A short introduction to INLA and R-INLA

with applications to extremes

Daniela Castro-Camilo & Thomas Opitz

EVA 2019 - Zagreb, Croatia

Table of contents

General `INLA` framework

General INLA framework

Bayes' theorem

$\Pr(\theta \mid \mathbf{y}) \propto \Pr(\mathbf{y} \mid \theta)\Pr(\theta)$

General INLA framework

Bayes' theorem

$\Pr(\theta \mid \mathbf{y}) \propto \Pr(\mathbf{y} \mid \theta)\Pr(\theta)$

- Mathematically simple.

General INLA framework

Bayes' theorem

$\Pr(\theta \mid \mathbf{y}) \propto \Pr(\mathbf{y} \mid \theta)\Pr(\theta)$

- Mathematically simple.
- Computationally complex.

General INLA framework

Bayes' theorem

$\Pr(\theta \mid \mathbf{y}) \propto \Pr(\mathbf{y} \mid \theta)\Pr(\theta)$

- Mathematically simple.
- Computationally complex.

- The integrated nested Laplace approximation (INLA; Rue et al., 2009) is a method to do approximate Bayesian inference for a specific class of models, called latent Gaussian models (LGMs).

General INLA framework

Bayes' theorem

$\Pr(\theta \mid \mathbf{y}) \propto \Pr(\mathbf{y} \mid \theta)\Pr(\theta)$

- Mathematically simple.
- Computationally complex.

- The integrated nested Laplace approximation (`INLA`; Rue et al., 2009) is a method to do approximate Bayesian inference for a specific class of models, called latent Gaussian models (LGMs).
- `INLA` is an alternative to simulation-based inference, such as Markov Chain Monte Carlo (MCMC).

General INLA framework

Bayes' theorem

$\Pr(\theta \mid \mathbf{y}) \propto \Pr(\mathbf{y} \mid \theta)\Pr(\theta)$

- Mathematically simple.
- Computationally complex.

- The integrated nested Laplace approximation (`INLA`; Rue et al., 2009) is a method to do approximate Bayesian inference for a specific class of models, called latent Gaussian models (LGMs).
- `INLA` is an alternative to simulation-based inference, such as Markov Chain Monte Carlo (MCMC).
- With `INLA`, posterior distributions of interest are numerically approximated using the Laplace approximation, therefore avoiding the usually complex updating schemes, long running times, and diagnostic convergence checks of MCMC.

General INLA framework

- Latent Gaussian models admit a hierarchical model formulation, whereby observations $\mathbf{y}$ are assumed to be conditionally independent given a latent Gaussian random field $\mathbf{x}$ and hyperparameters $\boldsymbol{\theta}_1$, i.e.,

$$\begin{aligned} \mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_1 &\sim \prod_{i=1}^{n} \mathrm{p}(y_i \mid x_i, \boldsymbol{\theta}_1), \\ \mathbf{x} \mid \boldsymbol{\theta}_2 &\sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}_2}, \mathbf{Q}_{\boldsymbol{\theta}_2}^{-1}), \\ \boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T &\sim \mathrm{p}(\boldsymbol{\theta}). \end{aligned}$$

- The latent Gaussian random field $\mathbf{x}$ describes the underlying dependence structure of the data. It has mean vector $\boldsymbol{\mu}_{\boldsymbol{\theta}_2}$ and precision matrix $\mathbf{Q}_{\boldsymbol{\theta}_2}$, which are controlled by the hyperparameters $\boldsymbol{\theta}_2$.
- We assume that y_i only depends on a linear predictor η_i, which has an additive structure with respect to some fixed covariates and random effects, i.e.,

$$\eta_i = \mu + \sum_{j=1}^{J} \beta_j z_{ji} + \sum_{k=1}^{K} f_k(w_{ki}), \quad i = 1, \ldots, n.$$

Likelihood, random effects, and priors in `INLA`

What models can we handle with INLA?

- The class of LGMs includes a wide variety of commonly applied statistical models.
- Large variety of response distributions for $\mathbf{y}$ and of random effects for η_i.

Likelihood		Random effects
Beta	Logistic	AR(p), AR1c
Binomial	Lognormal	Besag (CAR model)
Exponential	N. Binomial	Fractional Gaussian noise
Gamma	Poisson	Generic Matérn models
Gaussian	Skewnormal	Gaussian random effects
GEV	Student t	Measurement error model
GPD	Weibull	Random walks (orders 1 and 2)
G. Poisson	...	...

- More info: `inla.list.models("likelihood")` and `inla.list.models("latent")`.

Prior distributions

- Within the Bayesian framework, we put Gaussian prior distributions on fixed effects β_j and random effects $f_k(\cdot)$, which can be (relatively) uninformative or incorporate expert knowledge.

$\Rightarrow$ Linear predictor η_i is multivariate Gaussian.
$\Rightarrow$ When we have only few data, prior has a strong influence.

- We can estimate some hyperparameters controling
    - signal-to-noise ratio, smoothing $\Rightarrow$ precision parameters
    - range of dependence over space/time/..., e.g. range of Matérn correlation
    - shape of the response distribution, e.g. gamma shape, GEV/GPD shape.
- See `inla.list.models("prior")` for more info.

Examples

Linear model with two covariates

Model:	$y_i \mid (x, \theta) \sim \mathcal{N}(\eta_i(x), \sigma^2), \quad i = 1, \ldots, n$
Linear predictor:	$\eta_i(x) = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i}$
Latent Gaussian vector:	$x = (\beta_0, \beta_1, \beta_2)$

- Gaussian likelihood $\pi(y \mid \eta_i(x), \theta) = \frac{1}{\sqrt{2\pi \exp(-\theta)}} \exp\left(\frac{1}{2} \frac{y - \eta_i(x)}{\exp(-\theta)}\right)$
- $x \sim \mathcal{N}(0, Q^{-1})$ with $Q = \begin{pmatrix} \tau_0 & 0 & 0 \\ 0 & \tau_1 & 0 \\ 0 & 0 & \tau_2 \end{pmatrix}$

  For fixed effects, we usually fix the precision hyperparameters τ_j to a very low value, e.g. 10^{-3} ($\Rightarrow$ non-informative priors)
- Hyperparameter $\theta = \log(1/\sigma^2) \in \mathbb{R}$ (log-precision of Gaussian likelihood), for instance with an exponential prior with rate λ on σ such that

$$\pi(\theta) = \frac{\lambda}{2} \exp\left(-\lambda \exp(-\theta/2) - \theta/2\right), \quad \lambda > 0.$$

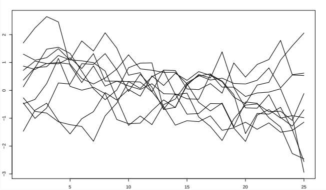
Examples

First-order Gaussian random walks

- A `rw1` is defined intrinsically over m classes through its “innovations”

$$x_{i+1} - x_i \sim \mathcal{N}(0, 1/\tau_{RW}), \quad i = 1, \ldots, m-1.$$

- Very useful prior models for capturing nonlinear covariate effects or time trends.
- It has one hyperparameter τ_{RW} measuring precision.
- To make it identifiable, we have to impose a constraint such as $\sum_{i=1}^{m} x_i = 0$.
- For easier interpretation, we may rescale to marginal distribution $\mathcal{N}(1/\tau_{RW})$.



10 realizations of a constrained `rw1` over $t = 1, \ldots, 25$, rescaled, $1/\tau_{RW} = 1$.

Estimation through `INLA`

Estimation through INLA

- We obtain the posterior estimations through Bayes' formula.

⚠ Usually, we cannot calculate the posterior estimations in closed form:

- posterior densities $\pi(\theta_j \mid y)$, $\pi(x_k \mid y)$, $\pi(\eta_i \mid y)$
- posterior mean estimates $\mathbb{E}(\theta_j \mid y)$, $\mathbb{E}(x_k \mid y)$, $\mathbb{E}(h(\eta_i) \mid y)$

⚠ Moreover, the latent Gaussian components x are often very high-dimensional.

$\Rightarrow$ use numerical approximation of complicated integrals:

- MCMC: iteratively simulate and update values $x^{(j)}, \theta^{(j)}$
  $\Rightarrow$ generate a large representative sample of posterior distribution.
  - Theoretical convergence guaranteed, but often too slow and unstable in practice.

  ⚠ mixing (i.e., exploration of the space of parameters) can be too slow.
- INLA: use astute numerical integration.

Laplace approximation

- Suppose that we want to calculate

$$\int_{-\infty}^{\infty} \exp(g(x))\,\mathrm{d}x,$$

where x has many components, such as the latent Gaussian components.

- If g has maximum at x_0 and its values decrease fast and smoothly around x_0, we can replace $g(x)$ by a curvature approximation around x_0:

$$g(x) \approx g(x_0) + 0.5(x - x_0)' H(g)(x_0)(x - x_0),$$

with Hessian matrix H containing second partial derivatives of g at x_0.

- Using this approximation, the function to integrate becomes an (unnormalized) Gaussian density:

$$\int_{-\infty}^{\infty} \exp(g(x))\,d\mathrm{x} \approx (2\pi)^{d/2} \left|H(g)(x_0)\right|^{-1/2} \exp(g(x_0)).$$

In practice, it remains to determine x_0 and $H(g)(x_0) \Rightarrow$ iterative Newton–Raphson algorithm.

A schematic overview of how INLA works

All hyperparameters are fixed: skip 1 and 2.2. Exactly 1 hyperparameter: skip 1.2.

1. Calculate posterior density of hyperparameter θ_j:

$$\pi(\theta_j \mid y) = \int\int \pi(\theta, x \mid y)\,\mathrm{d}x\mathrm{d}\theta_{-j}$$

   1.1 dx: use Laplace approximation for each configuration θ

   1.2 $\mathrm{d}\theta_{-j}$: use numerical integration

2. Calculate posterior density of latent components x_i or predictors η_i:

$$\pi(x_i \mid y) = \int\int \pi(x, \theta \mid y)\mathrm{d}x_{-i}\,\mathrm{d}\theta$$

   2.1 $\mathrm{d}x_{-i}$: use Laplace approximation (`strategy="laplace"`), but often too expensive
   - cheap but often less accurate: use conditional Gaussian from dx above (`strategy="gaussian"`)
   - `R-INLA` default is a simplified Laplace approximation (`strategy="simplified.laplace"`)

   2.2 $\mathrm{d}\theta$: use numerical integration

`R-INLA` proposes three variants for numerical integration in 1.2 and 2.2:

- grid of θ configurations (`int.strategy="grid"`, most costly)
- central composite design (`int.strategy="ccd"`, the default)
- only the mode of $\pi(\theta \mid y)$ (empirical Bayes, `int.strategy="eb"`)

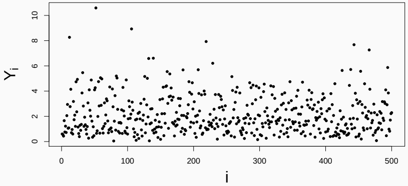
Extremes and `INLA`

GP model in INLA

- We will focus on the generalized Pareto (GP) distribution, which is, under mild conditions, the asymptotic distribution for threshold exceedances.

GP model in INLA

- We will focus on the generalized Pareto (GP) distribution, which is, under mild conditions, the asymptotic distribution for threshold exceedances.

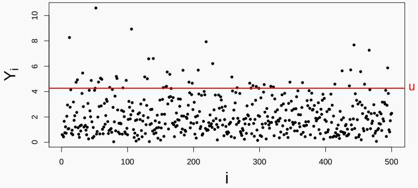


$$\Pr(Y - u \leq y \mid Y > u) \approx \mathrm{GP}, \qquad u \text{ large.}$$

$$\mathrm{GP}(y; \sigma, \xi) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \qquad y > u.$$

GP model in INLA

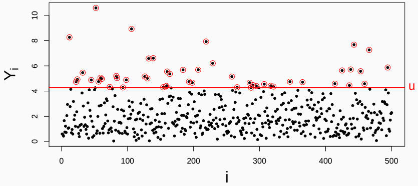
- We will focus on the generalized Pareto (GP) distribution, which is, under mild conditions, the asymptotic distribution for threshold exceedances.



$$\Pr(Y - u \leq y \mid Y > u) \approx \text{GP}, \qquad u \text{ large.}$$

GP model in INLA

- We will focus on the generalized Pareto (GP) distribution, which is, under mild conditions, the asymptotic distribution for threshold exceedances.

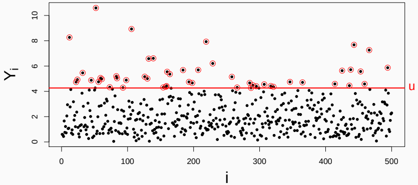


$$\Pr(Y - u \leq y \mid Y > u) \approx \text{GP}, \quad u \text{ large.}$$

$$\text{GP}(y; \sigma, \xi) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \quad y > u.$$

GP model in INLA

- We will focus on the generalized Pareto (GP) distribution, which is, under mild conditions, the asymptotic distribution for threshold exceedances.



$$\Pr(Y-u \le y \mid Y>u) \approx \mathrm{GP}, \qquad u \text{ large}.$$

$$\mathrm{GP}(y;\sigma,\xi) = 1-\left(1+\frac{\xi y}{\sigma}\right)^{-1/\xi}, \qquad y>u.$$

- In INLA, the linear predictor η controls the α-quantile of the GP distribution, $\Pr(y \le q_\alpha) = \alpha$, and $q_\alpha = \exp(\eta)$. The scale parameter σ is then a function of q_α and ξ:

$$\sigma = \frac{\xi q_\alpha}{(1-\alpha)^{-\xi}-1}.$$

- The shape ξ is treated as a hyperparameter and is represented as $\theta = \log\xi$, $\xi > 0$. The prior distribution is defined on θ.

Prior specification

- When little expert knowledge is available, a common practice is to assume non-informative priors.
- Alternatively, informative priors can be proposed using Penalized Complexity (PC) priors (Simpson et al., 2017).
- In this framework, model components (e.g., random effects) are considered to be flexible extensions of simpler base models. Priors are then developed in such a way that the components shrink towards their base models, thus preventing overfitting.

Prior specification

- PC prior for the GP shape parameter:
  - **Base model** f_{ξ_0}: $\xi = 0, \Rightarrow \text{GP}(y; \sigma, 0) =$ exponential distribution.
  - **Distance**: $d(f_\xi, f_{\xi_0}) = \sqrt{2\text{KLD}(f_\xi || f_{\xi_0})}$, where $\text{KLD}(f_\xi || f_{\xi_0}) = \frac{\xi^2}{1-\xi}$, $0 \leq \xi < 1$ is the Kullback-Leibler divergence.
  - **PC prior**: $d(f_\xi, f_{\xi_0}) = \sqrt{2}\xi/(1-\xi)^{1/2}$. Then, the PC prior for $\xi \in [0,1)$ is

$$\begin{aligned}
\pi(\xi) &= \lambda \exp\{-\lambda d(f_\xi, f_{\xi_0})\} \left| \frac{\mathrm{d}d(f_\xi, f_{\xi_0})}{\mathrm{d}\xi} \right| \\
&= \dots \\
&= \tilde{\lambda} \exp\left\{-\tilde{\lambda} \frac{\xi}{(1-\xi)^{1/2}}\right\} \frac{1-\xi/2}{(1-\xi)^{3/2}}, \quad \tilde{\lambda} = \sqrt{2}\lambda.
\end{aligned}$$

`R-INLA` in practice

Prediction with R-INLA

- The output of `inla(...)` contains the posterior estimates of fitted values for observations, but we often want to **predict where no observations are available**

- The simplest prediction strategy in R-INLA is to **add response data with NA values**

⇒ estimation and prediction can be done simultaneously in the Bayesian setting!

Model section by DIC

DIC is a generalization of AIC, that measures goodness of fit and penalizes model complexity. In the `R-INLA` library, we do `dic = TRUE` (*the lower, the better*).

Model selection and checking

Model section by DIC

DIC is a generalization of AIC, that measures goodness of fit and penalizes model complexity. In the `R-INLA` library, we do `dic = TRUE` (*the lower, the better*).

Model checking using PIT/CPO

$$\mathrm{CPO}_i = \Pr(y_i \mid \mathbf{y}_{-i}), \qquad \mathrm{PIT}_i = \Pr(y_i^\star \mid \mathbf{y}_{-i}).$$

- The CPO is the probability of an observed response based on the model fitted to the rest of the data.
- The PIT is the probability of a new response being less than the observed response using a model based on the rest of the data.

Practical examples

- We will now treat several examples using the Red Sea Surface Temperature data

$\longrightarrow$ open `INLAexamples.R` in `RStudio`.

Practical examples

- We will now treat several examples using the Red Sea Surface Temperature data $\longrightarrow$ open `INLAexamples.R` in `RStudio`.
- Our estimation procedure will consists of two steps. We construct and fit two separate models, the first for the probability of exceedances over the threshold (Bernoulli), the second for the positive exceedances over the threshold (GP).

Practical examples

- We will now treat several examples using the Red Sea Surface Temperature data $\longrightarrow$ open `INLAexamples.R` in `RStudio`.

- Our estimation procedure will consists of two steps. We construct and fit two separate models, the first for the probability of exceedances over the threshold (Bernoulli), the second for the positive exceedances over the threshold (GP).

- **Latent model 1:** AR1 time series model

$$\eta_i = \mu + \beta_T \times \texttt{time} + f(i), \quad i = 1, \ldots, n.$$

where

$$f(i) = \rho f(i-1) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, (1-\rho^2)/\tau_\varepsilon)$$

Two important hyperparameters: autocorrelation coefficient ρ, marginal precision τ_ε

Practical examples

- **Latent model 2:** Using y_{t-1} to predict y_t

$$\eta_i = \beta + \beta_T \times \texttt{time} + y_{i-1}, \quad i = 1, \ldots, n.$$

Practical examples

- **Latent model 2:** Using y_{t-1} to predict y_t

$$\eta_i = \beta + \beta_T \times \texttt{time} + y_{i-1}, \quad i = 1, \ldots, n.$$

- **Latent model 3:** Spatial modeling

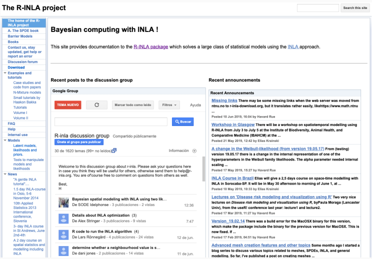
$$\eta_i = \beta_T \times \texttt{time} + f_1(w_{1i}) + f_2(w_{2i}), \quad i = 1, \ldots, n,$$

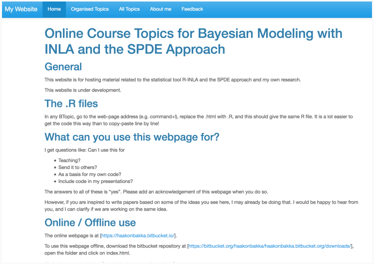
where $f_1(w_{1i})$ is a random walk of order 1, i.e., $f_1(w_{1i}) - f_1(w_{1i-1}) \sim \mathcal{N}(0, \tau^{-1})$, and $f_2(w_{2i})$ is a spatial effect with Matérn covariance structure (here with fixed range parameter).

Final remarks

Remarks - Help you can get from the web

- `www.r-inla.org`: homepage of the `INLA` team.
- `haakonbakka.bitbucket.io`: Bayesian Modeling with `INLA` and the SPDE approach.





Remarks - Help you can get from the web

- `http://people.bath.ac.uk/jjf23/brinla/`: `INLA` for Bayesian Regression.

INLA for Bayesian Regression Models

Julian Faraway

INLA stands for Integrated Nested Laplace Approximations. It is used for fitting Latent Gaussian models (LGM). LGMs include a wide range of commonly used regression models. Unlike MCMC which uses simulation methods, INLA uses approximation methods for Bayesian model fitting. Within the class of LGMs, INLA can fit models much faster than MCMC-based methods.

Visit the INLA website to learn much more including how to install the INLA R package.

I am the author of a forthcoming book entitled Bayesian Regression Models with Xiaofeng Wang and Ryan Yue

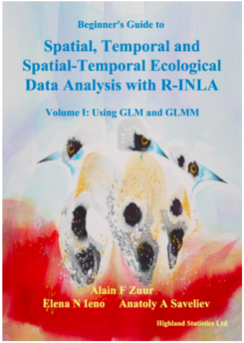
You will need to install the brinla R package to run many of the examples described here.

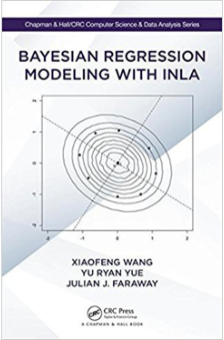
Worked Examples

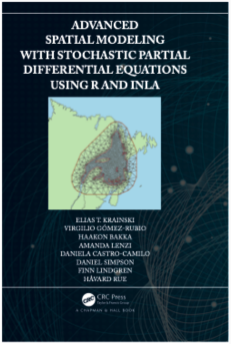
- Introduction - a simple example concerning Hubble's law.
- Linear regression - Chicago Insurance data
- One-way ANOVA - just one random effect
- Ridge regression - Ridge regression with meat spectroscopy data
- Gaussian Process regression - GP regression on fossil data
- Confidence bands for smoothness - GP regression to determine uncertainty regarding smoothness
- Non-stationary smoothing - GP regression with variable smoothing
- Generalized Extreme Values - fitting maximum annual river flows
- Define your own prior - using a half Cauchy prior for the SD of a random effect.
- Linear regression with bounded parameters - French presidential election example with slope parameters bounded in [0,1]

See also linear mixed models examples from my Extending Linear Models with R book.

Remarks - Books







Thank you!