

□

# ANÁLISIS DE DATOS ESPACIALES EN EL ÁMBITO DE LA EPIDEMIOLOGÍA

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

8, 10, 14 y 16 de septiembre de 2021

Grupo de Investigación en Estadística, Econometría y Salud (GRECS), Universidad de Girona  
CIBER de Epidemiología y Salud Pública (CIBERESP)

# INTRODUCCIÓN AL CURSO

1. Introducción al curso
2. Introducción a la epidemiología y la estadística espacial
3. Panorámica de los modelos mixtos
4. Panorámica de los modelos mixtos - Prácticas
5. **Introducción a INLA y R INLA**
6. R INLA - Prácticas

Miércoles 8

Viernes 10

## INTRODUCCIÓN AL CURSO

- 7. Mapas de enfermedades. Estandarización de razones de incidencia y mortalidad
  - 8. Mapas de enfermedades. Suavización de razones de incidencia y de mortalidad estandarizadas
  - 9. Mapas de enfermedades – Prácticas
  - 10. Estudios de asociación geográfica. Regresión ecológica espacial
  - 11. Regresión ecológica espacial - Prácticas
- Martes 14

# INTRODUCCIÓN AL CURSO

- 12. Agrupación de casos
- 13. Extensiones: BYM2, procesos puntuales, leaflet, pc priors
- 14. Extensiones – Prácticas

} Jueves 16

# INTRODUCCIÓN A INLA Y R INLA

1. Estadística Bayesiana
2. INLA
3. R INLA

# INTRODUCCIÓN A INLA Y R INLA

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

# ESTADÍSTICA BAYESIANA

- De una manera muy esquemática puede decirse que para los **frecuentistas** (estadística clásica), la probabilidad se considera como el límite de la frecuencia relativa cuando se realiza un experimento de manera repetida un número muy grande de veces en condiciones idénticas.
- Para los **Bayesianos** en cambio, la probabilidad es la medición fundamental de la incertidumbre y este concepto subjetivo de probabilidad debe construirse con juicio científico.

# ESTADÍSTICA BAYESIANA

- En un modelo Bayesiano, generalmente queremos las **distribuciones a posteriori** para nuestros modelos (por ejemplo, la distribución de los parámetros dados los datos), o **distribuciones predictivas a posteriori** (para extrapolación/predicción – la distribución de nuevos valores dados los observados).



# ESTADÍSTICA BAYESIANA

- La distribución a posteriori es igual a la probabilidad de observar los datos multiplicada por la distribución a priori de los parámetros (o **priors**), con una constante de normalización (de manera que la integral a posteriori es 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 más simplificada (sin tener en cuenta la constante de normalización)

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

# ESTADÍSTICA BAYESIANA

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

$\theta$  es el vector de **parámetros**.

$p(y|\theta)$  se conoce como **likelihood** (es el modelo).

$p(\theta)$  es la distribución a priori, o **priors**.

# ESTADÍSTICA BAYESIANA

- La elección de los priors a utilizar en cada caso es una elección subjetiva y que a menudo debe ser decidida basándose en el juicio de los expertos y del tipo de datos de que se disponga.
- Cuando una distribución a posteriori es de la misma familia que la distribución a priori utilizada se habla de distribuciones conjugadas.
- La ventaja de su uso es que los "prior" tienen buenas propiedades matemáticas para el cálculo de las distribuciones a posteriori.

## Priors conjugados

| 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 **aproximación frecuentista (estimación)** a menudo maximizamos la probabilidad de los datos (es decir, el **likelihood**) utilizando métodos numéricos, como el de Newton-Raphson u otros, para obtener una estimación puntual de un parámetro determinado (que vemos como no aleatorio – es decir, fijo - pero desconocido).
- En una **aproximación Bayesiana (computing o inferencia)** obtenemos una distribución a posteriori para el parámetro (que se considera como una variable aleatoria), para el que podemos proporcionar estadísticos de resumen (media, mediana o moda) y cuantiles para obtener directamente intervalos de credibilidad.

# ESTADÍSTICA BAYESIANA

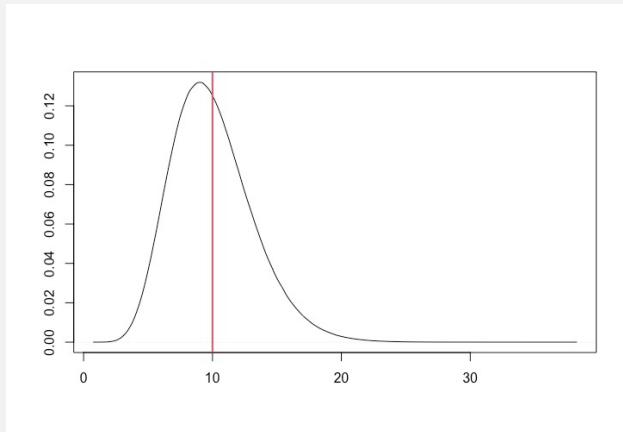
- El problema en la aproximación Bayesiana, es que mientras que la verosimilitud y la distribución a priori son fáciles de obtener,  $p(\theta|y)$  suele ser analíticamente inabordable (especialmente cuando no se utilizan priors conjugadas).

# BAYESIAN COMPUTING

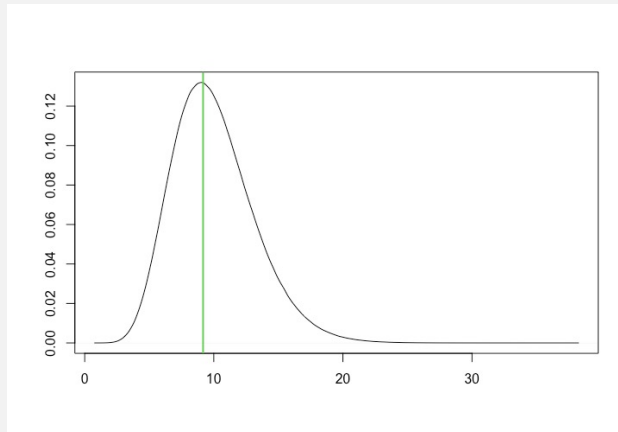
- Nos interesa obtener la distribución (marginal) a posteriori,  $p(\theta|y)$ :

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

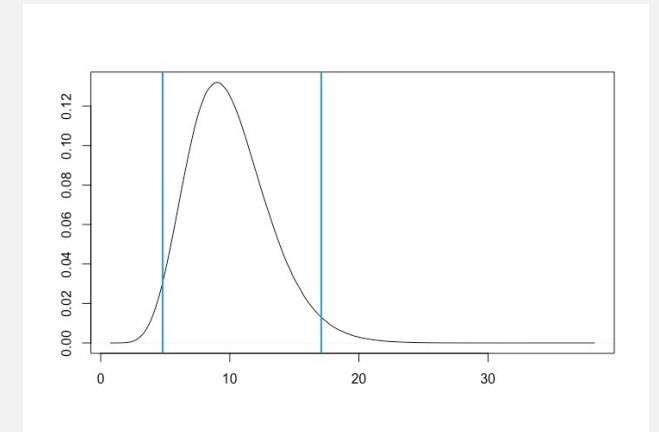
donde  $\theta_{(-i)}$  denota el vector  $\theta$  excluyendo la componente  $i$ .



Media



Mediana



Intervalo de credibilidad al 95%

# BAYESIAN COMPUTING

- En general, las integrales son intratables y se utilizan métodos numéricos como las **cadena Markovianas de Monte Carlo (MCMC)** para simular muestras de las distribuciones condicionales y calcular la distribución marginal de cada parámetro de interés.
- Una secuencia de variables aleatorias  $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots$  forma una cadena de Markov si  $\theta^{(i+1)} \longrightarrow p(\theta | \theta^{(i)})$
- Es decir, condicionado al valor de  $\theta^{(i)}$ ,  $\theta^{(i+1)}$  es independiente de  $\theta^{(i-1)}, \dots, \theta^{(0)}$

# BAYESIAN COMPUTING

- Existen varios algoritmos para diseñar cadenas de Markov.
- Entre ellos, el algoritmo 'Gibbs sampling' es uno de los más sencillos de las MCMC.
- Pero, también existen otros: Metropolis, Metropolis-Hastings, etc.



# BAYESIAN COMPUTING

## Gibbs sampling

- Sea  $\theta$ , un vector de parámetros desconocido  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$
1. Se escogen valores iniciales  $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}$  para los componentes.

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

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

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

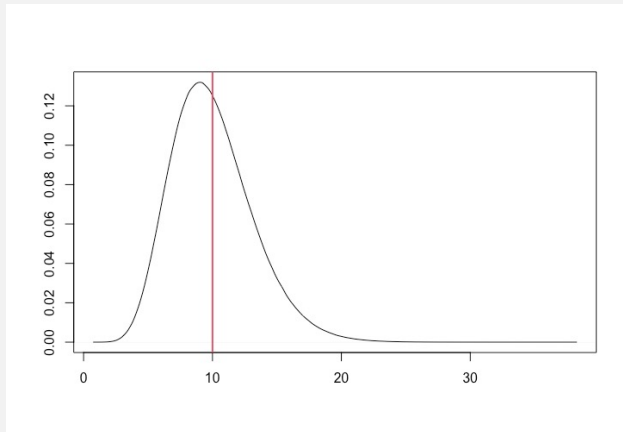
3. Se repite la etapa 2 muchas veces. Si el número de repeticiones es muy grande se obtendrá una muestra para  $p(\theta | x)$ .

# BAYESIAN COMPUTING

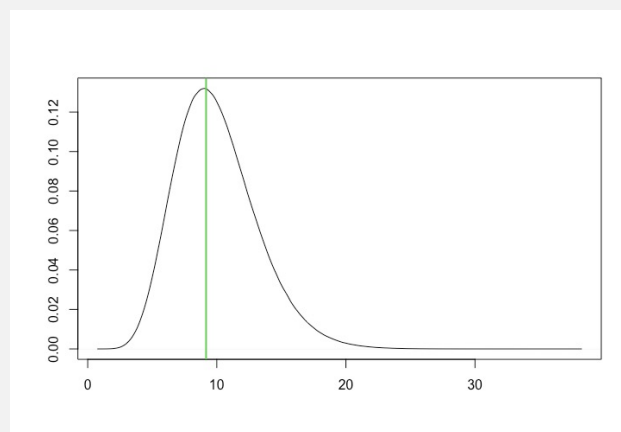
- Nos interesa obtener la distribución (marginal) a posteriori,  $p(\theta|y)$ :

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

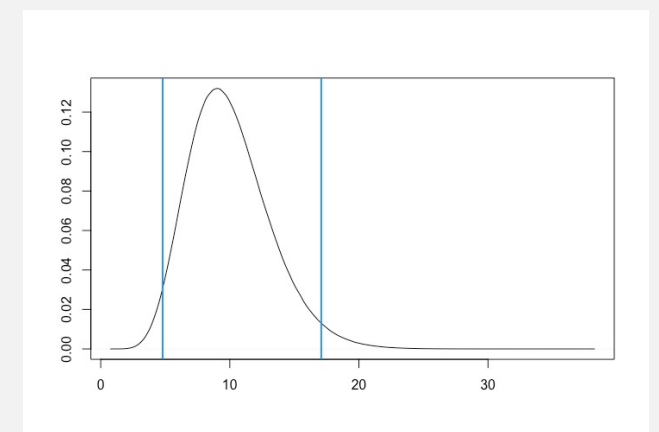
donde  $\theta_{(-i)}$  denota el vector  $\theta$  excluyendo la componente  $i$ .



Media



Mediana



Intervalo de credibilidad al 95%

# BAYESIAN COMPUTING

- Las MCMC se han desarrollado en programario como WinBUGS.
- MCMC es lento, no *escala* bien (es decir, los resultados no son invariantes a cambios de escala y/o tamaño de la muestra) y, para algunos modelos complejos, puede fallar (el modelo no convergirá). El programario más reciente (JAGS, Stan) ha intentado hacer frente a estos retos.

# BAYESIAN COMPUTING

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

# INTRODUCCIÓN A INLA Y R INLA

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

## INLA

- MCMC es un método asintóticamente exacto mientras que INLA es una aproximación.
- Empíricamente, el error MCMC y el error INLA suelen ser muy similares, como se ha demostrado en muchos estudios de simulación.

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 |

Regresión 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 |

Regresión de Poisson con efectos aleatorios (no estructurados) en la constante

## 5. Introducción a INLA y R INLA

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

- Cuando se trata de inferencias bayesianas para GMRF, es posible utilizar INLA (en lugar de 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)

- Supongamos que tenemos datos de incidencia de la COVID-19 en 6 áreas de salud, con la siguiente distribución geográfica:

|   |   |   |
|---|---|---|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |

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

- Nos interesa, entre otras cosas, estimar la velocidad de transmisión de la COVID-19 entre estas áreas de salud (es decir, la correlación).

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

- Nos interesa, entre otras cosas, estimar la velocidad de transmisión de la COVID-19 entre estas áreas de salud (es decir, la correlación).

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

- Es una matriz 'densa', con 36 parámetros desconocidos (¿podrían ser más de 36? ¿Qué pasaría si fuesen 36x2=72?).

## Gaussian Markov Random Field (GMRF)

- Para solucionar el Big n problem, los GMRF imponen a los GF el supuesto **de independencia condicional**. Por ejemplo, 'sólo' hay una correlación directa entre los vecinos.

|   |   |   |
|---|---|---|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |

- En este caso, estimaremos las correlaciones (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).
- Hemos pasado de 36 a 19 parámetros.

## Gaussian Markov Random Field (GMRF)

- En este caso, estimaremos las correlaciones (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).

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

- Se dice que es una matriz dispersa (**sparse**).

## Gaussian Markov Random Field (GMRF)

- Si, además de estar correlacionados solo los ‘vecinos contiguos’, la correlación es la misma para todos, la estructura se denomina **CAR (Conditional autoregressive)**.

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

## 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 modelo es un modelo latente Gaussiano si todos los parámetros tienen una distribución conjunta Gaussiana, es decir  $(\beta_0, \beta_1, \beta_2, \eta_i, \phi) \sim N(0, \Sigma)$ .
- Si suponemos independencia condicional de las observaciones de  $x_i$ , el modelo latente Gaussiano será 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)

- Se parte de modelos jerárquicos Bayesianos especificados en dos etapas.
- La **primera etapa** consiste en el modelo observacional  $\pi(y|x)$ , donde  $y$  denota el vector de observaciones y  $x$  son los parámetros desconocidos, los cuáles siguen un GMRF  $\pi(x|\theta)$ .

- Las distribuciones marginales a posteriori del GMRF,

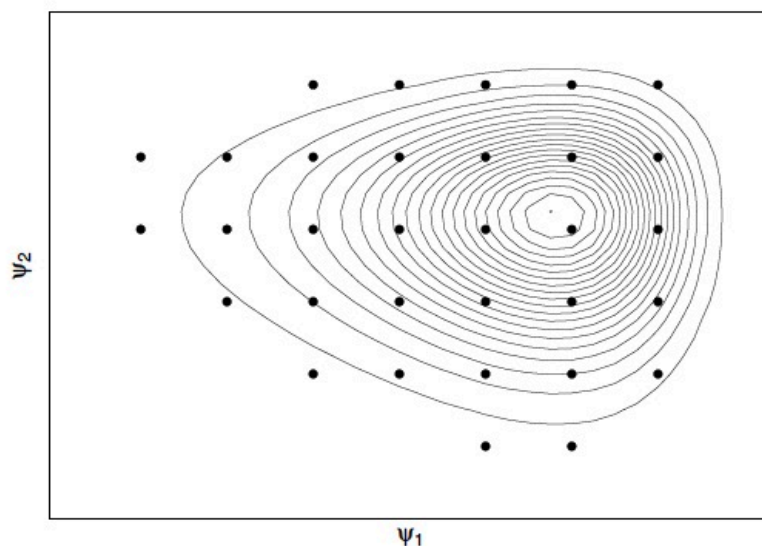
$$\pi(x_i|y) = \int_{\theta} \pi(x_i|\theta, y) \pi(\theta|y) d\theta$$

- Se aproximan utilizando la suma finita (evaluado en puntos de soporte  $\theta_k$  utilizando ponderaciones apropiadas  $\Delta_k$ )

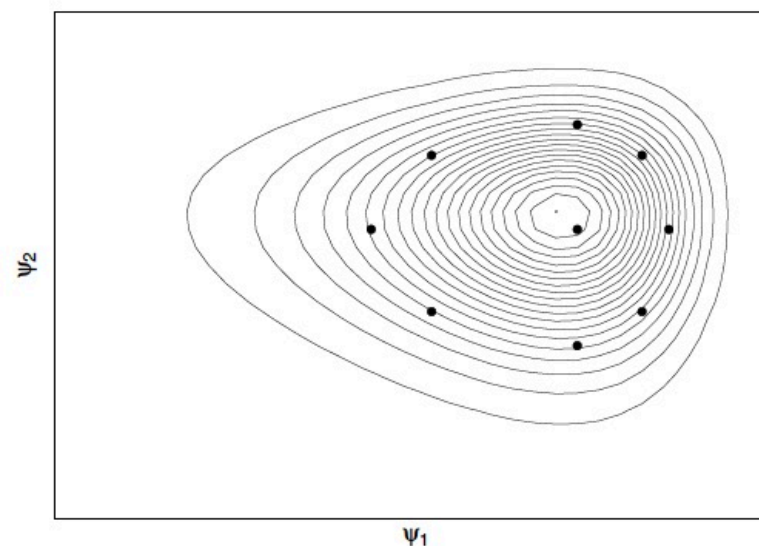
$$\pi(x_i|y) = \sum \pi(x_i|\theta_k, y) \pi(\theta_k|y) \Delta_k$$

donde  $\tilde{\pi}(x_i|\theta_k, y)$  y  $\tilde{\pi}(\theta_k|y)$  denotan aproximaciones de  $\pi(x_i|\theta_k, y)$  y  $\pi(\theta_k|y)$ , respectivamente.

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 **segunda etapa** viene dada por los hiperparámetros  $\theta$  y las distribuciones (marginales) a priori  $\pi(\theta)$  (**priors**).
- La distribución marginal a posteriori de los hiperparámetros,  $\pi(\theta|y)$ , se aproxima utilizando la aproximación de Laplace,

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

donde el denominador  $\tilde{\pi}_G(x|\theta, y)$  denota la aproximación Gaussiana de  $\pi(x, \theta, y)$  y  $x^*(\theta)$  es 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  $p(\theta_i | \mathbf{y}) = \int p(\theta_i, \psi | \mathbf{y}) d\psi = \int p(\theta_i | \psi, \mathbf{y}) p(\psi | \mathbf{y}) d\psi$

# INTRODUCCIÓN A INLA Y R INLA

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



