NTNU
Norwegian University of
Science and Technology

Lecture 11: Introduction to INLA

Andrea Riebler <andrea.riebler@math.ntnu.no>

01.03.2016

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? (people infected with a disease in each area, temperature, rainfall, fish weight ...)
- Count data? (people infected with a disease in each area)
- Point pattern? (E.g. air pollution measured at fixed stations)
- 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:

- Covariates
- Unstructured random effects (individual effects, group effects)
- Structured random effects (AR(1), regional effects, continuously indexed spatial 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:

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: 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.

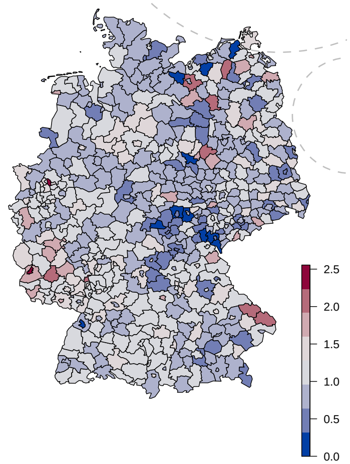
Information available:

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).



Stage 1: The data

First we decide on the likelihood for our data $\boldsymbol{y}$

- Our responses are counts
- We decide to model our responses as

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

- η_i is a linear function of the latent components

Stage 2: The latent model

The latent field $\boldsymbol{x}$ consists of two parts:

1. One fixed effect: the intercept μ
2.
   - The spatially structured effect $\boldsymbol{u}$.
   - The unstructured effect $\boldsymbol{v}$ which accounts for non-observed variability
   - The unknown effect $f(c_i)$ of the exposure covariate which assumes value c_i for district i.

These are combined for each location to give a linear predictor

$$\eta_i = \mu + u_i + v_i + f(c_i)$$

The latent field is $\boldsymbol{x} = (\mu, u_1, u_2, \ldots, u_n, v_1, v_2, \ldots, v_n, \{f(\cdot)\})$.

A spatially structured effect

To incorporate a spatial structure into a model, the so called Besag model is often used.

$$p(\boldsymbol{u} \mid \kappa_u) \propto \kappa_u^{(n-1)/2} \exp\left(-\frac{\kappa_u}{2}\sum_{i\sim j}(u_i - u_j)^2\right)$$
$$= \kappa_u^{(n-1)/2} \exp\left(-\frac{\kappa_u}{2}\boldsymbol{u}^T\mathbf{R}\boldsymbol{u}\right).$$

where R is called structure matrix and defined as

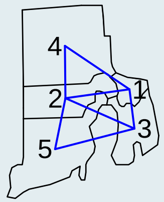
$$R_{ij} = \begin{cases} n_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise.} \end{cases}$$

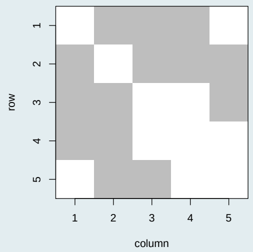
Here, $i \sim j$ denotes that i and j are neighbouring regions.

What does this mean?

Example: Five counties of the US state Rhode Island

The structure matrix **R** defines the neighborhood structure.





Adjacency matrix

3	-1	-1	- 1	0
-1	4	-1	-1	-1
-1	-1	3	0	-1
-1	-1	0	2	0
0	-1	-1	0	2

Structure matrix

With increasing number of regions **R** will be sparse, which allows to do many computations very efficient.

Gaussian Markov random field (GMRF)

- This model is an example for a Gaussian Markov random field (GMRF) model.
- If **R** has not full rank it is called an intrinsic GMRF.
- Other examples are a random walk of first order, a random walk of secoend order, an autoregressive model, ...

Stage 3: Hyperparameters

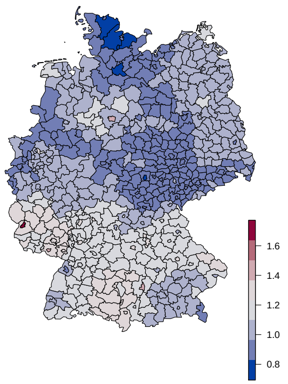
The structured and unstructured spatial effect as well as the smooth covariate effect will be each controlled by one parameter

- $\kappa_f, \kappa_u, \kappa_v$: The precisions (inverse variances) of the covariate effect, spatial effect and unstructured effect, respectively.

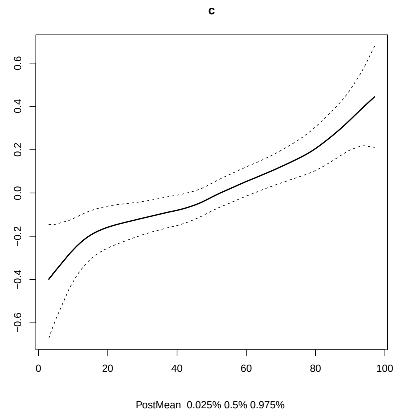
The hyperparameters are $\boldsymbol{\theta} = (\kappa_f, \kappa_u, \kappa_v)$, and must be given a prior $\pi(\kappa_f, \kappa_u, \kappa_v)$.

Quantities of interest

Structured spatial effect $exp(u_i)$



Covariate effect $f(c_i)$



Latent Gaussian models

This is 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, $\boldsymbol{x}|\theta \sim N(0, \mathbf{Q}^{-1})$

— 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,

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

where

μ : Intercept

β : 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

$\boldsymbol{u}$: Unstructured error terms

Loads of examples

— Generalized linear and additive (mixed) models

— Disease mapping

— Survival analysis

— Log-Gaussian Cox-processes

— 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 linear predictor in a 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.

Example: Smoothing binary time-series

- Have observed a sequence $y_1, y_2, \ldots, y_n$ of 0s and 1s
- Each time t has an associated covariate x_t
- We want to smooth the time series by inferring the sequence p_t, for $t = 1, 2, \ldots, n$, of probabilities for 1s at each time step

Example: Smoothing time series

Stage 1: We choose a Bernoulli distribution for the responses, so that

$$y_t|\eta_t \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\eta_t)}\right)$$

Stage 2: Covariates, AR(1) component, i.e. $a_t = \rho a_{t-1} + \epsilon_t$, and random noise are connected to likelihood by

$$\eta_t = \beta_0 + \beta_1 x_t + a_t + v_t$$

Stage 3: ρ: Dependence parameter in AR(1) process

σ_a^2: Marginal variance in AR(1) process

σ_v^2: Variance of unstructured term

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)
- 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:

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

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

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

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

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

What do we need to compute?

To be more precise, often we are interested in the posterior probability density of an element of the latent field

$$\pi(x_i|\boldsymbol{y})$$

or the posterior probability density of an element of the hyperparameters

$$\pi(\theta_j|\boldsymbol{y})$$

or some other statistics

$$\pi(f(\boldsymbol{x},\boldsymbol{\theta})|\boldsymbol{y})$$

Traditional approach: MCMC*

Based on sampling. Construct Markov chains 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

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

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

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

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.

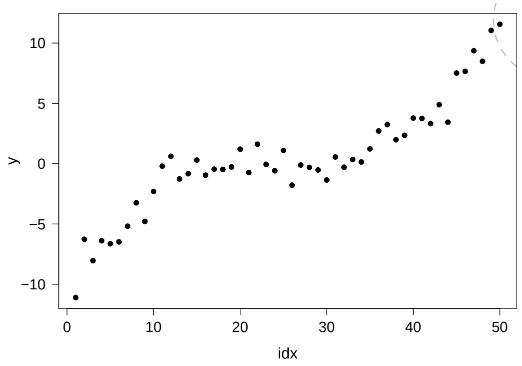
How does INLA work?

Observations

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

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.

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



Assumed hierarchical model

1. Data: Gaussian observations with known precision

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

2. Latent model: A Gaussian model for the smooth function[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

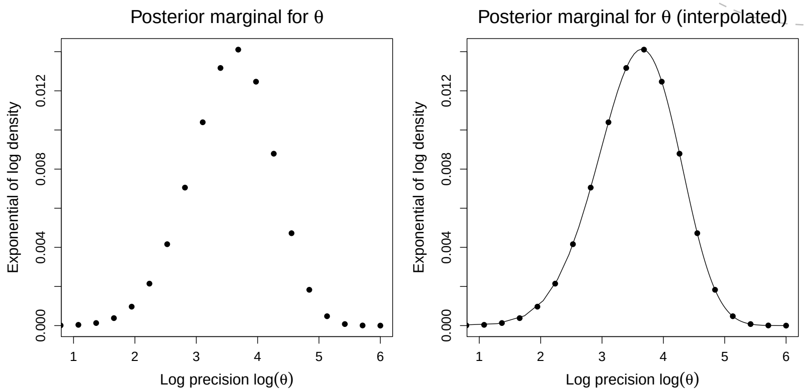
$$\boldsymbol{x}, \boldsymbol{y} \mid \theta \sim \mathcal{N}(\cdot, \cdot)$$

(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

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

Posterior marginal for hyperparameter

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



Derivation of posterior marginals (II)

From

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

we can compute

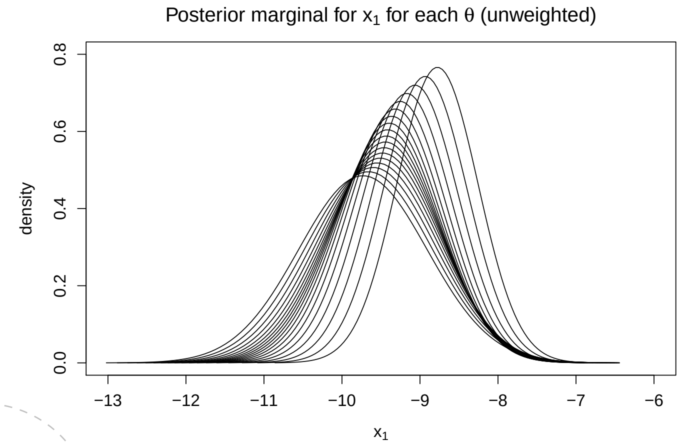
$$\pi(x_i \mid \boldsymbol{y}) = \int \underbrace{\pi(x_i \mid \theta, \boldsymbol{y})}_{\text{Gaussian}} \pi(\theta \mid \boldsymbol{y})\, d\theta$$

$$\approx \sum_k \pi(x_i \mid \theta_k, \boldsymbol{y}) \pi(\theta_k \mid \boldsymbol{y}) \Delta_k$$

where θ_k, $k = 1, \ldots, K$, correspond to representative points of $\theta \mid \boldsymbol{y}$ and Δ_k are the corresponding weights.

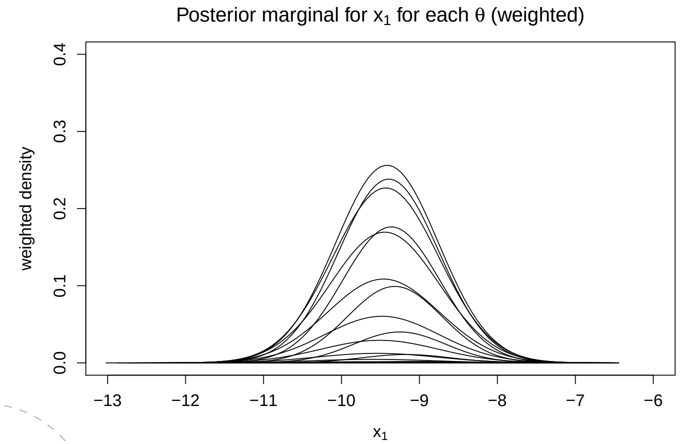
Posterior marginal for latent parameters

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



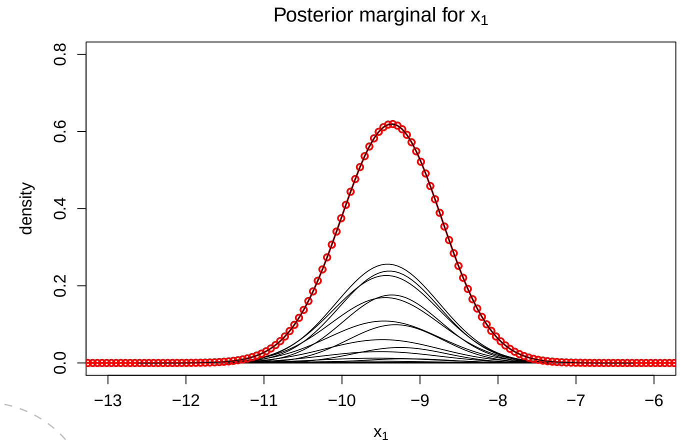
Posterior marginal for latent parameters

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



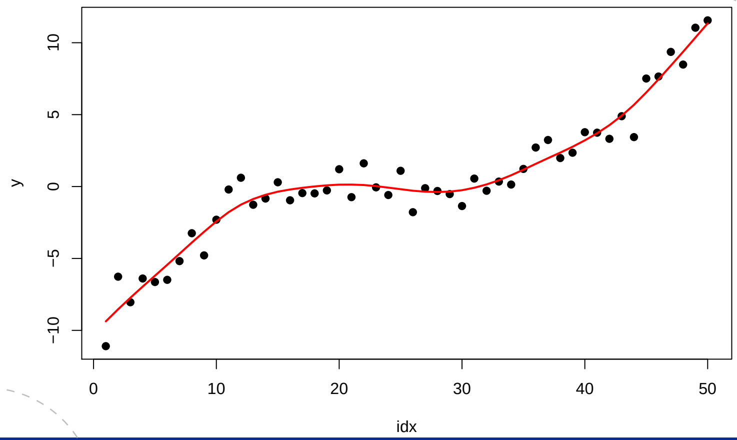
Posterior marginal for latent parameters

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:



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

- More than one hyperparameter
- Non-Gaussian observations

How, do things change?

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

Complications... Mostly practical

The non-Gaussian part of the model

- 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
- This means that all the really hard, high-dimensional integrals with respect to the latent field are easy, and only the integrals with respect to the hyperparameters remain
- If the number of hyperparameters is low, these integrals can be done efficiently numerically

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

The GMRF approximation (II)

Thus,

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

$$\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{Q} + \text{diag}(\boldsymbol{c}))\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right)$$

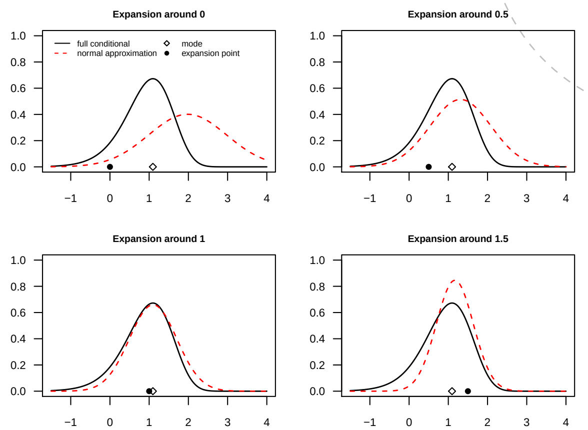
to get a Gaussian approximation with precision matrix $\boldsymbol{Q} + \text{diag}(\boldsymbol{c})$ and mean given by the solution of $(\boldsymbol{Q} + \text{diag}(\boldsymbol{c}))\boldsymbol{\mu} = \boldsymbol{b}$. The canonical parameterisation is

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

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!

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 (≤ 9)

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

How to use INLA?

INLA is implemented through the package `R-INLA` in the `R` software which

- is the most popular computing language in applied statistics
- is open source and *free*
- has a lot of packages that extend the functionality
- has a very user friendly `formula` interface

```
linear_model <- lm(weight ~ group)
```

Fits the linear model

$$\text{weight}_i = \mu + \text{group}_i + \epsilon_i$$