Stat Comput
DOI 10.1007/s11222-012-9366-0

Multilevel structured additive regression

Stefan Lang · Nikolaus Umlauf · Peter Wechselberger · Kenneth Harttgen · Thomas Kneib

Received: 16 May 2012 / Accepted: 31 October 2012

Abstract Models with structured additive predictor provide a very broad and rich framework for complex regression modeling. They can deal simultaneously with nonlinear covariate effects and time trends, unit- or cluster-specific heterogeneity, spatial heterogeneity and complex interactions between covariates of different type. In this paper, we propose a hierarchical or multilevel version of regression models with structured additive predictor where the regression coefficients of a particular nonlinear term may obey another regression model with structured additive predictor. In that sense, the model is composed of a hierarchy of complex structured additive regression models. The proposed model may be regarded as an extended version of a multilevel model with nonlinear covariate terms in every level of the hierarchy. The model framework is also the basis for generalized random slope modeling based on multiplicative random effects. Inference is fully Bayesian and based on Markov chain Monte Carlo simulation techniques. We provide an in depth description of several highly efficient sampling schemes that allow to estimate complex models with several hierarchy levels and a large number of observations within a couple of minutes (often even seconds). We demonstrate the practicability of the approach in a complex application on childhood undernutrition with large sample size and three hierarchy levels.

Keywords Bayesian hierarchical models · Gaussian random fields · Markov random fields · MCMC · Multiplicative random effects · P-splines

1 Introduction

The last years have seen enormous progress in Bayesian semiparametric regression modeling based on Markov chain Monte Carlo (MCMC) simulation for inference. Pioneering work has been done by Smith and Kohn (1996) and Smith and Kohn (1997) who developed uni- and bivariate smoothers based on adaptive knot selection. Related more recent approaches can be found in Chan et al. (2006) and Cottet et al. (2008). This paper is in the tradition of another branch of the literature based on Bayesian roughness penalty approaches, see e.g. Fahrmeir and Lang (2001), and Lang and Brezger (2004) for early references, and more recently Jullion and Lambert (2007) and Panagiotelis and Smith (2008).

A particularly broad and rich framework is provided by generalized structured additive regression (STAR) models introduced in Fahrmeir et al. (2004) and Brezger and Lang (2006). Models of similar complexity have been developed in a mostly frequentist setting by Simon Wood (see e.g. Wood 2003, 2006) and in Ruppert et al. (2003), Rigby and Stasinopoulos (2005) or Rue et al. (2009). STAR models assume that, given covariates, the distribution of response observations y_i, $i = 1, \dots, n$, belongs to an exponential family. The conditional mean $\mu_i = E(y_i)$ is linked to a semiparametric additive predictor η_i by $\mu_i = h(\eta_i)$ where h is a known response function. The predictor η_i is of the form

$$\eta_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \boldsymbol{x}_i'\boldsymbol{\gamma}, \quad i = 1, \dots, n, \qquad (1)$$

S. Lang (✉) · N. Umlauf · P. Wechselberger
Department of Statistics, University of Innsbruck,
Universitätsstraße 15, 6020 Innsbruck, Austria
e-mail: stefan.lang@uibk.ac.at

K. Harttgen
NADEL, ETH Zürich, Voltastrasse 24, 8092 Zurich, Switzerland

T. Kneib
Faculty of Economic Sciences, University of Göttingen,
Platz der Göttinger Sieben 5, 37073 Göttingen, Germany

Published online: 22 December 2012

where $f_1, \ldots, f_q$ are nonlinear functions of the (possibly multidimensional) covariates $z_1, \ldots, z_q$ and $\boldsymbol{x}'\boldsymbol{\gamma}$ is the usual linear part of the model. The functions f_j comprise usual nonlinear effects of continuous covariates as well as time trends and seasonal effects, two-dimensional surfaces, varying coefficient terms and cluster- or spatial effects. The nonlinear functions in (1) are modeled by a basis functions approach, i.e. a particular nonlinear function f of covariate z is approximated by a linear combination of basis or indicator functions:

$$f(z) = \sum_{k=1}^{K} \beta_k B_k(z). \tag{2}$$

The B_k's are known basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ is a vector of unknown regression coefficients to be estimated. Specific examples for the choice of basis functions and priors for the regression coefficients will be given in Sect. 2. Defining the $n \times K$ design matrix $\boldsymbol{Z}$ with elements $\boldsymbol{Z}[i,k] = B_k(z_i)$, the vector $\boldsymbol{f} = (f(z_1), \ldots, f(z_n))'$ of function evaluations can be written in matrix notation as $\boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\beta}$. Accordingly, for the predictor (1) we obtain

$$\boldsymbol{\eta} = \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{Z}_q\boldsymbol{\beta}_q + \boldsymbol{X}\boldsymbol{\gamma}. \tag{3}$$

In this paper, we propose a hierarchical or multilevel version of regression models with structured additive predictor. Multilevel STAR models assume that the regression coefficients $\boldsymbol{\beta}_j$ of a term f_j in (3) may themselves obey a regression model with structured additive predictor, i.e.

$$\boldsymbol{\beta}_j = \boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j = \boldsymbol{Z}_{j1}\boldsymbol{\beta}_{j1} + \cdots + \boldsymbol{Z}_{jq_j}\boldsymbol{\beta}_{jq_j} + \boldsymbol{X}_j\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j. \tag{4}$$

Here the terms $\boldsymbol{Z}_{j1}\boldsymbol{\beta}_{j1}, \ldots, \boldsymbol{Z}_{jq_j}\boldsymbol{\beta}_{jq_j}$ correspond to additional nonlinear functions $f_{j1}, \ldots, f_{jq_j}$, $\boldsymbol{X}_j\boldsymbol{\gamma}_j$ comprises additional linear effects, and

$$\boldsymbol{\varepsilon}_j \sim N\big(\boldsymbol{0}, \tau_j^2 \boldsymbol{I}\big) \tag{5}$$

is a vector of i.i.d. Gaussian random effects. See the case study on childhood undernutrition below and particular Sect. 2.5 for specific examples of multilevel STAR models. To keep the paper reasonable in length, we restrict ourselves to i.i.d. Gaussian random effects although more sophisticated structures like the Bayesian LASSO (Park and Casella 2008), Dirichlet process mixtures (Heinzl et al. 2012) or spike and slab priors (Frühwirth-Schnatter and Wagner 2011) can be implemented in a straightforward way. Moreover, a third level or even higher levels in the hierarchy are possible by assuming that the second level regression parameters $\boldsymbol{\beta}_{jl}, l = 1, \ldots, q_j$, obey again a STAR model. In that sense, the model is composed of a hierarchy of complex structured additive regression models.

The two main goals of this paper are

- to provide a rich Bayesian framework for multilevel additive modeling including generalizations of random slopes,
- to discuss several highly efficient MCMC sampling schemes that utilize the hierarchical structure and allow to estimate complex models with several hierarchy levels and a large number of observations within a couple of minutes (often even seconds).

We provide an implementation of the methodology within the software package `BayesX` together with the full `R` interface `R2BayesX`.

A typical application of the proposed models are multilevel data where a hierarchy of units or clusters grouped at different levels is given. As an example, we will analyze survey data on child undernutrition in India. Undernutrition among children is usually measured in the form of a Z-score (variable *zscore*) that determines the anthropometric status of the child relative to a reference population of children known to have grown well. A child whose Z-score is below −2 is typically regarded as undernourished. In our analysis, we will distinguish three levels: Children (level-1) are nested in districts (level-2) and districts are nested in states (level-3). In Sect. 5, we will present results for a probit model that models the probability that a child is undernourished, i.e. *zscore* < −2. The following three level hierarchical predictor is used:

$$\begin{aligned}
\text{level-1:}\quad \boldsymbol{\eta} &= \boldsymbol{f}_1(c\_age) + \boldsymbol{f}_2(c\_age)c\_sex + \boldsymbol{f}_3(ageb)\\
&\quad + \boldsymbol{f}_4(ageb)c\_sex + \boldsymbol{f}_5(educy)\\
&\quad + \boldsymbol{f}_6(educy)c\_sex + \boldsymbol{f}_7(ai) + \boldsymbol{f}_8(ai)c\_sex\\
&\quad + \boldsymbol{f}_9(dist) + \boldsymbol{f}_{10}(dist)c\_sex + \cdots + \boldsymbol{\varepsilon}\\
&= \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{Z}_9\boldsymbol{\beta}_9 + \boldsymbol{Z}_{10}\boldsymbol{\beta}_{10} + \cdots + \boldsymbol{\varepsilon}\\
\text{level-2:}\quad \boldsymbol{\beta}_9 &= \boldsymbol{f}_{9,1}(m\_ai) + \boldsymbol{f}_{9,2}(m\_educy) + \boldsymbol{f}_{9,3}(dist)\\
&\quad + \boldsymbol{f}_{9,4}(state) + \boldsymbol{\varepsilon}_9\\
&= \boldsymbol{Z}_{9,1}\boldsymbol{\beta}_{9,1} + \cdots + \boldsymbol{Z}_{9,4}\boldsymbol{\beta}_{9,4} + \boldsymbol{\varepsilon}_9\\
\text{level-2:}\quad \boldsymbol{\beta}_{10} &= \boldsymbol{f}_{10,1}(m\_ai) + \boldsymbol{f}_{10,2}(m\_educy)\\
&\quad + \boldsymbol{f}_{10,3}(dist) + \boldsymbol{f}_{10,4}(state) + \boldsymbol{\varepsilon}_{10}\\
&= \boldsymbol{Z}_{10,1}\boldsymbol{\beta}_{10,1} + \cdots + \boldsymbol{Z}_{10,4}\boldsymbol{\beta}_{10,4} + \boldsymbol{\varepsilon}_{10}\\
\text{level-3:}\quad \boldsymbol{\beta}_{9,4} &= \boldsymbol{f}_{9,4,1}(gdp) + \boldsymbol{\varepsilon}_{9,4} = \boldsymbol{Z}_{9,4,1}\boldsymbol{\beta}_{9,4,1} + \boldsymbol{\varepsilon}_{9,4}\\
\text{level-3:}\quad \boldsymbol{\beta}_{10,4} &= \boldsymbol{f}_{10,4,1}(gdp) + \boldsymbol{\varepsilon}_{10,4}\\
&= \boldsymbol{Z}_{10,4,1}\boldsymbol{\beta}_{10,4,1} + \boldsymbol{\varepsilon}_{10,4}
\end{aligned} \tag{6}$$

The level-1 equation consists of possibly nonlinear smooth effects of the child's age (variable *c_age*), the mother's age at birth (*ageb*), the mother's educational attainment measured through the years of education (*educy*) and an asset index (*ai*) measuring the household's wealth. The asset index is derived using a principal components analysis based on the possession of household assets and dwelling characteristics. The latter two covariates are measured as differences from the district mean education level and wealth index. Since a main scientific question is on possible gender differences we include interaction terms between the covariates and gender (*c_sex*) given in effect coding and with males as the reference category. District-specific spatial heterogeneity is modeled through the two level-2 equations containing the average asset index per district (*m_ai*) and the average education years per district (*m_educy*). Spatial heterogeneity beyond the available district specific covariates is modeled through spatially correlated (discrete) effects $f_{9,3}(\mathit{dist})$, $f_{10,3}(\mathit{dist})$ and state-specific spatial effects $f_{9,4}(\mathit{state})$, $f_{10,4}(\mathit{state})$ modeled through the level-3 equations of the model. The spatially correlated effects $f_{9,3}(\mathit{dist})$ and $f_{10,3}(\mathit{dist})$ are analogous to a nonlinear smooth time trend in time series modeling. The level-3 effects $f_{9,4,1}(\mathit{gdp})$, $f_{10,4,1}(\mathit{gdp})$ are nonlinear effects of the gross domestic product per capita within states. The second level-2 equation in combination with the second level-3 equation models a complex nonlinear random "slope" effect of gender.

In principle the model (6) can be reexpressed in a reduced form as a usual STAR model as in (1). Then the predictor would contain the nonlinear covariate effects of all hierarchy levels as well as an additive composition of the i.i.d district and state specific random effects. However, the hierarchical formulation provides several distinct advantages compared to the reduced form:

- From an interpretational perspective, the hierarchical formulation provides an interesting decomposition of the random effects.
- Most importantly, Bayesian inference based on MCMC simulations is almost revolutionized through the hierarchical formulation as it allows for well-behaved (in terms of mixing) and very fast samplers that would be impossible in the reduced formulation.
- Finally, models going beyond the i.i.d. random effects (5) (which is our goal for future research) circumvent a simple reexpression of model (6) in reduced form.

Note that the hierarchical formulation is in the spirit of hierarchical centering as in Bayesian linear mixed models, see Papaspiliopoulos et al. (2007) (and the references therein) for a general framework.

Multilevel STAR models are also the basis for generalized random slopes or multiplicative random effects of the form

$$(1+\alpha_{c_i})f(z_i) = f(z_i) + \alpha_{c_i} f(z_i), \tag{7}$$

where the possibly nonlinear function f of a covariate z is scaled by a cluster specific factor $(1+\alpha_c)$ with respect to clusters $c \in \{1,\dots,C\}$. Treating such models in full details is beyond the scope of this paper. An application of generalized random slope modeling is given in a marketing paper that analyzes the impact of price changes on a brand's sales using the technology presented here, see Lang et al. (2012).

The rest of the paper is organized as follows: Sect. 2 discusses modeling of covariate effects and corresponding priors. Sections 3 and 4 are devoted to MCMC inference. Section 5 presents the results for the case study on undernutrition in India. The final Sect. 6 concludes and points out directions for future research.

2 Effect modeling and priors

Effect modeling and priors depend on the covariate or term type. We distinguish two types of priors: "direct" or "basic" priors for the regression coefficients $\boldsymbol{\beta}_j$ (or $\boldsymbol{\beta}_{jl}$ in a second level equation) and compound priors (4). We first describe the general form of "basic" priors. Sections 2.2–2.4 give specific examples for effect modeling using specific design matrices and forms of the basic prior. Section 2.5 shows how the basic priors can be used as building blocks for the compound priors.

2.1 General form of basic priors

In a frequentist setting, overfitting of a particular function $\boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\beta}$ is avoided by defining a roughness penalty on the regression coefficients, see for instance Wood (2006) in the context of structured additive regression. In a Bayesian framework a standard smoothness prior is a (possibly improper) Gaussian prior of the form

$$p(\boldsymbol{\beta} \mid \tau^2) \propto \left(\frac{1}{\tau^2}\right)^{\mathrm{rk}(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta}\right) \cdot I(\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}), \tag{8}$$

where $I(\cdot)$ is the indicator function. The key components of the prior are the penalty matrix $\boldsymbol{K}$, the variance parameter τ_j^2 and the constraint $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$.

The structure of the penalty or prior precision matrix $\boldsymbol{K}$ depends on the covariate type and on prior assumptions about smoothness of f, see Sects. 2.2–2.4 for specific examples. With one notable exception for Gaussian random fields, the penalty matrix in our examples is rank deficient, i.e. $\mathrm{rk}(\boldsymbol{K}) < K$, resulting in a partially improper prior.

The amount of smoothness is governed by the variance parameter τ^2. A conjugate inverse Gamma prior is employed for τ^2 (as well as for the error variance parameter

σ^2 in models with Gaussian responses), i.e. $\tau^2 \sim IG(a,b)$ with small values such as $a = b = 0.001$ for the hyperparameters a and b resulting in an uninformative prior on the log scale. Alternative priors for τ^2 have been discussed in Gelfand (2006).

The term $I(\boldsymbol{A\beta} = \mathbf{0})$ imposes required identifiability constraints on the parameter vector. A straightforward choice is $\boldsymbol{A} = (1,\dots,1)$, i.e. the regression coefficients are centered around zero. A better choice in terms of interpretability and mixing of the resulting Markov chains is to use a weighted average of regression coefficients, i.e. $\boldsymbol{A} = (a_{11},\dots,a_{1K})$. As a standard we use $a_{1k} = \sum_{i=1}^{n} B_k(z_i)$, $k = 1,\dots,K$, resulting in the more natural constraint $\sum_{i=1}^{n} f(z_i) = 0$. Additional constraints such as sum to zero constraints $\sum_{i=1}^{n} f''(z_i) = 0$ on the derivatives can be defined by adding a second row to $\boldsymbol{A}$ and by setting $a_{2k} = \sum_{i=1}^{n} B'_k(z_i)$.

2.2 Continuous covariate effects

For a continuous covariate z, our basic approach for modeling a smooth function f are P-splines introduced in a frequentist setting by Eilers and Marx (1996) and in a Bayesian version by Lang and Brezger (2004). P-splines assume that the unknown functions can be approximated by a polynomial spline which can be written in terms of a linear combination of B-spline basis functions. Hence, the columns of the design matrix $\boldsymbol{Z}$ are given by the B-spline basis functions evaluated at the observations z_i. Lang and Brezger (2004) propose to use first or second order random walks as smoothness priors for the regression coefficients, i.e.

$$\beta_k = \beta_{k-1} + u_k, \quad \text{or} \quad \beta_k = 2\beta_{k-1} - \beta_{k-2} + u_k, \tag{9}$$

with Gaussian errors $u_k \sim N(0,\tau^2)$ and diffuse priors $p(\beta_1) \propto \text{const}$, or $p(\beta_1)$ and $p(\beta_2) \propto \text{const}$, for initial values. This prior is of the form (8) with penalty matrix given by $\boldsymbol{K} = \boldsymbol{D}'\boldsymbol{D}$, where $\boldsymbol{D}$ is a first or second order difference matrix. Locally adaptive variants of the basic P-splines approach have been proposed e.g. in Yue et al. (2012). The Bayesian P-splines approach can be generalized to two-dimensional smoothing for modeling interactions by assuming that the unknown surface is the tensor product of one-dimensional B-splines, see Lang and Brezger (2004) for details.

2.3 Spatial effects

Assume now that $\boldsymbol{z}$ represents the location a particular observation pertains to. If exact locations are available, $\boldsymbol{z} = (z^{(1)}, z^{(2)})'$ is two-dimensional and the components $z^{(1)}$ and $z^{(2)}$ correspond to the coordinates of the location. In this case the spatial effect $f(z^{(1)}, z^{(2)})$ could be modeled by two-dimensional extensions of P-splines as described in Lang and Brezger (2004). An alternative approach widely used in the geostatistics literature (e.g. Kamman and Wand 2003) is to model the spatial effect by stationary Gaussian random fields. Here $f(\boldsymbol{z}) = f(z^{(1)}, z^{(2)}) = \beta_{\boldsymbol{z}}$ is assumed to follow a zero mean stationary Gaussian field with variance τ^2 and isotropic covariance function $\text{Cov}(\beta_{\boldsymbol{z}}, \beta_{\boldsymbol{z}}') = C(\|\boldsymbol{z} - \boldsymbol{z}'\|)$. For a finite number of design points, the prior is of the form (8) with penalty matrix $\boldsymbol{K} = \boldsymbol{C}$ where $\boldsymbol{C}[k,s] = C(\|z_k - z_s\|)$, $1 \le k, s \le n$. The design matrix is given by $\boldsymbol{Z} = \boldsymbol{C}$. A widespread choice for the covariance is the Matern family of covariance functions. One of the practical problems with Gaussian random fields is that the number of parameters is equal or close to the number of observations n. For that reason the random field is often approximated by defining a representative subset of knots of the set of distinct locations, see Kamman and Wand (2003) for details. The R function `cover.design` in the package `fields` provides a convenient tool for obtaining the reduced design. However, as pointed out by Hennerfeind et al. (2006), Bayesian inference based on MCMC simulations can be extremely slow because the penalty matrix as well as the design matrix cross product $\boldsymbol{Z}'\boldsymbol{Z}$ are full matrices, i.e. the typical sparse matrix structure can not be exploited for efficient computation. We will circumvent the problem by using a reparametrization of the regression coefficients such that the resulting penalty and cross product matrix are diagonal, see Sect. 4 for details.

Another alternative for modeling smooth spatial effects are Markov random fields (MRF) as described e.g. in Brezger and Lang (2006). MRF's are particularly useful if a geographical map is given and exact locations are not available.

2.4 Modeling interactions through varying coefficients

In our case study on stunting in India we are particulary interested in gender differences, which are modeled by interactions with the covariate *c_sex*. Interactions as in (6) are specific varying coefficient terms (Hastie and Tibshirani 1993). More generally, suppose that the effect of a covariate $z^{(2)}$ is assumed to vary with respect to another covariate $z^{(1)}$. The interaction between $z^{(2)}$ and $z^{(1)}$ can be modeled by a predictor of the form

$$\eta = \cdots + z^{(1)} g\big(z^{(2)}\big) + \cdots,$$

where g is a function of $z^{(2)}$ which in turn is the effect modifier of $z^{(1)}$. If the effect modifier is the location either given as the coordinates or as a spatial index we have a space varying effect of $z^{(1)}$ (for instance Gamerman et al. 2003).

Independent of the specific type of the effect modifier, the interaction term $z^{(1)} g(z^{(2)})$ can be cast into the general framework by defining

$$f\big(z^{(1)}, z^{(2)}\big) = z^{(1)} g\big(z^{(2)}\big). \tag{10}$$

The overall design matrix $\boldsymbol{Z}$ is given by $\operatorname{diag}(z_1^{(1)}, \dots, z_n^{(1)})\mathbf{Z}^{(2)}$ where $\mathbf{Z}^{(2)}$ is the usual design matrix for P-Splines, tensor product P-splines, spatial effects etc.

Varying coefficient terms are also the key for MCMC based inference in the generalized random slope terms (7). It can be shown that for fixed scaling parameters or fixed regression coefficients, the term (7) is technically identical to a varying coefficients term and MCMC updating is done by repeatedly obeying this varying coefficients structure. Details can be found in Lang et al. (2012).

2.5 Compound priors

In many cases the compound prior (4) is used if a covariate $z_j \in \{1, \dots, K\}$ is a unit- or cluster index and z_{ij} indicates the cluster observation i pertains to. Then the design matrix $\mathbf{Z}_j$ is a $n \times K$ incidence matrix with $\mathbf{Z}_j[i,k] = 1$ if the i-th observation belongs to cluster k and zero otherwise. The $K \times 1$ parameter vector $\boldsymbol{\beta}_j$ is the vector of regression parameters, i.e. the k-th element in $\boldsymbol{\beta}$ corresponds to the regression coefficient of the k-th cluster. Using the compound prior (4) we obtain an additive decomposition of the cluster-specific effect. The covariates z_{jl}, $l = 1, \dots, q_j$, in (4) are cluster-specific covariates with possible nonlinear cluster effect. By allowing a full STAR predictor (as in the level-1 equation) a rather complex decomposition of the cluster effect $\boldsymbol{\beta}_j$ including interactions is possible. A special case arises if cluster-specific covariates are not available. Then the prior for $\boldsymbol{\beta}_j$ collapses to $\boldsymbol{\beta}_j = \boldsymbol{\varepsilon}_j \sim N(0, \tau_j^2\mathbf{I})$ and we obtain a simple i.i.d. Gaussian cluster-specific random effect with variance parameter τ_j^2.

Another special situation arises if the data are grouped according to some discrete geographical grid and the cluster index z_{ij} denotes the geographical region observation i pertains to. For instance, in our application on child undernutrition in Sect. 5 for every observation the district of the households residence is given. Then the compound prior (4) models a complex spatial heterogeneity effect with possibly nonlinear effects of region-specific covariates z_{jl}.

In a number of applications, geographical information and spatial covariates are given at different resolutions. For instance, in our case study on child undernutrition, the districts (level-2) are nested within states (level-3). This allows to model a spatial effect over two levels in the form

$$\boldsymbol{\beta}_j = \mathbf{Z}_{j1}\boldsymbol{\beta}_{j1} + \mathbf{Z}_{j2}\boldsymbol{\beta}_{j2} + \cdots + \boldsymbol{\varepsilon}_j,$$

$$\boldsymbol{\beta}_{j1} = \mathbf{Z}_{j11}\boldsymbol{\beta}_{j11} + \mathbf{Z}_{j12}\boldsymbol{\beta}_{j12} + \cdots + \boldsymbol{\varepsilon}_{j1}.$$

Here, the first covariate z_{j1} in the district-specific effect is another cluster indicator that indicates the state in which the districts are nested. Hence, $\mathbf{Z}_{j1}$ is another incidence matrix and $\boldsymbol{\beta}_{j1}$ is the vector of state-specific effects modeled through the level-3 equation.

We finally point out that the compound priors are not necessarily restricted to random effects modeling as described above. For instance, $\boldsymbol{Z}_j\boldsymbol{\beta}_j$ in (3) may comprise a smooth spatial term modeled by radial basis functions centered at the observed locations. The common assumption of a Gaussian random field for the regression coefficients $\boldsymbol{\beta}_j$ implies that parameters in close proximity are more alike than others. However, in many spatial applications the definition of locational similarity may be given by a bunch of similar locational characteristics (e.g. soil conditions) and less by spatial proximity in the narrow sense. This could be modeled using the compound prior (4) by regressing the coefficients $\boldsymbol{\beta}_j$ (nonparametrically) on location specific covariates.

3 MCMC Inference based on the original parametrization

We first discuss direct MCMC schemes based on the original parametrization of the previous sections. In Sect. 4, we provide an MCMC scheme which uses an alternative parametrization that results in diagonal precision matrices.

3.1 Gaussian responses

We first describe a Gibbs sampler for models with Gaussian errors. For the sake of simplicity, we restrict the presentation to a two level hierarchical model with one level-2 equation for the regression coefficients of the first term $\boldsymbol{Z}_1\boldsymbol{\beta}_1$. That is, the level-1 equation is $\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ with predictor (3) and errors $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{W}^{-1})$ with diagonal weight matrix $\boldsymbol{W} = \operatorname{diag}(w_1, \dots, w_n)$. The level-2 equation is of the form (4) with $j = 1$. Inference for models with more than two hierarchy levels or more level-2 equations is straightforward (and of course fully supported by our software), see also Sect. 5 for applications of three level models.

Based on usual conditional independence assumptions, the posterior is proportional to

$$L\big(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q, \boldsymbol{\gamma}, \sigma^2\big) \prod_{j=1}^{q} \big[p\big(\boldsymbol{\beta}_j \mid, \tau_j^2\big) p\big(\tau_j^2\big)\big] p(\boldsymbol{\gamma}) p\big(\sigma^2\big)$$

$$\prod_{j=1}^{q_1} \big[p\big(\boldsymbol{\beta}_{1j} \mid \tau_{1j}^2\big) p\big(\tau_{1j}^2\big)\big] p(\boldsymbol{\gamma}_1) p\big(\tau_1^2\big), \tag{11}$$

where $L(\cdot)$ denotes the likelihood which is the product of individual likelihood contributions.

The parameters are updated in blocks where each vector of regression coefficients $\boldsymbol{\beta}_j$ ($\boldsymbol{\beta}_{1l}$ in a second level of the hierarchy) of a particular term is updated in one (possibly large) block followed by updating the regression coefficients $\boldsymbol{\gamma}, \boldsymbol{\gamma}_1$ of linear effects and the variance components τ_j^2, τ_{1l}^2, σ^2. Simultaneously updating the regression coefficients $\boldsymbol{\beta}_j$ ($\boldsymbol{\beta}_{1l}$) and the corresponding variance component τ_j^2 (τ_{1l}^2) is possible and sometimes useful, see Rue and Held (2005) or Brezger and Lang (2006).

The full conditionals for the regression coefficients $\boldsymbol{\beta}_1$ with the compound prior (4) and the coefficients $\boldsymbol{\beta}_j$, $j = 2, \dots, q$, $\boldsymbol{\beta}_{1l}$, $l = 1, \dots, q_1$ with the basic prior (8) are all multivariate Gaussian. The respective posterior precision $\boldsymbol{\Sigma}^{-1}$ and mean $\boldsymbol{\mu}$ is given by

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1} &= \frac{1}{\sigma^2}\left(\boldsymbol{Z}_1'\boldsymbol{W}\boldsymbol{Z}_1 + \frac{\sigma^2}{\tau_1^2}\boldsymbol{I}\right),\\
\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \frac{1}{\sigma^2}\boldsymbol{Z}_1'\boldsymbol{W}\boldsymbol{r} + \frac{1}{\tau_1^2}\boldsymbol{\eta}_1 \quad (\boldsymbol{\beta}_1 \text{ compound prior}),\\
\boldsymbol{\Sigma}^{-1} &= \frac{1}{\sigma^2}\left(\boldsymbol{Z}_j'\boldsymbol{W}\boldsymbol{Z}_j + \frac{\sigma^2}{\tau_j^2}\boldsymbol{K}_j\right),\\
\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \frac{1}{\sigma^2}\boldsymbol{Z}_j'\boldsymbol{W}\boldsymbol{r} \quad (\boldsymbol{\beta}_j \text{ level-1 equation}),\\
\boldsymbol{\Sigma}^{-1} &= \frac{1}{\tau_1^2}\left(\boldsymbol{Z}_{1l}'\boldsymbol{Z}_{1l} + \frac{\tau_1^2}{\tau_{1l}^2}\boldsymbol{K}_{1l}\right),\\
\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \frac{1}{\tau_1^2}\boldsymbol{Z}_{1l}'\boldsymbol{r}_1 \quad (\boldsymbol{\beta}_{1l} \text{ level-2 equation}),
\end{aligned} \tag{12}$$

where $\boldsymbol{r}$ is the current partial residual and $\boldsymbol{r}_1$ is the "partial residual" of the level-2 equation. More precisely, $\boldsymbol{r}_1 = \boldsymbol{\beta}_1 - \tilde{\boldsymbol{\eta}}_1$ and $\tilde{\boldsymbol{\eta}}_1$ is the predictor of the level-2 equation excluding the current effect of z_{1l}.

MCMC updates of the regression coefficients take advantage of the following key features:

Sparsity Design matrices $\boldsymbol{Z}_j, \boldsymbol{Z}_{1l}$ as well as their cross products $\boldsymbol{Z}_j'\boldsymbol{W}\boldsymbol{Z}_j, \boldsymbol{Z}_{1l}'\boldsymbol{Z}_{1l}$ and associated penalty matrices $\boldsymbol{K}_j, \boldsymbol{K}_{1l}$ and posterior precision matrices in (12) are often sparse. The sparsity can be exploited for highly efficient computation of cross products (Sect. 3.3), Cholesky decompositions of posterior precision matrices and for fast solving of relevant linear equation systems. In some cases, appropriate reordering of the parameters is required. The parameters may be reordered according to the reverse Cuthill-McKee algorithm or the (approximate) minimum degree algorithm, see Davis (2006) for a recent reference.

Reduced complexity in the second or third stage of the hierarchy Updating the regression coefficients $\boldsymbol{\beta}_{1l}$, $l = 1, \dots, q_1$, in the second (or third level) is done conditionally on the parameter vector $\boldsymbol{\beta}_1$. This facilitates updating the parameters for two reasons. First the number of "observations" in the level-2 equation is equal to the length of the vector $\boldsymbol{\beta}_1$ and therefore much smaller than the actual number of observations n. Second the full conditionals for $\boldsymbol{\beta}_{1l}$ are Gaussian regardless of the response distribution in the first level of the hierarchy.

Number of different observations smaller than sample size In most cases the number m_j of different observations $z_{(1)}, \dots, z_{(m_j)}$ in $\boldsymbol{Z}_j$ (or m_{1l} in $\boldsymbol{Z}_{1l}$ in the level-2 equation) is much smaller than the total number n of observations. The fact that $m_j \ll n$ may be utilized to considerably speed up computations of the cross products $\boldsymbol{Z}_j'\boldsymbol{W}\boldsymbol{Z}_j$, $\boldsymbol{Z}_{1l}'\boldsymbol{Z}_{1l}$, the vectors $\boldsymbol{Z}_j'\boldsymbol{W}\boldsymbol{r}$, $\boldsymbol{Z}_{1l}'\boldsymbol{r}_1$ and finally the updated vectors of function evaluations $\boldsymbol{f}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$, $\boldsymbol{f}_{1l} = \boldsymbol{Z}_{1l}\boldsymbol{\beta}_{1l}$. Details will be given in Sect. 3.3. Note that efficient computation of cross products and function evaluations contributes at least as much to computational efficiency as the sparse matrix algorithms to solve relevant linear equation systems.

3.2 Non-Gaussian responses

The non-Gaussian case can often be traced back to Gaussian regression models via data augmentation as has been proposed for the first time in the seminal paper by Albert and Chib (1993) for parametric probit models. Since then other data augmentation schemes for logit models (Holmes and Held 2006; Frühwirth-Schnatter and Frühwirth 2010), Poisson regression (Frühwirth-Schnatter et al. 2009) and certain types of Gamma regression models (Frühwirth-Schnatter et al. 2009) have been developed. We very briefly illustrate the concept for probit models, i.e. $y_i \sim B(1, \Phi(\eta_i))$ where Φ is the cdf of a standard normal distribution. Introducing latent variables $U_i = \eta_i + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$, we obtain $y_i = 1$ if $U_i > 0$ and $y_i = 0$ if $U_i < 0$. The posterior of the model augmented by the latent variables depends now on the extra parameters U_i and additional sampling steps for updating the U_i's are required. Sampling the U_i's is relatively easy and fast because the full conditionals are truncated normal distributions, i.e. $U_i \mid \cdot \sim N(\eta_i, 1)$ truncated at the left by 0 if $y_i = 1$ and truncated at the right if $y_i = 0$. The advantage of defining a probit model through the latent variables U_i is that the full conditionals for the regression parameters are almost unchanged with the responses y_i in (12) replaced by the latent variables U_i. The other data augmentation approaches mentioned above work similar and are only slightly more complex.

In cases where data augmentation is not possible the regression parameters of the level-1 equation can be updated using Metropolis-Hastings steps with IWLS proposals as described for simple STAR models in Brezger and Lang (2006). The tricks for computationally improved MCMC sampling summarized in the previous subsection and detailed in the following subsections can still be used with minor modifications.

3.3 Efficient computation of $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$ and $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r}$

We describe efficient computation for a particular varying coefficient term

$$f(z) = f\big(z^{(1)}, z^{(2)}\big) = z^{(1)} g\big(z^{(2)}\big) \tag{13}$$

in the level-1 or level-2 equation with design matrix

$$\boldsymbol{Z} = \mathrm{diag}\big(z_1^{(1)}, \dots, z_n^{(1)}\big)\boldsymbol{Z}^{(2)} = \boldsymbol{D}\boldsymbol{Z}^{(2)}$$

where $\boldsymbol{D} = \mathrm{diag}(z_1^{(1)}, \dots, z_n^{(1)})$. Computation for a pure additive term, i.e. $\boldsymbol{D} = \boldsymbol{I}$, arises as a special case.

Denote by $z_{(1)}^{(2)} < z_{(2)}^{(2)} < \cdots < z_{(m)}^{(2)}$ the m ordered different observations of $z^{(2)}$. Compute the index vector $\boldsymbol{ind}$ with elements $\boldsymbol{ind}[i] \in \{1, \dots, m\}$ denoting the category of the i-th observation, i.e. if $z_i^{(2)} = z_{(j)}^{(2)}$ then $\boldsymbol{ind}[i] = j$. The index vector $\boldsymbol{ind}$ is required to match the sorted observations of $z^{(2)}$ with the response observations which can not be sorted directly because different model terms would result in different sorting.

We can now decompose the design matrix in $\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{P}\tilde{\boldsymbol{Z}}$, where

- $\tilde{\boldsymbol{Z}}$ is the $m \times K$ reduced design matrix for the different and sorted observations $z_{(1)}^{(2)}, \dots, z_{(m)}^{(2)}$, i.e. $\tilde{\boldsymbol{Z}}[s,k] = B_k(z_{(s)}^{(2)})$, $s = 1, \dots, m$, $k = 1, \dots, K$,
- $\boldsymbol{P}$ is a $n \times m$ permutation matrix, which reverts the sorting, i.e. $\boldsymbol{P}[i,s] = I(\boldsymbol{ind}(i) = s)$. Note that $\boldsymbol{P}$ is defined for presentation purposes and will not be computed explicitly.

For the vector of function evaluations we obtain $\boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{D}\boldsymbol{P}\tilde{\boldsymbol{Z}}\boldsymbol{\beta}$.

Computation of $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$ We get

$$\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} = \tilde{\boldsymbol{Z}}'\boldsymbol{P}'\boldsymbol{D}'\boldsymbol{W}\boldsymbol{D}\boldsymbol{P}\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Z}},$$

where $\tilde{\boldsymbol{W}} = \boldsymbol{P}'\boldsymbol{D}'\boldsymbol{W}\boldsymbol{D}\boldsymbol{P} = \mathrm{diag}(\tilde{w}_1, \dots, \tilde{w}_m)$ and the "reduced" weights $\tilde{w}_s$, $s = 1, \dots, m$, are given by

$$\tilde{w}_s = \sum_{i:\boldsymbol{ind}[i]=s} \big(z_i^{(1)}\big)^2 w_i. \tag{14}$$

The weights $\tilde{w}_s$ can be computed by first initializing $\tilde{w}_s = 0$ followed by a simple loop: For $i = 1, \dots, n$ add $(z_i^{(1)})^2 w_i$ to $\tilde{w}_{\boldsymbol{ind}[i]}$. Hence, the computation of the cross product $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$ is reduced to the computation of the cross product $\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Z}}$ where the dimension of $\tilde{\boldsymbol{Z}}$ is much more favorable in terms of computational costs than the dimension of the original design matrix $\boldsymbol{Z}$. Note that the reduced design matrix $\tilde{\boldsymbol{Z}}$ is still a sparse matrix. The sparsity can be exploited for efficient computation by using standard algorithms for sparse matrix multiplications as for example given in Davis (2006, Chap. 2.8). However, since $\tilde{\boldsymbol{Z}}$ usually remains constant during the MCMC run an even faster algorithm is possible:

Efficient computation of $\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Z}}$ We store $\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Z}}$ in sparse matrix format. Although the particular sparse matrix storage format differs from implementation to implementation there is always a vector, $\boldsymbol{C}$ say, that stores the nonzero entries of $\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Z}}$. Let n_z be the number of nonzero entries of $\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Z}}$, i.e. the dimension of $\boldsymbol{C}$. Suppose that the t-th entry $\boldsymbol{C}[t]$ of $\boldsymbol{C}$ corresponds to the element in the r-th row and l-th column of $\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Z}}$. Then we have

$$\boldsymbol{C}[t] = \sum_{s=1}^{m} \tilde{w}_s \tilde{\boldsymbol{Z}}[s,r]\tilde{\boldsymbol{Z}}[s,l],$$

where most of the products $\tilde{\boldsymbol{Z}}[s,r]\tilde{\boldsymbol{Z}}[s,l]$ are zero because either $\boldsymbol{Z}[s,r]$ or $\boldsymbol{Z}[s,l]$ or both are zero. We now store the nonzero products $\tilde{\boldsymbol{Z}}[s,r]\tilde{\boldsymbol{Z}}[s,l]$ required to compute $\boldsymbol{C}[t]$ in the auxiliary vector $\boldsymbol{h}_1$, the corresponding index s in the auxiliary vector $\boldsymbol{h}_2$ and the position of the first element in $\boldsymbol{h}_1$ corresponding to $\boldsymbol{C}[t]$ in the $(n_z + 1) \times 1$ index vector $\boldsymbol{h}_3$. The last element $\boldsymbol{h}_3[n_z + 1]$ in $\boldsymbol{h}_3$ is the dimension of $\boldsymbol{C}$. Then $\boldsymbol{C}[t]$ is efficiently computed as

$$\boldsymbol{C}[t] = \sum_{s=\boldsymbol{h}_3[t]}^{\boldsymbol{h}_3[t+1]-1} \tilde{w}_{\boldsymbol{h}_2[s]}\boldsymbol{h}_1[s].$$

Computation of $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r}$ For $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r}$ we obtain

$$\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r} = \tilde{\boldsymbol{Z}}'\boldsymbol{P}'\boldsymbol{D}'\boldsymbol{W}\boldsymbol{r} = \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{r}},$$

where the $m \times 1$ vector $\tilde{\boldsymbol{r}} = (\tilde{r}_1, \dots, \tilde{r}_m)'$ of "reduced" partial residuals is given by

$$\tilde{r}_s = \sum_{i:\boldsymbol{ind}[i]=s} z_i^{(1)} w_i r_i. \tag{15}$$

The $\tilde{r}_s$ are computed by first initializing $\tilde{r}_s = 0$ followed by the loop: For $i = 1, \dots, n$ add $z_i^{(1)} w_i r_i$ to $\tilde{r}_{\boldsymbol{ind}[i]}$. Once the reduced partial residual vector $\tilde{\boldsymbol{r}}$ is computed, the product $\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{r}}$ is obtained via sparse matrix-vector multiplications.

Remarks

1. *Indicator functions:* A particularly simple expression for $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$ and $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r}$ is obtained if the $B_k(z)$ are indicator functions, i.e. $B_k(z) \in \{0, 1\}$ and for a particular value z we have $B_k(z) = 1$ for exactly one $k \in \{1, \dots, K\}$. Typical examples are Markov random fields for modeling spatial heterogeneity or P-splines of degree 0 (simple random walk priors). Another example arises if the effect $\boldsymbol{Z}_1\boldsymbol{\beta}_1$ with compound prior for $\boldsymbol{\beta}_1$ models cluster- or individual-specific heterogeneity. In this case covariate $z_1 \in \{1, \dots, K\}$ corresponds to a cluster index and $\boldsymbol{Z}_1$ is an incidence matrix with elements either 0 or 1. In all examples the cross product $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$ reduces to the diagonal matrix $\tilde{\boldsymbol{W}} = \mathrm{diag}(\tilde{w}_1, \dots, \tilde{w}_m)$ and the product $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r}$ reduces to $\tilde{\boldsymbol{r}}$.

2. *Binning:* The efficiency of the formulae for computing $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$ and $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r}$ depends on the number m of different observations in the covariate vector $\boldsymbol{z}^{(2)}$. For large m, a simple device for increasing computational efficiency is to perform binning of the data. For continuous $z^{(2)}$ a very simple solution is rounding the data to a certain degree. Alternatively we may group the data according to an equidistant grid. Suppose that the support of the data is the interval $[a,b]$ and that we want to replace the observations $z_1^{(2)},\dots,z_n^{(2)}$ by a grid of m equally spaced design points

   $$a+\delta/2 = z_{(1)}^{(2)} < z_{(2)}^{(2)} < \dots < z_{(m)}^{(2)} = b-\delta/2.$$

   Here $\delta=(b-a)/m$ is the grid width. It is natural to replace a value $z^{(2)}$ by the design point which is closest in absolute value to $z^{(2)}$. Define for every value $z^{(2)}$ the index $h=\text{floor}((z^{(2)}-a)/\delta)$. Then we obtain $z_{\text{new}}^{(2)} = a+\delta/2+h\cdot\delta$.

   To give an example, computing time is reduced by approximately 40 to 70 percent (depending on the response distribution) for a simple model with one nonlinear function modeled by P-splines and 1000 different covariate observations.

3.4 Algorithm for updating regression parameters of nonlinear effects

On the basis of the preceding subsections we are now ready to describe an algorithm for updates of the regression parameters of nonlinear terms. We restrict the presentation to Gaussian responses. Adapting the algorithm for non-Gaussian responses using data augmentation or IWLS proposals as sketched in Sect. 3.2 is straightforward.

We describe a generic algorithm for updating an arbitrary vector of regression coefficients $\boldsymbol{\beta}$ regardless of the hierarchy level and its prior (compound prior (4) or basic prior (8)). This means that we need to implement only *one* algorithm for updating the regression coefficients of any hierarchy level. The input of the algorithm is a (pseudo) "response" vector $\tilde{\boldsymbol{y}}$, a diagonal matrix of weights $\tilde{\boldsymbol{W}}$, a predictor $\tilde{\boldsymbol{\eta}}$, a vector of regression coefficients $\boldsymbol{\beta}$, a vector of function evaluations $\boldsymbol{f}$, a (reduced) design matrix $\tilde{\boldsymbol{Z}}$ and its transpose $\tilde{\boldsymbol{Z}}'$, an index vector $\boldsymbol{ind}$, a cross product matrix $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$, a vector $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r}$, a penalty matrix $\boldsymbol{K}$ and a precision matrix $\boldsymbol{\Sigma}^{-1}$. The specific values passed to the algorithm depend on the respective model term, the hierarchy level and the prior. For instance, $\tilde{\boldsymbol{y}}=\boldsymbol{y}$, $\tilde{\boldsymbol{\eta}}=\boldsymbol{\eta}$, $\tilde{\boldsymbol{W}}=\boldsymbol{W}$ when updating a parameter vector of the level-1 equation and $\tilde{\boldsymbol{y}}=\boldsymbol{\beta}_1$, $\tilde{\boldsymbol{\eta}}=\boldsymbol{\eta}_1$, $\tilde{\boldsymbol{W}}=\boldsymbol{I}$ when updating a level-2 parameter vector. Some of the input vectors and matrices are modified by the algorithm. The algorithm is implemented using the following steps:

Algorithm ($\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{W}}$, $\tilde{\boldsymbol{\eta}}$, $\boldsymbol{\beta}$, $\boldsymbol{f}$, $\tilde{\boldsymbol{Z}}$, $\tilde{\boldsymbol{Z}}'$, $\boldsymbol{ind}$, $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$, $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r}$, $\boldsymbol{K}$, $\boldsymbol{\Sigma}^{-1}$)

1. Substract $\boldsymbol{f}$ from $\tilde{\boldsymbol{\eta}}$: $\tilde{\boldsymbol{\eta}}=\tilde{\boldsymbol{\eta}}-\boldsymbol{f}$ and compute the partial residual: $\boldsymbol{r}=\boldsymbol{y}-\tilde{\boldsymbol{\eta}}$.
2. Compute the cross product matrix $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}=\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Z}}$ and the vector $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{r}=\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{r}}$, based on the algorithms developed in Sect. 3.3. In models with Gaussian errors it is sufficient to compute the cross product $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$ once at the outset of the iterations because quantities involved remain constant. However, for non-Gaussian responses and some extensions as generalized random slope modeling defined in (7) the cross product has to be recomputed in every iteration of the sampler.
3. Compute the posterior precision matrix $\boldsymbol{\Sigma}^{-1}$, see formula (12), and its Cholesky decomposition: $\boldsymbol{\Sigma}^{-1}=\boldsymbol{L}\boldsymbol{L}'$.
4. Sample $\boldsymbol{\beta}$: First solve $\boldsymbol{L}'\boldsymbol{\beta}^*=\boldsymbol{u}$, where $\boldsymbol{u}$ is a vector of independent standard Gaussians. It follows that $\boldsymbol{\beta}^*\sim N(\boldsymbol{0},\boldsymbol{\Sigma})$. Compute the mean $\boldsymbol{\mu}$ by solving for $\boldsymbol{\mu}$ in (12) and add the mean $\boldsymbol{\mu}$ to the previously simulated $\boldsymbol{\beta}^*$. Finally correct the unconstraint vector $\boldsymbol{\beta}^*$ by

   $$\boldsymbol{\beta}=\boldsymbol{\beta}^*-\boldsymbol{\Sigma}\boldsymbol{A}'\left(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'\right)^{-1}\boldsymbol{A}\boldsymbol{\beta}.$$

   This is done at negligible computational cost using steps 5–9 of algorithm 2.6 in Rue and Held (2005).
5. Update the vector of function evaluations $\boldsymbol{f}=\boldsymbol{Z}\boldsymbol{\beta}=\boldsymbol{P}\tilde{\boldsymbol{Z}}\boldsymbol{\beta}$ (or $\boldsymbol{f}=\boldsymbol{D}\boldsymbol{P}\tilde{\boldsymbol{Z}}\boldsymbol{\beta}$ for varying coefficients terms). The first step is to compute the product $\tilde{\boldsymbol{f}}=\tilde{\boldsymbol{Z}}\boldsymbol{\beta}$ using sparse matrix - vector multiplications. Then the i-th element of $\boldsymbol{f}$ is given by $\boldsymbol{f}[i]=\tilde{\boldsymbol{f}}[\boldsymbol{ind}[i]]$ (or $\boldsymbol{f}[i]=z_i^{(1)}\tilde{\boldsymbol{f}}[\boldsymbol{ind}[i]]$ for varying coefficients terms) .
6. Update the predictor: $\tilde{\boldsymbol{\eta}}=\tilde{\boldsymbol{\eta}}+\boldsymbol{f}$

The generic algorithm is typically implemented as a function that takes the input vectors and matrices of the algorithm as arguments and modifies parts of these quantities. Since the algorithm updates parameter vectors of arbitrary hierarchy levels estimation of complex multilevel models is easily obtained by subsequently calling the function that implements the algorithm.

4 MCMC inference based on an alternative parametrization

In this section we develop an alternative to the sampling scheme outlined in Sect. 3. The new scheme is particularly useful for situations where the design and penalty matrix is not sparse as is for example the case for Gaussian random fields. The alternative sampling scheme works with a transformed parametrization such that the cross product of the design matrix and the penalty matrix of a nonlinear term are diagonal resulting in a diagonal posterior precision matrix. In the context of spline smoothing the resulting basis functions are known as the Demmler-Reinsch basis. For pure additive models based on P-splines the Demmler-Reinsch basis has been used for (frequentist) inference in Ruppert (2002).

We describe the alternative parametrization for a particular nonlinear function f with design matrix $\boldsymbol{Z} = \boldsymbol{P}\tilde{\boldsymbol{Z}}$ and parameter vector $\boldsymbol{\beta}$ with general prior (8).

Let $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} = \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{Z}} = \boldsymbol{R}\boldsymbol{R}'$ be the Cholesky decomposition of the cross product of the design matrix and let $\boldsymbol{Q}\boldsymbol{S}\boldsymbol{Q}'$ be the singular value decomposition of $\boldsymbol{R}^{-1}\boldsymbol{K}\boldsymbol{R}^{-T}$. The diagonal matrix $\boldsymbol{S} = \text{diag}(s_1, \dots, s_K)$ contains the eigenvalues of $\boldsymbol{R}^{-1}\boldsymbol{K}\boldsymbol{R}^{-T}$ in ascending order. The columns of the orthogonal matrix $\boldsymbol{Q}$ contain the corresponding eigenvectors. Columns 1 through $\text{rk}(\boldsymbol{K})$ form a basis for the vector space spanned by the columns of $\boldsymbol{R}^{-1}\boldsymbol{K}\boldsymbol{R}^{-T}$. The remaining columns are a basis of the nullspace.

Then the decomposition $\boldsymbol{\beta} = \boldsymbol{R}^{-T}\boldsymbol{Q}\bar{\boldsymbol{\beta}}$ yields

$$\boldsymbol{P}\tilde{\boldsymbol{Z}}\boldsymbol{\beta} = \boldsymbol{P}\tilde{\boldsymbol{Z}}\boldsymbol{R}^{-T}\boldsymbol{Q}\bar{\boldsymbol{\beta}} = \bar{\boldsymbol{Z}}\bar{\boldsymbol{\beta}},$$

where the transformed design matrix $\bar{\boldsymbol{Z}}$ is defined by $\bar{\boldsymbol{Z}} = \boldsymbol{P}\tilde{\boldsymbol{Z}}\boldsymbol{R}^{-T}\boldsymbol{Q}$. Note that $\bar{\boldsymbol{Z}}$ is a dense matrix in contrast to the sparse original design matrix $\boldsymbol{Z}$.

We now obtain for the cross product

$$\bar{\boldsymbol{Z}}'\boldsymbol{W}\bar{\boldsymbol{Z}} = \boldsymbol{Q}'\boldsymbol{R}^{-1}\tilde{\boldsymbol{Z}}'\boldsymbol{P}'\boldsymbol{W}\boldsymbol{P}\tilde{\boldsymbol{Z}}\boldsymbol{R}^{-T}\boldsymbol{Q} = \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$$

and for the penalty

$$\boldsymbol{\beta}'\boldsymbol{K}\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}'\boldsymbol{Q}'\boldsymbol{R}^{-1}\boldsymbol{K}\boldsymbol{R}^{-T}\boldsymbol{Q}\bar{\boldsymbol{\beta}} = \bar{\boldsymbol{\beta}}'\boldsymbol{S}\bar{\boldsymbol{\beta}},$$

with the new diagonal penalty matrix $\boldsymbol{S}$ given by the singular value decomposition of $\boldsymbol{R}^{-1}\boldsymbol{K}\boldsymbol{R}^{-T}$, see above.

Summarizing, we obtain the equivalent formulation $\boldsymbol{f} = \bar{\boldsymbol{Z}}\bar{\boldsymbol{\beta}}$ for the vector of function evaluations based on the transformed design matrix $\bar{\boldsymbol{Z}}$ and the transformed parameter vector $\bar{\boldsymbol{\beta}}$ with (possibly improper) Gaussian prior

$$\bar{\boldsymbol{\beta}} \mid \tau^2 \sim N\big(\boldsymbol{0}, \tau^2\boldsymbol{S}^{-}\big).$$

The advantage of the scheme is that the prior precision or penalty matrix $\boldsymbol{S}$ is diagonal resulting in a diagonal posterior precision matrix. More specifically, the full conditional for $\bar{\boldsymbol{\beta}}$ is Gaussian with k-th element μ_k, $k = 1, \dots, K$, of the mean vector $\boldsymbol{\mu}$ given by

$$\mu_k = \frac{1}{1 + \lambda s_k} \cdot u_k,$$

where $\lambda = \sigma^2/\tau^2$ and u_k is the k-th element of the vector $\boldsymbol{u} = \bar{\boldsymbol{Z}}'\boldsymbol{W}\boldsymbol{r}$ with $\boldsymbol{r}$ the partial residual. The covariance matrix $\boldsymbol{\Sigma}$ is diagonal with diagonal elements

$$\boldsymbol{\Sigma}[k,k] = \frac{\sigma^2}{1 + \lambda s_k}.$$

For MCMC simulation the matrix products $\boldsymbol{u} = \bar{\boldsymbol{Z}}'\boldsymbol{W}\boldsymbol{r}$ and $\boldsymbol{f} = \bar{\boldsymbol{Z}}\bar{\boldsymbol{\beta}}$ must be computed in every iteration of the sampler. The $n \times K$ design matrix $\bar{\boldsymbol{Z}}$ is a dense matrix that contains no zero elements. There is, however, a more efficient way to compute the required quantities than by direct matrix multiplication.

To compute $\boldsymbol{u}$ we first note that $\boldsymbol{u} = \bar{\boldsymbol{Z}}'\boldsymbol{W}\boldsymbol{r} = \boldsymbol{Q}'\boldsymbol{R}^{-1}\tilde{\boldsymbol{Z}}'\boldsymbol{P}'\boldsymbol{W}\boldsymbol{r}$. Since $\boldsymbol{P}'\boldsymbol{W}\boldsymbol{r} = \tilde{\boldsymbol{r}}$ is the reduced partial residual defined in Sect. 3.3 we get $\boldsymbol{u} = \boldsymbol{Q}'\boldsymbol{R}^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{r}}$. Hence $\boldsymbol{u}$ is obtained by first computing the product $\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{r}}$ using standard sparse matrix multiplications (or the even more efficient algorithm described in Sect. 3.3) and by multiplying the result with the $K \times K$ matrix $\boldsymbol{Q}'\boldsymbol{R}^{-1}$ which can be computed offline.

For computing the second product $\boldsymbol{f} = \bar{\boldsymbol{Z}}\bar{\boldsymbol{\beta}}$ we note that $\boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\beta}$ and $\boldsymbol{\beta} = \boldsymbol{R}^{-T}\boldsymbol{Q}\bar{\boldsymbol{\beta}}$. Hence $\boldsymbol{f}$ is obtained by first computing the untransformed $\boldsymbol{\beta}$ followed by step 5 of the algorithm described in Sect. 3.4.

The main advantage of the alternative transformation is that it provides fast MCMC inference even in situations where the posterior precision is relatively dense as is the case for many surface estimators. The prime example is a Gaussian random field which is almost intractable in the standard parametrization (see Hennerfeind et al. 2006). Using the approach described in this section MCMC inference for Gaussian random fields is extremely fast.

The main disadvantage of the sampling scheme is that it works only for fixed design, i.e. the design matrix $\boldsymbol{Z}$ and the weights $\boldsymbol{W}$ must be constant during an MCMC run. Otherwise the relatively costly singular value decomposition must be recomputed in every iteration of the sampler. This excludes MH updates with IWLS proposals as proposed in Brezger and Lang (2006).

5 Case study on child undernutrition in India

In this section we apply our methodology to data on the determinants of child undernutrition in India. The analysis is based on micro data from the second National Family Health Survey (NFHS-2) from India which was conducted in the years 1998 and 1999. The sample is representative of the population and collectes detailed health and anthropometric information on approximately 30000 children born in the 3 years preceding the survey.

Using the methodology of this paper we estimated the probit model (6) described in the introduction. The presentation is restricted to the most interesting covariates from a statistical point of view. Note, however, that all relevant covariates (e.g. the birth order or the household size) are included in our models but not discussed in this methodological paper.

For the nonlinear effects of continuous covariates, cubic P-splines with 20 inner knots have been specified. The

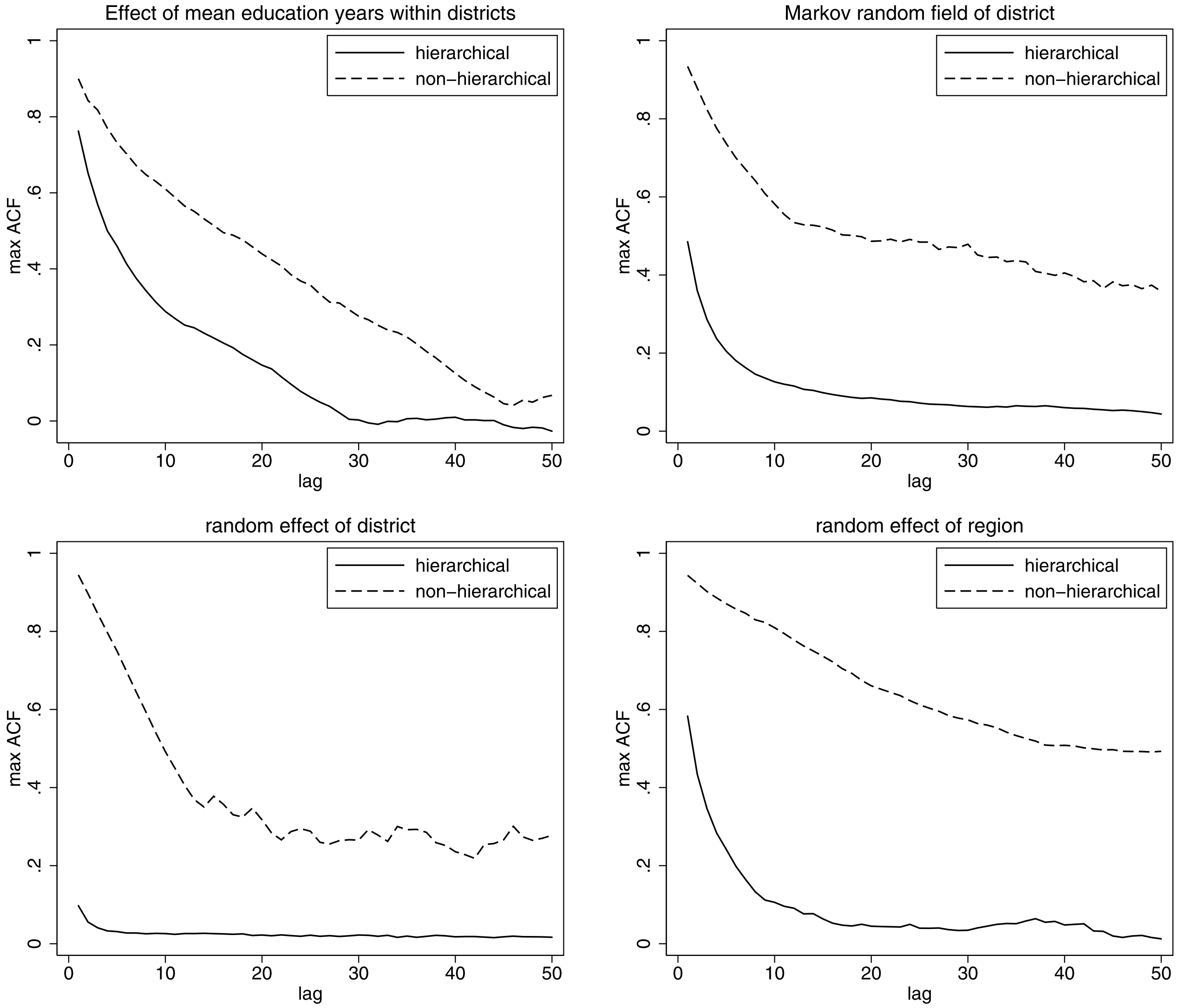

Fig. 1 Maximum autocorrelations for selected effects

smooth spatial effects $f_{9,3}(dist)$ and $f_{10,3}(dist)$ are modeled either by Markov random fields or Gaussian random fields with 50 representative knots (low rank approximation). The latter is estimated via the alternative parametrization outlined in Sect. 4 while all other terms can be estimated in the original parametrization. The results for both approaches to spatial smoothing are similar although Gaussian random fields shows a substantially lower deviance information criterion (DIC) (Spiegelhalter et al. 2002) with a difference of more than 50 points. Surprisingly the difference is due to a reduced deviance for the model based on Gaussian random fields while the equivalent degrees of freedom of both modeling variants are almost identical. This means that Gaussian random fields produce a better fit with less parameters.

5.1 Hierarchical versus non-hierarchical formulation

We first compare the hierarchical formulation of the model as outlined in this paper with a non-hierarchical version. In principle, a non-hierarchical reduced form could be estimated using the technology outlined primarily in Lang and Brezger (2004) and Brezger and Lang (2006). However, estimation of the full model (6) turned out to be not feasible because of very slow mixing and corresponding numerical problems. The comparison is therefore restricted to a main effects model with a reduced set of covariates. Estimation of the hierarchical version of this reduced model takes between 25 % and 50 % (depending on the operating system and the compiler used) of the non-hierarchical version (with the same number of iterations). Even more important is the by far superior mixing of sampled parameters as is demonstrated through Fig. 1. The figure shows for selected model terms the maximum autocorrelations of the corresponding parameters for lag sizes between 1 and 50. While for the hierarchical version the maximum autocorrelations decline rather quickly, we observe persistent autocorrelation with the non-hierarchical version. The autocorrelation functions of the hierarchical model suggest that 20000 to 30000 iterations after the burnin period should be sufficient to obtain 1000 nearly uncorrelated samples if every 20th to 30th sample is used. On the other hand the autocorrelation functions for the non-hierarchical version show that estimation of complex multilevel models using standard MCMC technology is not feasible.

To be on the safe side, the following results are based on 50000 iterations after a burnin period of 3000 iterations. On modern personal computers estimation takes between 10 and 20 minutes depending on the actual processor. Note that we have not run parallel chains which would reduce computing time even further (approximately 30–35 % of the computing time of a single chain on a usual quad core processor). Note also that in the model building phase 10000 iterations after the burnin period are enough to obtain sufficiently accurate preliminary results.

5.2 Results for nonlinear covariate effects

Figures 2 and 3 show estimated nonlinear effects of all hierarchy levels. The results rely on the modeling variant based on Gaussian random fields for the smooth spatial effect. Shown are the posterior means together with 95 % pointwise and simultaneous credible bands. The simultaneous credible intervals are based on a proposal by Krivobokova et al. (2010). Of the various interactions with gender, the varying effects with the child's age and mother's age at first birth are "significant" in the sense that at least the 95 % pointwise credible intervals do not fully cover the zero line. Therefore the presentation of interaction effects are restricted to *c_age* and *ageb*. We also completely omitted results for the gross national product per capita (*gnp*) in the level-3 equations as the effects are practically zero. Although this result is quite surprising, also other studies have failed to identify an effect of GDP per capita on child undernutrition in India using large scale household survey data (Subramanyam et al. 2011).

The age effect (left panel of Fig. 2) shows that the probability of being stunted in India rapidly increases between age 0 and about 20 months after which it oscillates. This is in line with findings from other studies and indicates that children are not born chronically malnourished but develop this as a result of disease and inadequate nutritional intake. The sudden improvement of the nutritional status around 24 months is an artifact of the reference standard as at this age, children switch from being compared to the better nourished reference children from the white, bottle-fed Fels study (Ohio Fels Research Institute), to the worse nourished reference children derived from a cross-section of the US population, see WHO (2002, pp. 4–6). The interaction with gender shows that females are less likely to be stunted than males up to the age of 20 months. This is in agreement with our expectations as male newborns are typically more vulnerable than females. More surprising is the fact that after 20 months the situation is reverted and female children are now more likely to be stunted than male children. This suggests that males have better access to limited (food) resources than females. This interesting finding supports the hypotheses among development economists that male children have a cultural advantage in South Asian countries because parents profit more from male offsprings (e.g. they are more beneficial after retirement), see e.g. Klasen (1996) and Somerfelt and Arnold (1999).

The effects of all other covariates in the study are much weaker than the age effect. An example is the effect of mother's age at first birth. This effect shows a U-form, i.e. children are most healthy if the mother's age at first birth is around 25 years. For younger and older mothers the probability of stunted children is increased (although the effect is not strong). The interaction effect provides evidence that the more problematic situation of old mother's is more risky for females than for males. The observation that "problematic situations" are riskier for females than males is also supported by some of the other interaction effects. Albeit not significant, they all point in the same direction that males are less affected by problematic situations (e.g. regarding the household wealth) than females.

For modeling the household's wealth and education effect we have used the multilevel structure of the data and estimated for both covariates external effects at district level by including the average wealth index and education years per district in the level-2 equation. At least for the wealth index such an external effect can be observed (top left panel of Fig. 3). Children who are born in a wealthier environment (district) are less likely to be stunted than children living in poor districts. There is, however, an additional household effect, see the bottom left panel of Fig. 3. Children in households which are wealthier than the district mean are less affected by stunting (and vice versa). Regarding education an external district effect is not significant although there is a tendency that children in districts with higher education level are less likely to be stunted. The individual education effect is comparably strong and shows that a higher education status goes along with better nourished children.

5.3 Hierarchical spatial random effect

Figures 4 and 5 show results for the spatial random effects modeled through the level-2 and level-3 equations. The kernel density estimates of Fig. 4 provide insight into the strength and importance of the various random effects. We first note that the interaction random effects are much weaker than the main random effects. Moreover, the district smooth effects and the uncorrelated district random effects are roughly of equal size and dominate the state random

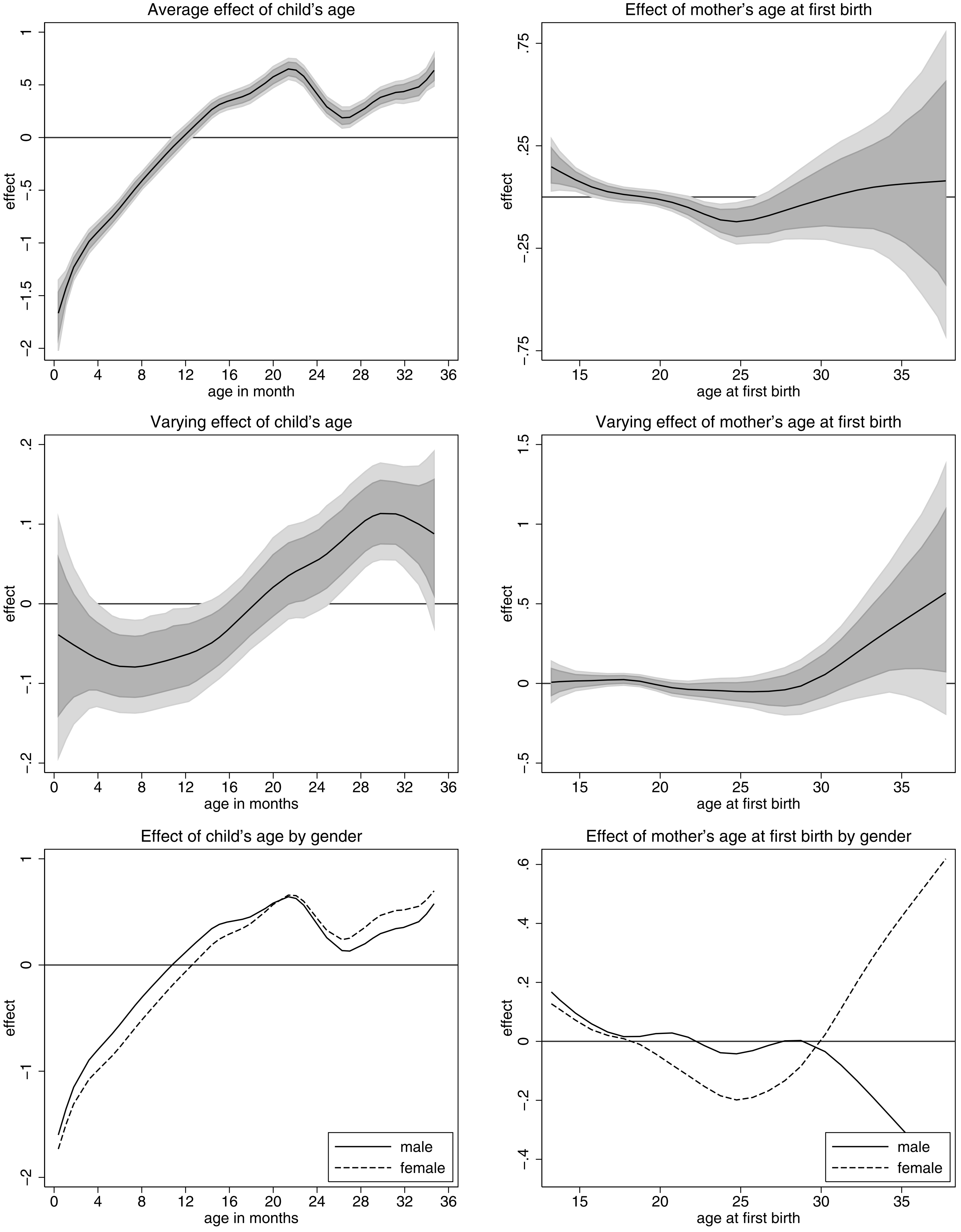

Fig. 2 Effect of child's age and mother's age at first birth by gender. Shown is the posterior mean together with 95 % pointwise and simultaneous credible intervals

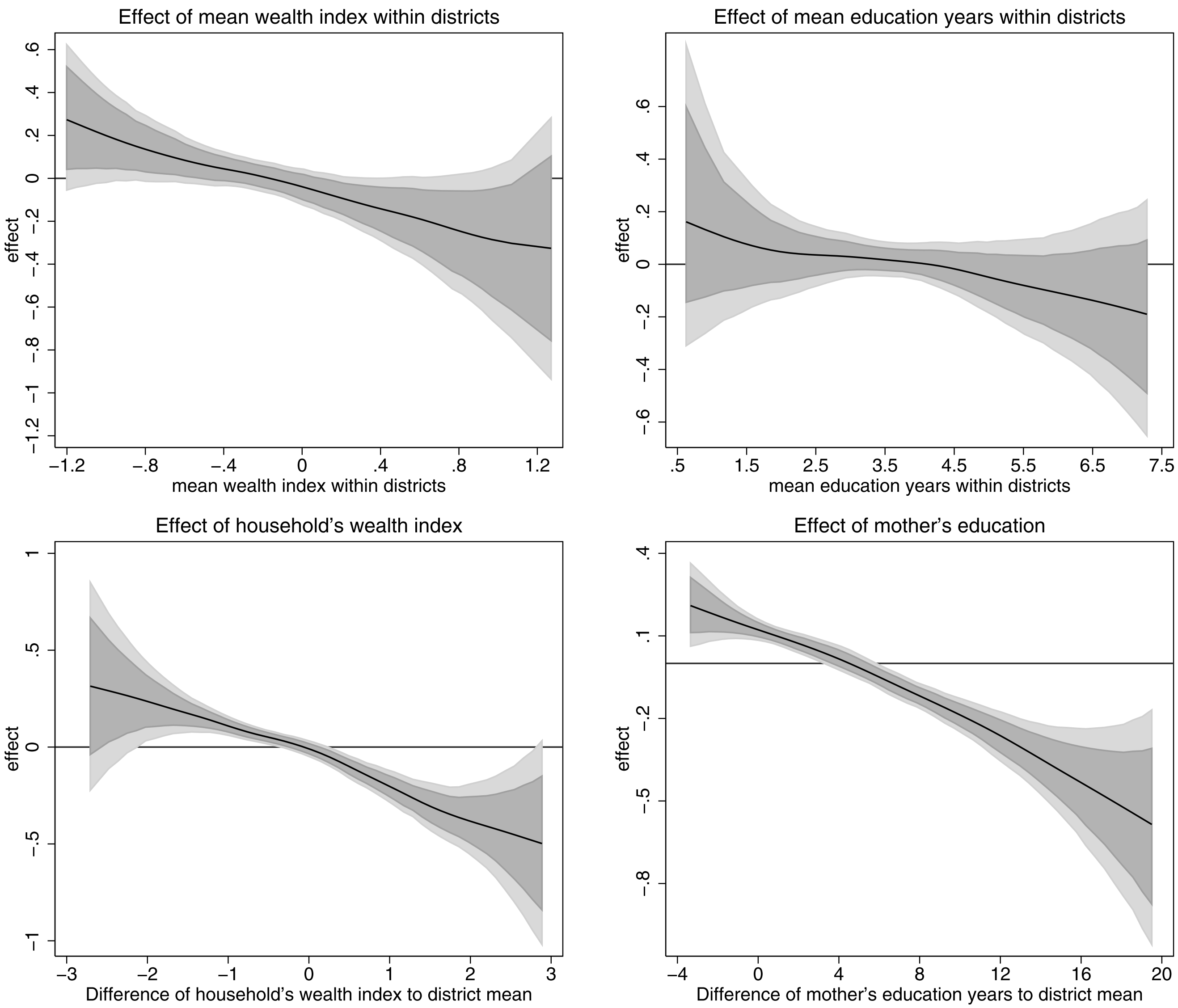

Fig. 3 Nonlinear effects of the wealth index and the education years. Shown is the posterior mean together with 95 % pointwise and simultaneous credible intervals

effects which are almost negligible. Figure 5 shows maps of the spatial heterogeneity not explained by covariates for males and females, respectively. Unexplained spatial heterogeneity is additively composed of the district smooth and uncorrelated random effect and the state random effect. Overall, unexplained heterogeneity is higher for females (see also in Fig. 4 the right bottom panel). Moreover, females exhibit a more pronounced spatial pattern with higher probabilities of stunting in the north-west and lower probabilities in the south and the north-east. For males we observe a similar pattern although the north-south patterns are less distinct.

5.4 Model choice

Some final remarks regarding model choice are in order. General tools for model choice are pointwise and simultaneous credible intervals for the nonlinear effects as well as Bayesian goodness of fit criteria, particularly the DIC. Also beneficial for model choice is the detailed hierarchical modeling of spatial heterogeneity. For instance, the kernel densities of Fig. 4 suggest that the interaction random effect can be restricted to a level-2 equation with a smooth and/or uncorrelated district effect. The spatial main effect could possibly be restricted to the level-2 equation omitting the level-3 states equation. To reduce the complexity of the full interaction model (6) we could in a first step exclude the smooth nonsignificant interactions (in terms of 95 % pointwise credible intervals) which slightly reduces the DIC by approximately 15 points. A further reduction of the DIC is obtained by more parsimonious random effects. The best model (in terms of the DIC) is given by a full main effects spatial ran-

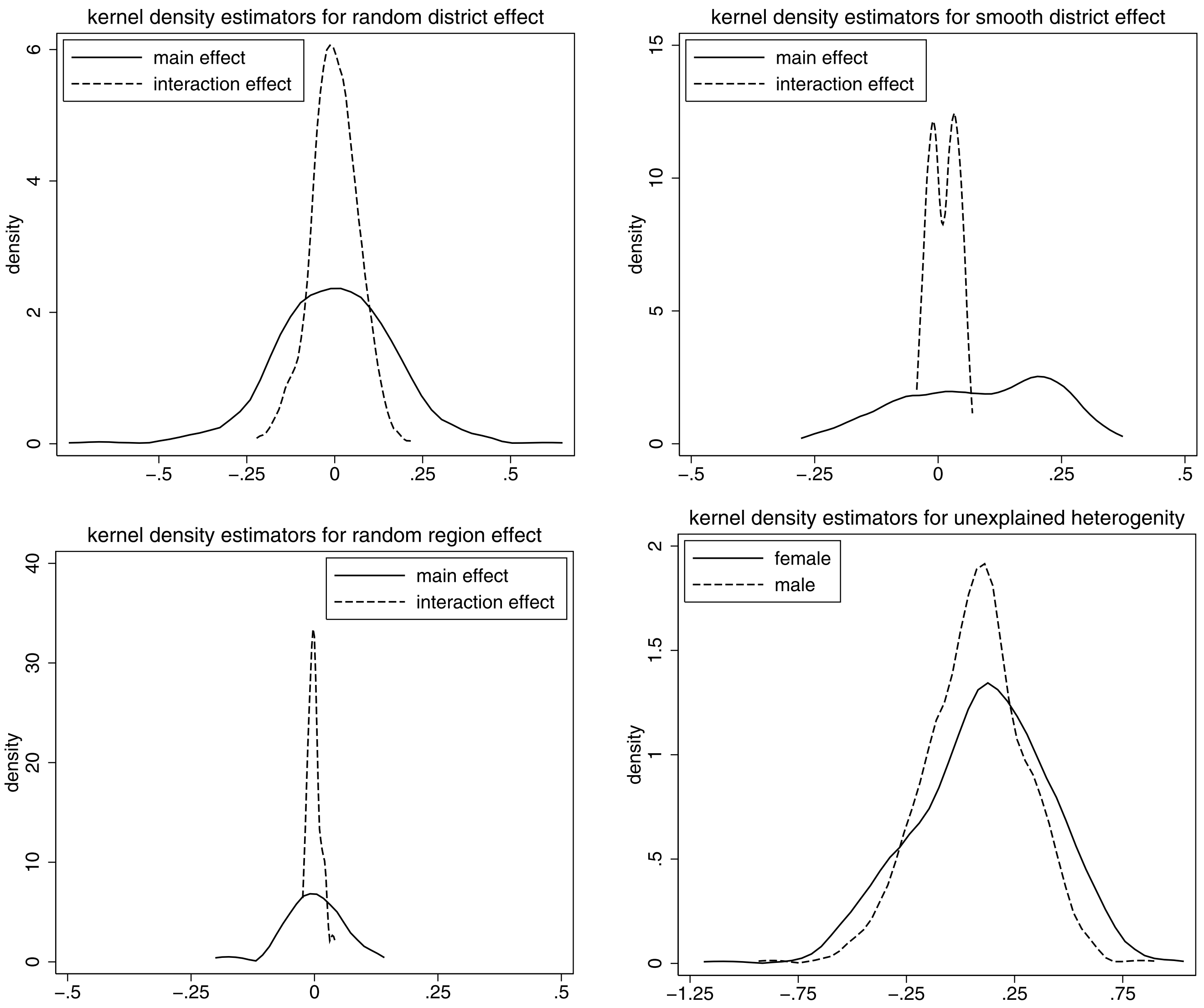

Fig. 4 Kernel densities of the spatial random effects

dom effect including the level-3 equation and a reduced spatial interaction with a simple i.i.d. Gaussian district random effect. In this model the DIC reduces by approximately 25 points compared to the full interaction model. Further reduction of the main effects spatial random effect to a level-2 equation shows almost identical DIC.

6 Conclusion

This paper proposes a multilevel version of STAR models by assuming that the regression coefficients of a particular nonlinear term obey another regression model with structured additive predictor. The proposed model may be regarded as an extended version of a multilevel model with nonlinear covariate terms in every level of the hierarchy. Our model framework also comprises proposals for generalizations of random slopes by assuming a common functional form that is scaled by cluster specific scaling factors. We have developed highly efficient MCMC schemes for simulation-based inference. The algorithms utilize the hierarchical structure of the models and rigorously exploit the sparsity of design matrices, cross products and penalty matrices. Thereby a considerable gain in numerical efficiency, reduction in computing time and improved mixing of Markov chains is achieved compared to non-hierarchical versions of the models.

The methodology of this paper is the basis for a number of extensions that we plan for future research:

- First of all, we plan to extend multilevel STAR models to *multivariate responses*, in particular multicategorical regression and seemingly unrelated regression.
- We also plan to model other parameters than the mean of the distribution in the spirit of generalized additive mod-

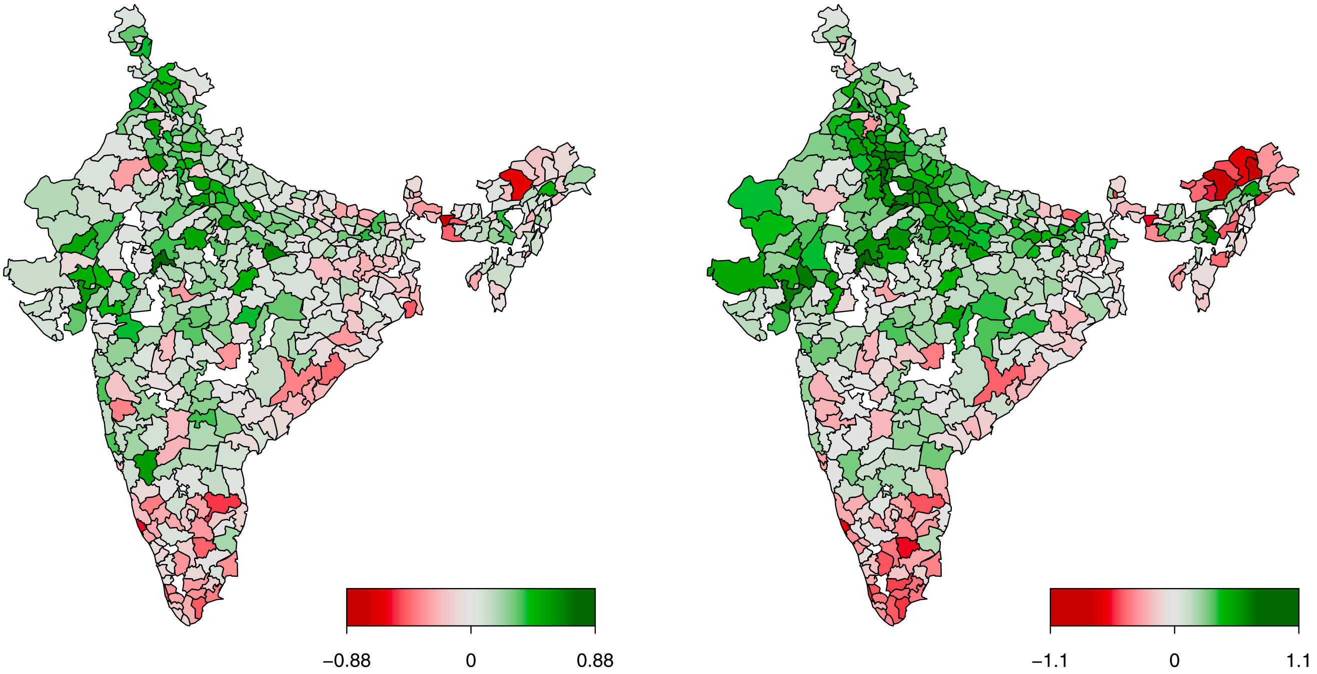

Fig. 5 Spatial heterogeneity not explained by covariates for males (*left panel*) and females (*right panel*)

els for location, scale and skewness (GAMLSS, Rigby and Stasinopoulos 2005).

- Another interesting (albeit rather challenging) field is to model hyperparameters in dependence of covariates, e.g. the variance parameter τ^2 in the general prior (8) or the weights in the penalty matrix of a Markov random field. Preferably, the specification of a full STAR model should be possible for these hyperparameters. This allows for modeling locally adaptive functions or complex covariate driven spatial neighborhood definitions.
- We finally want to develop methodology for automatic model choice and variable selection in the spirit of Belitz and Lang (2008) in a frequentist setting and Scheipl et al. (2012) in a Bayesian approach via spike and slab priors.

References

Albert, J., Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**, 669–679 (1993)

Belitz, C., Lang, S.: Simultaneous selection of variables and smoothing parameters in structured additive regression models. Comput. Stat. Data Anal. **53**, 61–81 (2008)

Brezger, A., Lang, S.: Generalized structured additive regression based on Bayesian P-splines. Comput. Stat. Data Anal. **50**, 967–991 (2006)

Chan, D., Kohn, R., Nott, D., Kirby, C.: Locally adaptive semiparametric estimation of the mean and variance functions in regression models. J. Comput. Graph. Stat. **15**, 915–936 (2006)

Cottet, R., Kohn, R., Nott, D.: Variable selection and model averaging in semiparametric overdispersed generalized linear models. J. Am. Stat. Assoc. **103**, 661–671 (2008)

Davis, T.A.: Direct Methods for Sparse Linear Systems. SIAM, Philadelphia (2006)

Eilers, P.H.C., Marx, B.D.: Flexible smoothing using B-splines and penalized likelihood. Stat. Sci. **11**, 89–121 (1996)

Fahrmeir, L., Kneib, T., Lang, S.: Penalized structured additive regression for space-time data: a Bayesian perspective. Stat. Sin. **14**, 731–761 (2004)

Fahrmeir, L., Lang, S.: Bayesian inference for generalized additive mixed models based on Markov random field priors. J. R. Stat. Soc., Ser. C, Appl. Stat. **50**, 201–220 (2001)

Frühwirth-Schnatter, S., Frühwirth, R.: Data augmentation and MCMC for binary and multinomial logit models. In: Kneib, T., Tutz, G. (eds.) Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir, pp. 111–132. Springer, Berlin (2010)

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., Rue, H.: Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. Stat. Comput. **19**, 479–492 (2009)

Frühwirth-Schnatter, S., Wagner, H.: Bayesian variable selection for random intercept modelling of Gaussian and non-Gaussian data. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (eds.) Bayesian Statistics, vol. 9, pp. 165–200. Oxford University Press, London (2011)

Gamerman, D., Moreira, A.R.B., Rue, H.: Space-varying regression models: Specifications and simulation. Comput. Stat. Data Anal. **42**, 513–533 (2003)

Gelfand, A.E.: Prior distributions for variance parameters in hierarchical models. Bayesian Anal. **1**, 515–534 (2006)

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

Heinzl, F., Kneib, T., Fahrmeir, L.: Additive mixed models with Dirichlet process mixture and P-spline priors. AStA Adv. Stat. Anal. **96**, 47–68 (2012)

Hennerfeind, A., Brezger, A., Fahrmeir, L.: Geoadditive survival models. J. Am. Stat. Assoc. **101**, 1065–1075 (2006)

Holmes, C.C., Held, L.: Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Anal. **1**, 145–168 (2006)

Jullion, A., Lambert, P.: Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. Comput. Stat. Data Anal. **51**, 2542–2558 (2007)

Kamman, E.E., Wand, M.P.: Geoadditive models. J. R. Stat. Soc., Ser. C, Appl. Stat. **52**, 1–18 (2003)

Klasen, S.: Nutrition, health, and mortality in Sub Saharan Africa: is there a gender bias? J. Dev. Stud. **32**, 913–933 (1996)

Krivobokova, T., Kneib, T., Claeskens, G.: Simultaneous confidence bands for penalized spline estimators. J. Am. Stat. Assoc. **105**, 852–863 (2010)

Lang, S., Brezger, A.: Bayesian P-splines. J. Comput. Graph. Stat. **13**, 183–212 (2004)

Lang, S., Steiner, W., Wechselberger, P.: Accommodating heterogeneity and functional flexibility in store sales models: a Bayesian semiparametric approach. Revised for Marketing Science (2012)

Panagiotelis, A., Smith, M.: Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. J. Econom. **143**, 291–316 (2008)

Papaspiliopoulos, O., Roberts, G.O., Sköld, M.: A general framework for the parametrization of hierarchical models. Stat. Sci. **22**, 59–73 (2007)

Park, T., Casella, G.: The Bayesian LASSO. J. Am. Stat. Assoc. **103**, 681–686 (2008)

Rigby, R.A., Stasinopoulos, D.M.: Generalized additive models for location, scale and shape. J. R. Stat. Soc., Ser. C, Appl. Stat. **54**, 507–554 (2005)

Rue, H., Held, L.: Gaussian Markov Random Fields. Chapman & Hall/CRC Press, London/CRC Press (2005)

Rue, H., Martino, S., Nicolas, C.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. B **71**, 319–392 (2009)

Ruppert, D.: Selecting the number of knots for penalized splines. J. Comput. Graph. Stat. **11**, 735–757 (2002)

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

Scheipl, F., Fahrmeir, L., Kneib, T.: Function selection in structured additive regression models based on spike-and-slab priors. J. Am. Stat. Assoc. (2012, to appear). doi:10.1080/01621459.2012.737742

Smith, M., Kohn, R.: Nonparametric regression using Bayesian variable selection. J. Econom. **75**, 317–343 (1996)

Smith, M., Kohn, R.: A Bayesian approach to nonparametric bivariate regression. J. Am. Stat. Assoc. **92**, 1522–1535 (1997)

Somerfelt, E., Arnold, F.: Sex differentials in the nutritional status of young children. In: United Nations (ed.) Too Young to Die, pp. 133–153. United Nations, New York (1999)

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit. J. R. Stat. Soc. B **65**, 583–639 (2002)

Subramanyam, M.A., Kawachi, I., Berkman, L.F., Subramanian, S.V.: Is economic growth associated with reduction in child undernutrition in India? PLoS Med. **8**, 1–15 (2011)

WHO: Global Database on Child Growth and Malnutrition. WHO, Department of Nutrition for Health and Development, Geneva (2002)

Wood, S.N.: Thin-plate regression splines. J. R. Stat. Soc. B **65**, 95–114 (2003)

Wood, S.N.: Generalized Additive Models: an Introduction with R. Chapman & Hall, London (2006)

Yue, Y., Speckman, P., Sun, D.: Priors for Bayesian adaptive spline smoothing. Ann. Inst. Stat. Math. **64**, 577–613 (2012)