ANALYSIS OF SPATIAL DATA IN EPIDEMIOLOGY

Prof. Dr. Maria A Barceló and Prof. Dr. Marc Saez

September 8, 10, 14 and 16, 2021

Research Group on Statistics, Econometrics and Health (GRECS), University of Girona
CIBER of Epidemiology and Public Health (CIBERESP)

GRECS
GRUP DE RECERCA EN ESTADÍSTICA ECONOMETRIA I SALUT

UdG

ciberesp
Centro de Investigación Biomédica en Red
Epidemiología y Salud Pública

COURSE INTRODUCTION

1. Course introduction
2. Introduction to epidemiology and spatial statistics
3. Overview of mixed models
4. Overview of mixed models - Practicals
5. **Introduction to INLA and R INLA**
6. R INLA - Practicals

Wednesday 8

Friday 10

Marc Saez Maria Antònia Barceló

COURSE INTRODUCTION

7. Disease mapping. Standardisation of incidence and mortality rates
8. Disease mapping. Smoothing standardised incidence and mortality rates
9. Disease mapping – Practicals
10. Geographical association studies. Spatial ecological regression
11. Spatial ecological regression - Practicals

Tuesday 14

Marc Saez Maria Antònia Barceló

COURSE INTRODUCTION

12. Clustering
13. Extensions: BYM2, point processes, leaflet, pc priors
14. Extensions – Practicals

Thursday 16

Marc Saez Maria Antònia Barceló

INTRODUCTION TO INLA AND R INLA

1. Bayesian statistics
2. INLA
3. R INLA

INTRODUCTION TO INLA AND R INLA

1. **Bayesian statistics**
2. INLA
3. R INLA

BAYESIAN STATISTICS

- Very schematically, it could be said that for **frequentists** (classical statisticians), probability is considered as the limit of the relative frequency when an experiment is performed repeatedly a very large number of times under identical conditions..
- For Bayesians, on the other hand, probability is the fundamental measure of uncertainty and this subjective concept of probability must be constructed with scientific judgment..

BAYESIAN STATISTICS

- In a Bayesian model, we generally want the a **posteriori distribution** for our models (e.g., the distribution of parameters given the data), or a **posteriori predictive distributions** (for extrapolation/prediction - the distribution of new values given the observed ones)

BAYESIAN STATISTICS

- The a posteriori distribution is equal to the probability of observing the data multiplied by the a priori distribution of the parameters (or **priors**), with a normalization constant (so that the a posteriori integral is equal to 1).

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int p(\mathbf{y}|\theta)p(\theta)d\theta}$$

- In a more simplified way (without considering the constant of normalisation)

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

BAYESIAN STATISTICS

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

θ is the vector of **parameters**.

$p(y|\theta)$ is known as ***likelihood*** (the model).

$p(\theta)$ is the a priori distribution, or **priors**.

BAYESIAN STATISTICS

- The choice of the priors to be used in each case is a subjective choice and must often be decided on the basis of expert judgment and the type of data available
- When the a posteriori distribution is from the same family as the a priori distribution used, we talk about conjugate distributions.
- The advantage of using these is that “priors” have good mathematical properties calculate the a posteriori distributions.

Conjugated priors

Versemblança	Paràmetre a estimar	Prior
Normal	Mitjana	Normal
Normal	Precisió (1/variància)	Gamma
Binomial	Probabilitat d’èxit	Beta
Poisson	Mitjana	Gamma

BAYESIAN STATISTICS

- In a **frequentist approximation (estimation)** we often maximize the probability of the data (that is, the ***likelihood***), using numerical methods, such as Newton-Raphson or others, to obtain a point estimate of a given parameter (which we see as non-random – in other words, fixed – but unknown).
- In a **Bayesian approximation (*computing* or inference)** we obtain an a posteriori distribution for the parameter (which is considered to be a random variable), for which we can provide summary statistics (mean, median or mode) and quantiles to directly obtain intervals of credibility.

BAYESIAN STATISTICS

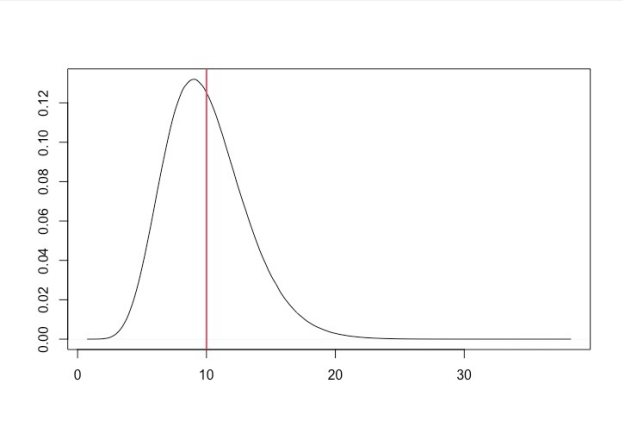
- The problem with Bayesian approximation is that while the likelihood and a priori distribution are easy to obtain, $p(\theta|y)$ tends to be analytically intractable (especially when we do not use conjugated priors).

BAYESIAN COMPUTING

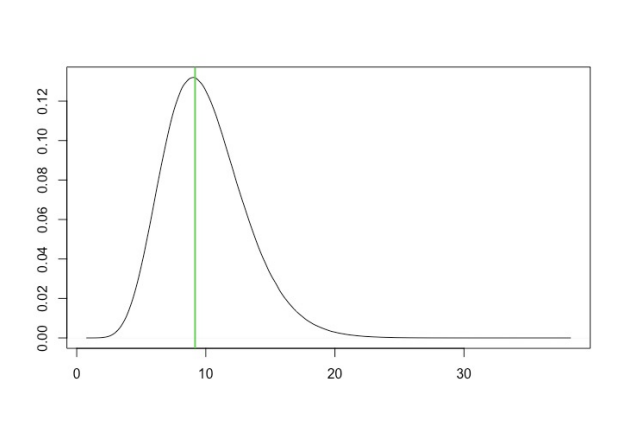
- We want to obtain the (marginal) a posteriori distribution, $p(\theta|y)$:

$$p(\theta_i|y) = \int \int \dots \int p(\theta|y)\, d\theta_{(-i)}$$

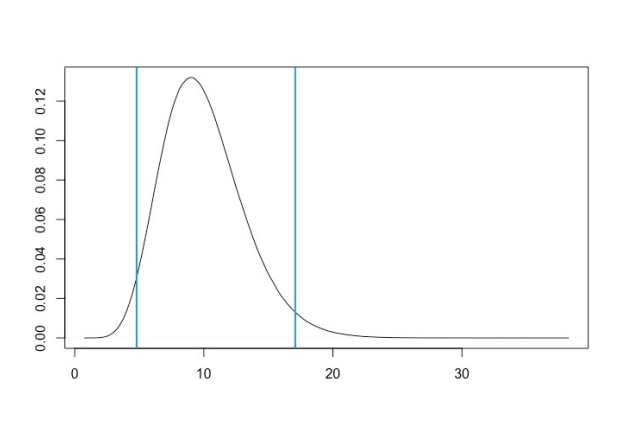
where $\theta_{(-i)}$ denotes vector θ excluding the component i.



Mean



Median



Interval of credibility at 95%

BAYESIAN COMPUTING

- In general, the integrals are intractable and numerical methods like **Markov chain Monte Carlo** (MCMC) methods are used to simulate samples of conditional distributions and to calculate the marginal distribution of each parameter of interest.
- A sequence of random variables $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$ form a Markov chain if $\theta^{(i+1)} \longrightarrow p(\theta | \theta^{(i)})$.
- That is, conditioned to the value $\theta^{(i)}$, $\theta^{(i+1)}$ is independent of $\theta^{(i-1)}$, ..., $\theta^{(0)}$.

BAYESIAN COMPUTING

- There are various algorithms to design Markov chains.
- Among them, the 'Gibbs sampling' algorithm is one of the simplest of the MCMC.
- However, there are also others: Metropolis, Metropolis-Hastings, etc.

BAYESIAN COMPUTING

Gibss sampling

- Let θ be a vector of unknown parameters $\theta = (\theta_1, \theta_2, ..., \theta_k)$

1. Initial values are chosen $\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_k^{(0)}$ for the components.
2. $\theta_1^{(1)}$ is sampled from $p(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, ... \theta_k^{(0)}, x)$

   $\theta_2^{(1)}$ is sampled from $p(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, ... \theta_k^{(0)}, x)$

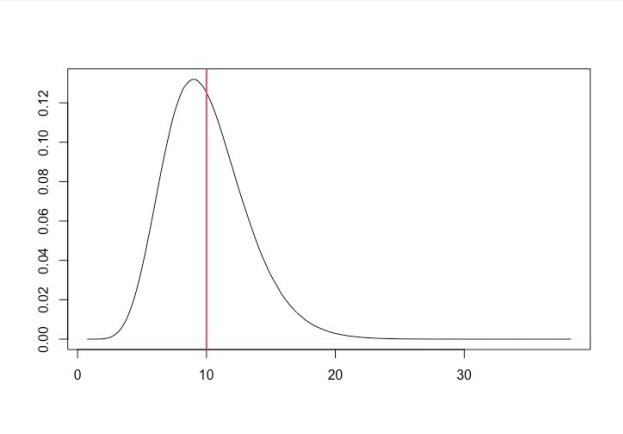
   $\theta_k^{(1)}$ is sampled from $p(\theta_k | \theta_1^{(1)}, \theta_2^{(1)}, ... \theta_{k-1}^{(1)}, x)$
3. Stage 2 is repeated many times. If the number of repetitions is very high we obtain a simple for $p(\theta | x)$.

BAYESIAN COMPUTING

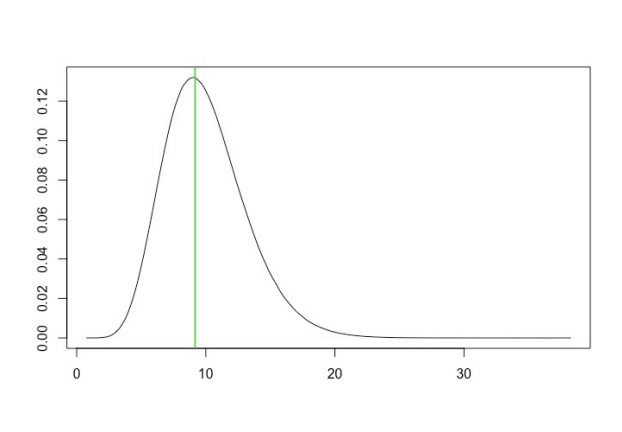
- We want to obtain the (marginal) a posteriori distribution, $p(\theta|y)$:

$$p(\theta_i|y) = \int \int \dots \int p(\theta|y)\, d\theta_{(-i)}$$

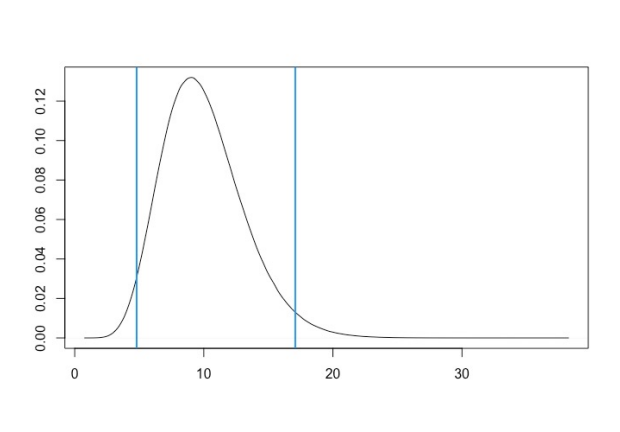
where $\theta_{(-i)}$ denotes the vector θ excluding the component i.



Mean



Median



Interval of credibility at 95%

BAYESIAN COMPUTING

- MCMC have been developed in software such as WinBUGS.
- MCMC are slow, do not scale well (in other words, results are not invariant to changes in scale and/or sample size) and, for some complex models, they can fail (the model will not converge). More recent programming (JAGS, Stan) has attempted to overcomes these challenges.

BAYESIAN COMPUTING

- Alternative **INLA (*Integrated Nested Laplace Approximations*)**.

INTRODUCTION TO INLA AND R INLA

1. Bayesian statistics
2. **INLA**
3. R INLA

INLA

- MCMC is an asymptotically exact method, while INLA is an approximation.
- Empirically, the MCMC error and the INLA error tend to be very similar, as has been demonstrated in many simulation studies.

Elapsed time in seconds

n	rjags	r-inla
100	4.19	0.176
500	18.141	0.359
5000	381.573	2.787
25000	2203.679	13.27
100000	8873.836	52.787

Simple linear regression

Elapsed time in seconds

n	rjags	r-inla
100	30.394	0.383
500	142.532	1.243
5000	1714.468	5.768
25000	8610.32	30.077
100000	got bored after 6 hours	166.819

Poisson regression with random effects (non-structured) in the constant

https://www.precision-analytics.ca/articles/a-gentle-inla-tutorial/

INLA

Random field, Gaussian field (GF), Gaussian Markov Random Field (GMRF)

- When using Bayesian inferences for GMRF, INLA can be used (instead of MCMC).

Many environmental phenomena, even if defined continuously over a region and in time, can be monitored and measured only at a limited number of spatial locations and time points. This is the case, for example, of air pollutant concentration, meteorological fields (temperature, precipitation, wind velocity, etc.) as well as geohydrological and oceanographic variables (soil moisture, wave height, etc.). In the geostatistical approach (see, for example, Cressie 1993; Gelfand et al. 2010; Cressie and Wikle 2011), data coming from monitoring networks are assumed to be realisations of a continuously indexed spatial process (*random field*) changing in time denoted by

$$Y(\boldsymbol{s},t) \equiv \left\{ y(\boldsymbol{s},t) : (\boldsymbol{s},t) \in \mathcal{D} \subseteq \mathbb{R}^2 \times \mathbb{R} \right\}.$$

Cameletti M, Lindgren F, Simpson D, Rue H. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Adv Stat Anal.* 2013; 97(2):109–131. doi: 10.1007/s10182-012-0196-3.

INLA

Random field, Gaussian field (GF), Gaussian Markov Random Field (GMRF)

- Gaussian field (GF).

These realisations are used to make inference about the process and to predict it at desired locations. Usually, we deal with a Gaussian field (GF) that is completely specified by its mean and spatio-temporal covariance function $\text{Cov}(y(\boldsymbol{s}, t), y(\boldsymbol{s}', t')) = \sigma^2 \mathcal{C}((\boldsymbol{s}, t), (\boldsymbol{s}', t'))$, defined for each $(\boldsymbol{s}, t)$ and $(\boldsymbol{s}', t')$ in $\mathbb{R}^2 \times \mathbb{R}$. Moreover, the process is second-order stationary if its mean is constant and the spatio-temporal covariance function depends on the locations and time points only through the spatial distance vector $\mathbf{h} = (\boldsymbol{s} - \boldsymbol{s}') \in \mathbb{R}^2$ and the temporal lag $l = (t - t') \in \mathbb{R}$.

Cameletti M, Lindgren F, Simpson D, Rue H. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Adv Stat Anal.* 2013; 97(2):109–131. doi: 10.1007/s10182-012-0196-3.

INLA

GF, Big n problema, Gaussian Markov Random Field (GMRF)

- Suppose we have COVID-19 incidence data for 6 health areas, with the following geographical distribution:

1	2	3
4	5	6
7	8	9

INLA

GF, Big n problem, Gaussian Markov Random Field (GMRF)

- Among other things, we want to estimate the speed of COVID-19 transmission in these health areas (that is, correlation).

INLA

GF, Big n problem, Gaussian Markov Random Field (GMRF)

- Among other things, we want to estimate the speed of COVID-19 transmission in these health areas (that is, correlation).

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} & \rho_{16} & \rho_{17} & \rho_{18} & \rho_{19} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} & \rho_{25} & \rho_{26} & \rho_{27} & \rho_{28} & \rho_{29} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} & \rho_{35} & \rho_{36} & \rho_{37} & \rho_{38} & \rho_{39} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 & \rho_{45} & \rho_{46} & \rho_{47} & \rho_{48} & \rho_{49} \\ \rho_{15} & \rho_{25} & \rho_{35} & \rho_{45} & 1 & \rho_{56} & \rho_{57} & \rho_{58} & \rho_{59} \\ \rho_{16} & \rho_{26} & \rho_{36} & \rho_{46} & \rho_{56} & 1 & \rho_{67} & \rho_{68} & \rho_{69} \\ \rho_{17} & \rho_{27} & \rho_{37} & \rho_{47} & \rho_{75} & \rho_{67} & 1 & \rho_{78} & \rho_{79} \\ \rho_{18} & \rho_{28} & \rho_{38} & \rho_{48} & \rho_{58} & \rho_{68} & \rho_{78} & 1 & \rho_{89} \\ \rho_{19} & \rho_{29} & \rho_{39} & \rho_{49} & \rho_{59} & \rho_{69} & \rho_{79} & \rho_{89} & 1 \end{pmatrix}$$

- It is a 'dense' matrix, with 36 unknown parameters (could it be more than 36? What if it were 36x2=72?).

INLA

Gaussian Markov Random Field (GMRF)

- To resolve the Big n problem, the GMRF impose the assumption of **conditional independence** on the GF. For example, there is 'only' a direct correlation between the neighbours.

1	2	3
4	5	6
7	8	9

- In this case, we will estimate the correlations (1,2), (1,4), (1,5), (2,3), (2,5), (2,6), (3,5), (3,6), (4,5), (4,7), (4,8), (5,6), (5,7), (5,8), (5,9), (6,8), (6,9), (7,8) and (8,9).

- We have gone from 36 to 19 parameters.

INLA

Gaussian Markov Random Field (GMRF)

- In this case, we will estimate the correlations (1,2), (1,4), (1,5), (2,3), (2,5), (2,6), (3,5), (3,6), (4,5), (4,7), (4,8), (5,6), (5,7), (5,8), (5,9), (6,8), (6,9), (7,8) and (8,9).

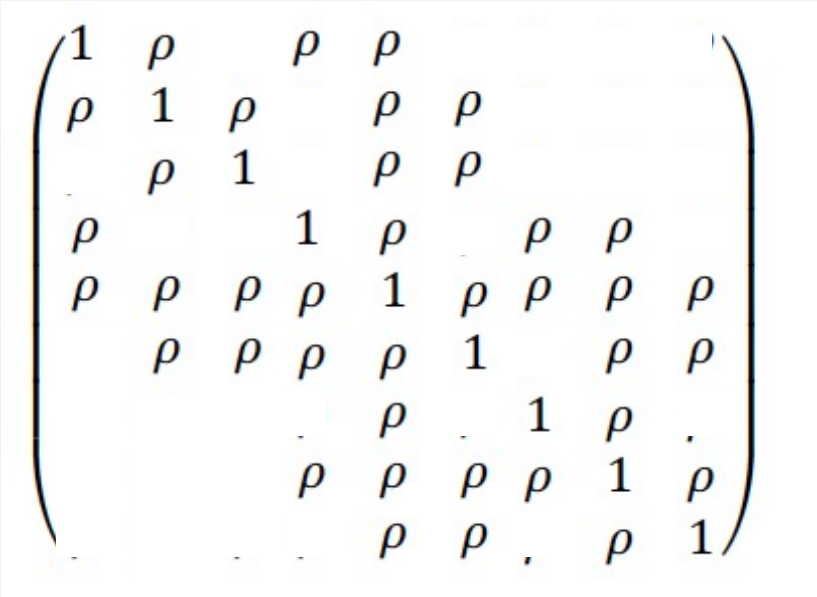
$$\begin{pmatrix} 1 & \rho_{12} & & \rho_{14} & \rho_{15} & & & & \\ \rho_{12} & 1 & \rho_{23} & & \rho_{25} & \rho_{26} & & & \\ & \rho_{23} & 1 & & \rho_{35} & \rho_{36} & & & \\ \rho_{14} & & & 1 & \rho_{45} & & \rho_{47} & \rho_{48} & \\ \rho_{15} & \rho_{25} & \rho_{35} & \rho_{45} & 1 & \rho_{56} & \rho_{57} & \rho_{58} & \rho_{59} \\ & \rho_{26} & \rho_{36} & \rho_{46} & \rho_{56} & 1 & & \rho_{68} & \rho_{69} \\ & & & & \rho_{75} & & 1 & \rho_{78} & \\ & & & \rho_{48} & \rho_{58} & \rho_{68} & \rho_{78} & 1 & \rho_{89} \\ & & & & \rho_{59} & \rho_{69} & & \rho_{89} & 1 \end{pmatrix}$$

- This is called a **sparse matrix**.

INLA

Gaussian Markov Random Field (GMRF)

- If, apart from only the 'adjacent neighbors' being correlated, the correlation is the same for all of them, the structure is known as **CAR (Conditional autoregressive).**



INLA

Gaussian Markov Random Field (GMRF)

GLMM

$$ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i}$$

$$\beta_{0i} = \beta_0 + \eta_i$$

$$Var(y_i|x_i) = \phi\mu_i(1-\mu_i)$$

- The model is a latent Gaussian model if all the parameters have a Gaussian joint distributions, that is $(\beta_0, \beta_1, \beta_2, \eta_i, \phi) \sim N(0, \Sigma)$.
- If we assume conditional independence of the observations of x_i, the latent Gaussian model will be a GMRF.

INLA

- The first "ingredient" of the INLA approach is the definition of conditional probability, which holds for any pair of variables (x, z) — and, technically, provided $p(z) > 0$

$$p(x \mid z) =: \frac{p(x, z)}{p(z)} \rightarrow p(x, z) = p(x \mid z)p(z)$$

$p(x \mid z)$ can be re-written as

$$p(z) = \frac{p(x, z)}{p(x \mid z)}$$

- In particular, a conditional version can be obtained further considering a third variable w as

$$p(z \mid w) = \frac{p(x, z \mid w)}{p(x \mid z, w)}$$

which is particularly relevant to the Bayesian case.

INLA

- The second "ingredient" is **Laplace approximation**.
- Main idea: approximate the integral

$$\int f(x)dx = \int \exp[\log f(x)]dx$$

by means of a Taylor's series expansion around the mode $x^* = \text{argmax}_x \log f(x)$:

$$\int f(x)dx \approx \int \exp\left[\log f(x^*) + \frac{(x-x^*)^2}{2} \left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*}\right]$$

- Setting $\sigma^{2*} = -1/\left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*}$ we can re-write

$$\int f(x)dx \approx f(x^*) \int \exp\left[-\frac{(x-x^*)^2}{2\sigma^{2*}}\right] dx$$

- Thus, under LA, $f(x) \approx \text{Normal}(x^*, \sigma^{2*})$.

INLA

Gaussian Markov Random Field (GMRF)

- We start from Bayesian hierarchical models specified in two stages.
- The **first stage** consists in the observational model $\pi(y|x)$, where y denotes the vector of observations and x are the unknown parameters, which follow a GMRF $\pi(x|\theta)$.

INLA

- The marginal a posteriori distributions of the GMRF,

$$\pi(x_i|y) = \int_\theta \pi(x_i|\theta, y)\pi(\theta|y)d\theta$$

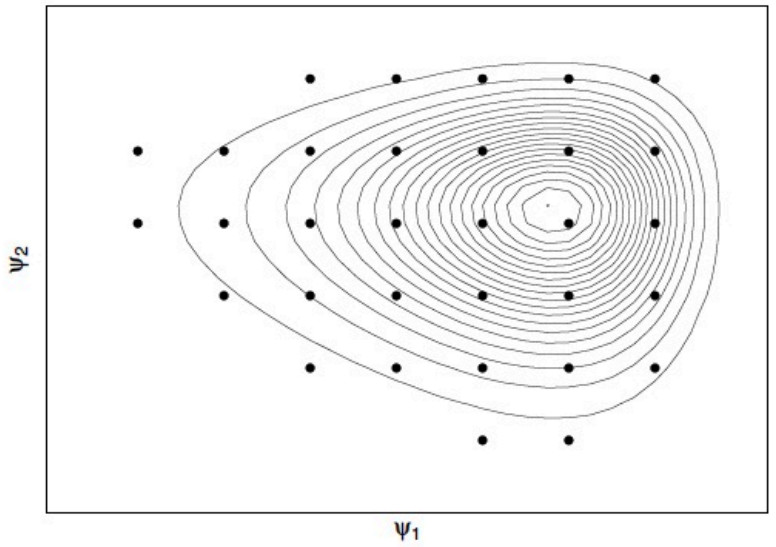
- They are estimated using the finite sum (evaluated at support points θ_k using appropriate weights Δ_k)

$$\pi(x_i|y) = \sum \pi(x_i|\theta_k, y)\, \pi(\theta_k|y)\, \Delta_k$$

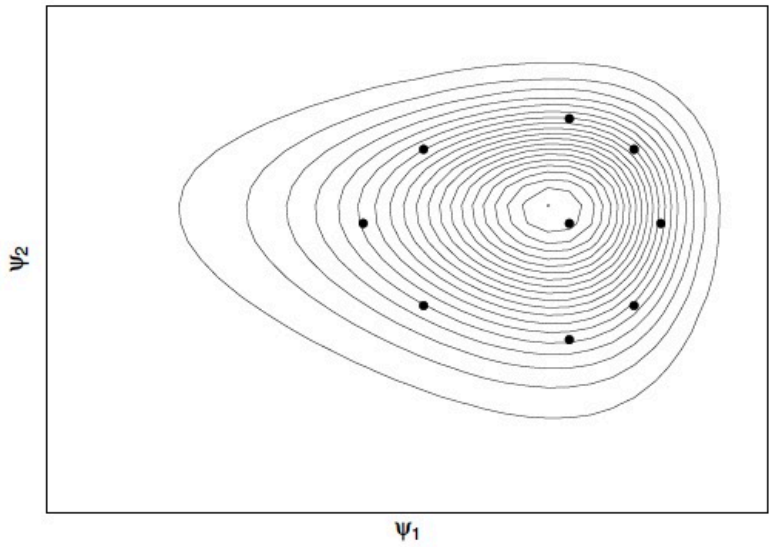
where $\tilde{\pi}(x_i|\theta_k, y)$ and $\tilde{\pi}(\theta_k|y)$ denotes approximations of $\pi(x_i|\theta_k, y)$ and $\pi(\theta_k|y)$, respectively.

INLA

Step 1. Explore the joint posterior for the hyperparameters $\tilde{p}(\boldsymbol{\psi} \mid \boldsymbol{y})$ and produce a grid of "good" integration points $\{\boldsymbol{\psi}^\star\}$ associated with the bulk of the mass, together with a corresponding set of area weights $\{\Delta^\star\}$:



Grid strategy



Central Composite Design strategy (CCD)

The CCD strategy is the default one in R-INLA: it produces a lower number of points which are however enough to capture the variability of the joint distribution (see [Martins et al., 2013]).

INLA

- The **second stage** is given by the hyperparameters θ and the (marginal) a priori distributions $\pi(\theta)$ (***priors***).
- The marginal a posteriori distribution of the hyperparameters, $\pi(\theta|y)$, is estimated using the Laplace approximation,

$$\tilde{\pi}(\theta|y) \propto \left(\frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x|\theta, y)} \middle| x \right) = x^*(\theta)$$

where the denominador $\tilde{\pi}_G(x|\theta, y)$ denotes the Gaussian approximation of $\pi(x, \theta, y)$ and $x^*(\theta)$ is the conditional mode.

INLA

Step 2. After the grid exploration, obtain the marginal posterior $\tilde{p}(\psi_k \mid \boldsymbol{y})$ using ar interpolation algorithm based on the values of the density $\tilde{p}(\boldsymbol{\psi} \mid \boldsymbol{y})$ evaluatec in the integration points $\{\boldsymbol{\psi}^\star\}$ (see Martins et al., 2013).

Step 3. For each integration point in $\boldsymbol{\psi}^\star$ and parameter θ_i, evaluate the approximatec marginal $\tilde{p}(\theta_i \mid \boldsymbol{\psi}^\star, \boldsymbol{y})$ for some selected values of θ_i.

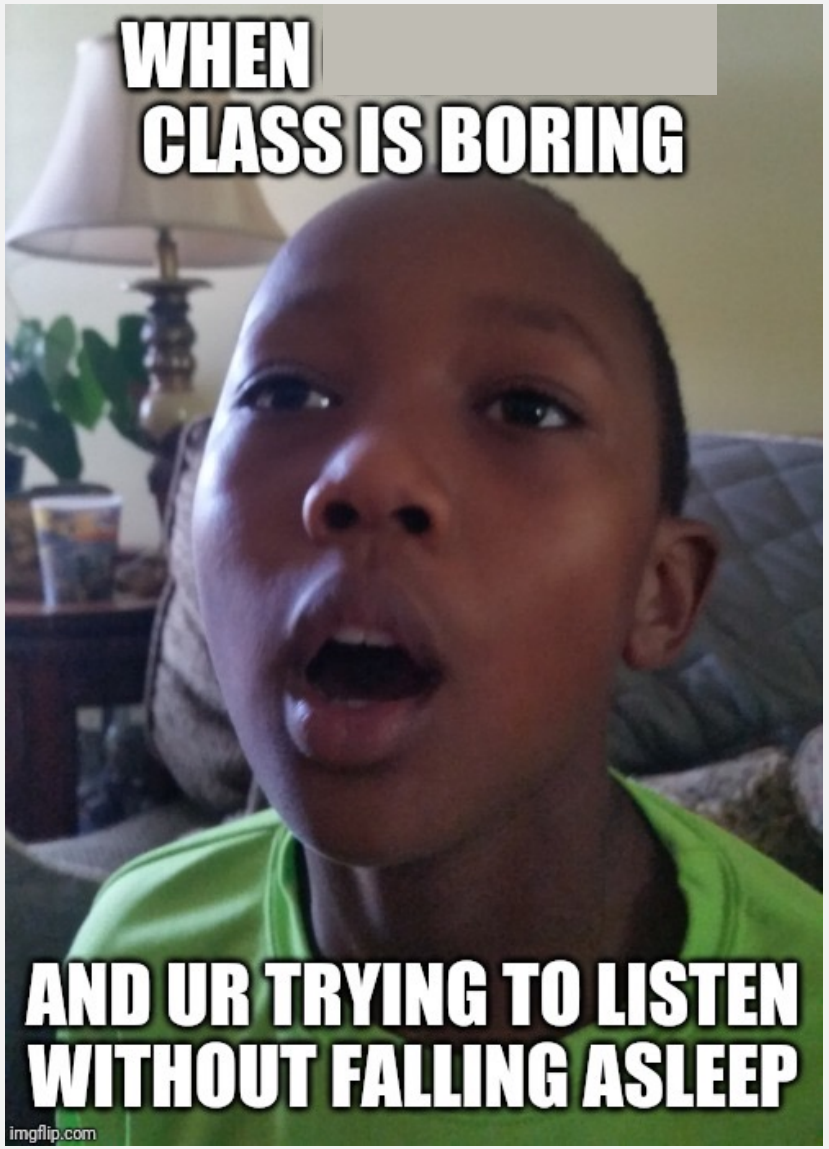
Step 4. For each i obtain the marginal posteriors $\tilde{p}(\theta_i \mid \boldsymbol{y})$ using **numerical inte-gration**[1]

$$\tilde{p}(\theta_i \mid \boldsymbol{y}) \approx \sum_{\boldsymbol{\psi}^\star} \tilde{p}(\theta_i \mid \boldsymbol{\psi}^\star, \boldsymbol{y}) \tilde{p}(\boldsymbol{\psi}^\star \mid \boldsymbol{y}) \, \Delta^\star$$

[1] Recall that $p(\theta_i \mid \boldsymbol{y}) = \int p(\theta_i, \boldsymbol{\psi} \mid \boldsymbol{y}) d\boldsymbol{\psi} = \int p(\theta_i \mid \boldsymbol{\psi}, \boldsymbol{y}) p(\boldsymbol{\psi} \mid \boldsymbol{y}) d\boldsymbol{\psi}$

INTRODUCTION TO INLA AND R INLA

1. Bayesian statistics
2. INLA
3. **R INLA**


WHEN
CLASS IS BORING
AND UR TRYING TO LISTEN
WITHOUT FALLING ASLEEP
imgflip.com