim Po(100)$, which is quite similar to the estimate in Kitagawa (1987), Figure 11. This estimate lies almost everywhere in the plotted confidence region, but it is very rough and shows mostly too much details. Similar results for the rainfall data are obtained for smoothing splines and state space models, see Fahrmeir and Tutz (1997), Sections 5.3 and 8.3.

Figure 2 gives some details of the samples of k and the knots for the estimate f_{nat} with prior $k \sim Po(30)$ in Figure 1 (solid line). The left part of Figure 2 shows the sample of k with values between 7 and 21, while in the middle there is the frequency of the accepted values of k. The mode is at $k = 12$, and we see, that in more then 50% of the iterations we use a model with 11 to 14 knots. The right part of Figure 2 depicts the frequency of candidate knots $t_{0i} \in \mathcal{T}_0$ of being used as knot t_j in the model.

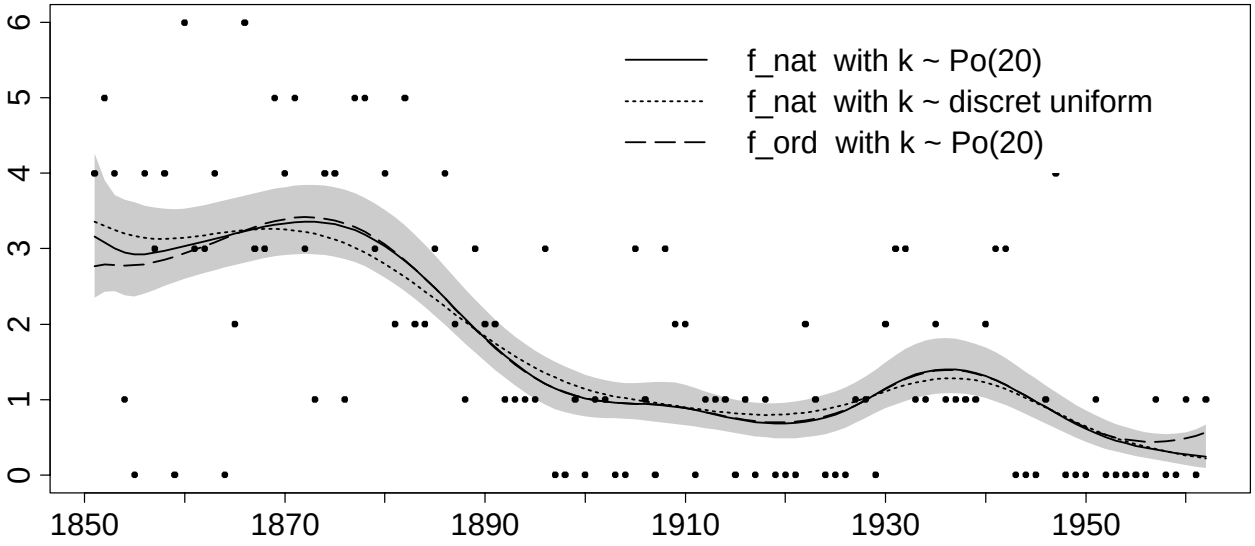

Figure 3: Coal mining data, estimates of f with different priors for k.

4.2 Coal mining disaster data

Here the response y_i is given as the number of disasters in British coal mines for the years 1850 to 1962, and is assumed to have a Poisson distribution with intensity modelled by $\mu_i = \exp(\, f(x_i)\,)$. Figure 3 shows the data, together with three different estimates for f. The estimated natural spline f_{nat} (solid line) and the estimated ordinary spline f_{ord} (dashed line) again only differ in the boundaries, but not as distinct as in the rainfall data in Figure 1. A possible explanation could be that the coal mining data are more informative than the rainfall data, and hence the boundary effects of the ordinary spline are minor. The shaded region is the pointwise one standard deviation confidence region of f_{nat} with $k \sim Po(20)$. The estimates of the natural spline f_{nat} with different priors for k, i.e., with $k \sim$ discrete uniform (dotted line) and $k \sim Po(20)$ (solid line), are similar, where the latter one shows more details. This result indicates, that the choice of the model prior here has less influence on the shape of the estimate than for the rainfall data. But it has some influence on the estimation of k. In the posterior distribution of k there is for $k \sim$ discrete uniform a mode of 6, while $k \sim Po(20)$ has a mode of 9, as can be seen in the middle of Figure 4. The left part of Figure 4 plots the sample of k. Here we see a better mixing of the chain in contrast to the sample of k for the rainfall data in Figure 2. The right part of Figure 4 depicts the frequency of

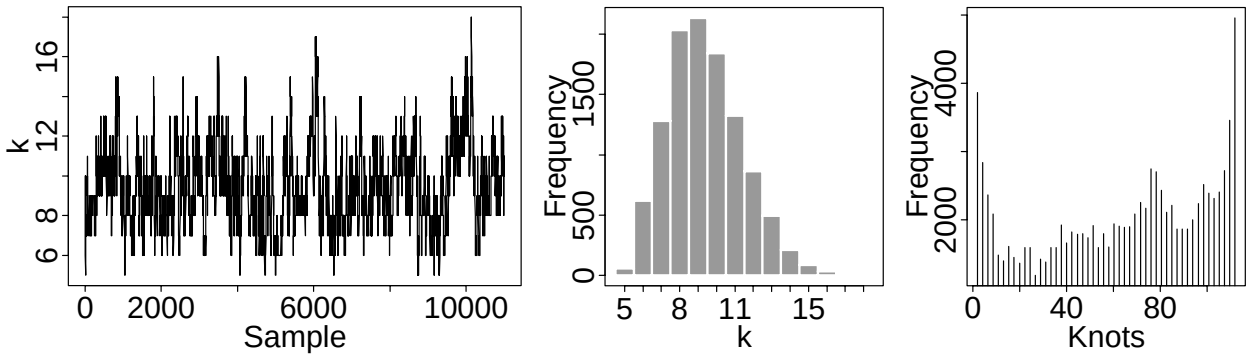

Figure 4: Coal mining data, sample of k and frequency of knots.

candidate knots $t_{0i} \in \mathcal{T}_0$ being used as knot t_j in the model.

4.3 Credit–scoring data

In credit business, banks are interested to predict the probability that a client with certain risk factors is to be considered to pay back his credits as agreed upon contract. In a data set of 1000 borrowers from a South German bank, Fahrmeir and Tutz (1997, Chapters 2.1 and 4.1) model the dichotomous response "creditability" ($y = 0$ for creditworthy, $y = 1$ for not creditworthy) in dependence of the following covariates:

x_{nr}, x_{gr} dummies for "no running account" and "good running account", respectively, with reference category "medium running account" (i.e. less than 200 DM)

x_d duration of credit in months, metrical

x_a amount of credit in DM, metrical

x_p payment of previous credits, categories "good" and "bad" (= reference category)

x_u intended use, categories "private" and "professional" (= reference category)

x_s marital status with reference category "living alone."

A parametric logistic model for the probability of being not creditworthy yields a non–significant effect of 0.000032 for the covariate "amount of credit" x_a. This unexpected result may be investi-

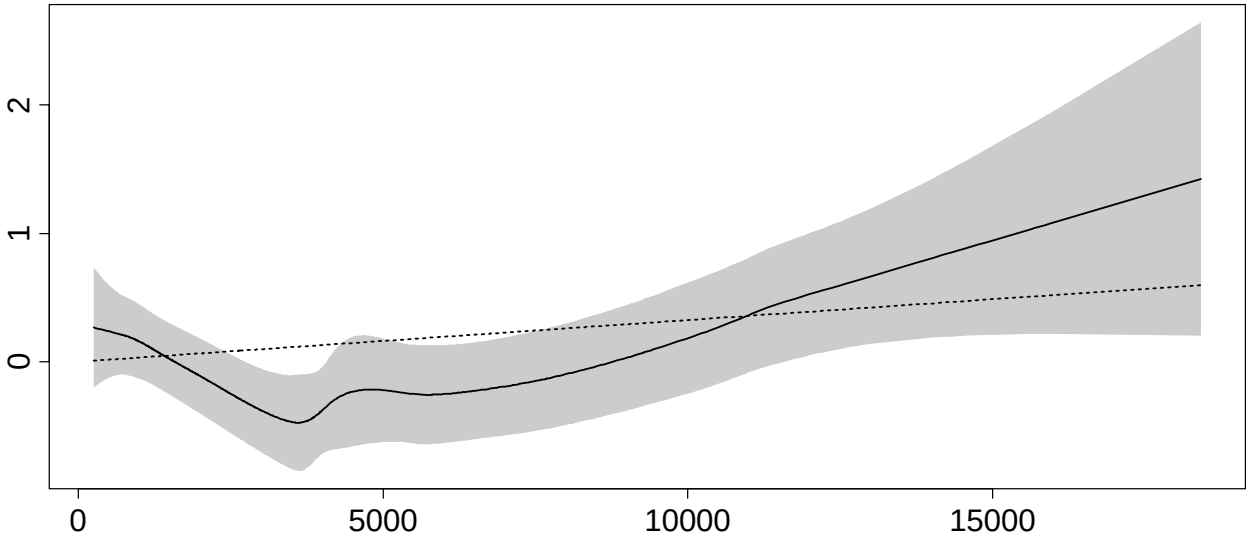

Figure 5: Credit–scoring, smooth and linear effects of "amount of credit".

gated more thoroughly by using the semiparametric model

$$\eta = x_{nr}\beta_1 + x_{gr}\beta_2 + x_d\beta_3 + x_p\beta_4 + x_u\beta_5 + x_s\beta_6 + f(x_a),$$

modelling the effect of the covariate "amount of credit" as smooth function $f(x_a)$. For the fixed effects the approach of Section 3 yields the following estimates, which are in agreement with the results of the maximum likelihood approach by Fahrmeir and Tutz (1997):

	mean	std
x_{nr}	0.647371	0.174005
x_{gr}	-1.301150	0.201329
x_d	0.041674	0.007775
x_p	-0.986134	0.231512
x_u	-0.423103	0.159075
x_s	-0.534271	0.158749

Figure 5 shows the estimate of the smooth effect of the covariate x_a together with the pointwise one standard deviation confidence region and the linear effect of x_a in the parametric logistic model (dotted line). The smooth effect is clear nonlinear with a bath–tub shape, indicating that both

high and low credits increase the risk of being not creditworthy, compared to medium credits. If we assume the influence of the covariate "amount of credit" to be linear, the estimated effect is near zero, falsely leading to the conclusion being non–significant.

5 Conclusions

The proposed Bayesian regression spline approach with adaptive knot placement using reversible jump Markov chain Monte Carlo gives a flexible but also smooth fit to the data with only few knots (e.g. the coal mining example: $k = 9$ in the mode). In comparison to overparameterized methods such as nonadaptive smoothing splines or state space models it is able to work out the shape of the curve in detail, but very smoothly without disturbing spikes, see the comments in Section 4.1 for the rainfall data. Since the smoothness of the curve is determined by the number of knots k, which is simultaneously chosen by the method, no further smoothing parameters have to be found by additional methods as, e.g., cross validation. Moreover, the results of the coal mining example in Section 4.2 indicate, that the prior of the model indicator k has only little influence on the smoothness of f, if there is enough information in the data. Similar results could be presented for the credit–scoring data in Section 4.3.

The choice between natural and ordinary splines should only be meaningful in situations with data getting sparse in the boundaries, as in event history analysis. Here natural splines may help to avoid boundary effects.

The following extensions to the approach proposed in Section 3 are fields for future research: First, the adjustment of the spline basis to account for jumps and discontinuities in the function f. Proposals for the B–spline basis are made, e.g., in Schumaker (1993) or Lyche and Strøm (1996). Further, due to the Bayesian approach using Markov chain Monte Carlo methods, extensions of the semiparametric model to generalized additive models (Hastie and Tibshirani, 1990) or the more general varying–coefficient models (Hastie and Tibshirani, 1993) are possible without much problems. With regard to Markov chain Monte Carlo methods, other approaches for updating fixed effects in the generalized linear model (Section 3.2) will be considered, since the applied method of Gamerman (1997) is intensive in computing time, though it has good mixing and convergence

properties. A possible approach is the slicing method, recently proposed by Neal (1997). Since in our approach the number of unknown parameters is varying, the recently defined Bayesian *Deviance Information Criterion* (DIC) of Spiegelhalter, Best and Carlin (1998) could help to measure the complexity and the fit of the model.

Acknowledgement: I like to thank Prof. Dr. L. Fahrmeir for his stimulating discussions and for supervising my research project. 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**(1): 3–66.

Dellaportas, P. and Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via gibbs sampling, *Applied Statistics* **42**(3): 443–459.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic bayesian curve fitting, *J. R. Statist. Soc. B* **60**(2): 333–350.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b–splines and penalties, *Statistical Science* **11**(2): 89–121.

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*, 3. edn, Springer–Verlag, New York.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion), *Technometrics* **31**(1): 3–39.

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

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995). Adaptive rejection metropolis sampling within gibbs sampling, *Applied Statistics* **44**(4): 455–472.

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

Gilks, W. R. and Wild, P. (1995). Adaptive rejection sampling for gibbs sampling, *Applied Statistics* **41**(2): 337–348.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination, *Biometrika* **82**(4): 711–732.

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

Green, P. J. and Yandell, B. S. (1985). Semi–parametric generalized linear models, *in* R. Gilchrist, B. Francis and J. Whittaker (eds), *Generalized Linear Models*, Springer Lecture Notes, Heidelberg, pp. 44–55.

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

Hastie, T. and Tibshirani, R. (1993). Varying–coefficient models, *J. R. Statist. Soc. B* **55**(4): 757–796.

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

Lyche, T. and Schumaker, L. L. (1973). Computation of smoothing and interpolating natural splines via local bases, *SIAM J. Numer. Anal.* **10**(6): 1027–1038.

Lyche, T. and Strøm, K. (1996). Knot insertion for natural splines, *Annals of Numerical Mathematics* **3**: 221–246.

Neal, R. M. (1997). Markov chain monte carlo methods based on 'slicing' the density function, *Technical Report 9722*, Department of Statistics, University of Toronto.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion), *J. R. Statist. Soc. B* **59**(4): 731–792.

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

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 Troung, Y. K. (1997). 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**(4): 1701–1762.