Bayesian Statistics with R-INLA

Instructor: Elias T Krainski <elias@r-inla.org>

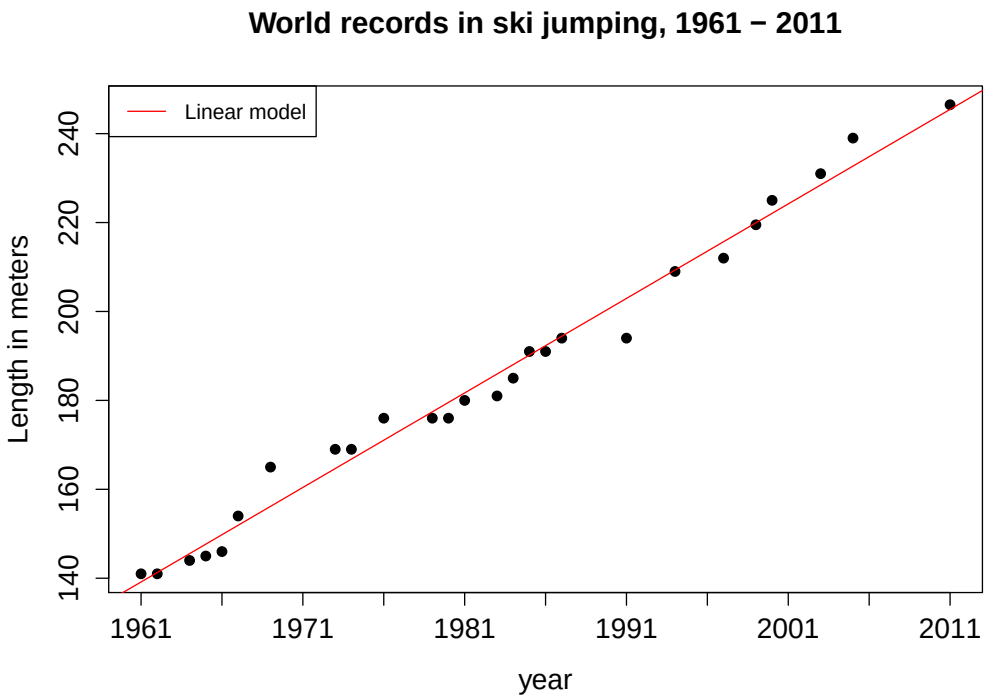
November, 2019

Outline

Example: Ski flying records

Assume a simple linear regression model with Gaussian observations $\boldsymbol{y} = (y_1, \ldots, y_n)$, where

$$\mathrm{E}(y_i) = \mu + \beta x_i, \quad \mathrm{Var}(y_i) = \tau^{-1}, \quad i = 1, \ldots, n$$



Estimates

Intercept: 137.03 (1.42195),
β: 2.13 (0.05)

Alternative: Bayesian hierarchical model

- Observation model $\boldsymbol{y} \mid \underbrace{\mu, \beta}_{\boldsymbol{x}}, \underbrace{\tau}_{\theta}$: Encodes information about observed data
- Latent model $\boldsymbol{x}$: The unobserved process
- Hyperprior for θ

Alternative: Bayesian hierarchical model

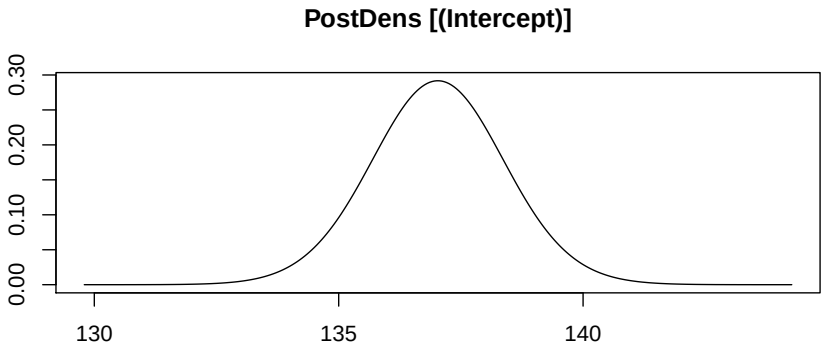
- Observation model $\boldsymbol{y} \mid \underbrace{\mu, \beta}_{\boldsymbol{x}}, \underbrace{\tau}_{\theta}$: Encodes information about observed data
- Latent model $\boldsymbol{x}$: The unobserved process
- Hyperprior for θ

From this we can compute the posterior distribution

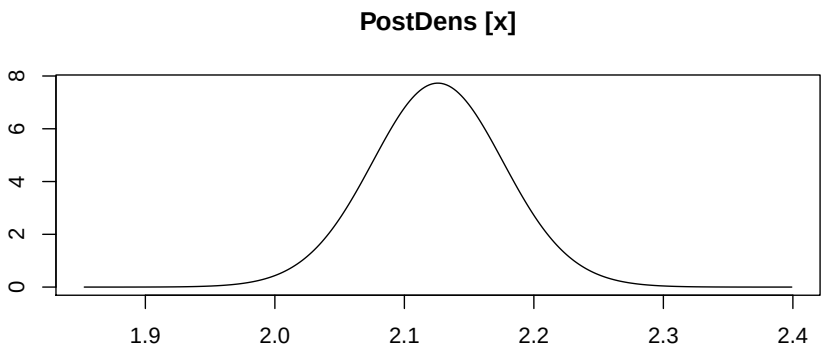
$$\mathrm{p}(\boldsymbol{x}, \theta \mid \boldsymbol{y}) \propto \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \theta)\, \mathrm{p}(\boldsymbol{x})\, \mathrm{p}(\theta)$$

and then the corresponding posterior marginal distributions.

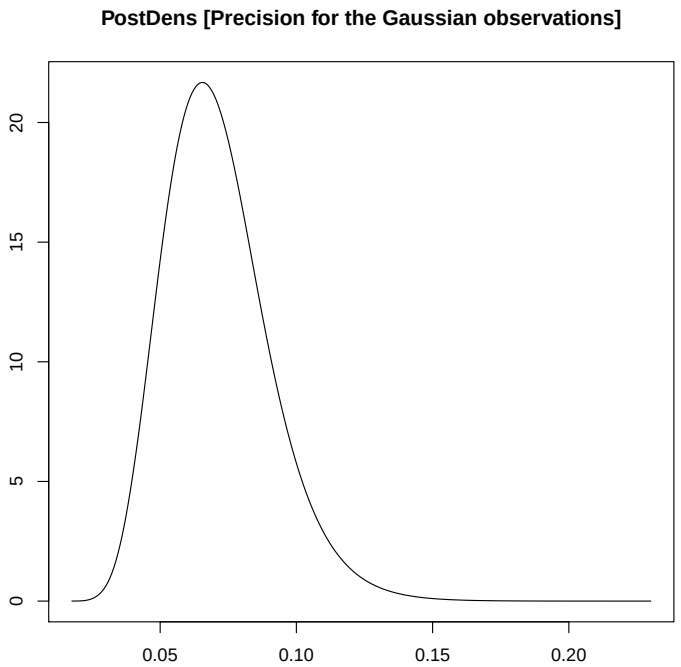
Results



Mean = 137.029 SD = 1.393



Mean = 2.126 SD = 0.053



Real-world datasets are usually much more complicated!

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

How is our data ($\boldsymbol{y}$) generated from the underlying components ($\boldsymbol{x}$) 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* $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$

Stage 1

How is our data ($\boldsymbol{y}$) generated from the underlying components ($\boldsymbol{x}$) 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* $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$

Stage 1

How is our data ($\boldsymbol{y}$) generated from the underlying components ($\boldsymbol{x}$) 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* $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$

Stage 1

How is our data ($\boldsymbol{y}$) generated from the underlying components ($\boldsymbol{x}$) 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* $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$

Stage 1

How is our data ($\boldsymbol{y}$) generated from the underlying components ($\boldsymbol{x}$) 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* $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$

Stage 2

The underlying unobserved components $\boldsymbol{x}$ 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 likelihood and the latent model typically have hyperparameters that control their behavior. The hyperparameters $\boldsymbol{\theta}$ can include:

Stage 3

The likelihood and the latent model typically have hyperparameters that control their behavior. The hyperparameters $\boldsymbol{\theta}$ 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 likelihood and the latent model typically have hyperparameters that control their behavior. The hyperparameters $\boldsymbol{\theta}$ can include:

Examples likelihood:

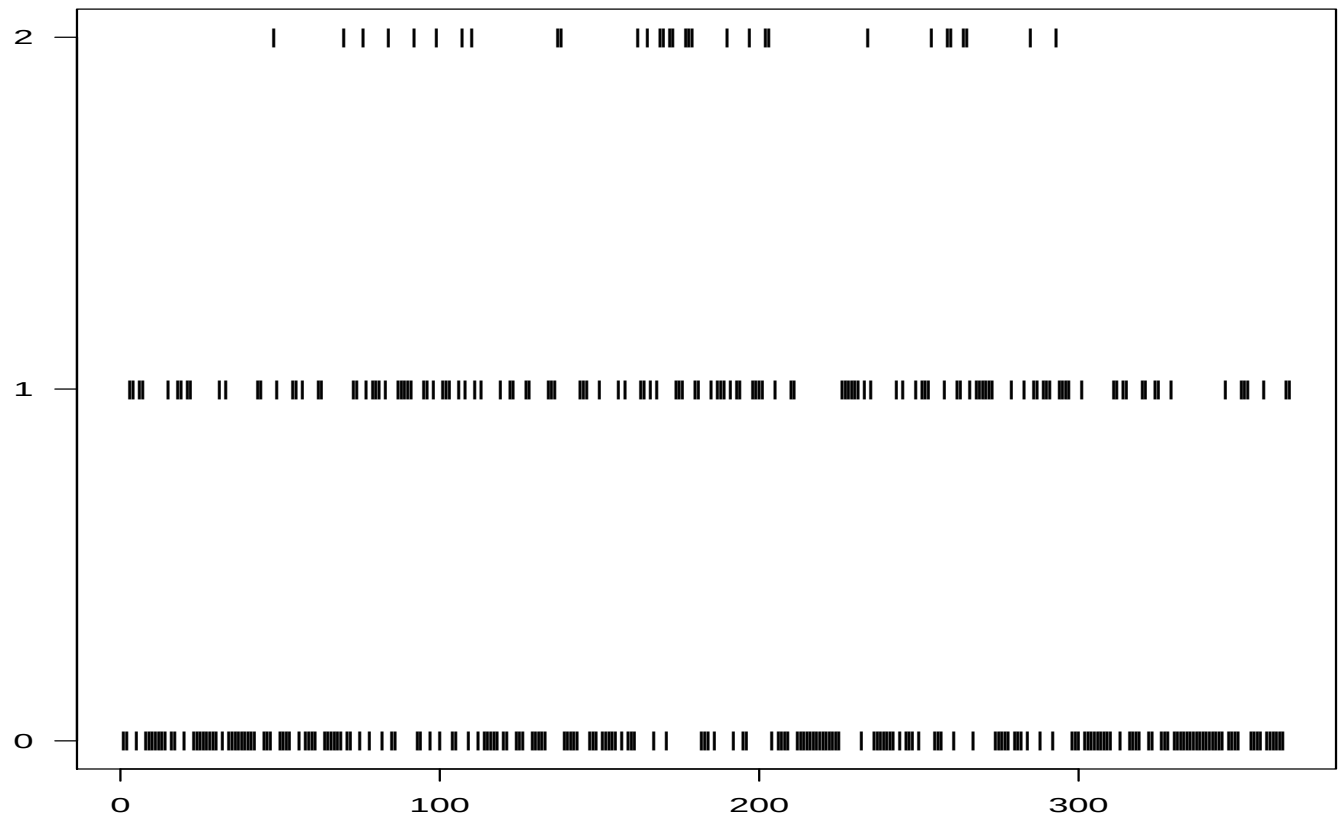
- Variance of observation noise
- Dispersion parameter in the negative binomial model
- Probability of a zero (zero-inflated models)

Examples latent model:

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

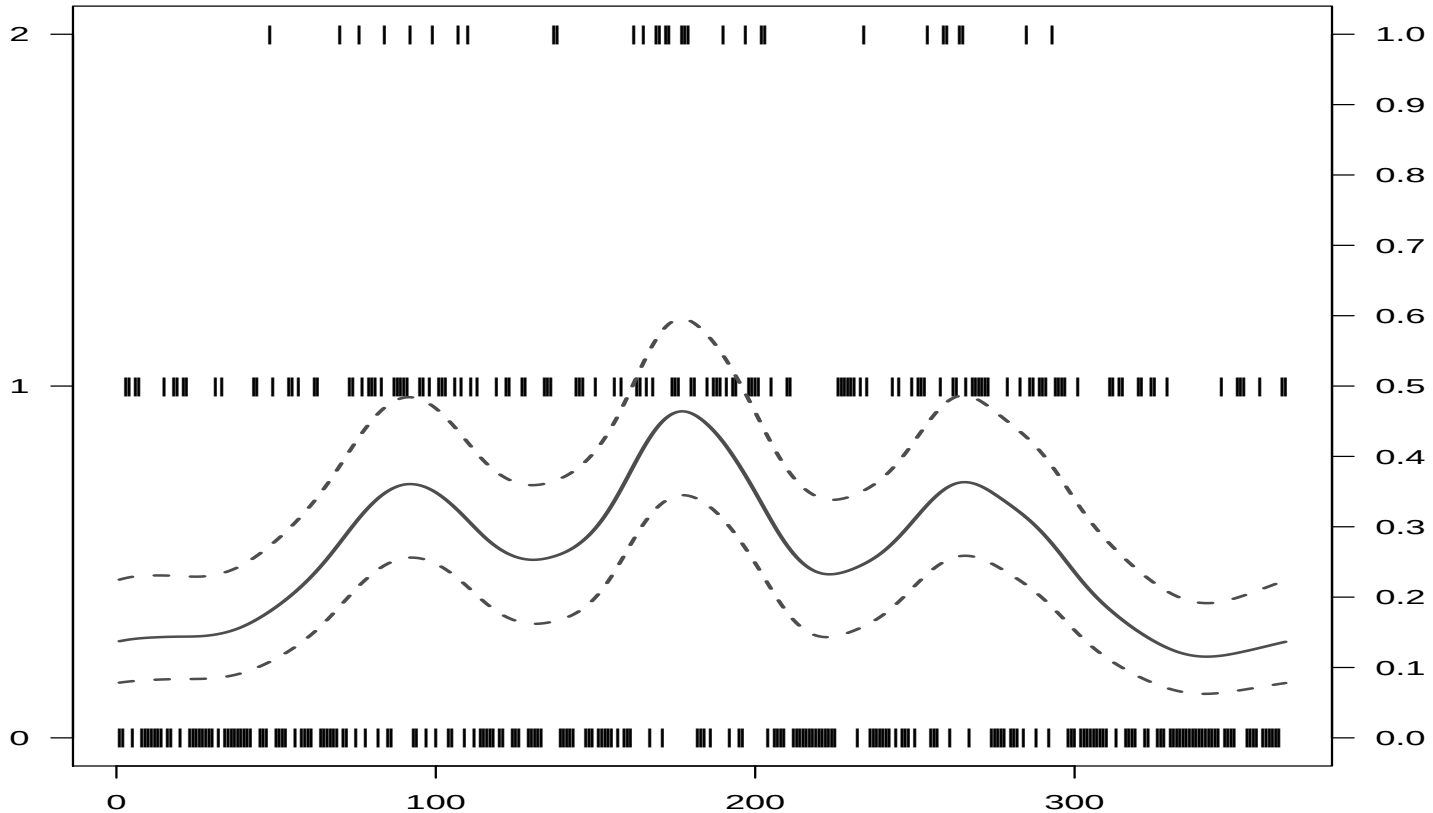
Example: Tokyo rainfall data

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 $i = 1, 2, \ldots, 366$

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

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

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

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:

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

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:

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

with $\boldsymbol{R} =$

$$\begin{pmatrix} 2 & -1 & & & & & & -1 \\ -1 & 2 & -1 & & & & & \\ & -1 & 2 & -1 & & & & \\ & & & \ddots & & & & \\ & & & & -1 & 2 & -1 & \\ & & & & & -1 & 2 & -1 \\ -1 & & & & & & -1 & 2 \end{pmatrix}$$

Stage 3: Hyperparameters

The structured time effect is controlled by one precision (inverse variance) parameter θ.

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

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

- However, θ depends on $\mathbf{R}$, 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

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, $\boldsymbol{x}|\boldsymbol{\theta} \sim N(0, \mathbf{Q}^{-1})$
- The expected value is $\mathbf{0}$
- The *precision* matrix (inverse covariance matrix) is $\mathbf{Q}$

The general set-up

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

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

where

μ : Intercept

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

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

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

$\boldsymbol{v}$: Unstructured error terms

Further Examples

- Dynamic linear models
- Stochastic volatility models (famously difficult with MCMC)
- Generalised linear (mixed) models
- Generalised additive (mixed) models
- Spline smoothing
- Semiparametric regression
- Space-varying (semiparametric) regression models
- Disease mapping
- Log-Gaussian Cox-processes
- Model-based geostatistics (*)
- Spatio-temporal models
- Survival analysis
- +++

Specification of the latent field

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

Specification of the latent field

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

Specification of the latent field

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

Flexibility through f-functions

The functions $\{f_\gamma\}$ in the linear predictor make it possible to capture very different types of random effects in the same framework:

- $f(\texttt{time})$: For example, an AR(1) process, RW1 or RW2
- $f(\texttt{spatial location})$: For example, a Matérn field
- $f(\texttt{covariate})$: For example, a RW1 or RW2 on the covariate values
- $f(\texttt{time}, \texttt{spatial location})$ 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.

A small point to think about

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

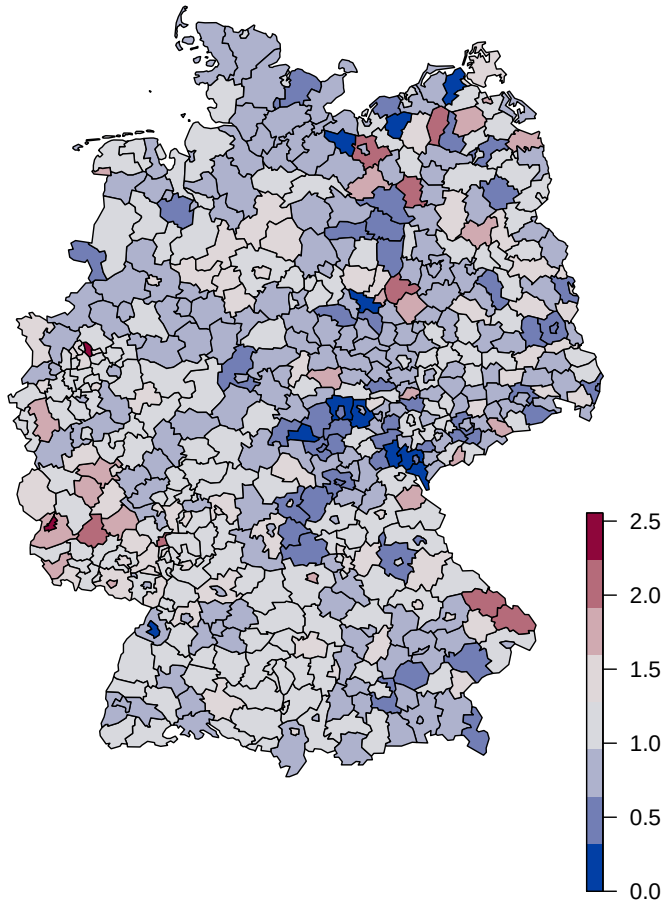
- Fixed effects are also random
- They only differ in the prior we put on them

Example: Disease mapping in Germany

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

Information available:

- 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 location i
- $\boldsymbol{s}_i$: spatial location i (here, district).



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: η_i is a linear function of the latent components: a covariate c_i, a spatially structured effect f_u, an unstructured effect $\boldsymbol{v}$ 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: η_i is a linear function of the latent components: a covariate c_i, a spatially structured effect f_u, an unstructured effect $\boldsymbol{v}$ 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: η_i is a linear function of the latent components: a covariate c_i, a spatially structured effect f_u, an unstructured effect $\boldsymbol{v}$ likelihood by

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

Stage 3: τ_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: η_i is a linear function of the latent components: a covariate c_i, a spatially structured effect f_u, an unstructured effect $\boldsymbol{v}$ likelihood by

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

Stage 3: τ_f: Precision parameter for the structured effect

τ_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: η_i is a linear function of the latent components: a covariate c_i, a spatially structured effect f_u, an unstructured effect $\boldsymbol{v}$ likelihood by

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

Stage 3: τ_f: Precision parameter for the structured effect

τ_v: Precision parameter for the unstructured effect

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

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:

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

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

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

or $\pi(\theta_j \mid \boldsymbol{y})$.

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

Tasks

Compute from

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\, \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \prod_{i \in \mathcal{I}} \pi(y_i \mid x_i)$$

the posterior marginals:

$$\pi(x_i \mid \boldsymbol{y}), \qquad \text{for some or all } i$$

and/or

$$\pi(\theta_i \mid \boldsymbol{y}), \qquad \text{for some or all } i$$

Our approach: Approximate Bayesian Inference

- Can we compute (approximate) marginals directly without using MCMC?

Our approach: Approximate Bayesian Inference

- Can we compute (approximate) marginals directly without using MCMC?
- YES!

Our approach: Approximate Bayesian Inference

- Can we compute (approximate) marginals directly without using MCMC?
- YES!
- Gain
  - **Huge speedup & accuracy**
  - **The ability to treat latent Gaussian models properly ;-)**

Main ideas (I)

Main ideas are simple and based on the identity

$$\pi(z) = \frac{\pi(x,z)}{\pi(x|z)} \qquad \text{leading to} \qquad \widetilde{\pi}(z) = \frac{\pi(x,z)}{\widetilde{\pi}(x|z)}$$

Main ideas (I)

Main ideas are simple and based on the identity

$$\pi(z) = \frac{\pi(x,z)}{\pi(x|z)} \qquad \text{leading to} \qquad \widetilde{\pi}(z) = \frac{\pi(x,z)}{\widetilde{\pi}(x|z)}$$

When $\widetilde{\pi}(x|z)$ is the Gaussian-approximation, this is the Laplace-approximation.

Main ideas (II)

Construct the approximations to

1. $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
2. $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

Main ideas (II)

Construct the approximations to

1. $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
2. $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

then we integrate

$$\pi(x_i|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})\, \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\, d\boldsymbol{\theta}$$

Main ideas (II)

Construct the approximations to

1. $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
2. $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

then we integrate

$$\pi(x_i|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})\, \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\, d\boldsymbol{\theta}$$

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

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.

* Markov chain Monte Carlo

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?

Recall: What is our model framework?

Latent Gaussian models

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta} \sim \prod_i \pi(y_i|\eta_i,\boldsymbol{\theta}) \qquad \text{Any of several distributions!}$$

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0},\mathbf{Q}(\boldsymbol{\theta})^{-1}) \qquad \text{Gaussian (GMRF)!}$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) \qquad \text{Any distribution!}$$

where the precision matrix $\mathbf{Q}(\boldsymbol{\theta})$ 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 posterior distribution

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$$

to approximate the posterior marginals

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

directly.

Let us consider a toy example to illustrate the ideas.

How does INLA work?

Observations

$$y_i = m(i) + \epsilon_i, \qquad i = 1, \ldots, n$$

Here, we assume that $m(i)$ is a smooth function wrt i and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_0)$ with *known* precision τ_0.

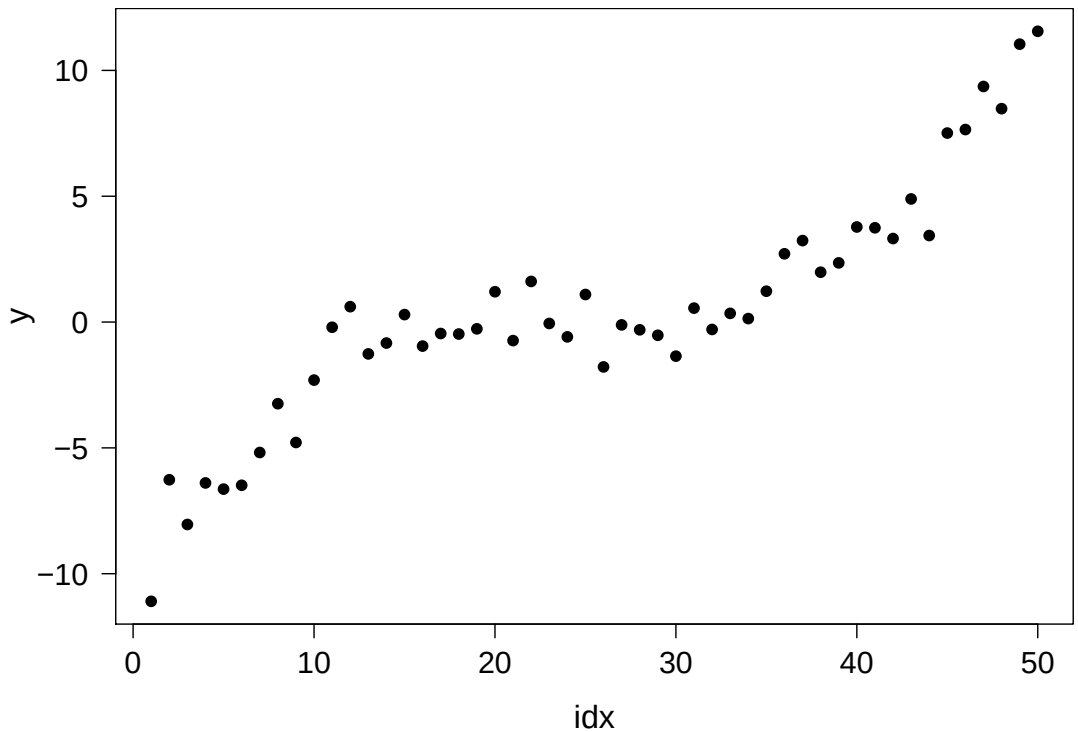
How does INLA work?

Observations

$$y_i = m(i) + \epsilon_i, \qquad i = 1, \ldots, n$$

Here, we assume that $m(i)$ is a smooth function wrt i and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_0)$ with *known* precision τ_0.

```
1 n = 50
2 idx = 1:n
3 # generate something
      smooth representing m
4 fun = 100*((idx-n/2)/n)^3
5 # add some noise
6 y = fun + rnorm(n, mean
      =0, sd=1)
7 plot(idx, y)
```



Assumed hierarchical model

1. Data: Gaussian observations with known precision

$$y_i \mid x_i, \theta \sim \mathcal{N}(x_i, \tau_0)$$

[1] `model="rw2"`

Assumed hierarchical model

1. Data: Gaussian observations with known precision

$$y_i \mid x_i, \theta \sim \mathcal{N}(x_i, \tau_0)$$

2. Latent model: A random walk of second order[1]

$$\pi(\boldsymbol{x} \mid \theta) \propto \theta^{(n-2)/2} \exp\left(-\frac{\theta}{2}\sum_{i=3}^{n}(x_i - 2x_{i-1} + x_{i-2})^2\right)$$

[1] `model="rw2"`

Assumed hierarchical model

1. Data: Gaussian observations with known precision

$$y_i \mid x_i, \theta \sim \mathcal{N}(x_i, \tau_0)$$

2. Latent model: A random walk of second order[1]

$$\pi(\boldsymbol{x} \mid \theta) \propto \theta^{(n-2)/2} \exp\left(-\frac{\theta}{2}\sum_{i=3}^{n}(x_i - 2x_{i-1} + x_{i-2})^2\right)$$

3. Hyperparameter: The smoothing parameter θ which we assign a $\Gamma(a, b)$ prior

$$\pi(\theta) \propto \theta^{a-1} \exp(-b\theta), \quad \theta > 0$$

[1] `model="rw2"`

Derivation of posterior marginals (I)

Since

$$\boldsymbol{x}, \boldsymbol{y} \mid \theta \sim \mathcal{N}(\cdot, \cdot)$$

(derived using $\pi(\boldsymbol{x}, \boldsymbol{y} \mid \theta) \propto \pi(\boldsymbol{y} \mid \boldsymbol{x}, \theta)\, \pi(\boldsymbol{x} \mid \theta)$),
we can compute (numerically) all marginals, using that

Derivation of posterior marginals (I)

Since

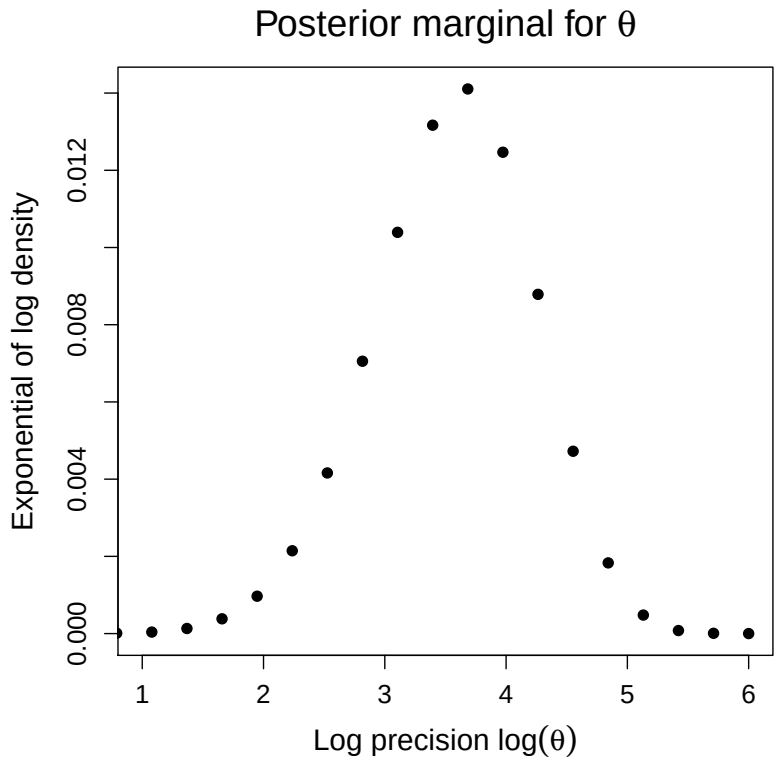
$$\boldsymbol{x}, \boldsymbol{y} \mid \theta \sim \mathcal{N}(\cdot, \cdot)$$

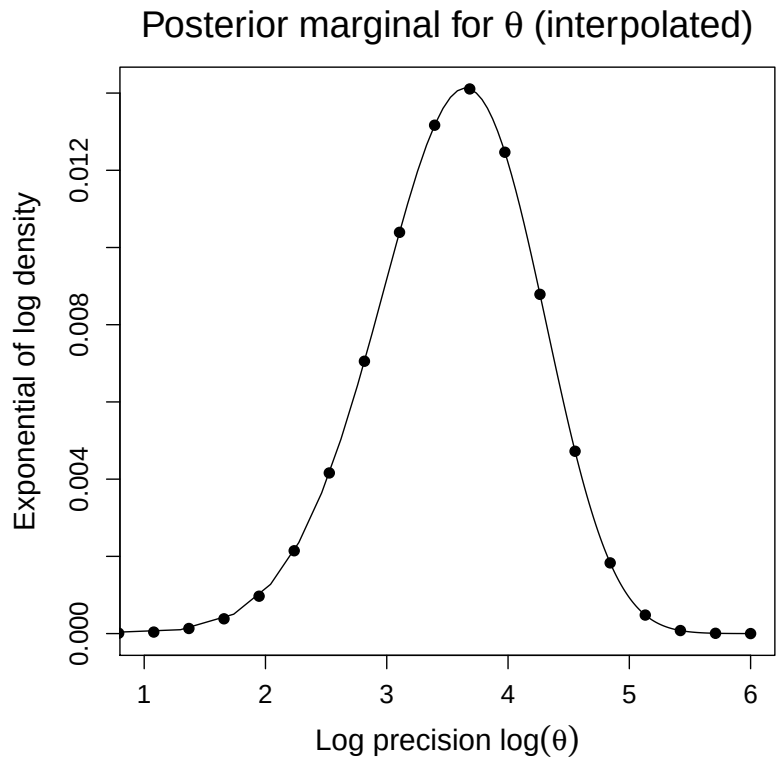
(derived using $\pi(\boldsymbol{x}, \boldsymbol{y} \mid \theta) \propto \pi(\boldsymbol{y} \mid \boldsymbol{x}, \theta)\, \pi(\boldsymbol{x} \mid \theta)$),
we can compute (numerically) all marginals, using that

$$\pi(\theta \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{x}, \theta \mid \boldsymbol{y})}{\pi(\boldsymbol{x} \mid \boldsymbol{y}, \theta)} \propto \frac{\overbrace{\pi(\boldsymbol{x}, \boldsymbol{y} \mid \theta)}^{\text{Gaussian}}\ \pi(\theta)}{\underbrace{\pi(\boldsymbol{x} \mid \boldsymbol{y}, \theta)}_{\text{Gaussian}}}$$

Posterior marginal for hyperparameter

Select a grid of points $t_1, \ldots, t_k$ to represent the density $\theta \mid \boldsymbol{y}$. (Here, the points are chosen to be equi-distant).





Derivation of posterior marginals (II)

From

$$\boldsymbol{x} \mid \boldsymbol{y}, \theta \sim \mathcal{N}(\cdot, \cdot)$$

Derivation of posterior marginals (II)

From

$$\boldsymbol{x} \mid \boldsymbol{y}, \theta \sim \mathcal{N}(\cdot, \cdot)$$

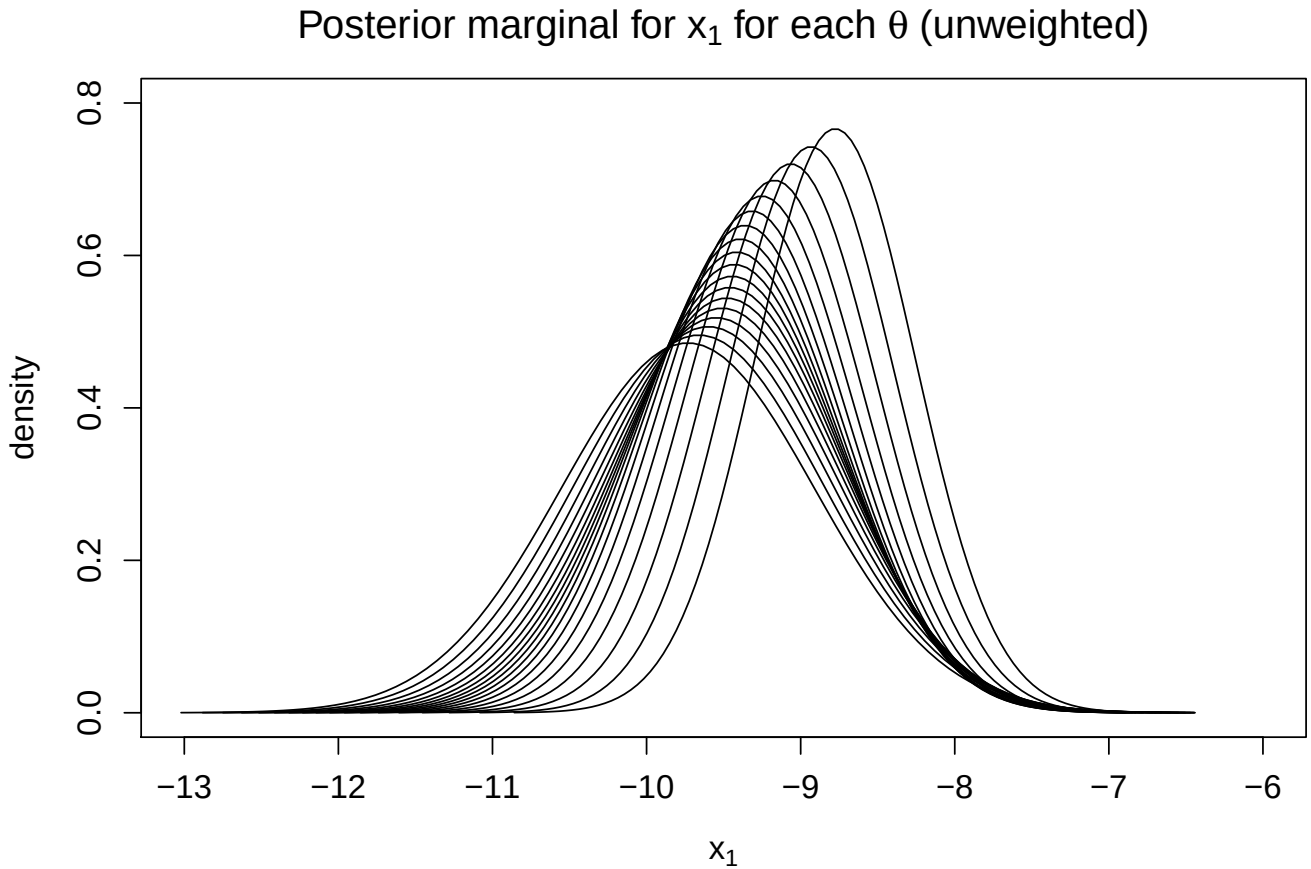
we can compute

$$\begin{aligned}\pi(x_i \mid \boldsymbol{y}) &= \int \underbrace{\pi(x_i \mid \boldsymbol{y}, \theta)}_{\text{Gaussian}} \, \pi(\theta \mid \boldsymbol{y}) \, d\theta \\ &\approx \sum_k \pi(x_i \mid \boldsymbol{y}, t_k) \pi(t_k \mid \boldsymbol{y}) \Delta_k\end{aligned}$$

where t_k, $k = 1, \ldots, K$, correspond to representative points of $\theta \mid \boldsymbol{y}$ and Δ_k are the corresponding weights (equal to 1 if points are equi-distant).

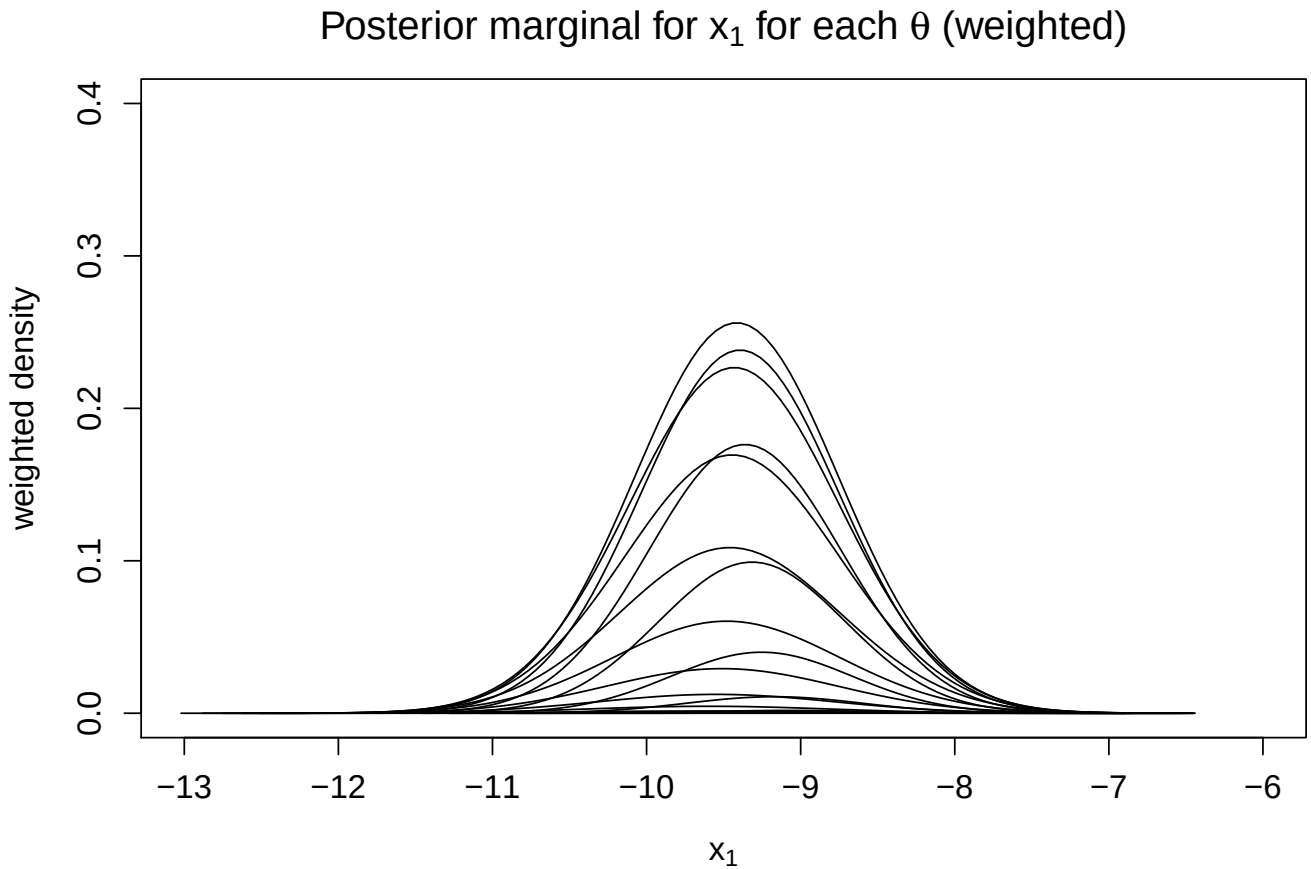
Posterior marginal for latent parameters

Compute the conditional marginal posterior for each x_i given t_k. Here, shown for x_1.



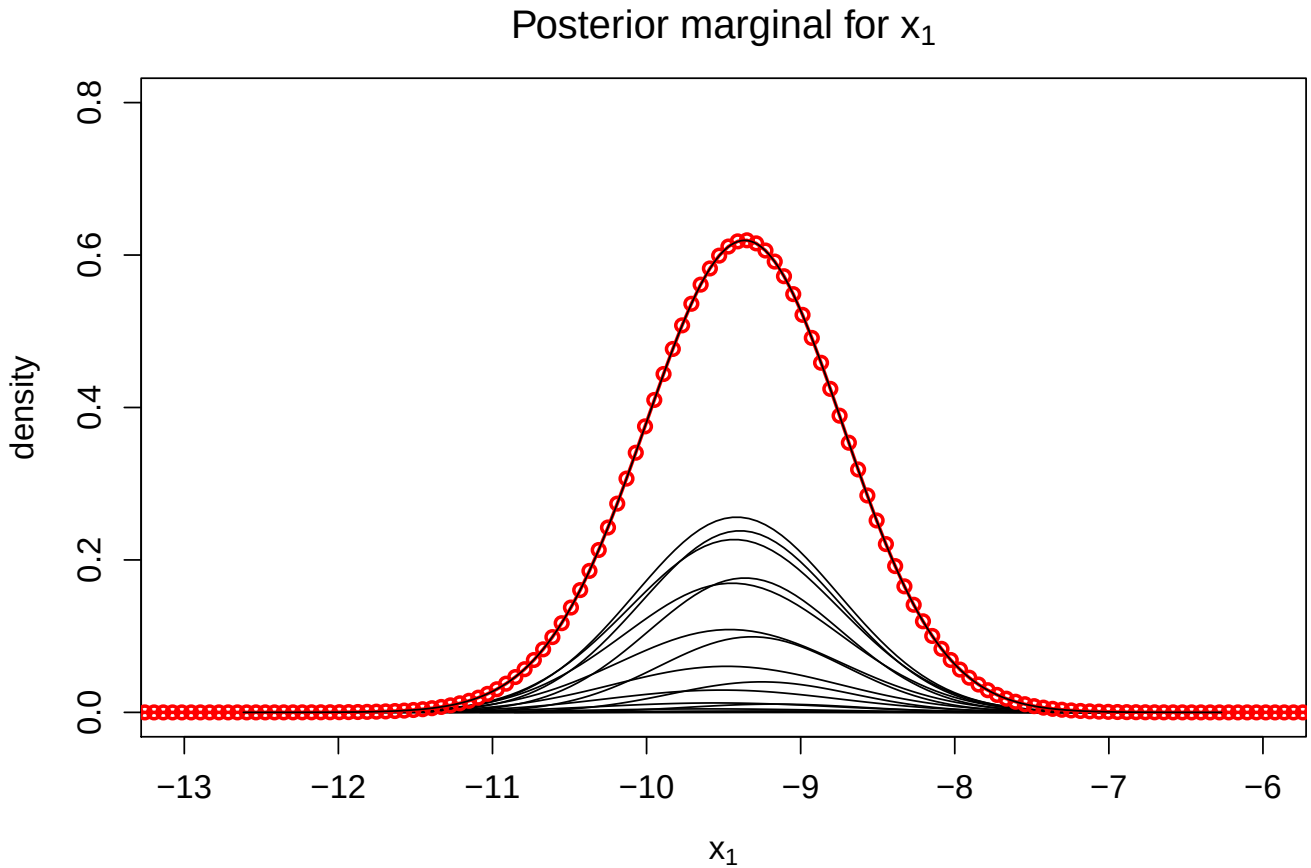
Posterior marginal for latent parameters

Weigh the resulting (conditional) marginal posterior by the density associated with each θ_k.



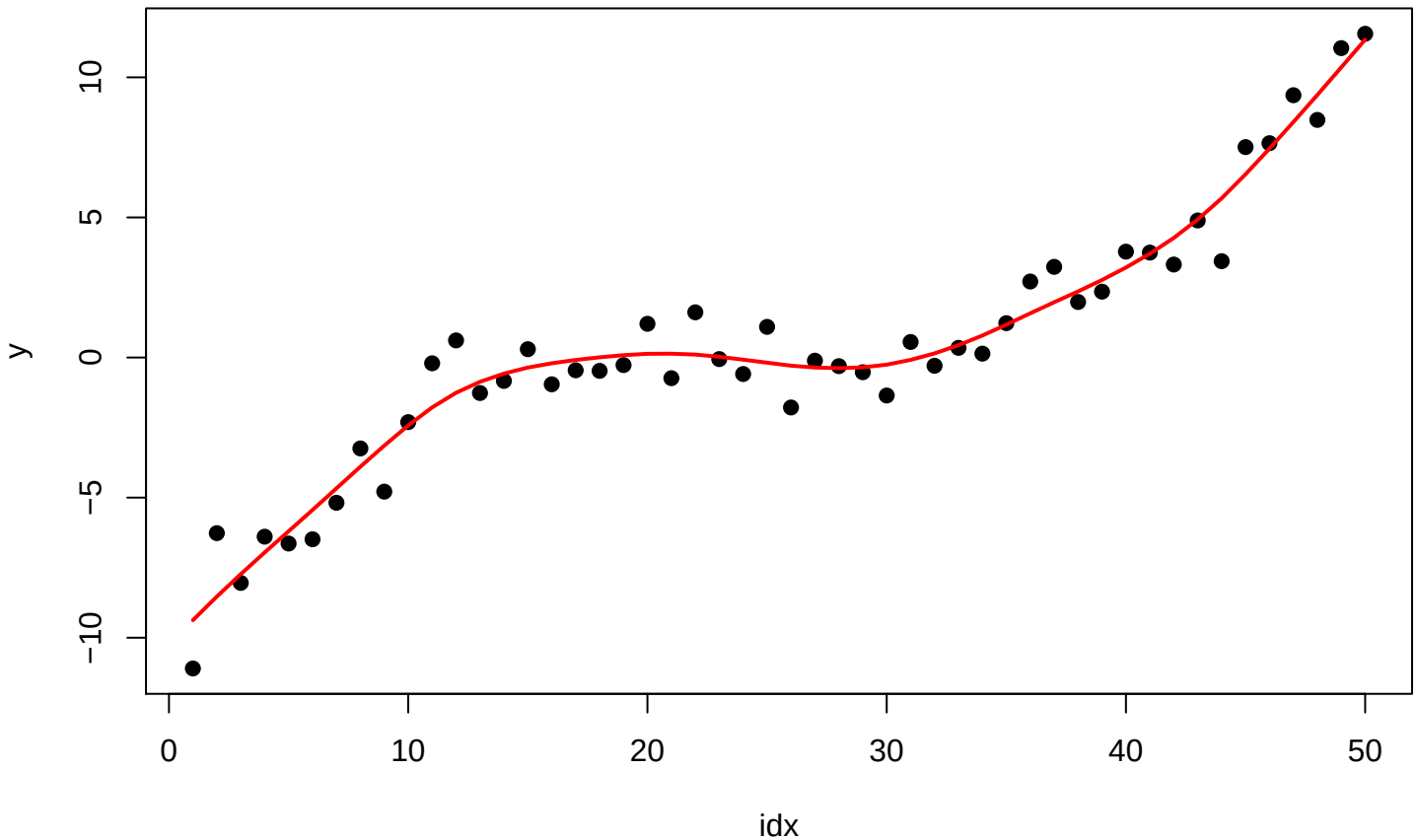
Posterior marginal for latent parameters

Numerically sum over all conditional densities to obtain the posterior marginal for each x_i.



Fitted spline

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



R-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
           ))

plot(idx, y, pch=19)
lines(result$summary.random[[1]]$mean, col=2, lwd=2)
```

Extensions

This is the basic idea behind INLA. It is quite simple.

Extensions

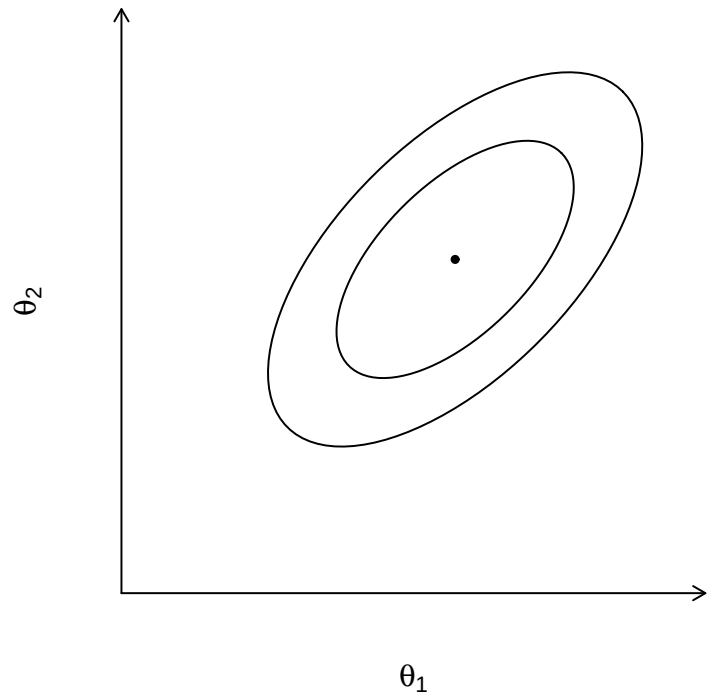
This is the basic idea behind INLA. It is quite simple.

However, we need to extend this basic idea so we can deal with

1. More than one hyperparameter
2. Non-Gaussian observations

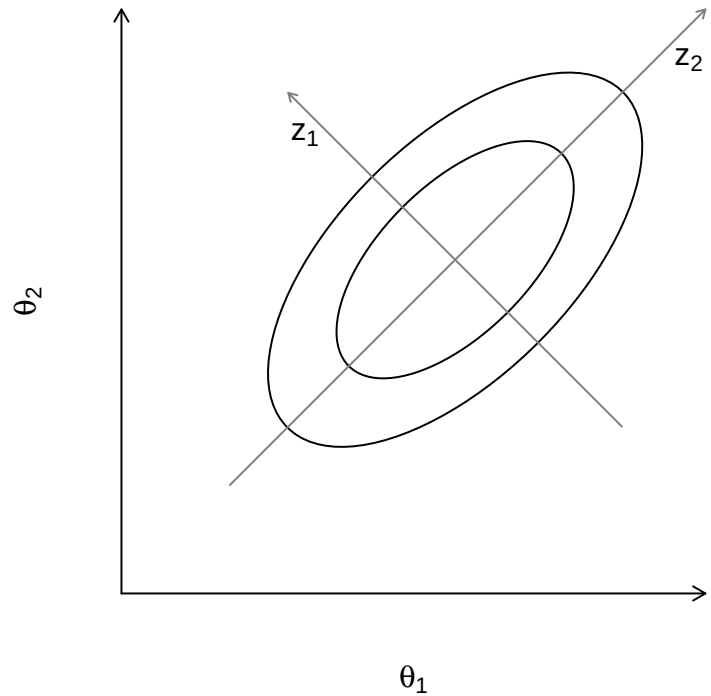
1. More than one hyperparameter

- Locate the mode



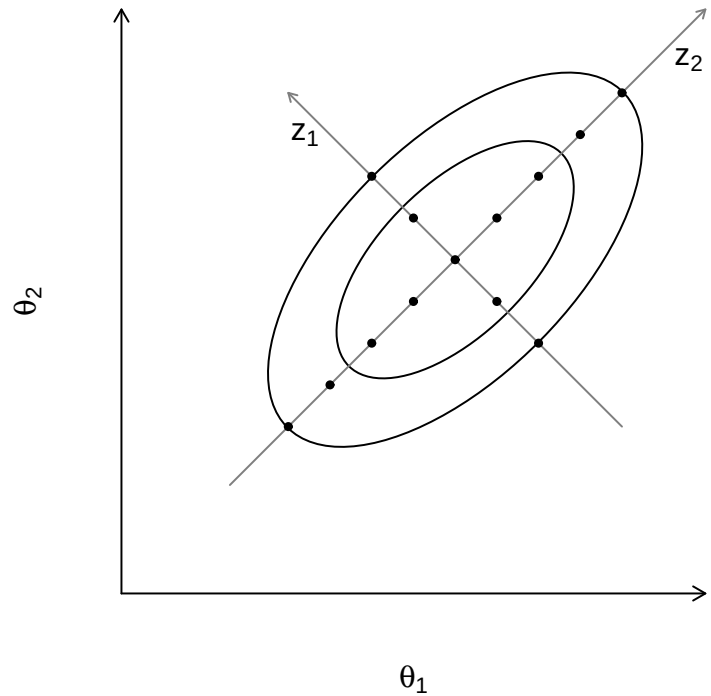
1. More than one hyperparameter

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



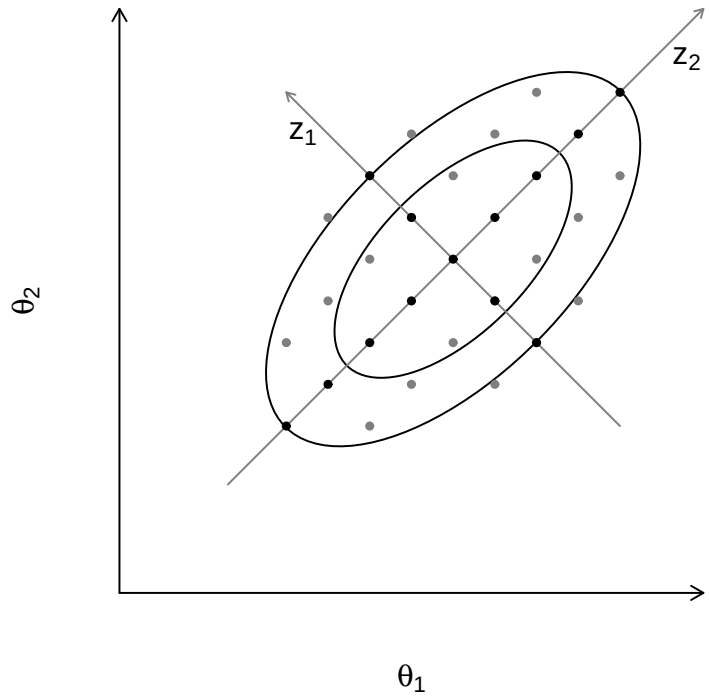
1. More than one hyperparameter

- 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



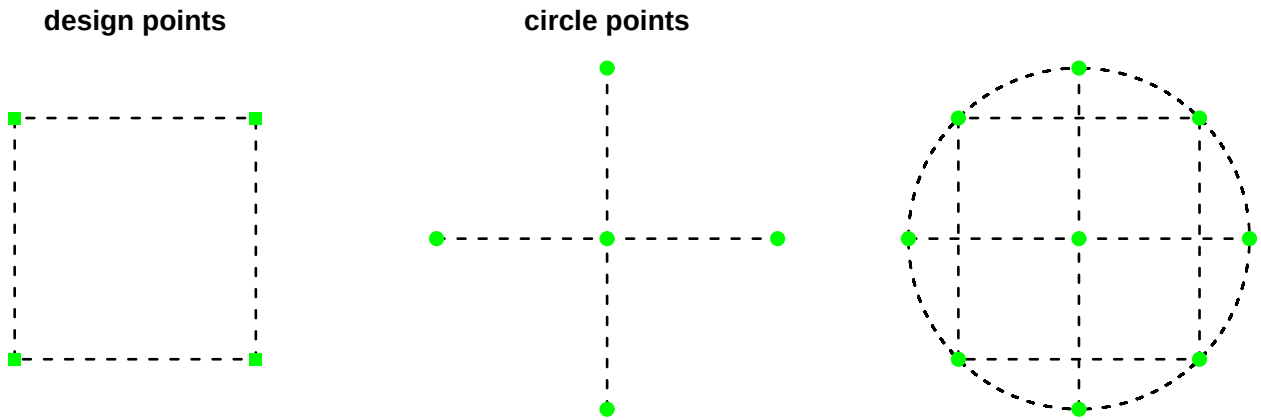
All points found have equal area weight Δ_k.

Alternatives for moderate number of hyperparameters

Integrating out the hyperparameter for moderate m (6 to 12) is expensive as the number of evaluation points is exponential in m.

Alternatives:

- Extreme: use just the modal configuration (empirical Bayes)
- Use a central composite design (CCD), e.g. for $m = 2$



2. Non-Gaussian observations

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

2. Non-Gaussian observations

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

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

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

The GMRF (Laplace) approximation

Let $\boldsymbol{x}$ denote a GMRF with precision matrix $\boldsymbol{Q}$ and mean $\boldsymbol{\mu}$. Approximate

$$\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \sum_{i=1}^{n} \log \pi(y_i|x_i)\right)$$

by using a second-order Taylor expansion of $\log \pi(y_i|x_i)$ around $\boldsymbol{\mu}_0$, say.

Recall

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 = a + bx - \frac{1}{2}cx^2$$

with $b = f'(x_0) - f''(x_0)x_0$ and $c = -f''(x_0)$. (Note: a is not relevant).

The GMRF approximation (II)

Thus,

$$\tilde{\pi}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \sum_{i=1}^{n}(a_i + b_i x_i - 0.5 c_i x_i^2)\right)$$
$$\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{Q} + \text{diag}(\boldsymbol{c}))\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right)$$

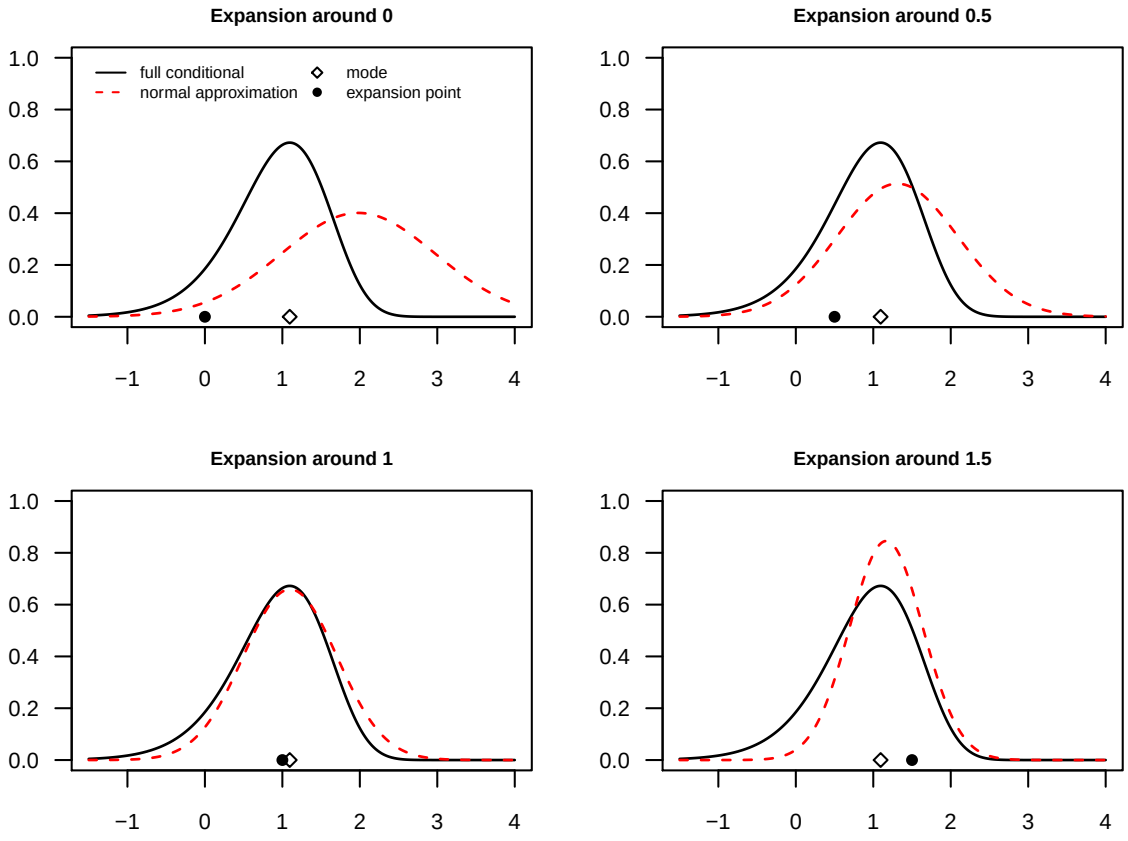
to get a Gaussian approximation with precision matrix $\boldsymbol{Q} + \text{diag}(\boldsymbol{c})$ and mean given by the solution of $(\boldsymbol{Q} + \text{diag}(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$. The canonical parameterisation is

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

which corresponds to

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

Illustration



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

What do we get ...

$$\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{x}=\boldsymbol{x}^\star(\boldsymbol{\theta})}$$

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

What do we get ...

$$\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{x}=\boldsymbol{x}^\star(\boldsymbol{\theta})}$$

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

However, why is it called integrated nested Laplace approximation?

What do we get ...

$$\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\tilde{\pi}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{x}=\boldsymbol{x}^\star(\boldsymbol{\theta})}$$

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

However, why is it called integrated nested Laplace approximation?

There is another step that changes:

$$\pi(x_i \mid \boldsymbol{y}) \approx \sum_k \underbrace{\pi(x_i \mid \boldsymbol{y}, t_k)}_{\text{Not Gaussian!}} \tilde{\pi}(t_k \mid \boldsymbol{y}) \Delta_k$$

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

Three possible approximations:

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

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

with mean $\mu_i(\boldsymbol{\theta})$ and marginal variance $\sigma_i^2(\boldsymbol{\theta})$.
However, errors in location and/or lack of skewness possible

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

Three possible approximations:

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

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

   with mean $\mu_i(\boldsymbol{\theta})$ and marginal variance $\sigma_i^2(\boldsymbol{\theta})$.
   However, errors in location and/or lack of skewness possible

2. Laplace approximation

3. Simplified Laplace approximation

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

$$\tilde{\pi}_{\mathsf{LA}}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_{\mathsf{GG}}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^{\star}_{-i}(x_i,\boldsymbol{\theta})}$$

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

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

$$\tilde{\pi}_{\mathsf{LA}}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_{\mathsf{GG}}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^{\star}_{-i}(x_i,\boldsymbol{\theta})}$$

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 $\tilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$
- 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 $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$
- 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.

The integrated nested Laplace approximation (INLA)

Step I Approximate $\pi(\boldsymbol{\theta}|y)$ using the Laplace approximation and select good evaluation points $\boldsymbol{t}_k$.

Step II For each $\boldsymbol{t}_k$ and i approximate $\pi(x_i|\boldsymbol{y}, \boldsymbol{t}_K)$ using the Laplace or simplified Laplace approximation for selected values of x_i .

Step III For each i, sum out $\boldsymbol{t}_k$

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

How can we assess the error in the approximations?

Tool 1: Compare a sequence of improved approximations

1. Gaussian approximation
2. Simplified Laplace
3. Laplace

No big differences $\rightarrow$ good approximation.

How can we assess the error in the approximations?

Tool 2: Estimate the "effective" number of parameters as defined in the Deviance Information Criteria:

$$p_{\mathrm{D}}(\boldsymbol{\theta}) = \overline{D}(\boldsymbol{x};\boldsymbol{\theta}) - D(\overline{\boldsymbol{x}};\boldsymbol{\theta})$$

and compare this with the number of observations.

Low ratio is good.

This criteria has theoretical justification.

Limitations

- The dimension of the latent field $\boldsymbol{x}$ can be large (10^2–10^6)
- But the dimension of the hyperparameters $\boldsymbol{\theta}$ must be small (≤ 15)

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

Model choice

Chose/compare various model is important but difficult

- Bayes factors (general available)
- Deviance information criterion (DIC) (hierarchical models)

Marginal likelihood

Marginal likelihood is the normalising constant for $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$,

$$\widetilde{\pi}(\boldsymbol{y}) = \int \left. \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\theta)\pi(\boldsymbol{y}|\boldsymbol{x},\theta)}{\widetilde{\pi}_{\mathsf{G}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^{\star}(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{1}$$

Marginal likelihood

Marginal likelihood is the normalising constant for $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$,

$$\widetilde{\pi}(\boldsymbol{y}) = \int \left. \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\theta)\pi(\boldsymbol{y}|\boldsymbol{x},\theta)}{\widetilde{\pi}_{\mathsf{G}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^\star(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{1}$$

I many hierarchical GMRF models the prior is intrinsic/improper, so this is difficult to use.

Deviance Information Criteria

Based on the *deviance*

$$D(\boldsymbol{x};\boldsymbol{\theta}) = -2\sum_i \log(y_i \mid x_i, \boldsymbol{\theta})$$

and

$$DIC = 2 \times \text{Mean}\left(D(\boldsymbol{x};\boldsymbol{\theta})\right) - D(\text{Mean}(\boldsymbol{x});\boldsymbol{\theta}^*)$$

This is quite easy to compute

Bayesian Cross-validation

Easy to compute using the INLA-approach

$$\pi(y_i \mid \boldsymbol{y}_{-i}) = \int_{\boldsymbol{\theta}} \left\{ \int_{x_i} \pi(y_i \mid x_i, \boldsymbol{\theta})\, \pi(x_i \mid \boldsymbol{y}_{-i}, \boldsymbol{\theta})\, dx_i \right\} \pi(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})\, d\boldsymbol{\theta}$$

where

$$\pi(x_i \mid \boldsymbol{y}_{-i}, \boldsymbol{\theta}) \propto \frac{\pi(x_i|\boldsymbol{y}, \boldsymbol{\theta})}{\pi(y_i|x_i, \boldsymbol{\theta})}$$

Require a one-dimensional integral for each i and $\boldsymbol{\theta}$.

Automatic detection of "surprising" observations

Compute

$$\text{Prob}(y_i^{\text{new}} \leq y_i \mid \boldsymbol{y}_{-i})$$

Look for unusual large or small values