BayesX and INLA - Opponents or Partners?

Thomas Kneib

Institut für Mathematik
Carl von Ossietzky Universität Oldenburg

Monia Mahling

Institut für Statistik
Ludwig-Maximilians-Universität München

CARL VON OSSIETZKY universität OLDENBURG
FAKULTÄT V
MATHEMATIK UND NATURWISSENSCHAFTEN

Trondheim, 15.5.2009

Outline

- Conditionally Gaussian hierarchical models.
- MCMC inference in conditionally Gaussian models.
- BayesX.
- Credit Scoring Data.
- Summary and Discussion.

Conditionally Gaussian Hierarchical Models

- Hierarchical models with conditionally Gaussian priors for regression coefficients define a large class of flexible regression models.
- We will consider regression models with predictors of the form

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + f_1(\boldsymbol{z}_{i1}) + \ldots + f_r(\boldsymbol{z}_{ir}),$$

where $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are potentially high-dimensional vectors of covariates and parameters, while the generic functions $f_1, \ldots, f_r$ represent different types of nonlinear regression effects.

- Examples:
  - Nonlinear, smooth effects of continuous covariates x where $f_j(\boldsymbol{z}_j) = f(x)$.
  - Interaction surfaces of two continuous covariates or coordinates x_1, x_2 where $f_j(\boldsymbol{z}_j) = f(x_1, x_2)$.
  - Spatial effects based on discrete spatial, i.e. regional information $s \in \{1, \ldots, S\}$ where $f_j(\boldsymbol{z}_j) = f_{\text{spat}}(s)$.
  - Varying coefficient models where $f_j(\boldsymbol{z}_j) = x_1 f(x_2)$.
  - Random effects where $f_j(\boldsymbol{z}_j) = x b_c$ with a cluster index c.

- Model the generic functions with basis function approaches:

$$f_j(\boldsymbol{z}_j) = \sum_{k=1}^{K} \gamma_{jk} B_{jk}(\boldsymbol{z}_j).$$

- Yields a vector-matrix representation of the predictor:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_r\boldsymbol{\gamma}_r$$

- Conditionally Gaussian priors:

$$\boldsymbol{\beta}|\boldsymbol{\vartheta}_0 \sim \mathrm{N}(\boldsymbol{b}, \boldsymbol{B}) \quad \text{and} \quad \boldsymbol{\gamma}_j|\boldsymbol{\vartheta}_j \sim \mathrm{N}(\boldsymbol{g}_j, \boldsymbol{G}_j)$$

where $\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{\vartheta}_0)$, $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{\vartheta}_0)$, $\boldsymbol{g}_j = \boldsymbol{g}_j(\boldsymbol{\vartheta}_j)$, $\boldsymbol{G}_j = \boldsymbol{G}_j(\boldsymbol{\vartheta}_j)$.

- Most prominent examples of conditionally Gaussian priors in the context of estimating smooth effects are of the (intrinsic) Gaussian Markov random field type where

$$p(\boldsymbol{\gamma}_j|\delta_j^2) \propto \left(\frac{1}{\delta_j^2}\right)^{\frac{\mathrm{rank}(\boldsymbol{K}_j)}{2}} \exp\left(-\frac{1}{2\delta_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right),$$

i.e. $\boldsymbol{g}_j = \mathbf{0}$ and $\boldsymbol{G}_j^{-1} = \delta_j^2\boldsymbol{K}_j$.

- Example 1: Bayesian P-Splines

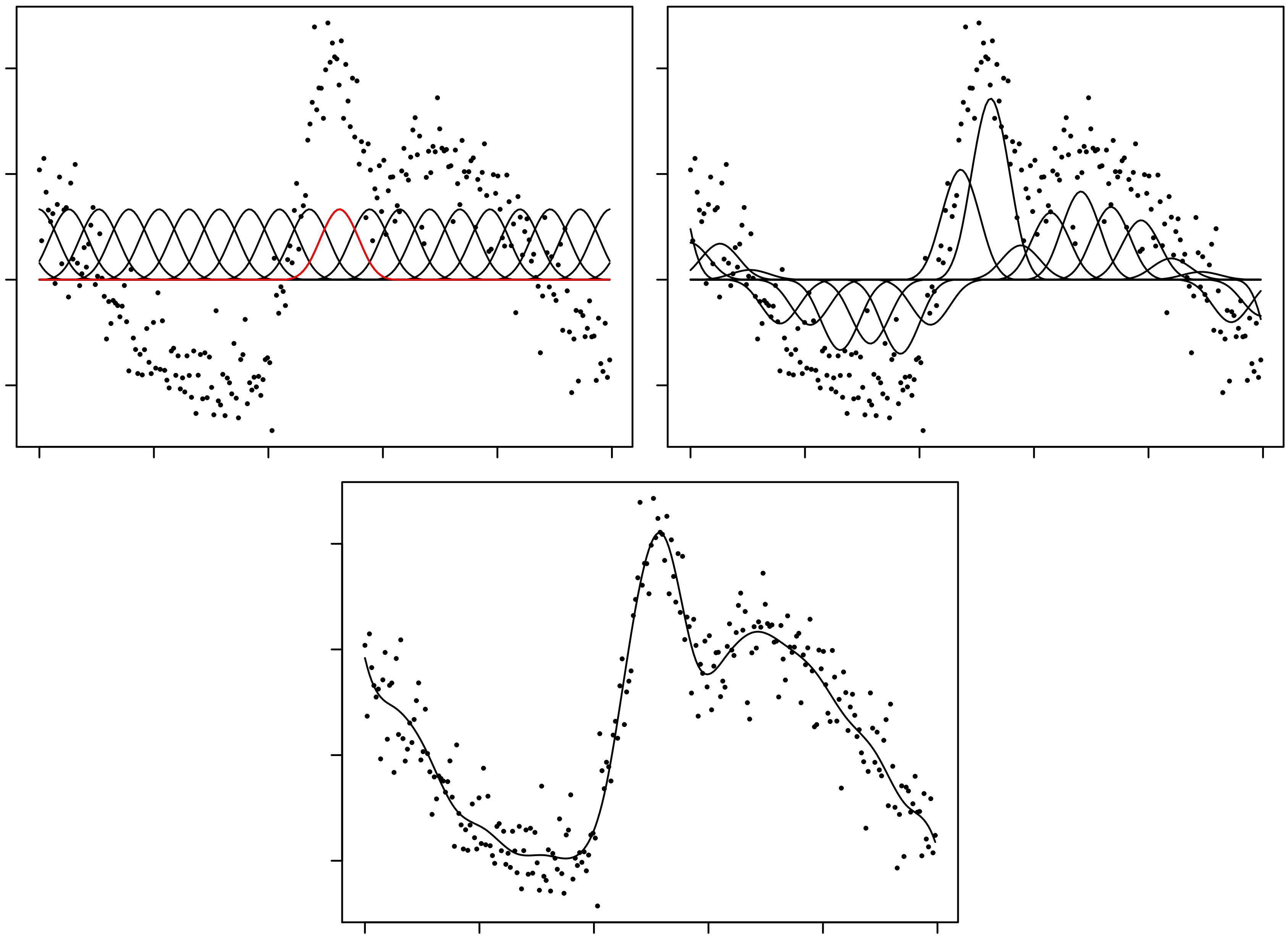
$$f(x) = \sum_{k=1}^{K} \gamma_k B_k(x).$$

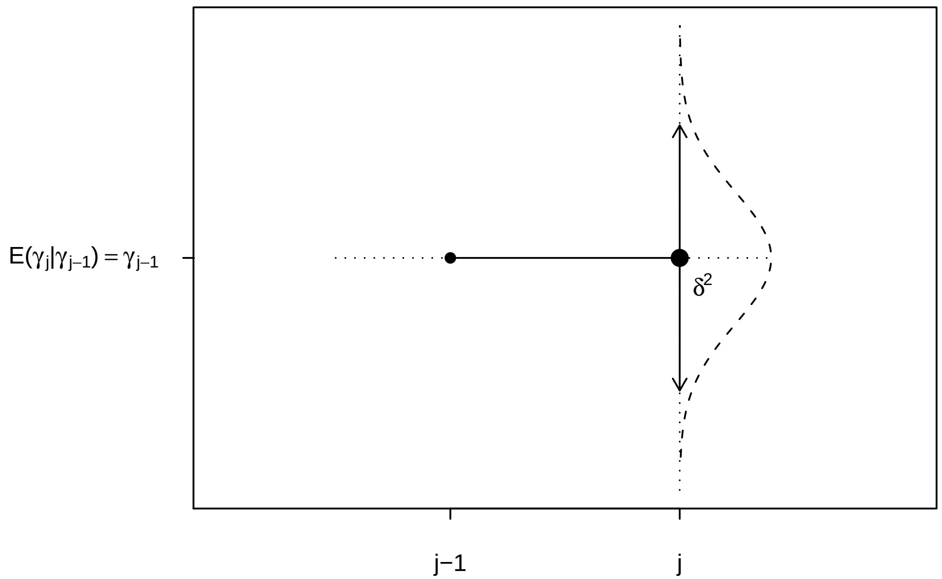
where $B_k(x)$ are B-spline basis functions of degree l and $\boldsymbol{\gamma}$ follows a random walk prior such as

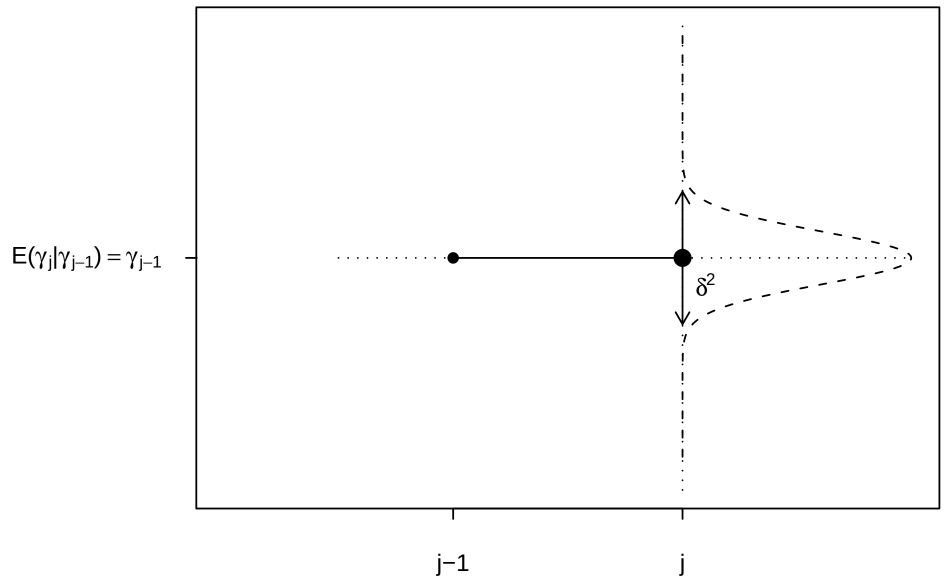
$$\gamma_k = \gamma_{k-1} + u_k, \quad u_k|\delta^2 \sim \mathrm{N}(0, \delta^2)$$

or

$$\gamma_k = 2\gamma_{k-1} - \gamma_{k-2} + u_k, \quad u_k|\delta^2 \sim \mathrm{N}(0, \delta^2).$$







- Usually, an inverse gamma prior is assigned to the smoothing variance:

$$\delta^2 \sim \mathrm{IG}(a, b).$$

- Bayesian P-splines include simple random walks as special cases (degree zero, knots at each distinct observed covariate value).

- Bayesian P-splines can be made more adaptive by replacing the homoscedastic random walk with a heteroscedastic version:

$$\gamma_k = \gamma_{k-1} + u_k, \quad u_k|\delta_k^2 \sim \mathrm{N}(0, \delta_k^2).$$

- Joint distribution of the regression coefficients becomes

$$p(\boldsymbol{\gamma}|\boldsymbol{\delta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{D}\boldsymbol{\Delta}\boldsymbol{D}\boldsymbol{\gamma}\right)$$

  where $\boldsymbol{\Delta} = \mathrm{diag}(\delta_2^2, \ldots, \delta_k^2)$.

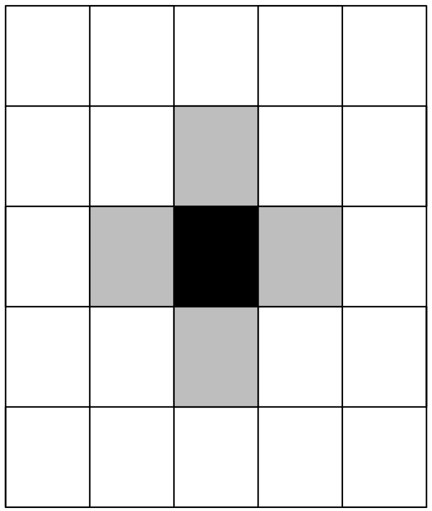
- Different types of hyperpriors for $\boldsymbol{\Delta}$:
  - I.i.d. hyperpriors, e.g. δ_k^2 i.i.d. $\mathrm{IG}(a, b,)$.
  - Functional hyperpriors, e.g. $\delta_k^2 = g(k)$ with a smooth function $g(k)$ modeled again as a P-spline.

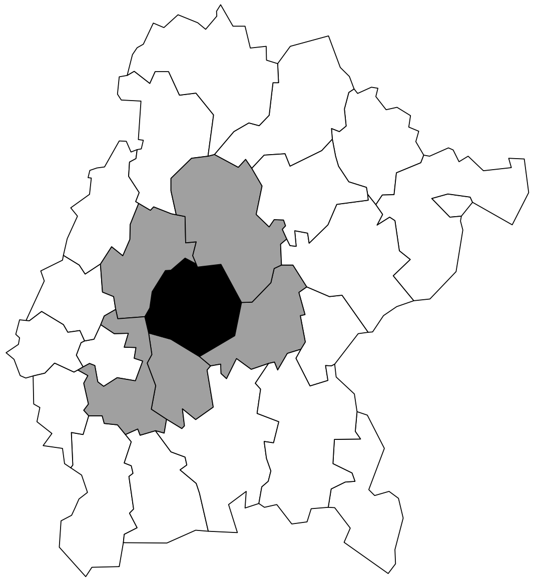
- Conditional on $\boldsymbol{\Delta}$ the prior for $\boldsymbol{\gamma}$ remains of the same type and an MCMC updates would not require changes.

- Example 2: Markov random fields for regional spatial effects:

$$\gamma_s|\gamma_r, r \in N(s) \sim N\left(\frac{1}{|N(s)|}\sum_{r\in N(s)}\gamma_r,\ \frac{\delta^2}{|N(s)|}\right).$$

- Based on the notion of spatial adjacency:





- Again, a hyperprior can be assigned to the smoothing variance but the joint distribution of the spatial effects remains conditionally Gaussian.

- For regularised estimation of high-dimensional regression effects $\boldsymbol{\beta}$ we are considering conditionally independent priors, i.e.

$$\boldsymbol{\beta}|\boldsymbol{\vartheta}_0 \sim \mathrm{N}(\boldsymbol{b}, \boldsymbol{B})$$

  with $\boldsymbol{b} = \boldsymbol{0}$ and $\boldsymbol{B} = \mathrm{diag}(\tau_1^2, \ldots, \tau_q^2)$.

- While allowing for different variances, hyperpriors for τ_j^2 will typically be identical.

- Example 1: Bayesian ridge regression

$$\beta_j|\tau_j^2 \sim \mathrm{N}(0, \tau_j^2), \quad \tau_j^2 \sim \mathrm{IG}(a, b).$$

- Note that the log-prior $\log p(\beta_j|\tau_j^2)$ equals the ridge penalty β_j^2 up to an additive constant.
- Induces a marginal t-distribution with $2a$ degrees of freedom and scale parameter $\sqrt{a/b}$.

- Informative priors provide the Bayesian analogon to frequentist regularisation.
- Example: Multiple linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

- For high-dimensional covariate vectors, least squares estimation becomes increasingly unstable.

  $\Rightarrow$ Add a penalty term to the least squares criterion, for example a ridge penalty

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

- Closed form solution: Penalised least squares estimate

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

- Bayesian version of the linear model:

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

- Yields the posterior

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

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

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}$$

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

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

- The posterior mode coincides with the penalised least squares estimate (for given smoothing parameter).
- More generally:
  - Penalised likelihood
$$l_{\text{pen}}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \text{pen}(\boldsymbol{\beta}).$$
  - Posterior:
$$p(\boldsymbol{\beta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}).$$
- In terms of the prior distribution

Penalty $\equiv$ log-prior.

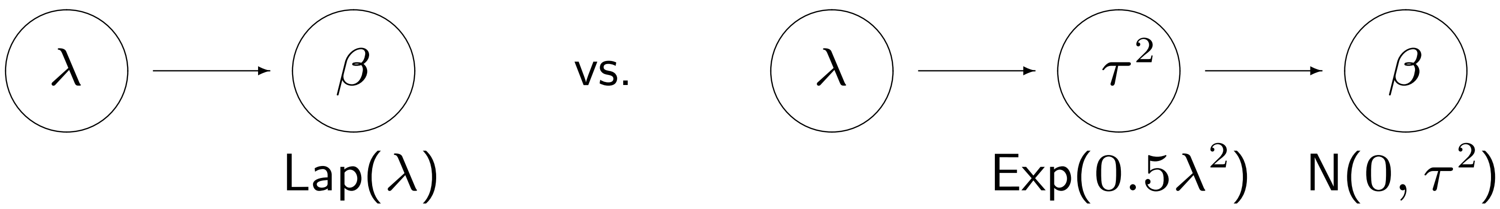
- Example 2: Bayesian lasso prior:

$$\beta_j|\tau_j^2, \lambda \sim \mathrm{N}(0, \tau_j^2), \quad \tau_j^2 \sim \mathrm{Exp}\left(\frac{\lambda^2}{2}\right).$$

- Marginally, β_j follows a Laplace prior

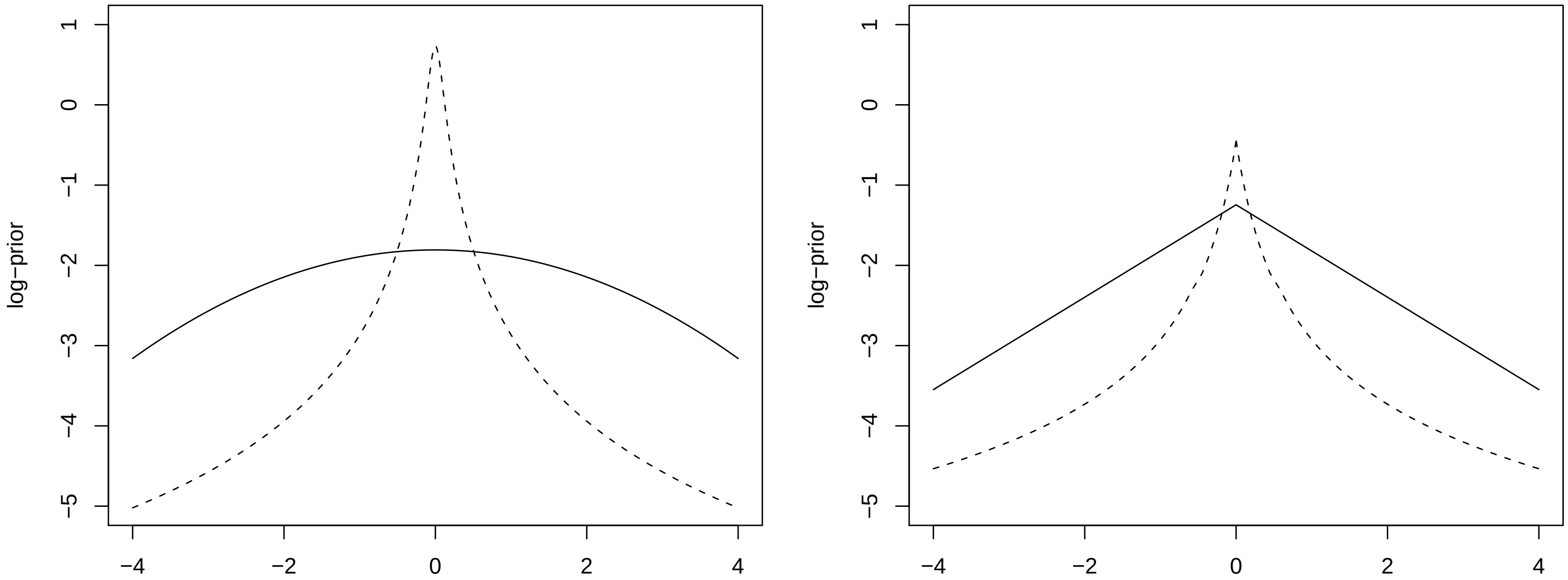
$$p(\beta_j) \propto \exp(-\lambda|\beta_j|).$$

- Hierarchical (scale mixture of normals) representation:



- A further hyperprior can be assigned to the smoothing parameter such as a gamma distribution $\lambda^2 \sim \mathrm{Ga}(a, b)$.

- Marginal Bayesian ridge and marginal Bayesian lasso:



- Example 3: General L_p priors

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

with $0 < p < 2$ (power exponential prior).

- Note that

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

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

MCMC Inference in Conditionally Gaussian models

- The general structure of conditionally Gaussian models enables the construction of general MCMC samplers.
- The conditionally Gaussian prior makes inference tractable in situations which are difficult with direct estimation (such as the lasso).
- Suitable hyperpriors enable inference and uncertainty assessment for all model parameters.
- MCMC fully exploits the hierarchical nature of the models through the consideration of full conditionals.

- For (latent) Gaussian responses, we obtain Gibbs sampling steps for the regression coefficients.

- For example, $\boldsymbol{\beta}|\cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ with

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \frac{1}{\sigma^2} \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\eta}_{-\boldsymbol{\beta}}) + \boldsymbol{B}^{-1}\boldsymbol{b}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B}^{-1}\right)^{-1},$$

- For non-Gaussian responses, construct adaptive proposal densities based on iteratively weighted least squares approximations to the full conditionals.

- For example, $\boldsymbol{\beta}$ is proposed from a multivariate Gaussian distribution with expectation and covariance matrix

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{X}'\boldsymbol{W}(\tilde{\boldsymbol{y}} - \boldsymbol{\eta}_{-\boldsymbol{\beta}}) + \boldsymbol{B}^{-1}\boldsymbol{b}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \boldsymbol{B}^{-1}\right)^{-1}.$$

where $\boldsymbol{W}$ and $\tilde{\boldsymbol{y}}$ are the usual generalised linear model weights and working responses.

- Full conditionals for hyperparameters are independent of the observation model.
- Bayesian ridge:

$$\tau_j^2 | \cdot \sim \mathrm{IG}\left(a + \frac{q}{2}, b + \frac{1}{2}\beta_j^2\right)$$

- Bayesian lasso:

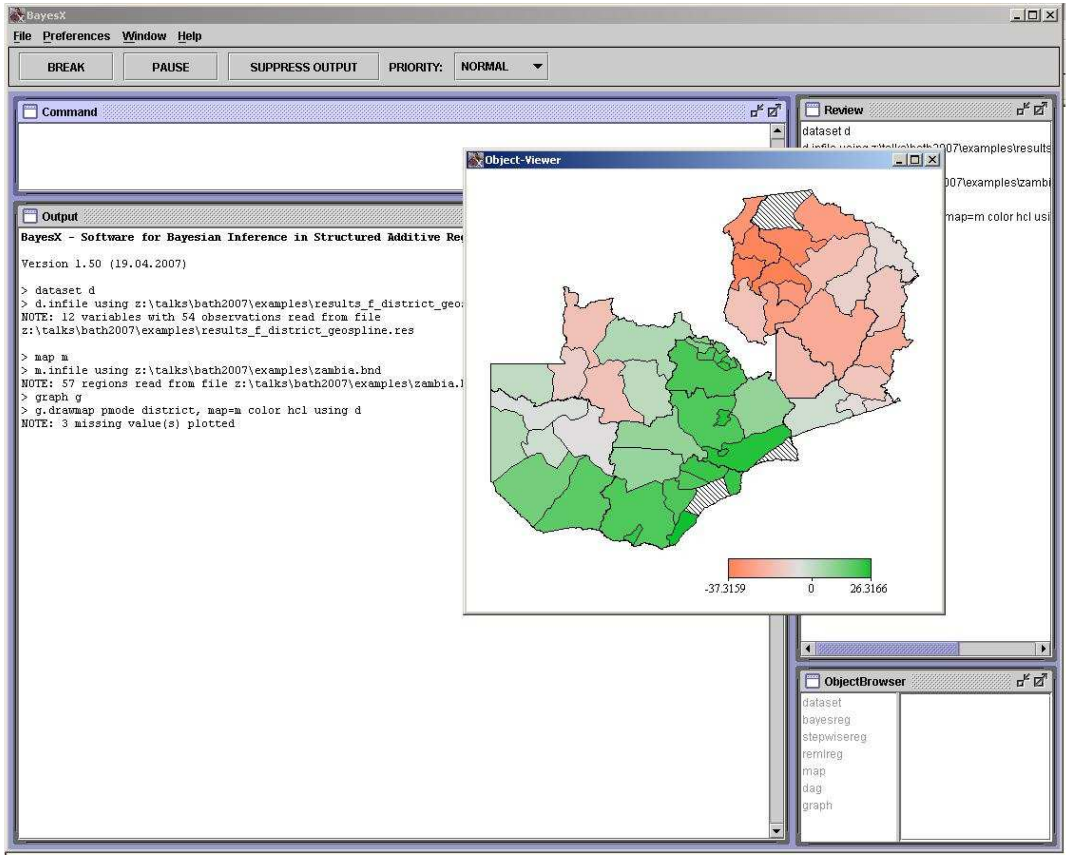
$$\left.\frac{1}{\tau_j^2}\right| \cdot \sim \mathsf{InvGauss}\left(\frac{|\lambda|}{|\beta_j|}, \lambda^2\right), \quad \lambda^2 | \cdot \sim \mathrm{Ga}\left(a + q, b + \frac{1}{2}\sum_{j=1}^{q}\tau_j^2\right).$$

- Smoothing variances:

$$\delta_j^2 | \cdot \sim \mathrm{IG}\left(a_j + \frac{\mathrm{rank}(\boldsymbol{K}_j)}{2}, b_j + \frac{1}{2}\boldsymbol{\gamma}_j \boldsymbol{K}_j \boldsymbol{\gamma}_j\right).$$

BayesX

- Markov chain Monte Carlo approaches for conditionally Gaussian regression models are implemented in BayesX.



- Available from

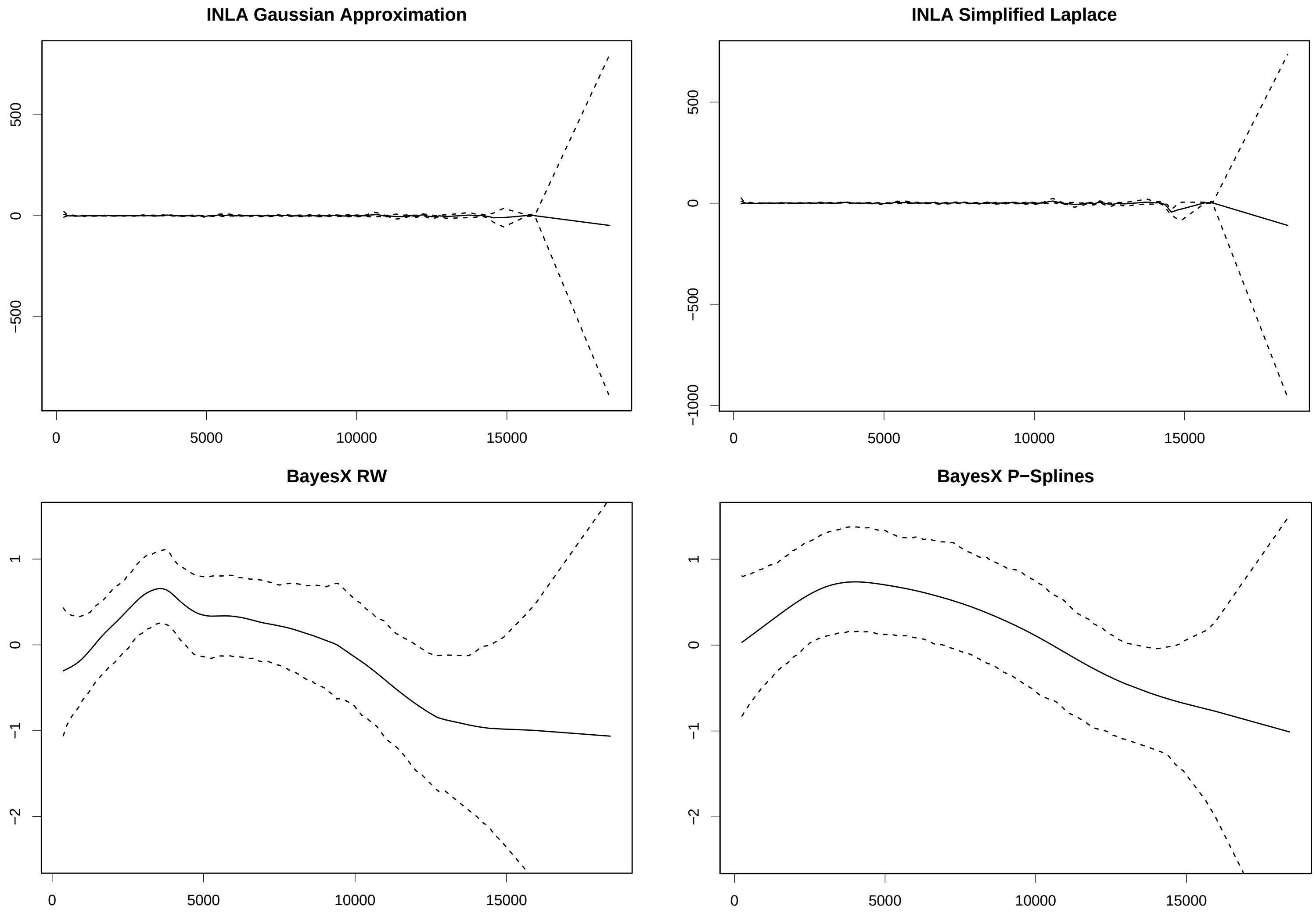
`http://www.stat.uni-muenchen.de/~bayesx`

- Numerical efficient implementation employing sparse matrix operations.
- Also contains mixed model based inference for the same class of models (comparable to INLAs Gaussian approximation).

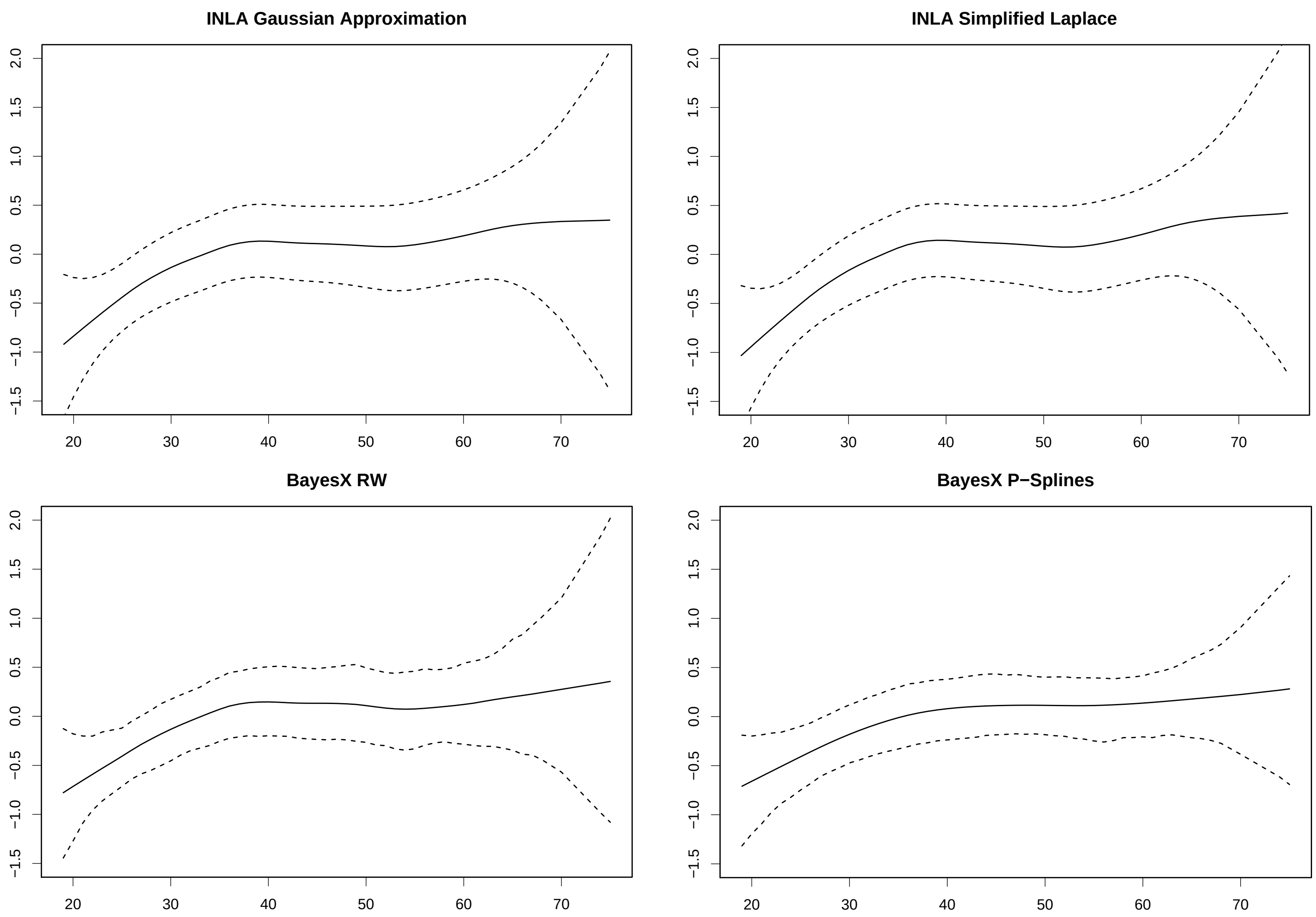
Credit Scoring Data

- Data on the defaults of 1,000 consumer credits from a German bank.
- Response variable is a binary indicator y_i that specifies whether the credit has been paid back ($y_i = 1$, credit-worthy) or not ($y_i = 0$, not credit-worthy).
- Covariates include age of the client, credit amount and duration of the credit.
- Consider binary logit models with nonparametric effects of these three covariates.
- Compare different approximations available in INLA with MCMC-based estimation in BayesX.

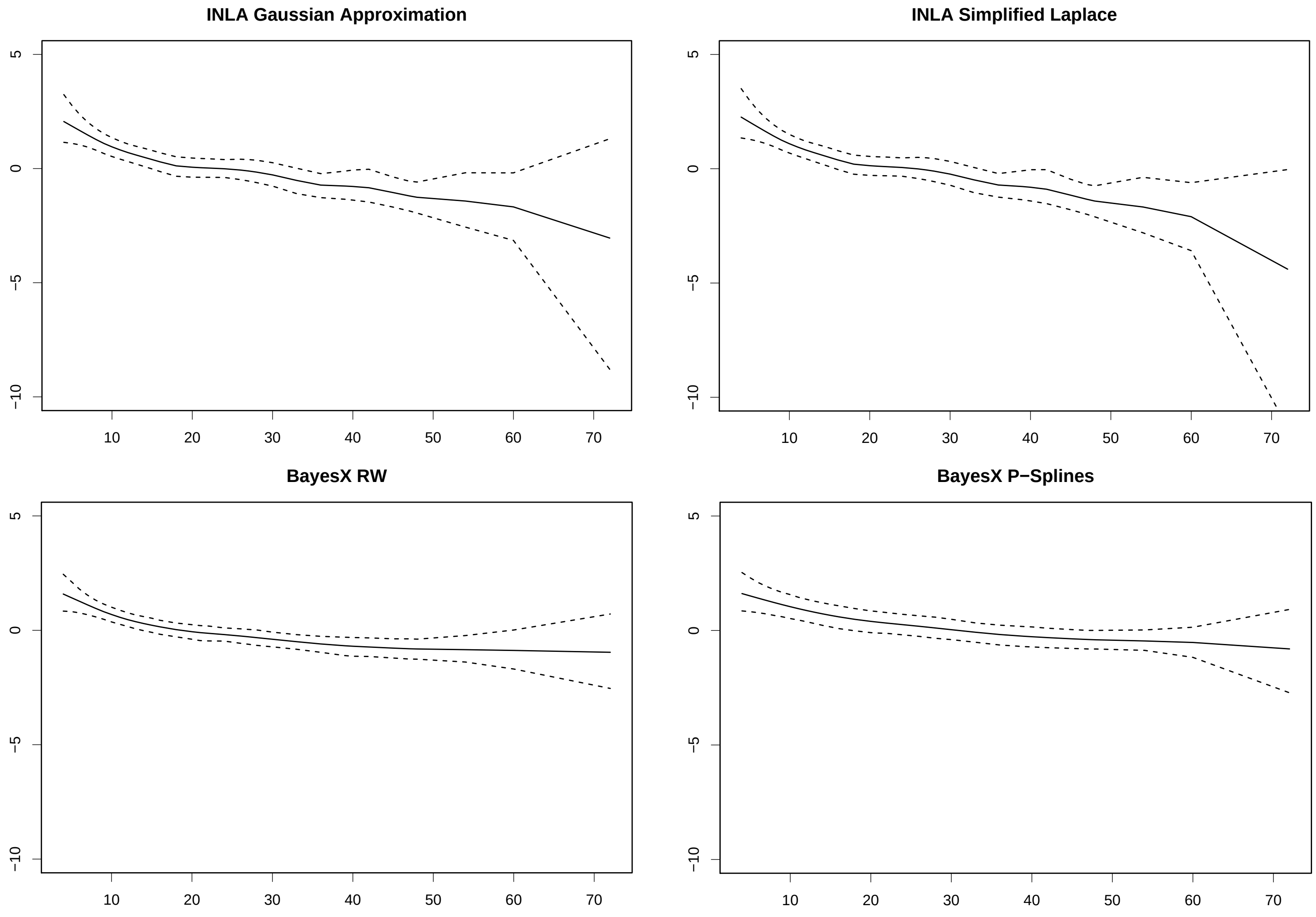
- Effects of amount obtained with the complete data:



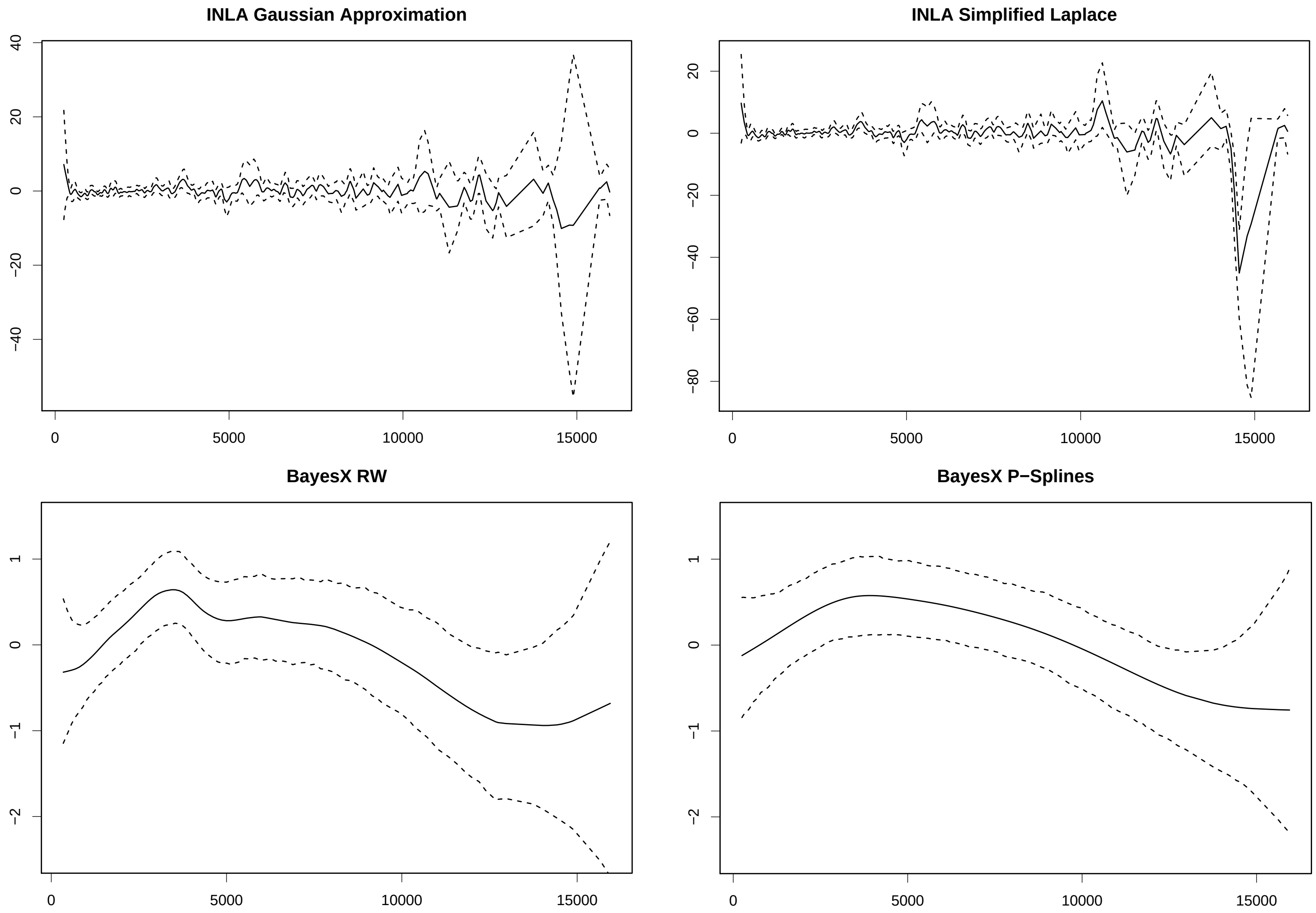
- Effects of age obtained with one outlier excluded:



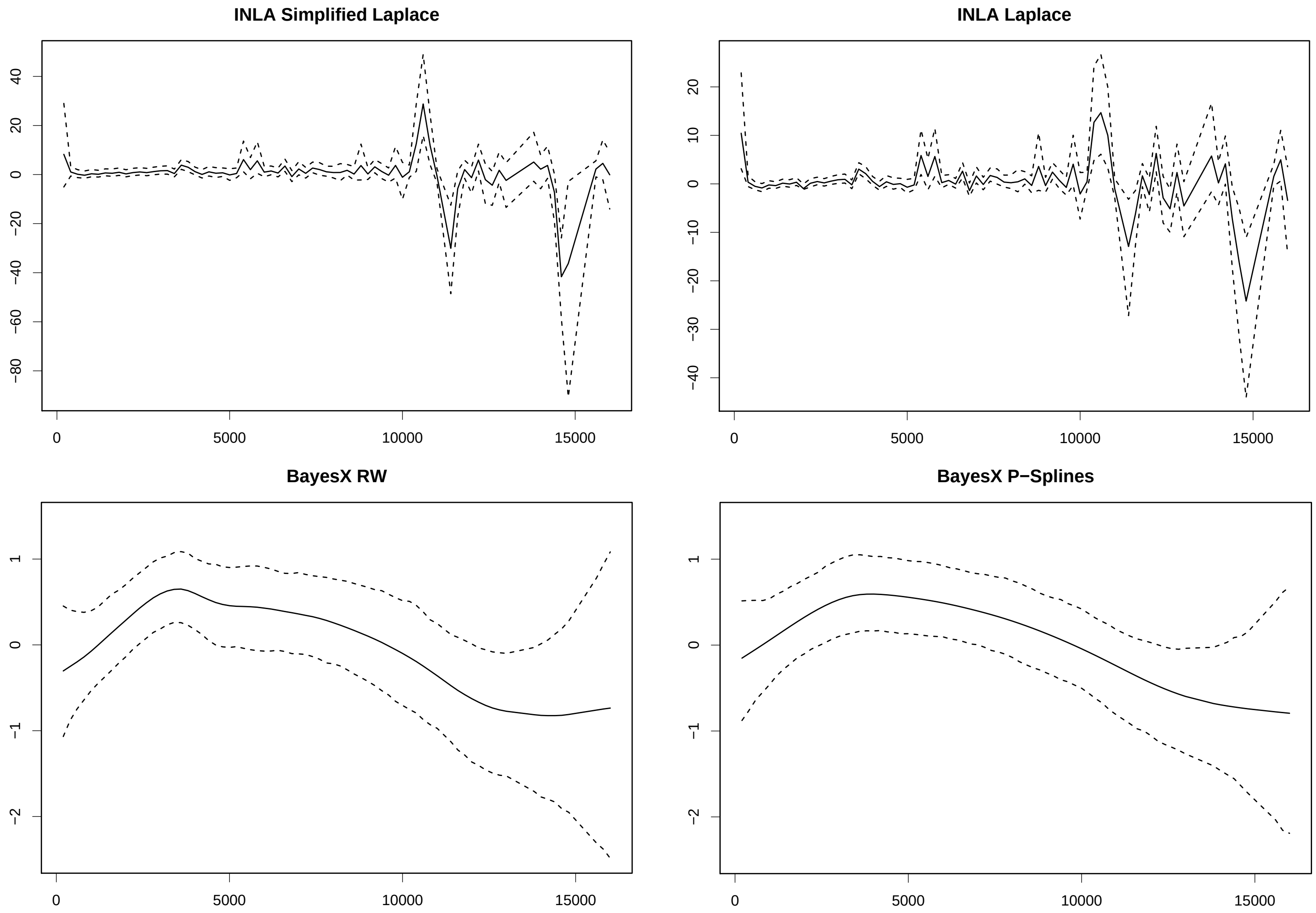
- Effects of duration obtained with one outlier excluded:



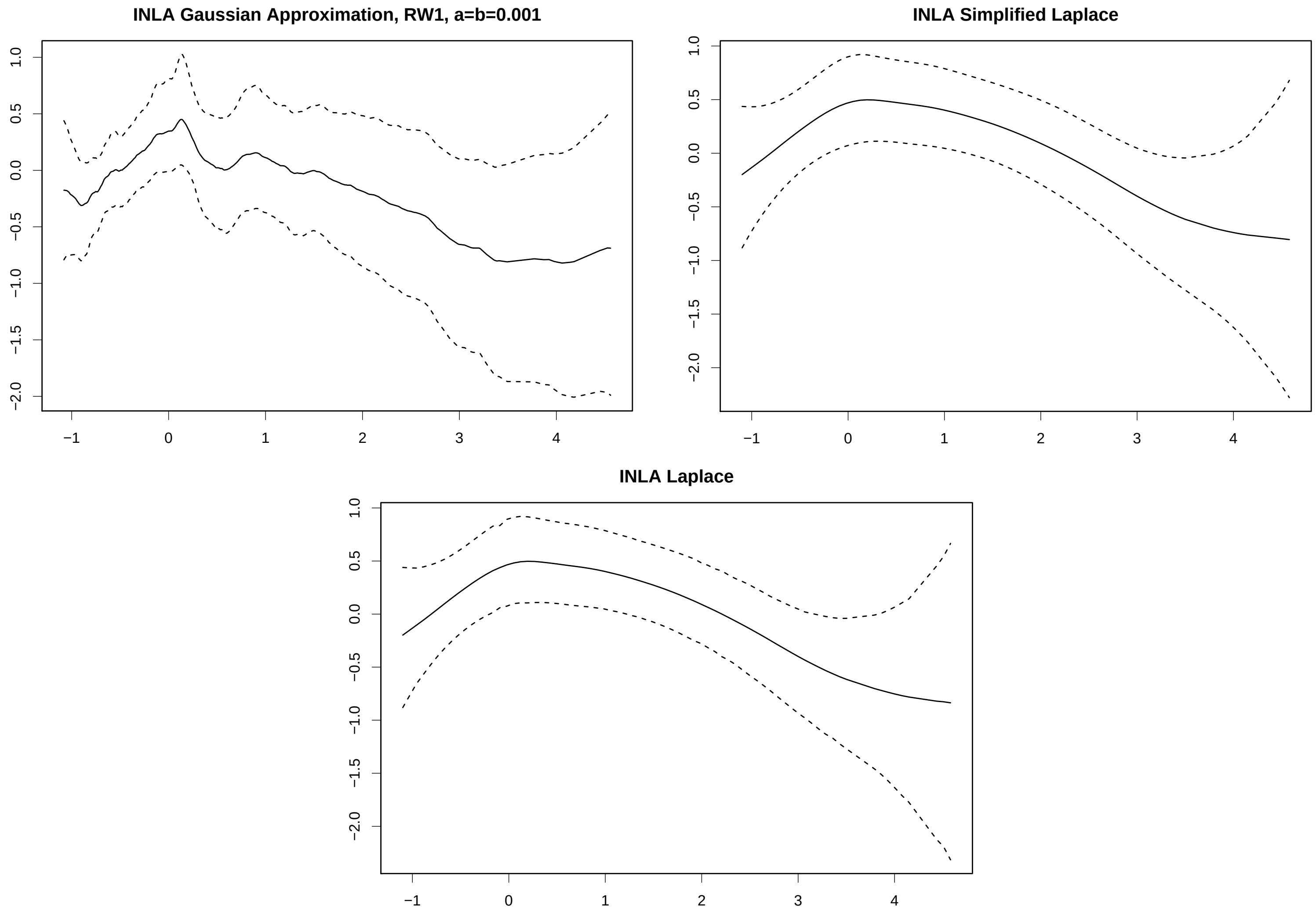
- Effects of amount obtained with one outlier excluded:



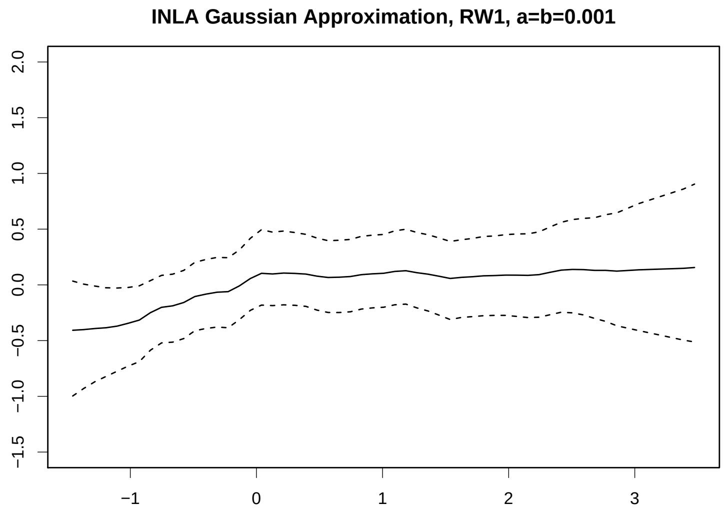
- Effects of amount based on rounded data with one outlier excluded:

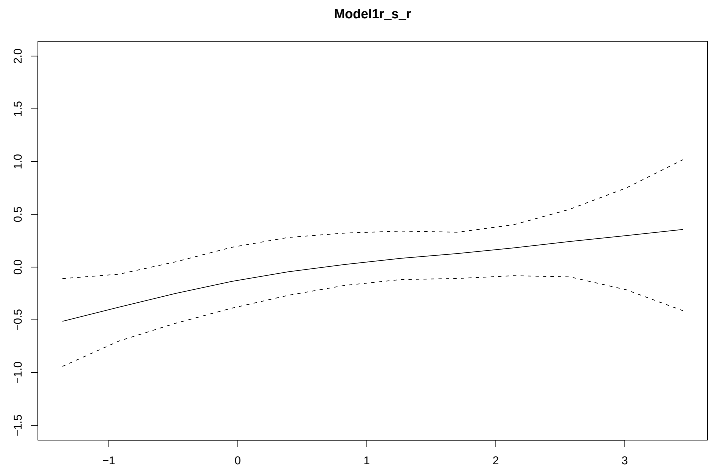


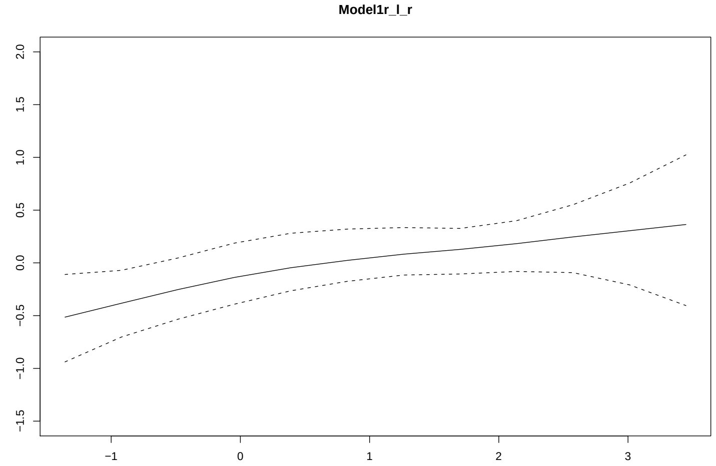
- Effects of amount after standardising covariates with one outlier excluded:



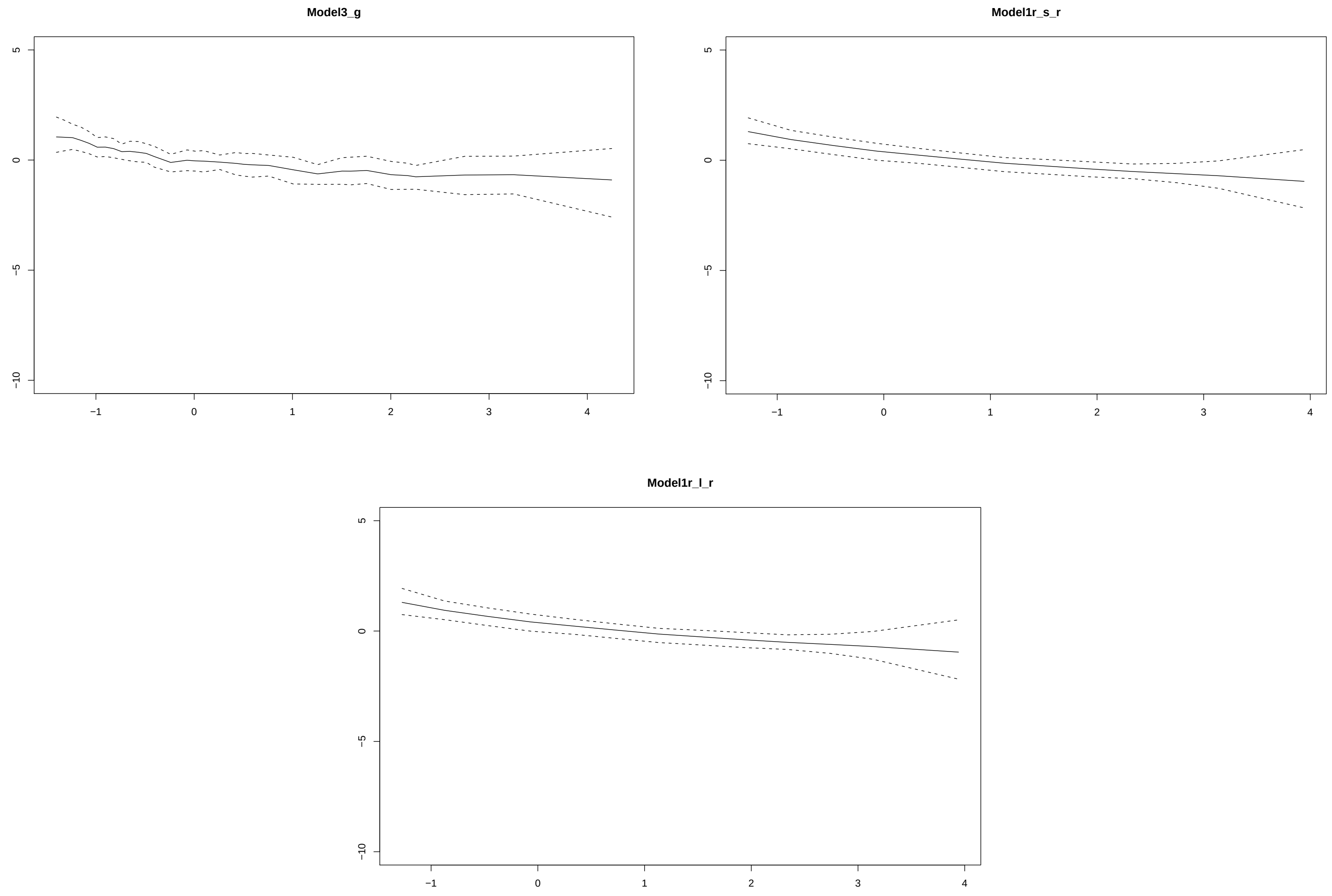
- Effects of age after standardising covariates with one outlier excluded:





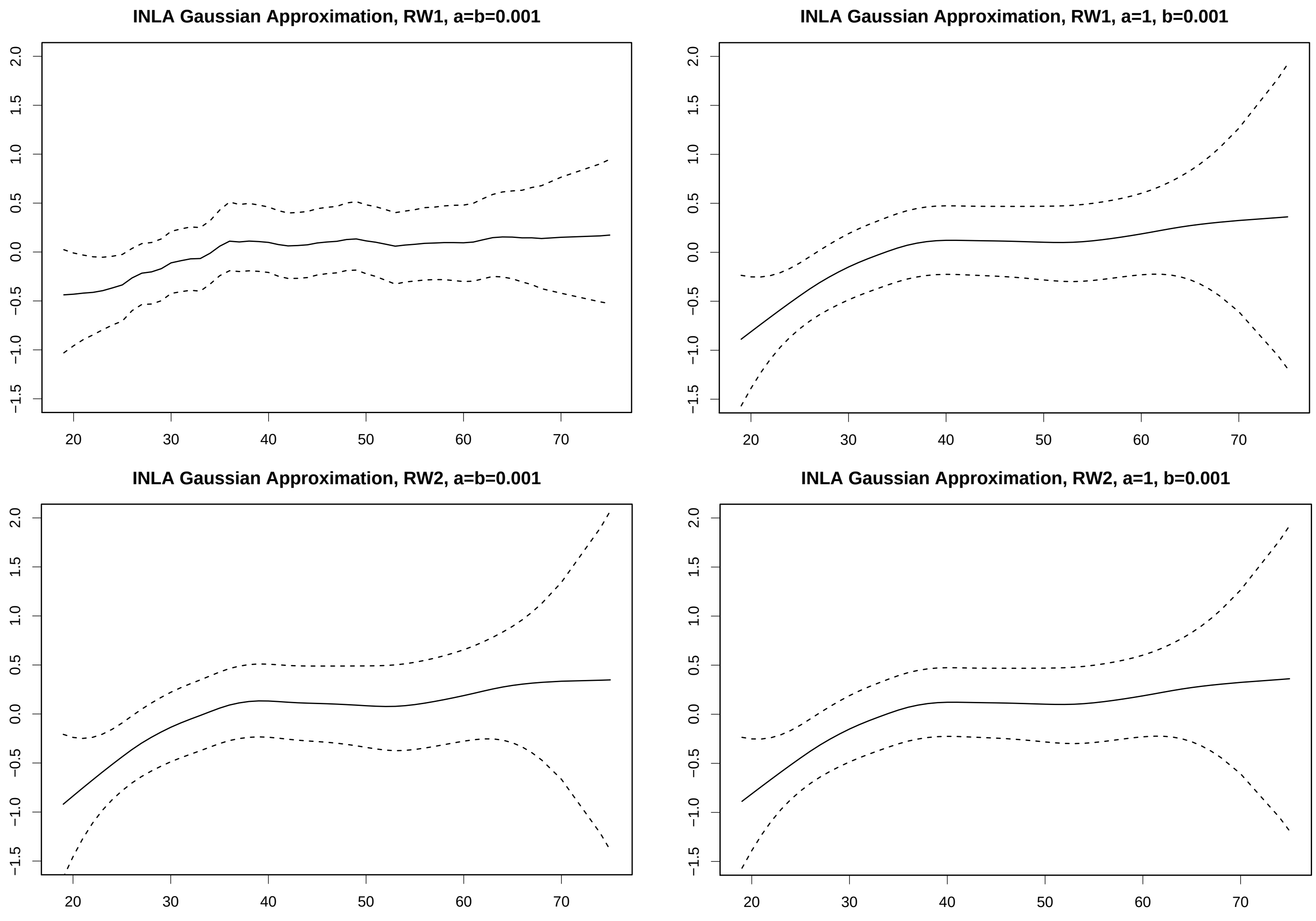


- Effects of duration after standardising covariates with one outlier excluded:

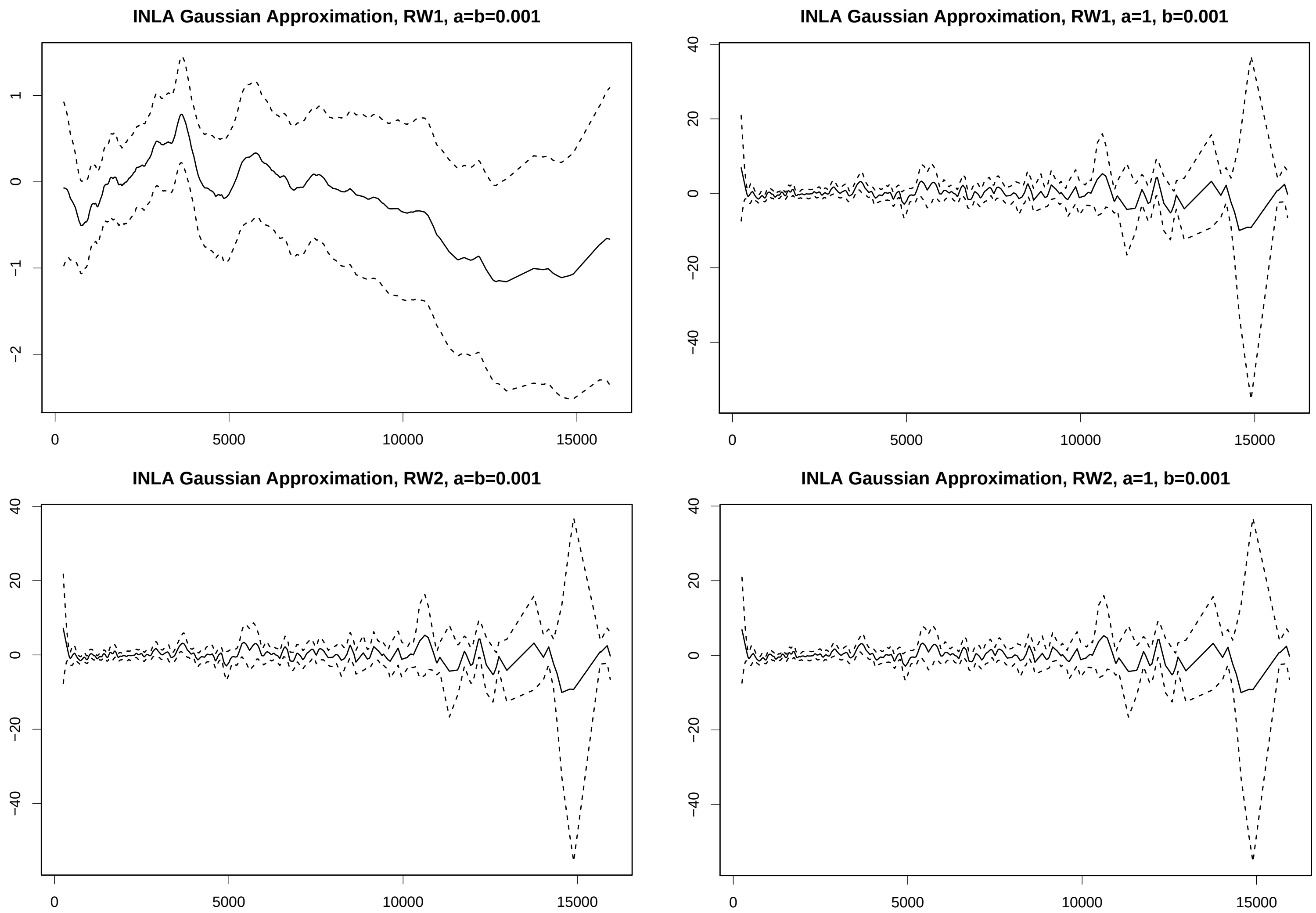


- Computing times for some selected models (in seconds, very rough estimates):
  - INLA with Gaussian approximation: 200s.
  - INLA with simplified Laplace: 240s.
  - INLA with Laplace (amount rounded): 2540s.
  - BayesX with RW prior and 12,000 iterations: 60s.
  - BayesX with RW prior and 103,000 iterations: 510s.
  - BayesX with P-spline prior and 12,000 iterations: 90s.
  - BayesX with P-spline prior and 103,000 iterations: 790s.

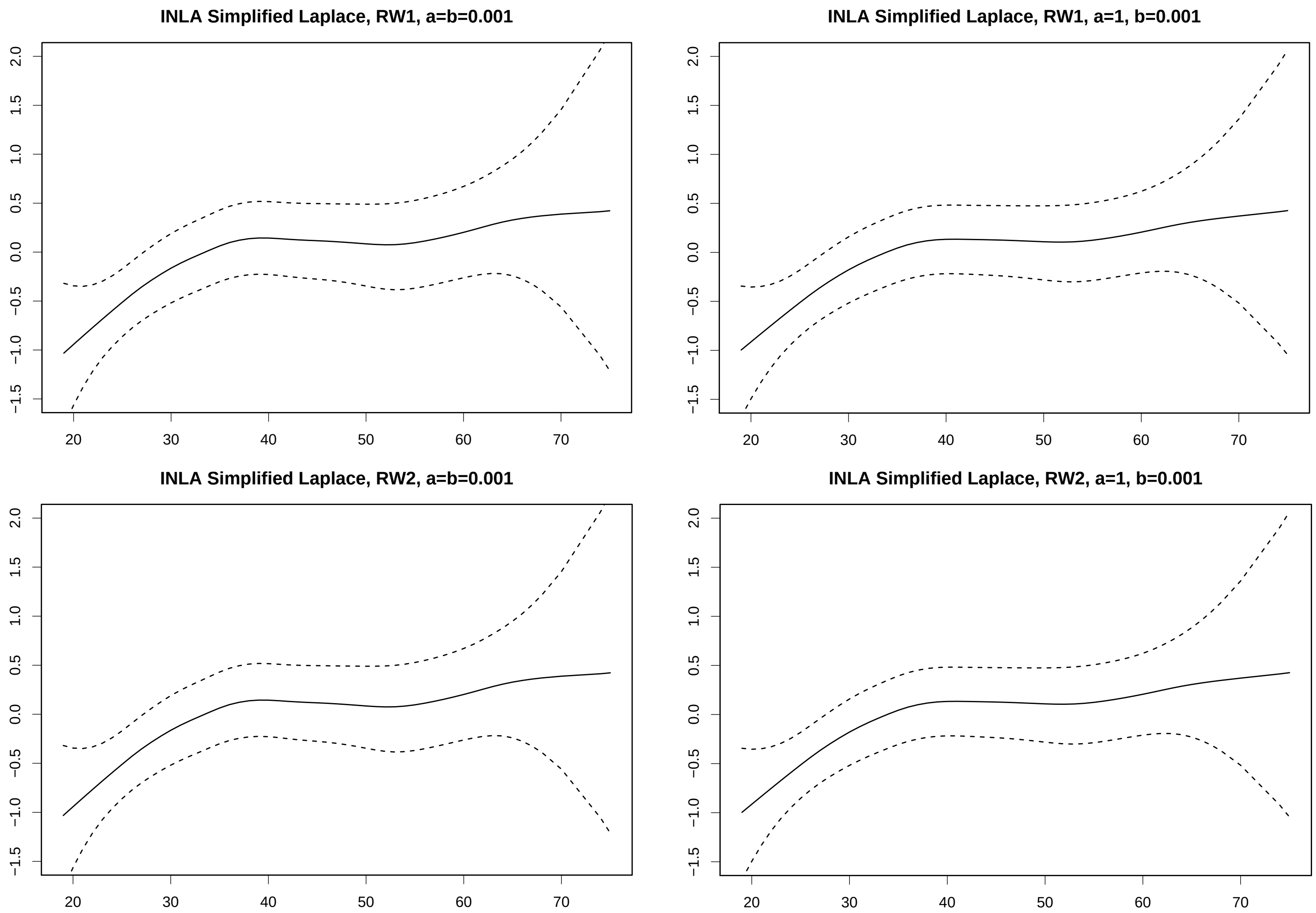
- Effects of age obtained with one outlier excluded: Different random walk orders and hyperparameters for Gaussian Approximation



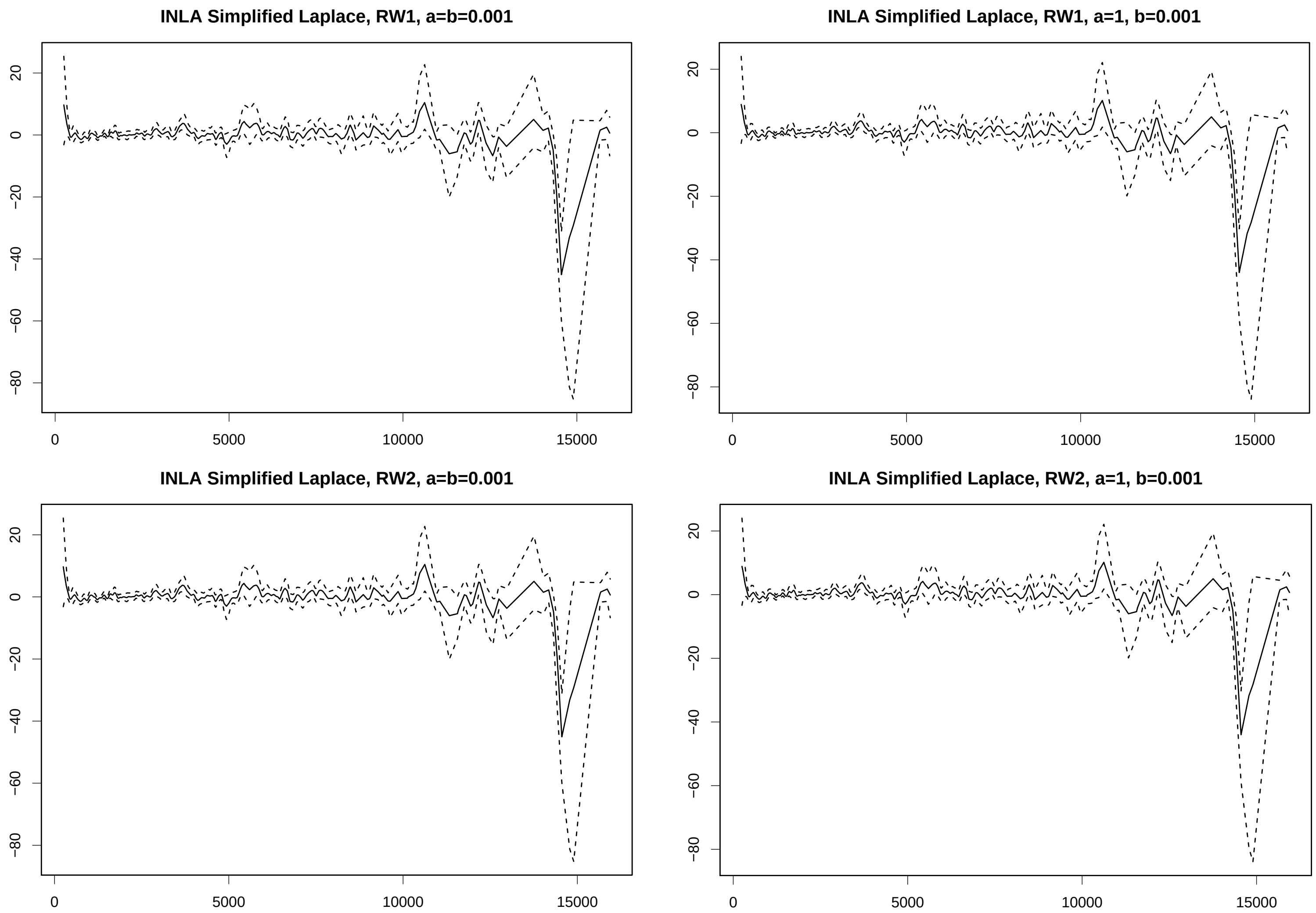
- Effects of amount obtained with one outlier excluded: Different random walk orders and hyperparameters for Gaussian Approximation



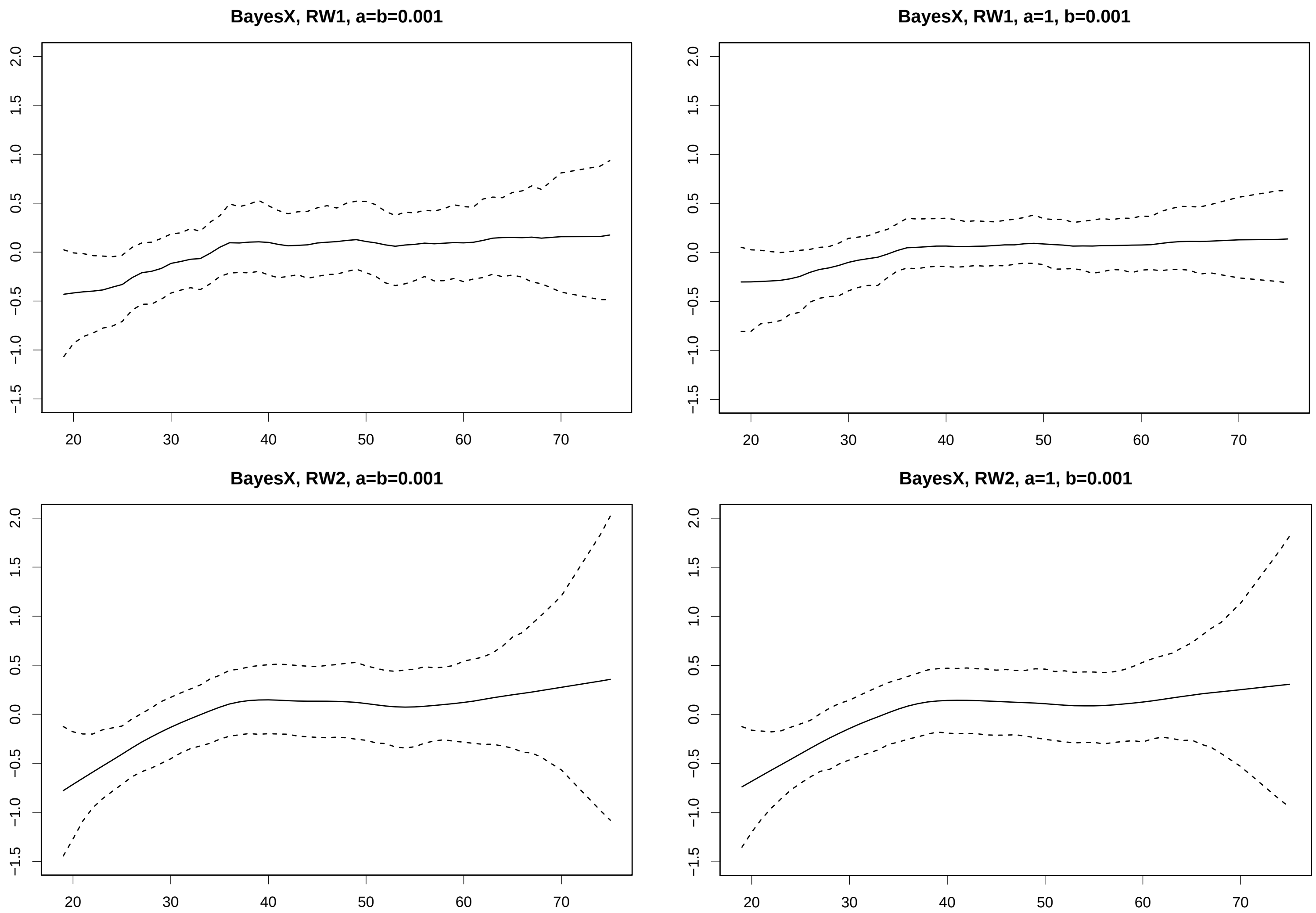
- Effects of age obtained with one outlier excluded: Different random walk orders and hyperparameters for Simplified Laplace



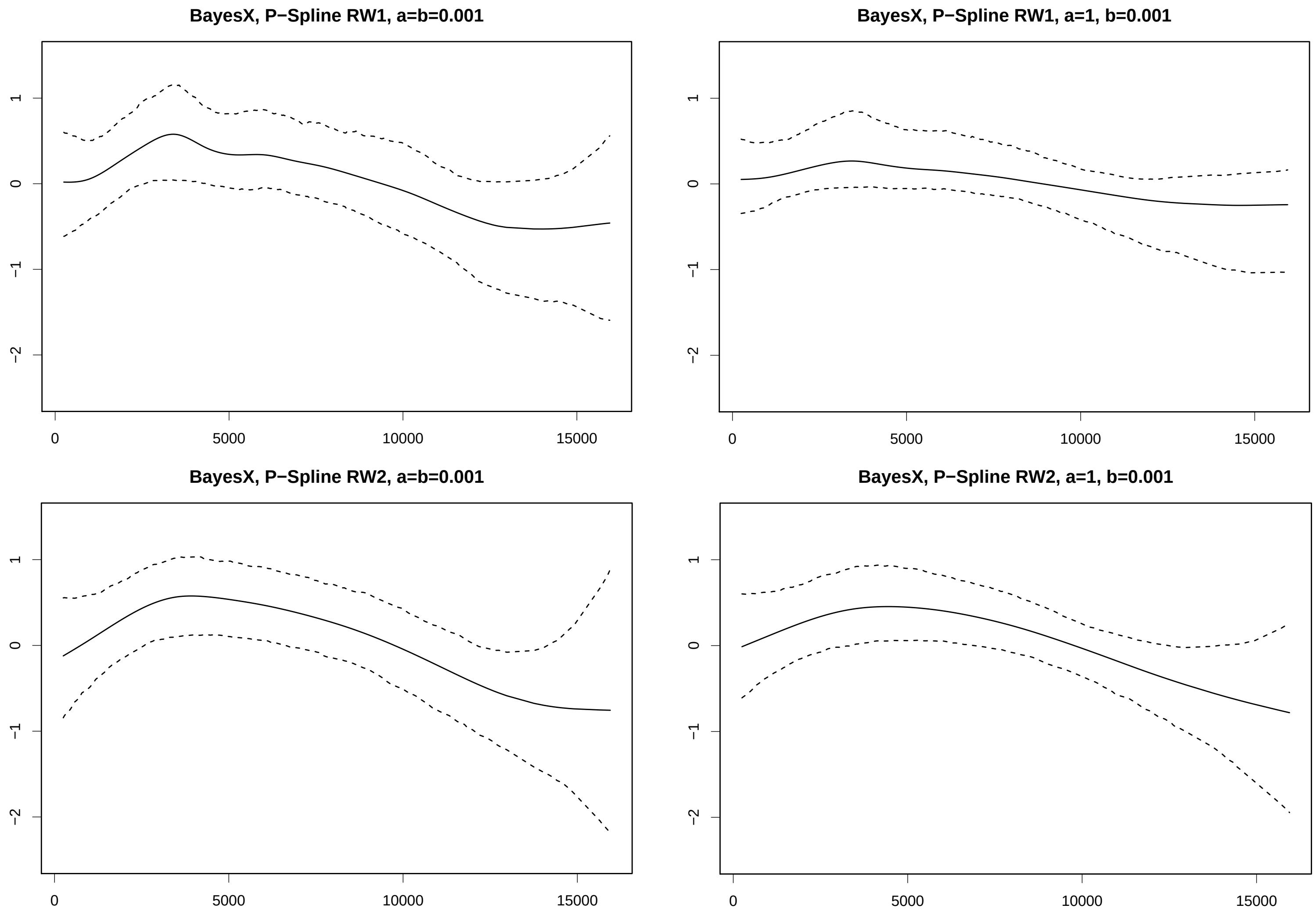
- Effects of amount obtained with one outlier excluded: Different random walk orders and hyperparameters for Simplified Laplace



- Effects of age obtained with one outlier excluded: Different random walk orders and hyperparameters for BayesX



- Effects of amount obtained with one outlier excluded: Different random walk orders and hyperparameters for BayesX



Summary and Discussion

- Conditionally Gaussian models provide a rich class of regression models.
- BayesX and INLA provide comparable estimates in well-behaved examples but results may differ substantially in difficult situations.
- In particular, covariates with outliers seem to yield highly variable estimates with INLA.
- Differences in computing times not always as expected (full Laplace approximation may be slow).
- In particular, covariates with a large number of different covariate values yield long computing times.

- Suggestions for improving INLA:
  - Provide characterisations for “difficult” data sets?
  - Implement Bayesian P-splines instead of random walk priors (faster and more stable)?
  - Revise default prior choice for hyperparameters?
- Further questions:
  - Flexibility in terms of hyperprior choices (further hierarchical levels)?
  - Partial impropriety of the conditionally Gaussian priors and model choice quantities.