Day 2: Bayesian Additive Regression

Exercise 1 (Childhood Malnutrition)

Childhood malnutrition is one of the most challenging problems in developing and transition countries. The World Health Organization (WHO) conducts representative household surveys (demographic and health surveys) in developing countries on a regular basis. The primary goal of the statistical analysis is to determine the effect of certain socioeconomic variables of the child, the mother, and the household on the child's nutritional condition. In this exercise, we look at an exemplary profile of a data set for Zambia. The data set `zambia.dat` contains information reflecting the malnutrition status of children (variable `zscore`, measuring chronic malnutrition, i.e. stunting) along with a set of explanatory variables, in particular `bmi` (body mass index of the mother), `age` (age of the child), `survey` (year of the survey) and `district` (district where the mother lives). The corresponding spatial information is given in the boundary file `zambia.bnd` (in terms of polygons for the district boundaries) and the graph file `zambia.gra` (in terms of the neighborhood structure defined by common boundaries).

Variable	Explanation
`z`	child's Z-score
`age`	child's age in months
`bmi`	mother's body mass index
`survey`	year of the survey
`district`	district of residence in Zambia (55 districts)

We will use the R package `bamlss` for our analyses, but there are various other packages that support additive regression models, such `mgcv`, `vgam`, `gjrm`, etc.

- Store both the data set and the map in appropriate `R` objects. For the map, you can use the functions `BayesX::read.bnd` and `read.gra`.
- Visually explore the relation of the available covariates with the child's Z-score.
- Start by constructing an additive model with nonlinear effects of `bmi` and `age`. To use P-splines in the specification, use function `s(x, bs="ps")`. The dimension of the basis is defined by option `k`, the degree of the splines and the difference penalty are jointly specified by option `m`.
- Add a spatial effect using `s(d, bs="mrf")` or a random effect using `s(d, bs="re")` to investigate whether unobserved spatial heterogeneity remains after adjusting for the covariates.
- Investigate whether there are differences in the age effects between the three surveys using a varying coefficient model. The interaction variable `int` of a varying coefficients term can be specified by adding the option `by=int`, e.g. `s(x, by=int)`.

Exercise 2 (Forest Health Data)

The data set `foresthealth.dat` contains the following information on the forest health status of beeches at 83 observation plots in a northern Bavarian forest district collected in yearly visual forest health inventories between 1983 and 2004:

Variable	Explanation
`year`	calendar time in years
`x`, `y`	coordinates of the observation plots
`id`	id for the observation plot
`inclination`	inclination of slope in percent
`elevation`	elevation above sea level in meters
`soil`	depth of soil layer in centimeters
`fertilisation`	application of fertilisation (1=yes, -1=no)
`age`	average age of the stand in years
`canopy`	density of forest canopy in percent
`stand`	type of stand (1=deciduous forest, -1=mixed forest)
`def`	binary response for defoliation (1=yes, 0=no)

- Estimate a logistic additive regression model for defoliation as the response variable and nonlinear effects of age and canopy density.
- Change the link function to a probit model.
- Add a random effect for the plot id.
- Add a spatial effect based on the coordinates of the observation plots
- Try to improve your model by adding some further explanatory variables.
- Liesel bonus exercises: Make changes to the model and inference algorithm.
  - Use one joint HMC sampler for all parameters instead of blocked sampling.
  - Swap out the priors for the inverse smoothing parameter τ^2 of the spatial effect with an InverseGamma$(0.01, 0.01)$ prior and set up a Gibbs sampler for it.
  - Transform the inverse smoothing parameter τ^2 of your nonlinear function for `age` to the real line using `lsl.Var.transform(tfb.Exp())`, and sample it with an HMC sampler.