Bayesian Regularisation Priors

Thomas Kneib

Department of Statistics
Ludwig-Maximilians-University Munich

21.1.2008

Outline

- Regularising Geoadditive Regression Models
  (with Ludwig Fahrmeir)

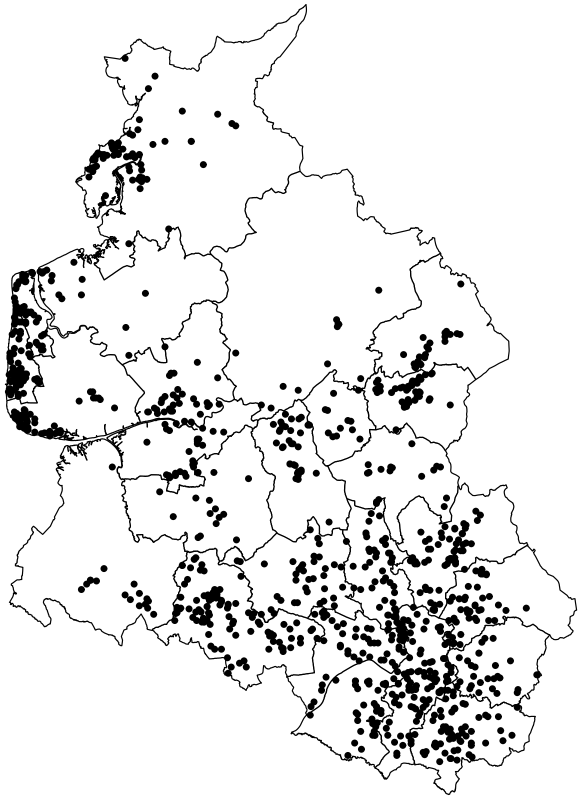
- Regularisation Priors for High-Dimensional Predictors
  (with Ludwig Fahrmeir, Susanne Konrath & Fabian Scheipl)

Leukemia Survival Data

- Survival time of adults after diagnosis of acute myeloid leukemia.
- 1,043 cases diagnosed between 1982 and 1998 in Northwest England.
- 16 % (right) censored.
- Continuous and categorical covariates:

age	age at diagnosis,
wbc	white blood cell count at diagnosis,
sex	sex of the patient,
tpi	Townsend deprivation index.

- Spatial information in different resolution.



- Classical Cox proportional hazards model:

$$\lambda(t;x) = \lambda_0(t)\exp(x'\gamma).$$

- Baseline-hazard $\lambda_0(t)$ is a nuisance parameter and remains unspecified.

- Estimate γ based on the partial likelihood.

- Questions / Limitations:
  - Estimate the baseline simultaneously with covariate effects.
  - Flexible modelling of covariate effects (e.g. nonlinear effects, interactions).
  - Spatially correlated survival times.
  - Non-proportional hazards models / time-varying effects.

$\Rightarrow$ Geoadditive hazard regression models.

Geoadditive hazard regression

- Replace usual parametric predictor with a flexible semiparametric predictor

$$\lambda(t;\cdot) = \lambda_0(t)\exp[f_1(age) + f_2(wbc) + f_3(tpi) + f_{spat}(s_i) + \gamma_1 sex]$$

  and absorb the baseline

$$\lambda(t;\cdot) = \exp[f_0(t) + f_1(age) + f_2(wbc) + f_3(tpi) + f_{spat}(s_i) + \gamma_1 sex]$$

  where

  – $f_0(t) = \log(\lambda_0(t))$ is the log-baseline-hazard,

  – f_1, f_2, f_3 are nonparametric functions of age, white blood cell count and deprivation, and

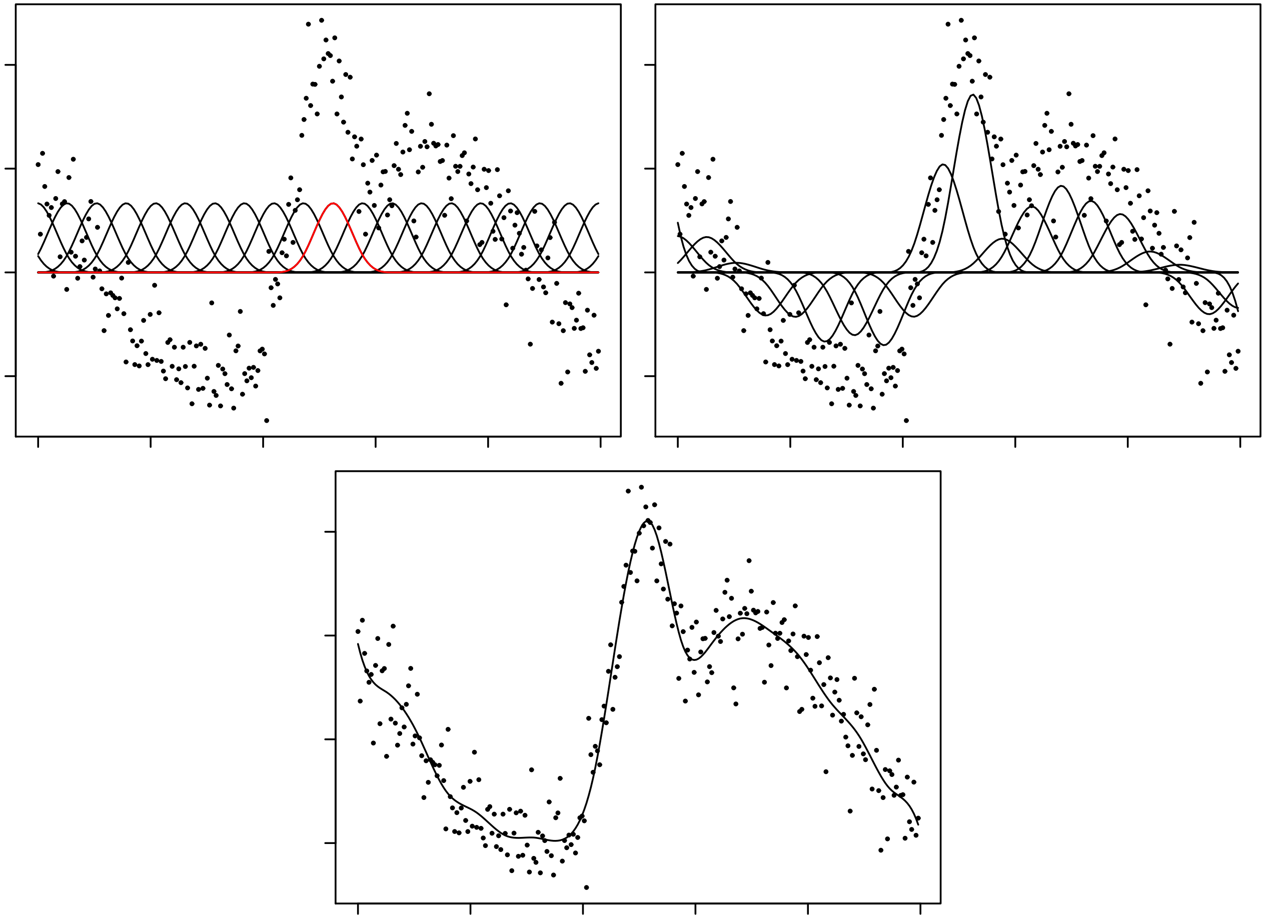
  – f_{spat} is a spatial function.

- Time-varying effects such as $g_1(t)sex$ can be included if needed.

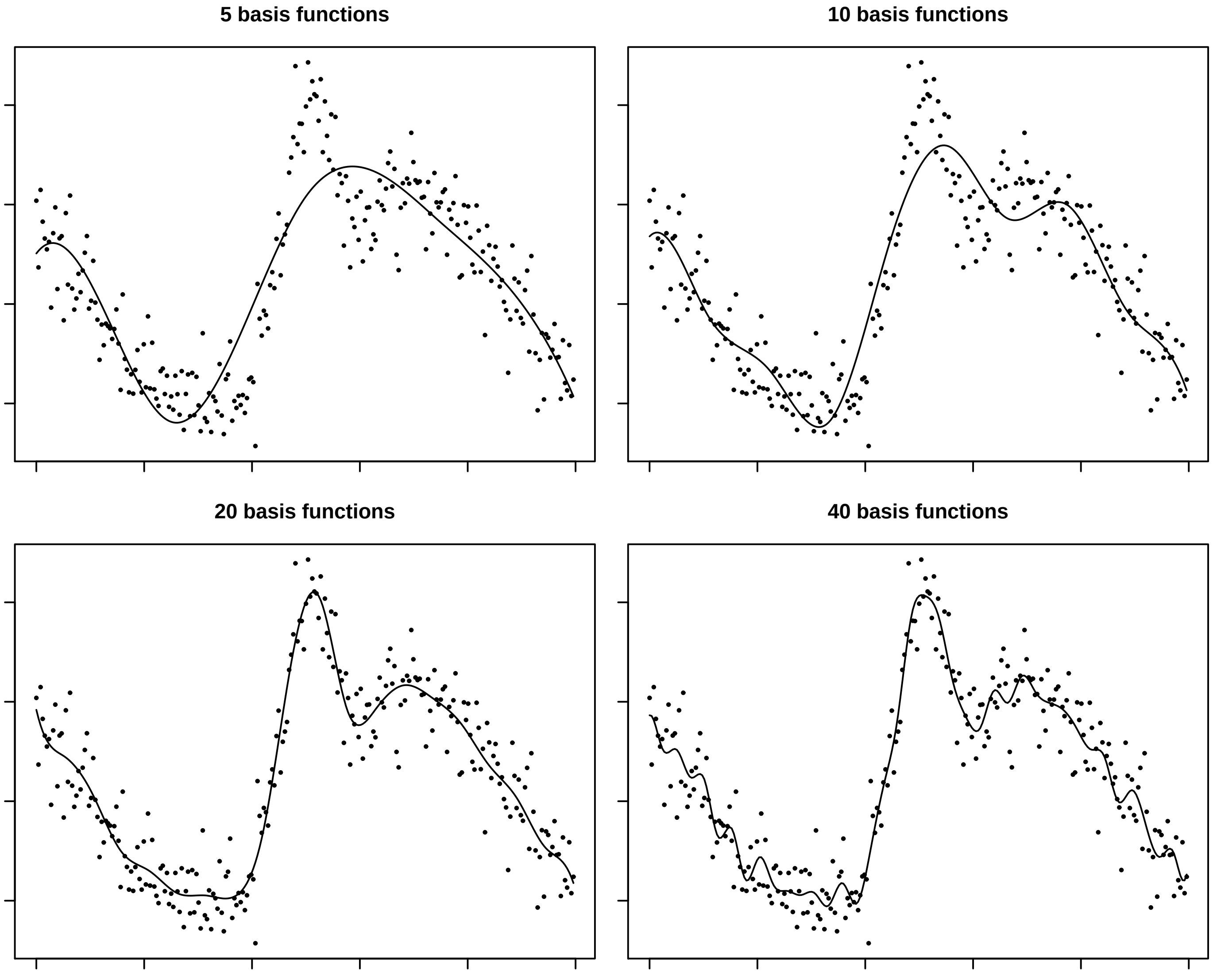
Penalised Splines

- Approximate a function $f(x)$ or $g(t)$ 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 basis functions:



- 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

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

- Joint prior distribution is multivariate Gaussian

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

- The penalty corresponds to the log-prior.

Spatial Effects

- **Regional data:** 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

$$\beta_s = \frac{1}{|N(s)|} \sum_{r \in N(s)} \beta_r + u_s, \qquad u_s \sim N\left(0, \frac{\tau^2}{|N(s)|}\right)$$

- Again the joint prior distribution is multivariate Gaussian

$$p(\beta) \propto \exp\left(-\frac{1}{2\tau^2}\beta' K \beta\right)$$

where K is an adjacency matrix and

$$\mathrm{pen}(\beta) = -\log(p(\beta)).$$

- **Individual data:** Estimate a separate parameter β_s for each distinct location $s = (s_x, s_y)$.

- Smoothness assumption: The correlation of the spatial effect between two points s_1 s_2 can be described in terms of a parametric correlation function, e.g.

$$\rho(s_1, s_2) = \rho(||s_1 - s_2||) = \exp(-\alpha||s_1 - s_2||).$$

- More precisely: $\{\beta_s, s \in \mathbb{R}^2\}$ is assumed to follow a zero-mean stationary Gaussian random field.

- Well-known as Kriging in geostatistics.

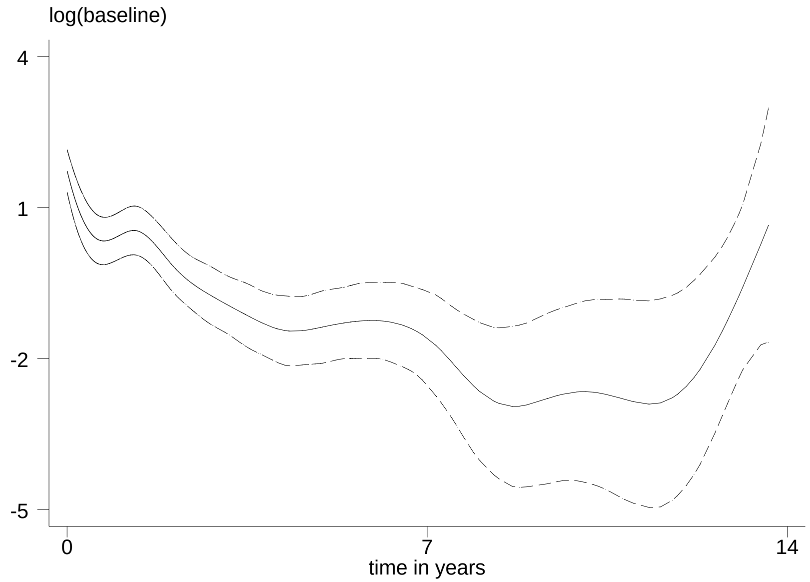
- Results in a multivariate Gaussian prior for the spatial effects.

Bayesian Inference

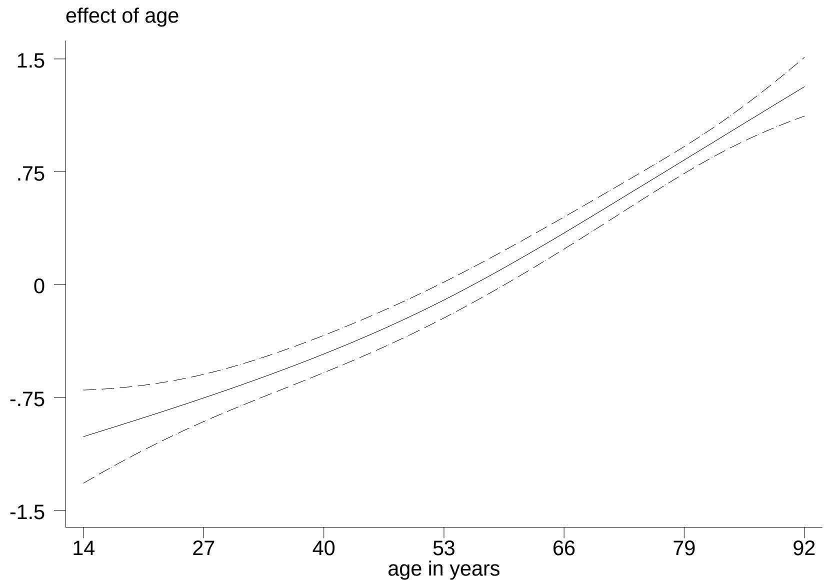
- Unifying framework:
  - All vectors of function evaluations can be written as the product of a design matrix X_j and a vector of regression coefficients β_j, i.e. $f_j = X_j\beta_j$.
  - Regularisation penalties are quadratic forms $\lambda_j\beta_j'K_j\beta_j$ corresponding to Gaussian priors
$$p(\beta|\tau^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j'K_j\beta_j\right).$$
  - The variance τ_j^2 is a transformation of the smoothing parameter λ_j.
- The unifying framework allows to devise equally general inferential procedures.
- Implemented in the stand-alone software BayesX.

- 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 IWLS-proposals).
  - Gibbs sampler for the variances (inverse gamma with updated parameters).

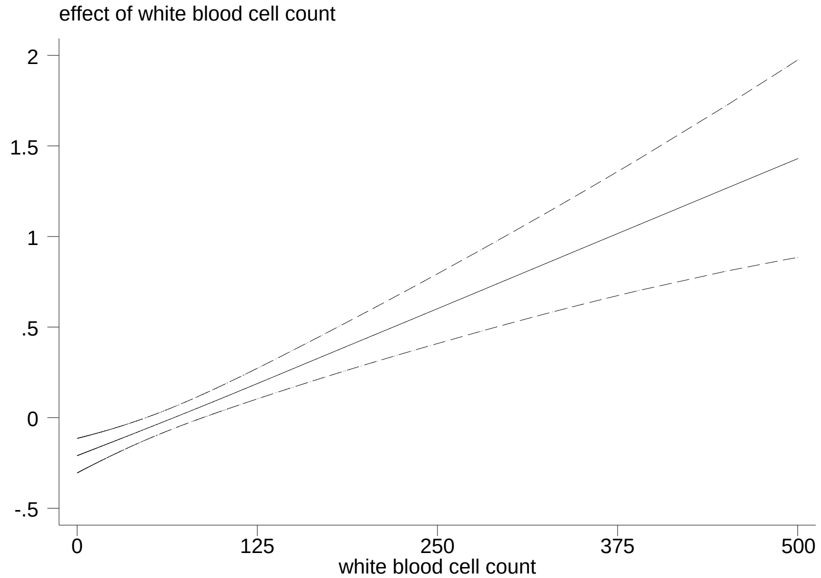
Results



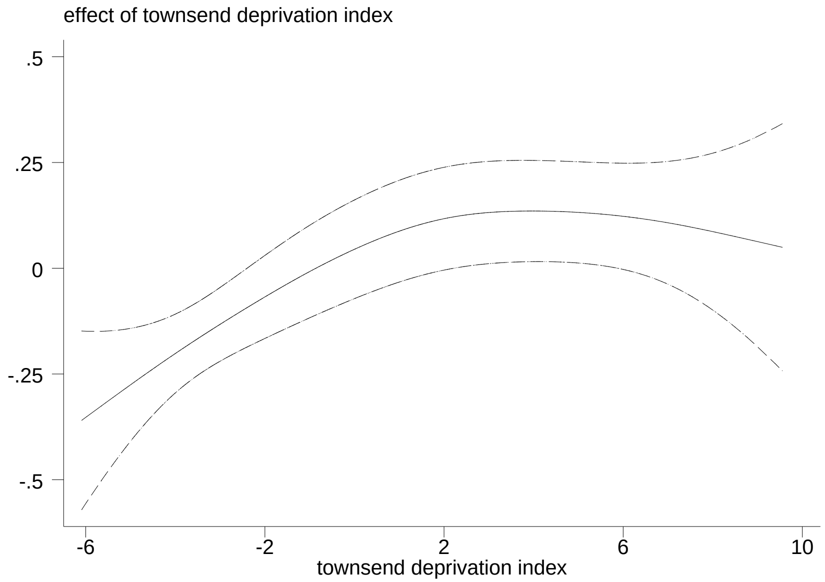
Log-baseline hazard.



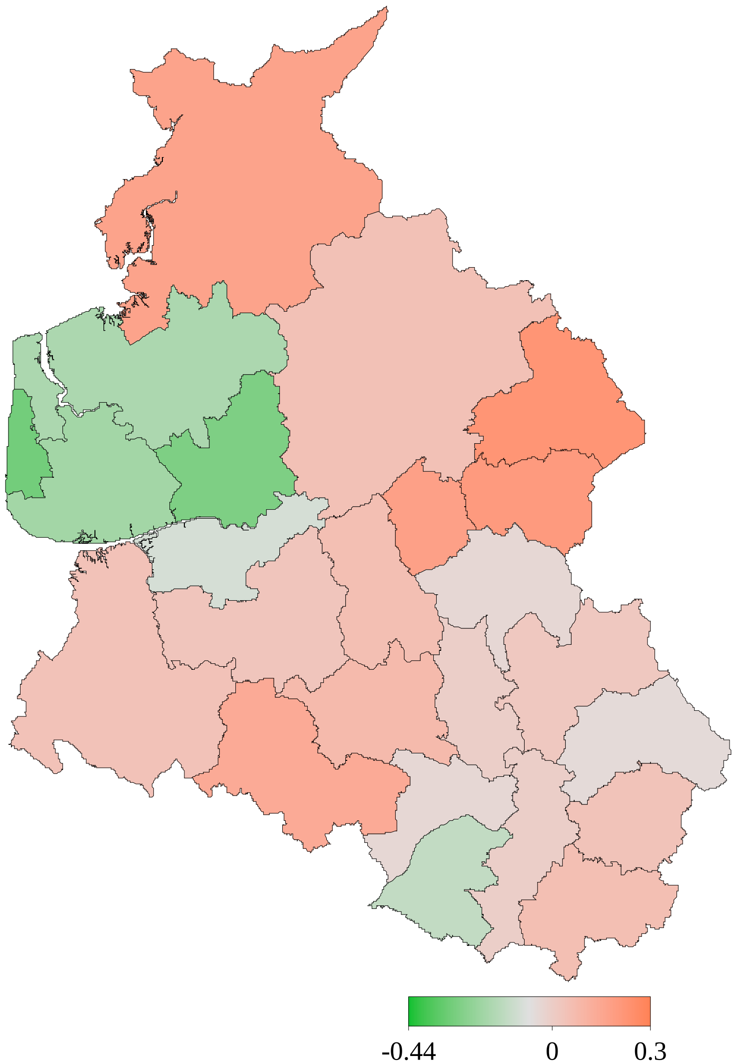
Effect of age at diagnosis.



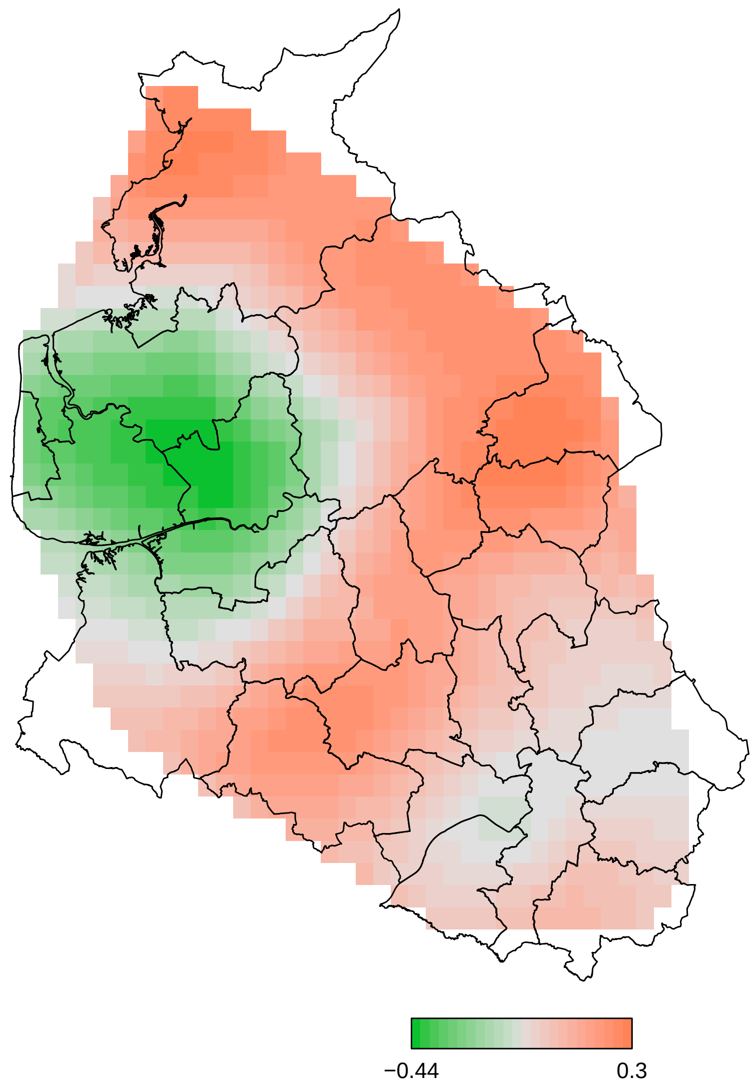
Effect of white blood cell count.



Effect of deprivation.



District-level analysis



Individual-level analysis

Summary I

- Geoadditive hazard regression provides a flexible model class for analysing survival times.
- The software also supports more general censoring schemes, including left and interval censoring.
- Boosting-based methods for model choice and variable selection are currently under development.
- Similar models are available in the context of generalised linear models and categorical regression.

Penalisation Approaches for High-Dimensional Predictors

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

- Examples include gene expression data but also social science and economic applications.

- Most well-known approach: Ridge regression in the Gaussian model

$$y = X\beta + \varepsilon$$

- Estimation of β becomes numerically unstable for a large number of covariates

  $\Rightarrow$ Add a quadratic penalty to the least squares criterion:

$$LS_{\text{pen}}(\beta) = (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \to \min_{\beta}.$$

- Closed form solution: Penalised least squares (PLS) estimate

$$\hat{\beta} = (X'X + \lambda I)^{-1} X'y$$

- The PLS estimate is biased, but has a reduced variance compared to the least squares estimate.

- Suitable choices of the smoothing parameter (for example by cross validation) should yield a reduced mean squared error.

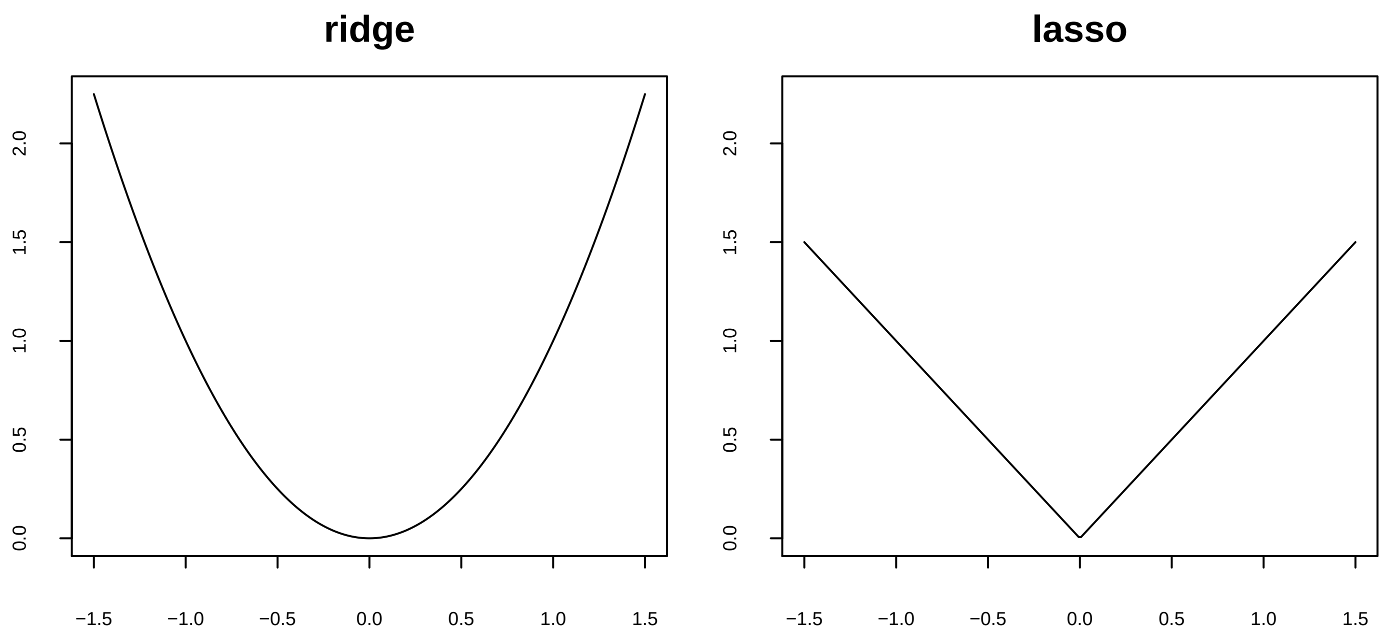
- Essential for deriving the PLS estimate: The penalty term is differentiable with respect to β.

- Drawback: Ridge regression typically does not induce enough sparsity.

  $\Rightarrow$ Consider penalties that have a spike in zero.

- LASSO: Replace quadratic penalty with absolute value penalty:

$$LS_{\text{pen}}(\beta) = (y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \rightarrow \min_{\beta} .$$



- No closed form solution available, but efficient algorithms exist for purely linear models.

- LASSO imposes more sparseness and is able to set coefficients equal to zero.

- Other types of regularisation penalties:

  – L_p-penalties:

$$\text{pen}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^p, \qquad 0 \leq p \leq 2.$$

  – Bridge-penalty:

$$\text{pen}(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2.$$

- Algorithms exist for linear models but become increasingly complex when considering non-Gaussian responses or combinations with geoadditive regression terms.

  $\Rightarrow$ Can we benefit from a Bayesian formulation?

Regularisation Priors

- Bayesian linear model:

$$y = X\beta + \varepsilon, \qquad \beta \sim N(0, \tau^2 I).$$

- Yields the posterior

$$p(\beta|y) \propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right) \exp\left(-\frac{1}{2\tau^2}\beta'\beta\right)$$

- Maximising the posterior is equivalent to minimising the penalised least squares criterion

$$(y - X\beta)'(y - X\beta) + \lambda\beta'\beta$$

where the smoothing parameter is given by the noise to signal ratio

$$\lambda = \frac{\sigma^2}{\tau^2}.$$

- Posterior mode for Gaussian prior is equivalent to the PLS (ridge) estimate.

- The analogy carries over to more general types of priors:

Penalty	**Prior density**	**Distribution**
Ridge	$p(\beta_j) \propto \exp(-\lambda\beta_j^2)$	Gauss
LASSO	$p(\beta_j) \propto \exp(-\lambda\|\beta_j\|)$	Laplace
L_p	$p(\beta_j) \propto \exp(-\lambda\|\beta_j\|^p)$	Powered exponential
Bridge	$p(\beta_j) \propto \exp(-\lambda_1\|\beta_j\|) + \exp(-\lambda_2\beta_j^2)$	Mixture

- Instead of maximising the posterior, consider simulation based estimation of the posterior mean.

- Advantages of MCMC simulation:
  - Modular framework allows for immediate combination with nonparametric or spatial effects.
  - Hyperpriors for further model parameters yield a fully automated estimation scheme.
  - Credible intervals for all parameters are available.

- Difficulty: Constructing appropriate proposal densities.
  - The Gaussian prior is conjugate for Gaussian responses and yields a Gibbs sampling scheme.
  - For non-Gaussian responses and Gaussian priors, adaptive proposal densities have been constructed based on iteratively weighted least squares proposals.
  - For non-Gaussian priors, new proposal densities have to be developed, e.g. random walk proposals.
  - Difficult due to the spike at zero.

Scale Mixtures of Normals

- Popular idea in robust Bayesian approaches if the Gaussian distribution seems to be questionable: Specify a hierarchical model, where

$$y|\sigma^2 \sim N(\mu, \sigma^2), \qquad \sigma^2 \sim IG(a, b).$$

- Marginally, y follows a t-distribution but sampling can be based on Gaussian responses with inverse gamma hyperprior on the variance.

- Similarly, several regularisation priors can be written as scale mixtures of normals, i.e.

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

where

$$\beta_j|\tau_j^2 \sim N(0, \tau_j^2) \qquad \text{and} \qquad \tau_j^2|\lambda \sim p(\tau_j^2|\lambda).$$

- For the LASSO:

$$\tau_j^2 | \lambda \sim Exp\left(\frac{\lambda^2}{2}\right).$$

- 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.

  $\Rightarrow$ IWLS updates 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.

- Example: Powered exponential prior

$$\exp(-|\beta_j|^p) \propto \int_0^\infty \exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right) \frac{1}{\tau_j^6} s_{p/2}\left(\frac{1}{2\tau_j^2}\right) d\tau_j^2$$

where $s_p(\cdot)$ is the density of the positive stable distribution with index p.

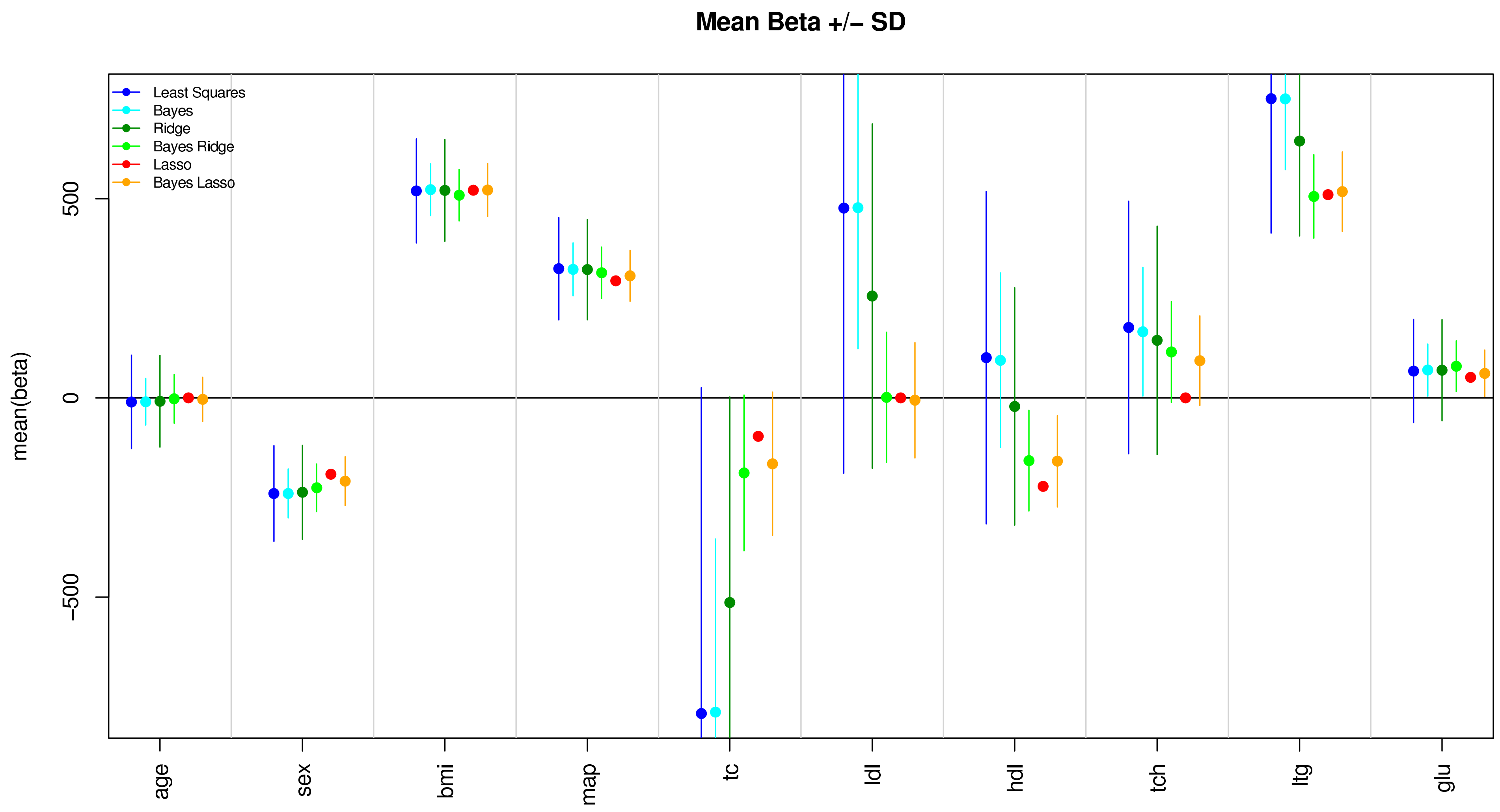
Example

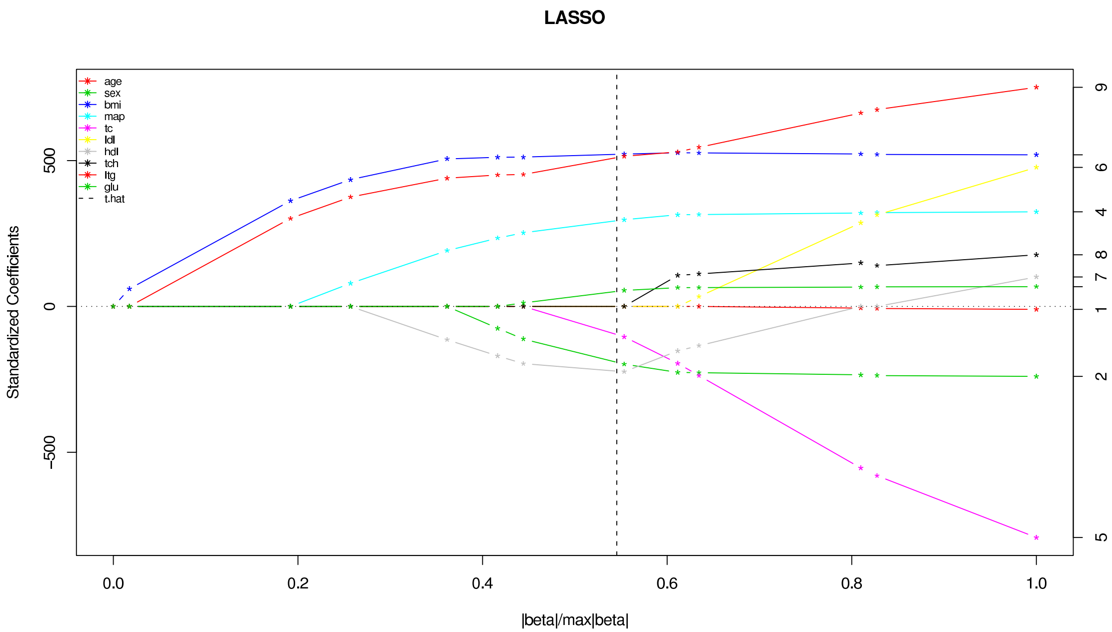
- Diabetes data also used in the LARS-paper by Efron et al. (2004).
- 442 observations on a measure of disease progression (response) shall be related to the covariates

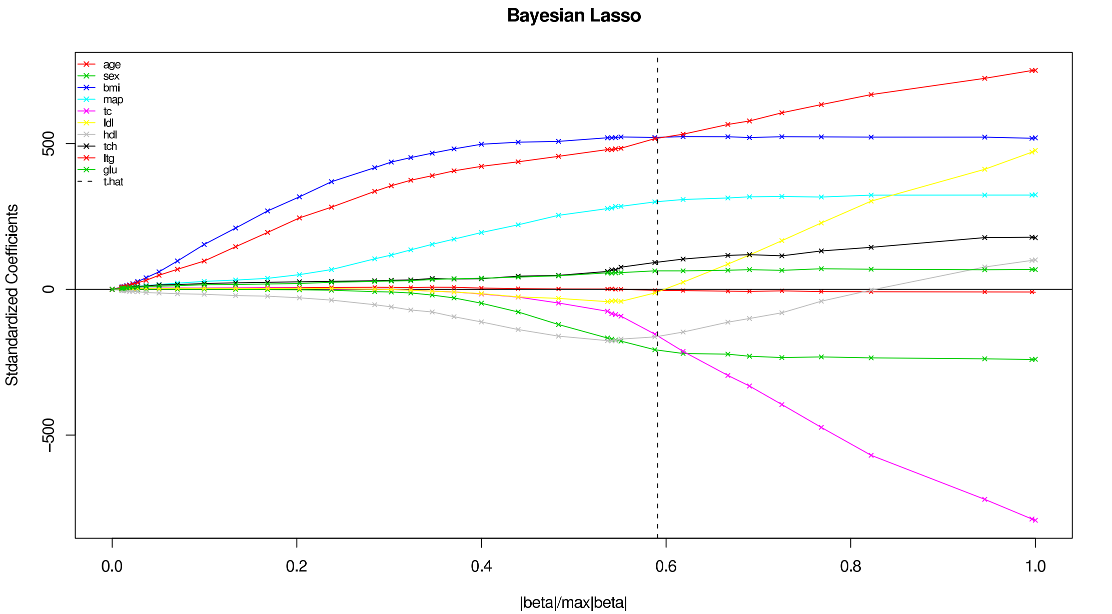
age	age of the patient
sex	gender
bmi	body mass index
map	average blood preasure
tc, ldl, hdl, tch, ltg, glu	blood serum measurements

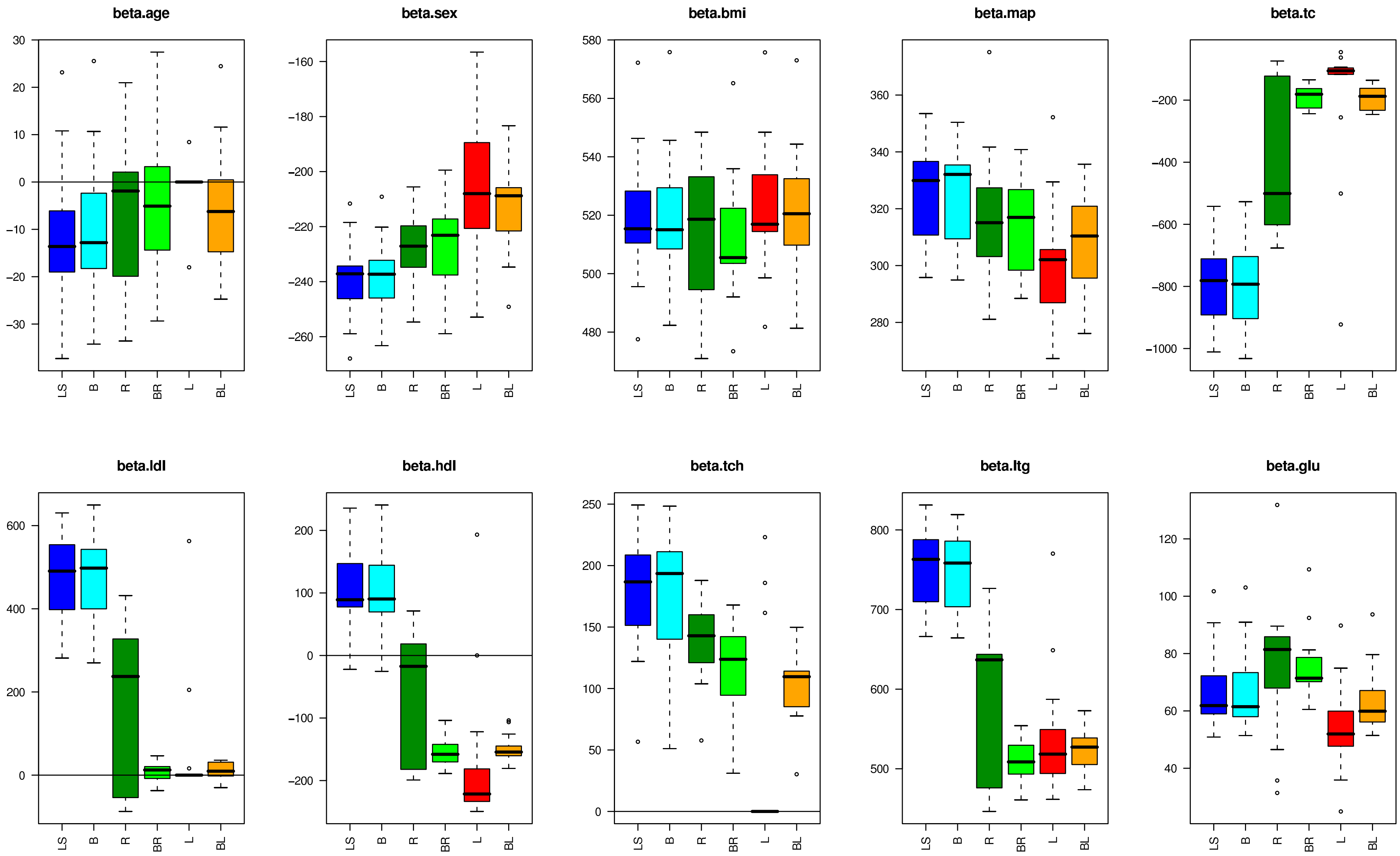
- Covariates are standardised and the response is centered.

- Compare six competing approaches:
  - Ordinary least squares (LS),
  - Bayes with noninformative prior (B),
  - Ridge regression (R),
  - Bayesian ridge regression (BR),
  - Frequentist LASSO (L),
  - Bayesian LASSO (BL).
- Boxplots are based on 13-fold cross-validation (408 training cases and 34 test cases).


Mean Beta +/– SD
Least Squares
Bayes
Ridge
Bayes Ridge
Lasso
Bayes Lasso
mean(beta)
500
0
−500
age
sex
bmi
map
tc
ldl
hdl
tch
ltg
glu


LASSO
age
sex
bmi
map
tc
ldl
hdl
tch
ltg
glu
t.hat
Standardized Coefficients
500
0
−500
0.0
0.2
0.4
0.6
0.8
1.0
|beta|/max|beta|
9
6
4
8
7
1
2
5


Bayesian Lasso
age
sex
bmi
map
tc
ldl
hdl
tch
ltg
glu
t.hat
Stdandardized Coefficients
500
0
−500
0.0
0.2
0.4
0.6
0.8
1.0
|beta|/max|beta|


beta.age
beta.sex
beta.bmi
beta.map
beta.tc
beta.ldl
beta.hdl
beta.tch
beta.ltg
beta.glu
LS
B
R
BR
L
BL

Summary II

- Bayesian formulation allows to
  - represent complex penalties in terms of Gaussian penalties via scale mixtures,
  - re-use efficient algorithms derived for Gaussian priors,
  - provides the full posterior, i.e. measures of uncertainty like credible intervals.
- Disadvantage: Small coefficients are no longer set to zero.
- Possible remedy: Mixed discrete-continuous distributions with a point mass in zero.
- Simpler approximation: Two-component continuous mixture, where one component is concentrated around zero (despite being continuous).
- Find out more:

`http://www.stat.uni-muenchen.de/~kneib`