Regularising Geoadditive Regression Models

Thomas Kneib

Department of Statistics
Ludwig-Maximilians-University Munich

18.1.2008

Outline

- Geoadditive Regression: Models and Applications
  (with Ludwig Fahrmeir & Stefan Lang)

- Regularisation Priors
  (with Ludwig Fahrmeir, Susanne Konrath & Fabian Scheipl)

- Model Choice and Variable Selection in Geoadditive Regression Models
  (with Torsten Hothorn & Gerhard Tutz)

Childhood Malnutrition in Zambia

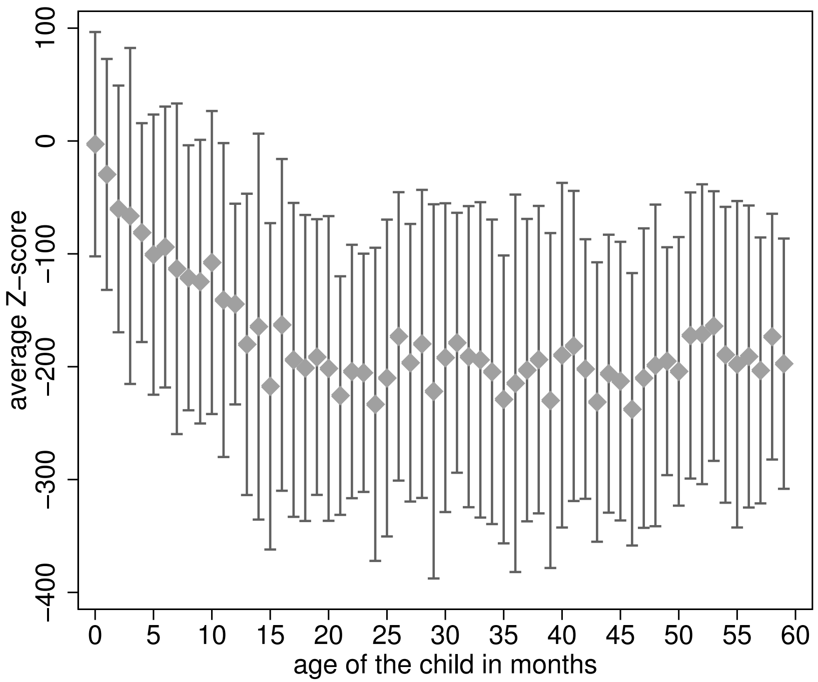
- Data obtained from MEASURE Demographic and Health Surveys (DHS).
- Conducted more than 200 surveys in 75 countries to advance global understanding of health and population trends in developing countries.
- Nationally representative data on fertility, family planning, maternal and child health, as well as child survival, HIV/AIDS, malaria, and nutrition.
- In the following: Z-score for chronic undernutrition (insufficient height for age, stunting) in Zambia:

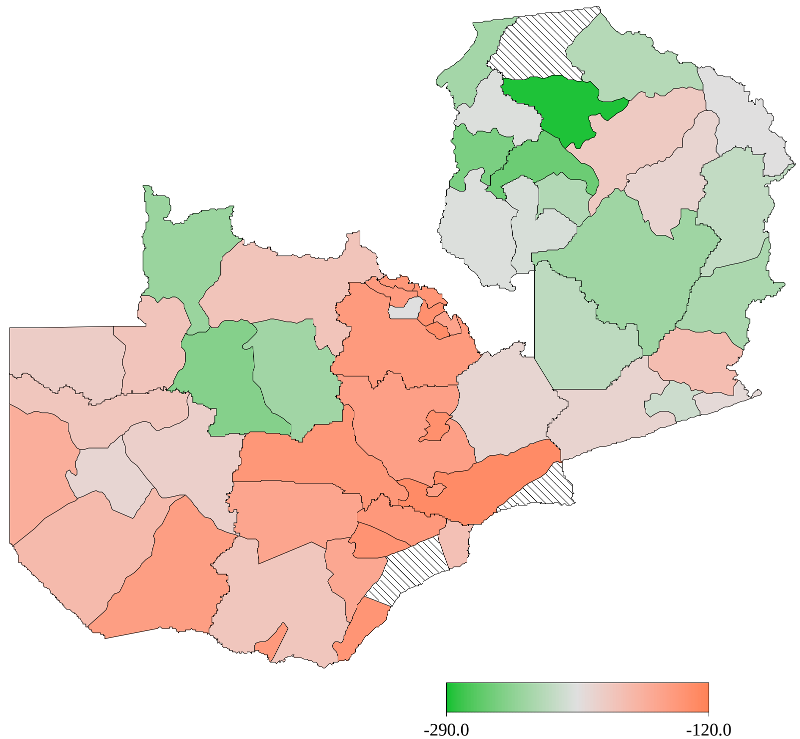
$$Z_i = \frac{\text{height}_i - \text{median height}}{\text{standard deviation}}$$

- Median and standard deviation are obtained from a reference population.

- The Z-score shall be related to covariates including age of the child, duration of breastfeeding, age of the mother at birth, body mass index of the mother, etc.

- Descriptive analyses hint at the presence of nonlinear and spatial effects in the data.





⇒ Usual linear models are not appropriate.

- Replace the linear model by a geoadditive model

$$Z = f_1(agec, bf) + f_2(agem) + f_3(height) + f_4(bmi) + f_{\mathsf{spat}}(region) + u'\gamma + \varepsilon.$$

where

– $f_1(agec, bf)$ is an interaction effect between age of the child and duration of breastfeeding,

– f_2, f_3, f_4 are nonlinear effects of the age, height and body mass index of the mother,

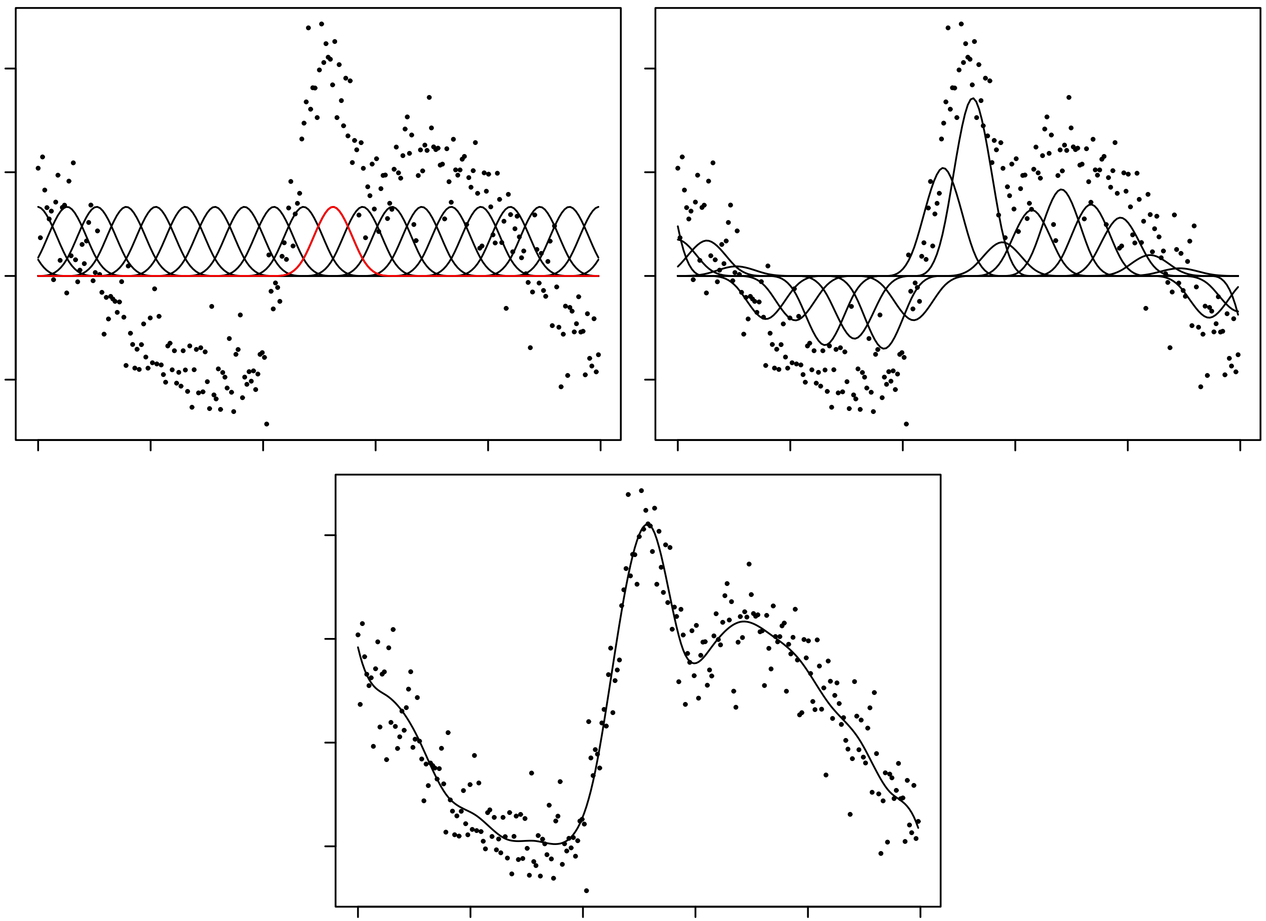
– f_{spat} is a spatial effect, and

– $u'\gamma$ is a linear predictor capturing parametric effects (of categorical covariates).

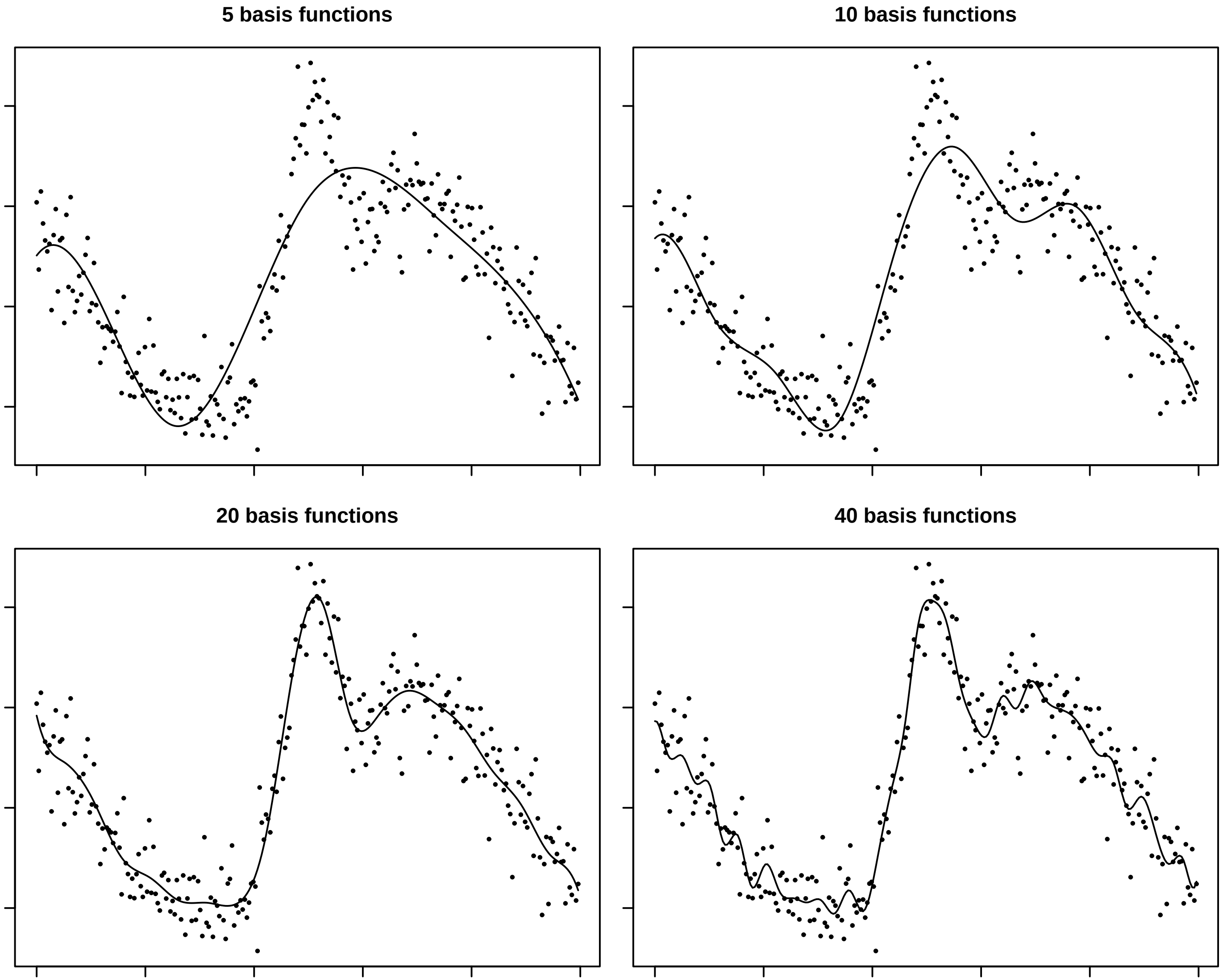
Model Components and Priors

- **Smooth model components:** Approximate a function $f(x)$ by a linear combination of B-spline basis functions

$$f(x) = \sum_j \beta_j B_j(x)$$



- B-spline fit for different numbers of knots:



- Unconstrained estimation crucially depends on the number of basis functions.

  $\Rightarrow$ Add a regularisation term to the likelihood that enforces smoothness.

- Popular approach: Squared derivative penalty, e.g.

$$\text{pen}(f) = \lambda \int (f''(x))^2 dx$$

- Easy approximation for B-splines: Difference penalties, e.g.

$$\text{pen}(\beta) = \lambda \sum_j (\beta_j - \beta_{j-1})^2 = \lambda \beta' K \beta$$

- Smoothing parameter λ governs the impact of the penalty (should be estimated).

- Corresponds to random walk prior in a Bayesian setting:

$$\beta_j = \beta_{j-1} + u_j, \qquad u_j \sim N(0, \tau^2).$$

- **Spatial effects:** Estimate a separate parameter β_s for each region.

- Estimation becomes unstable if the number of regions is large relative to the sample size.

  $\Rightarrow$ Regularised estimation to enforce spatial smoothness.

- Effects of neighboring regions (common boundary) should be similar.

- Define a penalty term based on differences between neighboring parameters:

$$\text{pen}(\beta) = \lambda \sum_s \sum_{r \in N(s)} (\beta_s - \beta_r)^2$$

  where $N(s)$ denotes the set of neighbors of region s.

- In a stochastic formulation equivalent to a Markov random field prior.

Bayesian Inference

- Unifying framework:
  - All vectors of function evaluations can be written as the product of a design matrix X_j and a vector of regression coefficients β_j, i.e. $f_j = X_j\beta_j$.
  - Regularisation penalties are quadratic forms $\lambda_j\beta_j'K_j\beta_j$ corresponding to Gaussian priors

$$p(\beta|\tau^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j'K_j\beta_j\right).$$

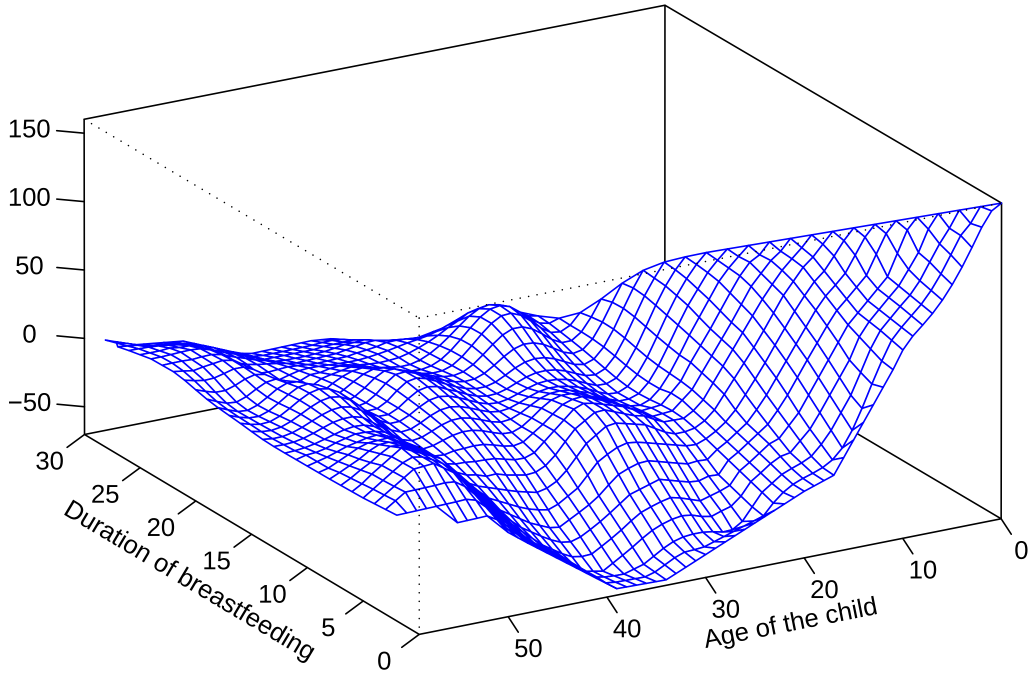
  - The variance τ_j^2 is a transformation of the smoothing parameter λ_j.
  - In many cases, the penalty matrix K_j is rank-deficient.

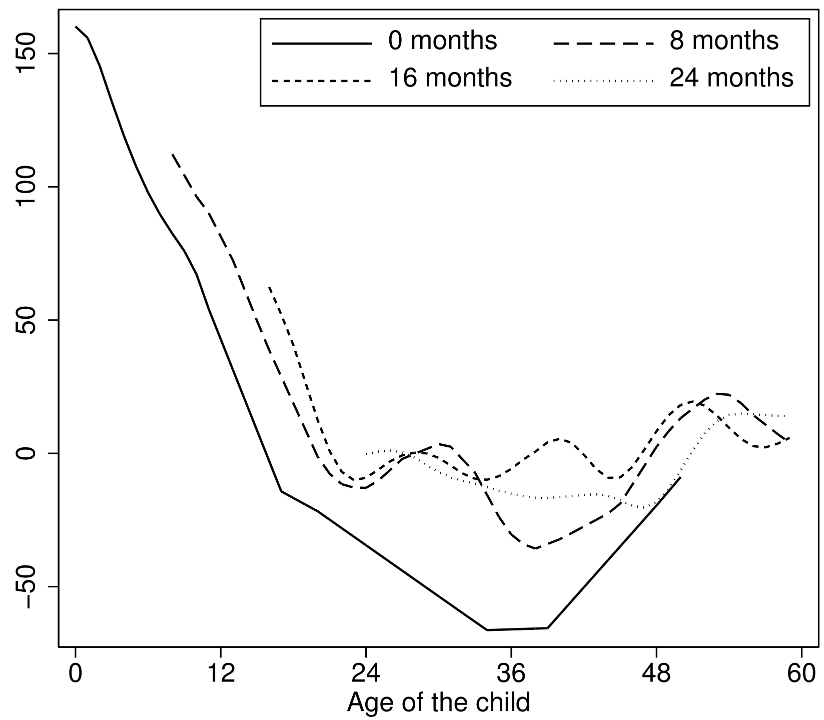
- The unifying framework allows to devise equally general inferential procedures.

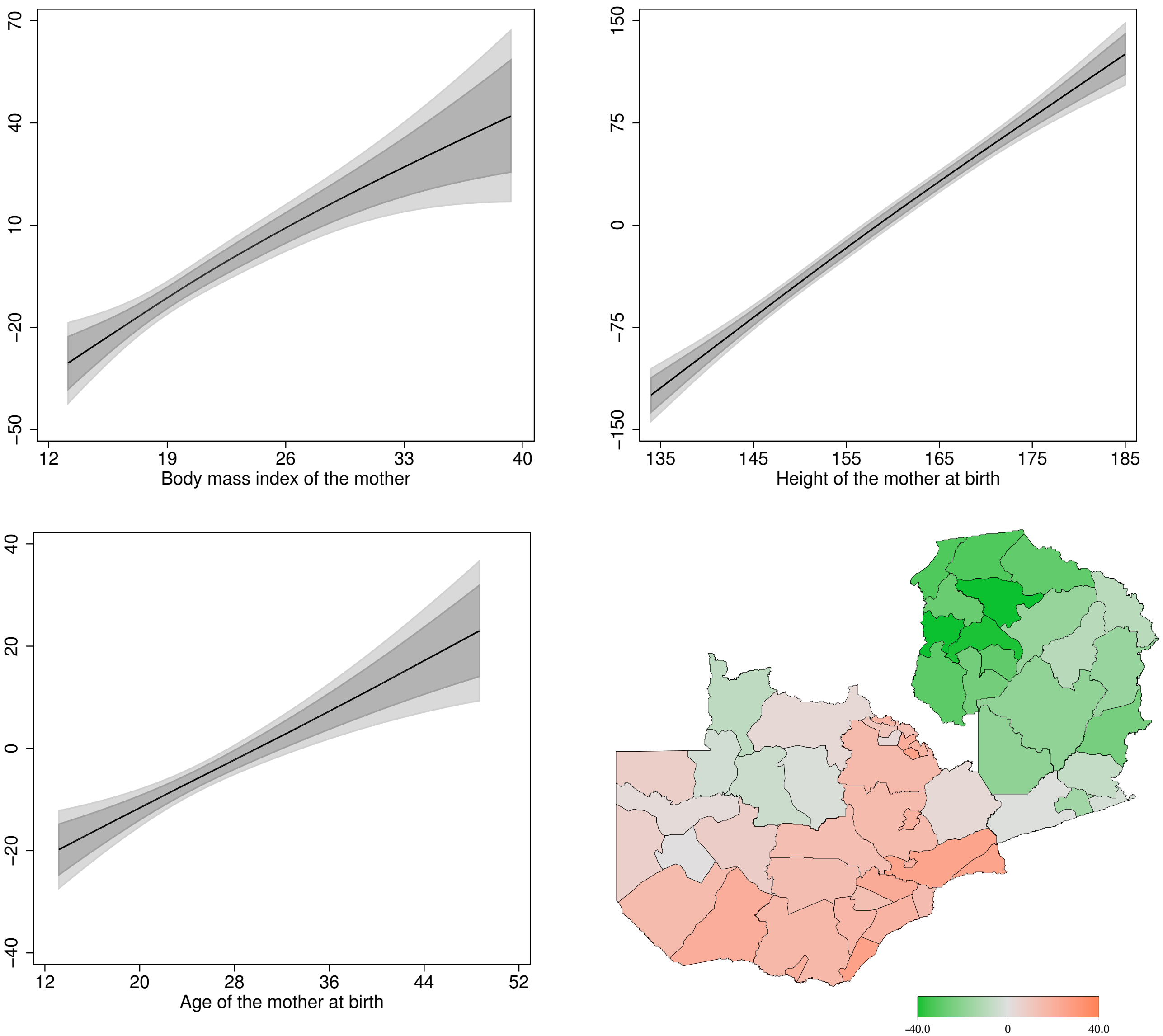
- Mixed model based empirical Bayes inference:
  - Consider the variances / smoothing parameters as unknown constants to be estimated by mixed model methodology.
  - Decompose the vector of regression coefficients into (unpenalised) fixed effects and (penalised) random effects.
  - Penalised likelihood estimation of the regression coefficients in the mixed model (posterior modes).
  - Marginal likelihood estimation of the variance and smoothing parameters (Laplace approximation).

- Fully Bayesian inference based on Markov Chain Monte Carlo simulation techniques:
  - Assign inverse gamma priors to the variance / smoothing parameters.
  - Metropolis-Hastings update for the regression coefficients (based on iteratively weighted least squares-proposals).
  - Gibbs sampler for the variances (inverse gamma with updated parameters).

Results


150
100
50
0
−50
30
25
20
15
10
5
0
Duration of breastfeeding
50
40
30
20
10
0
Age of the child


0 months
8 months
16 months
24 months
150
100
50
0
−50
0
12
24
36
48
60
Age of the child


70
40
10
−20
−50
12
19
26
33
40
Body mass index of the mother
150
75
0
−75
−150
135
145
155
165
175
185
Height of the mother at birth
40
20
0
−20
−40
12
20
28
36
44
52
Age of the mother at birth
-40.0
0
40.0

Bayesian Regularisation Priors

- Regularisation in regression models with a large number of covariates: Enforce sparse models where most of the regression coefficients are (close to) zero.

- Examples: Gene expression data but also social science and economic applications.

- Most well-known approach: Ridge regression.

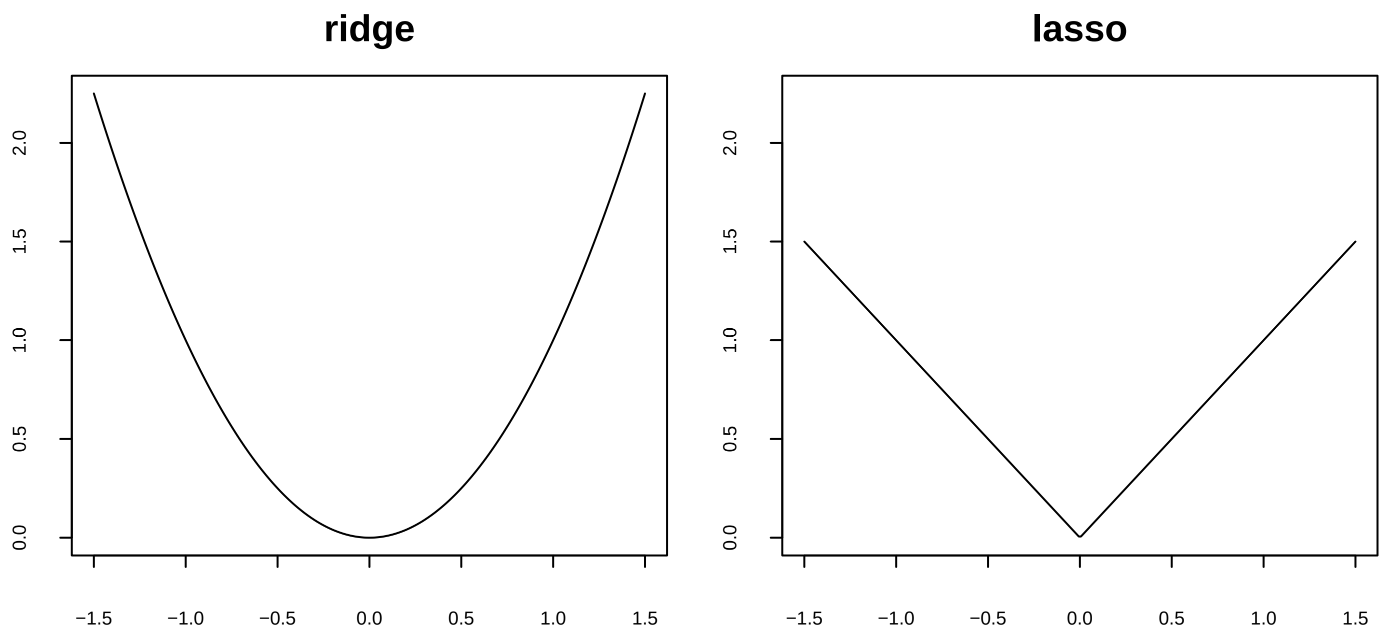
- Add a quadratic penalty to the log-likelihood:

$$l_{\text{pen}}(\beta) = l(\beta) - \lambda \sum_{j=1}^{p} \beta_j^2 \to \max_{\beta}.$$

- Ridge regression fits into the framework of geoadditive regression models but does not induce enough sparsity.

- LASSO penalty: Replace quadratic penalty with absolute value penalty:

$$l_{\text{pen}}(\beta) = l(\beta) - \lambda \sum_{j=1}^{p} |\beta_j| \rightarrow \max_{\beta} .$$



- LASSO imposes more sparsity but the solution is computationally more demanding, in particular in combination with geoadditive regression terms and for non-Gaussian models.

- Ridge and LASSO correspond to prior distributions in a Bayesian interpretation:

Ridge = Gaussian prior

$$p(\beta_j|\lambda) \propto \exp\left(-\lambda\beta_j^2\right)$$

LASSO = Laplace prior

$$p(\beta_j|\lambda) \propto \exp\left(-\lambda|\beta_j|\right)$$

- Convenient feature of the Laplace prior: Can be written as a scale mixture of Gaussians

$$p(\beta_j|\lambda) = \int_0^\infty p(\beta_j|\tau_j^2)p(\tau_j^2|\lambda)d\tau_j^2$$

where

$$\beta_j|\tau_j^2 \sim N(0, \tau_j^2) \qquad \text{and} \qquad \tau_j^2|\lambda \sim Exp\left(\frac{\lambda^2}{2}\right)$$

- Bayesian interpretation: Hierarchical prior formulation.

$$\lambda \longrightarrow \underset{\text{Lap}(\lambda)}{\beta} \qquad \text{vs.} \qquad \lambda \longrightarrow \underset{\text{Exp}(0.5\lambda^2)}{\tau^2} \longrightarrow \underset{\text{N}(0,\tau^2)}{\beta}$$

- Advantage: Estimation based on MCMC recurs to the computationally simpler case of ridge regression with an additional update step for the variances.

  $\Rightarrow$ Update schemes developed in geoadditive regression become available.

- Easily combined with nonparametric or spatial effects.

- Also applicable for non-Gaussian regression models.

- The concept extends to other types of priors that can be written as scale mixture of normals.

Model Choice and Variable Selection in Geoadditive Regression

- Bayesian regularisation priors can be seen as an indirect approach to variable selection for high-dimensional predictors.
- Drawbacks (if model choice and variable selection are of direct interest):
  - Coefficients will be close to zero but not equal to zero.
  - No model choice for spatial effects, nonparametric components, etc.
- Boosting procedures have proven to be a useful (non-Bayesian) tool for model choice and variable selection.
- Principal idea of boosting: Repeated fitting of base-learning procedures to updated negative gradients of a loss function ("residuals").

- Componentwise boosting algorithm for geoadditive regression:
  - Choose a suitable loss function, e.g. the log-likelihood.
  - Define separate base-learners for all model components (possibly even more than one base-learner).
  - Iteratively apply all base-learners in sequence and update only the best-fitting component.
  - Compute updated residuals.

- Boosting implements both variable selection and model choice:
  - Variable selection: Stop the boosting procedure after an appropriate number of iterations (for example based on AIC reduction).
  - Model choice: Consider concurring base-learning procedures for the same covariate, e.g. linear vs. non-linear modeling.

- Base-learning procedures in geoadditive regression: Penalised least squares fits

$$X_j(X_j'X_j + \lambda_j K_j)^{-1}X_j'.$$

  with fixed smoothing parameters λ_j

- Crucial point: Make the base-learners comparable in terms of their complexity (otherwise biased selection results).

- General complexity measures: equivalent degrees of freedom

$$\mathrm{df}(\lambda_j) = \mathrm{trace}(X_j(X_j'X_j + \lambda_j K_j)^{-1}X_j').$$

- Choose the smoothing parameters such that

$$\mathrm{df}(\lambda_j) = 1.$$

- Requires reparameterisation for some effects (e.g. penalised splines).

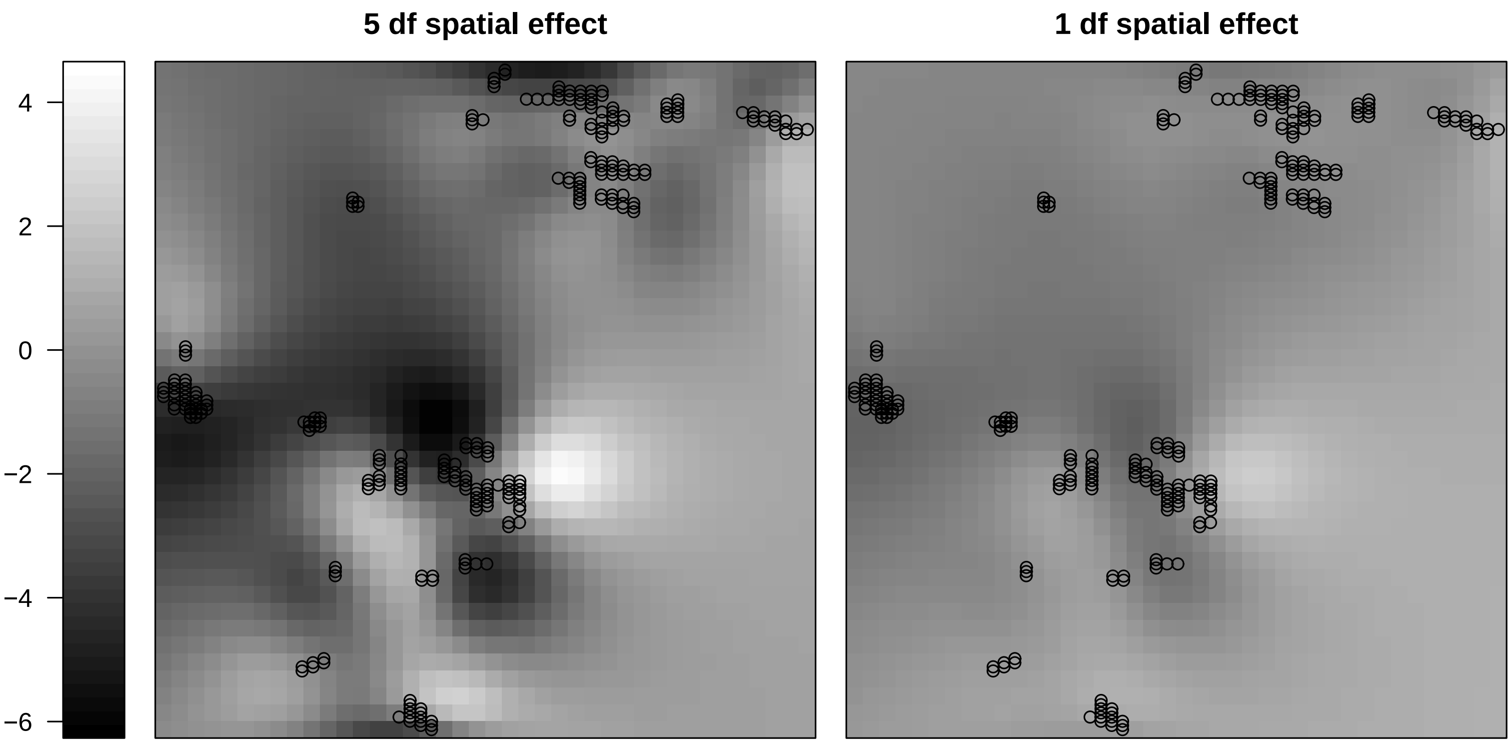
Habitat Suitability Analyses

- Identify factors influencing habitat suitability for breeding bird communities.
- Variable of interest: Counts of subjects from different species collected at 258 observation plots in a Northern Bavarian forest district.
- Research questions:
  a) Which covariates influence habitat suitability (31 covariates in total)? Does spatial correlation have an impact on variable selection?
  b) Are there non-linear effects of some of the covariates?
  c) Are effects varying spatially?
- All questions can be addressed with the boosting approach.
- In the following only results on a).

- Selection frequencies in a spatial Poisson-GLM:

	GST	DBH	AOT	AFS	DWC	LOG	SNA	COO
non-spatial GLM	0	0	0	0.06	0.3	0	0.01	0
spatial with 5 df	0	0.02	0	0.01	0.05	0	0.01	0
spatial with 1 df	0	0	0	0.06	0.15	0	0	0
	COM	CRS	HRS	OAK	COT	PIO	ALA	MAT
non-spatial GLM	0.03	0.04	0.03	0.05	0.06	0	0.04	0.06
spatial with 5 df	0	0.01	0	0	0	0	0.01	0.05
spatial with 1 df	0.03	0.02	0.02	0.04	0.05	0	0.03	0.04
	GAP	AGR	ROA	LCA	SCA	HOT	CTR	RLL
non-spatial GLM	0.03	0	0	0.1	0.07	0	0	0
spatial with 5 df	0.01	0	0.01	0.01	0.01	0	0	0
spatial with 1 df	0.03	0	0	0.07	0.06	0	0	0
	BOL	MSP	MDT	MAD	COL	AGL	SUL	spatial
non-spatial GLM	0	0.06	0	0	0.05	0	0	0
spatial with 5 df	0	0	0	0	0.03	0	0	0.76
spatial with 1 df	0	0.04	0	0	0.04	0	0	0.3

- Spatial effects for high and low degrees of freedom:



- Spatial correlation has non-negligible influence on variable selection.

- Making terms comparable in terms of complexity is essential to obtain valid results.

Summary

- Geoadditive regression is a useful extension of classical regression models.
- Can be adapted to
  - Categorical regression models (Forest health, Brand choice).
  - Survival Modelling (Leukemia, Childhood mortality).
- Variable selection and model choice algorithms are under development.
- Accompanying software exists (BayesX, mboost).
- Bayesian approaches provide full inferential details (measures of uncertainty, credible intervals).
- Boosting algorithms implement model choice and variable selection but provide only point estimates.

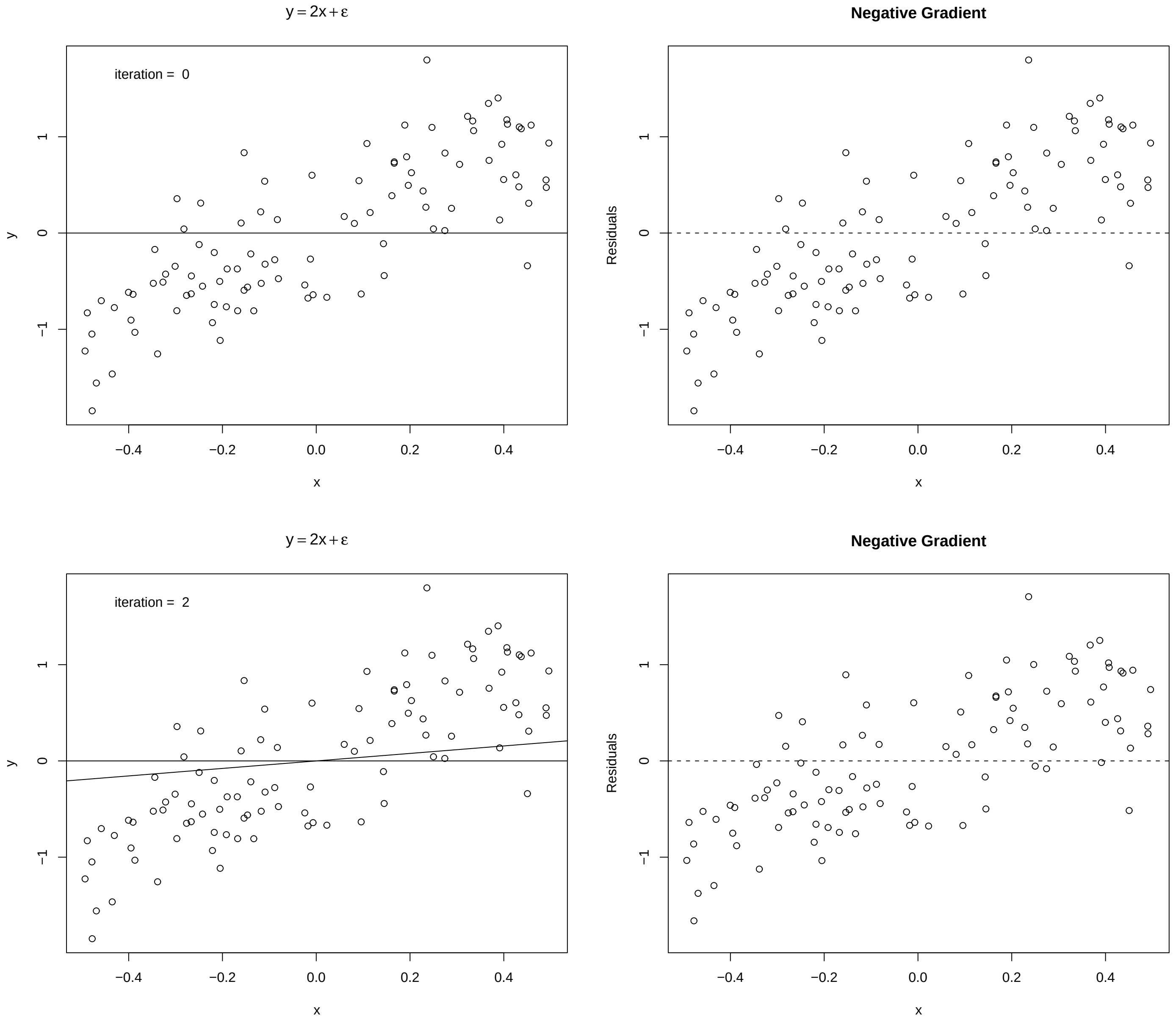
- Some ongoing projects:
  - Measurement error in semiparametric regression models
    (with Ciprian Crainiceanu, Johns-Hopkins University Baltimore; Susanne Breitner, GSF - National Research Center for Environment and Health Munich)
  - Interval censored multi-state models
    (with Martin Daumer, Sylvia Lawry Centre for Multiple Sclerosis Research; Ludwig Fahrmeir, LMU Munich)
  - Geoadditive Analysis of the Determinants of Gender Bias in Mortality in India
    (with Jan Priebe, Georg-August-University Göttingen)
  - Flexible Semiparametric Regression for the Analysis of Human Sleep
    (with Stefanie Kalus & Alexander Yassouridis, Max-Planck-Institute for Psychiatry, Munich)
  - Semiparametric Discrete Choice Models for the Analysis of Consumer Choice Behaviour
    (with Bernhard Baumgartner, University of Regensburg; Winfried Steiner, Clausthal University of Technology)

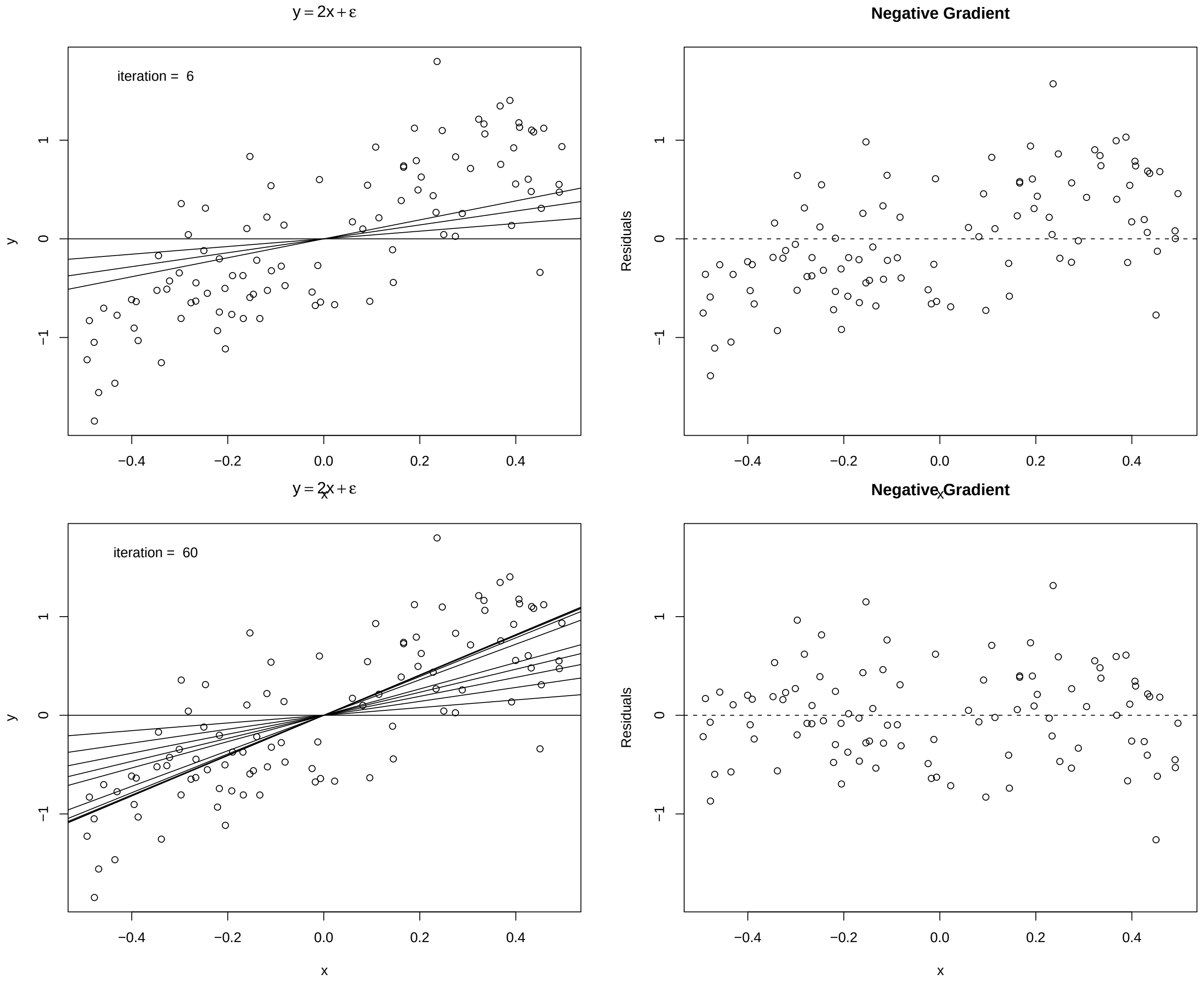
Boosting Example

- Linear model with quadratic loss function $\rho(y, \eta) = |y - \eta|^2$.
  - The gradient of the loss function yields the least squares residuals.
  - Base-learner: Least-squares fit $\hat{g}$.
  - In each iteration, update η via

$$\hat{\eta}^{[m]} = \hat{\eta}^{[m-1]} + 0.1\hat{g}$$

    i.e. multiply the current fit with a reduction factor.


y = 2x + ε
iteration = 0
y
x
Negative Gradient
Residuals
x
y = 2x + ε
iteration = 2
y
x
Negative Gradient
Residuals
x


y = 2x + ε
iteration = 6
Negative Gradient
Residuals
y = 2x + ε
iteration = 60
Negative Gradient
Residuals
x
y