Generalized structured additive regression based on Bayesian P-splines

Andreas Brezger, Stefan Lang [*]

Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany

Abstract

Generalized additive models (GAM) for modeling nonlinear effects of continuous covariates are now well established tools for the applied statistician. A Bayesian version of GAM's and extensions to generalized structured additive regression (STAR) are developed. One or two dimensional P-splines are used as the main building block. Inference relies on Markov chain Monte Carlo (MCMC) simulation techniques, and is either based on iteratively weighted least squares (IWLS) proposals or on latent utility representations of (multi)categorical regression models. The approach covers the most common univariate response distributions, e.g. the binomial, Poisson or gamma distribution, as well as multicategorical responses. For the first time, Bayesian semiparametric inference for the widely used multinomial logit model is presented. Two applications on the forest health status of trees and a space-time analysis of health insurance data demonstrate the potential of the approach for realistic modeling of complex problems. Software for the methodology is provided within the public domain package *BayesX*.

Key words: geoadditive models, IWLS proposals, multicategorical response, structured additive predictors, surface smoothing

[*] Corresponding author. Tel. +49-89-2180-6404; fax: +49-89-2180-5040
Email addresses: `andib@stat.uni-muenchen.de` (Andreas Brezger), `stefan.lang@stat.uni-muenchen.de` (Stefan Lang).

[1] This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen". We are very grateful to two anonymous referees for their many valuable suggestions to improve the first version of the paper.

1 Introduction

Generalized additive models (GAM) provide a powerful class of models for modeling nonlinear effects of continuous covariates in regression models with non-Gaussian responses. A considerable number of competing approaches is now available for modeling and estimating nonlinear functions of continuous covariates. Prominent examples are smoothing splines (e.g. Hastie and Tibshirani, 1990), local polynomials (e.g. Fan and Gijbels, 1996), regression splines with adaptive knot selection (e.g. Friedman and Silverman, 1989; Friedman, 1991; Stone, Hansen, Kooperberg and Truong, 1997) and P-splines (Eilers and Marx, 1996; Marx and Eilers, 1998). Currently, smoothing based on mixed model representations of GAM's and extensions is extremely popular, see e.g. Lin and Zhang (1999), Currie and Durban (2002), Wand (2003) and the book by Ruppert, Wand and Carroll (2003). Indeed, the approach is very promising and has several distinct advantages, e.g. smoothing parameters can be estimated simultaneously with the regression functions.

Bayesian approaches are currently either based on regression splines with adaptive knot selection (e.g. Smith and Kohn, 1996; Denison, Mallick and Smith, 1998; Biller, 2000; Di Matteo, Genovese and Kass, 2001; Biller and Fahrmeir, 2001; Hansen and Kooperberg, 2002), or on smoothness priors (Hastie and Tibshirani (2000) and own work). Inference is based on Markov chain Monte Carlo inference techniques. A nice introduction into MCMC can be found in Green (2001).

This paper can be seen as the final in a series of articles on Bayesian semiparametric regression based on smoothness priors and MCMC simulation techniques, see particularly Fahrmeir and Lang (2001a), Fahrmeir and Lang (2001b) and Lang and Brezger (2004). A very general class of models with *structured additive predictor* (STAR) is proposed and MCMC algorithms for posterior inference are developed. STAR models include a number of model classes well known from the literature as special cases. Examples are generalized additive mixed models (e.g. Lin and Zhang, 1999), geoadditive models (Kammann and Wand, 2003), varying coefficient models (Hastie and Tibshirani, 1993) or geographically weighted regression (Fotheringham, Brunsdon and Charlton, 2002). The approach covers the most common univariate response distributions (Gaussian, binomial, Poisson, gamma) as well as models for multicategorical responses.

In a first paper Fahrmeir and Lang (2001a) develop univariate Generalized additive mixed models based on random walk and Markov random field priors. Inference is based on MCMC simulation based on conditional prior proposals suggested by Knorr-Held (1999) in the context of dynamic models. Fahrmeir and Lang (2001b) extend the methodology to models with multicategorical responses. As an alternative to conditional prior proposals, more efficient MCMC techniques that rely on latent utility representations are proposed, see also Albert and Chib (1993), Chen and Dey (2000) and Holmes and Held (2003). In Lang and Brezger (2004) Bayesian versions of P-splines for modeling nonlinear effects of continuous covariates and time scales are proposed. Additionally, two dimensional P-splines for modeling interactions between continuous covariates are developed. P-splines usually perform better in terms of bias and mean squared error than the simple random walk priors used in Fahrmeir and Lang (2001a) and Fahrmeir and Lang (2001b). In fact, Bayesian P-splines contain random walk priors as a special case. MCMC inference makes use of matrix operations for band or more generally sparse matrices but is restricted to Gaussian responses. The plan of this paper is the following:

- Models with STAR predictor are described in a more general form than the previous papers and a unified notation is used. We thereby start with generalized additive models and gradually extend the approach to incorporate unit- or cluster specific heterogeneity, spatial effects and interactions between covariates of different types.
- We extend the inference techniques for Gaussian responses in Lang and Brezger (2004) to situations with fundamentally non-Gaussian responses. We develop a number of highly efficient updating schemes with iteratively weighted least squares (IWLS) used for fitting generalized linear models as the main building block. The proposed updating schemes are much more efficient in terms of mixing of the Markov chains and computing time than conditional prior proposals used in Fahrmeir and Lang (2001a). Related algorithms have been proposed by Gamerman (1997) and Lenk and DeSarbo (2000) for estimating Bayesian generalized linear mixed models. Compare also Rue (2001) and Knorr-Held and Rue (2002) who develop efficient MCMC updating schemes for spatial smoothing of poisson responses.
- For the first time, we present semiparametric Bayesian inference for multinomial logit models.
- For categorical responses we review the state of the art of MCMC inference techniques based on latent utility representations. A comparison with the direct sampling schemes (see the second point) is made.

Our Bayesian approach for semiparametric regression has the following advantages compared to existing methodology:

- *Extendability to more complex formulations*
  A main advantage of a Bayesian approach based on MCMC techniques is its flexibility and extendability to more complex formulations. Our approach may also be used as a starting point for Bayesian inference in other model classes or more specialized settings. For example Hennerfeind, Brezger and Fahrmeir (2003) build on the inference techniques of this paper for developing geoadditive survival models.

- *Inference for functions of the parameters*
  Another important advantage of inference based on MCMC is easy prediction for unobserved covariate combinations including credible intervals, and the availability of inference for functions of the parameters (again including credible intervals). We will give specific examples in our second application.
- *Estimating models with a large number of parameters and observations*
  In Fahrmeir, Kneib and Lang (2004) we compare the relative merits of the full Bayesian approach presented here, and empirical Bayesian inference based on mixed model technology where the smoothing parameters are estimated via restricted maximum likelihood. Although the standard methodology from the literature has been improved the algorithms are still of the order p^3 where p is the total number of parameters. Similar problems arise if the smoothing parameters are estimated via GCV, see Wood (2000). In our Bayesian approach based on MCMC techniques we can use a divide and conquer strategy similar to backfitting. The difference to backfitting is, however, that we are able to estimate the smoothing parameters simultaneously with the regression parameters with almost negligible additional effort. We are therefore able to handle problems with more than 1000 parameters and 200000 observations.

The methodology of this paper is included in the public domain program *BayesX*, a software package for Bayesian inference. It may be downloaded including a detailed manual from
`http://www.stat.uni-muenchen.de/~lang/bayesx`.
As a particular advantage *BayesX* can estimate reasonable complex models and handle fairly large data sets.

We will present examples of STAR models in two applications. In our first application we analyze longitudinal data on the health status of beeches in northern Bavaria. Important influential factors on the health state of trees are e.g. the age of the trees, the canopy density at the stand, calendar time as a surrogate for changing environmental conditions, and the location of the stand. The second application is a space-time analysis of hospital treatment costs based on data from a German private health insurance company.

The remainder of this paper is organized as follows: The next section describes Bayesian GAM's based on one or two dimensional P-splines and discusses extensions to STAR models. Section 3 gives details about MCMC inference. In Section 4 we present two applications on the health status of trees and hospital treatment costs. Section 5 concludes and discusses directions for future research.

2 Bayesian STAR models

2.1 GAM's based on Bayesian P-splines

Suppose that observations (y_i, x_i, v_i), $i = 1, \dots, n$, are given, where y_i is a response variable, $x_i = (x_{i1}, \dots, x_{ip})'$ is a vector of continuous covariates and $v_i = (v_{i1}, \dots, v_{iq})'$ are further (mostly categorical) covariates. Generalized additive models (Hastie and Tibshirani, 1990) assume that, given x_i and v_i the distribution of y_i belongs to an exponential family, i.e.

$$p(y_i \,|\, x_i, v_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\right) c(y_i, \phi), \tag{1}$$

where $b(\cdot)$, $c(\cdot)$, θ_i and ϕ determine the respective distributions. A list of the most common distributions and their specific parameters can be found e.g. in Fahrmeir and Tutz (2001), page 21. The mean $\mu_i = E(y_i \,|\, x_i, v_i)$ is linked to a semiparametric additive predictor η_i by

$$\mu_i = h(\eta_i), \quad \eta_i = f_1(x_{i1}) + \dots + f_p(x_{ip}) + v_i'\gamma. \tag{2}$$

Here, h is a known response function and $f_1, \dots, f_p$ are unknown smooth functions of the continuous covariates and $v_i'\gamma$ represents the strictly linear part of the predictor. Note that the mean levels of the unknown functions f_j are not identifiable. To ensure identifiability, the functions f_j are constrained to have zero means. This can be incorporated into estimation via MCMC by centering the functions f_j about their means in every iteration of the sampler. To avoid that the posterior is changed the subtracted means are added to the intercept (included in $v_i'\gamma$).

For modeling the unknown functions f_j we follow Lang and Brezger (2004), who present a Bayesian version of P-splines introduced in a frequentist setting by Eilers and Marx (1996) and Marx and Eilers (1998). The approach assumes that the unknown functions can be approximated by a polynomial spline of degree l and with equally spaced knots

$$\zeta_{j0} = x_{j,min} < \zeta_{j1} < \ldots < \zeta_{j,k_j-1} < \zeta_{jk_j} = x_{j,max}$$

over the domain of x_j. The spline can be written in terms of a linear combination of $M_j = k_j + l$ B-spline basis functions (De Boor, 1978). Denoting the m-th basis function by B_{jm}, we obtain

$$f_j(x_j) = \sum_{m=1}^{M_j} \beta_{jm} B_{jm}(x_j).$$

By defining the $n \times M_j$ design matrices X_j with the elements in row i and column m given by $X_j(i,m) = B_{jm}(x_{ij})$, we can rewrite the predictor (2) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p + V\gamma. \tag{3}$$

Here, $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})'$, $j = 1, \ldots, p$, correspond to the vectors of unknown regression coefficients. The matrix V is the usual design matrix for linear effects. To overcome the well known difficulties involved with regression splines, Eilers and Marx (1996) suggest a relatively large number of knots (usually between 20 to 40) to ensure enough flexibility, and to introduce a roughness penalty on adjacent regression coefficients to regularize the problem and avoid overfitting. In their frequentist approach they use penalties based on squared r-th order differences. Usually first or second order differences are enough. In our Bayesian approach, we replace first or second order differences with their stochastic analogues, i.e. first or second order random walks defined by

$$\beta_{jm} = \beta_{j,m-1} + u_{jm}, \quad \text{or} \quad \beta_{jm} = 2\beta_{j,m-1} - \beta_{j,m-2} + u_{jm}, \tag{4}$$

with Gaussian errors $u_{jm} \sim N(0, \tau_j^2)$ and diffuse priors $\beta_{j1} \propto \mathit{const}$, or β_{j1} and $\beta_{j2} \propto \mathit{const}$, for initial values, respectively. The amount of smoothness is controlled by the variance parameter τ_j^2 which corresponds to the inverse smoothing parameter in the traditional approach. By defining an additional hyperprior for the variance parameters the amount of smoothness can be estimated simultaneously with the regression coefficients. We assign the conjugate prior for τ_j^2 which is an inverse gamma prior with hyperparameters a_j and b_j, i.e. $\tau_j^2 \sim IG(a_j, b_j)$. Common choices for a_j and b_j are $a_j = 1$ and b_j small, e.g. $b = 0.005$ or $b_j = 0.0005$. Alternatively we may set $a_j = b_j$, e.g. $a_j = b_j = 0.001$. Based on experience from extensive simulation studies we use $a_j = b_j = 0.001$ as our standard choice. Since the results may considerably depend on the choice of a_j and b_j some sort of sensitivity analysis is strongly recommended. For instance, the models under consideration could be re-estimated with (a small) number of different choices for a_j and b_j.

In some situations, a global variance parameter τ_j^2 may be not appropriate, for example if the underlying function is highly oscillating. In such cases the assumption of a global variance parameter τ_j^2 may be relaxed by replacing the errors $u_{jm} \sim N(0, \tau_j^2)$ in (4) by $u_{jm} \sim N(0, \tau_j^2/\delta_{jm})$. The weights δ_{jm} are additional hyperparameters and assumed to follow independent gamma distributions $\delta_{jm} \sim G(\frac{\nu}{2}, \frac{\nu}{2})$. This is equivalent to a t-distribution with ν degrees of freedom for β_j (see e.g. Knorr-Held (1996) in the context of dynamic models). As an alternative, *locally adaptive dependent* variances as proposed in Lang, Fronk and Fahrmeir (2002) and Jerak and Lang (2003) could be used as well. Our software is capable of estimating such models, but we do not investigate them in the following. However, estimation is straightforward, see Lang and Brezger (2004), Lang, Fronk and Fahrmeir (2002) and Jerak and Lang (2003) for details.

2.2 Modeling interactions

In many situations, the simple additive predictor (2) may be not appropriate because of interactions between covariates. In this section we describe interactions between categorical and continuous covariates, and between two continuous covariates. In the next section, we also discuss interactions between space and categorical covariates. For simplicity, we keep the notation of the predictor as in (2) and assume for the rest of the section that x_j is now two dimensional, i.e. $x_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)})'$.

Interactions between categorical and continuous covariates can be conveniently modeled within the varying coefficient framework introduced by Hastie and Tibshirani (1993). Here, the effect of covariate $x_{ij}^{(1)}$ is assumed to vary smoothly over the range of the second covariate $x_{ij}^{(2)}$, i.e.

$$f_j(x_{ij}) = g\left(x_{ij}^{(2)}\right) x_{ij}^{(1)}. \tag{5}$$

The covariate $x_{ij}^{(2)}$ is called the effect modifier of $x_{ij}^{(1)}$. The design matrix X_j is given by $diag(x_{1j}^{(1)}, \dots, x_{nj}^{(1)})X_j^{(2)}$ where $X_j^{(2)}$ is the usual design matrix for splines composed of the basis functions evaluated at the observations $x_{ij}^{(2)}$.

If both interacting covariates are continuous, a more flexible approach for modeling interactions can be based on two dimensional surface fitting. Here, we concentrate on two dimensional P-splines described in Lang and Brezger (2004), see also Wood (2003) for a recent approach based on thin plate splines. We assume that the unknown surface $f_j(x_{ij})$ can be approximated by the tensor product of one dimensional B-splines, i.e.

$$f_j(x_{ij}^{(1)}, x_{ij}^{(2)}) = \sum_{m_1=1}^{M_{1j}} \sum_{m_2=1}^{M_{2j}} \beta_{j,m_1 m_2} B_{j,m_1}\left(x_{ij}^{(1)}\right) B_{j,m_2}\left(x_{ij}^{(2)}\right). \tag{6}$$

The design matrix X_j is now $n \times (M_{1j} \cdot M_{2j})$ dimensional and consists of products of basis functions. Priors for $\beta_j = (\beta_{j,11}, \dots, \beta_{j,M_{1j}M_{2j}})'$ are based on spatial smoothness priors common in spatial statistics (see e.g. Besag and Kooperberg (1995)). Based on previous experience, we prefer a two dimensional first order random walk constructed from the four nearest neighbors. It is usually defined by specifying the conditional distributions of a parameter given its neighbors, i.e.

$$\beta_{jm_1m_2} \,|\, \cdot \sim N\left(\frac{1}{4}(\beta_{jm_1-1,m_2} + \beta_{jm_1+1,m_2} + \beta_{jm_1,m_2-1} + \beta_{jm_1,m_2+1}), \frac{\tau_j^2}{4}\right) \tag{7}$$

for $m_1 = 2, \ldots, M_{1j} - 1$, $m_2 = 2, \ldots, M_{2j} - 1$ and appropriate changes for corners and edges. Again, we restrict the unknown function f_j to have mean zero to guarantee identifiability.

Sometimes it is desirable to decompose the effect of the two covariates $x_j^{(1)}$ and $x_j^{(2)}$ into two main effects modeled by one dimensional functions and a two dimensional interaction effect. Then, we obtain

$$f_j(x_{ij}) = f_j^{(1)}\left(x_{ij}^{(1)}\right) + f_j^{(2)}\left(x_{ij}^{(2)}\right) + f_j^{(1|2)}\left(x_{ij}^{(1)}, x_{ij}^{(2)}\right). \tag{8}$$

In this case, additional identifiability constraints have to be imposed on the three functions, see Lang and Brezger (2004).

2.3 Unobserved heterogeneity

So far, we have considered only continuous and categorical covariates in the predictor. In this section, we relax this assumption by allowing that the covariates x_j in (2) or (3) are not necessarily continuous. We still pertain the assumption of the preceding section that covariates x_j may be one or two dimensional. Based on this assumptions the models can be considerably extended within a unified framework. We are particularly interested in the handling of unobserved unit- or cluster specific and spatial heterogeneity. Models that can deal with spatial heterogeneity are also called *geoadditive models* (Kammann and Wand (2003)).

Unit- or cluster specific heterogeneity

Suppose that covariate x_j is an index variable that indicates the unit or cluster a particular observation belongs to. An example are longitudinal data where x_j is an individual index. In this case, it is common practice to introduce unit- or cluster specific i.i.d. Gaussian random intercepts or slopes, see e.g. Diggle, Haegerty, Liang and Zeger (2002). Suppose x_j can take the values $1, \ldots, M_j$. Then, an i.i.d. random intercept can be incorporated into our framework of structured additive regression by assuming $f_j(m) = \beta_{jm} \sim N(0, \tau_j^2)$, $m = 1, \ldots, M_j$. The design matrix X_j is now a 0/1 incidence matrix with dimension $n \times M_j$. In order to introduce random slopes we assume $x_j = \left(x_j^{(1)}, x_j^{(2)}\right)$

as in Section 2.2. Then, a random slope with respect to index variable $x_j^{(2)}$ is defined as $f_j(x_{ij}) = g\left(x_{ij}^{(2)}\right) x_{ij}^{(1)}$ with $g\left(x_{ij}^{(2)}\right) = \beta_{jm} \sim N(0, \tau_j^2)$. The design matrix X_j is given by $diag\left(x_{1j}^{(1)}, \ldots, x_{nj}^{(1)}\right) X_j^{(2)}$ where $X_j^{(2)}$ is again a 0/1 incidence matrix. Note the close similarity between random slopes and varying coefficient models. In fact, random slopes may be regarded as varying coefficient terms with unit- or cluster variable $x_j^{(2)}$ as the effect modifier.

Spatial heterogeneity

To consider spatial heterogeneity, we may introduce a *spatial effect* f_j of location x_j to the predictor. Depending on the application, the spatial effect may be further split up into a spatially correlated (structured) and an uncorrelated (unstructured) effect, i.e. $f_j = f_{str} + f_{unstr}$. The correlated effect f_{str} aims at capturing spatially dependent heterogeneity and the uncorrelated effect f_{unstr} local effects.

For data observed on a regular or irregular lattice a common approach for the correlated spatial effect f_{str} is based on Markov random field (MRF) priors, see e.g. Besag, York and Mollie (1991). Let $s \in \{1, \ldots, S_j\}$ denote the pixels of a lattice or the regions of a geographical map. Then, the most simple Markov random field prior for $f_{str}(s) = \beta_{str,s}$ is defined by

$$\beta_{str,s} \mid \beta_{str,u}, u \neq s \sim N\left(\sum_{u \in \partial_s} \frac{1}{N_s} \beta_{str,u}, \frac{\tau_{str}^2}{N_s}\right), \tag{9}$$

where N_s is the number of adjacent regions or pixels, and ∂_s denotes the regions which are neighbors of region s. Hence, prior (9) can be seen as a two dimensional extension of a first order random walk. More general priors than (9) are described in Besag et al. (1991). The design matrix X_{str} is a $n \times S_j$ incidence matrix whose entry in the i-th row and s-th column is equal to one if observation i has been observed at location s and zero otherwise.

Alternatively, the structured spatial effect f_{str} could be modeled by two dimensional surface estimators as described in Section 2.2. In most of our applications, however, the MRF proves to be superior in terms of model fit.

For the unstructured effect f_{unstr} we may again assume i.i.d. Gaussian random effects with the location as the index variable.

Similar to continuous covariates and index variables we can again define varying coefficient terms, now with the location index as the effect modifier, see e.g. Fahrmeir, Lang, Wolff and Bender (2003) and Gamerman, Moreira and Rue (2003) for applications. Models of this kind are known in the geography literature as *geographically weighted regression* (Fotheringham, Brunsdon and Charlton (2002)).

2.4 General structure of the priors

As we have pointed out, it is always possible to express the vector of function evaluations $f_j = (f_{j1}, \dots, f_{jn})$ of a covariate effect as the matrix product of a design matrix X_j and a vector of regression coefficients β_j, i.e. $f_j = X_j\beta_j$. It turns out that the smoothness priors for the regression coefficients β_j can be cast into a general form as well. It is given by

$$\beta_j \mid \tau_j^2 \propto \frac{1}{(\tau_j^2)^{rk(K_j)/2}} \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right), \tag{10}$$

where K_j is a *penalty matrix* which depends on the prior assumptions about *smoothness of* f_j and the *type of covariate.*

For the variance parameter an inverse gamma prior (the conjugate prior) is assumed, i.e. $\tau_j^2 \sim IG(a_j, b_j)$.

The general structure of the priors particularly facilitates the description and implementation of MCMC inference in the next section.

3 Bayesian inference via MCMC

Bayesian inference is based on the posterior of the model which is given by

$$\begin{aligned} p(\alpha \mid y) \propto\ & L(y, \beta_1, \tau_1^2, \dots, \beta_p, \tau_p^2, \gamma) \\ & \prod_{j=1}^{p} \frac{1}{(\tau_j^2)^{rk(K_j)/2}} \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right) \prod_{j=1}^{p} (\tau_j^2)^{-a_j-1} \exp\left(-\frac{b_j}{\tau_j^2}\right), \end{aligned}$$

where α is the vector of all parameters in the model. The likelihood $L(\cdot)$ is a product of the individual likelihoods (1). Since the posterior is analytically intractable we make use of MCMC simulation techniques. Models with Gaussian responses are already covered in Lang and Brezger (2004). Here, the main focus is on methods applicable for general distributions from an exponential family. We first develop in Section 3.1 several sampling schemes based on iteratively weighted least squares (IWLS) used for estimating generalized linear models (Fahrmeir and Tutz, 2001). For many models with (multi)categorical responses alternative sampling schemes can be developed by considering their latent utility representations (Section 3.2). In either case, MCMC simulation is based on drawings from full conditionals of blocks of parameters, given the rest and the data. We use the blocks $\beta_1, \dots, \beta_p, \tau_1^2, \dots, \tau_p^2, \gamma$ (sampling scheme 1) or $(\beta_1, \tau_1^2), \dots, (\beta_p, \tau_p^2), \gamma$ (sampling scheme 2).

An alternative to MCMC techniques is proposed in Fahrmeir, Kneib and Lang (2004). Here, mixed model representations and inference techniques are used for estimation. The drawback is that models with a large number of parameters and/or observations as well as multivariate responses can not be handled by the approach.

3.1 Updating by iteratively weighted least squares (IWLS) proposals

The basic idea is to combine Fisher scoring or IWLS (e.g. Fahrmeir and Tutz, 2001) for estimating regression parameters in generalized linear models, and the Metropolis-Hastings algorithm. More precisely, the goal is to approximate the full conditionals of regression parameters β_j and γ by a Gaussian distribution, obtained by accomplishing *one* Fisher scoring step in every iteration of the sampler. Suppose we want to update the regression coefficients β_j of the function f_j with current state β_j^c of the chain. Denote η^c the current predictor based on the current regression coefficients β_j^c. Then, according to IWLS, a new value β_j^p is proposed by drawing a random number from the multivariate Gaussian proposal distribution $q(\beta_j^c, \beta_j^p)$ with precision matrix and mean

$$P_j = X_j' W(\eta^c) X_j + \frac{1}{\tau_j^2} K_j, \quad m_j = P_j^{-1} X_j' W(\eta^c)(\tilde{y}(\eta^c) - \tilde{\eta}^c). \tag{11}$$

The matrix $W(\eta^c) = diag(w_1(\eta^c), \dots, w_n(\eta^c))$ and the vector $\tilde{y}(\eta^c)$ contain the usual weights and working observations for IWLS with $w_i^{-1}(\eta_i^c) = b''(\theta_i)\{g'(\mu_i)\}^2$ and $\tilde{y}(\eta_i^c) = \eta_i^c + (y_i - \mu_i)g'(\mu_i)$. The weights and the working observations depend on the current predictor η^c which in turn depends on the current state β_j^c. The vector $\tilde{\eta}^c$ is the part of the predictor associated with all remaining effects in the model. The proposed new value β_j^p is accepted with probability

$$\alpha(\beta_j^c, \beta_j^p) = \frac{L(y, \dots, \beta_j^p, \dots, \gamma^c) p(\beta_j^p \mid (\tau_j^2)^c) q(\beta_j^p, \beta_j^c)}{L(y, \dots, \beta_j^c, \dots, \gamma^c) p(\beta_j^c \mid (\tau_j^2)^c) q(\beta_j^c, \beta_j^p)}. \tag{12}$$

Computation of the acceptance probability requires the evaluation of the normalizing constant of the IWLS proposal which is given by $|P_j|^{0.5}$. The determinant of P_j can be computed without significant additional effort as a by-product of the Cholesky decomposition. The computation of the likelihood $L(y, \dots, \beta_j^p, \dots, \gamma^c)$ and the proposal density $q(\beta_j^p, \beta_j^c)$ is based on the current predictor η^c where $X_j\beta_j^c$ is exchanged by $X_j\beta_j^p$, i.e. $\eta^c = \eta^c + X_j(\beta_j^p - \beta_j^c)$.

In order to emphasize implementation aspects and details we use here and elsewhere a pseudo code like notation. Note that the computation of $q(\beta_j^p, \beta_j^c)$ requires to recompute P_j and m_j. If the proposal is accepted we set $\beta_j^c = \beta_j^p$, otherwise we keep the current β_j^c and exchange $X_j\beta_j^p$ in η^c by $X_j\beta_j^c$.

A slightly different sampling scheme uses the current posterior mode approximation m_j^c rather than β_j^c for computing the IWLS weight matrix W and the transformed responses $\tilde{y}$ in (11). More precisely, we first replace $X_j\beta_j^c$ in the current predictor η^c by $X_jm_j^c$, i.e. $\eta^c = \eta^c + X_j(m_j^c - \beta_j^c)$. The vector m_j^c is the mean of the proposal distribution used in the last iteration of the sampler. We proceed by drawing a proposal β_j^p from the Gaussian distribution with covariance and mean (11). The difference of using m_j^c rather than β_j^c in η^c is that the proposal is *independent* of the current state of the chain, i.e. $q(\beta_j^c, \beta_j^p) = q(\beta_j^p)$. Hence, it is not required to recompute P_j and m_j when computing the proposal density $q(\beta_j^p, \beta_j^c)$ in (12). The computation of the likelihood $L(y, \ldots, \beta_j^p, \ldots, \gamma^c)$ is again based on the current predictor η^c where $X_jm_j^c$ is exchanged by $X_j\beta_j^p$, i.e. $\eta^c = \eta^c + X_j(\beta_j^p - m_j^c)$. If the proposal is accepted we set $\beta_j^c = \beta_j^p$, otherwise we keep the current β_j^c and exchange $X_j\beta_j^p$ in η^c by $X_j\beta_j^c$. The last step is to set $m_j^c = m_j$.

The advantage of the updating scheme based on the current mode approximation m_j^c is that acceptance rates are considerably higher compared to the sampling scheme based on the current β_j^c. This is particularly important for updating spatial effects based on Markov random field priors because of the usually high dimensional parameter vector β_j.

It turns out that convergence to the stationary distribution can be slow for both algorithms because of inappropriate starting values for the β_j. As a remedy, we initialize the Markov chain with posterior mode estimates which are obtained from a backfitting algorithm with fixed and usually large values for the variance parameters. Too small variances often result in initial regression parameter estimates close to zero and the subsequent MCMC algorithm is stuck for a considerable number of iterations. Large variances may lead to quite wiggled estimates as starting values but guarantees fast convergence of the MCMC sampler. In principle, posterior mode estimates could be obtained directly without backfitting, see e.g. Marx and Eilers (1998). Note, however, that for direct computation of posterior modes high dimensional linear equation systems have to be solved because in our models the total number of parameters is usually large. Models with more than 1000 parameters are quite common. The reason is that quite often Markov random fields or random effects for modeling spatial effects or unobserved heterogeneity are involved.

Note, that the posterior precision matrix P_j in (11) is a band matrix or can be at least transformed into a matrix with band structure. For one dimensional P-splines, the band size is $max\{\text{degree of spline}, \text{order of differences}\}$, for two dimensional P-splines the band size is $M_j \cdot l + l$, and for i.i.d. random effects the posterior precision matrix is diagonal. For a Markov random field, the precision matrix is not a priori a band matrix but sparse. It can be transformed into a band matrix (with differing band size in every row) by reordering the regions using the reverse Cuthill Mc-Kee algorithm (see George and Liu (1981) p. 58 ff). Hence, random numbers from the (high dimensional) proposal distributions can be efficiently drawn by using matrix operations for sparse matrices, in particular Cholesky decompositions. In our implementation we use the *envelope method* for Cholesky decompositions of sparse matrices as described in George and Liu (1981), see also Rue (2001) and Lang and Brezger (2004).

Updating of the variance parameters τ_j^2 is straightforward because their full conditionals are inverse gamma distributions with parameters

$$a_j' = a_j + \frac{rank(K_j)}{2} \quad \text{and} \quad b_j' = b_j + \frac{1}{2}\beta_j' K_j \beta_j. \tag{13}$$

We finally summarize the second proposed IWLS updating scheme based on the current mode approximations m_j^c and m_γ^c. The first IWLS updating scheme based on the current β_j^c and γ_j^c is very similar and therefore omitted. In what follows, the order of evaluations is stressed because it is crucial for computational efficiency.

Sampling scheme 1 (IWLS proposals based on current mode):

For implementing the sampling scheme described below the quantities β_j^c, β_j^p, $(\tau_j^2)^c$, γ^c, m_j^c, X_j, m_j, P_j and η^c must be created and enough memory must be allocated to store them.

(1) *Initialization:*
Compute the posterior modes m_j^c and γ^c for $\beta_1, \ldots, \beta_p$ and γ given fixed variance parameters $\tau_j^2 = c_j$, (e.g. $c_j = 10$). The mode is computed via backfitting within Fisher scoring. Use the posterior mode estimates as the current state β_j^c, γ^c of the chain. Set $(\tau_j^2)^c = c_j$. Store the current predictor in the vector η^c.

(2) *For $j = 1, \ldots, p$ update β_j:*

- Compute the likelihood $L(y, \ldots, \beta_j^c, \ldots, \gamma^c)$.
- Exchange $X_j\beta_j^c$ in the current predictor η^c by $X_j m_j^c$, i.e. $\eta^c = \eta^c + X_j(m_j^c - \beta_j^c)$.
- Draw a proposal β_j^p from the Gaussian proposal density $q(\beta_j^c, \beta_j^p)$ with mean m_j and precision matrix P_j given in (11).
- Exchange $X_j m_j^c$ in the current predictor η^c by $X_j\beta_j^p$, i.e. $\eta^c = \eta^c + X_j(\beta_j^p - m_j^c)$.
- Compute the likelihood $L(y, \ldots, \beta_j^p, \ldots, \gamma^c)$.

- Compute $q(\beta_j^p, \beta_j^c)$, $q(\beta_j^c, \beta_j^p)$, $p(\beta_j^c \mid (\tau_j^2)^c)$ and $p(\beta_j^p \mid (\tau_j^2)^c)$.
- Accept β_j^p as the new state of the chain β_j^c with probability (12). If the proposal is rejected exchange $X_j\beta_j^p$ in the current predictor η^c by $X_j\beta_j^c$, i.e. $\eta^c = \eta^c + X_j(\beta_j^c - \beta_j^p)$.
- Set $m_j^c = m_j$.

(3) *Update fixed effects parameters:*
Update fixed effects parameters by similar steps as for updating of β_j.

(4) *For $j = 1, \ldots, p$ update variance parameters τ_j^2:*
Variance parameters are updated by drawing from the inverse gamma full conditionals with hyperparameters given in (13). Obtain $(\tau_j^2)^c$.

Usually convergence and mixing of Markov chains is excellent with both variants of IWLS proposals. If, however, the effect of two covariates $x_j^{(1)}$ and $x_j^{(2)}$ is decomposed into main effects and a two dimensional interaction effect as in (8), severe convergence problems for the variance parameter of the interaction effect are the rule. To overcome the difficulties, we follow Knorr-Held and Rue (2002) who propose to construct a *joint proposal* for the parameter vector β_j and the corresponding variance parameter τ_j^2, and to simultaneously accept/reject (β_j, τ_j^2). We illustrate the updating scheme with IWLS proposals based on the current state of the chain β_j^c. We first sample $(\tau_j^2)^p$ from a proposal distribution for τ_j^2, and subsequently draw from the IWLS proposal for the corresponding regression parameters given the proposed $(\tau_j^2)^p$. The proposal distribution for τ_j^2 may depend on the current state $(\tau_j^2)^c$ of the variance, but *must be independent* of β_j^c. As suggested by Knorr-Held and Rue (2002), we construct the proposal by multiplying the current state $(\tau_j^2)^c$ by a random variable z with density proportional to $1 + 1/z$ on the interval $[1/f, f]$, where $f > 1$ is a tuning constant. The density is independent of the regression parameters and the joint proposal for (β_j, τ_j^2) is the product of the two proposal densities. We tune f in the burn in period to obtain acceptance probabilities of 30-60%. The acceptance probability is given by

$$\alpha(\beta_j^c, (\tau_j^2)^c, \beta_j^p, (\tau_j^2)^p) = \frac{L(y, \ldots, \beta_j^p, (\tau_j^2)^p, \ldots, \gamma^c)}{L(y, \ldots, \beta_j^c, (\tau_j^2)^c, \ldots, \gamma^c)} \frac{p(\beta_j^p \mid (\tau_j^2)^p) p((\tau_j^2)^p)}{p(\beta_j^c \mid (\tau_j^2)^c) p((\tau_j^2)^c)} \frac{q(\beta_j^p, \beta_j^c)}{q(\beta_j^c, \beta_j^p)}. \tag{14}$$

Note that the proposal ratio of the variance parameter cancels out. Summarizing, we obtain the following sampling scheme:

Sampling scheme 2: IWLS proposals, update β_j's and τ_j^2's in one block:

(1) *Initialization:*
Compute the posterior mode for $\beta_1, \ldots, \beta_p$ and γ given fixed variance

parameters $\tau_j^2 = c_j$, (e.g. $c_j = 10$). Use the posterior mode estimates as the current state β_j^c, γ^c of the chain. Set $(\tau_j^2)^c) = c_j$. Store the current predictor in the vector η^c.

(2) *For $j = 1, \dots, p$ update β_j, τ_j^2:*

- Compute the likelihood $L(y, \dots, \beta_j^c, (\tau_j^2)^c, \dots, \gamma^c)$.
- *Propose new τ_j^2:*
  Sample a random number z with density proportional to $1 + 1/z$ on the interval $[1/f, f]$, $f > 1$. Set $(\tau_j^2)^p = z \cdot (\tau_j^2)^c$ as the proposed new value for the jth variance parameter.
- Draw a proposal β_j^p from $q(\beta_j^c, \beta_j^p)$ with mean $m_j((\tau_j^2)^p)$ and precision matrix $P_j((\tau_j^2)^p)$ defined in (11).
- Exchange $X_j\beta_j^c$ in the current predictor η^c by $X_j\beta_j^p$, i.e. $\eta^c = \eta^c + X_j(\beta_j^p - \beta_j^c)$.
- Compute the likelihood $L(y, \dots, \beta_j^p, (\tau_j^2)^p, \dots, \gamma^c)$.
- Compute $q(\beta_j^c, \beta_j^p)$, $p(\beta_j^c \,|\, (\tau_j^2)^c)$, $p(\beta_j^p \,|\, (\tau_j^2)^p)$, $p((\tau_j^2)^c)$ and $p((\tau_j^2)^p)$.
- Based on the current predictor η^c compute again P_j and m_j defined in (11) and use these quantities to compute $q(\beta_j^p, \beta_j^c)$.
- Accept $\beta_j^p, (\tau_j^2)^p$ with probability (14). If the proposals are rejected exchange $X_j\beta_j^p$ in the current predictor η^c by $X_j\beta_j^c$, i.e. $\eta^c = \eta^c + X_j(\beta_j^c - \beta_j^p)$.

(3) *Update fixed effects parameters*

We conclude this section with three remarks:

- *Suppressing the computation of weights:*
  A natural source for saving computing time is to avoid the (re)computation of the IWLS weight matrix W in every iteration of the sampler. As a consequence the computation of a number of quantities required for updating the regression coefficients β_j can be omitted. Besides the weight matrix W these are the quantities $X_j'WX_j$, $X_j'W$ in (11) and the ratio of the normalizing constant of the proposal in sampling scheme 1. Moreover the posterior precision must be computed only once per iteration (sampling scheme 1 only). A possible strategy is to recompute the weights only every t-th iteration. It is even possible to keep the weights fixed after the burn in period. Our experience suggests that for most distributions the acceptance rates and the mixing of the chains is almost unaffected by keeping the weights fixed.
- *Multinomial logit models:*
  Our sampling schemes for univariate response distributions can be readily extended to the widely used multinomial logit model. Suppose that the response is multicategorical with k categories, i.e. $y_i = (y_{i1}, \dots, y_{ik})'$ where $y_{ir} = 1$ if the r-th category has been observed and zero otherwise. The multinomial logit model assumes that given covariates and parameters the responses y_i are multinomial distributed, i.e. $y_i \,|\, x_i, v_i \sim MN(1, \pi_i)$ where

$\pi_i = (\pi_{i1}, \dots, \pi_{ik})'$. The covariates enter the model by assuming

$$\pi_{ir} = \frac{\exp(\eta_{ir})}{1 + \sum_{l=1}^{k-1} \exp(\eta_{il})}, \qquad r = 1, \dots, k-1,$$

where

$$\eta_{ir} = f_{1r}(x_{i1r}) + \cdots + f_{pr}(x_{ipr}) + v_{ir}'\gamma_r, \qquad r = 1, \dots, k-1,$$

are structured additive predictors of the covariates (as described in Section 2.3). Note that our formulation allows category specific covariates. For identifiability reasons one category must be chosen as the reference category, without loss of generality we use the last category, i.e. $\pi_{ik} = 1 - \sum_{r=1}^{k-1} \pi_{ir}$. Abe (1999) (see also Hastie and Tibshirani (1990), Ch. 8.1) describes IWLS in combination with backfitting to estimate a multinomial logit model with additive predictors. Here, the transformed responses

$$\tilde{y}_{ir} = \eta_{ir} + \frac{1}{\pi_{ir}(1 - \pi_{ir})}(y_{ir} - \pi_{ir})$$

and weights

$$w_{ir} = \pi_{ir}(1 - \pi_{ir})$$

are used for subsequent backfitting to obtain estimates of the unknown functions. We can use the same transformed responses and weights for our IWLS proposals and the sampling algorithms described above readily extend to multicategorical logit models. E.g. sampling scheme 2 successively updates parameters in the order $(\beta_{11}, \tau_{11}^2)$, $(\beta_{12}, \tau_{12}^2)$,. . ., $(\beta_{1p}, \tau_{1p}^2)$,$\gamma_1$,. . .,$(\beta_{k-1,1}$,$\tau_{k-1,1}^2)$, . . ., $(\beta_{k-1,p}$, $\tau_{k-1,p}^2)$, γ_{k-1}, where β_{jr}, τ_{jr}^2 correspond to the regression parameters and variance parameter of the j-th nonlinear function f_{jr} of category r. Again, computing time may be saved by avoiding the computation of the weights in every iteration of the sampler.

- Note that sampling schemes 2 is not always preferable to sampling scheme 1. The advantage of sampling scheme 1 is that it is easier to implement, requires no tuning parameter and yields higher acceptance rates. The latter is often important for updating spatial effects based on Markov random fields. If there are a large number of regions, sampling scheme 2 might be inappropriate because of too small acceptance rates.

3.2 Inference based on latent utility representations of categorical regression models

For models with categorical responses alternative sampling schemes based on latent utility representations can be developed. The seminal paper by Albert and Chib (1993) develops algorithms for probit models with ordered categorical responses. The case of probit models with unordered multicategorical responses is dealt with e.g. in Chen and Dey (2000) or Fahrmeir and Lang (2001b). Recently, another important data augmentation approach for binary and multinomial logit models has been presented by Holmes and Held (2003). The adaption of these sampling schemes to the models discussed in this paper is more or less straightforward. We briefly illustrate the concept for binary data, i.e. y_i takes only the values 0 or 1. We first assume a probit model. Conditional on the covariates and the parameters, y_i follows a Bernoulli distribution $y_i \sim B(1, \mu_i)$ with conditional mean $\mu_i = \Phi(\eta_i)$ where Φ is the cumulative distribution function of a standard normal distribution. Introducing latent variables

$$U_i = \eta_i + \epsilon_i, \tag{15}$$

with $\epsilon_i \sim N(0, 1)$, we define $y_i = 1$ if $U_i \geq 0$ and $y_i = 0$ if $U_i < 0$. It is easy to show that this corresponds to a binary probit model for the y_i's. The posterior of the model augmented by the latent variables depends now on the extra parameters U_i. Correspondingly, additional sampling steps for updating the U_i's are required. Fortunately, sampling the U_i's is relatively easy and fast because the full conditionals are truncated normal distributions. More specifically, $U_i \,|\, \cdot \sim N(\eta_i, 1)$ truncated at the left by 0 if $y_i = 1$ and truncated at the right if $y_i = 0$. Efficient algorithms for drawing random numbers from a truncated normal distribution can be found in Geweke (1991) or Robert (1995). The advantage of defining a probit model through the latent variables U_i is that the full conditionals for the regression parameters β_j (and γ) are Gaussian with precision matrix and mean given by

$$P_j = X_j' X_j + \frac{1}{\tau_j^2} K_j, \quad m_j = P_j^{-1} X_j' (U - \tilde{\eta}).$$

Hence, the efficient and faster sampling schemes developed for Gaussian responses can be used with slight modifications. Updating of β_j and γ can be done exactly as described in Lang and Brezger (2004) using the current values U_i^c of the latent utilities as (pseudo) responses.

For binary logit models, the sampling schemes become more complicated and less efficient (regarding computing time). A logit model can be expressed in terms of latent utilities by assuming $\epsilon_i \sim N(0, \lambda_i)$ in (15) with $\lambda_i = 4\psi_i^2$, where ψ_i follows a Kolmogorov-Smirnov distribution (Devroye, 1986). Hence, ϵ_i is a scale mixture of normal form with a marginal logistic distribution (Andrews and Mallows, 1974). The main difference to the probit case is that additional parameters λ_i must be sampled. Holmes and Held (2003) propose to joint update U_i, λ_i by first drawing from the marginal distribution $p(U_i \,|\, \beta_1, \ldots, \beta_p, \gamma, y_i)$ of the $U_i's$ followed by drawing from $p(\lambda_i \,|\, U_i, \beta_1, \ldots, \beta_p, \gamma)$.

The marginal densities of the $U_i' s$ are truncated logistic distributions while $p(\lambda_i \,|\, U_i, \beta_1, \ldots, \beta_p)$ is not of standard form. Detailed algorithms for sampling from both distributions can be found in Holmes and Held (2003), appendix A3 and A4. Similar to probit models the full conditionals for the regression parameters β_j are Gaussian with precision matrix and mean given by

$$P_j = X_j' \Lambda^{-1} X_j + \frac{1}{\tau_j^2} K_j, \quad m_j = P_j^{-1} X_j' \Lambda^{-1} (U - \tilde{\eta}) \tag{16}$$

with weight matrix $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$. This updating scheme is considerably slower than the scheme for probit models. It is also much slower than the IWLS schemes discussed in Section 3.1. The reason is that drawing random numbers from $p(\lambda_i \,|\, U_i, \beta_1, \ldots, \beta_p, \gamma)$ is based on rejection sampling and therefore time consuming. Moreover, the matrix products $X_j' \Lambda^{-1} X_j$ in (16) must be recomputed in every iteration of the sampler. The advantage of the updating scheme is, however, that the acceptance rates will always be unity regardless of the number of parameters. This may be a particular advantage when estimating high dimensional Markov random fields.

3.3 Future prediction with Bayesian P-splines

In our second application on health insurance data it is necessary to get estimates of a function f_j outside the range of x_j. More specifically, we are interested in a one year ahead prediction of a time trend. Future prediction with Bayesian P-splines is obtained in a similar way as described in Besag, Green, Higdon and Mengersen (1995) for simple random walks. The spline can be defined outside the range of x_j by defining additional equidistant knots and by computing the corresponding B-spline basis functions. Samples of the additional regression parameters $\beta_{j,M_j+1}, \beta_{j,M_j+2}, \ldots$ are obtained by continuing the random walks in (4). E.g. for a second order random walk samples $\beta_{j,M_j+1}^{(t)}$, $t = 1, 2, 3, \ldots$ are obtained through $\beta_{j,M_j+1}^{(t)} \sim N(2\beta_{j,M_j}^{(t)} - \beta_{j,M_j-1}^{(t)}, (\tau_j^2)^{(t)})$, i.e. the samples of β_{j,M_j}, β_{j,M_j-1} and τ_j^2 are inserted into (4). Samples of additional parameters $\beta_{j,M_j+2}, \ldots$ are computed accordingly.

3.4 Model selection

The models proposed in this paper are quite general and the model building process can be quite challenging. Currently, an automated procedure for model selection is not available. However, a few recommendations are possible:

- Keep the basic model as simple as possible, start e.g. with a main effects

model that contains no interaction effect. Our experience with many *BayesX* users is that they try to incorporate everything that is theoretically possible.
- Simulation studies (Lang and Fahrmeir, 2001) suggest that both an unstructured and a structured effect should be incorporated when modeling a spatial effect. If one of both effects is very small it can be omitted in a final run.
- Different models could be compared via the DIC (Spiegelhalter, Best, Carlin and van der Linde, 2002).

4 Applications

4.1 Longitudinal study on forest health

In this longitudinal study on the health status of trees, we analyze the influence of calendar time t, age of trees A (in years) at the beginning of the observation period, canopy density CP (in percent) and location L of the stand on the defoliation degree of beeches. Data have been collected in yearly forest damage inventories carried out in the forest district of Rothenbuch in northern Bavaria from 1983 to 2001. There are 80 observation points with occurrence of beeches spread over an area extending about 15 km from east to west and 10 km from north to south. The degree of defoliation is used as an indicator for the state of a tree. It is measured in three ordered categories, with $y_{it} = 1$ for "bad" state of tree i in year t, $y_{it} = 2$ for "medium" and $y_{it} = 3$ for "good". A detailed data description can be found in Göttlein and Pruscha (1996).

We use a three-categorical ordered probit model based on a latent semiparametric model $U_{it} = \eta_{it} + \epsilon_{it}$ with predictor

$$\eta_{it} = f_1(t) + f_2(A_i) + f_{1|2}(t, A_i) + f_3(CP_{it}) + f_{str}(L_i). \tag{17}$$

In a frequentist setting, the model is an example for a mixed or random effects model with a spatially correlated tree or location specific random effect f_{str}.

The calendar time trend $f_1(t)$ and the age effect $f_2(A)$ are modeled by cubic P-splines with a second order random walk penalty. The interaction effect between calendar time and age $f_{1|2}(t, A)$ is modeled by a two dimensional cubic P-splines on a 12 by 12 grid of knots. Since canopy density is measured only in 11 different values (0%, 10%,. . . ,100%) we use a simple second order random walk prior (i.e. a P-spline of degree 0) for $f_3(CP)$. For the spatial effect $f_{str}(L)$ we experimented with both a two dimensional P-spline (model 1) and a Markov random field prior (model 2). Following Fahrmeir and Lang (2001b), the neighborhood ∂_s of trees for the Markov random field includes all trees u

with Euclidian distance $d(s, u) \leq 1.2$ km. In terms of the DIC, the model based on the Markov random field is preferable. An unstructured spatial effect f_{unstr} is excluded from the predictor for the following two reasons. First, a look at the map of observation points (see Figure 3) reveals some sites with only one neighbor, making the identification of a structured and an unstructured effect difficult if not impossible. The second reason is that for each of the 80 sites only 19 observations on the same tree are available with only minor changes of the response category. In fact, there are only a couple of sites where all three response categories have been observed.

The data have been already analyzed in Fahrmeir and Lang (2001b) (for the years 1983-1997 only). Here, nonlinear functions have been modeled solely by random walk priors. Also, the modeling of the interaction between calendar time and age is less sophisticated.

Since the results for model 1 and 2 differ only for the spatial effect, we present for the remaining covariates only estimates based on model 2. All results are based on the choice $a_j = b_j = 0.001$ for the hyperparameters of the variances. A sensitivity analysis revealed that the results are robust to other choices of a_j and b_j. Figure 1 shows the nonlinear main effects of calendar time and age of the tree as well as the effect of canopy density. The interaction effect between calendar time and age is depicted in Figure 2. The spatial effect is shown in Figure 3. Results based on a two dimensional P-spline can be found in the left panel, and for a Markov random field in the right panel. Shown are posterior probabilities based on a nominal level of 95%.

As we might have expected younger trees are in healthier state than the older ones. We also see that trees recover after the bad years around 1986, but after 1994 health status declines to a lower level again. The interaction effect between time and age is remarkably strong. In the beginning of the observation period young trees are affected higher than the average from bad environmental conditions. Thereafter, however, they recover better than the average. The distinct monotonic increase of the effect of canopy densities $\geq 30\%$ gives evidence that beeches get more shelter from bad environmental influences in stands with high canopy density. The spatial effect based on the two dimensional P-spline and the Markov random field are very similar. The Markov random field is slightly rougher (as could have been expected). Note that the spatial effect is quite strong and therefore not negligible.

4.2 Space-time analysis of health insurance data

In this section we analyze space-time data from a German private health insurance company. In a consulting case the main interest was on analyzing

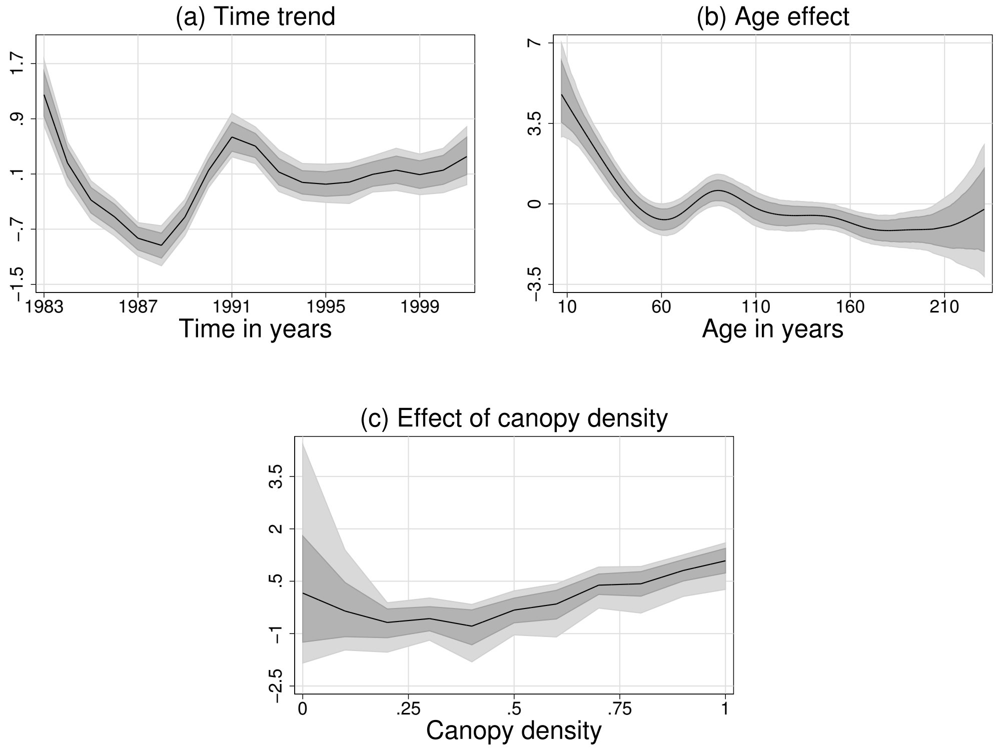

Fig. 1. Forest health data. Nonlinear main effects of calendar time, age of the tree and canopy density. Shown are the posterior means together with 95% and 80% pointwise credible intervals.

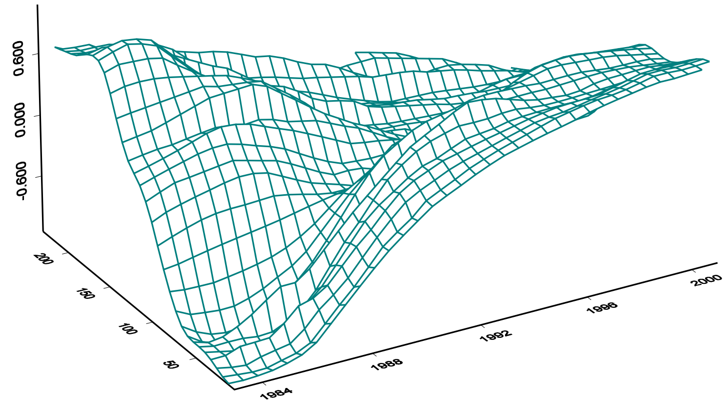

Fig. 2. Forest health data. Nonlinear interaction between calendar time and age of the tree. Shown are the posterior means.

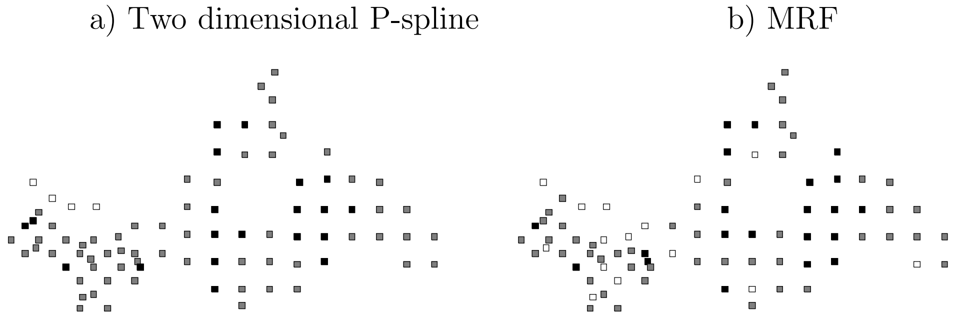

Fig. 3. Forest health data. Panel a) shows the spatial effect based on two dimensional P-splines. Panels b) displays the spatial effect based on Markov random fields. Shown are posterior probabilities for a nominal level of 95%. Black denotes locations with strictly negative credible intervals, white denotes locations with strictly positive credible intervals.

the dependence of treatment costs on covariates with a special emphasis on modeling the spatio-temporal development. The data set contains individual observations for a sample of 13.000 males (with about 160.000 observations) and 1.200 females (with about 130.000 observations) in West Germany for the years 1991-1997. The variable of primary interest is the treatment cost C in hospitals. Except some categorical covariates characterizing the insured person we analyzed the influence of the continuous covariates age (A) and calendar time (t) as well as the influence of the district (D) where the policy holder lives. We carried out separate analysis for men and women. We also distinguish between 3 types of health services, "accommodation", "treatment with operation" and "treatment without operation". In this demonstrating example, we present only results for males and "treatment with operation". Since the treatment costs are nonnegative and considerably skewed we assume that the costs for individual i at time t given covariates x_{it} are gamma distributed, i.e. $C_{it} \,|\, x_{it} \sim Ga(\mu_{it}, \phi)$ where ϕ is a scale parameter and the mean μ_{it} is defined as

$$\mu_{it} = \exp(\eta_{it}) = \exp(\gamma_0 + f_1(t) + f_2(A_{it}) + f_3(D_{it})).$$

For the effects of age and calendar time we assumed cubic P-splines with 20 knots and a second order random walk penalty. To distinguish between spatially smooth and small scale regional effects, we further split up the spatial effect f_3 into a spatially structured and a unstructured effect, i.e.

$$f_3(D_{it}) = f_{str}(D_{it}) + f_{unstr}(D_{it}).$$

For the unstructured effect f_{unstr} we assume i.i.d. Gaussian random effects. For the spatially structured effect we tested both a Markov random field prior and a two dimensional P-spline on a 20 by 20 knots grid. In terms of the DIC the model based on the MRF prior is preferable. Therefore results based on the two dimensional P-spline are omitted. Both sampling schemes 1 and 2 may be used for posterior inference in this situation.

The estimation of the scale parameter ϕ deserves special attention because MCMC inference is not trivial. In analogy to the variance parameter in Gaussian response models, we assume an inverse gamma prior with hyperparameters a_ϕ and b_ϕ for ϕ, i.e. $\phi \sim IG(a_\phi, b_\phi)$. Using this prior the full conditional for ϕ is given by

$$p(\phi \,|\, \cdot) \propto \left(\frac{1}{\Gamma(\phi)\phi^{\phi}}\right)^{n} \phi^{a_\phi - 1} \exp(-\phi b'_\phi)$$

with

$$b'_\phi = b_\phi + \sum_{i,t} (\log(\mu_{it}) - \log(C_{it}) + C_{it}/\mu_{it}).$$

This distribution is not of standard form. Hence, the scale parameter must be updated by Metropolis-Hastings steps. We update ϕ by drawing a random number ϕ^p from an inverse gamma proposal distribution with a variance s^2 and a mean equal to the current state of the chain ϕ^c. The variance s^2 is a tuning parameter and must be chosen appropriately to guarantee good mixing properties. We choose s^2 such that the acceptance rates are roughly between 30 and 60 percent.

It turns out that the results are unsensitive to the choice of hyperparameters a_j and b_j. The presentation of results is therefore restricted to the standard choice $a_j = b_j = 0.001$ for the hyperparameters of the variances.

Figure 4 shows the time trend f_1 (panel a) and the age effect f_2 (panel b). Shown are the posterior means together with 80% and 95% pointwise credible intervals. The effect for the year 1998 is future prediction explaining the growing uncertainty for the time effect in this year. Note also the large credible intervals of the age effect for individuals of age 90 and above. The reason is small sample sizes for these age groups. To gain more insight into the size of the effects, panels c) and d) display the marginal effects $f_j^{marginal}$ which are defined as $f_j^{marginal}(x_j) = \exp(\gamma_0 + f_j(x_j))$, i.e. the mean of treatment costs with the values of the remaining covariates fixed such that their effect is zero. The marginal effects (including credible intervals) can be easily estimated in a MCMC sampling scheme by computing (and storing) $f_j^{marginal}(x_j)$ in every iteration of the sampler from the current value of $f_j(x_j)$ and the intercept γ_0. Posterior inference is then based on the samples of $f_j^{marginal}(x_j)$. For the ease of interpretation, a horizontal line is included in the graphs indicating the marginal effect for $f_j = 0$, i.e. $\exp(\gamma_0) \approx 940DM$. Finally, panels e) and f) show the first derivatives of both effects (again including credible intervals). They may be computed by the usual formulas for derivatives of polynomial splines, see De Boor (1978).

Figure 5 displays the structured spatial effect f_{str} based on a Markov random field prior. The posterior mean of f_{str} can be found in panel a), the marginal effect is depicted in panel b). Panels c) and d) show posterior probabilities based on nominal levels of 80% and 95%. Note the large size of the spatial effect with a marginal effect ranging from 730-1200 German marks. It is clear that it is of great interest for health insurance companies to detect regions with large deviations of treatment costs compared to the average. The unstructured spatial effect f_{unstr} is negligible compared to the structured effect and therefore omitted.

Finally, note that results based on a two dimensional P-spline for the structured spatial effect are similar. The main difference is, that the estimated spatial effect is smoother.

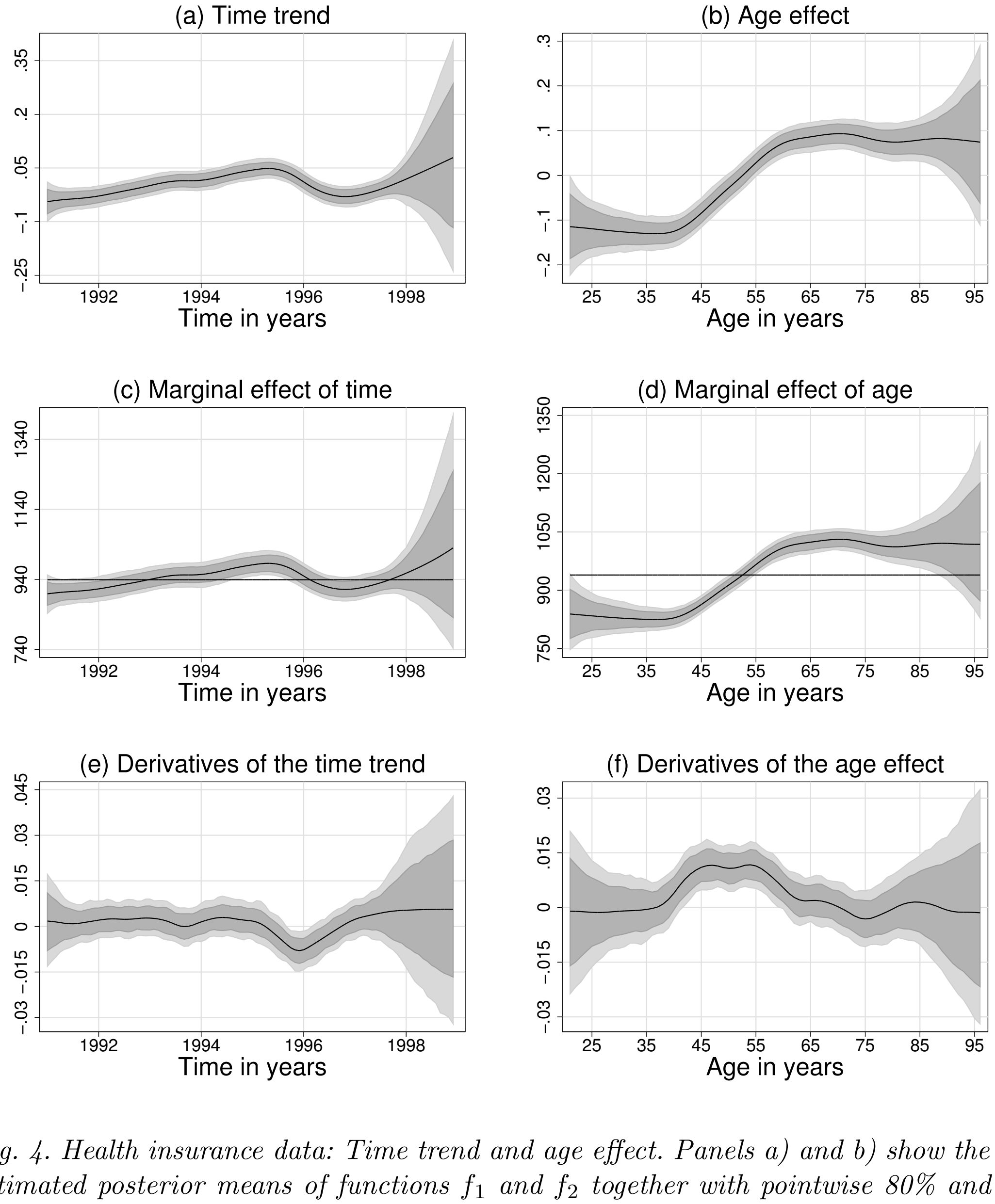

Fig. 4. Health insurance data: Time trend and age effect. Panels a) and b) show the estimated posterior means of functions f_1 and f_2 together with pointwise 80% and 95% pointwise credible intervals. Panels c) and d) depict the respective marginal effects and panels e) and f) the first derivatives f_1' and f_2'.

5 Conclusions

This paper proposes semiparametric Bayesian inference for regression models with responses from an exponential family and with structured additive predictors. Our approach allows estimation of nonlinear effects of continuous covariates and time scales as well as the appropriate consideration of unobserved unit- or cluster specific and spatial heterogeneity. Many well known regression models from the literature appear to be special cases of our approach, e.g. dynamic models, generalized additive mixed models, varying coefficient models, geoadditive models or the famous and widely used BYM-model for disease mapping (Besag, York and Mollie (1991)). The proposed sampling schemes work well and automatically for the most common response distributions. Software is provided in the public domain package *BayesX*.

Our current research is mainly focused on model choice and variable selection. Presently, model choice is based primarily on pointwise credible intervals for regression parameters and the DIC. A first step for more sophisticated variable selection is to replace pointwise credible intervals by simultaneous probability statements as proposed by Besag, Green, Higdon and Mengersen (1995) and more recently by Knorr-Held (2003). For the future, we plan to develop Bayesian inference techniques that allow estimation and model choice (to some extent) simultaneously.

References

Abe, M., 1999: A generalized additive model for discrete-choice data. *Journal of Business & Economic Statistics*, 17, 271-284.

Albert, J. and Chib, S., 1993: Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669-679.

Andrews, D.F. and Mallows, C.L., 1974: Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B*, 36, 99-102.

Besag, J. E. Green, P. J. Higdon, D. and Mengersen, K., 1995: Bayesian computation and stochastic systems (with discussion). *Statistical Science*, 10, 3–66.

Besag, J. and Kooperberg, C., 1995: On conditional and intrinsic autoregressions. *Biometrika*, 82, 733-746.

Besag, J., York, J. and Mollie, A., 1991: Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Biller, C., 2000: Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics*, 9, 122-140.

Biller, C. and Fahrmeir, L., 2001: Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling*, 2, 195-211.

De Boor, C., 1978: *A practical guide to splines.* Spriner-Verlag, New York.

Chen, M. H. and Dey, D. K., 2000: Bayesian analysis for correlated ordinal data models. In: Dey, D. K., Ghosh, S. K. and Mallick, B. K., 2000: *Generalized linear models: A Bayesian perspective.* Marcel Dekker, New York.

Cleveland, W. and Grosse, E., 1991: Computational methods for local regression. *Statistics and Computing*, 1991, 1, 47-62.

Currie, I. and Durban, M., 2002: Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, 4, 333-349.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M., 1998: Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society B*, 60, 333-350.

Devroye, L., 1986: *Non-uniform random variate generation.* Springer-Verlag, New York.

Diggle, P.J., Haegerty, P., Liang, K.Y., and Zeger, S.L., 2002: *Analysis of longitudinal data*, Clarendon Press, Oxford.

Di Matteo, I., Genovese, C.R. and Kass, R.E., 2001: Bayesian curve-fitting with free-knot splines, *Biometrika*, 2001, 88, 1055–1071.

Eilers, P.H.C. and Marx, B.D., 1996: Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statistical Science*, 11, 89-121.

Fahrmeir, L., Kneib, Th. and Lang, S., 2004: Penalized additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, 14, 715-745.

Fahrmeir, L. and Lang, S., 2001a: Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society C*, 50, 201-220.

Fahrmeir, L. and Lang, S., 2001b: Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the Institute of Statistical Mathematics*, 53, 10-30

Fahrmeir, L., Lang, S., Wolff, J. and Bender, S. (2003): Semiparametric Bayesian time-space analysis of unemployment duration. *Journal of the German Statistical Society*, 87, 281-307.

Fahrmeir, L. and Tutz, G., 2001: *Multivariate statistical modelling based on generalized linear models*, Springer–Verlag, New York.

Fan, J. and Gijbels, I., 1996: *Local polynomial modelling and its applications.* Chapman and Hall, London.

Fotheringham, A.S., Brunsdon, C., and Charlton, M.E., 2002: *Geographically weighted regression: The analysis of spatially varying relationships.* Chichester: Wiley.

Friedman, J. H., 1991: Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19, 1-141.

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

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

Gamerman, D., Moreira, A.R.B., and Rue, H., 2003: Space-varying regression models: specifications and simulation. *Computational Statistics and Data Analysis*, 42, 513-533.

George, A. and Liu, J.W., 1981: *Computer solution of large sparse positive definite systems.* Prentice–Hall.

Geweke, J. 1991: Efficient simulation from the multivariate normal and Student-t distribution subject to linear constraints. In: *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface,* 571-578, Alexandria.

Göttlein, A. and Pruscha, H., 1996: Der Einfluß von Bestandskenngrößen, Topographie, Standord und Witterung auf die Entwicklung des Kronenzustandes im Bereich des Forstamtes Rothenbuch. *Forstwissenschaftliches Centralblatt*, 114, 146-162.

Green, P.J., 2001: A primer in Markov Chain Monte Carlo. In: Barndorff-Nielsen, O.E., Cox, D.R. and Klüppelberg, C. (eds.), *Complex stochastic systems.* Chapmann and Hall, London, 1-62.

Hansen, M. H., Kooperberg, C., 2002: Spline adaptation in extended linear models. *Statistical Science*, 17, 2-51.

Hastie, T. and Tibshirani, R., 1990: *Generalized additive models.* Chapman and Hall, London.

Hastie, T. and Tibshirani, R., 1993: Varying-coefficient Models. *Journal of the Royal Statistical Society B*, 55, 757-796.

Hastie, T. and Tibshirani, R., 2000: Bayesian Backfitting. *Statistical Science*, 15, 193-223.

Hennerfeind, A., Brezger, A., and Fahrmeir, L., 2003: Geoadditive survival models. SFB 386 Discussion paper 333, University of Munich.

Holmes, C.C., and Held, L., 2003: On the simulation of Bayesian binary and polychotomous regression models using auxiliary variables. Technical report. Available at: `http://www.stat.uni-muenchen.de/~leo`

Jerak, A. and Lang, S., 2003: Locally adaptive function estimation for binary regression models. Discussion Paper 310, SFB 386. Revised for *Biometrical Journal.*

Kamman, E. E. and Wand, M. P., 2003: Geoadditive models. *Journal of the Royal Statistical Society C*, 52, 1-18.

Knorr-Held, L., 1996: *Hierarchical modelling of discrete longitudinal data.* Shaker Verlag.

Knorr-Held, L., 1999: Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics*, 26, 129-144.

Knorr-Held, L., 2003: Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics*, 13, 20-35.

Knorr-Held, L. and Rue, H., 2002: On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29, 597-614.

Lang, S. and Brezger, A., 2004: Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13, 183-212.

Lang, S. and Fahrmeir, L., 2001: Bayesian generalized additive mixed models. A simulation study. Discussion Paper 230, SFB 386.

Lang, S., Fronk, E.-M. and Fahrmeir, L., 2002: Function estimation with locally adaptive dynamic models. *Computational Statistics*, 17, 479-500.

Lenk, P. and DeSarbo, W.S., 2000: Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65, 93-119.

Lin, X. and Zhang, D., 1999: Inference in generalized additive mixed models by using smoothing splines. *Journal auf the Royal Statistical Society B* , 61, 381-400.

Marx, B.D. and Eilers, P.H.C., 1998: Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, 28, 193-209.

Robert, C.P., 1995: Simulation of truncated normal variables. *Statistics and Computing*, 5, 121-125.

Rue, H., 2001: Fast sampling of Gaussian Markov random fields with applications. *Journal of the Royal Statistical Society B*, 63, 325-338.

Ruppert, D., Wand, M.P. and Carroll, R.J., 2003: *Semiparametric Regression.* Cambridge University Press.

Smith, M. and Kohn, R., 1996: Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75, 317-343.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A., 2002: Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 65, 583 - 639.

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

Wand, M.P., 2003: Smoothing and mixed models, *Computational Statistics*, 18, 223-249.

Wood, S.N., 2000: Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society B*, 62, 413-428.

Wood, S.N., 2003: Thin plate regression splines. *Journal of the Royal Statistical Society B*, 65, 95-114.

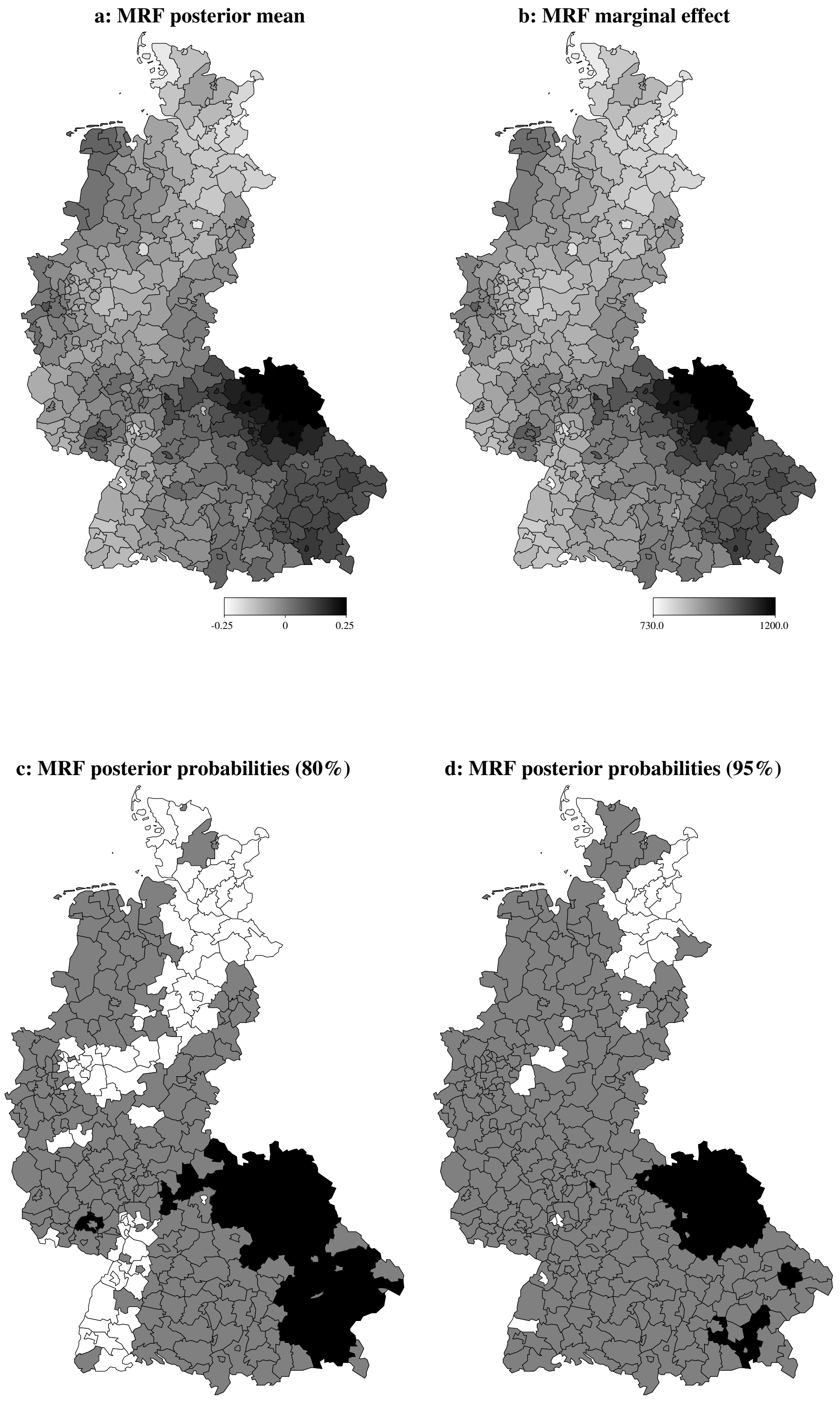
a: MRF posterior mean
b: MRF marginal effect
-0.25
0
0.25
730.0
1200.0
c: MRF posterior probabilities (80%)
d: MRF posterior probabilities (95%)

Fig. 5. Health insurance data: Structured spatial effect f_{str} based on Markov random field priors. The posterior mean of f_{str} is shown in panel a) and the marginal effect in panel b). Panels c) and d) display posterior probabilities for nominal levels of 80% and 95%. Black denotes regions with strictly positive credible intervals and white regions with strictly negative credible intervals.