

□

# ANÀLISI DE DADES ESPACIALS EN L'ÀMBIT DE L'EPIDEMIOLOGIA

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

8, 10, 14 i 16 de setembre de 2021

Grup de Recerca en Estadística, Econometria i Salut (GRECS), Universitat de Girona  
CIBER d'Epidemiologia i Salut Pública(CIBERESP)

# INTRODUCCIÓ AL CURS

1. Introducció al curs
2. Introducció a l'epidemiologia i l'estadística espacial
3. Panoràmica del models mixtos
4. Panoràmica del models mixtos - Pràctiques
5. **Introducció a INLA i R INLA**
6. R INLA - Pràctiques

Dimecres 8

Divendres 10

# INTRODUCCIÓ AL CURS

- 7. Mapes de malalties. Estandardització de raons d'incidència i mortalitat
- 8. Mapes de malalties. Suavització de raons d'incidència i de mortalitat estandarditzades
- 9. Mapes de malalties – Pràctiques
- 10. Estudis d'associació geogràfica. Regressió ecològica espacial
- 11. Regressió ecològica espacial - Pràctiques

Dimarts 14

# INTRODUCCIÓ AL CURS

- 12. Agrupació de casos
- 13. Extensions: BYM2, processos puntuals, leaflet, pc priors
- 14. Extensions – Pràctiques

} Dijous 16

# INTRODUCCIÓ A INLA I R INLA

1. Estadística Bayesiana
2. INLA
3. R INLA

# INTRODUCCIÓ A INLA I R INLA

1. **Estadística Bayesiana**
2. INLA
3. R INLA

# ESTADÍSTICA BAYESIANA

- D'una manera molt esquemàtica es pot dir que per als **freqüentistes** (estadística clàssica), la probabilitat es considera com el límit de la freqüència relativa quan es realitza un experiment de manera repetida un nombre molt gran de vegades en condicions idèntiques.
- Per als **Bayesians** en canvi, la probabilitat és la mesura fonamental de la incertesa i aquest concepte subjectiu de probabilitat ha de construir-se amb judici científic.

# ESTADÍSTICA BAYESIANA

- En un model Bayesià, generalment volem les **distribucions a posteriori** per als nostres models (per exemple, la distribució dels paràmetres donades les dades), o **distribucions predictives a posteriori** (per a extrapolació/ predicció – la distribució de nous valors donats els observats).



# ESTADÍSTICA BAYESIANA

- La distribució a posteriori és igual a la probabilitat d'observar les dades multiplicada per la distribució a priori dels paràmetres (o **priors**), amb una constant de normalització (de manera que la integral a posteriori és igual a 1).

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

- De forma més simplificada (sense tenir en compte la constant de normalització)

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$$

# ESTADÍSTICA BAYESIANA

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

$\theta$  és el vector de **paràmetres**.

$p(y|\theta)$  és conegut com a **likelihood** (és el model).

$p(\theta)$  és la distribució a priori, o **priors**.

# ESTADÍSTICA BAYESIANA

- L'elecció de les priors a utilitzar en cada cas és una elecció subjectiva i que sovint ha de ser decidida basant-se el judici dels experts i del tipus de dades de què es disposi.
- Quan una distribució a posteriori és de la mateixa família que la distribució a priori utilitzada es parla de distribucions conjugades.
- L'avantatge del seu ús és que els "prior" tenen bones propietats matemàtiques per al càlcul de les distribucions a posteriori.

## Priors conjugats

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

# ESTADÍSTICA BAYESIANA

- En una **aproximació freqüentista (estimació)** sovint maximitzem la probabilitat de les dades (és a dir, el **likelihood**), utilitzant mètodes numèrics, com el de Newton-Raphson o d'altres, per obtenir una estimació puntual d'un paràmetre determinat (que veiem com a no aleatori – és a dir, fix - però desconegut).
- En una **aproximació Bayesiana (computing o inferència)** obtenim una distribució a posteriori per al paràmetre (que es considera com una variable aleatòria), per al qual podem proporcionar estadístics de resum (mitjana, mediana o moda) i quantils per obtenir directament intervals de credibilitat.

# ESTADÍSTICA BAYESIANA

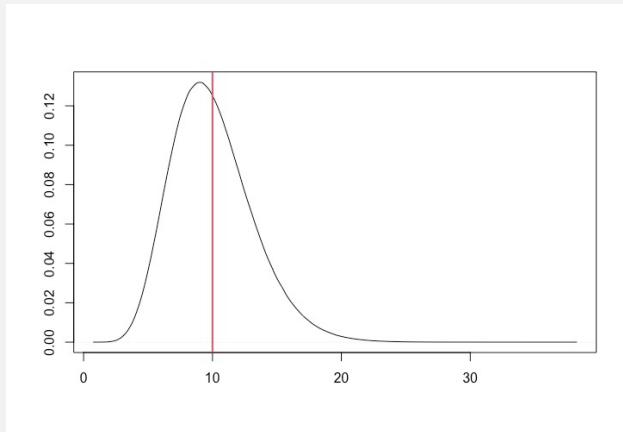
- El problema en l'aproximació Bayesiana, és que mentre la versemblança i la distribució a priori són fàcils d'obtenir,  $p(\theta|y)$  sol ser analíticament inabordable (especialment quan no s'utilitzen priors conjugades).

# BAYESIAN COMPUTING

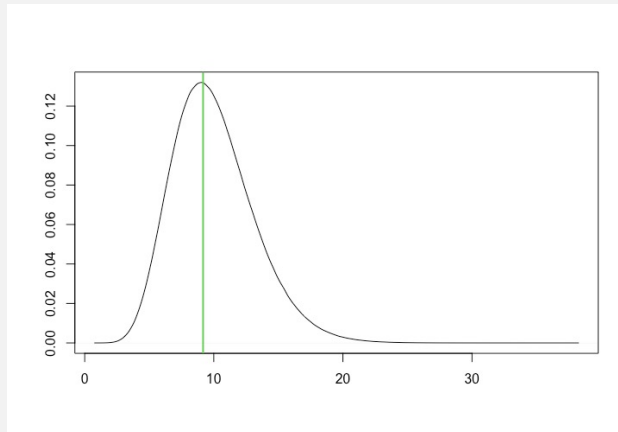
- Ens interessa obtenir la distribució (marginal) a posteriori,  $p(\theta|y)$ :

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

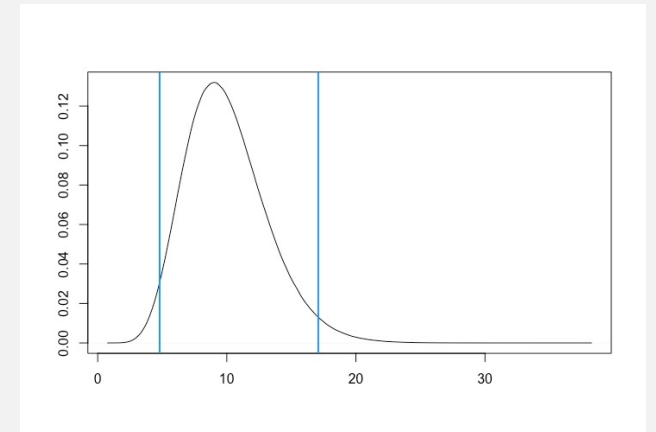
on  $\theta_{(-i)}$  denota el vector  $\theta$  excloent el component  $i$ .



Mitjana



Mediana



Interval de credibilitat al 95%

# BAYESIAN COMPUTING

- En general, les integrals són intractables i s'utilitzen mètodes numèrics com les **cadenaes Markovianes de Monte Carlo (MCMC)** per fer mostres de les distribucions condicionals i calcular la distribució marginal de cada paràmetre d'interès.
- Una seqüència de variables aleatòries  $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots$  forma una cadena de Markov si
$$\theta^{(i+1)} \longrightarrow p(\theta | \theta^{(i)})$$
- És a dir, condicionat al valor de  $\theta^{(i)}$ ,  $\theta^{(i+1)}$  és independent de  $\theta^{(i-1)}, \dots, \theta^{(0)}$

# BAYESIAN COMPUTING

- Existeixen varis algorismes per dissenyar cadenes de Markov.
- Entre ells, l'algorisme 'Gibbs sampling' és un dels més senzills de les MCMC.
- Però, també n'hi ha d'altres: Metropolis, Metropolis-Hastings, etc.



# BAYESIAN COMPUTING

## Gibbs sampling

➤ Sigui  $\theta$ , un vector de paràmetres desconegut  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$

1. S'escolleixen valors inicials  $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}$  per als components.

2. Es mostreja  $\theta_1^{(1)}$  a partir de  $p(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, x)$

Es mostreja  $\theta_2^{(1)}$  a partir de  $p(\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, x)$

Es mostreja  $\theta_k^{(1)}$  a partir de  $p(\theta_k | \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{k-1}^{(1)}, x)$

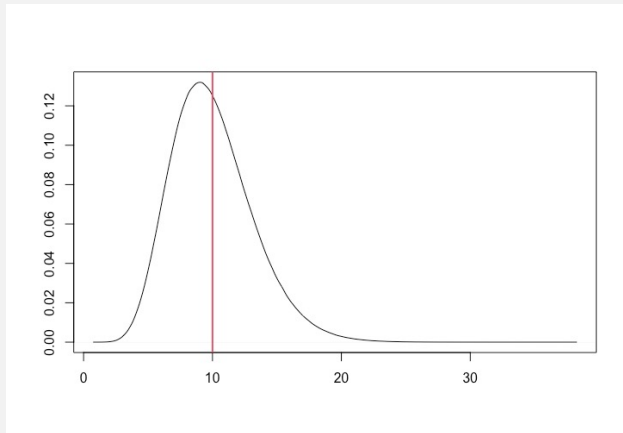
3. Es repeteix l'etapa 2 moltes vegades. Si el número de repeticions és molt gran s'obtindrà una mostra per a  $p(\theta | x)$ .

# BAYESIAN COMPUTING

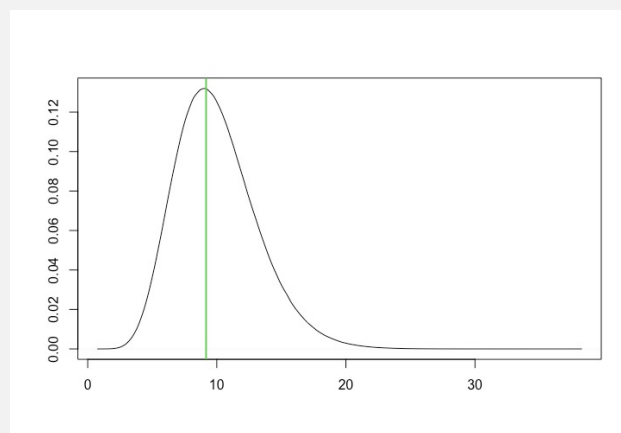
- Ens interessa obtenir la distribució (marginal) a posteriori,  $p(\theta|y)$ :

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

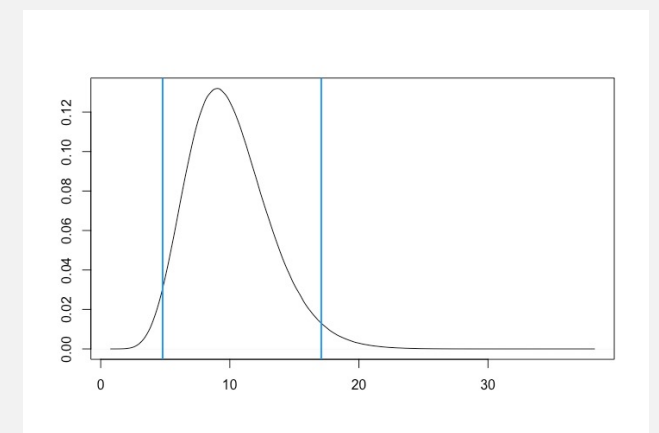
on  $\theta_{(-i)}$  denota el vector  $\theta$  excloent el component  $i$ .



Mitjana



Mediana



Interval de credibilitat al 95%

# BAYESIAN COMPUTING

- Les MCMC s'han desenvolupat en programari com WinBUGS.
- MCMC és lent, no *escala* bé (és a dir, els resultats no són invariants a canvis d'escala i/o mida de la mostra) i, per a alguns models complexos, pot fallar (el model no convergirà). El programari més recent (JAGS, Stan) ha intentat fer front a aquests reptes.

# BAYESIAN COMPUTING

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

# INTRODUCCIÓ A INLA I R INLA

1. Estadística Bayesiana
2. **INLA**
3. R INLA

## INLA

- MCMC és un mètode asimptòticament exacte mentre que l'INLA és una aproximació.
- Empíricament, l'error MCMC i l'error INLA solen ser molt similars, com s'ha demostrat en molts estudis de simulació.

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

Regressió lineal simple

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

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

Regressió de Poisson amb efectes aleatoris (no estructurats) en la constant

## 4. Introducció a INLA I R INLA

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

- Quan es tracta d'inferències bayesianes per a GMRF, és possible fer ús de l'INLA (enlloc de la 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

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

## 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  $\text{Cov}(y(s, t), y(s', t')) = \sigma^2 \mathcal{C}((s, t), (s', t'))$ , defined for each  $(s, t)$  and  $(s', t')$  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  $\mathbf{h} = (s - s') \in \mathbb{R}^2$  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](https://doi.org/10.1007/s10182-012-0196-3).



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

- Supposem que tenim dades d'incidència de la COVID-19 en 6 àrees de salut, amb la següent distribució geogràfica:

1	2	3
4	5	6
7	8	9

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

- Ens interessa, entre d'altres, estimar la velocitat de transmissió de la COVID-19 entre elles (és a dir la correlació).

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

- Ens interessa, entre d'altres, estimar la velocitat de transmissió de la COVID-19 entre elles (és a dir la correlació).

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

- És una matriu 'densa', amb 36 paràmetres desconeguts (podrien ser més de 36? Què passaria si fossin 36x2=72?).

## Gaussian Markov Random Field (GMRF)

- Per solucionar el Big n problem, els GMRF imposen, als GF el supòsit **d'independència condicional**. Per exemple, 'només' hi ha una correlació directa entre els veïns.

1	2	3
4	5	6
7	8	9

- En aquest cas, estimarem les correlacions (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) i (8,9).
- Hem passat de 36 a 19 paràmetres.

## Gaussian Markov Random Field (GMRF)

- En aquest cas, estimarem les correlacions (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) i (8,9).

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

- Es diu que és una matriu dispersa (**sparse**).

# INLA

## Gaussian Markov Random Field (GMRF)

- Si, a més de que només estan correlacionats els 'veïns contigus' la correlació és la mateixa per a tots, l'estructura es diu **CAR (Conditional autoregressive)**.

$$\begin{pmatrix} 1 & \rho & & \rho & \rho & & & & \\ \rho & 1 & \rho & & \rho & \rho & & & \\ & \rho & 1 & & \rho & \rho & & & \\ \rho & & & 1 & \rho & & \rho & \rho & \\ \rho & \rho & \rho & \rho & 1 & \rho & \rho & \rho & \rho \\ & \rho & \rho & \rho & \rho & 1 & & \rho & \rho \\ & & & & \rho & & 1 & \rho & . \\ & & & \rho & \rho & \rho & \rho & 1 & \rho \\ & & & & \rho & \rho & . & \rho & 1 \end{pmatrix}$$

## Gaussian Markov Random Field (GMRF)

GLMM

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

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

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

- El model és un model latent Gaussià si tots els paràmetres tenen una distribució conjunta Gaussiana, és a dir  $(\beta_0, \beta_1, \beta_2, \eta_i, \phi) \sim N(0, \Sigma)$ .
- Si suposem independència condicional de les observacions de  $x_i$ , el model latent Gaussià serà un 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  $p(z) > 0$

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

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

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

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

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

which is particularly relevant to the Bayesian case.



- 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  
 $x^* = \operatorname{argmax}_x \log f(x)$ :

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

- Setting  $\sigma^{2*} = -1 / \frac{\partial^2 \log f(x)}{\partial x^2} \Big|_{x=x^*}$  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,  $f(x) \approx \text{Normal}(x^*, \sigma^{2*})$ .

## Gaussian Markov Random Field (GMRF)

- Es parteix de models jeràrquics Bayesianes especificats en dues etapes.
- La **primera etapa** consisteix en el model observacional  $\pi(y|x)$ , on  $y$  denota el vector d'observacions i  $x$  son els paràmetres desconeguts, els quals segueixen un GMRF  $\pi(x|\theta)$ .

- Les distribucions marginals a posteriori del GMRF,

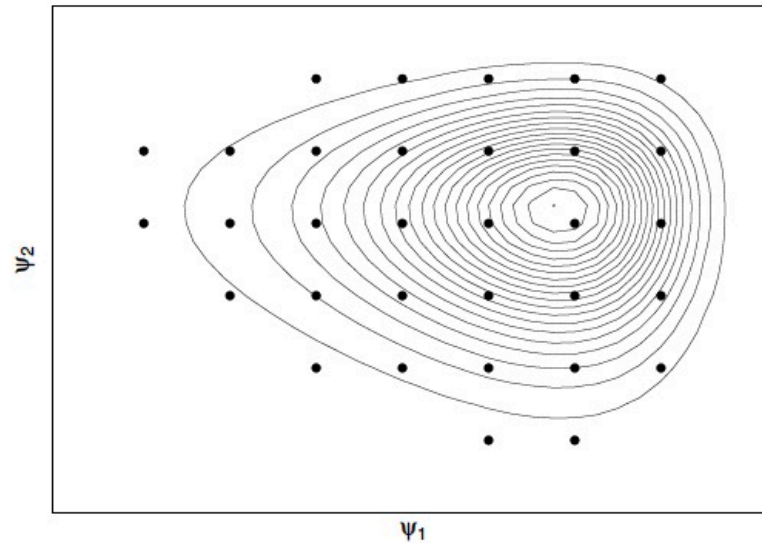
$$\pi(x_i|y) = \int_{\theta} \pi(x_i|\theta, y) \pi(\theta|y) d\theta$$

- S'aproximen utilitzant la suma finita (avaluat en punts de suport  $\theta_k$  utilitzant ponderacions apropiades  $\Delta_k$ )

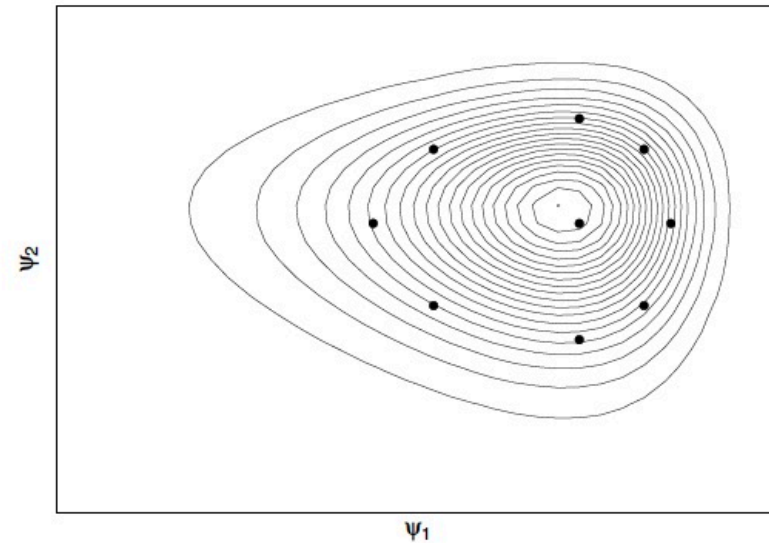
$$\pi(x_i|y) = \sum \pi(x_i|\theta_k, y) \pi(\theta_k|y) \Delta_k$$

On  $\tilde{\pi}(x_i|\theta_k, y)$  i  $\tilde{\pi}(\theta_k|y)$  denoten aproximacions de  $\pi(x_i|\theta_k, y)$  i  $\pi(\theta_k|y)$ , respectivament.

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



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

- La **segona etapa** ve donada model pels hiperparàmetres  $\theta$  i les distribucions (marginals) a priori  $\pi(\theta)$  (**priors**).
- La distribució marginal a posteriori dels hiperparàmetres,  $\pi(\theta|y)$ , s'aproxima utilitzant l'aproximació de Laplace,

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

on el denominador  $\tilde{\pi}_G(x|\theta, y)$  denota l'aproximació Gaussiana de  $\pi(x, \theta, y)$  i  $x^*(\theta)$  és la moda condicional.

**Step 2.** After the grid exploration, obtain the marginal posterior  $\tilde{p}(\psi_k | \mathbf{y})$  using an interpolation algorithm based on the values of the density  $\tilde{p}(\psi | \mathbf{y})$  evaluated in the integration points  $\{\psi^*\}$  (see Martins et al., 2013).

**Step 3.** For each integration point in  $\psi^*$  and parameter  $\theta_i$ , evaluate the approximate marginal  $\tilde{p}(\theta_i | \psi^*, \mathbf{y})$  for some selected values of  $\theta_i$ .

**Step 4.** For each  $i$  obtain the marginal posteriors  $\tilde{p}(\theta_i | \mathbf{y})$  using **numerical integration**<sup>1</sup>

$$\tilde{p}(\theta_i | \mathbf{y}) \approx \sum_{\psi^*} \tilde{p}(\theta_i | \psi^*, \mathbf{y}) \tilde{p}(\psi^* | \mathbf{y}) \Delta^*$$

<sup>1</sup>Recall that  $p(\theta_i | \mathbf{y}) = \int p(\theta_i, \psi | \mathbf{y}) d\psi = \int p(\theta_i | \psi, \mathbf{y}) p(\psi | \mathbf{y}) d\psi$

# INTRODUCCIÓ A INLA I R INLA

1. Estadística Bayesiana
2. INLA
- 3. R INLA**



