Bayesian inference and INLA Introduction

King Abdullah University of Science and Technology
جامعة الملك عبدالله
للعلوم والتقنية

Janet.vanNiekerk@kaust.edu.sa

July 2024
The repo for this workshop is here
https://tinyurl.com/INLAUP24

QR code for repository

https://tinyurl.com/INLAUP24

Please run the Libraries.R file to ensure you have the relevant libraries for this workshop

Outline

BayesComp group at KAUST

Bayesian inference

noise

hyper parameter | nuisance

link

Data $\boldsymbol{y}$ (with covariates $\boldsymbol{Z}$), depend on $\boldsymbol{X}$ and $\boldsymbol{\theta}$ such that, $\boxed{E[Y]} = h(\boldsymbol{A}(\boldsymbol{Z})\boxed{\boldsymbol{X}})$.

ϕ

Bayes' theorem:

Structure

$$\underbrace{q(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{y})}_{\text{Posterior}} \propto \underbrace{L(\boldsymbol{y}|\boxed{\boldsymbol{X}, \boldsymbol{\theta}})}_{\text{Likelihood}} \times \underbrace{\boxed{p(\boldsymbol{X}, \boldsymbol{\theta})}}_{\text{Prior}}$$

uncertainty quantification

Computational aspects

- Analytical methods - conjugacy (pre-computer era)
- Approximate methods - Laplace (can be inaccurate)
- Exact methods - MCMC (very slow for complex models or large data)

Now, due to computing resources approximate methods are gaining popularity - INLA, VB, EP etc.

Methodology

R library

INLA - 2009 [Rue et al., 2009]

2021+ [Van Niekerk et al., 2023]

HPC [Gaedke-Merzhäuser et al., 2023]

Model definition - GAMM

anything

Suppose we have response data $\boldsymbol{y}_{n\times 1}$ (conditionally independent) with density function $\pi(y|\boldsymbol{X},\boldsymbol{\theta})$ and link function $h(.)$, that is linked to some covariates $\boldsymbol{Z}$ through linear predictors

random effects (structured)

$$\boldsymbol{\eta}_n = \beta_0 + \boldsymbol{Z}_\beta\boldsymbol{\beta} + \sum f^k(\boldsymbol{Z}_f) = \boldsymbol{AX}$$

random

The inferential aim is to estimate the latent field $\boldsymbol{X}_m = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$, and $\boldsymbol{\theta}$.

fixed

fixed effects

GAMM → LGM

Latent Gaussian model

$\sim N(\mu, \Sigma)$

covariance matrix

inverse

precision matrix

Assume

$$\boldsymbol{X}|\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{Q}(\boldsymbol{\theta})^{-1})$$

where $\boldsymbol{Q}(\boldsymbol{\theta})$ is a sparse matrix ($\boldsymbol{X}$ is a GMRF).

anything

Default ≡ "PC" Penalizing Complexity Prior

$p(\boldsymbol{X}, \boldsymbol{\theta}) = p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ can be non-Gaussian.

Why is INLA so accurate and so fast?

Integrated ~~Nested~~ Laplace Approximation

VB correction

- LGM structure ✓
- Sparse precision matrix ✓
- Specialized matrix algebra for sparse matrices
- NEW: VB (low-rank) correction [van Niekerk and Rue, 2024]

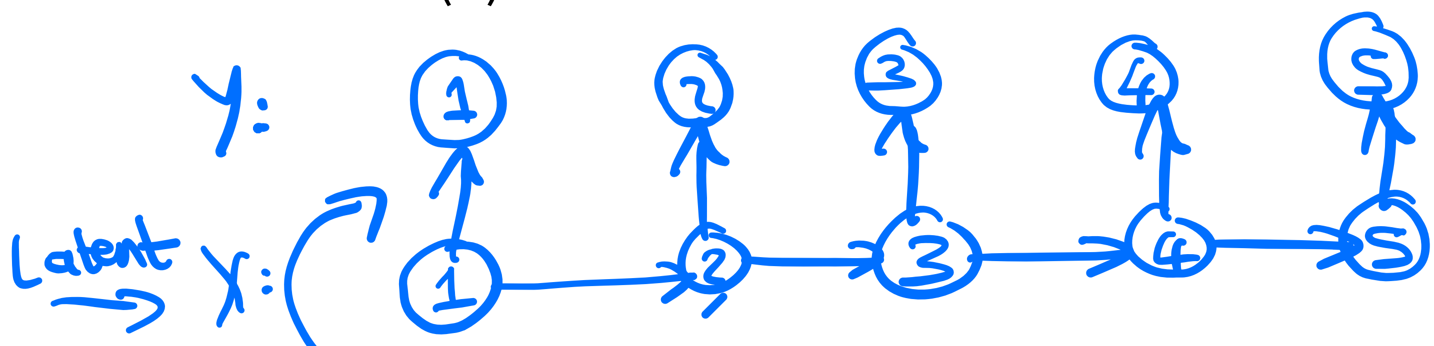
Use precision matrix instead of covariance matrix → natural occurrence

$X \sim N(\mu, \Sigma)$ → Q sparse

$$f(X) \propto \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\right)$$

How common are sparse $\boldsymbol{Q}(\theta)$?

Consider an AR(1) model..



} LGM ✓

GAMM

$Y \sim N(\mu, \boxed{\sigma^2})$, $\mu_i = u_i$

p(x) ~ Normal

"T large"

$N(0, \boxed{\tau})$

conditional

(1) $x_i \,|\, x_{i-1} \stackrel{d}{=} \boxed{\phi}\, x_{i-1} + \boxed{\varepsilon_t}$

$$f(x_1, x_2, \ldots, x_T) = f(x_1)\, f(x_2|x_1)\, f(x_3|x_2) \ldots f(x_T|x_{T-1})$$

AR(1) example

$$\propto \exp\left(-\frac{1}{2}\tau x_1^2\right) \exp\left(-\frac{1}{2}\tau\left(x_2 - \phi x_1\right)^2\right) \to x_2^2 - 2\phi x_1 x_2 + \phi^2 x_1^2$$

$$\exp\left(-\frac{1}{2}\tau\left(x_3 - \phi x_2\right)^2\right) \cdots \exp\left(-\frac{1}{2}\tau\left(x_T - \phi x_{T-1}\right)^2\right)$$

$$= \exp\left(-\frac{1}{2}\tau \begin{bmatrix} x_1 & x_2 & \cdots & x_T \end{bmatrix} \begin{bmatrix} 1+\phi^2 & -\phi & & \\ -\phi & 1+\phi^2 & -\phi & \\ & -\phi & \ddots & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{bmatrix}\right)$$

$$= \exp\left(-\frac{1}{2}\tau [\underline{x} - \underline{0}]^T \Sigma^{-1} [\underline{x} - \underline{0}]\right)$$

$T \times T$ $\quad \Sigma^{-1} = Q$

non-zero $\quad T + 2(T-1) \ll T^2 : Q$

T^2 nonzero $: \Sigma$

AR(1) example

$$\begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & f_1 & f_2 & f_3 \end{bmatrix} \begin{pmatrix} \bullet & & & & & \\ & \bullet & & & & \\ & & \bullet & & & \\ & & & \boxed{} & & \\ & & & & \boxed{} & \\ & & & & & \boxed{} \end{pmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ f_1 \\ f_2 \\ f_3 \end{bmatrix}$$

Posterior approximations by INLA

LGM's: X huge $\gg n$ but θ small

$$\pi(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{y}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{X}|\boldsymbol{\theta})\prod_{i=1}^{n}\pi(y_i|(\boldsymbol{AX})_i,\boldsymbol{\theta})$$

$$\boxed{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})} \propto \left.\frac{\pi(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{y})}{\boxed{\tilde{\pi}_G(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{y})}}\right|_{\boldsymbol{X}=\boldsymbol{\mu}(\boldsymbol{\theta})}$$

$$\tilde{\pi}(\theta_j|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}$$

$$\tilde{\pi}(X_j|\boldsymbol{y}) = \int \boxed{\tilde{\pi}(X_j|\boldsymbol{\theta},\boldsymbol{y})}\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

I_{INLA}

$\tilde{\pi}(X_j|\boldsymbol{\theta},\boldsymbol{y})$ depends on the approximation used, for Gaussian it is straightforward for the Laplace approximation we do another Gaussian approximation to $\cancel{\tilde{\pi}(\boldsymbol{X}_{-j}|\boldsymbol{\theta},\boldsymbol{y})}$.

Modern INLA

The Gaussian approximation $\pi_G(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{y})$ to $\pi(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{y})$ is calculated from a second order expansion of the likelihood around the mode of $\pi(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{y})$, $\boldsymbol{\mu}(\boldsymbol{\theta})$ as follows

$$\log\left(\pi(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{y})\right) \propto -\frac{1}{2}\boldsymbol{X}^\top \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{X} + \sum_{i=1}^{n}\left(b_i(\boldsymbol{A}\boldsymbol{X})_i - \frac{1}{2}c_i(\boldsymbol{A}\boldsymbol{X})_i^2\right)$$

$(Q(\theta)+A^\top DA)\boxed{\mu} = b$

$$= -\frac{1}{2}\boldsymbol{X}^\top \underbrace{\left(\boldsymbol{Q}(\boldsymbol{\theta}) + \boldsymbol{A}^\top\boldsymbol{D}\boldsymbol{A}\right)}_{\text{Precision matrix}}\boldsymbol{X} - \boxed{\boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{X}}$$

where $\boldsymbol{b}$ is an n-dimensional vector with entries $\{b_i\}$ and $\boldsymbol{D}$ is a diagonal matrix with n entries $\{c_i\}$. Note that both $\boldsymbol{b}$ and $\boldsymbol{D}$ depend on $\boldsymbol{\theta}$, so the Gaussian approximation is for a fixed $\boldsymbol{\theta}$.

$\therefore X|\theta,y \sim \underline{N}(\mu + \overset{VB}{\delta}, Q(\theta)+A^\top DA)$

Modern INLA

$\therefore x_i \sim N(\mu_i, \sigma_i^2$

$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \cdot & \cdot \\ \cdot & \sigma_2^2 & \cdot \\ \cdot & \cdot & \sigma_3^2 \end{bmatrix} \right)$

The process is iterated to find $\boldsymbol{b}$ and $\boldsymbol{D}$ that gives the Gaussian approximation at the mode, $\boldsymbol{\mu}(\boldsymbol{\theta})$, so that

$$\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{y} \sim N\left(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{Q}_{\boldsymbol{X}}^{-1}(\boldsymbol{\theta})\right).$$

The graph of the Gaussian approximation consists of two components,

1. $\mathcal{G}_p$: the graph obtained from the prior of the latent field through $\boldsymbol{Q}(\boldsymbol{\theta})$
2. $\mathcal{G}_d$: the graph obtained from the data based on the non-zero entries of $\boldsymbol{A}^\top \boldsymbol{A}$

Modern INLA

Next, the marginal conditional posteriors of the elements of $\boldsymbol{X}$ is calculated from the joint Gaussian approximation as

$$\boldsymbol{X}_j|\boldsymbol{\theta},\boldsymbol{y} \sim N\left((\boldsymbol{\mu}(\boldsymbol{\theta}))_j\,,\,(\boldsymbol{Q}_{\boldsymbol{X}}^{-1}(\boldsymbol{\theta}))_{jj}\right).$$

and the marginals

$$\tilde{\pi}(\boldsymbol{X}_j|\boldsymbol{y}) = \int \boxed{\pi_G(\boldsymbol{X}_j|\boldsymbol{\theta},\boldsymbol{y})}\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \approx \sum_{k=1}^{K} \pi_G(\boldsymbol{X}_j|\boldsymbol{\theta}_k,\boldsymbol{y})\underline{\tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y})\delta_k}.$$

Many options

OG: Nested Laplace

strategy = "laplace"

strategy = "gaussian"

uncorrected

Conditional posterior of η_i

Post fit

$\eta = \beta_0 + \beta_1 \ldots + f_1 + \ldots f_2 \ldots$

In order to calculate $\tilde{\pi}(\eta_i|\boldsymbol{y})$, we first calculate $\tilde{\pi}(\eta_i|\boldsymbol{\theta},\boldsymbol{y})$. We postulate a Gaussian density for $\eta_i|\boldsymbol{\theta},\boldsymbol{y}$ such that $\tilde{\pi}(\eta_i|\boldsymbol{\theta},\boldsymbol{y}) = \pi_G(\eta_i|\boldsymbol{\theta},\boldsymbol{y})$, with mean

$$E(\boldsymbol{\eta}|\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{A}E(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{\theta})$$

and covariance matrix

$$\text{Cov}(\boldsymbol{\eta}|\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{A}^\top,$$

VB corrected marginal posterior of η_i

$$
\begin{aligned}
\eta_j|\boldsymbol{\theta},\boldsymbol{y} &\sim N(\mu_j(\boldsymbol{\theta}),\sigma_j^2(\boldsymbol{\theta}))\\
\mu_j(\boldsymbol{\theta}) &= (\boldsymbol{A}\boldsymbol{\mu}^*(\boldsymbol{\theta}))_j\\
\tilde{\pi}(\eta_j|\boldsymbol{y}) &\approx \sum_{k=1}^{K}\pi_G(\eta_j|\boldsymbol{\theta}_k,\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y})\delta_k.
\end{aligned}
$$

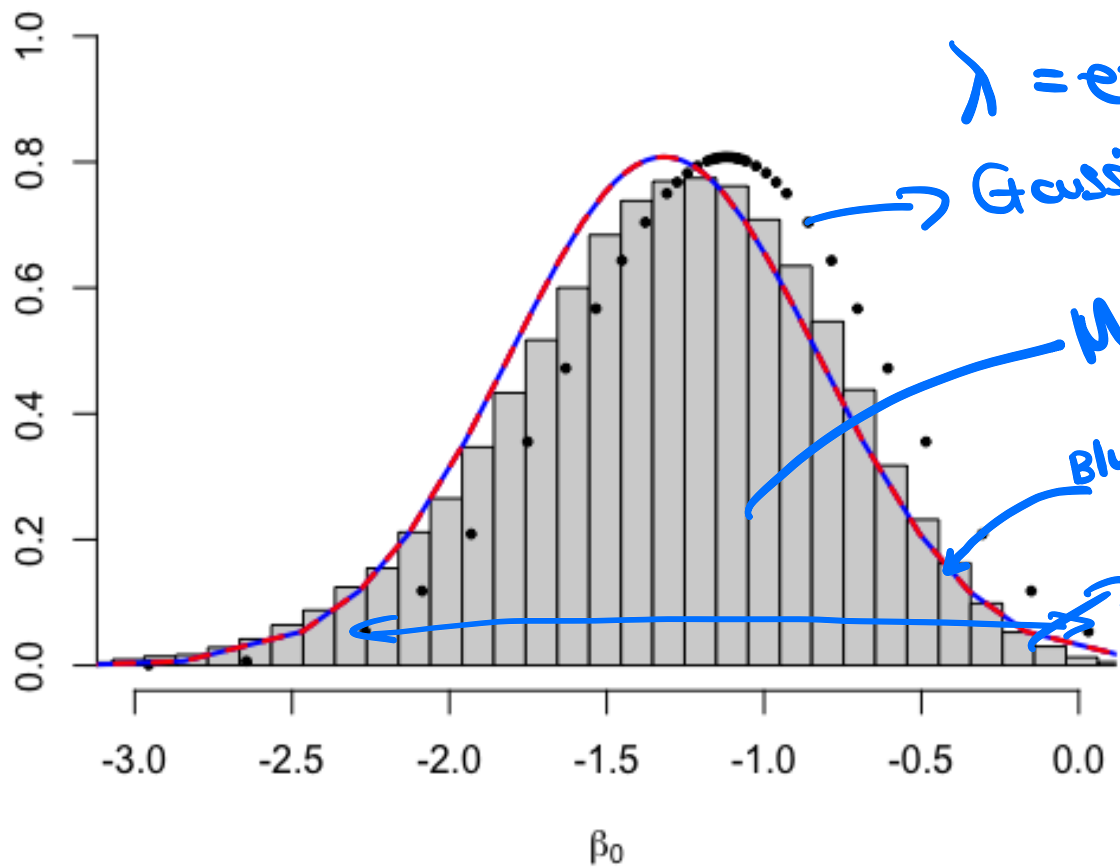
Example (small data)

Poisson $\xrightarrow[\infty]{\Delta}$ Normal

link

$$Y_i|\beta_0 = -1, \beta_1 = -0.5 \quad \sim \quad \text{Poisson}(\exp(\beta_0 + \beta_1 X_i))$$

$\lambda = \exp(\beta_0 + \beta_1 X_i)$



Gaussian (uncorrected)

MCMC

Blue "Laplace"

Red "VB"

Example (large data)

$$Y_i|\beta_0 = -1, \beta_1 = -0.5 \quad \sim \quad \text{Poisson}(\exp(\beta_0 + \beta_1 X_i))$$

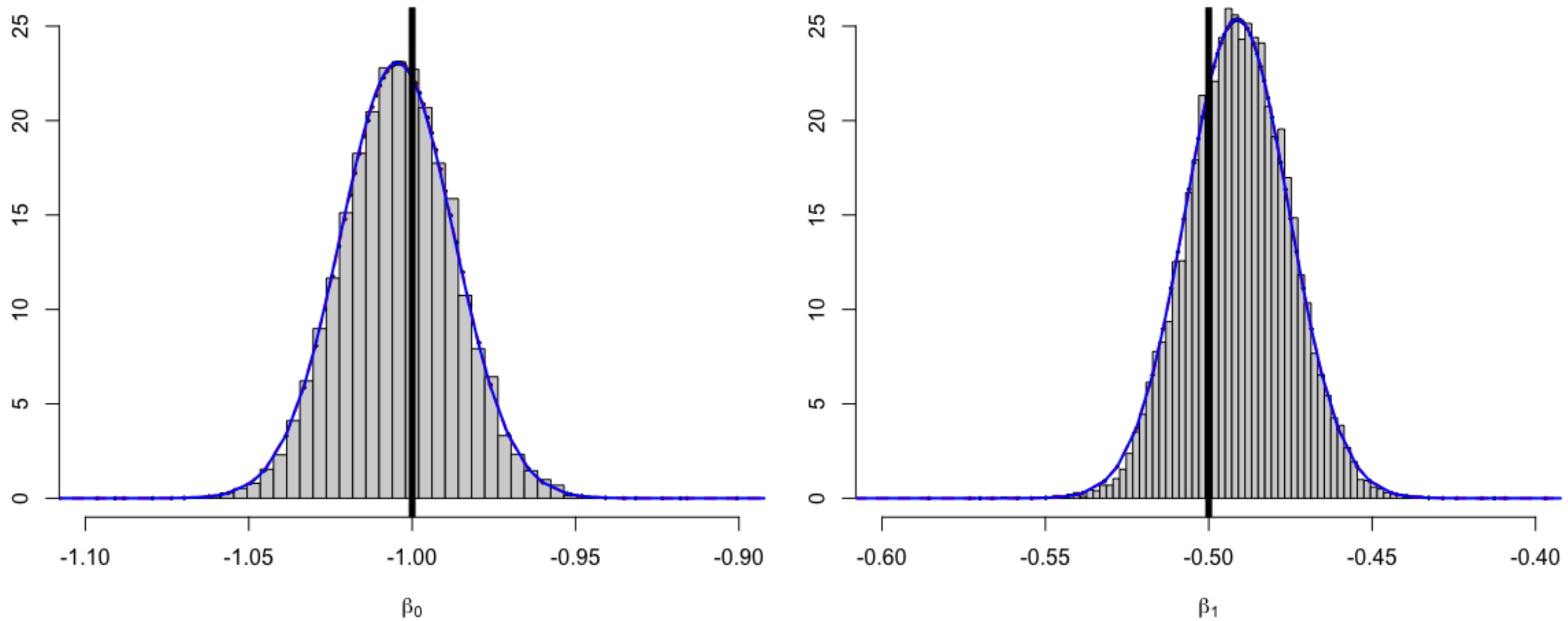


Figure: Marginal posterior of β_0 (center) and β_1 (right) from the Laplace method (points), VBC (solid line) and INLA (broken line) approximations

Tokyo example

The Tokyo dataset in the R INLA library contains information on the number of times the daily rainfall measurements in Tokyo was more than $1mm$ on a specific day t for two consecutive years. In order to model the annual rainfall pattern, a stochastic spline model with fixed precision is used to smooth the data

RW1 or RW2

2 $\in (0,1)$ $\eta \in \boxed{\mathbb{R}}$

$$
\begin{aligned}
y_i \mid \mathcal{X} &\sim Bin\left(n_i, p_i = \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\right)\\
(\alpha_{i+1} - 2\alpha_i + \alpha_{i-1}) \mid \tau &\overset{\text{iid}}{\sim} N(0, \tau^{-1}),
\end{aligned}
$$

where $i = 1, 2, \ldots, 366$ on a torus, and $n_{60} = 1$ else $n_i = 2$.

$\alpha_1 \longrightarrow \alpha_{366}$

Results

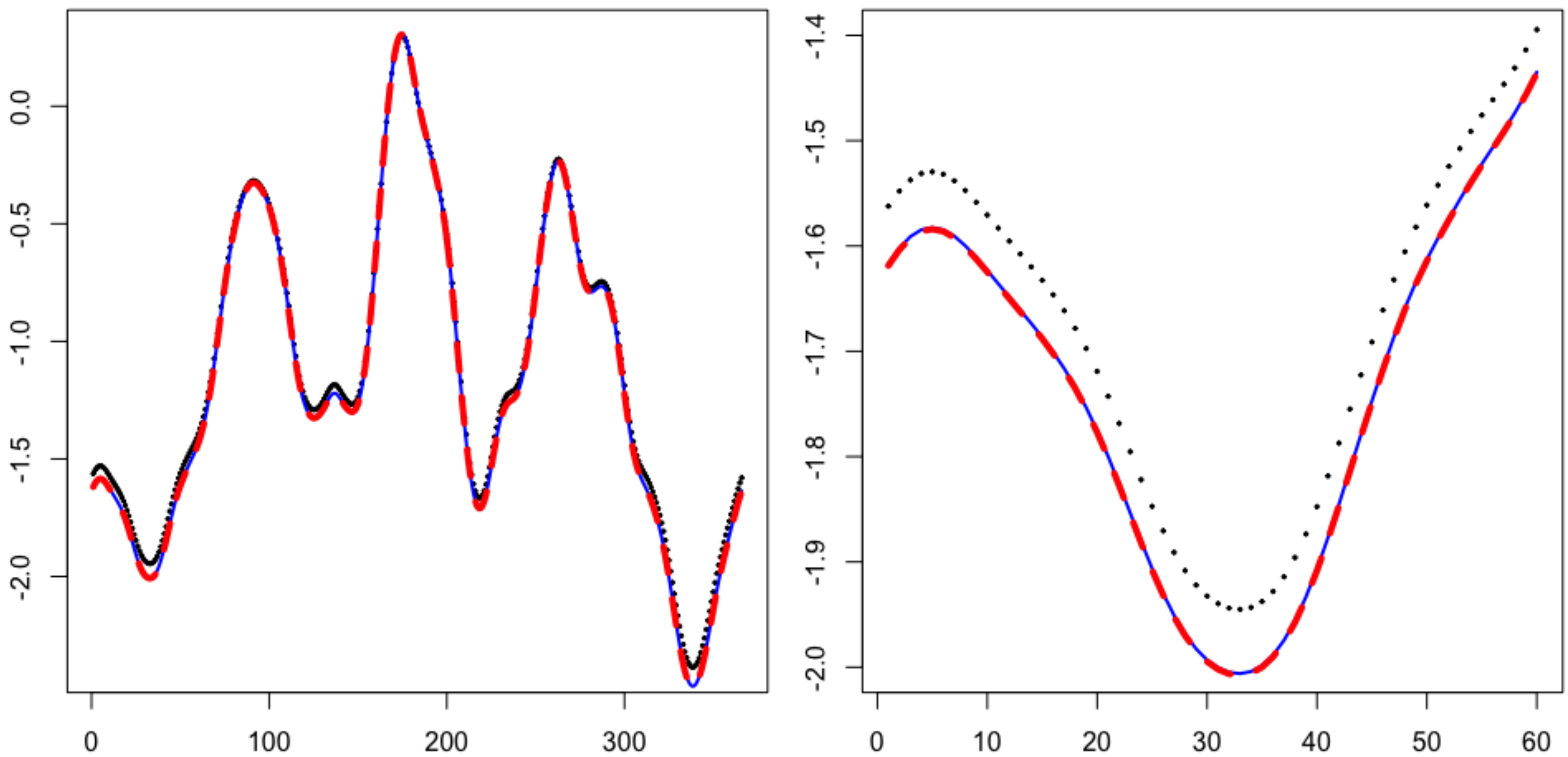
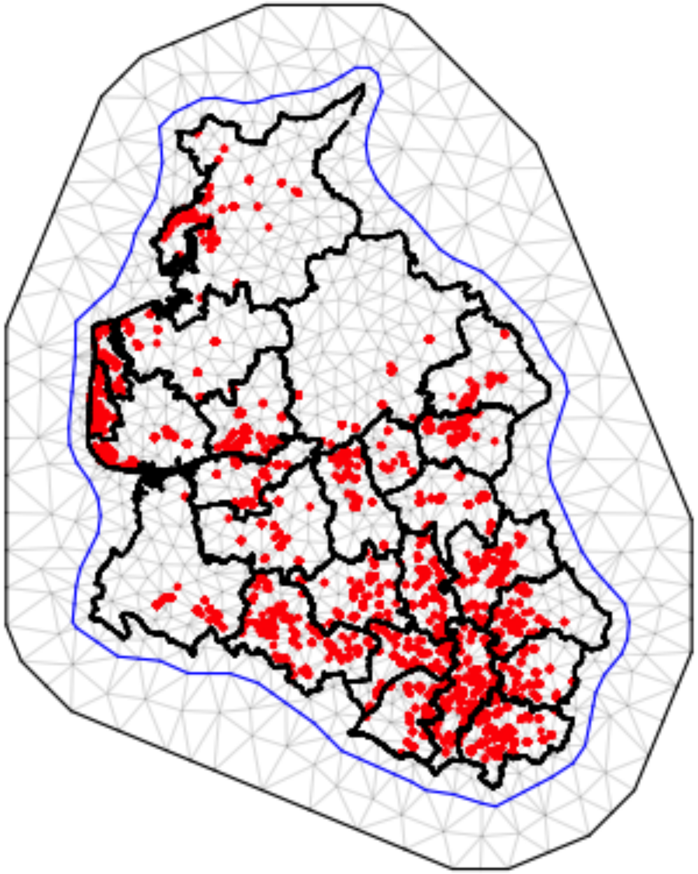


Figure: Posterior mean of α (left) (zoomed for the first two months (right)) from the Laplace method (points), VBC (solid line) and INLA (broken line)

Spatial survival example

Consider the R dataset `Leuk` that features the survival times of 1043 patients with acute myeloid leukemia (AML) in Northwest England between 1982 to 1998.



Cox spatial model

$$h(t, \boldsymbol{s}) = h_0(t) \exp(\beta \boldsymbol{X} + \boldsymbol{u}(\boldsymbol{s})),$$

with

$$\eta_i(s) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WBC}_i + \beta_3 \text{TPI}_i + \boxed{u(s).}$$

which implies a latent field of size $m = \boxed{39158.}$

Results

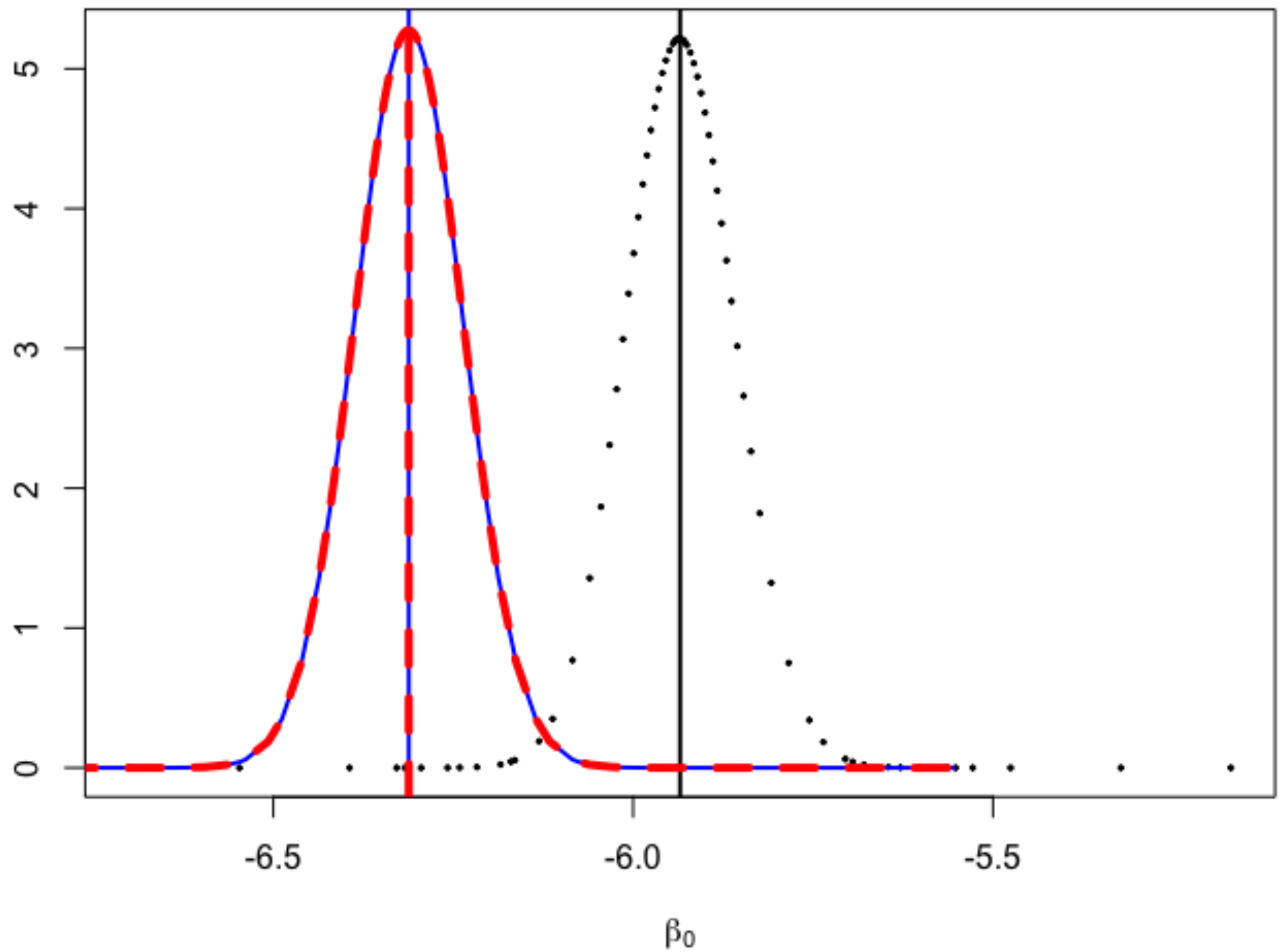


Figure: Marginal posteriors from the Laplace method (points), VBC (solid line) and INLA (broken line)

Results

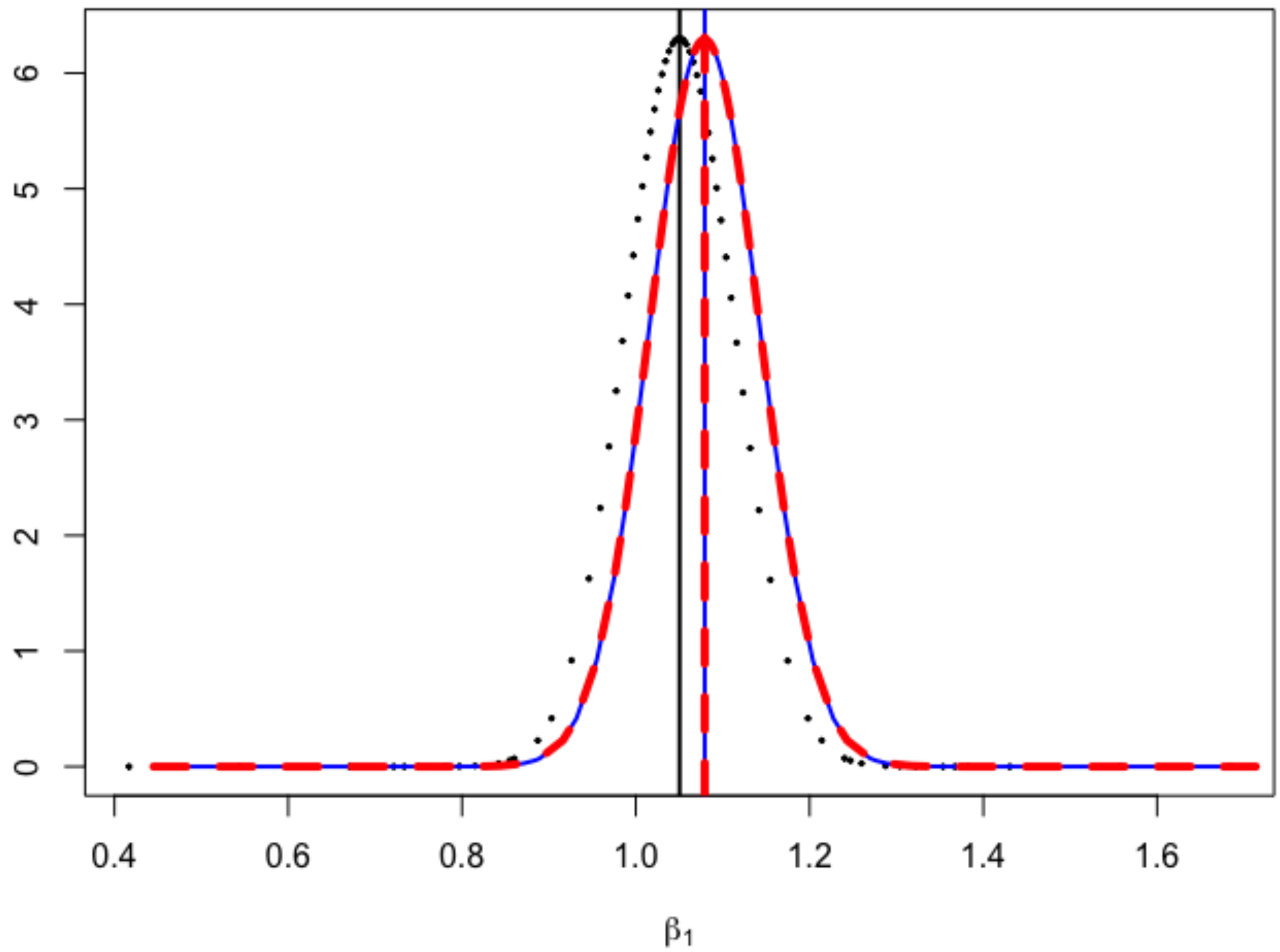


Figure: Marginal posteriors from the Laplace method (points), VBC (solid line) and INLA (broken line)

Results

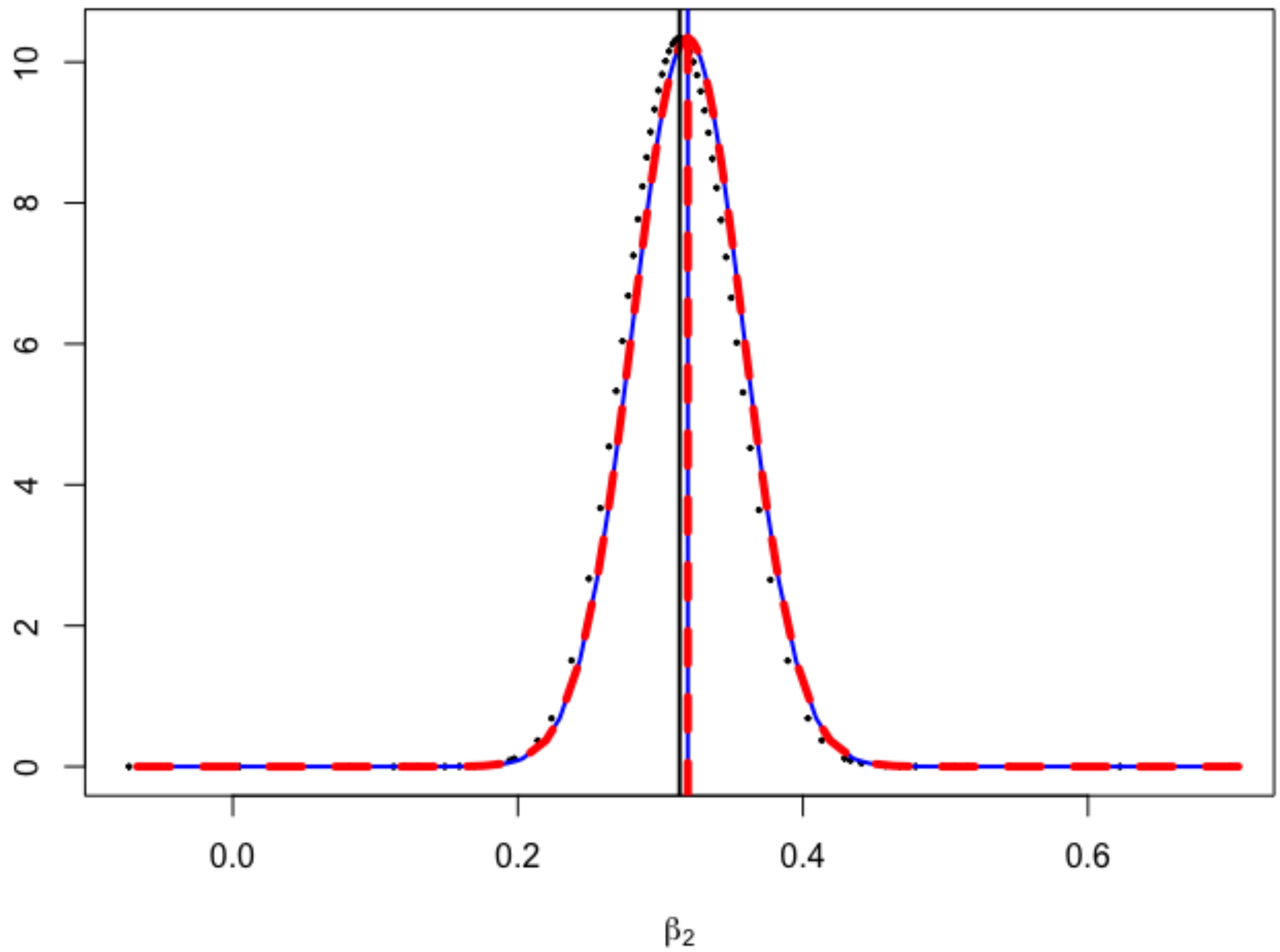


Figure: Marginal posteriors from the Laplace method (points), VBC (solid line) and INLA (broken line)

Cox proportional hazards model

We simulate survival data for n patients using the following very simple Cox proportional hazards model

$$h_i(t) = h_0(t)\exp(\beta x_i) = 1.2t^{0.2}\exp\left(0.1x_i\right), \quad 1 = 1, 2, \ldots, n,$$

where x is a scaled and centered continuous covariate, and the baseline hazard, $h_0(t)$ is estimated using a scaled random walk order one model with 50 bins. We also consider four different values of n which are $n = 10^2$, to 10^5.

Cox proportional hazards model

n	Augmented size	classic INLA (s)	modern INLA (s)
10^2	1 327	1.6	0.1
10^3	12 657	1.3	0.4
10^4	131 807	10.2	2.3
10^5	1 302 413	113.3	22.5

Table: Results from simulation of Cox proportional hazards model

cs-fMRI model

Functional magnetic resonance imaging (fMRI) is a noninvasive neuro-imaging technique used to localize regions of specific brain activity during certain tasks.
For T timepoints and N vertices per hemisphere resulting in data $\boldsymbol{y}_{TN\times 1}$ with the latent Gaussian model as follows:

smooth signals

"noise"

$$
\begin{aligned}
\boldsymbol{y}|\beta,\boldsymbol{b},\boldsymbol{\theta} &\sim N(\boldsymbol{\mu}_y,\boldsymbol{V}), \quad \boldsymbol{\mu}_y = \underbrace{\sum_{k=0}^{K}\boldsymbol{X}_k\beta_k}_{} + \underbrace{\sum_{j=1}^{J}\boldsymbol{Z}_j\boldsymbol{b}_j}_{} \\
\beta_k &= \boldsymbol{\Psi}_k\boldsymbol{w}_k \quad \text{(SPDE prior on } \beta_k\text{)} \\
\boldsymbol{w}_k|\boldsymbol{\theta} &\sim N(\boldsymbol{0},\boldsymbol{Q}^{-1}_{\tau_k,\kappa_k}) \\
\boldsymbol{b}_j &\sim N(\boldsymbol{0},\delta\boldsymbol{I}) \quad \text{(Diffuse priors for } \boldsymbol{b}_j\text{)} \\
\boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}),
\end{aligned}
$$

where we have K task signals and J nuisance signals.

cs-fMRI model

The data consists of a 3.5-min fMRI for each subject, consisting of 284 volumes, where each subject performs 5 different motor tasks interceded with a 3 second visual cue. Each hemisphere of the brain contained 32492 surface vertices. From these, 5000 are resampled to use for the analysis. This results in a response data vector $\boldsymbol{y}$ of size **2 523 624**, with an SPDE model defined on a mesh with 8795 triangles.

The inference based on the modern formulation of INLA was computed in 148 seconds.

cs-fMRI model

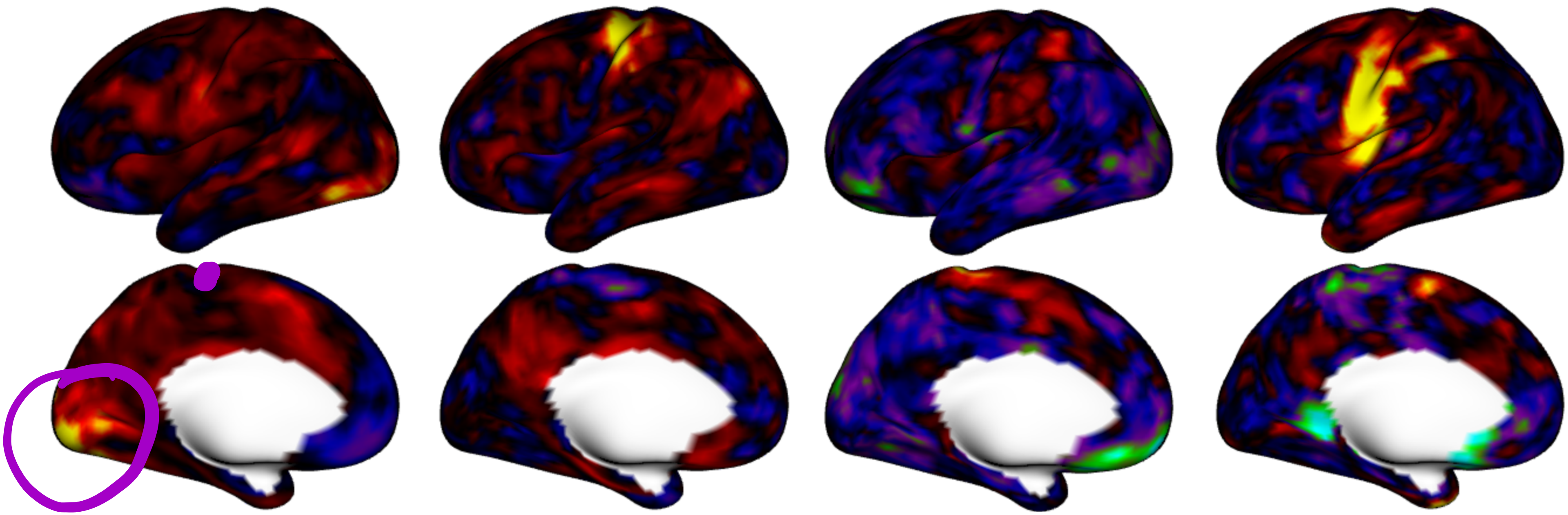


Figure: Activation areas for the different tasks in the left hemisphere - visual cue, right hand motor, right foot motor, tongue motor task (from left to right)

Further details

www.r-inla.org

New default setting in INLA (VB) (previously inla.mode = "experimental")

- INLA can fit many different statistical models and complex models can be built using multiple "building blocks"/random effects.
- Remove the linear predictors from the latent field $\rightarrow$ accurate posterior inference with VB correction (I - VB - LA)
- New applications that aren't feasible with INLA 1.0

Gaedke-Merzhäuser, L., van Niekerk, J., Schenk, O., and Rue, H. (2023).
Parallelized integrated nested Laplace approximations for fast Bayesian inference.
Statistics and Computing, 33(1):25.

Rue, H., Martino, S., and Chopin, N. (2009).
Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(2):319–392.

Van Niekerk, J., Krainski, E., Rustand, D., and Rue, H. (2023).
A new avenue for Bayesian inference with INLA.
Computational Statistics & Data Analysis, 181:107692.

van Niekerk, J. and Rue, H. (2024).
Low-rank variational Bayes correction to the Laplace method.
Journal of Machine Learning Research, 25(62):1–25.

Thank you • شكرا

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology