INLA 2.0 - into the future

King Abdullah University of Science and Technology

جامعة الملك عبدالله للعلوم والتقنية

May 2022

Outline

Model definition

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

$$\boldsymbol{\eta}_{n\times 1}=\beta_0\mathbf{1}+\boldsymbol{\beta}\boldsymbol{X}+\sum_{k=1}^{K}f^k(\boldsymbol{u}_k)$$

The inferential aim is to estimate the latent field $\mathcal{X}_{m_*\times 1}=\{\beta_0,\boldsymbol{\beta},\boldsymbol{f}\}$. Define the augmented latent field

$$\mathcal{X}_{m\times 1}=\{\boldsymbol{\eta},\beta_0,\boldsymbol{\beta},\boldsymbol{f}\}.$$

Posterior approximations

$$
\begin{aligned}
\pi(\mathcal{X}, \boldsymbol{\theta}, \boldsymbol{y}) &= \pi(\boldsymbol{\theta})\pi(\mathcal{X}|\boldsymbol{\theta})\prod_{i=1}^{n}\pi(y_i|\mathcal{X}_i, \boldsymbol{\theta}) \\
\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \left.\frac{\pi(\mathcal{X}, \boldsymbol{\theta}, \boldsymbol{y})}{\pi_G(\mathcal{X}|\boldsymbol{\theta}, \boldsymbol{y})}\right|_{\mathcal{X}=\boldsymbol{\mu}(\boldsymbol{\theta})} \\
\tilde{\pi}(\theta_j|\boldsymbol{y}) &= \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j} \\
\tilde{\pi}(\mathcal{X}_j|\boldsymbol{y}) &= \int \tilde{\pi}(\mathcal{X}_j|\boldsymbol{\theta}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},
\end{aligned}
$$

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

Thoughts and ideas

- For large data, $\mathcal{X}$ is large - can we remove $\boldsymbol{\eta}$, but still produce cheap and accurate inference for $\boldsymbol{\eta}$.
- For stability we can remove the "noisy" linear predictors (non-singularity)

Modern framework

The latent field is defined as

$$\boldsymbol{\mathcal{X}} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\},$$

and the n linear predictors are defined as

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{\mathcal{X}}, \tag{1}$$

with $\boldsymbol{A}$ a sparse design matrix that links the linear predictors to the latent field.

Modern framework

From this formulation

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

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

Modern framework

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

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

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 $\boldsymbol{\theta}$.

Modern framework

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

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

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 framework

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

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

and the marginals

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

Conditional posterior of η_i

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

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

and covariance matrix

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

Marginal posterior of η_i

Now let $\boldsymbol{C}$ be a sparse selected inverse of $\boldsymbol{Q}_{\mathcal{X}} = \boldsymbol{Q} + \boldsymbol{A}^\top \boldsymbol{D}\boldsymbol{A}$ based on the graph $\mathcal{G}_{\mathcal{X}} = \{\mathcal{G}_p, \mathcal{G}_d\}$

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

so that $\mathrm{E}(\eta_j|\boldsymbol{y}) = \mu_j$ and $\mathrm{Var}(\eta_j|\boldsymbol{y}) = \sigma_j^2$.

VB correction to latent field posteriors

VB mean correction paper: `https://arxiv.org/abs/2111.12945`

Let

$$\boldsymbol{\mu}^*(\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\theta}) + \boldsymbol{M\lambda},$$

with

$$\arg_{\boldsymbol{\lambda}} \min \Big(E_{\mathcal{X}|\boldsymbol{y},\boldsymbol{\theta} \sim N(\boldsymbol{\mu}(\boldsymbol{\theta})+\boldsymbol{M\lambda}, \boldsymbol{Q}_{\mathcal{X}}^{-1}(\boldsymbol{\theta}))} [-\log l(\boldsymbol{\mathcal{X}}|\boldsymbol{y})] \\ + \frac{1}{2} (\boldsymbol{\mu}(\boldsymbol{\theta}) + \boldsymbol{M\lambda})^\top \boldsymbol{Q}(\boldsymbol{\theta}) (\boldsymbol{\mu}(\boldsymbol{\theta}) + \boldsymbol{M\lambda}) \Big)$$

VB corrected marginal posterior of η_i

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

Cox proportional hazards model

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

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

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:

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

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

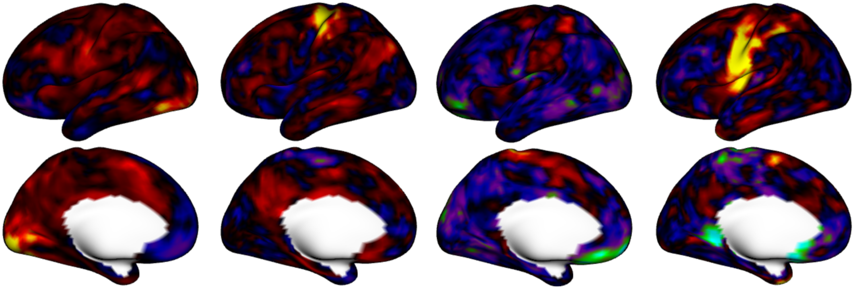

Figure: Activation areas for the different tasks in the left hemisphere - visual cue, right hand motor, right foot motor, tongue motor task (from left to right)

SPDE model

Consider an SPDE model with 948 mesh nodes and data n.

n	Classic INLA (s)	Modern INLA (s)
10^2	2.6	2.7
10^3	2.9	2.5
10^4	13.4	3.7

Table: Number of data points and the computing time (in seconds) considering the classic and modern INLA formulations

Stable prediction for SPDE model

Consider an SPDE model on a mesh with 946 nodes, conditional on 10^4 observations.

grid layout	size	Classic INLA (s)	Modern INLA (s)
250×150	37500	5.39	2.59
500×300	150000	17.70	4.68
1000×600	600000	156.48	13.69

Table: Number of predictions (grid layout and size) and the computing time (in seconds).

Discussion

inla(..., inla.mode = "experimental")

- INLA 2.0
- Remove the linear predictors from the latent field $\rightarrow$ accurate posterior inference with VB correction
- New applications that aren't feasible with INLA 1.0

Thank you • شكرا

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