BayesX

Software for Bayesian Inference in Structured Additive Regression Models

Version 3.0.2

Methodology Manual

Developed by

Christiane Belitz
Andreas Brezger
Nadja Klein (University of Göttingen)
Thomas Kneib (University of Göttingen)
Stefan Lang (University of Innsbruck)
Nikolaus Umlauf (University of Innsbruck)

With contributions by

Daniel Adler
Paul Cochrane
Jan Fahrenholz
Eva-Maria Fronk
Felix Heinzl
Andrea Hennerfeind
Manuela Hummel
Alexander Jerak
Susanne Konrath
Petra Kragler
Cornelia Oberhauser
Leyre Estíbaliz Osuna Echavarría
Daniel Sabanés Bové
Achim Zeileis

Supported by

Ludwig Fahrmeir (mentally)
Leo Held (mentally)
German Research Foundation (DFG)

Acknowledgements

The development of *BayesX* has been supported by grants from the German Research Foundation (DFG), Collaborative Research Center 386 "Statistical Analysis of Discrete Structures".

Special thanks go to (in alphabetical order of first names):

Dieter Gollnow for computing and providing the map of Munich (a really hard job);
Leo Held for advertising the program;
Ludwig Fahrmeir for his patience with finishing the program and for carefully reading and correcting the manual;
Ngianga-Bakwin Kandala for being the first user of the program (a really hard job);
Samson Babatunde Adebayo for carefully reading and correcting the manual;
Ursula Becker for carefully reading and correcting the manual;

Licensing agreement

BayesX is available at http://www.bayesx.org

1 Introduction

In this manual we provide a brief review of the methodological background for the four regression tools currently implemented in *BayesX*. The first two regression tools (*bayesreg objects* and *mcmcreg objects*) rely on Markov chain Monte Carlo (MCMC) simulation techniques and yields fully Bayesian posterior mean or posterior mode estimates. While *bayesreg objects* provide access to exponential family structured additive regression as well as survival times and multi-state models, *mcmcreg objects* implement distributional and quantile structured additive regression models as well as multilevel extensions of structure additive regression. The third regression tool (*remlreg objects*) is based on the mixed model representation of penalised regression models with inference being based on penalised maximum likelihood and marginal likelihood (a generalisation of restricted maximum likelihood) estimation. The fourth regression tool (*stepwisereg objects*) simultaneously performs model choice and estimation with inference being based on penalised likelihood. MCMC techniques are partly used for computing interval estimates. All regression tools allow to estimate structured additive regression (STAR) models (Belitz & Lang (2008), Brezger & Lang (2006), Fahrmeir, Kneib & Lang (2004)) with complex semiparametric predictors. STAR models cover a number of well known model classes as special cases, including *generalized additive models* (Hastie & Tibshirani 1990), *generalized additive mixed models* (Lin & Zhang 1999), *geoadditive models* (Kammann & Wand 2003), *varying coefficient models* (Hastie & Tibshirani 1993), and *geographically weighted regression* Fotheringham, Brunsdon & Charlton (2002). Besides models for responses from univariate exponential families, BayesX also supports non-standard regression situations such as models for categorical responses with either ordered or unordered categories, uni- and multivariate distributional regression in the spirit of generalised additive models for location, scale and shape with parametric response distributions beyond the exponential family framework, Bayesian quantile regression, continuous time survival data, or continuous time multi-state models. To provide a first impression of structured additive regression, Sections 2 to 6 describe STAR models for exponential family regression. Section 7 extends structured additive regression to the analysis of survival times and multi-state data while Section 8 considers extensions for distributional regression, quantile regression and multilevel specifications. Full details on STAR methodology can be found in the following references:

Structured additive regression based on MCMC simulation

- Brezger, A., Lang, S. (2006): Generalized Structured Additive Regression based on Bayesian P-Splines. *Computational Statistics and Data Analysis*, **50**, 967–991.
- Brezger, A., Lang, S. (2008) Simultaneous Probability Statements for Bayesian P-splines. *Statistical Modelling*, **8**, 141–168.
- Brezger, A., Steiner, W. (2008): Monotonic Regression based on Bayesian P-Splines: an Application to Estimating Price Response Functions from Store-level Scanner Data. *Journal of Economic and Business Statistics*, **26**, 90–104.
- Fahrmeir, L., Kneib, T., Lang, S., Marx, B. (2013): *Regression: Models, Methods and Applications*, New York: Springer-Verlag.
- Fahrmeir, L., Lang, S. (2001): Bayesian Inference for Generalized Additive Mixed Models based on Markov Random Field Priors. *Journal of the Royal Statistical Society C (Applied Statistics)*, **50**, 201–220.
- Fahrmeir, L., Lang, S. (2001): Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics*, **53**, 10–30.
- Hennerfeind, A., Brezger, A., Fahrmeir, L. (2006): Geoadditive Survival Models. *Journal of the American Statistical Association*, **101**, 1065–1075.

- Klein, N., Kneib, T., Lang, S. (2013): Bayesian Structured Additive Distributional Regression. Under revision for *Annals of Applied Statistics*.
- Klein, N., Kneib, T., Lang, S. (2014): Bayesian Generalized Additive Models for Location, Scale and Shape for Zero-Inflated and Overdispersed Count Data. To appear in *Journal of the American Statistical Association*, doi:10.1080/01621459.2014.912955.
- Klein, N., Kneib, T., Klasen, S., Lang, S.(2014): Bayesian Structured Additive Distributional Regression for Multivariate Responses. To appear in *Journal of the Royal Statistical Society C*, doi:10.1111/rssc.12090.
- Kneib, T., Hennerfeind, A. (2006) Bayesian Semiparametric Multi-State Models. *Statistical Modelling*, **8**, 169–198..
- Lang, S., Brezger, A. (2004): Bayesian P-Splines *Journal of Computational and Graphical Statistics*, **13**, 183–212.
- Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K. and Kneib, T. (2014): Multilevel Structured Additive Regression, *Statistics and Computing*, **24**, 223–238

Presumably the best starting point is the paper by Brezger & Lang (2006) or the monograph by Fahrmeir et al. (2013).

Structured additive regression based on mixed model methodology

- Fahrmeir, L., Kneib, T., Lang, S. (2004): Penalized Structured Additive Regression for Space-Time Data: a Bayesian Perspective. *Statistica Sinica*, **14**, 715–745.
- Kneib, T. (2006): Mixed Model based Inference in Structured Additive Regression. Dr. Hut Verlag, München. Available online from http://edoc.ub.uni-muenchen.de/archive/00005011/
- Kneib, T. (2006): Geoadditive Hazard Regression for Interval Censored Survival Times. *Computational Statistics and Data Analysis*, **51**, 777–792.
- Kneib, T., Fahrmeir, L. (2007): A Mixed Model Approach for Geoadditive Hazard Regression. *Scandinavian Journal of Statistics*, **34**, 207–228.
- Kneib, T., Fahrmeir, L. (2006): Structured Additive Regression for Multicategorical Space-Time Data: A Mixed Model Approach. *Biometrics*, **62**, 109–118.
- Kneib, T., Hennerfeind, A. (2006): Bayesian Semiparametric Multi-State Models. *Statistical Modelling*, **8**, 169–198.

Presumably the best starting point is the paper by Fahrmeir, Kneib & Lang (2004) or the monograph by Kneib (2006).

Structured additive regression including model selection

- Belitz, C. (2007): Model Selection in Generalised Structured Additive Regression Models. Dr. Hut Verlag, München.

- Belitz, C., Lang, S. (2008) Simultaneous Selection of Variables and Smoothing Parameters in Structured Additive Regression Models. *Computational Statistics and Data Analysis*, **53** , 61-81.

Presumably the best starting point is the paper by Belitz & Lang (2008).

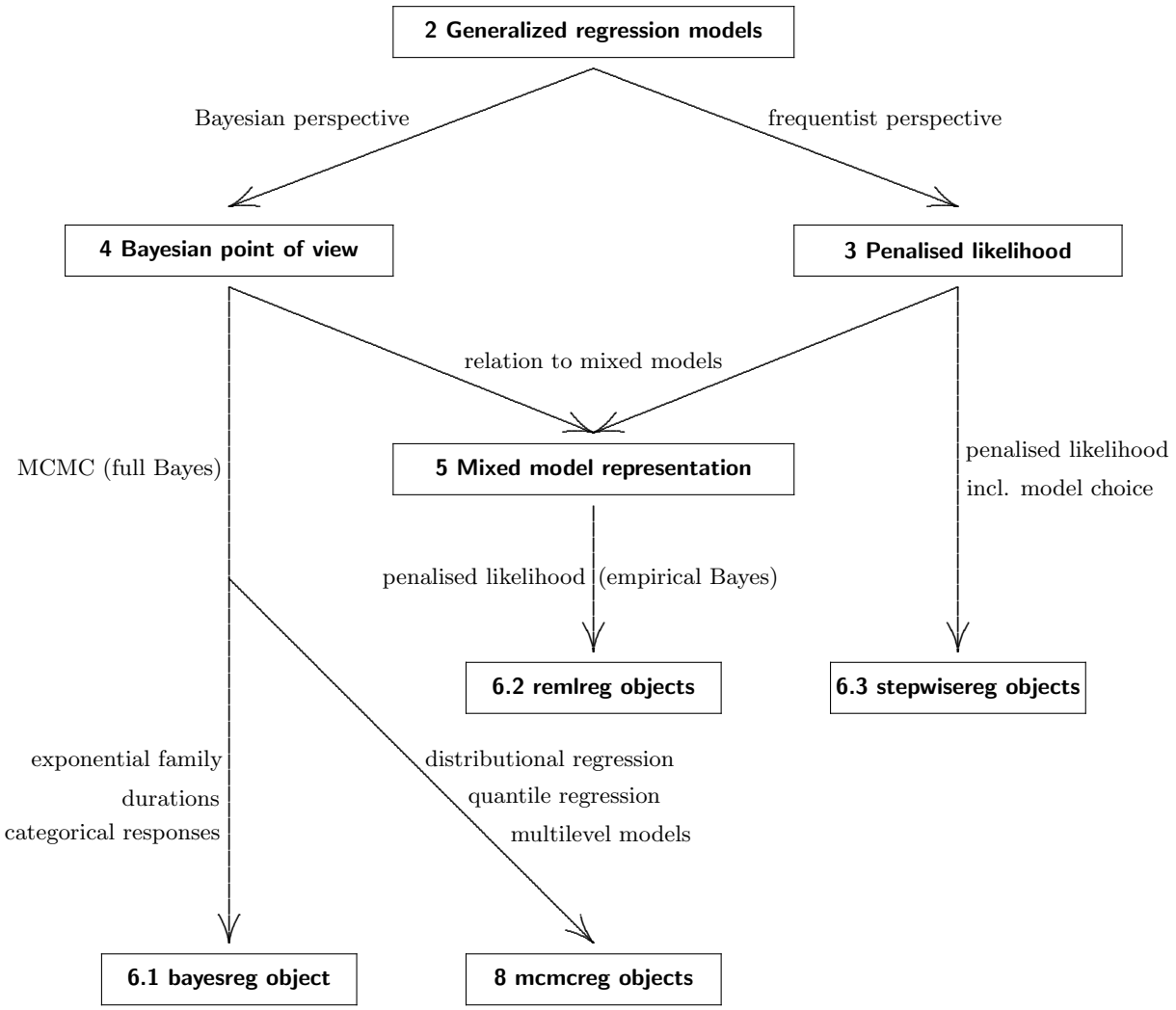


Figure 1: Guidelines for reading this manual.

Guideline for the reader

The rest of this manual is organized as follows:

The next section describes the general structure of STAR models for distributions of the response variable belonging to an exponential family. The following Sections 3 – 6 discuss alternative approaches for specifying and estimating the different model terms in STAR models. Section 3 describes the models from a more classical penalized least squares perspectives. A Bayesian point of view is taken in Section 4. The close connection to mixed models is highlighted in Section 5. Section 6 gives a brief outline of the various inference techniques for exponential family STAR models. Fully Bayesian inference via MCMC simulation techniques for exponential family responses, categorical responses and duration times is the topic of Section 6.1. Inference based on mixed model technology is sketched in Section 6.2. Simultaneous selection of relevant model terms and estimation of the parameters is described in Section 6.3. Section 7 discusses extensions for duration times and multi-state models while Section 8 provides details on distributional regression, quantile regression and multilevel specifications.

For most users of BayesX it is sufficient to read only parts of this manual. Some recommendations are given in Figure 1.

2 Generalized regression models

Generalized linear models assume that, given covariates $\mathbf{u}$ and unknown parameters $\boldsymbol{\gamma}$, the distribution of the response variable y belongs to an exponential family, i.e.

$$p(y \mid \mathbf{u}) = \exp\left(\frac{y\theta - b(\theta)}{\phi}\right) c(y, \phi) \tag{1}$$

where $b(\cdot)$, $c(\cdot)$, θ and ϕ determine the specific response distribution. A list of the most common distributions and their parameters can be found for example in Fahrmeir & Tutz (2001), page 21. The mean $\mu = E(y|\mathbf{u}, \boldsymbol{\gamma})$ is linked to a linear predictor η by

$$\mu = h(\eta) \qquad \eta = \mathbf{u}'\boldsymbol{\gamma}, \tag{2}$$

where h is a known response function and $\boldsymbol{\gamma}$ are unknown regression parameters.

In most practical regression situations, however, we are facing at least one of the following problems:

- For the *continuous covariates* in the data set, the assumption of a strictly linear effect on the predictor may be not appropriate.
- Observations may be *spatially correlated*.
- Observations may be *temporally correlated*.
- Complex interactions may be required to model the joint effect of some of the covariates adequately.
- Heterogeneity among individuals or units may be not sufficiently described by covariates. Hence, unobserved *unit or cluster specific heterogeneity* has to be considered appropriately.

To overcome these difficulties, we replace the strictly linear predictor in (2) by a structured additive predictor

$$\eta = f_1(x_1) + \ldots + f_j(x_j) + \ldots + f_p(x_p) + \mathbf{u}'\boldsymbol{\gamma}, \tag{3}$$

where x_j denote covariates of different type and dimension, and f_j are (not necessarily smooth) functions of the covariates. The functions f_j comprise usual nonlinear effects of continuous covariates, time trends and seasonal effects, two-dimensional surfaces, varying coefficient models, i.i.d. random intercepts and slopes as well as spatial effects. STAR-models cover a number of special cases well known from the literature, in particular *Generalized additive models (GAM)*, *Generalized additive mixed models (GAM)*, *Geoadditive models*, *Multilevel models*, *Varying coefficient models (VCM)*, *ANOVA type interaction models* and *geographically weighted regression*.

3 Penalized least squares

In BayesX, the nonlinear functions f_j are modeled by a basis functions approach, i.e. a particular nonlinear function f is approximated by a linear combination of basis functions:

$$f(x) = \sum_{k=1}^{K} \beta_k B_k(x)$$

The B_k are known basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ is a vector of unknown regression coefficients to be estimated. Note that the term basis function in our understanding is not limited to basis functions known from nonparametric smoothing such as B-splines but also refers to non-standard basis functions such as indicator functions for regions or clusters. To ensure enough flexibility, typically a large number of basis functions is defined. To avoid overfitting, a roughness penalty on the regression coefficients is additionally specified. We use quadratic penalties of the form $\boldsymbol{\beta}'\mathbf{P}(\boldsymbol{\lambda})\boldsymbol{\beta}$ where $\mathbf{P}(\boldsymbol{\lambda})$ is a penalty matrix. The penalty depends on one or multiple smoothing parameters $\boldsymbol{\lambda}$ that govern the amount of smoothness imposed on the function f. Most penalty matrices are of the particular simple form $\mathbf{P}(\boldsymbol{\lambda}) = \lambda\mathbf{K}$ where λ is a scalar smoothing parameter. For *stepwisereg objects* more complicated penalties are sometimes possible. They are an additive combination of penalty matrices. An example is $\mathbf{P}(\boldsymbol{\lambda}) = \lambda_1\mathbf{K}_1 + \lambda_2\mathbf{K}_2$ where λ_1 and λ_2 are smoothing parameters and $\mathbf{K}_1$ and $\mathbf{K}_2$ are penalty matrices.

The choice of basis functions $B_1, \ldots, B_K$ and penalty $\mathbf{P}(\boldsymbol{\lambda})$ depends on our prior assumptions about the smoothness of f as well as the type and dimension of x. We will give specific examples below. Defining the $n \times K$ design matrix $\mathbf{X}$ with elements $X[i,k] = B_k(x_i)$ the vector $\mathbf{f} = (f(x_1), \ldots, f(x_n))'$ of function evaluations can be written in matrix notation as $\mathbf{f} = \mathbf{X}\boldsymbol{\beta}$. Accordingly, for model (3) we obtain

$$\boldsymbol{\eta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\mathbf{U}$ is the design matrix for linear effects, $\boldsymbol{\gamma}$ is the vector of regression coefficients for linear effects, and $\boldsymbol{\varepsilon}$ are the vectors of observations and errors. In the next subsections we will give specific examples for modeling the unknown functions f_j or in other words for the choice of basis functions and penalty matrices. We start with modeling the effect of continuous covariates using splines.

3.1 Continuous covariates

3.1.1 P(enalized)-splines

Suppose first that a particular component x of the covariate is univariate and continuous. There is a considerable amount of literature on basis functions approaches in combination with a (quadratic) roughness penalty for continuous covariates. BayesX applies the P-splines approach introduced by Eilers & Marx (1996). The approach assumes that the unknown functions can be approximated by a polynomial spline of degree l and with equally spaced knots

$$x_{min} = \zeta_0 < \zeta_1 < \cdots < \zeta_{m-1} < \zeta_m = x_{max}$$

over the domain of x. The spline can be written in terms of a linear combination of $K = m + l$ B-spline basis functions. The columns of the design matrix $\mathbf{X}$ are given by the B-spline basis functions evaluated at the observations x_i. To overcome the well known difficulties involved with regression splines, Eilers & 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 based on squared r-th order differences, i.e.

$$\boldsymbol{\beta}'\lambda\mathbf{K}\boldsymbol{\beta} = \lambda \sum_{k=r+1}^{K} (\Delta^r \beta_k)^2.$$

The penalty matrix is given by $\mathbf{K} = \mathbf{D}_r'\mathbf{D}_r$ where $\mathbf{D}_r$ is a r-th order difference matrix. Typically, second or third order differences are used. The limiting behavior $\lambda \to \infty$ depends both on the order of the spline and the order of the penalty. If the order of the spline is equal to or higher than the order of the penalty, which is typically the case, then a polynomial fit of degree $r - 1$ is obtained in the limit.

The approach can be extended to impose monotonicity or more general shape constraints. We follow an approach proposed by Bollaerts, Eilers & van Mechelen (2006). A sufficient condition for a decreasing spline is given by $\beta_k \leq \beta_{k-1}$, i.e. a parameter β_k is less than its predecessor β_{k-1}. The simple but powerful idea is to impose the required constraint by expanding the penalty by an additional term. More specifically they propose the composed penalty

$$\mathbf{P}(\boldsymbol{\lambda}) = \boldsymbol{\beta}' \left(\lambda_1 \mathbf{K}_1 + \lambda_2 \mathbf{K}_2\right) \boldsymbol{\beta}$$

where λ_1 and $\mathbf{K}_1$ are the usual smoothing parameter and penalty matrix for P-splines. The additional penalty matrix $\mathbf{K}_2$ is a diagonal matrix with entries 1 whenever the condition $\beta_k \leq \beta_{k-1}$ fails and 0 otherwise. For increasing functions, $\mathbf{K}_2$ has to be adapted accordingly. The parameter λ_2 is not estimated but set large enough to enforce monotonic functions.

3.1.2 Tensor product P-splines

Assume now that $\mathbf{x}$ is two-dimensional, i.e. $\mathbf{x} = \left(x^{(1)}, x^{(2)}\right)'$ with continuous components $x^{(1)}$ and $x^{(2)}$. The aim is to extend the univariate P-spline from the preceding section to two dimensions. A common approach is to approximate the unknown surface $f(x)$ by the tensor product of one dimensional B-splines, i.e.

$$f\left(x^{(1)}, x^{(2)}\right) = \sum_{k=1}^{K_1} \sum_{s=1}^{K_2} \beta_{ks} B_{1,k}(x^{(1)}) B_{2,s}(x^{(2)}), \tag{4}$$

where $B_{11}, \ldots, B_{1K_1}$ are the basis functions in $x^{(1)}$ direction and $B_{21}, \ldots, B_{2K_2}$ in $x^{(2)}$ direction. The $n \times K = n \times K_1 K_2$ design matrix $\mathbf{X}$ now consists of products of basis functions.

Several alternatives are available for the penalty matrix $\mathbf{P}(\boldsymbol{\lambda})$:

a) *Penalty based on first differences:* The two-dimensional generalization of a penalty based on first differences is given by combining row- and column wise quadratic differences

$$\sum_{k=2}^{K_1} \sum_{s=1}^{K_2} (\beta_{ks} - \beta_{k-1,s})^2 = \boldsymbol{\beta}'(\mathbf{I}_{K_2} \otimes \mathbf{D}_1)'(\mathbf{I}_{K_2} \otimes \mathbf{D}_1)\boldsymbol{\beta}$$
$$\sum_{k=1}^{K_1} \sum_{s=2}^{K_2} (\beta_{ks} - \beta_{k,s-1})^2 = \boldsymbol{\beta}'(\mathbf{D}_2 \otimes \mathbf{I}_{K_1})'(\mathbf{D}_2 \otimes \mathbf{I}_{K_1})\boldsymbol{\beta}$$

to the penalty

$$\boldsymbol{\beta}'\mathbf{P}(\boldsymbol{\lambda})\boldsymbol{\beta} = \boldsymbol{\beta}'\lambda \left[(\mathbf{I}_{K_2} \otimes \mathbf{D}_1)'(\mathbf{I}_{K_2} \otimes \mathbf{D}_1) + (\mathbf{D}_2 \otimes \mathbf{I}_{K_1})'(\mathbf{D}_2 \otimes \mathbf{I}_{K_1})\right] \boldsymbol{\beta}.$$

Another way of expressing the penalty is given by

$$\boldsymbol{\beta}'\mathbf{P}(\lambda)\boldsymbol{\beta} = \boldsymbol{\beta}'\lambda \left[\mathbf{I}_{K_2} \otimes \mathbf{K}_1 + \mathbf{K}_2 \otimes \mathbf{I}_{K_1}\right] \boldsymbol{\beta}, \tag{5}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are the respective one dimensional penalty matrices. In the limit $\lambda \to \infty$ a constant fit is obtained.

b) *Penalty based on second differences:* In a similar way two-dimensional penalties based on higher order differences are constructed. A second order difference penalty is obtained if $\mathbf{K}_1$ and $\mathbf{K}_2$ in (5) correspond to penalty matrices based on second rather than first differences. Similar to one dimensional P-splines the limit $\lambda \to \infty$ results in linear effects in $x^{(1)}$ and $x^{(2)}$ with an additional interaction effect, i.e.

$$f\left(z^{(1)}, z^{(2)}\right) = c_0 + c_1\, x^{(1)} + c_2\, x^{(2)} + c_3\, x^{(1)} x^{(2)}.$$

c) *Anisotropic penalty:* The two penalties considered so far are not capable of different penalization in $x^{(1)}$ and $x^{(2)}$ direction, respectively. Anisotropic penalties are obtained by assuming separate smoothing parameters λ_1 and λ_2 in $x^{(1)}$ and $x^{(2)}$ direction. The penalty is then given by

$$\boldsymbol{\beta}'\mathbf{P}(\boldsymbol{\lambda})\boldsymbol{\beta} = \boldsymbol{\beta}'\left[\lambda_1\mathbf{I}_{K_2}\otimes\mathbf{K}_1 + \lambda_2\mathbf{K}_2\otimes\mathbf{I}_{K_1}\right]\boldsymbol{\beta}. \tag{6}$$

The resulting fit in the limit $\lambda_1 \to \infty$ and $\lambda_2 \to \infty$ depends on the penalty used to construct $\mathbf{K}_1$ and $\mathbf{K}_2$. If $\mathbf{K}_1$ and $\mathbf{K}_2$ correspond to a first order difference penalty a constant fit is obtained in the limit. Second order difference penalties result in a linear fit for $f\left(x^{(1)}, x^{(2)}\right)$.

d) *Penalties with main effects in the limit:* Sometimes it is desirable to decompose the effect of the two covariates $x^{(1)}$ and $x^{(2)}$ into two main effects modeled by one dimensional functions and a two-dimensional interaction effect, i.e.

$$f\left(x^{(1)}, x^{(2)}\right) = f_1\left(x^{(1)}\right) + f_2\left(x^{(2)}\right) + f_{1|2}\left(x^{(1)}, x^{(2)}\right). \tag{7}$$

Usually a two-dimensional surface smoother together with two additional one dimensional P-splines (or other smoothers) are estimated. This approach is possible with *bayesreg objects* and *remlreg objects*. *stepwisereg objects* take, however, a different approach. We specify a two-dimensional surface based on tensor product P-splines and compute the decomposition of the resulting surface into main effects and the interaction effect *after* estimation. Moreover, we specify a penalty that allows for a main effects only model as a special case. This allows to discriminate between a simple main effects model and a more complicated two way interactions model. A penalty that guarantees a main effects model in the limit is defined by the Kronecker product of the two penalty matrices for one dimensional P-splines, i.e.

$$\boldsymbol{\beta}'\mathbf{P}(\lambda)\boldsymbol{\beta} = \boldsymbol{\beta}'\lambda\mathbf{K}_1\otimes\mathbf{K}_2\boldsymbol{\beta}. \tag{8}$$

The drawback of this penalty is that the limit $\lambda \to \infty$ yields *unpenalized* main effects, i.e. wiggly functions. We therefore use a modified penalty which is effectively a combination of the two penalties (6) and (8). More specifically we define

$$\boldsymbol{\beta}'\mathbf{P}(\boldsymbol{\lambda})\boldsymbol{\beta} = \boldsymbol{\beta}'\left[\frac{\lambda_1}{K_1}\mathbf{I}_{K_2}\otimes\mathbf{K}_1 + \frac{\lambda_2}{K_2}\mathbf{K}_2\otimes\mathbf{I}_{K_1} + \lambda_3\mathbf{K}_1\otimes\mathbf{K}_2\right]\boldsymbol{\beta}, \tag{9}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are penalty matrices corresponding to one dimensional P-splines based on first or second order differences. This penalty has the following nice properties:

- The limit $\lambda_3 \to \infty$ results in a mere main effects model. The main effects are one dimensional P-splines with smoothing parameters λ_1 and λ_2.
- The limit $\lambda_3 \to 0$ yields the anisotropic penalty (6) as a special case.
- The limit $\lambda_1 \to 0$ and $\lambda_2 \to 0$ yields the Kronecker product penalty (8) as a special case.
- The limit $\lambda_1 \to \infty$, $\lambda_2 \to \infty$ and $\lambda_3 \to \infty$ results in a main effects model with linear or constant main effects depending on the difference order used to construct $\mathbf{K}_1$ and $\mathbf{K}_2$.

3.2 Spatial effects

In this subsection we assume that x represents the location a particular observation pertains to. The location is typically given in two ways. If exact locations are available $x = (x^{(1)}, x^{(2)})'$ is two-dimensional and the components $x^{(1)}$ and $x^{(2)}$ correspond to the coordinates of the location. In this case, the spatial effect $f(x^{(1)}, x^{(2)})$ could be modeled by two-dimensional surface estimators as described in the preceding section.

In many applications, however, exact locations are not available. Typically, a geographical map is available and $x \in \{1, \ldots, K\}$ is an index that denotes the region (e.g. district) an observation pertains to. A common approach is to assume $f(x) = \beta_x$, i.e. separate parameters $\beta_1, \ldots, \beta_K$ for each region are estimated. The $n \times K$ design matrix $\mathbf{X}$ is an incidence matrix whose entry in the i-th row and k-th column is equal to one if observation i has been observed at location k and zero otherwise. To prevent overfitting a penalty based on squared differences is defined that guarantees that parameters of neighboring regions are similar. Typically two regions are assumed to be neighbors if they share a common boundary although other neighborhood definitions are possible. The penalty is defined as

$$\boldsymbol{\beta}' \lambda \mathbf{K} \boldsymbol{\beta} = \lambda \sum_{k=2}^{K} \sum_{s \in N(k), s<k} (\beta_k - \beta_s)^2,$$

where $N(k)$ denotes all sites that are neighbors of site k. The elements of the penalty matrix are given by

$$\mathbf{K}[s,r] = \lambda \begin{cases} -1 & k \neq s, k \sim s, \\ 0 & k \neq s, k \nsim s, \\ |N(k)| & k = s. \end{cases} \tag{10}$$

Depending on the prior belief on smoothness of the spatial effect several alternatives to penalty (10) are available. If a very smooth effect is assumed, the two-dimensional smoothers discussed in the preceding section could be used as an alternative. Since exact locations are not available the centroids of the regions could be used instead.

3.3 Unit- or cluster specific heterogeneity

Typically, unit- or cluster specific random effects are introduced to account for unobserved heterogeneity. In its simplest form, a random intercept β_x with $\beta_x \sim N(0, \tau^2)$ is introduced. Here, $x \in \{1, \ldots, K\}$ is an index variable that denotes the cluster a particular observation pertains to. This is equivalent to a penalized least squares approach with function $f(x) = \beta_x$, penalty matrix $\mathbf{I}$ and smoothing parameter $\lambda = \sigma^2/\tau^2$. The $n \times K$ design matrix $\mathbf{X}$ is a 0/1 incidence matrix whose entry in the i-th row and k-th column is equal to one if observation i belongs to the k-th cluster and zero otherwise. Random slopes could be treated in the same way, see the next subsection.

A particular cluster variable is a spatial index that indicates the region an observation pertains to. Usually a spatially correlated effect as described in the preceding subsection is specified. However, in some situations a smooth spatial effect is not justified because of local, spatial heterogeneity. In this case, the assumption of spatial dependence of neighboring parameters is not meaningful. Instead, the simple (ridge type) penalty

$$\boldsymbol{\beta}' \lambda \mathbf{K} \boldsymbol{\beta} = \lambda \boldsymbol{\beta}' \boldsymbol{\beta} = \lambda \sum_{k=1}^{K} \beta_k^2$$

with penalty matrix $\mathbf{K} = \mathbf{I}$ may be defined. This penalty does not assume any spatial dependence but prevents highly variable estimates induced by small samples for some regions or sites.

Note that more than one random intercept with respect to different cluster variables are possible. In many cases there exists a hierarchical ordering of clusters. Models with such hierarchical clusters are also called multilevel models.

3.4 Varying coefficients

Suppose now that the effect of a continuous covariate $x^{(2)}$ is assumed to vary with respect to a categorical covariate $x^{(1)}$. For notational convenience, we restrict the discussion to binary covariates $x^{(1)}$. The generalization to (multi)categorical covariates is straightforward. The interaction between $x^{(2)}$ and $x^{(1)}$ can be modeled by a predictor of the form

$$\eta = \ldots + f_1(x^{(2)}) + g(x^{(2)})x^{(1)} + \ldots,$$

where f_1 and g are smooth functions (modeled by P-splines). The interpretation of the two functions f_1 and g depends on the coding of the binary variable $x^{(1)}$. If $x^{(1)}$ is in dummy-coding, the function f_1 corresponds to the effect of $x^{(2)}$ subject to $x^{(1)} = 0$, and g is the difference effect for observations with $x^{(1)} = 1$. If $x^{(1)}$ is in effect-coding, the function f_1 can be interpreted as an average effect of $x^{(2)}$, where g and $-g$ represent the deviation from f_1 for $x^{(1)} = 1$ and $x^{(1)} = -1$, respectively. It turns out that the coding of $x^{(2)}$ is not only important for interpretation but sometimes also crucial for inference (in *bayesreg objects* and *stepwisereg objects*). Estimation for *bayesreg and stepwisereg objects* described in the next section is based on an iterative backfitting type procedure. Hence dependence between f_1 and g should be minimized to avoid convergence problems. Hence, effect coding for $x^{(2)}$ is an effective yet simple device to avoid convergence problems.

Models with interaction effects of the form $g(x^{(2)})\,x^{(1)}$ are known as varying coefficient models because the effect of $x^{(1)}$ varies smoothly with respect to the continuous covariate $x^{(2)}$. Covariate $x^{(2)}$ is called the effect modifier of $x^{(1)}$. The approach can be easily extended to a two-dimensional effect modifier with components $x^{(2)}$ and $x^{(3)}$. The interaction effect is then given by $g(x^{(2)}, x^{(3)})\,x^{(1)}$ where $g(x^{(2)}, x^{(3)})$ is a two-dimensional surface which is modeled by the tensor product P-splines discussed in section 3.1.2. Another modification arises if the effect modifier is the location either given as the coordinates or as a spatial index. In this case we have a space varying effect of $x^{(1)}$. Models of this kind are also known as geographically weighted regression, see Fotheringham, Brunsdon & Charlton (2002). A final modification is obtained for a unit or cluster index as effect modifier. The effect of $x^{(1)}$ is now assumed to be unit- or cluster-specific and typically referred to as a random slope.

Independent of the specific type of the effect modifier, the interaction term $g\left(x^{(2)}\right)\,x^{(1)}$ (or $g\left(x^{(2)}, x^{(3)}\right)\,x^{(1)}$) can be cast into our general framework by defining

$$f\left(x^{(1)}, x^{(2)}\right) = g\left(x^{(2)}\right)\,x^{(1)} \quad \text{or} \quad f\left(x^{(1)}, x^{(2)}, x^{(3)}\right) = g\left(x^{(2)}, x^{(3)}\right)\,x^{(1)}. \tag{11}$$

The overall design matrix $\mathbf{X}$ is given by $\text{diag}(x_1^{(1)}, \ldots, x_n^{(1)})\,\mathbf{X}^{(1)}$ where $\mathbf{X}^{(1)}$ is the usual design matrix for P-Splines, tensor product P-splines, spatial-, or cluster-specific effects.

4 Bayesian point of view

For Bayesian inference, the unknown functions $f_1, \ldots, f_p$ in predictor (3), more exactly corresponding vectors of function evaluations, and the fixed effects parameters $\boldsymbol{\gamma}$ are considered as random variables and must be supplemented by appropriate prior assumptions.

In the absence of any prior knowledge, diffuse priors are the appropriate choice for fixed effects parameters, i.e.

$$p(\gamma_j) \propto const$$

Another common choice, not yet supported by *BayesX*, are informative multivariate Gaussian priors with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$.

Priors for the unknown functions $f_1, \ldots, f_p$ depend on the *type of the covariates* and on *prior beliefs about the smoothness of* f_j. In the following we express the vector of function evaluations $\mathbf{f}_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))'$ of a function f_j as the matrix product of a design matrix $\mathbf{X}_j$ and a vector of unknown parameters $\boldsymbol{\beta}_j$, i.e.

$$\mathbf{f}_j = \mathbf{X}_j \boldsymbol{\beta}_j. \tag{12}$$

Then, we obtain the predictor (3) in matrix notation as

$$\boldsymbol{\eta} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \cdots + \mathbf{X}_p \boldsymbol{\beta}_p + \mathbf{U}\boldsymbol{\gamma}, \tag{13}$$

where $\mathbf{U}$ corresponds to the usual design matrix for fixed effects.

A prior for a function f_j is defined by specifying a suitable design matrix $\mathbf{X}_j$ and a prior distribution for the vector $\boldsymbol{\beta}_j$ of unknown parameters. The general form of the prior for $\boldsymbol{\beta}_j$ is given by

$$p(\boldsymbol{\beta}_j | \tau_j^2) \propto \frac{1}{(\tau_j^2)^{rank(\mathbf{K}_j)/2}} \exp\left(-\frac{1}{2\tau_j^2} \boldsymbol{\beta}_j' \mathbf{K}_j \boldsymbol{\beta}_j \right), \tag{14}$$

where $\mathbf{K}_j$ is a *penalty matrix*. In most cases $\mathbf{K}_j$ will be rank deficient and therefore the prior for $\boldsymbol{\beta}_j$ is partially improper.

The variance parameter τ_j^2 is equivalent to the inverse smoothing parameter in a penalized likelihood approach and controls the trade off between flexibility and smoothness.

In the following we will describe specific priors for different types of covariates and functions f_j.

4.1 Continuous covariates

Several alternatives have been proposed for specifying smoothness priors for continuous covariates or time scales. These are *random walk priors* or more generally *autoregressive priors* (see Fahrmeir & Lang (2001) and Fahrmeir & Lang (2001), *Bayesian P-splines* (Lang & Brezger 2004) and *Bayesian smoothing splines* (Hastie & Tibshirani 2000). *BayesX* supports random walk priors and P-splines.

4.1.1 Random walks

Suppose first that x is a time scale or continuous covariate with equally spaced ordered observations

$$x^{(1)} < x^{(2)} < \cdots < x^{(K)}.$$

Here $K \leq n$ denotes the number of *different* observed values for x in the data set. A common approach in dynamic or state space models is to estimate one parameter β_k for each distinct $x^{(k)}$, i.e. $f(x^{(k)}) = \beta_k$, and penalize too abrupt jumps between successive parameters using random walk priors. For example, first and second order random walk models are given by

$$\beta_k = \beta_{k-1} + u_k \quad \text{and} \quad \beta_k = 2\beta_{k-1} - \beta_{k-2} + u_k \tag{15}$$

with Gaussian errors $u_k \sim N(0, \tau^2)$ and diffuse priors $p(\beta_1) \propto const$, and $p(\beta_1)$ and $p(\beta_2) \propto const$, for initial values, respectively. Both specifications act as smoothness priors that penalize too rough functions f. A first order random walk penalizes abrupt jumps $\beta_k - \beta_{k-1}$ between successive states while a second order random walk penalizes deviations from the linear trend $2\beta_{k-1} - \beta_{k-2}$. The joint distribution of the regression parameters $\boldsymbol{\beta}$ is easily computed as the product of conditional densities defined by (15) and can be brought into the general form (14). The penalty matrix is of

the form $\mathbf{K} = \mathbf{D}'\mathbf{D}$ where $\mathbf{D}$ is a first or second order difference matrix. For example, for a random walk of first order the penalty matrix is given by:

$$\mathbf{K} = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

The design matrix $\mathbf{X}$ is a simple 0/1 matrix where the number of columns equals the number of parameters, i.e. the number of distinct covariate values. If for the i-th observation $x_i = x^{(k)}$ the element in the i-th row and k-th column of $\mathbf{X}$ is one and zero otherwise.

In case of non-equally spaced observations slight modifications of the priors defined in (15) are necessary, see Fahrmeir & Lang (2001) for details.

4.1.2 P-splines

A second approach for effects of continuous covariates, that is closely related to random walk models, is based on P-splines introduced by Eilers & Marx (1996). The approach assumes that an unknown smooth function f of a covariate x can be approximated by a polynomial spline of degree l defined by a set of equally spaced knots $x_{min} = \zeta_0 < \zeta_1 < \cdots < \zeta_{m-1} < \zeta_m = x_{max}$ within the domain of x. Such a spline can be written in terms of a linear combination of $K = m + l$ B-spline basis functions B_k, i.e.

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

In this case, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ corresponds to the vector of unknown regression coefficients and the $n \times K$ design matrix $\mathbf{X}$ consists of the basis functions evaluated at the observations x_i, i.e. $\mathbf{X}[i,k] = B_k(x_i)$. The crucial point is the choice of the number of knots. For a small number of knots, the resulting spline may not be flexible enough to capture the variability of the data. For a large number of knots, estimated curves tend to overfit the data and, as a result, too rough functions are obtained. As a remedy, Eilers & Marx (1996) suggest a moderately large number of equally spaced knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on first or second order differences of adjacent B-Spline coefficients to guarantee sufficient smoothness of the fitted curves. This leads to penalized likelihood estimation with penalty terms

$$\lambda \sum_{k=r+1}^{K} (\Delta^r \beta_k)^2, \quad r = 1, 2 \tag{16}$$

where λ is the smoothing parameter. First order differences penalize abrupt jumps $\beta_k - \beta_{k-1}$ between successive parameters while second order differences penalize deviations from the linear trend $2\beta_{k-1} - \beta_{k-2}$. In a Bayesian approach we use the stochastic analogue of difference penalties, i.e. first or second order random walks, as priors for the regression coefficients. Note that simple first or second order random walks can be regarded as P-splines of degree $l = 0$ and are therefore included as a special case. More details about Bayesian P-splines can be found in Lang & Brezger (2004) and Brezger & Lang (2006).

4.1.3 Tensor product P-splines

Assume now that x is two-dimensional, i.e. $x = \left(x^{(1)}, x^{(2)}\right)'$ with continuous components $x^{(1)}$ and $x^{(2)}$. Similar to section 3.1.2 the aim is to extend the univariate P-spline to two dimensions.

In *BayesX* Bayesian surface fitting is based on two-dimensional P-splines described in more detail in Lang & Brezger (2004) and Brezger & Lang (2006). The unknown surface $f(x^{(1)}, x^{(2)})$ is approximated by the tensor product of two one-dimensional B-splines, i.e.

$$f\left(x^{(1)}, x^{(2)}\right) = \sum_{k=1}^{K_1} \sum_{s=1}^{K_2} \beta_{ks} B_{1,k}(x^{(1)}) B_{2,s}(x^{(2)}),$$

where $B_{11}, \ldots, B_{1K_1}$ are the basis functions in $x^{(1)}$ direction and $B_{21}, \ldots, B_{2K_2}$ in $x^{(2)}$ direction. The $n \times K = n \times K_1 K_2$ design matrix $\mathbf{X}$ now consists of products of basis functions.

Priors for $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{K_1,K_2})'$ can be based on spatial smoothness priors common in spatial statistics, e.g. two-dimensional first order random walks. The most commonly used prior specification based on the four nearest neighbors is defined by

$$\beta_{ks}|\cdot \sim N\left(\frac{1}{4}(\beta_{k-1,s} + \beta_{k+1,s} + \beta_{k,s-1} + \beta_{k,s+1}), \frac{\tau^2}{4}\right) \tag{17}$$

and appropriate edge corrections. This prior as well as higher order bivariate random walks can be easily brought into the general form (14).

4.2 Spatial effects

4.2.1 Markov random fields

Suppose that the index $s \in \{1, \ldots, S\}$ represents the location or site in connected geographical regions. For simplicity we assume that the regions are labelled consecutively. A common way to introduce a spatially correlated effect is to assume that neighboring sites are more alike than arbitrary sites. Thus, for a valid prior definition a set of neighbors for each site s must be defined. For geographical data one usually assumes that two sites s and s' are neighbors if they share a common boundary.

The simplest (but most frequently used) spatial smoothness prior for the function evaluations $f(s) = \beta_s$ is given by

$$\beta_s|\beta_{s'},\ s \neq s', \tau^2 \sim N\left(\frac{1}{N_s} \sum_{s' \in \partial_s} \beta_{s'}, \frac{\tau^2}{N_s}\right), \tag{18}$$

where N_s is the number of adjacent sites and $s' \in \partial_s$ denotes that site s' is a neighbor of site s. Hence, the (conditional) mean of β_s is an unweighted average of function evaluations of neighboring sites. The prior is a direct generalization of a first order random walk to two-dimensions and is called a Markov random field (MRF).

The $n \times S$ design matrix $\mathbf{X}$ is a 0/1 incidence matrix. Its value in the i-th row and the s-th column is 1 if the i-th observation is located in site or region s, and zero otherwise. The $S \times S$ penalty matrix $\mathbf{K}$ has the form of an adjacency matrix.

4.2.2 Kriging

If exact locations $\mathbf{s} = (s_x, s_y)$ are available, we can use two-dimensional surface estimators to model spatial effects. One option are two-dimensional P-splines, described in subsubsection 3.1.2. Another option are Gaussian random field (GRF) priors, originating from geostatistics. These can also be interpreted as two-dimensional surface smoothers based on radial basis functions and have been employed by Kammann & Wand (2003) to model the spatial component in Gaussian regression

models. The spatial component $f(\mathbf{s}) = \beta_{\mathbf{s}}$ is assumed to follow a zero mean stationary Gaussian random field $\{\beta_{\mathbf{s}} : \mathbf{s} \in \mathbb{R}^2\}$ with variance τ^2 and isotropic correlation function $\text{Cov}(\beta_{\mathbf{s}}, \beta_{\mathbf{s}+\vec{h}}) = C(||\vec{h}||)$. This means that correlations between sites that are $||\vec{h}||$ units apart are the same, regardless of direction and the sites location. For a finite array $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_S\}$ of sites as in image analysis, the prior for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_S)'$ is of the general form (14) with $\mathbf{K} = \mathbf{C}^{-1}$ and

$$\mathbf{C}(i, j) = C(||\mathbf{s}_i - \mathbf{s}_j||), 1 \leq i, j \leq S.$$

The design matrix X is again a 0/1 incidence matrix.

Several proposals for the choice of the correlation function $C(r)$ have been made. In the kriging literature, the Matérn family $C(r; \rho, \nu)$ is highly recommended. For prechosen values $\nu = m + 1/2$, $m = 0, 1, 2, \ldots$ of the smoothness parameter ν simple correlation functions $C(r; \rho)$ are obtained, e.g.

$$C(r; \rho) = \exp(-|r/\rho|)(1 + |r/\rho|)$$

with $\nu = 1.5$. The parameter ρ controls how fast correlations die out with increasing $r = ||h||$. It can be determined in a preprocessing step or may be estimated jointly with the variance components by restricted maximum likelihood. A simple rule, that also ensures scale invariance of the estimates, is to choose ρ as

$$\hat{\rho} = \max_{i,j} ||\mathbf{s}_i - \mathbf{s}_j||/c.$$

The constant $c > 0$ is chosen in such a way, that $C(c)$ is small, e.g. 0.001. Therefore the different values of $||\mathbf{s}_i - \mathbf{s}_j||/\hat{\rho}$ are spread out over the r-axis of the correlation function. This choice of ρ has proved to work well in our experience.

4.2.3 Discrete vs. continuous spatial smoothing

Although we described them separately, approaches for exact locations can also be used in the case of connected geographical regions, e.g. based on the centroids of the regions. Conversely, we can also apply MRFs to exact locations if neighborhoods are defined based on a distance measure. In general, it is not clear which of the different approaches leads to the "best" fits. For data observed on a discrete lattice MRFs seem to be most appropriate. If the exact locations are available, surface estimators may be more natural, particularly because predictions for unobserved locations are available. However, in some situations surface estimators lead to an improved fit compared to MRF's even for discrete lattices and vice versa. A general approach that can handle both situations is given by Müller, Stadtmüller & Tabnak (1997).

From a computational point of view MRF's and P-splines are preferable to GRF's because their posterior precision matrices are band matrices or can be transformed into a band matrix like structure. This special structure considerably speeds up computations, at least for inference based on MCMC techniques. For inference based on mixed models, the main difference between GRFs and MRFs, considering their numerical properties, is the dimension of the penalty matrix. For MRFs the dimension of K equals the number of different regions S and is therefore independent from the sample size. On the other side, for GRFs, the dimension of K is given by the number of distinct locations, which is usually close to the sample size. Therefore, the number of regression coefficients used to describe a MRF is usually much smaller than for a GRF and therefore the estimation of GRFs is computationally much more expensive. To overcome this difficulty Kammann & Wand (2003) propose low-rank kriging to approximate stationary Gaussian random fields. Note first, that we can define GRFs equivalently based on a design matrix $\mathbf{X}$ with entries $\mathbf{X}[i, j] = C(||\mathbf{s}_i - \mathbf{s}_j||)$ and penalty matrix $\mathbf{K} = \mathbf{C}$. To reduce the dimensionality of the estimation problem we define a subset of knots $\mathcal{D} = \{\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_M\}$ of the set of distinct locations $\mathcal{C}$. These knots can be chosen to

be "representative" for the set of distinct locations $\mathcal{C}$ based on a space filling algorithm. Therefore consider the distance measure

$$d(\mathbf{s}, \mathcal{D}) = \left(\sum_{\boldsymbol{\kappa} \in \mathcal{D}} ||\mathbf{s} - \boldsymbol{\kappa}||^p \right)^{\frac{1}{p}},$$

with $p < 0$, between any location $\mathbf{s} \in \mathcal{D}$ and a possible set of knots $\mathcal{C}$. Obviously this distance measure is zero for all knots. Using a simple swapping algorithm to minimize the overall coverage criterion

$$\left(\sum_{\mathbf{s} \in \mathcal{C}} d(\mathbf{s}, \mathcal{D})^q \right)^{\frac{1}{q}}$$

with $q > 0$ (compare Johnson, Moore & Ylvisaker (1990) and Nychka & Saltzman (1998) for details) yields an optimal set of knots $\mathcal{D}$. Based on these knots we define the approximation $\mathbf{f} = \mathbf{X}\boldsymbol{\beta}$ with the $n \times M$ design matrix $\mathbf{X}[i, j] = C(||\mathbf{s}_i - \boldsymbol{\kappa}_j||)$, penalty matrix $\mathbf{K} = \mathbf{C}$, and $\mathbf{C}[i, j] = C(||\boldsymbol{\kappa}_i - \boldsymbol{\kappa}_j||)$. The number of knots M allows to control the trade-off between accuracy of the approximation (M close to the sample size) and numerical simplification (M small).

4.3 Unit- or cluster-specific heterogeneity

In many situations we observe the problem of heterogeneity among clusters of observations caused by unobserved covariates. Neglecting unobserved heterogeneity may lead to considerably biased estimates for the remaining effects as well as false standard error estimates. Suppose now $x \in \{1, \ldots, K\}$ is a cluster variable indicating the cluster a particular observation belongs to. A common approach to overcome the difficulties of unobserved heterogeneity is to introduce additional Gaussian i.i.d. effects $f(x) = \beta_x$ with

$$\beta_x \sim N(0, \tau^2), \quad x = 1, \ldots, K. \tag{19}$$

The design matrix $\mathbf{X}$ is again a $n \times K$-dimensional 0/1 incidence matrix that represents the grouping structure of the data, while the penalty matrix is simply the identity matrix, i.e. $\mathbf{K} = \mathbf{I}$. From a classical perspective, (19) defines i.i.d. *random effects*. However, from a Bayesian point of view all unknown parameters are assumed to be random and hence the notation "random effects" in this context is misleading. Hence, one may also think of (19) as an approach for modelling an unsmooth function.

Prior (19) may also be used for a more sophisticated modelling of spatial effects. In some situation it may be useful to split up the spatial effect f_{spat} into a spatially correlated (structured) part f_{str} and a spatially uncorrelated (unstructured) part f_{unstr}, i.e.

$$f_{spat} = f_{str} + f_{unstr}.$$

A rationale is that a spatial effect is usually a surrogate of many unobserved influential factors, some of which obeying a strong spatial structure while others are present only locally. By estimating a structured and an unstructured component we aim at distinguishing between the two kinds of influential factors, see also Besag, York & Mollié (1991). For the smooth spatial part we can assume any of the spatial priors discussed in subsection 4.2. For the uncorrelated part we may assume prior (19).

4.4 Varying coefficients

The models considered so far are not appropriate for modelling interactions between covariates. A common approach is based on varying coefficient models introduced by Hastie & Tibshirani (1993)

in the context of smoothing splines. Varying coefficient terms have already been discussed in the previous chapter, see section 3.4.

4.5 Regularization Priors for highdimensional covariates

In this section the vector $\boldsymbol{\beta}_s = (\beta_{p+1}, \ldots \beta_{p+q})$ is a subvector of the unknown regression coefficients $\boldsymbol{\gamma}$ for fixed (linear) effects. A desirable feature for variable selection is to shrink small effects to zero but to shrink important effects only moderately to prevent them from large bias. At the opposite to the unregularized regression coefficients where flat priors $p(\boldsymbol{\gamma}) \propto 1$ are assumed, we need informative priors appropriate for shrinkage. Several alternatives have been proposed for specifying shrinkage priors for the regression coefficients, see e.g. Griffin and Brown (2005). Three shrinkage priors, which are hierarchically represented as scale mixtures of normals, are implemented in *BayesX*: the *ridge-*, the *lasso-* and the *Normal Mixture of Inverse Gamma (NMIG)-prior*. Common for all approaches is an informative, zero mean conditional Gaussian distribution of the regression coefficients

$$\beta_j|\sigma^2, \tau_j^2 \overset{i.i.d.}{\sim} N\left(0, \sigma^2\tau_j^2\right), \quad j = p+1, \ldots, p+q$$

together with special mixing distributions for the variances $\boldsymbol{\tau}_s^2 = \left(\tau_{p+1}^2, \ldots, \tau_{p+q}^2\right)$. The general form of the prior for $\boldsymbol{\beta}_s$ is given by

$$p\left(\boldsymbol{\beta}_s|\sigma^2, \tau_s^2\right) \propto \frac{1}{\det\left(\boldsymbol{K}_s\right)^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}_s'\boldsymbol{K}_s^{-1}\boldsymbol{\beta}_s\right), \tag{20}$$

where $\boldsymbol{K}_s = \operatorname{diag}\left(\sigma^2\tau_{p+1}^2, \ldots, \sigma^2\tau_{p+q}^2\right)$ denotes the diagonal matrix of the variances.

Ridge-Prior

In this case the priors of the variances τ_j^2 are point masses given the shrinkage parameter λ

$$\tau_j^2|\lambda \overset{i.i.d.}{\sim} \delta_{1/2\lambda}\left(\tau_j^2\right), \quad j = p+1, \ldots, p+q. \tag{21}$$

The symbol $\delta_a(x)$ denotes the Kronecker function which is 1 if $x = a$ and 0 if $x \neq a$. The shrinkage parameter λ determines the amount of the shrinkage of the regression coefficients and is equipped with a Gamma distribution

$$\lambda \sim Ga(a, b); \quad a, b > 0$$

which results in a marginal (respective τ_j^2) Gaussian distribution of the regression coefficients $\beta_j|\sigma^2, \lambda \overset{i.i.d.}{\sim} N\left(0, \lambda/2\sigma^2\right)$. The logarithm of the common marginal prior for $\beta_j|\sigma^2, \lambda$ corresponds to the ridge penalty and for a given value of λ posterior mode estimation coincides with penalized likelihood estimation.

The *adaptive* version allows the tuning each variance patameter via

$$\tau_j^2|\lambda_j \sim \delta_{1/2\lambda_j}\left(\tau_j^2\right),$$
$$\lambda_j \sim Ga(a_j, b_j).$$

Lasso-Prior

In this case the variances τ_j^2 are (conditional) exponential distributed given the squared shrinkage parameter λ^2

$$\tau_j^2|\lambda^2 \overset{i.i.d.}{\sim} Exp\left(\lambda^2/2\right), \quad j = p+1, \ldots, p+q. \tag{22}$$

The prior of the squared shrinkage parameter is also a gamma distribution

$$\lambda^2 \sim Ga\,(a,b)\,; \quad a,b>0$$

and the corresponding marginal priors of the regression coefficients are i.i.d. Laplace distributions $\beta_j|\sigma^2,\lambda \overset{i.i.d.}{\sim} Lap\,(0,\lambda/\sigma)$ so that the logarithm of the common marginal prior for β (respective τ_j^2) corresponds to the Lasso penalty (Park and Casella, 2008).

The *adaptive* version allows the tuning each variance patameter via

$$\tau_j^2|\lambda_j^2 \sim Exp\left(\lambda_j^2/2\right),$$
$$\lambda_j^2 \sim Ga\,(a_j,b_j)\,.$$

Normal-Mixture of inverse Gamma-Prior

The variance parameters τ_j^2, $j=p+1,\ldots,p+q$ are, in contrast to the ridge and the lasso, specified through a mixture distribution modeled by the product of the two components

$$\begin{aligned} &I_j|\nu_0,\nu_1,\omega \sim (1-\omega)\,\delta_{\nu_o}(\cdot)+\omega\delta_{\nu_1}(\cdot)\,,\\ &t_j^2|a,b \sim IG\,(a,b)\,. \end{aligned} \tag{23}$$

The first component in (23) is an indicator variable with point mass at the values $\nu_0>0$ and $\nu_1>0$ denoted by the corresponding Kronecker symbols. Therein the parameter ν_0 should have a positive value close to zero and the value of ν_1 is set to 1 by default. The parameter ω controls how likely the binary variable I_j equals ν_1 or ν_0, and therefore it takes on the role of a complexity parameter that controls the size of the models. The assumptions in (23) are leading to a continuous bimodal distribution for the variance parameters $\tau_j^2 := I_j t_j^2$, given ν_0, ν_1, ω, a, b with representation as a mixture of scaled inverse Gamma distributions

$$\pi\left(\tau_j^2|\nu_0,\nu_1,\omega,a,b\right) = (1-\omega)\cdot IG\left(\tau_j^2|a,\nu_0 b\right)+\omega\cdot IG\left(\tau_j^2|a,\nu_1 b\right).$$

We assume a beta prior for the parameter ω, i.e. $\omega \sim Beta\,(a_\omega,b_\omega)$, with $a_\omega=b_\omega=1$ as default, which expresses an indifferent prior knowledge about the model complexity.

The *adaptive* version enables

$$\begin{aligned} &I_j|\nu_0,\nu_1,\omega_j \sim (1-\omega_j)\,\delta_{\nu_o}(\cdot)+\omega_j\delta_{\nu_1}\\ &\omega_j \sim Beta\,(a_{\omega,j},b_{\omega,j})\,. \end{aligned}$$

More Details can be found in Kneib, Konrath, Fahrmeir (2009) for the Gaussian case and in Konrath, Kneib and Fahrmeir (2008) for exponential family and hazard regression.

5 Mixed Model representation

In this section we show how STAR models can be represented as generalized linear mixed models (GLMM) after appropriate reparametrization, an idea dating back to Green (1987) in the context of smoothing splines. A broader overview is given in the book by Ruppert, Wand & Carroll (2003), while Fahrmeir, Kneib & Lang (2004) specifically discuss the mixed model representation of structured additive regression models. In fact, model (2) with the structured additive predictor (13) can always be expressed as a GLMM. This provides the key for simultaneous estimation of the functions f_j, $j=1,\ldots,p$ and the variance parameters τ_j^2 in the empirical Bayes approach described in subsection 6.2 and used for estimation by *remlreg objects*. To rewrite the model as a GLMM, the general model formulation again proves to be useful. We proceed as follows:

The vectors of regression coefficients $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, are decomposed into an *unpenalized* and a *penalized part*. Suppose that the j-th coefficient vector has dimension $K_j \times 1$ and the corresponding penalty matrix $\mathbf{K}_j$ has rank k_j. Then we define the decomposition

$$\boldsymbol{\beta}_j = \mathbf{X}_j^{unp}\boldsymbol{\beta}_j^{unp} + \mathbf{X}_j^{pen}\boldsymbol{\beta}_j^{pen}, \tag{24}$$

where the columns of the $K_j \times (K_j - k_j)$ matrix $\mathbf{X}_j^{unp}$ contain a basis of the nullspace of $\mathbf{K}_j$ and $\mathbf{X}_j^{pen}$ contains the orthogonal deviation form this nullspace. Therefore, decomposition (24) effectively separates the unpenalised part in $\boldsymbol{\beta}_j$ from the penalised part. For example, for penalised splines with k_j-th order difference penalty, a polynomial of order $k_j - 1$ forms the null space and is therefore captured in $\boldsymbol{\beta}_j^{unp}$. The $K_j \times k_j$ matrix $\mathbf{X}_j^{pen}$ can be derived as $\mathbf{X}_j^{pen} = \mathbf{L}_j(\mathbf{L}_j'\mathbf{L}_j)^{-1}$ where the $K_j \times k_j$ matrix $\mathbf{L}_j$ is determined by the decomposition of the penalty matrix $\mathbf{K}_j$ into $\mathbf{K}_j = \mathbf{L}_j\mathbf{L}_j'$. A requirement for the decomposition is that $\mathbf{L}_j'\mathbf{X}_j^{unp} = \mathbf{0}$ and $\mathbf{X}_j^{unp}\mathbf{L}_j' = \mathbf{0}$ hold. Hence the parameter vector $\boldsymbol{\beta}_j^{unp}$ represents the part of $\boldsymbol{\beta}_j$ which is not penalized by $\mathbf{K}_j$ whereas the vector $\boldsymbol{\beta}_j^{pen}$ represents the deviation of $\boldsymbol{\beta}_j$ from the nullspace of $\mathbf{K}_j$.

In general, the decomposition $\mathbf{K}_j = \mathbf{L}_j\mathbf{L}_j'$ is obtained from the spectral decomposition $\mathbf{K}_j = \boldsymbol{\Gamma}_j\boldsymbol{\Omega}_j\boldsymbol{\Gamma}_j'$. The $(k_j \times k_j)$ diagonal matrix $\boldsymbol{\Omega}_j$ contains the positive eigenvalues ω_{jm}, $m = 1, \ldots, k_j$, of K_j in descending order, i.e. $\Omega_j = \text{diag}(\omega_{j1}, \ldots, \omega_{j,k_j})$. $\boldsymbol{\Gamma}_j$ is a $(K_j \times k_j)$ orthogonal matrix of the corresponding eigenvectors. From the spectral decomposition we can choose $\mathbf{L}_j = \boldsymbol{\Gamma}_j\boldsymbol{\Omega}_j^{\frac{1}{2}}$. In some cases a more favorable decomposition can be found. For instance, for P-splines a simpler choice for $\mathbf{L}_j$ is given by $\mathbf{L}_j = \mathbf{D}'$ where $\mathbf{D}$ is the first or second order difference matrix. Of course, for prior (19) of subsection 4.3 and in general for proper priors a decomposition of $\mathbf{K}_j$ is not necessary. In this case the unpenalized part vanishes completely.

The matrix $\mathbf{X}_j^{unp}$ is the identity vector $\mathbf{1}$ for P-splines with first order random walk penalty and Markov random fields. For P-splines with second order random walk penalty $\mathbf{X}_j^{unp}$ is a two column matrix where the first column again equals the identity vector while the second column is composed of the (equidistant) knots of the spline.

From the decomposition (24) we get

$$\frac{1}{\tau_j^2}\boldsymbol{\beta}_j'\mathbf{K}_j\boldsymbol{\beta}_j = \frac{1}{\tau_j^2}(\boldsymbol{\beta}_j^{pen})'\boldsymbol{\beta}_j^{pen}$$

and from the general prior (14) for $\boldsymbol{\beta}_j$ it follows that

$$p(\beta_{jm}^{unp}) \propto const, \qquad m = 1, \ldots, K_j - k_j$$

and

$$\boldsymbol{\beta}_j^{pen} \sim N(\mathbf{0}, \tau_j^2\mathbf{I}). \tag{25}$$

Finally, by defining the matrices $\tilde{\mathbf{U}}_j = \mathbf{X}_j\mathbf{X}_j^{unp}$ and $\tilde{\mathbf{X}}_j = \mathbf{X}_j\mathbf{X}_j^{pen}$, we can rewrite the predictor (13) as

$$\begin{aligned}
\boldsymbol{\eta} &= \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{U}\boldsymbol{\gamma} \\
&= \sum_{j=1}^{p}(\tilde{\mathbf{U}}_j\boldsymbol{\beta}_j^{unp} + \tilde{\mathbf{X}}_j\boldsymbol{\beta}_j^{pen}) + \mathbf{U}\boldsymbol{\gamma} \\
&= \tilde{\mathbf{U}}\boldsymbol{\beta}^{unp} + \tilde{\mathbf{X}}\boldsymbol{\beta}^{pen}.
\end{aligned}$$

The design matrix $\tilde{\mathbf{X}}$ and the vector $\boldsymbol{\beta}^{pen}$ are composed of the matrices $\tilde{\mathbf{X}}_j$ and the vectors $\boldsymbol{\beta}_j^{pen}$, respectively. More specifically, we obtain $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1\ \tilde{\mathbf{X}}_2\ \ldots\ \tilde{\mathbf{X}}_p)$ and the stacked vector $\boldsymbol{\beta}^{pen} = ((\boldsymbol{\beta}_1^{pen})', \ldots, (\boldsymbol{\beta}_p^{pen})')'$. Similarly the matrix $\tilde{\mathbf{U}}$ and the vector $\boldsymbol{\beta}^{unp}$ are given by $\tilde{\mathbf{U}} = (\tilde{\mathbf{U}}_1\ \tilde{\mathbf{U}}_2\ \ldots\ \tilde{\mathbf{U}}_p\ \mathbf{U})$ and $\boldsymbol{\beta}^{unp} = ((\boldsymbol{\beta}_1^{unp})', \ldots, (\boldsymbol{\beta}_p^{unp})', \boldsymbol{\gamma}')'$.

Finally, we obtain a GLMM with fixed effects $\boldsymbol{\beta}^{unp}$ and random effects $\boldsymbol{\beta}^{pen} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$ where $\boldsymbol{\Lambda} = \text{diag}(\tau_1^2, \ldots, \tau_1^2, \ldots, \tau_p^2, \ldots, \tau_p^2)$. Hence, we can utilize GLMM methodology for simultaneous estimation of smooth functions and the variance parameters τ_j^2, see the next section.

The mixed model representation also enables us to examine the identification problem inherent to nonparametric regression from a different perspective. For most types of nonparametric effects the design matrix $\tilde{\mathbf{U}}_j$ for the unpenalized part contains the identity vector. Provided that there is at least one such nonlinear effect and that $\boldsymbol{\gamma}$ contains an intercept, the matrix $\tilde{\mathbf{U}}$ has not full column rank. Hence, all identity vectors in $\tilde{\mathbf{U}}$ except for the intercept have to be deleted to guarantee identifiability.

6 Inference

BayesX provides four alternative approaches for Bayesian inference. *Bayesreg objects* (chapter 7 of the reference manual) estimate exponential family, categorical and duration time STAR models using MCMC simulation techniques described in subsection 6.1. *MCMCreg objects* (chapter 10 of the reference manual) provide the counterpart for distributional and quantile regression as well as multilevel models (see section 8). *Remlreg objects* (chapter 8 of the reference manual) use mixed model representations of STAR models for empirical Bayesian inference, see subsection 6.2. *Stepwisereg objects* (chapter 9 of the reference manual) simultaneously perform model selection and estimation of parameters, see subsection 6.3.

6.1 Full Bayesian inference based on MCMC techniques

This subsection may be skipped if you are not interested in using the regression tool for full Bayesian inference based on MCMC simulation techniques (*bayesreg objects*).

For full Bayesian inference, the unknown variance parameters τ_j^2 are also considered as random variables supplemented with suitable hyperprior assumptions. In *BayesX*, highly dispersed (but proper) inverse Gamma priors $p(\tau_j^2) \sim IG(a_j, b_j)$ are assigned to the variances. The corresponding probability density function is given by

$$\tau_j^2 \propto (\tau_j^2)^{-a_j-1} \exp\left(-\frac{b_j}{\tau_j^2}\right).$$

Using proper priors for τ_j^2 (with $a_j > 0$ and $b_j > 0$) ensures propriety of the joint posterior despite the partial impropriety of the priors for the β_j. A common choice for the hyperparameters are small values for a_j and b_j, e.g. $a_j = b_j = 0.001$ which is also the default in *BayesX*.

In some situations, the estimated nonlinear functions f_j may depend considerably on the particular choice of hyperparameters a_j and b_j. This may be the case for very low signal to noise ratio and/or small sample size. It is therefore highly recommended to estimate all models under consideration using a (small) number of *different* choices for a_j and b_j to assess the dependence of results on minor changes in the model assumptions. In that sense, the variation of hyperparameters can be used as a tool for model diagnostics.

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

$$p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \tau_1^2, \ldots, \tau_p^2, \boldsymbol{\gamma} \,|\, \mathbf{y}) \quad \propto \quad L(\mathbf{y}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \boldsymbol{\gamma}) \prod_{j=1}^{p} \left(p(\boldsymbol{\beta}_j|\tau_j^2) p(\tau_j^2)\right) \tag{26}$$

where $L(\cdot)$ denotes the likelihood which, under the assumption of conditional independence, is the product of individual likelihood contributions.

In many practical situations (and in particular for most structured additive regression models) the posterior distribution is numerically intractable. A technique that overcomes this problem are Markov Chain Monte Carlo (MCMC) simulation methods that allow to draw random samples from the posterior. From these random samples, characteristics of the posterior such as posterior means, standard deviations or quantiles can be estimated by their empirical analogues. Instead of drawing samples directly from the posterior (which is impossible in most cases anyway) MCMC devices a way to construct a Markov chain with the posterior as stationary distribution. Hence, the iterations of the transition kernel of this Markov chain converge to the posterior yielding a sample of dependent random numbers. Usually the first part of the sample (the burn-in phase) is discarded since the algorithm needs some time to converge. In addition, some thinning is typically applied to the Markov chain to reduce autocorrelations. In *BayesX* the user can specify options for the number of burn-in iterations, the thinning parameter and the total number of iterations, see chapter 7 of the reference manual for more details.

BayesX provides a number of different sampling schemes, specifically tailored to the distribution of the response. The first sampling scheme is suitable for Gaussian responses. The second sampling scheme is particularly useful for categorical responses and uses the sampling scheme for Gaussian responses as a building block. The third sampling scheme is based on iteratively weighted least squares proposals and is used for general responses from an exponential family. A further sampling scheme, not described in this manual, is based on conditional prior proposals.

6.1.1 Gaussian responses

Suppose first that the distribution of the response variable is Gaussian, i.e. $y_i|\eta_i, \sigma^2 \sim N(\eta_i, \sigma^2/c_i)$, $i = 1, \ldots, n$ or $\mathbf{y}|\boldsymbol{\eta}, \sigma^2 \sim N(\boldsymbol{\eta}, \sigma^2\mathbf{C}^{-1})$ where $\mathbf{C} = \text{diag}(c_1, \ldots, c_n)$ is a known weight matrix. In this case, full conditionals for fixed effects as well as nonlinear functions f_j are multivariate Gaussian and, as a consequence, a Gibbs sampler can be employed. To be more specific, the full conditional $\boldsymbol{\gamma}|\cdot$ for fixed effects with diffuse priors is Gaussian with mean

$$E(\boldsymbol{\gamma}|\cdot) = (\mathbf{U}'\mathbf{C}\mathbf{U})^{-1}\mathbf{U}'\mathbf{C}(\mathbf{y} - \tilde{\boldsymbol{\eta}}) \tag{27}$$

and covariance matrix

$$Cov(\boldsymbol{\gamma}|\cdot) = \sigma^2(\mathbf{U}'\mathbf{C}\mathbf{U})^{-1} \tag{28}$$

where $\mathbf{U}$ is the design matrix of fixed effects and $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta} - \mathbf{U}\boldsymbol{\gamma}$ is the part of the additive predictor associated with the remaining effects in the model. Similarly, the full conditional for the regression coefficients $\boldsymbol{\beta}_j$ of a function f_j is Gaussian with mean

$$\mathbf{m}_j = E(\boldsymbol{\beta}_j|\cdot) = \left(\frac{1}{\sigma^2}\mathbf{X}_j'\mathbf{C}\mathbf{X}_j + \frac{1}{\tau_j^2}\mathbf{K}_j\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}_j'\mathbf{C}(\mathbf{y} - \boldsymbol{\eta}_{-j}), \tag{29}$$

where $\boldsymbol{\eta}_j = \boldsymbol{\eta} - \mathbf{X}_j\boldsymbol{\beta}_j$, and covariance matrix

$$Cov(\boldsymbol{\beta}_j|\cdot) = \mathbf{P}_j^{-1} = \left(\frac{1}{\sigma^2}\mathbf{X}_j'\mathbf{C}\mathbf{X}_j + \frac{1}{\tau_j^2}\mathbf{K}_j\right)^{-1}. \tag{30}$$

Although the full conditional is Gaussian, drawing random samples in an efficient way is not trivial, since linear equation systems with a high dimensional precision matrix $\mathbf{P}_j$ must be solved in every iteration of the MCMC scheme. Following Rue (2001), random numbers from $p(\boldsymbol{\beta}_j|\cdot)$ can be obtained as follows: Compute the Cholesky decomposition $\mathbf{P}_j = \mathbf{L}\mathbf{L}'$ and solve $\mathbf{L}'\boldsymbol{\beta}_j = \mathbf{z}$, where $\mathbf{z}$ is a vector of independent standard Gaussians. It follows that $\boldsymbol{\beta}_j \sim N(\mathbf{0}, \mathbf{P}_j^{-1})$. Afterwards compute the mean $\mathbf{m}_j$ by solving $\mathbf{P}_j\mathbf{m}_j = \frac{1}{\sigma^2}\mathbf{X}_j'\mathbf{C}(\mathbf{y} - \boldsymbol{\eta}_{-j})$. This is achieved by first solving

$\mathbf{L}\boldsymbol{\nu} = \frac{1}{\sigma^2}\mathbf{X}_j'\mathbf{C}(\mathbf{y}-\tilde{\boldsymbol{\eta}})$ by forward substitution followed by backward substitution $\mathbf{L}'\mathbf{m}_j = \boldsymbol{\nu}$. Finally, adding $\mathbf{m}_j$ to the previously simulated $\boldsymbol{\beta}_j$ yields $\boldsymbol{\beta}_j \sim N(\mathbf{m}_j, \mathbf{P}_j^{-1})$.

In most cases, the posterior precision matrices $\mathbf{P}_j$ can be brought into a band matrix like structure with bandsize depending on the prior. If f_j corresponds to a spatially correlated effect for regional data, the posterior precision matrix is usually a sparse matrix but not a band matrix. In this case, the regions of a geographical map must be *reordered*, using the *reverse Cuthill-McKee algorithm*, to obtain a band matrix like precision matrix. Random samples from the full conditional can now be drawn in a very efficient way using Cholesky decompositions for band matrices or band matrix like matrices. In our implementation, we use the *envelope method* for band matrix like matrices as described in George & Liu (1981).

The full conditionals for the variance parameters τ_j^2, $j = 1, \ldots, p$, and σ^2 are all inverse Gamma distributions with parameters

$$a_j' = a_j + \frac{\text{rank}(\mathbf{K}_j)}{2} \quad \text{and} \quad b_j' = b_j + \frac{1}{2}\boldsymbol{\beta}_j'\mathbf{K}_j\boldsymbol{\beta}_j \tag{31}$$

for τ_j^2. For σ^2 we obtain

$$a_\sigma' = a_\sigma + \frac{n}{2} \quad \text{and} \quad b_\sigma' = b_\sigma + \frac{1}{2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \tag{32}$$

where $\boldsymbol{\varepsilon}$ is the usual vector of residuals.

Note that prior to estimation the response variable is standardized in *BayesX* to avoid numerical problems with too large or too small values of the response. All results are, however, retransformed into the original scale.

The sampling scheme for Gaussian responses can be summarized as follows:

Sampling scheme 1:

1. *Initialization:*
   Compute the posterior mode for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ and $\boldsymbol{\gamma}$ given fixed (usually small) smoothing parameters $\lambda_j = \sigma^2/\tau_j^2$, by default *BayesX* uses $\lambda_j = 0.1$. This value may be changed by the user. The mode is computed via backfitting. Use the posterior mode estimates as the initial state $\boldsymbol{\beta}_j^c$, $(\tau_j^2)^c$, $\boldsymbol{\gamma}^c$ of the chain.

2. *Update regression parameters* $\boldsymbol{\gamma}$
   Update regression parameters $\boldsymbol{\gamma}$ by drawing from the Gaussian full conditional with mean and covariance matrix specified in (27) and (28).

3. *Update regression parameters* $\boldsymbol{\beta}_j$
   Update $\boldsymbol{\beta}_j$ for $j = 1, \ldots, p$ by drawing from the Gaussian full conditionals with mean and covariance matrix given in (29) and (30).

4. *Update variance parameters* τ_j^2 *and* σ^2
   Update variance parameters by drawing from inverse gamma full conditionals with parameters given in (31) and (32).

6.1.2 Categorical Response

For most models with categorical responses, efficient sampling schemes can be developed based on latent utility representations. The seminal paper by Albert & Chib (1993) describes algorithms for probit models with ordered categorical responses. The case of probit models with unordered categorical responses is dealt with e.g. in Fahrmeir & Lang (2001). Recently, a similar data augmentation approach for logit models has been presented by Holmes & Held (2006). The adaption

of these sampling schemes to STAR models used in *BayesX* 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, i.e. $y_i \sim B(1, \mu_i)$ with conditional mean $\mu_i = \Phi(\eta_i)$ where Φ is the cumulative distribution function of a standard normal distribution. Introducing latent variables

$$L_i = \eta_i + \epsilon_i, \tag{33}$$

with $\epsilon_i \sim N(0, 1)$, we can equivalently define the binary probit model by $y_i = 1$ if $L_i > 0$ and $y_i = 0$ if $L_i < 0$. The latent variables are augmented as additional parameters and, as a consequence, an additional sampling step for updating the L_is is required. Fortunately, sampling the L_is is relatively easy and fast because the full conditionals are truncated normal distributions. More specifically, $L_i|\cdot \sim N(\eta_i, 1)$ truncated at the left by zero if $y_i = 1$ and truncated by zero at the right if $y_i = 0$. The advantage of defining a probit model through the latent variables L_i is that the full conditionals for the regression parameters $\boldsymbol{\beta}_j$ (and $\boldsymbol{\gamma}$) are again Gaussian with precision matrix and mean given by

$$\mathbf{P}_j = \mathbf{X}_j'\mathbf{X}_j + \frac{1}{\tau_j^2}\mathbf{K}_j, \quad \mathbf{m}_j = \mathbf{P}_j^{-1}\mathbf{X}_j'(\mathbf{L} - \tilde{\boldsymbol{\eta}}). \tag{34}$$

Hence, the efficient and fast sampling schemes for Gaussian responses can be used with slight modifications. Updating of $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}$ can be done exactly as described in sampling scheme 1 using the current values $\mathbf{L}^c$ of the latent utilities as (pseudo) responses and setting $\sigma^2 = 1$, $\mathbf{C} = \mathbf{I}$.

For binary logit models, the sampling schemes become slightly more complicated. A logit model can be expressed in terms of latent utilities by assuming $\epsilon_i \sim N(0, \lambda_i)$ in (33) 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 & Mallows 1974). The full conditionals for the L_is are still truncated normals with $L_i|\cdot \sim N(\eta_i, \lambda_i)$ but additional drawings from the conditional distributions of λ_i are necessary, see Holmes & Held (2006) for details.

Similar updating schemes may be developed for multinomial probit models with unordered categories and cumulative threshold models for ordered categories of the response, see Fahrmeir & Lang (2001) for details. *BayesX* supports both types of models. The cumulative threshold model is, however, restricted to three response categories. For multinomial logit models updating schemes based on latent utilities are not available in *BayesX*.

6.1.3 General uni- or multivariate response from an exponential family

Let us now turn our attention to general responses from an exponential family. In this case the full conditionals are no longer Gaussian, so that more refined algorithms are needed.

BayesX supports several updating schemes based on *iteratively weighted least squares (IWLS) proposals* as proposed by Gamerman (1997) in the context of generalized linear mixed models. As an alternative *conditional prior proposals* as proposed by Knorr-Held (1999) for estimating dynamic models are also available.

The basic idea behind IWLS proposals is to combine Fisher scoring or IWLS (e.g. Fahrmeir & 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 $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}$ 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 $\boldsymbol{\beta}_j$ of the function f_j with current state $\boldsymbol{\beta}_j^c$ of the chain. Then, according to IWLS, a new value $\boldsymbol{\beta}_j^p$ is proposed by drawing a random number from the multivariate Gaussian proposal distribution

$q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)$ with precision matrix and mean

$$\mathbf{P}_j = \mathbf{X}_j'\mathbf{W}(\boldsymbol{\beta}_j^c)\mathbf{X}_j + \frac{1}{\tau_j^2}\mathbf{K}_j, \quad \mathbf{m}_j = \mathbf{P}_j^{-1}\mathbf{X}_j'\mathbf{W}(\boldsymbol{\beta}_j^c)(\tilde{\mathbf{y}}(\boldsymbol{\beta}_j^c) - \boldsymbol{\eta}_{-j}). \tag{35}$$

Here, $\mathbf{W}(\boldsymbol{\beta}_j^c) = \text{diag}(w_1, \ldots, w_n)$ is the usual weight matrix for IWLS with weights $w_i^{-1}(\boldsymbol{\beta}_j^c) = b''(\theta_i)(g'(\mu_i))^2$ obtained from the current state $\boldsymbol{\beta}_j^c$. The vector $\boldsymbol{\eta}_{-j} = \boldsymbol{\eta} - \mathbf{X}_j\boldsymbol{\beta}_j$ is the part of the predictor associated with all remaining effects in the model. The working observations $\tilde{y}_i$ are defined as

$$\tilde{y}_i(\boldsymbol{\beta}_j^c) = \eta_i + (y_i - \mu_i)g'(\mu_i).$$

The sampling scheme can be summarized as follows:

Sampling scheme 2 (IWLS-proposals):

1. *Initialization*
   Compute the posterior mode for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ and $\boldsymbol{\gamma}$ given fixed smoothing parameters $\lambda_j = 1/\tau_j^2$. By default, *BayesX* uses $\lambda_j = 0.1$ but the value may be changed by the user. The mode is computed via backfitting within Fisher scoring. Use the posterior mode estimates as the initial state $\boldsymbol{\beta}_j^c$, $(\tau_j^2)^c$, $\boldsymbol{\gamma}^c$ of the chain.

2. *Update* $\boldsymbol{\gamma}$
   Draw a proposed new value $\boldsymbol{\gamma}^p$ from the Gaussian proposal density $q(\boldsymbol{\gamma}^c, \boldsymbol{\gamma}^p)$ with mean

   $$\mathbf{m}_\gamma = (\mathbf{U}'\mathbf{W}(\boldsymbol{\gamma}^c)\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}(\boldsymbol{\gamma}^c)(\mathbf{y} - \tilde{\boldsymbol{\eta}})$$

   and precision matrix

   $$P_\gamma = \mathbf{U}'\mathbf{W}(\boldsymbol{\gamma}^c)\mathbf{U}.$$

   Accept $\boldsymbol{\gamma}^p$ as the new state of the chain $\boldsymbol{\gamma}^c$ with acceptance probability

   $$\alpha = \frac{L(\mathbf{y}, \ldots, \boldsymbol{\gamma}^p)}{L(\mathbf{y}, \ldots, \boldsymbol{\gamma}^c)} \frac{q(\boldsymbol{\gamma}^p, \boldsymbol{\gamma}^c)}{q(\boldsymbol{\gamma}^c, \boldsymbol{\gamma}^p)},$$

   otherwise keep $\boldsymbol{\gamma}^c$ as the current state.

3. *Update* $\boldsymbol{\beta}_j$
   Draw for $j = 1, \ldots, p$ a proposed new value $\boldsymbol{\beta}_j^p$ from the Gaussian proposal density $q(\boldsymbol{\gamma}^c, \boldsymbol{\gamma}^p)$ with mean and precision matrix given in (35). Accept $\boldsymbol{\beta}_j$ as the new state of the chain $\boldsymbol{\beta}_j^c$ with probability

   $$\alpha = \frac{L(\mathbf{y}, \ldots, \boldsymbol{\beta}_j^p, (\tau_j^2)^c, \ldots, \boldsymbol{\gamma}^c)}{L(\mathbf{y}, \ldots, \boldsymbol{\beta}_j^c, (\tau_j^2)^c, \ldots, \boldsymbol{\gamma}^c)} \frac{p(\boldsymbol{\beta}_j^p \mid (\tau_j^2)^c)}{p(\boldsymbol{\beta}_j^c \mid (\tau_j^2)^c)} \frac{q(\boldsymbol{\beta}_j^p, \boldsymbol{\beta}_j^c)}{q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)},$$

   otherwise keep $\boldsymbol{\beta}_j^c$ as the current state.

4. *Update* τ_j^2
   Update variance parameters by drawing from inverse gamma full conditionals with parameters given in (31).

A slightly different updating scheme computes the mean and the precision matrix of the proposal distribution based on the current posterior mode $\mathbf{m}_j^c$ (from the last iteration) rather than the current $\boldsymbol{\beta}_j^c$, i.e. (35) is replaced by

$$\mathbf{P}_j = \mathbf{X}_j'\mathbf{W}(\mathbf{m}_j^c)\mathbf{X}_j + \frac{1}{\tau_j^2}\mathbf{K}_j, \quad \mathbf{m}_j = \mathbf{P}_j^{-1}\mathbf{X}_j'\mathbf{W}(\mathbf{m}_j^c)(\tilde{\mathbf{y}}(\boldsymbol{\beta}_j^c) - \boldsymbol{\eta}_{-j}). \tag{36}$$

The difference of using $\mathbf{m}_j^c$ rather than $\boldsymbol{\beta}_j^c$ is that the proposal is *independent* of the current state of the chain, i.e. $q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p) = q(\boldsymbol{\beta}_j^p)$. Hence, it is not required to recompute $\mathbf{P}_j$ and $\mathbf{m}_j$ when computing the proposal density $q(\boldsymbol{\beta}_j^p, \boldsymbol{\beta}_j^c)$.

Usually acceptance rates are significantly higher compared to sampling scheme 2. This is particularly useful for updating spatial effects based on Markov random fields where, in many cases, sampling scheme 2 yields quite low acceptance rates.

The sampling scheme can be summarized as follows:

Sampling scheme 3 (IWLS-proposals based on current mode):

1. *Initialization*
   Compute the posterior mode for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ and $\boldsymbol{\gamma}$ given fixed smoothing parameters $\lambda_j = 1/\tau_j^2$. By default, *BayesX* uses $\lambda_j = 0.1$ but the value may be changed by the user. The mode is computed via backfitting within Fisher scoring. Use the posterior mode estimates as the initial state $\boldsymbol{\beta}_j^c$, $(\tau_j^2)^c$, $\boldsymbol{\gamma}^c$ of the chain. Define $\mathbf{m}_j^c$ and $\mathbf{m}_\gamma^c$ as the current mode.

2. *Update* $\boldsymbol{\gamma}$
   Draw a proposed new value $\boldsymbol{\gamma}^p$ from the Gaussian proposal density $q(\boldsymbol{\gamma}^c, \boldsymbol{\gamma}^p)$ with mean

   $$\mathbf{m}_\gamma = (\mathbf{U}'\mathbf{W}(\mathbf{m}_\gamma^c)\mathbf{U})^{-1}\mathbf{U}'\mathbf{W}(\mathbf{m}_\gamma^c)(\mathbf{y} - \tilde{\boldsymbol{\eta}})$$

   and precision matrix

   $$P_\gamma = \mathbf{U}'\mathbf{W}(\mathbf{m}_\gamma^c)\mathbf{U}.$$

   Accept $\boldsymbol{\gamma}^p$ as the new state of the chain $\boldsymbol{\gamma}^c$ with acceptance probability

   $$\alpha = \frac{L(\mathbf{y}, \ldots, \boldsymbol{\gamma}^p)}{L(y, \ldots, \boldsymbol{\gamma}^c)} \frac{q(\boldsymbol{\gamma}^p, \boldsymbol{\gamma}^c)}{q(\boldsymbol{\gamma}^c, \boldsymbol{\gamma}^p)},$$

   otherwise keep $\boldsymbol{\gamma}^c$ as the current state.

3. *Update* $\boldsymbol{\beta}_j$
   Draw for $j = 1, \ldots, p$ a proposed new value $\boldsymbol{\beta}_j^p$ from the Gaussian proposal density $q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)$ with mean and precision matrix given in (36). Accept $\boldsymbol{\beta}_j^p$ as the new state of the chain $\boldsymbol{\beta}_j^c$ with probability

   $$\alpha = \frac{L(y, \ldots, \boldsymbol{\beta}_j^p, (\tau_j^2)^c, \ldots, \boldsymbol{\gamma}^c)}{L(y, \ldots, \boldsymbol{\beta}_j^c, (\tau_j^2)^c, \ldots, \boldsymbol{\gamma}^c)} \frac{p(\boldsymbol{\beta}_j^p \mid (\tau_j^2)^c)}{p(\boldsymbol{\beta}_j^c \mid (\tau_j^2)^c)} \frac{q(\boldsymbol{\beta}_j^p, \boldsymbol{\beta}_j^c)}{q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)},$$

   otherwise keep $\boldsymbol{\beta}_j^c$ as the current state.

4. *Update* τ_j^2
   Update variance parameters by drawing from inverse gamma full conditionals with parameters given in (31).

6.1.4 Inference of the Shrinkage Components

If shrinkage priors are involved the posterior (26) have to be completed to take account for the additional shrinkage components $\beta_{p+1}, \ldots, \beta_{p+q}, \tau_{p+1}^2, \ldots, \tau_{p+q}^2$ and λ. Due to the similarity of the prior (20) to the prior (14) the full conditionals of the shrinkage effects $\boldsymbol{\beta}_s = (\beta_{p+1}, \ldots \beta_{p+q})$ are similar to the full conditionals of the previous subsections 6.1.1 to 6.1.3 if the penalty matrix $K_j \big/ \tau_j^2$ is replaced by the penalty matrix of the shrinkage priors $\boldsymbol{K}_s^{-1}$. For example for Gaussian responses

6.1.1 the full conditional for the regression coefficients $\boldsymbol{\beta}_s$ is also Gaussian with covariance (inverse precision) matrix and mean similar to the formulas (29) and (30):

$$Cov\left(\boldsymbol{\beta}_s|\cdot\right) = \boldsymbol{P}_s^{-1} = \left(\frac{1}{\sigma^2}\boldsymbol{X}_s'\boldsymbol{C}\boldsymbol{X}_s + \boldsymbol{K}_s^{-1}\right)^{-1}, \quad \mathbf{m}_s = \boldsymbol{P}_s^{-1}\frac{1}{\sigma^2}\boldsymbol{X}_s'\boldsymbol{C}\left(\mathbf{y} - \boldsymbol{\eta}_{-s}\right)$$

where $\boldsymbol{X}_s$ is the design matrix corresponding to the shrinkage effects in $\boldsymbol{\beta}_s$ and $\boldsymbol{\eta}_s = \boldsymbol{\eta} - \boldsymbol{X}_s\boldsymbol{\beta}_s$.

For categorical responses 6.1.2 and for general uni- or multivariate response from an exponential family 6.1.3 the full conditionals for the regression coefficients β_s are achieved in the same way.

For all response types the full conditionals of the variances $\tau_s^2 = \left(\tau_{p+1}^2, \ldots \tau_{p+q}^2\right)$ of the shrinkage effects and the complexity parameters λ, ω are known densities so that the corresponding updates for the Markov chain are available via Gibbs steps.

Ridge-Prior

For the ridge prior the full conditionals of the variance parameters simplify to $\tau_j^2|\cdot \sim 1/2\lambda$, $j = p+1, \ldots, p+q$. For the shrinkage parameter λ we get a gamma density with parameters

$$a' = \frac{q}{2} + a \quad \text{and} \quad b' = \sum_{j=p+1}^{p+q} \beta_j^2/\sigma^2 + b$$

and for the *adaptive* version

$$a' = \frac{1}{2} + a_j \quad \text{and} \quad b' = \beta_j^2/\sigma^2 + b_j.$$

Lasso-Prior

For the variance parameters τ_j^2, $j = p+1, .., p+q$ of the Bayesian lasso the full conditionals are inverse Gaussian distributions

$$\frac{1}{\tau_j^2}|\cdot \sim InvGauss\left(\frac{\sqrt{\sigma^2\lambda^2}}{|\beta_j|}, \lambda^2\right).$$

The full conditional for the quadratic shrinkage parameter $\lambda^2|\cdot$ is a Gamma distribution with parameters:

$$a' = q + a \quad \text{and} \quad b' = \frac{1}{2}\sum_{j=p+1}^{p+q} \tau_j^2 + b.$$

For the *adaptive* version we get

$$\frac{1}{\tau_j^2}|\cdot \sim InvGauss\left(\frac{\sqrt{\sigma^2\lambda_j^2}}{|\beta_j|}, \lambda_j^2\right).$$

and

$$a' = 1 + a_j \quad \text{and} \quad b' = \frac{1}{2}\tau_j^2 + b_j.$$

Normal-Mixture of inverse Gamma-Prior

The diagonal matrix K_s now contains the diagonal elements $\tau_j^2 = I_j t_j^2$, $j = p+1, \ldots, p+q$. The full conditionals for the binary indicator variables are Bernoulli distributions with probabilities

$$p_{1,j} = (1 + A_{I,j}/B_{I,j})^{-1}$$

and

$$\frac{A_{I,j}}{B_{I,j}} = \frac{(1-\omega)}{\omega}\frac{\sqrt{\nu_1}}{\sqrt{\nu_0}}\exp\left\{-\frac{1}{2\sigma^2\nu_0 t_j^2}\beta_j^2 + \frac{1}{2\sigma^2\nu_1 t_j^2}\beta_j^2\right\}.$$

The full conditionals for the second variance parameter component t_j^2 are inverse gamma densities with parameters

$$a' = \frac{1}{2} + a \quad \textit{and} \quad b' = b + \frac{\beta_j^2}{2\sigma^2 I_j}.$$

while the full conditional for the complexity parameter ω is a $Beta\,(a_\omega + n.\nu_1; b_\omega + n.\nu_0)$ distribution with $n.\nu_0 := \#\{j : I_j = \nu_0\}, n.\nu_1 := \#\{j : I_j = \nu_1\}$.

For the *adaptive* version we get

$$\frac{A_{I,j}}{B_{I,j}} = \frac{(1-\omega_j)}{\omega_j}\frac{\sqrt{\nu_1}}{\sqrt{\nu_0}}\exp\left\{-\frac{1}{2\sigma^2\nu_0 t_j^2}\beta_j^2 + \frac{1}{2\sigma^2\nu_1 t_j^2}\beta_j^2\right\}$$

and ω_j is a $Beta\,(a_{\omega,j} + 1; a_{\omega,j} + 1)$ distribution.

The scale parameter σ^2 is for all shrinkage priors inverse gamma with parameters

$$a'_\sigma = a_\sigma + \frac{n}{2} + \frac{q}{2} \quad \textit{und} \quad b'_\sigma = b_\sigma + \frac{1}{2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} + \frac{1}{2}\boldsymbol{\beta}'(\boldsymbol{K}_s/\sigma^2)^{-1}\boldsymbol{\beta}$$

and the corresponding variances matrix $\boldsymbol{K}_s/\sigma^2 = \text{diag}\left(\tau_{p+1}^2, \ldots, \tau_{p+q}^2\right)$ for the ridge, lasso and NMIG.

6.2 Empirical Bayes inference based on mixed model methodology

This section may be skipped if you are not interested in using the regression tool based on mixed model methodology (*remlreg objects*).

For empirical Bayes inference, the variances τ_j^2 are considered as unknown constants to be estimated from their marginal likelihood. In terms of the GLMM representation outlined in section 5, the posterior is given by

$$p(\boldsymbol{\beta}^{unp}, \boldsymbol{\beta}^{pen} \,|\, \mathbf{y}) \quad \propto \quad L(\mathbf{y}, \boldsymbol{\beta}^{unp}, \boldsymbol{\beta}^{pen}) \prod_{j=1}^{p} \left(p(\boldsymbol{\beta}_j^{pen}|\tau_j^2)\right) \tag{37}$$

where $p(\boldsymbol{\beta}_j^{pen}|\tau_j^2)$ is defined in (25).

Based on the GLMM representation, regression and variance parameters can be estimated using iteratively weighted least squares (IWLS) and (approximate) marginal or restricted maximum likelihood (REML) developed for GLMMs. Estimation is carried out iteratively in two steps:

1. Obtain updated estimates $\hat{\boldsymbol{\beta}}^{unp}$ and $\hat{\boldsymbol{\beta}}^{pen}$ given the current variance parameters as the solutions of the system of equations

$$\begin{pmatrix} \tilde{\mathbf{U}}'\mathbf{W}\tilde{\mathbf{U}} & \tilde{\mathbf{U}}'\mathbf{W}\tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}'\mathbf{W}\tilde{\mathbf{U}} & \tilde{\mathbf{X}}'\mathbf{W}\tilde{\mathbf{X}} + \tilde{\boldsymbol{\Lambda}}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{unp} \\ \boldsymbol{\beta}^{pen} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{U}}'\mathbf{W}\tilde{\mathbf{y}} \\ \tilde{\mathbf{X}}'\mathbf{W}\tilde{\mathbf{y}} \end{pmatrix}. \tag{38}$$

The $(n \times 1)$ vector $\tilde{\mathbf{y}}$ and the $n \times n$ diagonal matrix $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$ are the usual working observations and weights in generalized linear models, see subsubsection 6.1.3.

2. Updated estimates for the variance parameters $\hat{\tau}_j^2$ are obtained by maximizing the (approximate) marginal / restricted log likelihood

$$\begin{aligned} l^*(\tau_1^2, \ldots, \tau_p^2) &= -\tfrac{1}{2}\log(|\boldsymbol{\Sigma}|) - \tfrac{1}{2}\log(|\tilde{\mathbf{U}}\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{U}}|) \\ &\quad -\tfrac{1}{2}(\tilde{\mathbf{y}} - \tilde{\mathbf{U}}\hat{\boldsymbol{\beta}}^{unp})'\boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{U}}\hat{\boldsymbol{\beta}}^{unp}) \end{aligned} \tag{39}$$

with respect to the variance parameters $\tau_1^2, \ldots, \tau_p^2$. Here, $\boldsymbol{\Sigma} = \mathbf{W}^{-1} + \tilde{\mathbf{X}}\boldsymbol{\Lambda}\tilde{\mathbf{X}}'$ is an approximation to the marginal covariance matrix of $\tilde{\mathbf{y}}|\boldsymbol{\beta}^{pen}$.

The two estimation steps are iterated until convergence. In *BayesX*, the marginal likelihood (39) is maximized by a computationally efficient alternative to the usual Fisher scoring iterations as described e.g. in Harville (1977), see Fahrmeir, Kneib & Lang (2004) for details.

Convergence problems of the above algorithm may occur, if one of the parameters τ_j^2 is small. In this case the maximum of the marginal likelihood may be on the boundary of the parameter space so that Fisher scoring fails in finding the marginal likelihood estimates $\hat{\tau}^2$. Therefore, the estimation of small variances τ_j^2 is stopped if the criterion

$$c(\tau_j^2) = \frac{||\tilde{\mathbf{X}}_j\hat{\boldsymbol{\beta}}_j^{pen}||}{||\hat{\boldsymbol{\eta}}||} \tag{40}$$

is smaller than the user-specified value `lowerlim`. This usually corresponds to small values of the variances τ_j^2 but defines "small" in a data driven way.

Models for categorical responses are in principle estimated in the same way as presented above but have to be embedded into the framework of multivariate generalized linear models, see Kneib & Fahrmeir (2006) for details.

6.3 Simultaneous selection of model terms and estimation of unknown parameters

This section may be skipped if you are not interested in using (*stepwisereg objects*).

A main building block of the algorithms of *stepwisereg objects* are smoothers of the form

$$S(\mathbf{y}, \boldsymbol{\lambda}) = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{P}(\boldsymbol{\lambda}))^{-1}\mathbf{X}'\mathbf{y},$$

where $\mathbf{X}$ and $\mathbf{P}(\boldsymbol{\lambda})$ are design and penalty matrices corresponding to the smooth covariate effects discussed in the preceding section. For fixed smoothing parameter(s) $\hat{\boldsymbol{\beta}}$ is the minimizer of the penalized least squares criterion

$$PLS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{P}(\boldsymbol{\lambda})\boldsymbol{\beta}.$$

Consecutively applying smoothers S_j corresponding to the j-th function f_j in (3) to the current partial residual reveals the well known backfitting algorithm to minimize the overall PLS-criterion

$$PLS = \left(\mathbf{y} - \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\beta}_j - \mathbf{U}\boldsymbol{\gamma}\right)'\left(\mathbf{y} - \sum_{j=1}^{p}\mathbf{X}_j\boldsymbol{\beta}_j - \mathbf{U}\boldsymbol{\gamma}\right) + \sum_{j=1}^{p}\boldsymbol{\beta}_j'\mathbf{P}_j(\boldsymbol{\lambda}_j)\boldsymbol{\beta}_j.$$

The complexity of the fit may be determined by the equivalent degrees of freedom *df* as a measure of the effective number of parameters. In concordance with linear models the degrees of freedom of the fit are defined as

$$\text{df} = \text{trace}(\mathbf{H}),$$

where $\mathbf{H}$ is the prediction matrix that projects the observations $\mathbf{y}$ on their fitted values $\hat{\mathbf{y}}$, i.e. $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. In complex models with many effects the trace of $\mathbf{H}$ is difficult and computationally intensive to compute. Therefore *df* is typically approximated by the sum of the degrees of freedom of individual smoothers, i.e.

$$\mathrm{df} = \sum_{j=1}^{p} df_j + q,$$

where q is the number of linear effects in the model and df_j is in most cases computed as

$$df_j = trace(\mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j + \mathbf{P}_j(\boldsymbol{\lambda}_j))^{-1}\mathbf{X}_j') - 1. \tag{41}$$

The substraction of one from the trace is necessary because terms are usually centered around zero to guarantee identifiability and as a consequence one degree of freedom is lost. For two-dimensional P-splines as well as unit- or cluster specific effects the approximation is not valid and modifications are necessary, see Belitz & Lang (2008).

Our variable selection procedure described below aims at minimizing a goodness of fit criterion. The following options are available:

- **Test- and validation sample**
  Provided that enough data are available the best strategy is to divide the data into a test- and validation sample. The test data set is used to estimate the parameters of the models. The fit of different models is assessed via the validation data set. In the case of a continuous response, typically the mean squared prediction error is minimized.

- **Goodness of fit criteria** AIC, AIC_c, BIC where AIC_c is the bias corrected version of the AIC as proposed by Hurvich, Simonoff & Tsai (1998).

- **5 or 10 fold cross validation**
  Cross validation is recommended if the approximation to the degrees of freedom is likely to be erroneous e.g. because of strong correlations among covariates.

The basic algorithm for simultaneous selection of model terms and estimation of parameters works as follows:

1. **Initialization**
   Define for every possible nonlinear term f_j, $j = 1, \ldots, p$, a discrete number M_j of decreasing smoothing parameters $\lambda_{j1} > \ldots > \lambda_{jM_j}$. The smoothing parameters are chosen such that they correspond to certain equivalent degrees of freedom.

2. **Start model**
   Choose a start model with current predictor

   $$\hat{\boldsymbol{\eta}} = \hat{\mathbf{f}}_1 + \ldots + \hat{\mathbf{f}}_p.$$

   where $\hat{\mathbf{f}}_j$ is the vector of function evaluations at the observations. Choose a goodness of fit criterion C.

3. **Iteration**

   a) For $j = 1, \ldots, p$:

For $m = 0, \ldots, M_j$:
Compute the fits

$$\begin{aligned}\hat{\mathbf{f}}_{jm} &:= \begin{cases} \mathbf{0} & m = 0 \\ \mathbf{S}_j(\mathbf{y} - \hat{\boldsymbol{\eta}}_{[j]}, \lambda_{jm}) & m = 1, \ldots, M_j \end{cases} \\ &= \begin{cases} \mathbf{0} & m = 0 \\ (\mathbf{X}_j'\mathbf{X}_j + \mathbf{P}(\lambda_{jm}))^{-1}\mathbf{X}_j'(\mathbf{y} - \hat{\boldsymbol{\eta}}_{[j]}) & m = 1, \ldots, M_j \end{cases}\end{aligned}$$

and the corresponding predictors $\hat{\boldsymbol{\eta}}_{jm} := \hat{\boldsymbol{\eta}}_{[j]} + \hat{\mathbf{f}}_{jm}$. Here, $\hat{\boldsymbol{\eta}}_{[j]}$ is the current predictor with the j-th fit $\hat{\mathbf{f}}_j$ removed.

Compute the updated estimate

$$\hat{\mathbf{f}}_j = \text{argmin } C(\hat{\mathbf{f}}_{jm}),$$

i.e. among the fits $\hat{\mathbf{f}}_{jm}$ for the j-th component, choose the one that minimizes the goodness of fit criteria C.

b) The linear effects part $\mathbf{u}'\boldsymbol{\gamma}$ typically consists of the intercept γ_0 and dummy variables for the categorical covariates. For the moment suppose that $\mathbf{u}$ contains dummies representing only one categorical variable. Then we compare the fits $\hat{\gamma}_0 = \overline{y - \eta_{[lin]}}, \gamma_1 = 0, \ldots, \gamma_q = 0$ (covariate removed from the model) and $\hat{\boldsymbol{\gamma}} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'(\mathbf{y} - \hat{\boldsymbol{\eta}}_{[lin]})$ where $\hat{\boldsymbol{\eta}}_{[lin]}$ is the current predictor with the linear effects removed and $\overline{y - \eta_{[lin]}}$ is the mean of the elements of the partial residual vector $\mathbf{y} - \hat{\boldsymbol{\eta}}_{[lin]}$. If more than one categorical covariate is available the procedure is repeated for every variable.

4. **Termination**
The iteration cycle in 3. is repeated until the model, regression and smoothing parameters do not change anymore.

If a two-dimensional surface with penalty (9) is specified the basic algorithm 1 must be adapted, see Belitz & Lang (2008) for details.

7 Survival analysis and multi-state models

We will now describe some of the capabilities of *BayesX* for the estimation of survival time and multi-state models. Discrete time duration and multi-state models can be estimated by categorical regression models after some data augmentation as outlined in subsection 7.1. Continuous time survival models can estimated using either the piecewise exponential model or structured hazard regression, an extension of the well known Cox model (subsection 7.2). For the latter, extensions allowing for interval censored survival times are described in subsection 7.3. Finally, subsection 7.4 contains information on continuous time multi-state models.

While discrete time models and the piecewise exponential models can be estimated with all three regression objects available in *BayesX*, continuous time survival and multi-state models are only supported by *bayesreg* and *remlreg objects*. More details on estimating continuous time survival models based on MCMC can be found in Hennerfeind, Brezger & Fahrmeir (2006). Kneib & Fahrmeir (2006) and Kneib (2006) present inference based on mixed model methodology. Kneib & Hennerfeind (2006) introduce both MCMC and mixed model based inference in multi-state models.

7.1 Discrete time duration data

In applications, duration data are often measured on a discrete time scale or can be grouped in suitable intervals. In this section we show how data of this kind can be written as categorical regression models. Estimation is then based on methodology for categorical regression models as described in the previous sections. We start by assuming that there is only one type of failure event, i.e. we consider the case of survival times.

Let the duration time scale be divided into intervals $[a_0, a_1), [a_1, a_2), \ldots, [a_{q-1}, a_q), [a_q, a_\infty)$. Usually $a_0 = 0$ is assumed and a_q denotes the final follow up time. Identifying the discrete time index t with interval $[a_{t-1}, a_t)$, duration time T is considered as discrete, where $T = t \in \{1, \ldots, q+1\}$ denotes end of duration within the interval $t = [a_{t-1}, a_t)$. In addition to duration T, a sequence of possibly time-varying covariate vectors u_t is observed. Let $u_t^* = (u_1, \ldots, u_t)$ denote the history of covariates up to interval t. Then the discrete hazard function is given by

$$\lambda(t; u_t^*) = P(T = t \mid T \geq t, u_t^*), \quad t = 1, \ldots, q,$$

that is the conditional probability for the end of duration in interval t, given that the interval is reached and the history of the covariates. Discrete time hazard functions can be specified in terms of binary response models. Common choices are binary logit, probit or grouped Cox models.

For a sample of individuals i, $i = 1, \ldots, n$, let T_i denote duration times and C_i right censoring times. Duration data are usually given by $(t_i, \delta_i, u_{it_i}^*)$, $i = 1, \ldots, n$, where $t_i = \min(T_i, C_i)$ is the observed discrete duration time, $\delta_i = 1$ if $T_i \leq C_i$, $\delta_i = 0$ else is the censoring indicator, and $u_{it_i}^* = (u_{it}, t = 1, \ldots, t_i)$ is the observed covariate sequence. We assume that censoring is noninformative and occurs at the end of the interval, so that the risk set R_t includes all individuals who are censored in interval t.

We define binary event indicators y_{it}, $i \in R_t$, $t = 1, \ldots, t_i$, by

$$y_{it} = \begin{cases} 1 & \text{if } t = t_i \text{ and } \delta_i = 1 \\ 0 & \text{else.} \end{cases}$$

Then the duration process of individual i can be considered as a sequence of binary decisions between remaining in the transient state $y_{it} = 0$ or leaving for the absorbing state $y_{it} = 1$, i.e. end of duration at t. For $i \in R_t$, the hazard function for individual i can be modelled by binary response models

$$P(y_{it} = 1 \mid u_{it}^*) = h(\eta_{it}), \tag{42}$$

with appropriate predictor η_{it} and response function $h : \mathbf{R} \to [0, 1]$. Traditionally, a linear predictor is assumed, i.e.

$$\eta_{it} = \gamma_0(t) + u_{it}'\gamma, \tag{43}$$

where the sequence $\gamma_0(t)$, $t = 1, \ldots, q$, represents the baseline effect. In *BayesX* the linear predictor can be replaced by a structured additive predictor

$$\eta_{it} = f_0(t) + f_1(x_{it1}) + \cdots + f_p(x_{itp}) + u_{it}'\gamma, \tag{44}$$

where, again, the x_j denote generic covariates of different types and dimension, and f_j are (not necessarily smooth) functions of the covariates. The baseline effect $f_0(t)$ can be modelled as a penalised spline or a random walk.

To fix ideas, we describe the necessary manipulations that yield a data set suitable for binary regression models with an example. Suppose the first few observations of a data set are given as follows:

t	δ	x1	x2
4	0	0	2
3	1	1	0
⋮	⋮	⋮	⋮

The first individual is censored ($\delta = 0$) and the observed duration time is 4. The second individual is uncensored with duration time 3. Now we augment the data set as follows:

y	indnr	t	δ	x1	x2
0	1	1	0	0	2
0	1	2	0	0	2
0	1	3	0	0	2
0	1	4	0	0	2
0	2	1	1	1	0
0	2	2	1	1	0
1	2	3	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮

The first individual is now represented by 4 observations because the observed duration time is 4. The event indicator y always equals 0 because the corresponding observations is censored. For the second individual we obtain 3 observations and the event indicator jumps at time t=3 from 0 to 1. Now we can estimate a logit or probit model with y as the response and covariates t, x1, x2.

So far we have only considered situations with one type of failure. Suppose now that we may distinguish several types of failure. For example, Fahrmeir & Lang (2001) distinguished between full- and part time jobs as events ending the duration of unemployment. Models of this kind are often referred to as competing risks models.

Let $R \in \{1, \ldots, m\}$ denote distinct events of failure. Then the cause-specific discrete hazard function resulting from cause or risk r is given by

$$\lambda_r(t|u_t, x_t) = P(T = t, R = r|T \geq t, u_t, x_t).$$

Modelling $\lambda_r(t|u_t, x_t)$ may be based on multicategorial regression models. For example, assuming a multinomial logit model yields

$$\lambda_r(t|u_t) = \frac{\exp(\eta_r)}{1 + \sum_{s=1}^{m} \exp(\eta_s)}$$

with structured additive predictors

$$\eta_r = f_{0r}(t) + f_{1r}(x_{t1}) + \cdots + f_{pr}(x_{tp}) + u_t'\gamma_r. \tag{45}$$

An alternative would be the multinomial probit model.

Again, we demonstrate the necessary data manipulations with an example. Suppose we have data with 2 terminating events R=1 and R=2. The first few observations of a data set are given as follows:

t	δ	R	x1	x2
4	0	1	0	2
3	1	2	1	0
5	1	1	0	3
⋮	⋮	⋮	⋮	⋮

The first individual is censored ($\delta = 0$) and the observed duration time is 4. The second individual is uncensored with duration time 3 and terminating event R=2. The third individual is uncensored with duration time 5 and terminating event R=1. We augment the data set as follows:

y	indnr	t	δ	x1	x2
0	1	1	0	0	2
0	1	2	0	0	2
0	1	3	0	0	2
0	1	4	0	0	2
0	2	1	1	1	0
0	2	2	1	1	0
2	2	3	1	1	0
0	3	1	1	0	3
0	3	2	1	0	3
0	3	3	1	0	3
0	3	4	1	0	3
1	3	5	1	0	3
⋮	⋮	⋮	⋮	⋮	⋮

For the first individual we create 4 observations because the observed duration time is 4. The event indicator y always equals 0 because the observation is censored. For the second individual we obtain 3 observations and the event indicator jumps at time t=3 from 0 to 2. For the third individual the event indicator jumps at time 5 from 0 to 1. Now we can estimate a multinomial logit or probit model with y as the response, reference category 0, and covariates t, x1, x2.

7.2 Continuous time survival analysis for right censored survival times

In applications where the duration time t is measured on a continuous time scale, grouping the data for a discrete time analysis is possible, but causes a loss of information. In this section we introduce the continuous time Cox model and describe the two alternatives *BayesX* offers for the estimation of such models. The first alternative is to assume that all time-dependent values are piecewise constant, which leads to the so called piecewise exponential model (p.e.m.). Data augmentation is needed here, but estimation is then based on methodology for Poisson regression, and the inclusion of time-varying effects does not imply any difficulties. The second alternative is to estimate the log-baseline effect by a P-spline of arbitrary degree. This approach is less restrictive and does not demand data augmentation.

Let $u_t^* = \{u_s, 0 \leq s \leq t\}$ denote the history of possibly time-varying covariates up to time t. Then the continuous hazard function is given by

$$\lambda(t; u_t^*) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t, u_t^*)}{\Delta t},$$

i.e. by the conditional instantaneous rate of end of duration at time t, given that time t is reached and the history of the covariates. In the Cox model the individual hazard rate is modelled by

$$\lambda_i(t) = \lambda_0(t) \cdot \exp(\eta_{it}) = \exp(f_0(t) + \eta_{it}) \tag{46}$$

where $\lambda_0(t)$ is the baseline hazard, ($f_0(t) = \log(\lambda_0(t))$ is the log-baseline hazard) and η_{it} is an appropriate predictor. Traditionally the predictor is linear and the baseline hazard is unspecified.

In *BayesX*, however, a structured additive predictor may be assumed and the baseline effect is estimated jointly with the covariate effects either by a piecewise constant function (in case of a p.e.m.) or by a P-spline.

7.2.1 Piecewise exponential model (p.e.m.)

The basic idea of the p.e.m. is to assume that all values that depend on time t are piecewise constant on a grid

$$(0, a_1], (a_1, a_2], \ldots, (a_{s-1}, a_s], \ldots, (a_{t-1}, a_t], (a_t, \infty),$$

where a_t is the largest of all observed duration times $t_i, i = 1, \ldots, n$. This grid may be equidistant or, for example, constructed according to quantiles of the observed survival times. The assumption of a p.e.m. is quite convenient since estimation can be based on methodology for Poisson regression models. For this purpose the data set has to be modified as described below.

Let δ_i be an indicator of non-censoring (i.e. $\delta_i = 1$ if observation i is uncensored, 0 else) and $\gamma_{0s}, s = 1, 2, \ldots$ the piecewise constant log-baseline effect. We define an indicator variable y_{is} as well as an offset Δ_{is} as follows:

$$y_{is} = \begin{cases} 1 & t_i \in (a_{s-1}, a_s],\ \delta_i = 1 \\ 0 & \text{else.} \end{cases}$$

$$\Delta'_{is} = \begin{cases} a_s - a_{s-1}, & a_s < t_i \\ t_i - a_{s-1}, & a_{s-1} < t_i \leq a_s \\ 0, & a_{s-1} \geq t_i \end{cases}$$

$$\Delta_{is} = \log \Delta'_{is} \quad (\Delta_{is} = -\infty \text{ if } \Delta'_{is} = 0).$$

The likelihood contribution of observation i in the interval $(a_{s-1}, a_s]$ is

$$L_{is} = \exp\left(y_{is}(\gamma_{0s} + \eta_{is}) - \exp(\Delta_{is} + \gamma_{0s} + \eta_{is})\right).$$

As this likelihood is proportional to a Poisson likelihood with offset δ_{is}, estimation can be performed using Poisson regression with response variable y, (log-)offset Δ and a as a continuous covariate. Due to the assumption of a piecewise constant hazard rate the estimated log-baseline is a step function on the defined grid. To obtain a smooth step function, a random walk (corresponding to a zero degree P-spline) is specified for the parameters γ_{0s}.

In practice this means that the data set has to be modified in such a way that for every individual i there is an observation row for each interval $(a_{s-1}, a_s]$ between a_0 and the final duration time t_i. Instead of the indicator of non-censoring δ_i the modified data set contains the indicator y_{is} and instead of duration time t_i the variable a_s as well as the offset Δ_{is} (covariates are duplicated). To give a short example, consider an equidistant grid with interval width 0.1 and observations

t	δ	x1	x2
0.25	1	0	3
0.12	0	1	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Then the data set has to be augmented to

y	indnr	a	δ	Δ	x1	x2
0	1	0.1	1	log(0.1)	0	3
0	1	0.2	1	log(0.1)	0	3
1	1	0.3	1	log(0.05)	0	3
0	2	0.1	0	log(0.1)	1	5
0	2	0.2	0	log(0.02)	1	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Now a Poisson model with offset Δ, response y, random walk prior for covariate a, and appropriate priors for x1 and x2 can be estimated.

7.2.2 Specifying a P-spline prior for the log-baseline

The p.e.m. can be seen as a model where the log-baseline $f_0(t)$ in (46) is modelled by a P-spline (see (4.1)) of degree 0, which is quite convenient as it simplifies the calculation of the likelihood, but may be too restrictive since the baseline effect is estimated by a step-function. A more general way of estimating the nonlinear shape of the baseline effect is to assume a P-spline prior of arbitrary degree instead. Unlike the p.e.m. such a model can not be estimated within the context of GAMs, but specific methods for extended Cox models are also implemented in *BayesX* (for *bayesreg* and *remlreg objects*. The individual likelihood for continuous survival data is given by

$$L_i \quad = \quad \lambda_i(t_i)^{\delta_i} \cdot \exp\left(-\int_0^{t_i} \lambda_i(u)du\right).$$

Inserting (46) yields

$$L_i \quad = \quad \exp(f_0(t_i) + \eta_{it_i})^{\delta_i} \cdot \exp\left(-\int_0^{t_i} \exp(f_0(u) + \eta_{iu})du\right).$$

If the degree of the P-spline prior for $f_0(t)$ is greater than one, the integral can no longer be calculated analytically. For linear P-splines the integral can still be solved but the formulae become quite cumbersome. Therefore *BayesX* makes use of the trapezoidal rule for a numerical approximation.

7.3 Continuous time survival analysis for interval censored survival times

Usually, the Cox model and extensions are developed for right-censored observations. More formally spoken, if the true survival time is given by T and C is a censoring time, only $\tilde{T} = \min(T, C)$ is observed along with the censoring indicator $\delta = \mathbb{1}_{(T \leq C)}$. Many applications, however, confront the analyst with more complicated data structures involving more general censoring schemes. For example, interval censored survival times T are not observed exactly but are only known to fall into an interval $[T_{lo}, T_{up}]$. If $T_{lo} = 0$ such survival times are also referred to as being left censored. Furthermore, each of the censoring schemes may appear in combination with left truncation of the corresponding observation, i.e. the survival time is only observed if it exceeds the truncation time T_{tr}. Accordingly, some survival times are not observable and the likelihood has to be adjusted appropriately. Figure 2 illustrates the different censoring schemes we will consider in the following: The true survival time is given by T which is observed for individuals 1 and 2. While individual 1 is not truncated, individual 2 is left truncated at time T_{tr}. Similarly, individuals 3 and 4 are right-censored at time C and individuals 5 and 6 are interval censored with interval $[T_{lo}, T_{up}]$ and the same pattern for left truncation.

In a general framework an observation can now be uniquely described by the quadruple $(T_{tr}, T_{lo}, T_{up}, \delta)$, with

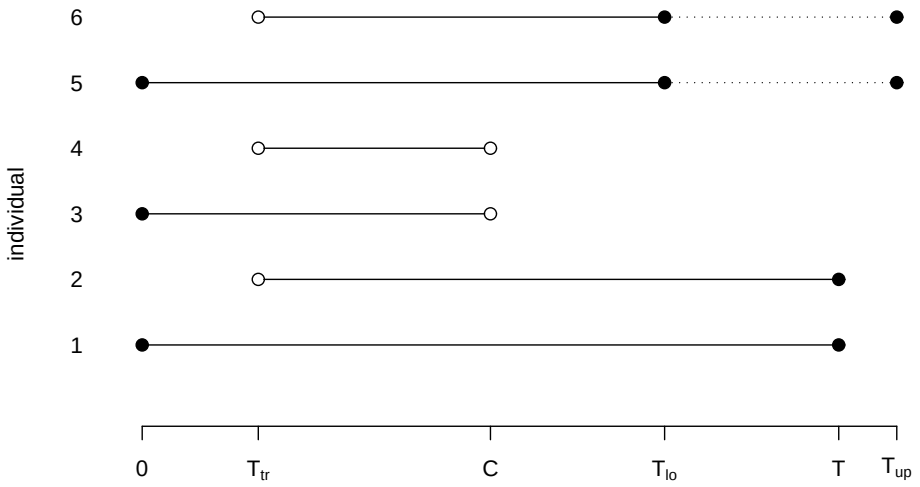


Figure 2: Illustration of different censoring schemes.

$T_{lo} = T_{up} = T$, $\delta = 1$	if the observation is uncensored,
$T_{lo} = T_{up} = C$, $\delta = 0$	if the observation is right censored,
$T_{lo} < T_{up}$, $\delta = 0$	if the observation is interval censored.

For left truncated observations we have $T_{tr} > 0$ while $T_{tr} = 0$ for observations which are not truncated.

Based on these definitions we can now construct the likelihood contributions for the different censoring schemes in terms of the hazard rate $\lambda(t)$ and the survivor function $S(t) = \exp(\int_0^t \lambda(u)du)$. Under the common assumption of noninformative censoring and conditional independence, the likelihood is given by

$$L = \prod_{i=1}^{n} L_i, \tag{47}$$

where

$$L_i = \lambda(T_{up})S(T_{up})/S(T_{tr}) = \lambda(T_{up}) \exp\left(-\int_{T_{tr}}^{T_{up}} \lambda(t)dt\right)$$

for an uncensored observation,

$$L_i = S(T_{up})/S(T_{tr}) = \exp\left(-\int_{T_{tr}}^{T_{up}} \lambda(t)dt\right)$$

for a right censored observation and

$$L_i = (S(T_{lo}) - S(T_{up}))/S(T_{tr}) = \exp\left(-\int_{T_{tr}}^{T_{lo}} \lambda(t)dt\right)\left(1 - \exp\left(-\int_{T_{lo}}^{T_{up}} \lambda(t)dt\right)\right)$$

for an interval censored observation. Note that for explicit evaluation of the likelihood (47) some numerical integration technique has to be employed, since none of the integrals can in general be solved analytically.

The above notation also allows for the easy inclusion of piecewise constant, time-varying covariates via some data augmentation. Noting that

$$\int_{T_{tr}}^{T} \lambda(t)dt = \int_{T_{tr}}^{t_1} \lambda(t)dt + \int_{t_1}^{t_2} \lambda(t)dt + \ldots + \int_{t_{p-1}}^{t_p} \lambda(t)dt + \int_{t_p}^{T} \lambda(t)dt$$

for $T_{tr} < t_1 < \ldots < t_q < T$, we can replace an observation $(T_{tr}, T_{lo}, T_{up}, \delta)$ by a set of new observations $(T_{tr}, t_1, t_1, 0)$, $(t_1, t_2, t_2, 0)$, $\ldots (t_{p-1}, t_p, t_p, 0)$, $(t_p, T_{lo}, T_{up}, \delta)$ without changing the likelihood. Therefore, observations with time-varying covariates can be split up into several observations, where the values $t_1 < \ldots < t_p$ are defined by the changepoints of the covariate and the covariate is now time-constant on each of the intervals. In theory, other paths for a covariate $x(t)$ than piecewise constant ones are also possible, if $x(t)$ is known for $T_{tr} \leq t \leq T_{lo}$. In this case the the likelihood (47) can also be evaluated numerically but a general path $x(t)$ may lead to complicated data structures.

Figure 3 illustrates the data augmentation step for a left truncated, uncensored observation and a covariate $x(t)$ that takes the three different values x_1, x_2 and x_3 on the three intervals $[T_{tr}, t_1], [t_1, t_2]$ and $[t_2, T_{up}]$. Here, the original observation $(T_{tr}, T_{up}, T_{up}, 1)$ has to be replaced by $(T_{tr}, t_1, t_1, 0)$, $(t_1, t_2, t_2, 0)$ and $(t_2, T_{up}, T_{up}, 1)$.

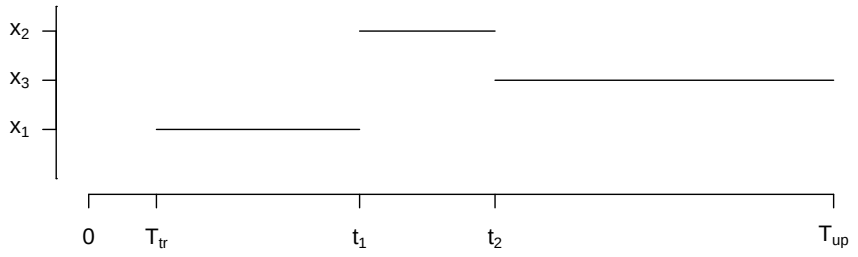


Figure 3: Illustration of time-varying covariates.

Currently, interval censored survival times can only be handled with *remlreg objects*.

7.4 Continuous-time multi-state models

Multi-state models are a flexible tool for the analysis of time-continuous phenomena that can be characterized by a discrete set of states. Such data structures naturally arise when observing a discrete response variable for several individuals or objects over time. Some common examples are depicted in Figure 4 in terms of their reachability graph for illustration. For recurrent events (Figure 4 (a)), the observations evolve through time moving repeatedly between a fixed set of states. Other model classes involve absorbing states, for example disease progression models (Figure 4 (b)), that are used to describe the chronological development of a certain disease. If the severity of this disease can be grouped into $q-1$ ordered stages of increasing severity, a reasonable model might look like this: Starting from disease state 'j', an individual can only move to contiguous states, i.e. either the disease gets worse and the individual moves to state '$j+1$', or the disease attenuates and the individual moves to state '$j-1$'. In addition, death is included as a further, absorbing state 'q', which can be reached from any of the disease states. A model with several absorbing states is the competing risks model (Figure 4 (c)) where, for example, different causes of death are analysed simultaneously.

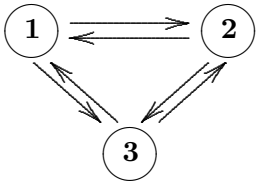
A multi-state model is fully described by a set of hazard rates $\lambda_{hi}(t)$ where h, $h = 1, \ldots, k$, indexes the type of the transition and i, $i = 1, \ldots, n$, indexes the individuals. Since the hazard rates describe durations between transitions, we specify them in analogy to hazard rate models for continuous time survival analysis. To be more specific, $\lambda_{hi}(t)$ is modelled in a multiplicative Cox-type way as

$$\lambda_{hi}(t) = \exp(\eta_{hi}(t)),$$

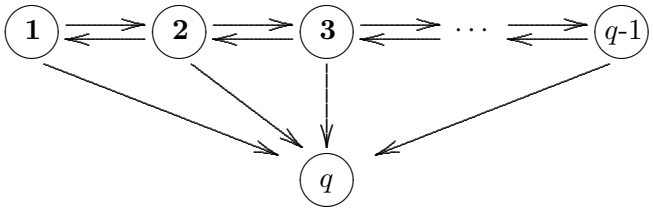
where

$$\eta_{hi}(t) = g_{h0}(t) + \sum_{l=1}^{L} g_{hl}(t) u_{il}(t) + \sum_{j=1}^{J} f_{hj}(x_{ij}(t)) + v_i(t)'\gamma_h + b_{hi} \tag{48}$$

(a) Recurrent events



(b) Disease progression



(c) Competing risks

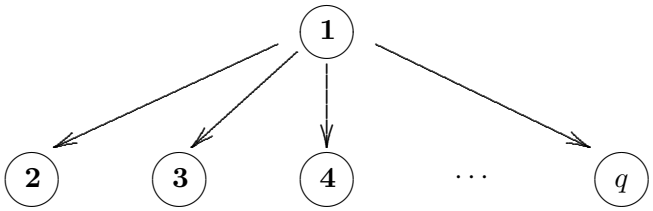


Figure 4: Reachability graphs of some common multi-state models.

is an additive predictor consisting of the following components:

- A time-varying, nonparametric baseline effect $g_{h0}(t)$ common for all observations.
- Covariates $u_{il}(t)$ with time-varying effects $g_{hl}(t)$.
- Nonparametric effects $f_{hj}(x_{ij}(t))$ of continuous covariates $x_{ij}(t)$.
- Parametric effects γ_h of covariates $v_i(t)$.
- Frailty terms b_{hi} to account for unobserved heterogeneity.

For each individual i, $i = 1, \ldots, n$, the likelihood contribution in a multi-state model can be derived from a counting process representation of the multi-state model. Let $N_{hi}(t)$, $h = 1, \ldots, k$ be a set of counting processes counting transitions of type h for individual i. Consequently, $h = 1, \ldots, k$ indexes the observable transitions in the model under consideration and the jumps of the counting processes $N_{hi}(t)$ are defined by the transition times of the corresponding multi-state process for individual i.

From classical counting process theory (see e.g. Andersen et al. (1993), Ch. VII.2), the intensity processes $\alpha_{hi}(t)$ of the counting processes $N_{hi}(t)$ are defined as the product of the hazard rate for type h transitions $\lambda_{hi}(t)$ and a predictable at-risk indicator process $Y_{hi}(t)$, i.e.

$$\alpha_{hi}(t) = Y_{hi}(t)\lambda_{hi}(t),$$

where the hazard rates are constructed in terms of covariates as in (48). The at-risk indicator $Y_{hi}(t)$ takes the value one if individual i is at risk for a type h transition at time t and zero otherwise. For example, in the multi-state model of Figure 4a), an individual in state 2 is at risk for both transitions to state 1 and state 3. Hence, the at-risk indicators for both the transitions '2 to 1' and '2 to 3' will be equal to one as long as the individual remains in state 2.

Under mild regularity conditions, the individual log-likelihood contributions can now be obtained from counting process theory as

$$l_i = \sum_{h=1}^{k} \left[\int_0^{T_i} \log(\lambda_{hi}(t)) dN_{hi}(t) - \int_0^{T_i} \lambda_{hi}(t) Y_{hi}(t) dt \right], \tag{49}$$

where T_i denotes the time until which individual i has been observed. The likelihood contributions can be interpreted similarly as with hazard rate models for survival times (and in fact coincide with these in the case of a multi-state process with only one transition to an absorbing state). The first term corresponds to contributions at the transition times since the integral with respect to the counting process in fact equals a simple sum over the transition times. Each of the summands is then given by the log-intensity for the observed transition evaluated at this particular time point. In survival models this term simply equals the log-hazard evaluated at the survival time for uncensored observations. The second term reflects cumulative intensities integrated over accordant waiting periods between two successive transitions. The integral is evaluated for all transitions the corresponding person is at risk at during the current period. In survival models there is only one such transition (the transition from 'alive' to 'dead') and the integral is evaluated from the time of entrance to the study to the survival or censoring time.

More details on multi-state models, including an exemplary analysis on human sleep, can be found in Kneib & Hennerfeind (2006).

8 Multilevel structured additive distributional and quantile regression

8.1 Distributional regression

Structured additive regression models assume that the distribution of the response variable y, given covariates $\mathbf{x}$ and $\mathbf{u}$, belongs to an exponential family. The conditional mean $\mu_i = E(y_i|\mathbf{x}, \mathbf{u})$ is linked to a structured additive predictor

$$\eta_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \mathbf{u}_i'\boldsymbol{\gamma}, \quad i = 1, \ldots, n,$$

by $\mu_i = h(\eta_i)$, see Chapter 2. In matrix notation we obtain for the predictor

$$\boldsymbol{\eta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{U}\boldsymbol{\gamma},$$

see again Chapter 2 for details. For the most basic model with Gaussian responses and the identity response function we have

$$y_i \sim \mathcal{N}(\eta_i, \sigma^2) = \mathcal{N}(f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \mathbf{u}_i'\boldsymbol{\gamma}, \sigma^2)$$

or

$$\mathbf{y} = \mathcal{N}(\boldsymbol{\eta}, \sigma^2\mathbf{I}) = \mathcal{N}(\mathbf{X}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{U}\boldsymbol{\gamma}, \sigma^2\mathbf{I}).$$

Here and in other convential STAR models only the conditional mean is modeled in dependence of covariates.

A more flexible approach is given by distributional regression as introduced in Klein et al. (2014b) and Klein, Kneib & Lang (2014). On the one hand, the class of distributions that can be estimated with distributional regression is no longer restricted to the exponential family. On the other hand, distributional regression allows to model not only the conditional mean of the response variable but the whole set of distribution parameters $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K$, i.e. all K parameters of the response distribution can be related to a set of predictor variables, which of course may vary between the different parameters. Using response functions $h_1, \ldots, h_K$ each of these parameters can be linked to a structured additive predictor via

$$\boldsymbol{\vartheta}_k = h_k(\boldsymbol{\eta}_k) = h_k(\mathbf{X}_{1k}\boldsymbol{\beta}_{1k} + \ldots + \mathbf{X}_{pk}\boldsymbol{\beta}_{pk} + \mathbf{U}_k\boldsymbol{\gamma}_k), \quad k = 1, \ldots, K.$$

Usually, the response functions are chosen to ensure appropriate restrictions on the parameter spaces. We use, for example, the exponential function to ensure positivity of the scale parameter.

Presumably the most simple distributional regression model is obtained with Gaussian responses where both the mean μ and the variance σ^2 is modeled in terms of covariates. Thus, we consider a regression model with

$$\begin{aligned} \boldsymbol{\mu} &= h_1(\boldsymbol{\eta}_1) &&= \boldsymbol{\eta}_1, \\ \boldsymbol{\sigma}^2 &= h_2(\boldsymbol{\eta}_2) &&= \exp(\boldsymbol{\eta}_2). \end{aligned}$$

A possible generalization of the normal distribution is given by the three-parameter Student's t distribution with location parameter μ, scale parameter $\sigma^2 > 0$ and degrees of freedom $\texttt{df} > 0$. The probability density function is given by

$$f(y|\mu, \sigma, df) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sigma\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{df}{2}\right)\sqrt{df}} \cdot \left(1 + \frac{(y-\mu)^2}{\sigma^2 df}\right)^{-\frac{df+1}{2}},$$

where $\Gamma(x) = \int_0^\infty u^{x-1}\exp(-u)\mathrm{d}u$ for $x > 0$ is the gamma function. Similar to the normal distribution the t distribution is symmetric and bell-shaped but it has heavier tails, offering a robust alternative to the normal distribution. For `df` $\to \infty$ it collapses to the normal distribution.

Regarding the positivity of σ^2 and `df` we consider the model

$$\begin{aligned} \boldsymbol{\mu} &= h_1(\boldsymbol{\eta}_1) &&= \boldsymbol{\eta}_1, \\ \boldsymbol{\sigma}^2 &= h_2(\boldsymbol{\eta}_2) &&= \exp(\boldsymbol{\eta}_2), \\ \texttt{df} &= h_3(\boldsymbol{\eta}_3) &&= \exp(\boldsymbol{\eta}_3). \end{aligned}$$

A popular two parameter distribution for modeling skewed distributions is the gamma distribution with mean parameter $\mu > 0$ and shape parameter $\sigma > 0$. The probability density function is given by

$$f(y_i|\mu_i, \sigma_i) = \left(\frac{\sigma_i}{\mu_i}\right)^{\sigma_i} \cdot \frac{y_i^{\sigma_i - 1}}{\Gamma(\sigma_i)} \cdot \exp\left(-\frac{\sigma_i}{\mu_i} \cdot y_i\right),$$

The mean of the gamma distribution corresponds to μ, the variance is given by μ^2/σ. Setting up a regression model both the mean and the shape parameter are linked to a STAR predictor via the exponential function due to the positivity constraints:

$$\begin{aligned} \boldsymbol{\mu} &= h_1(\boldsymbol{\eta}_1) &&= \exp(\boldsymbol{\eta}_1), \\ \boldsymbol{\sigma} &= h_2(\boldsymbol{\eta}_2) &&= \exp(\boldsymbol{\eta}_2). \end{aligned}$$

A comprehensive list of all distributional regression models available in BayesX can be found in Tables 10.6 to 10.9 of the Reference Manual.

8.2 Quantile Regression

Distributional regression assumes a specific parametric probability distribution of the response (like the normal, lognormal or gamma distribution) and models some or all of its parameters in dependence of covariates. Quantile regression, in contrast, is a distribution-free approach, trying to directly model the different quantiles of the response as a function of covariates.

In linear quantile regression (see Koenker (2005)), we assume

$$q_{\varphi,i} = \beta_{\varphi,0} + \beta_{\varphi,1}x_{i1} + \ldots + \beta_{\varphi,p}x_{ip}$$

where q_φ, for $\varphi \in (0,1)$, is the φ-quantile of the response distribution. Estimation of the quantile-specific regression coefficients $\boldsymbol{\beta}_\varphi$ relies on minimizing the asymmetrically weighted error (AWE) criterion

$$\hat{\boldsymbol{\beta}}_\varphi = \text{argmin}_{\boldsymbol{\beta}_\varphi} \left\{ \sum_{i=1}^n \rho_\varphi(y_i - \mathbf{x}_i'\boldsymbol{\beta}_\varphi) \right\}, \tag{50}$$

with the loss function ρ_φ defined by

$$\rho_\varphi(u) = \begin{cases} u\varphi & \text{if } u \geq 0 \\ u(\varphi - 1) & \text{if } u < 0, \end{cases}$$

which is also known as the check function. Since there exists no closed form solution for this minimization problem, estimates are typically obtained based on linear programming and modifications

of the simplex algorithm, see Koenker (2005) for details. The distribution of the response is implicitly determined by the estimated quantiles q_φ provided that quantiles for a reasonable dense grid of φ-values are estimated. Generalizations to structured additive predictors are conceptually straightforward. However, estimation is highly challenging and almost impossible for complex hierarchical models, revealing the limits of frequentist quantile regression.

Bayesian structured quantile regression requires a distributional assumption for the responses to be able to set up a likelihood. Following Waldmann et al. (2013) we will assume independent and identically distributed observations following an asymmetric Laplace distribution with location parameter $\eta_{i,\varphi}$ (specified in the usual structured additive fashion), scale parameter σ^2 and skewness parameter φ,

$$y_i|\eta_{i,\varphi},\sigma^2,\varphi \overset{\text{iid}}{\sim} \text{ALD}(\eta_{i,\varphi},\sigma^2,\varphi).$$

Then, the density of the responses is given by

$$p(y_i|\eta_{i,\varphi},\sigma^2,\varphi) = \frac{\varphi(1-\varphi)}{\sigma^2}\exp\left(-\frac{\rho_\varphi(y_i-\eta_i\varphi)}{\sigma^2}\right).$$

Maximizing the corresponding posterior (for fixed σ^2 and φ) obviously is equivalent to minimizing the AWE criterion (50) in case of a linear predictor. However, in contrast to frequentist quantile regression the linear predictor can be replaced by a hierarchical structured additive predictor without any further difficulties, see Waldmann et al. (2013) for details.

Since the check function ρ_φ is non-differentiable, inference based on Markov chain Monte Carlo (MCMC) simulations at a first glance seems to be complicated. However, the asymmetric Laplace distribution can be represented as a scaled mixture of normals

$$Y_i = \eta_i + \xi W_i + \delta Z_i\sqrt{\sigma^2 W_i}$$

with $\xi = \frac{1-2\varphi}{\varphi(1-\varphi)}$ and $\delta^2 = \frac{2}{\varphi(1-\varphi)}$. $W_i \sim \text{Exp}(\frac{1}{\sigma^2})$ and $Z_i \sim \mathcal{N}(0,1)$ are independent random variables following an exponential distribution with mean σ^2 and a standard normal distribution, respectively. Thus, using offsets ξW_i and weights $\delta\sqrt{\sigma^2 W_i}$ the Bayesian quantile regression problem can be interpreted as a conditionally Gaussian regression model after imputing W_i as a part of the MCMC sampler.

8.3 Multilevel models

Recently Lang et al. (2014) have proposed a multilevel version of STAR models to cope with the hierarchical nature of many data sets. Suppose that covariate $x_j \in \{1,\ldots,K\}$ is a unit- or cluster index and x_{ij} indicates the cluster observation i pertains to. Then the design matrix $\mathbf{X}_j$ is a $n \times K$ incidence matrix with $\mathbf{X}_j[i,k] = 1$ if the i-th observation belongs to cluster k and zero else. The $K \times 1$ parameter vector $\boldsymbol{\beta}_j$ is the vector of regression parameters, i.e. the k-th element in $\boldsymbol{\beta}_j$ corresponds to the regression coefficient of the k-th cluster. We now define a second level equation

$$\boldsymbol{\beta}_j = \boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j = \mathbf{X}_{j1}\boldsymbol{\beta}_{j1} + \ldots + \mathbf{X}_{jp_j}\boldsymbol{\beta}_{jp_j} + \mathbf{U}_j\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j, \tag{51}$$

where the terms $\mathbf{X}_{j1}\boldsymbol{\beta}_{j1},\ldots,\mathbf{X}_{jp_j}\boldsymbol{\beta}_{jp_j}$ correspond to additional nonlinear functions $f_{j1},\ldots,f_{jp_j}$ and $\mathbf{U}_j\boldsymbol{\gamma}_j$ comprises additional linear effects of cluster level covariates. The "errors" $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0},\tau_j^2\mathbf{I})$ comprise a vector of i.i.d. Gaussian random effects. Using the compound prior (51) we obtain an additive decomposition of the cluster specific 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 $\boldsymbol{\beta}_j = \boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \tau_j^2 \mathbf{I})$ and we obtain a simple i.i.d. Gaussian cluster specific random effect with variance parameter τ_j^2.

A third or fourth level in the hierarchy is possible by assuming that the second or third level regressions contain additional cluster-specific random effects whose parameters are again modeled through STAR predictors of cluster level covariates.

References

Albert, J. & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, 669–679.

Andersen, P. K., Borgan, Ø, Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer Verlag.

Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of Normal Distributions. *Journal of the Royal Statistical Society B*, **36**, 99–102.

Albert, J. & Chib, S. (1993). Simple and Multiple P-Spline Regression with Shape Constraints. *British Journal of Mathematical and Statistical Psychology*, **59**, 451–469.

Belitz, C. (2007). *Model Selection in Generalized Structured Additive Regression Models.* PhD Thesis, University of Munich.

Belitz, C. & Lang, S. (2008). Simultaneous Selection of Variables and Smoothing Parameters in Structured Additive Regression Models. *Computational Statistics and Data Analysis*, **53** , 61-81.

Besag, J., York, J. & Mollié, A. (1991). Bayesian Image Restoration with two Applications in Spatial Statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Brezger, A. & Lang, S. (2006). Generalized Additive Regression based on Bayesian P-Splines. *Computational Statistics and Data Analysis* **50**, 967–991.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* New York: Springer Verlag.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible Smoothing using B-Splines and Penalized Likelihood (with comments and rejoinder). *Statistical Science*, **11**, 89–121.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized Structured Additive regression for Space-Time Data: A Bayesian Perspective. *Statistica Sinica*, **14**, 715–745.

013)fahkne13 Fahrmeir, L., Kneib, T., Lang, S. & Marx, B.2013 *Regression: Models, Methods and Applications.* New York: Springer-Verlag.

Fahrmeir, L. & 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. & Lang, S. (2001b). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics*, **53**, 10–30.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models.* New York: Springer-Verlag.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. E. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* Chichester: Wiley.

Gamerman, D. (1997). Efficient Sampling from the Posterior Distribution in Generalized Linear Models. *Statistics and Computing*, **7**, 57–68.

Gelfand, A. E., Sahu, S. K. & Carlin, B. P. (1996). Efficient Parametrizations for Generalized Linear Mixed Models. In: Bernardo, J. M., Berger, J. O., Dawid, A. P. & Smith, A. F. M. (eds.), *Bayesian Statistics 5*, 165–180. Oxford University Press.

George, A. & Liu, J.W. (1981). *Computer Solution of Large Sparse Positive Definite Systems.* Series in computational mathematics, Prentice-Hall.

Green, P. J. (1987). Penalized Likelihood for General Semiparametric Regression Models. *International Statistical Review*, **55**, 245–259.

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

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

Griffin, J. E., and Brown, P. J. (2005). Alternative Prior Distributions for Variable Selection with very many more Variables than Observations. Technical report, University of Warwick, Dept. of Statistics.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to related Problems. *Journal of the American Statistical Association*, **72**, 320–338.

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

Hastie, T. & Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society B*, **55**, 757–796.

Hastie, T. & Tibshirani, R. (2000). Bayesian Backfitting. *Statistical Science*, **15**, 193–223.

Hastie, T., Tisbshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer-Verlag.

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive Survival Models. *Journal of the American Statistical Association*, **101**, 1065–1075.

Holmes, C., Held, L. (2006). Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis*, **1**, 145–168.

Hurvich, C. M., Simonoff, J. S. & Tsai, C. L. (1998). Smoothing Parameter Selection in Nonparametric Regression using an improved Akaike Information Criterion. *Journal of the Royal Statistical Society B*, **60**, 271–293.

Ishwaran, H., and Rao, S. J. (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics*, **33**, 730-773.

Johnson, M.E., Moore, L.M. & Ylvisaker, D. (1990). Minimax and Maximin Designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Kammann, E. E. & Wand, M. P. (2003). Geoadditive Models. *Journal of the Royal Statistical Society C*, **52**, 1–18.

Klein, N., Denuit, M., Kneib, T. & Lang, S.(2014). Nonlife Ratemaking and Risk Management with Bayesian additive Models for Location, Scale and Shape. *Insurance: Mathematics and Economics*, **55**, 225–249.

Klein, N., Kneib, T., Lang, S.(2013). Bayesian Structured Additive Distributional Regression. *Under revision for Annals of Applied Statistics.*

Klein, N., Kneib, T. & Lang, S.(2014). Bayesian Generalized Additive Models for Location, Scale and Shape for Zero-Inflated and Overdispersed Count Data. To appear in *Journal of the American Statistical Association*, doi:10.1080/01621459.2014.912955.

Klein, N., Kneib, T., Klasen, S. & Lang, S.(2014). Bayesian Structured Additive Distributional Regression for Multivariate Responses. To appear in *Journal of the Royal Statistical Society C*, doi:10.1111/rssc.12090.

Koenker, R.(2005). *Quantile Regression.* Cambridge University Press, New York.

Kneib, T. (2006). Geoadditive Hazard Regression for Interval Censored Survival Times. *Computational Statistics and Data Analysis*, **51**, 777–792

Kneib, T. & Hennerfeind, A. (2006). Bayesian Semiparametric Multi-State Models. *Statistical Modelling*, **8**, 169–198.

Kneib, T. & Fahrmeir, L. (2006). Structured Additive Regression for Categorical Space-Time Data: A Mixed Model approach. *Biometrics*, **62**, 109–118.

Kneib, T. & Fahrmeir, L. (2007). A Mixed Model Approach to Structured Hazard Regression. *Scandinavian Journal of Statistics*, **34**, 207–228.

Kneib, T., Konrath, S. und Fahrmeir, L. (2009). High-dimensional Structured Additive Regression Models: Bayesian Regularisation, Smoothing and Predictive Performance. Department of Statistics, Technical Report No. 46, LMU Munich.

Knorr-Held, L. (1999). Conditional Prior Proposals in Dynamic Models. *Scandinavian Journal of Statistics*, **26**, 129–144.

Konrath, S., Kneib, T., Fahrmeir, L. (2008). Bayesian Regularisation in Structured Additive Regression Models for Survival Data. Department of Statistics, Technical Report No.35, LMU Munich.

Lang, S. & Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K. & Kneib, T. (2014). Multilevel Structured Additive Regression. *Statistics and Computing*, **24**, 223–238.

Lin, X. & Zhang, D. (1999). Inference in Generalized Additive Mixed Models by using Smoothing Splines. *Journal of the Royal Statistical Society B*, **61**, 381–400.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

Müller, H. G., Stadtmüller, U. & Tabnak, F. (1997). Spatial Smoothing of Geographically Aggregated Data, with Applications to the Construction of Incidence Maps. *Journal of the American Statistical Association* **92**, 61–71.

Nychka, D. & Saltzman, N. (1998). *Design of Air-Quality Monitoring Networks.* Lecture Notes in Statistics, 132, 51–76.

Osuna, L. (2004) *Semiparametric Bayesian Count Data Models.* Dr. Hut Verlag, München.

Park, T., and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, **482**, 681-686.

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. & Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society B*, **65**, 583–639.

Waldmann, E. and Kneib, T. and Lang, S. and Yue, Y.(2013) Bayesian Semiparametric Additive Quantile Regression. *Statistical Modelling*, **13**, 223–252.

Index