

# Bayesian Statistics with R-INLA

University of Zurich, March, 2022

Instructor: Sara Martino

## Introduction

## Bayesian hierarchical models

## Latent Gaussian models

## Deterministic inference

Introduction Bayesian hierarchical models Latent Gaussian models Deterministic inference

ooooooooooooo oooooooooooooo

ooooooooooooo

oo

# Plan for this 2-day course

## Today

- **9:00-10:45** Introduction and basics concepts of INLA
- **11:00-12:30** Practical session I

## Lunch

- **13:30-15:00** R-INLA: Basics
- **15:15-17:00** Practical session II

# Plan for this 2-day course

## Tomorrow

- **9:00-10:45??**
- **11:00-12:30** Practical session III

## Lunch

- **13:30-15:00 ??**
- **15:15-16:00** Practical session IV



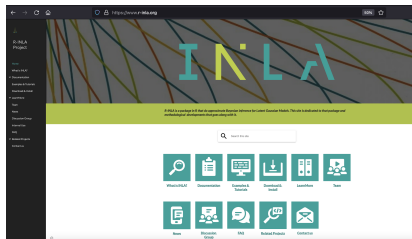
*INLA is a fast method to do Bayesian inference with latent Gaussian models and INLA is an R-package that implements this method with a flexible and simple interface.*





# Where?

The software, information, examples and help can be found at <http://www.r-inla.org>



- paper
- tutorials
- discussion group
- ...

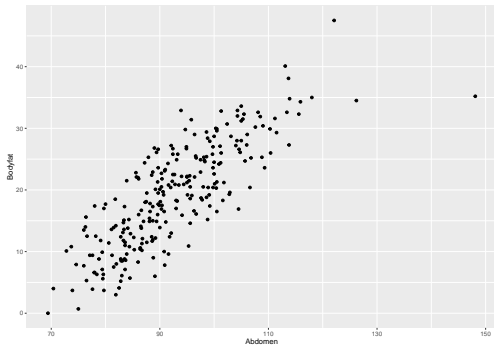
So... Why should you use R-INLA?

- What type of problems can we solve?
- What type of models can we use?
- When can we use it?

*To give proper answers to these questions, we need to start at the very beginning*

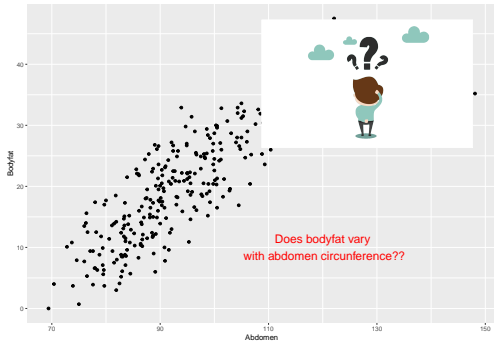
## The core

- We have observed something.



## The core

- We have observed something.
- We have questions.



## The core

- We have observed something.
- We have questions.
- We want answers!

- Bayesian or frequentist?

- Bayesian or frequentist?
- How do we model the data?

# How do we find answers?

We need to make choices:

- Bayesian or frequentist?
- How do we model the data?
- How do we compute the answer?



- Bayesian or frequentist?
- How do we model the data?
- How do we compute the answer?

## How do we find answers?

We need to make choices:

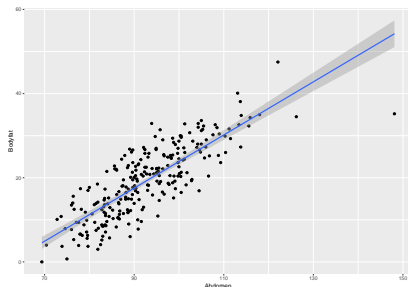
- Bayesian or frequentist?
- How do we model the data?
- How do we compute the answer?

These questions are **not** independent.

## A simple example

Assume a simple linear regression model with Gaussian observations  $y = (y_1, \dots, y_n)$ , where

$$E(y_i) = \alpha + \beta x_i, \text{Var}(y_i) = \tau^{-1}, \quad i = 1, \dots, n$$



Estimates:

	Estimate	Std.Error
(Intercept)	-39.280	2.66
Abdomen	0.631	0.029
Residual sd	4.877	

# A Bayesian hierarchical model

- Observation model

$$y \mid \underbrace{\mu, \beta}_x, \underbrace{\tau}_\theta$$

Encodes information about observed data

- Latent model  $x$ : The unobserved process
- Hyperprior for  $\theta$

# A Bayesian hierarchical model

- Observation model

$$y \mid \underbrace{\mu, \beta}_x, \underbrace{\tau}_\theta$$

Encodes information about observed data

- Latent model  $x$ : The unobserved process
- Hyperprior for  $\theta$

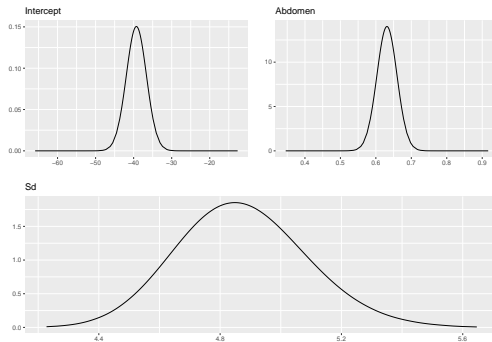
From this we can compute the **posterior distribution**

$$\pi(x, \theta | y) \propto \pi(y | x, \theta) \pi(x) \pi(\theta)$$

and then the corresponding **posterior marginal distributions**.

## Results

- Assign priors to  $\alpha, \beta, \tau$
- Use Bayes theorem to compute posterior distributions





Using a Bayesian framework:

- Build (hierarchical) models to account for potentially complicated dependency structures in the data.
- Attribute uncertainty to model parameters and latent variables using priors.

Two main challenges:

1. Need computationally efficient methods to calculate posteriors.
2. Select priors in a sensible way (see tomorrow)



## Bayesian hierarchical models

INLA can be used with Bayesian hierarchical models where we model in different stages or levels:

- **Stage 1:** What is the distribution of the responses?
- **Stage 2:** What is the distribution of the underlying unobserved (latent) components?
- **Stage 3:** What are our prior beliefs about the parameters controlling the components in the model?

## Stage 1: The data generating process

How is our **data** ( $\mathbf{y}$ ) generated from the **underlying components** ( $\mathbf{x}$ ) and **hyperparameters** ( $\boldsymbol{\theta}$ ) in the model:

- Gaussian response? (temperature, rainfall, fish weight ...)

It is also important how data are collected!

This information is placed into our **likelihood**  $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$

## Stage 1: The data generating process

How is our **data** ( $\mathbf{y}$ ) generated from the **underlying components** ( $\mathbf{x}$ ) and **hyperparameters** ( $\boldsymbol{\theta}$ ) in the model:

- Gaussian response? (temperature, rainfall, fish weight ...)
- Count data? (people infected with a disease in each area)

It is also important how data are collected!

This information is placed into our **likelihood**  $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$

## Stage 1: The data generating process

How is our **data** ( $\mathbf{y}$ ) generated from the **underlying components** ( $\mathbf{x}$ ) and **hyperparameters** ( $\boldsymbol{\theta}$ ) in the model:

- Gaussian response? (temperature, rainfall, fish weight ...)
- Count data? (people infected with a disease in each area)
- Point pattern? (locations of trees in a forest)

It is also important how data are collected!

This information is placed into our **likelihood**  $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$

## Stage 1: The data generating process

How is our **data** ( $\mathbf{y}$ ) generated from the **underlying components** ( $\mathbf{x}$ ) and **hyperparameters** ( $\boldsymbol{\theta}$ ) in the model:

- Gaussian response? (temperature, rainfall, fish weight ...)
- Count data? (people infected with a disease in each area)
- Point pattern? (locations of trees in a forest)
- Binary data? (yes/no response, binary image)

It is also important how data are collected!

This information is placed into our **likelihood**  $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$

## Stage 1: The data generating process

How is our **data** ( $\mathbf{y}$ ) generated from the **underlying components** ( $\mathbf{x}$ ) and **hyperparameters** ( $\boldsymbol{\theta}$ ) in the model:

- Gaussian response? (temperature, rainfall, fish weight ...)
- Count data? (people infected with a disease in each area)
- Point pattern? (locations of trees in a forest)
- Binary data? (yes/no response, binary image)
- Survival data? (recovery time, time to death)

It is also important how data are collected!

This information is placed into our **likelihood**  $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$

## Stage 1: The data generating process

We assume that *given* the **underlying components** ( $\mathbf{x}$ ) and **hyperparameters** ( $\boldsymbol{\theta}$ ) the data are independent on each other

$$\pi(y|x, \theta) = \prod_{i \in \mathcal{I}} \pi(y_i | x_{\mathcal{I}_y}, \theta)$$

## Stage 1: The data generating process

We assume that *given* the **underlying components** ( $\mathbf{x}$ ) and **hyperparameters** ( $\boldsymbol{\theta}$ ) the data are independent on each other

$$\pi(y|x, \theta) = \prod_{i \in \mathcal{I}} \pi(y_i | x_{\mathcal{I}_i}, \theta)$$

This implies that all the dependence structure in the data is explained in Stage II !!



## Stage 2: The dependence structure

The underlying **unobserved components**  $\mathbf{x}$  are called **latent components** and can be:

- Fixed effects for covariates
- Unstructured random effects (individual effects, group effects)
- Structured random effects (AR(1), regional effects, ...)

These are linked to the responses in the likelihood through linear predictors.

## Stage 3: The hyperparameters

The likelihood and the latent model typically have hyperparameters that control their behavior.

The **hyperparameters**  $\theta$  can include:

## Stage 3: The hyperparameters

The likelihood and the latent model typically have hyperparameters that control their behavior.

The **hyperparameters**  $\theta$  can include:

**Examples likelihood:**

- Variance of observation noise
- Dispersion parameter in the negative binomial model
- Probability of a zero (zero-inflated models)

## Stage 3: The hyperparameters

The likelihood and the latent model typically have hyperparameters that control their behavior.

The **hyperparameters**  $\theta$  can include:

**Examples likelihood:**

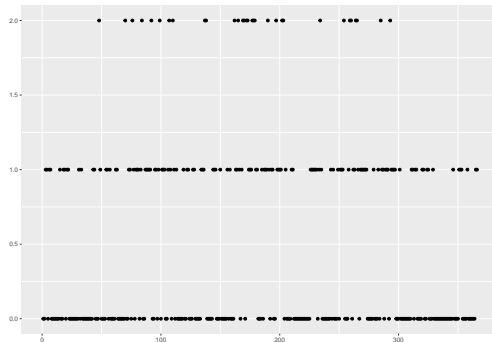
- Variance of observation noise
- Dispersion parameter in the negative binomial model
- Probability of a zero (zero-inflated models)

**Examples latent model:**

- Variance of unstructured effects
- Correlation of multivariate effects
- Range and variance of spatial effects
- Autocorrelation parameter

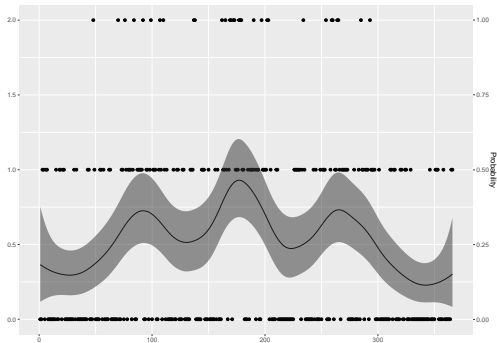
## Example: Tokyo rainfall data

Rainfall over 1 mm in the Tokyo area for each calendar day during two years (1983-84) are registered.



## Tokyo rainfall data

Rainfall over 1 mm in the Tokyo area for each calendar day during two years (1983-84) are registered.



## Stage 1: The data

$$y_i \mid p_i \sim \text{Binomial}(n_i, p_i),$$

for  $i = 1, 2, \dots, 366$

$$n_i = \begin{cases} 1, & \text{for 29 February} \\ 2, & \text{other days} \end{cases}$$

$$y_i = \begin{cases} \{0, 1\}, & \text{for 29 February} \\ \{0, 1, 2\}, & \text{other days} \end{cases}$$

Linear predictor

$$\text{logit}(p_i) = x_i \quad \Leftrightarrow \quad p_i = \frac{1}{1 + \exp(-x_i)}$$

- probability of rain on day  $i$  depends on  $x_i$
- the likelihood has no hyperparameters  $\theta$





## Stage 2: The latent model

It seems natural borrow strength over time and assume a cyclic smooth random effect, e.g. a **cyclic random walk of first or second order**. A random walk of first order (CRW1) is defined as:

$$\begin{aligned}\pi(x|\theta) &\propto \exp \left\{ -\frac{\theta}{2} \left[ (x_1 - x_{366})^2 + \sum_{i=2}^{366} (x_i - x_{i-1})^2 \right] \right\} \\ &= \exp \left\{ -\frac{\theta}{2} x^T R x \right\}\end{aligned}$$

$$R = \begin{bmatrix} 2 & -1 & & & & & -1 \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & & \ddots & & & \\ & & & & -1 & 2 & -1 \\ & & & & -1 & 2 & -1 \\ -1 & & & & & -1 & 2 \end{bmatrix}$$

## Stage 3: Hyperparameters

The structured time effect is controlled by one **precision (inverse variance) parameter  $\theta$** .

- A larger value of  $\theta$  means less variation in  $x$ , i.e. a smoother effect.
- $\theta$  is related to the variation in  $p_i$ .
- $\theta > 0$ : people commonly assume

$$\theta \sim \text{Ga}(\text{shape} = a, \text{rate} = b)$$

- However,  $\theta$  depends on  $R$ , so it is hard to define values for  $a$  and  $b$ . You could do this by defining reasonable lower and upper quantiles. (We talk about this tomorrow)

# Latent Gaussian models

# Latent Gaussian models

This was just one example of a very useful class of models called **Latent Gaussian models**.

- The characteristic property is that the **latent part** of the hierarchical model is **Gaussian**,  $\mathbf{x}|\boldsymbol{\theta} \sim N(0, Q^{-1})$
- The expected value is  $\mathbf{0}$
- The *precision* matrix (inverse covariance matrix) is  $Q$

## The general set-up

The set up contains GLMs, GLMMs, GAMs, GAMMs, and more. The mean of the observation  $i$ ,  $\mu_i$ , is connected to the linear predictor,  $\eta_i$ , through a link function  $g$ ,

$$\eta_i = g(\mu_i) = \mu + \mathbf{z}_i^\top \boldsymbol{\beta} + \sum_{\gamma} w_{\gamma,i} f_{\gamma}(c_{\gamma,i}) + v_i, \quad i = 1, 2, \dots, n$$

where

$\mu$  : Intercept

$\boldsymbol{\beta}$  : Fixed effects of covariates  $\mathbf{z}$

$\{f_{\gamma}(\cdot)\}$  : Non-linear/smooth effects of covariates  $\mathbf{c}$

$\{w_{\gamma,i}\}$  : Known weights defined for each observed data point

$\mathbf{v}$  : Unstructured error terms

## Loads of examples

- Generalized linear and additive (mixed) models
- Disease mapping
- Survival analysis
- Log-Gaussian Cox-processes
- Geostatistics
- Spatio and spatio-temporal models
- Stochastic volatility models
- Measurement error models
- And more!

# Specification of the latent field

- Collect all parameters (random variables) in the **latent field**  
 $x = \{\mu, \beta, \{f_\gamma(\cdot)\}, \eta\}.$

## Specification of the latent field

- Collect all parameters (random variables) in the **latent field**  $x = \{\mu, \beta, \{f_\gamma(\cdot)\}, \eta\}$ .
- A latent Gaussian model is obtained by assigning Gaussian priors to all elements of  $x$ .



## Specification of the latent field

- Collect all parameters (random variables) in the **latent field**  $x = \{\mu, \beta, \{f_\gamma(\cdot)\}, \eta\}$ .
- A latent Gaussian model is obtained by assigning Gaussian priors to all elements of  $x$ .
- Very flexible due to many different forms of the unknown functions  $\{f_\gamma(\cdot)\}$ :

## Specification of the latent field

- Collect all parameters (random variables) in the **latent field**  $x = \{\mu, \beta, \{f_\gamma(\cdot)\}, \eta\}$ .
- A latent Gaussian model is obtained by assigning Gaussian priors to all elements of  $x$ .
- Very flexible due to many different forms of the unknown functions  $\{f_\gamma(\cdot)\}$ :
- **Hyperparameters** account for variability and length/strength of dependence

- $f(\text{time})$ :, For example, an AR(1) process, RW1 or RW2
- $f(\text{spatial location})$ :, For example, a Mat'ern field
- $f(\text{covariate})$ :, For example, a RW1 or RW2 on the covariate values
- $f(\text{time, spatial location})$  can be a spatio-temporal effect
- And much more

# Additivity

- One of the most useful features of the framework is the additivity.
- Effects can easily be removed and added without difficulty.
- Each component might add a new latent part and might add new hyperparameters, but the modelling framework and computations stay the same.

# Additivity

- One of the most useful features of the framework is the additivity.
- Effects can easily be removed and added without difficulty.
- Each component might add a new latent part and might add new hyperparameters, but the modelling framework and computations stay the same.

**OBS:** The *linear* predictor needs to stay linear!! So effects can be added but not multiplied (will say more tomorrow..)

## A small point to think about

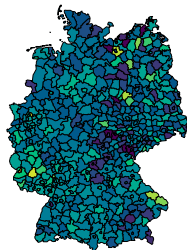
From a Bayesian point of view fixed effects and random effects are all the same.

- Fixed effects are also random
- They only differ in the prior we put on them

## Example: disease mapping

We observed larynx cancer mortality counts for males in 544 district of Germany from 1986 to 1990 and want to make a model.

- $y_i$ : The count at location  $i$ .
- $E_i$ : An offset; expected number of cases in district  $i$ .
- $c_i$ : A covariate (level of smoking consumption) at  $i$
- $s_i$ : spatial location  $i$ .



## Bayesian disease mapping

- **Stage 1:** We choose a Poisson distribution for the responses, so that

$$y_i \mid \eta_i \sim \text{Poisson}(E_i \exp(\eta_i))$$



## Bayesian disease mapping

- **Stage 1:** We choose a Poisson distribution for the responses, so that

$$y_i \mid \eta_i \sim \text{Poisson}(E_i \exp(\eta_i))$$

- **Stage 2:**  $\eta_i$  is a linear function of the latent components: a covariate  $c_i$ , a spatially structured effect  $f_u$ , an unstructured effect  $v$  likelihood by

$$\eta_i = \mu + \beta c_i + f_u(s_i) + v_i$$

## Bayesian disease mapping

- **Stage 1:** We choose a Poisson distribution for the responses, so that

$$y_i \mid \eta_i \sim \text{Poisson}(E_i \exp(\eta_i))$$

- **Stage 2:**  $\eta_i$  is a linear function of the latent components: a covariate  $c_i$ , a spatially structured effect  $f_u$ , an unstructured effect  $v$  likelihood by

$$\eta_i = \mu + \beta c_i + f_u(s_i) + v_i$$

- **Stage 3:**

## Bayesian disease mapping

- **Stage 1:** We choose a Poisson distribution for the responses, so that

$$y_i \mid \eta_i \sim \text{Poisson}(E_i \exp(\eta_i))$$

- **Stage 2:**  $\eta_i$  is a linear function of the latent components: a covariate  $c_i$ , a spatially structured effect  $f_u$ , an unstructured effect  $v$  likelihood by

$$\eta_i = \mu + \beta c_i + f_u(s_i) + v_i$$

- **Stage 3:**
  - $\tau_f$ : Precision parameter for the structured effect

## Bayesian disease mapping

- **Stage 1:** We choose a Poisson distribution for the responses, so that

$$y_i \mid \eta_i \sim \text{Poisson}(E_i \exp(\eta_i))$$

- **Stage 2:**  $\eta_i$  is a linear function of the latent components: a covariate  $c_i$ , a spatially structured effect  $f_u$ , an unstructured effect  $v$  likelihood by

$$\eta_i = \mu + \beta c_i + f_u(s_i) + v_i$$

- **Stage 3:**
  - $\tau_f$ : Precision parameter for the structured effect
  - $\tau_v$ : Precision parameter for the unstructured effect



## Bayesian disease mapping

- **Stage 1:** We choose a Poisson distribution for the responses, so that

$$y_i \mid \eta_i \sim \text{Poisson}(E_i \exp(\eta_i))$$

- **Stage 2:**  $\eta_i$  is a linear function of the latent components: a covariate  $c_i$ , a spatially structured effect  $f_u$ , an unstructured effect  $v$  likelihood by

$$\eta_i = \mu + \beta c_i + f_u(s_i) + v_i$$

- **Stage 3:**
  - $\tau_f$ : Precision parameter for the structured effect
  - $\tau_v$ : Precision parameter for the unstructured effect

The latent field is  $\mathbf{x} = (\mu, \beta, \{f_u(\cdot)\}, v_1, v_2, \dots, v_n)$ , the hyperparameters are  $\boldsymbol{\theta} = (\tau_f, \tau_v)$ , and must be given a prior.

So... which model fit the INLA framework??

1. Latent **Gaussian** model
2. The latent field has a sparse precision matrix (Markov properties)
3. The data are conditionally independent given the latent field
4. The predictor is linear

## Quiz!

Assume that, given  $\eta = (\eta_1, \dots, \eta_n)$  the observations  $y = (y_1, \dots, y_n)$  are independent and Poisson distributed with parameter  $\lambda_i = \exp(\eta_i)$  i.e.

$$y_i | \eta_i = \text{Poisson}(\lambda_i); i = 1, \dots, n$$

1.  $\eta_i = \alpha + \beta x_i + U_i$  where

$$\alpha, \beta \sim \mathcal{N}(0, 1)$$

$$U_i \sim \mathcal{N}(0, 1) \text{ for } i = 1, \dots, n$$

2.  $\eta_i = \alpha + \beta x_i + V_i$  where

$$\alpha, \beta \sim \mathcal{N}(0, 1)$$

$$U_i \sim \text{Bernoulli}(0.4) \text{ for } i = 1, \dots, n$$

3.  $\eta_i = \alpha + \beta x_i$  where

$$\alpha, \beta \sim \mathcal{N}(0, 1)$$

4.  $\eta_i = \alpha + \beta x_i + U_i V_i$  where

$$\alpha, \beta \sim \mathcal{N}(0, 1)$$

$$U_i \sim \mathcal{N}(0, 1) \text{ for } i = 1, \dots, n$$



# Deterministic inference

## Computations

So...

Now we have a modelling framework...

But how do we get our answers?

## What do we care about?

It depends on the problem!

- A single element of the latent field (e.g. the sign or quantiles of a fixed effect)

## What do we care about?

It depends on the problem!

- A single element of the latent field (e.g. the sign or quantiles of a fixed effect)
- A linear combination of elements from the latent field (the average over an area of a spatial effect, the difference of two effects)

## What do we care about?

It depends on the problem!

- A single element of the latent field (e.g. the sign or quantiles of a fixed effect)
- A linear combination of elements from the latent field (the average over an area of a spatial effect, the difference of two effects)
- A single hyperparameter (the correlation)

## What do we care about?

It depends on the problem!

- A single element of the latent field (e.g. the sign or quantiles of a fixed effect)
- A linear combination of elements from the latent field (the average over an area of a spatial effect, the difference of two effects)
- A single hyperparameter (the correlation)
- A non-linear combination of hyper parameters (animal models)

## What do we care about?

It depends on the problem!

- A single element of the latent field (e.g. the sign or quantiles of a fixed effect)
- A linear combination of elements from the latent field (the average over an area of a spatial effect, the difference of two effects)
- A single hyperparameter (the correlation)
- A non-linear combination of hyper parameters (animal models)
- Predictions at unobserved locations

## What do we care about?

The most important quantity in Bayesian statistics is **the posterior distribution**:

$$\overbrace{\pi(x, \theta \mid y)}^{\text{Posterior}} \propto \overbrace{\pi(\theta)\pi(x \mid \theta)}^{\text{Prior}} \overbrace{\prod_{i \in \mathcal{I}} \pi(y_i \mid x_i, \theta)}^{\text{Likelihood}}$$

from which we can derive the quantities of interest, such as

$$\begin{aligned} \pi(x_i \mid y) &\propto \int \int \pi(x, \theta \mid y) dx_{-i} d\theta \\ &= \int \pi(x_i \mid \theta, y) \pi(\theta \mid y) d\theta \end{aligned}$$

or  $\pi(\theta_j \mid y)$ .

These are very high dimensional integrals and are typically not analytically tractable.



## Traditional approach: MCMC

MCMC is based on sampling with the goal to **construct a Markov chain with the target posterior as stationary distribution.**

- Extensively used within Bayesian inference since the 1980's.
- Flexible and general, sometimes the only thing we can do!
- A generic tool is available with **JAGS/OpenBUGS**.
- Tools for specific models are of course available, e.g. **BayesX** and **stan**.
- Standard MCMC sampler are generally easy-ish to program and are in fact implemented in readily available software
- However, depending on the complexity of the problem, their efficiency might be limited.

## Approximate inference

Bayesian inference can (almost) never be done exactly. Some form of approximation must always be done.

- MCMC “works” for everything, but it can be incredibly slow
- Is it possible to make a quicker, more specialized inference scheme which only needs to work for this limited class of models? (specifically LGM)

## Recall: What is our model framework?

### Latent Gaussian models

$$y|x, \theta \sim \prod \pi(y_i|x_i, \theta)$$

$$x|\theta \sim \mathcal{N}(0, Q(\theta)) \quad \text{Gaussian!!!}$$

$$\theta \sim \pi(\theta) \quad \text{Not Gaussian}$$

where the precision matrix  $Q(\theta)$  is sparse. Generally these “sparse” Gaussian distributions are called **Gaussian Markov random fields** (GMRFs).

The sparseness can be exploited for very quick computations for the Gaussian part of the model through numerical algorithms for sparse matrices.

## The INLA idea

Use the properties of the LGM we have defined to approximate the posterior **marginals**

$$\pi(x_i \mid \mathbf{y}) \quad \text{and} \quad \pi(\theta_j \mid \mathbf{y})$$

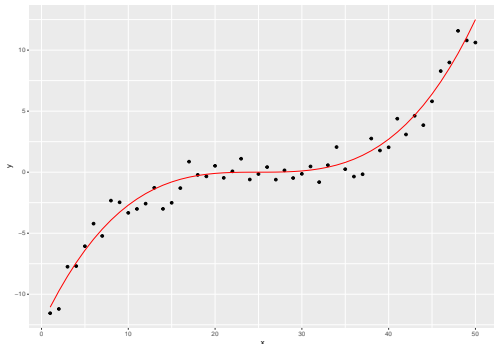
directly.

Let us consider a **toy example to illustrate the ideas.**

## How does INLA work? A toy example

Smoothing noisy observations - Data

We observe some smooth function but our measures are noisy  
(but we know the size of such noise!)



**Goal:** Recover the smooth function observed with noise!



## Smoothing noisy observations - LGM

- **Data** Gaussian Observations with known precision

$$y_i|x_i \sim \mathcal{N}(x_i, 1)$$

- **Latent Model:** A Gaussian model for the smooth function (RW2 model)

$$\pi(\mathbf{x}|\theta) \propto \theta^{(n-2)/n} \exp \left\{ -\frac{\theta}{2} \sum_{i=2}^n (x_i - 2x_{i-1} + x_{i-2})^2 \right\}$$

- **Hyperparameter** The precision of the smooth function  $\theta$ . We assign a Gamma prior

$$\pi(\theta) \propto \theta^{a-1} \exp(-b\theta)$$

# Smoothing noisy observations - Goal

Find approximations for:

1. The posterior marginal for the hyperparameter  $\pi(\theta|\mathbf{y})$
2. The posterior marginals for the elements of the latent field  $\pi(x_i|\mathbf{y})$



## Approximating $\pi(\theta|\mathbf{y})$

We have that

$$\pi(\mathbf{x}, \theta, \mathbf{y}) = \pi(\mathbf{x}|\theta, \mathbf{y})\pi(\theta|\mathbf{y})\pi(\mathbf{y})$$

so

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\pi(\mathbf{x}|\theta, \mathbf{y})\pi(\mathbf{y})} \propto \frac{\pi(\mathbf{y}, \mathbf{x}|\theta) \pi(\theta)}{\pi(\mathbf{x}|\theta, \mathbf{y})}$$

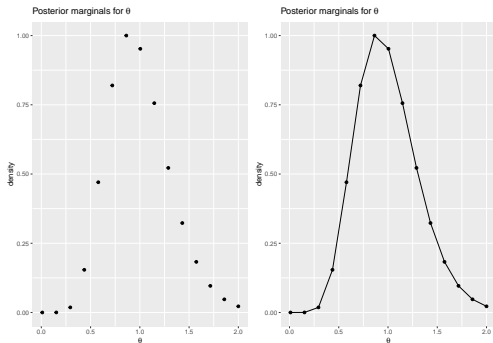
\$ Since the likelihood is Gaussian, then  $\pi(\mathbf{y}, \mathbf{x}|\theta)$  is also Gaussian. We have then:

$$\pi(\theta|\mathbf{y}) \propto \frac{\overbrace{\pi(\mathbf{y}, \mathbf{x}|\theta)}^{\text{Gaussian}} \pi(\theta)}{\underbrace{\pi(\mathbf{x}|\theta, \mathbf{y})}_{\text{Gaussian}}}$$

This is valid for any  $\mathbf{x}$

# Posterior marginal for the hyperparameter

Select a grid of points to represent the density  $\pi(\theta|\mathbf{x})$



## Approximating $\pi(x_i|y, \theta)$

Again we have that

$$\mathbf{x}, \mathbf{y} | \theta \sim \mathbf{N}(\cdot, \cdot)$$

so also  $\pi(x_i | \theta, \mathbf{y})$  is Gaussian!!\

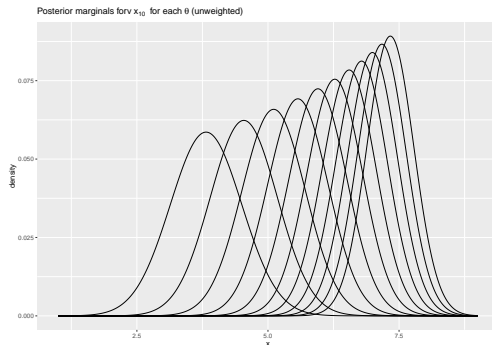
We compute

$$\begin{aligned} \pi(x_i | \mathbf{y}) &= \int \pi(x_i | \theta, \mathbf{y}) \pi(\theta | \mathbf{y}) d\theta \\ &\approx \sum_k \pi(x_i | \theta_k, \mathbf{y}) \pi(\theta_k | \mathbf{y}) \Delta_k \end{aligned}$$

where  $\theta_k, k = 1, \dots, K$  are the representative points of  $\pi(\theta | \mathbf{y})$  and  $\Delta_k$  are the corresponding weights

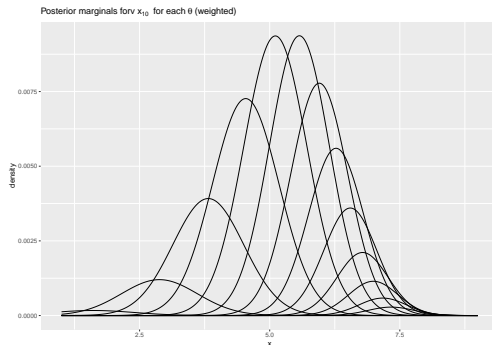
# Posterior marginals for latent field I

Compute the conditional posterior marginal for  $x_i$  given each  $\theta_k$



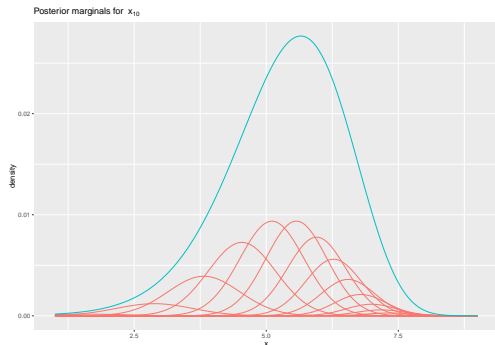
## Posterior marginals for latent field II

Weight the conditional posterior marginal for  $\pi(x_i|\theta_k, \mathbf{y})$  by  $\pi(\theta_k|\mathbf{y})\Delta_{\mathbf{k}}$



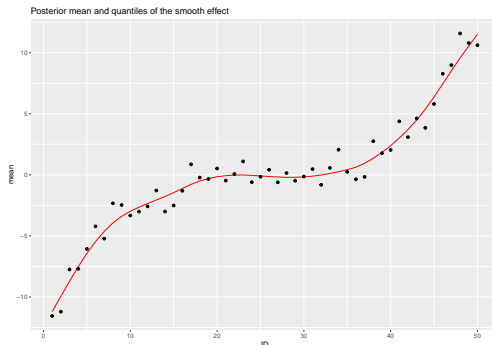
# Posterior marginals for latent field III

Sum to get the posterior marginal for  $x_i|\mathbf{y}$



# Fitted Spline

The posterior marginals are used to calculate summary statistics, like means, variances and credible intervals:



## R-INLA code

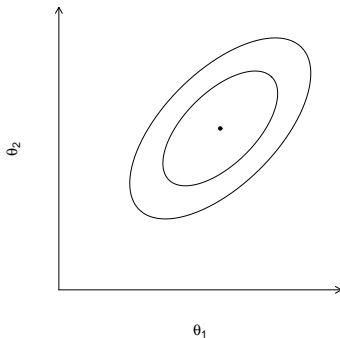
```
formula = y ~ -1 + f(idx, model="rw2", constr=FALSE,  
  hyper=list(prec=list(prior="loggamma", param=c(a,b))))  
  
result = inla(formula,  
  data = data.frame(y=y, idx=idx),  
  control.family = list(initial = log(tau_0), fixed=TRUE))
```





# 1. More than one hyperparameter

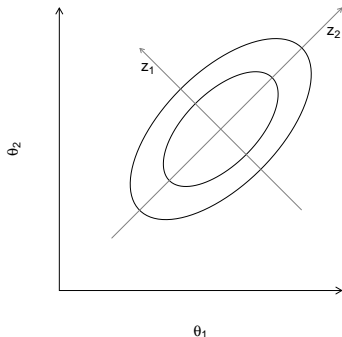
**Main use:** Select good evaluation points  $\theta_k$  for the numerical integration when approximating  $\tilde{\pi}(x_i|y)$



# 1. More than one hyperparameter

**Main use:** Select good evaluation points  $\theta_k$  for the numerical integration when approximating  $\tilde{\pi}(x_i|y)$

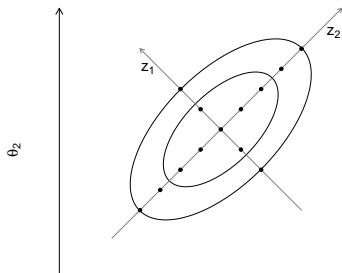
- Locate the mode
- Compute the Hessian to construct principal components



# 1. More than one hyperparameter

**Main use:** Select good evaluation points  $\theta_k$  for the numerical integration when approximating  $\tilde{\pi}(x_i|y)$

- Locate the mode
- Compute the Hessian to construct principal components
- Grid-search to locate bulk of the probability mass



# 1. More than one hyperparameter

- Locate the mode
- Compute the Hessian to construct principal components
- Grid-search to locate bulk of the probability mass
- For each point  $k$  in the grid compute:
  - $\tilde{\pi}(\theta^k|y)$
  - $\tilde{\pi}(x_i|\theta^k, y)$
  - $\Delta_k$

## 2. Non-Gaussian observations

In application we may choose likelihoods other than a Gaussian.  
How does this change things?

$$\pi(\theta \mid \mathbf{y}) \propto \frac{\overbrace{\pi(\mathbf{x}, \mathbf{y} \mid \theta)}^{\text{Non-Gaussian, BUT KNOWN}} \pi(\theta)}{\underbrace{\pi(\mathbf{x} \mid \mathbf{y}, \theta)}_{\text{Non-Gaussian and UNKNOWN}}}$$

- In many cases  $\pi(\mathbf{x} \mid \mathbf{y}, \theta)$  is very close to a Gaussian distribution, and can be replaced with a **Laplace approximation**.



## The GMRF approximation (II)

Thus,

$$\begin{aligned}\tilde{\pi}(\mathbf{x}|\theta, \mathbf{y}) &\propto \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \sum_{i=1}^n (\mathbf{a}_i + \mathbf{b}_i\mathbf{x}_i - 0.5\mathbf{c}_i\mathbf{x}_i^2)\right) \\ &\propto \exp\left(-\frac{1}{2}\mathbf{x}^\top (\mathbf{Q} + \text{diag}(\mathbf{c}))\mathbf{x} + \mathbf{b}^\top \mathbf{x}\right)\end{aligned}$$

which is Gaussian with precision matrix  $\mathbf{Q} + \text{diag}(\mathbf{c})$  and mean given by the solution of  $(\mathbf{Q} + \text{diag}(\mathbf{c}))\boldsymbol{\mu} = \mathbf{b}$

The canonical parameterisation is

$$\mathcal{N}_C(\mathbf{b}, \mathbf{Q} + \text{diag}(\mathbf{c}))$$

which corresponds to

$$\mathcal{N}((\mathbf{Q} + \text{diag}(\mathbf{c}))^{-1}\mathbf{b}, (\mathbf{Q} + \text{diag}(\mathbf{c}))^{-1}).$$



# The GMFR approximation - One dimensional example

Assume

$y|\lambda \sim \text{Poisson}(\lambda)$  Likelihood

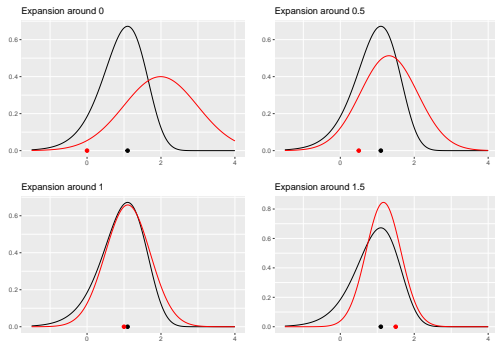
$\lambda = \exp(x)$  Likelihood

$x \sim \mathcal{N}(0, 1)$  Latent Model

we have that

$$\pi(x|y) \propto \pi(y|x)\pi(x) \propto \exp\left\{-\frac{1}{2}x^2 + \underbrace{xy - \exp(x)}_{\text{non-gaussian part}}\right\}$$

# The GMRF approximation



If  $\mathbf{y} \mid \mathbf{x}, \theta$  is Gaussian "the approximation" is exact! }

## What do we get ...

$$\tilde{\pi}(\theta \mid \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \mathbf{y} \mid \theta) \pi(\theta)}{\tilde{\pi}_{\mathbf{G}}(\mathbf{x} \mid \mathbf{y}, \theta)} \bigg|_{\mathbf{x}=\mathbf{x}^*(\theta)}$$

\* find the mode of  $\tilde{\pi}(\theta \mid \mathbf{y})$  (optimization) \* explore  $\tilde{\pi}(\theta \mid \mathbf{y})$  to find grid points  $t_k$  for numerical integration.

What do we get ...

$$\tilde{\pi}(\theta | \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \mathbf{y} | \theta) \pi(\theta)}{\tilde{\pi}_{\mathbf{G}}(\mathbf{x} | \mathbf{y}, \theta)} \Bigg|_{\mathbf{x}=\mathbf{x}^*(\theta)}$$

\* find the mode of  $\tilde{\pi}(\theta | \mathbf{y})$  (optimization) \* explore  $\tilde{\pi}(\theta | \mathbf{y})$  to find grid points  $t_k$  for numerical integration.

However, why is it called **integrated nested Laplace approximation**?

## What do we get ...

$$\tilde{\pi}(\theta | \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \mathbf{y} | \theta) \pi(\theta)}{\tilde{\pi}_{\mathbf{G}}(\mathbf{x} | \mathbf{y}, \theta)} \bigg|_{\mathbf{x}=\mathbf{x}^*(\theta)}$$

\* find the mode of  $\tilde{\pi}(\theta | \mathbf{y})$  (optimization) \* explore  $\tilde{\pi}(\theta | \mathbf{y})$  to find grid points  $t_k$  for numerical integration.

However, why is it called **integrated nested Laplace approximation**?

There is another step that changes:

$$\pi(x_i | \mathbf{y}) \approx \sum_{\mathbf{k}} \underbrace{\pi(x_i | \mathbf{y}, \theta^{\mathbf{k}})}_{\text{Not Gaussian!}} \tilde{\pi}_{\mathbf{G}}(\theta^{\mathbf{k}} | \mathbf{y}) \Delta_{\mathbf{k}}$$

# Approximating $\pi(x_i|\mathbf{y}, \theta)$

Three possible approximations:

However, errors in location and/or lack of skewness possible

# Approximating $\pi(x_i|\mathbf{y}, \theta)$

Three possible approximations:

- 1. **Gaussian distribution** derived from  $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$ , i.e.

$$\tilde{\pi}(x_i|\theta, \mathbf{y}) = \mathcal{N}(\mathbf{x}_i; \mu_i(\theta), \sigma_i^2(\theta))$$

with mean  $\mu_i(\theta)$  and marginal variance  $\sigma_i^2(\theta)$ .

However, errors in location and/or lack of skewness possible

# Approximating $\pi(x_i|\mathbf{y}, \theta)$

Three possible approximations:

- 1. **Gaussian distribution** derived from  $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$ , i.e.

$$\tilde{\pi}(x_i|\theta, \mathbf{y}) = \mathcal{N}(\mathbf{x}_i; \mu_i(\theta), \sigma_i^2(\theta))$$

with mean  $\mu_i(\theta)$  and marginal variance  $\sigma_i^2(\theta)$ .

However, errors in location and/or lack of skewness possible



# Approximating $\pi(x_i|\mathbf{y}, \theta)$

Three possible approximations:

- 1. **Gaussian distribution** derived from  $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$ , i.e.

$$\tilde{\pi}(x_i|\theta, \mathbf{y}) = \mathcal{N}(\mathbf{x}_i; \mu_i(\theta), \sigma_i^2(\theta))$$

with mean  $\mu_i(\theta)$  and marginal variance  $\sigma_i^2(\theta)$ .

However, errors in location and/or lack of skewness possible

- 2. **Laplace approximation**

# Approximating $\pi(x_i|\mathbf{y}, \theta)$

Three possible approximations:

- 1. **Gaussian distribution** derived from  $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$ , i.e.

$$\tilde{\pi}(x_i|\theta, \mathbf{y}) = \mathcal{N}(\mathbf{x}_i; \mu_i(\theta), \sigma_i^2(\theta))$$

with mean  $\mu_i(\theta)$  and marginal variance  $\sigma_i^2(\theta)$ .

However, errors in location and/or lack of skewness possible

- 2. **Laplace approximation**

# Approximating $\pi(x_i|\mathbf{y}, \theta)$

Three possible approximations:

- 1. **Gaussian distribution** derived from  $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$ , i.e.

$$\tilde{\pi}(x_i|\theta, \mathbf{y}) = \mathcal{N}(\mathbf{x}_i; \mu_i(\theta), \sigma_i^2(\theta))$$

with mean  $\mu_i(\theta)$  and marginal variance  $\sigma_i^2(\theta)$ .

However, errors in location and/or lack of skewness possible

- 2. **Laplace approximation**
- 3. **Simplified Laplace approximation**

The approximation is very good but expensive as  $n$  factorizations of  $(n - 1) \times (n - 1)$  matrices are required to get the  $n$  marginals.

## Laplace approximation of $\pi(x_i|\theta, \mathbf{y})$

$$\tilde{\pi}_{\text{LA}}(x_i|\theta, \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\tilde{\pi}_{\text{GG}}(\mathbf{x}_{-i}|\mathbf{x}_i, \theta, \mathbf{y})} \Big|_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(\mathbf{x}_i, \theta)}$$

The approximation is very good but expensive as  $n$  factorizations of  $(n-1) \times (n-1)$  matrices are required to get the  $n$  marginals.

Computational modifications exist:

1. Approximate the modal configuration of the GMRF approximation.
2. Reduce the size  $n$  by only involving the neighbors

# Simplified Laplace approximation

Faster alternative to the Laplace approximation

- based on a **series expansion up to third order of the numerator and denominator of  $\tilde{\pi}_{\text{LA}}(x_i|\theta, \mathbf{y})$**
- corrects the Gaussian approximation for error in location and lack of skewness.

## Simplified Laplace approximation

Faster alternative to the Laplace approximation

- based on a **series expansion up to third order of the numerator and denominator of  $\tilde{\pi}_{\text{LA}}(x_i|\theta, \mathbf{y})$**
- corrects the Gaussian approximation for error in location and lack of skewness.

This is **default option when using INLA** but this choice can be modified.

## INLA: Overview

- **Step I** Approximate  $\pi(\theta|y)$  using the Laplace approximation and select good evaluation points  $\theta_k$ .
- **Step II** For each  $\theta_k$  and  $i$  approximate  $\pi(x_i|\theta_k, y)$  using the Laplace or simplified Laplace approximation for selected values of  $x_i$
- **Step III** For each  $i$ , sum out  $\theta_k$

$$\tilde{\pi}(x_i|y) = \sum_k \tilde{\pi}(x_i|\theta_k, y) \times \tilde{\pi}(\theta_k|y) \times \Delta_k.$$

Build a log spline corrected Gaussian to represent  $\tilde{\pi}(x_i|y)$ .



## INLA: Why does it work?

- The full conditional  $\pi(x|y, \theta)$  is “almost” Gaussian
- The latent field  $x$  is a GMRF
  - GMRF  $\rightarrow$  sparse precision matrix!!
  - Easy to solve and store
- Smart numerical methods
- Parallel implementation

## Limitations

- The dimension of the latent field  $x$  can be large ( $10^2 - 10^6$ )
- The dimension of the hyperparameters  $\theta$  must be small ( $\leq 9$ )

In other words, each random effect can be big, but there cannot be too many random effects unless they share parameters.

- Marginal posterior densities of all (hyper-)parameters
- Posterior mean, median, quantiles, std.-deviation, etc.
- The approach can be used for predictions, model assessment, ...
- Joint posterior marginal not available... but it is possible to sample from  $\tilde{\pi}(x, \theta|y)$

Thank you for your attention!

If you have any doubts or questions, please write :  
[sara.martino@math.ntnu.no](mailto:sara.martino@math.ntnu.no)

