

Bayesian Statistics with R-INLA - Part 2

Geilo, January, 2023

Sara Martino



NTNU

Norwegian University of
Science and Technology

Outline

Recap

Markov Properties

Deterministic inference for Gaussian models

Extending the method

Recap

Latent Gaussian models

Models of the kind:

$$\theta \sim \pi(\theta)$$

$$\mathbf{x}|\theta \sim \pi(\mathbf{x}|\theta) = \mathcal{N}(0, \mathbf{Q}^{-1}(\theta))$$

$$\mathbf{y}|\mathbf{x}, \theta \sim \prod_i \pi(y_i|\eta_i, \theta)$$

occurs in many, seemingly unrelated, statistical models.

Main Characteristics

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

Markov Properties

Precision Matrix

We always talk about the precision matrix

$$\mathbf{Q}(\theta) = \Sigma(\theta)^{-1}$$

* Interpretation of the elements of \mathbf{Q}

- The precision matrix gives direct information about the conditional independence structure of the model

$$x_i \perp x_j | \mathbf{x}_{-ij} \iff Q_{ij} = 0$$

Example - AR1 model

Consider an autoregressive process of order 1

$$\begin{aligned}x_t &= \phi x_{t-1} + \epsilon_t, & \epsilon_t &= \mathcal{N}(0, 1), & |\phi| &< 1 \\x_1 &\sim \mathcal{N}(0, 1/(1 - \phi^2))\end{aligned}$$

The joint distribution is:

$$\begin{aligned}\pi(\mathbf{x}) &= \pi(x_1)\pi(x_2|x_1)\dots\pi(x_n|x_{n-1}) \\&= \frac{1}{(2\pi)^{n/2}} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x}\right)\end{aligned}$$

Example - AR1 model

The precision matrix \mathbf{Q} is:

$$\mathbf{Q} = \begin{pmatrix} 1 & -\phi & & & & \\ -\phi & 1 + \phi^2 & -\phi & & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\phi & 1 + \phi^2 & -\phi \\ & & & & -\phi & 1 \end{pmatrix}$$

Example - AR1 model

The precision matrix \mathbf{Q} is:

$$\mathbf{Q} = \begin{pmatrix} 1 & -\phi & & & \\ -\phi & 1 + \phi^2 & -\phi & & \\ & & \ddots & \ddots & \ddots \\ & & & -\phi & 1 + \phi^2 & -\phi \\ & & & & -\phi & 1 \end{pmatrix}$$

The tridiagonal form is due to the fact that x_i and x_j are conditionally independent for $|i - j| > 1$, given the rest.

Example - AR1 models

Variance covariace function:

$$\Sigma = \begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 & \dots & \dots & \phi^{n-1} \\ \phi & 1 & \phi & \phi^2 & \dots & \dots & \phi^{n-2} \\ & & \ddots & \ddots & \ddots & & \\ \phi^{n-2} & \phi^{n-3} & \dots & \dots & \phi & 1 & \phi \\ \phi^{n-1} & \phi^{n-2} & \dots & \dots & \phi^2 & \phi & 1 \end{pmatrix}$$

Precision Matrix

There are two main advantages in using (sparse) precision matrix over the variance/covariance matrix:

1. Building models through conditioning (“Hierarchical models”)
2. Computational

Building models through conditioning

If

- $\mathbf{x} \sim \mathcal{N}(0, \mathbf{Q}_x^{-1})$
- $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(0, \mathbf{Q}_y^{-1})$

Building models through conditioning

If

- $\mathbf{x} \sim \mathcal{N}(0, \mathbf{Q}_x^{-1})$
- $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(0, \mathbf{Q}_y^{-1})$

then

$$\mathbf{Q}_{(\mathbf{x}, \mathbf{y})} = \begin{bmatrix} \mathbf{Q}_x + \mathbf{Q}_y & -\mathbf{Q}_y \\ -\mathbf{Q}_y & \mathbf{Q}_y \end{bmatrix}$$

We don't get so nice expressions with Covariance matrix

Computational benefits

- Models we have seen gives a sparse precision matrix
- These are much faster to compute with, than dense matrices
- Special case: Kalman-filter algorithms

Computational benefits

- Models we have seen gives a sparse precision matrix
- These are much faster to compute with, than dense matrices
- Special case: Kalman-filter algorithms

Tasks:

- Factorize \mathbf{Q} into $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$ (Cholesky)
- Solve $\mathbf{Q}\mathbf{x} = \mathbf{b}$, $\mathbf{L}\mathbf{x} = \mathbf{b}$ or $\mathbf{L}^T\mathbf{x} = \mathbf{b}$
- Compute $\text{diag}(\mathbf{Q}^{-1})$

Numerical algorithms for sparse matrices: scaling properties

Cholesky factorization of “sparse” SPD¹ matrix

- Time: $\mathcal{O}(n)$
- Space: $\mathcal{O}(n^{3/2})$
- Space-time: $\mathcal{O}(n^2)$

¹Symmetric and positive definite

Numerical algorithms for sparse matrices: scaling properties

Cholesky factorization of “sparse” SPD¹ matrix

- Time: $\mathcal{O}(n)$
- Space: $\mathcal{O}(n^{3/2})$
- Space-time: $\mathcal{O}(n^2)$

This is to be compared with general $\mathcal{O}(n^3)$ algorithm for the Cholesky factorization of a SPD dense matrix.

¹Symmetric and positive definite

Gaussian Markov random fields

- Gaussians with a sparse precision matrix are called *Gaussian Markov random fields* (GMRFs)
- Good computational properties through numerical algorithms for sparse matrices
- Very useful for doing MCMC-based inference as well

Summary

Three main ingredients in INLA

- Gaussian Markov random fields
- Latent Gaussian models
- Laplace approximations

Summary

Three main ingredients in INLA

- Gaussian Markov random fields
- Latent Gaussian models
- Laplace approximations

which together (and +++++...) gives a very very nice tool for Bayesian inference:

- quick
- accurate (relative error)
- good scaling properties
- ++++

Deterministic inference for Gaussian models

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**:

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

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

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

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

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**

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

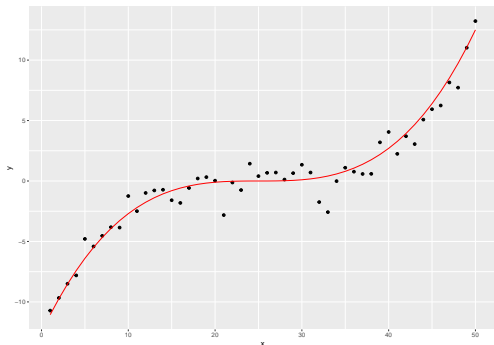
directly.

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 - Model

Assume:

$$y_i = f(i) + \epsilon_i; i = 1, \dots, n$$

$$\epsilon_i \sim N(0, 1)$$

$$f(i) = x_i \text{ smooth function of } i$$

- Only one hyperparameter
- Gaussian likelihood

Is this a Latent Gaussian model?

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 θ . 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

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

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

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

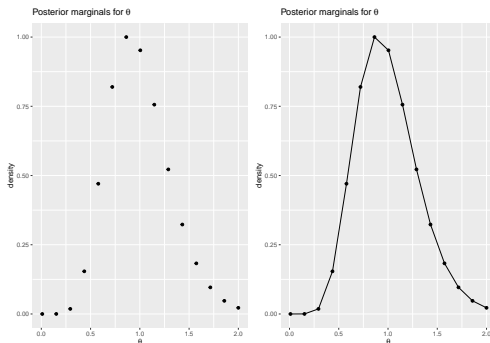
Since the likelihood is Gaussian, then $\pi(\mathbf{y}, \mathbf{x}|\theta)$ is also Gaussian.
We have then:

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

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!!

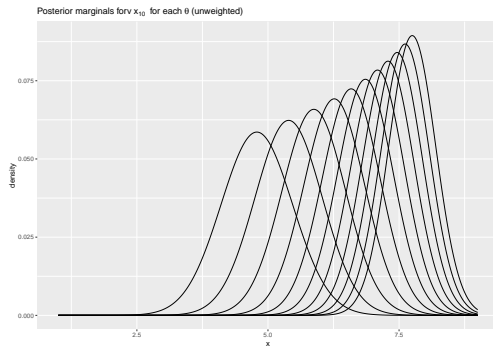
We compute

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

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

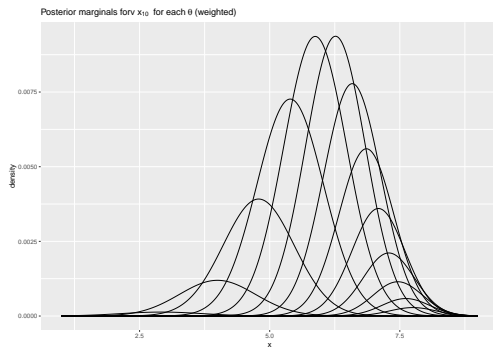
Posterior marginals for latent field I

Compute the conditional posterior marginal for x_i given each θ_k



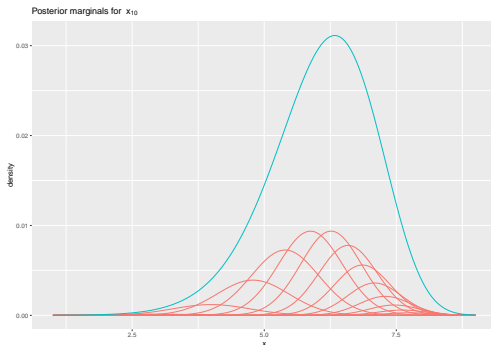
Posterior marginals for latent field II

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



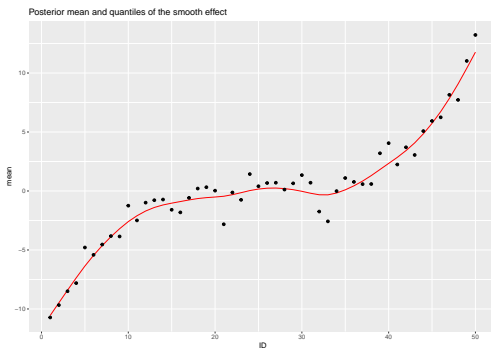
Posterior marginals for latent field III

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))
```

This exercise is contained in the
01_Practical_implement_INLA.html file.

Extending the method

Extending the method

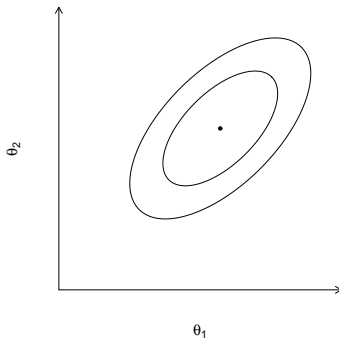
This is the basic idea behind INLA. It is quite simple.

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

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

1. More than one hyperparameter

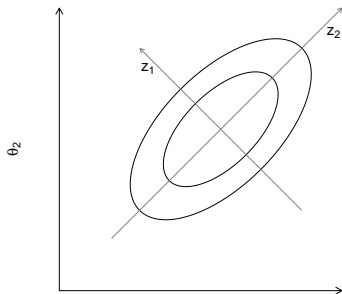
Main use: Select good evaluation points θ_k for the numerical integration when approximating $\tilde{\pi}(x_i|y)$



1. More than one hyperparameter

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

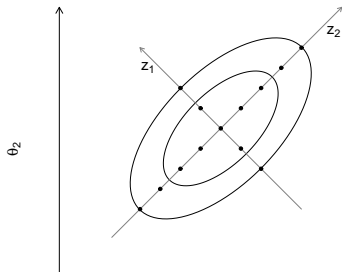
- Locate the mode
- Compute the Hessian to construct principal components



1. More than one hyperparameter

Main use: Select good evaluation points θ_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:
 - $\tilde{\pi}(\theta^k|y)$
 - $\tilde{\pi}(x_i|\theta^k, y)$
 - Δ_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 (Laplace) approximation

Let \mathbf{x} denote a GMRF with precision matrix \mathbf{Q} and mean μ .

Approximate

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

by using a second-order Taylor expansion of $\log \pi(y_i|x_i)$ around μ_0 , say.

- Recall

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

with

$$b = f'(x_0) - f''(x_0)x_0$$

$$c = -f''(x_0)$$

.

(Note: a is not relevant).

The GMRF approximation (II)

Thus,

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

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

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

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

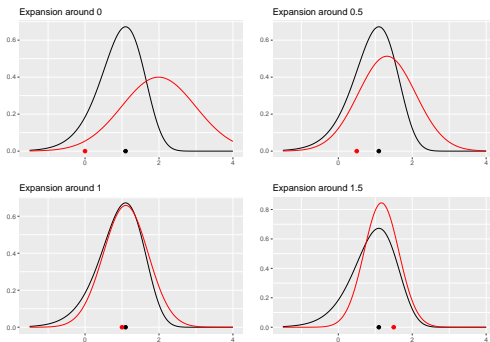
$$\lambda = \exp(x) \text{ Likelihood}$$

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

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

What do we get ...

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

- 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 ...

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

- 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.

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

What do we get ...

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

- 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:

$$\pi(x_i | \mathbf{y}) \approx \sum_k \underbrace{\pi(x_i | \mathbf{y}, \theta^k)}_{\text{Not Gaussian!}} \tilde{\pi}_G(\theta^k | \mathbf{y}) \Delta_k$$

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

Three possible approximations:

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.

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

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.

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

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.

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

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.

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

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.

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

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**

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

Use again the same idea!

Based on the identity

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

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

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

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

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

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

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

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: When does it work

We consider models of the kind

$$\theta \sim \pi(\theta)$$

$$\mathbf{x}|\theta \sim \pi(\mathbf{x}|\theta) = \mathcal{N}(0, \mathbf{Q}^{-1}(\theta))$$

$$\mathbf{y}|\mathbf{x}, \theta \sim \prod_i \pi(y_i|\eta_i, \theta)$$

where

- \mathbf{x} can be large but endowed with Markov properties so that $\mathbf{Q}(\theta)$ is sparse
- the size of θ is small (say <15)
- η is a predictor that depends *linearly* on the other elements of \mathbf{x}
- The main inferential interest lies in the posterior marginals $\pi(x_i|\mathbf{y})$, $\pi(\theta_j|\mathbf{y})$ rather than in the joint $\pi(\mathbf{x}, \theta|\mathbf{y})$ (...but joint inference is possible through sampling!)

INLA: Overview

- **Step I** Approximate $\pi(\theta|y)$ using the Laplace approximation and select good evaluation points θ_k .
- **Step II** For each θ_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 θ_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 θ must be small (≤ 9)

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

INLA: summary

- These are the basic ideas
- The rest are *just* details...but there are a lot of them!

INLA features

INLA fully incorporates posterior uncertainty with respect to hyperparameters \Rightarrow tool for full Bayesian inference

- 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 $\tilde{\pi}(x, \theta|y)$