Cap 3: Bayesian Inference and INLA

Estadística e Inferencia Bayesiana

Estadística clásica – Probabilidad frecuentista

Estadística bayesiana – Distribuciones de probabilidad para determinar la incertidumbre de un modelo.

Se utiliza el **Teorema de Bayes.**

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y},\boldsymbol{\theta})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Distribución posterior de θ (parámetros de la distribución)

Para la **inferencia bayesiana** se requiere especificar una **prior** ("creencia inicial") previamente, que se representa como una distribución de probabilidad.

Se calcula una **verosimilitud**, que se basa en la distribución de probabilidad de los datos condicionada a los valores de los parámetros, y describe la probabilidad de observar los datos que se conocen dados los parámetros de los modelos.

Algunos métodos para calcular distribuciones posteriores son demandantes de muchos **recursos computacionales (ej. MCMC)**

Inferencia bayesiana con MCMC

Métodos de Monte Carlo de cadenas de Markov. Generan una muestra de valores de una cadena de Markov convergente: cuya distribución estacionaria es la distribución posterior ($\pi(\boldsymbol{\theta}|\boldsymbol{y})$)

$$\{\boldsymbol{\theta}^{(g)},\ g=1,\ldots,G\}$$

usar la media muestral para estimar la media posterior

$$\widehat{E(\theta_i|\boldsymbol{y})}=\frac{1}{G}\sum_{i=1}^{G}\theta_i^{(g)},$$

usar la varianza muestral para estimar la varianza

$$\widehat{Var(\theta_i|\boldsymbol{y})}=\frac{1}{G-1}\sum_{i=1}^{G}(\theta_i^{(g)}-\widehat{E(\theta_i|\boldsymbol{y})})^2.$$

INLA - Integrated nested Laplace approximation

(Aproximación Integrada de Laplace)

Se utiliza para realizar inferencias en modelos bayesianos, especialmente modelos jerárquicos, espaciales y espacio-temporales **(Modelos Gaussianos)** debido a su eficiencia computacional.

Utiliza una combinación de aproximaciones analíticas e integración numérica para obtener aproximaciones de distribuciones posteriores de los parámetros, siendo mucho más rápida que los modelos de Cadenas de Markov con Mote Carlo.

Para utilizarlo en R se utiliza el paquete "R-INLA", pero previo a su instalación se debe instalar la última versión de R (https://cran.r-project.org/bin/windows/base/). El paquete "R-INLA" se descarga desde web, mediante el siguiente comando:

```
install.packages("INLA", repos = "https://inla.r-inla-download.org/R/stable", dep = TRUE)
```

INLA – Modelos Gaussianos latentes

Los datos u observaciones pertenecen a la familia de las exponenciales (e.g. Gaussiana, Poisson, Binomial, …).

$$y_i|\boldsymbol{x}, \boldsymbol{\theta} \sim \pi(y_i|x_i, \boldsymbol{\theta}) \; i = 1, \ldots, n$$

$$\mu_i = g^{-1}(\eta_i)$$ **(media de la distribución)**

Se tiene el campo Gaussiano latente, donde los parámetros (x) siguen una distribución Gaussiana

$$\boldsymbol{x}|\boldsymbol{\theta} \sim N(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{Q}(\boldsymbol{\theta})^{-1})$$

Además, dicha distribución depende de los hiperparámetros (ϴ) (no necesariamente Gaussiana)

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$$

Se tiene un predictor línea que da cuenta de los efectos de las covariables.

$$\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}(u_{ji})$$

Intercepto

Efecto de las covariables

Efectos aleatorios definidos en términos de algunas covariables

$$\pi(y_i|x_i, \boldsymbol{\theta})$$

Distribución de probabilidad de los datos dado el vector de parámetros (Función de verosimilitud)

INLA – Modelos Gaussianos latentes

Se calculan los marginales posteriores del campo Gaussiano latente y de los hiperparámetros

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},\ \pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{\ j}$$

Se utiliza esta forma para construir las aproximaciones anidadas

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \int \tilde{\pi}(x_i|\boldsymbol{\theta}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},\ \tilde{\pi}(\theta_j|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}$$

Esta aproximación puede integrarse numéricamente con respecto a los hiperparámetros

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta_k}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta_k}|\boldsymbol{y}) \times \Delta_k$$

$$\tilde{\pi}(\theta_j|\boldsymbol{y}) = \sum_l \tilde{\pi}(\boldsymbol{\theta_l^*}|\boldsymbol{y}) \times \Delta_l^*$$

Cap 4: The R-INLA package

R-INLA

0. Identificar las distribuciones y el predictor lineal

$Y_i|\eta_i, \sigma^2 \sim N(\eta_i, \sigma^2), i = 1, \ldots, n$

$\eta_i = \beta_0 + \beta_1 \times x_{1i} + \beta_2 \times x_{2i} + u_i,\ u_i \sim N(0, \sigma_u^2)$

1. Escribir el predictor lineal como una fórmula en R

```
formula <- y ~ x1 + x2 + f(id, model = "iid")
```

iid: nombre del modelo

x1, x2 covariables

id: Vector indicador que especifica el elemento de los efectos aleatorios

2. Llamar a la función inla()

```
res <- inla(formula, family = "gaussian", data = d)
```

names(inla.models()$likelihood): lista de posibles alternativas de familias
inla.doc("nombredefamilia"): detalles sobre familias individuales
names(inla.models()$prior): lista con las opciones de cada una de las priors
inla.doc("nombredeprior"): documentación sobre una prior específica

Los priors se pueden especificar o dejar en default (que sería una distribución gaussiana de promedio 0 y precisión 0).
prior.fixed <- list(mean.intercept = <>, prec.intercept = <>, mean = <>, prec = <>)
res <- inla(formula, data = d, control.fixed = prior.fixed)

R-INLA - Ejemplo de ajuste de un Modelo

Ejemplo: teniendo como información el número de operaciones y muertes en 12 hospitales, el objetivo es evaluar el desempeño de cada hospital con respecto a las tasas de mortalidad quirúrgica.

1. El modelo asume una probabilidad binomial para el número de muertes en cada hospital con una tasa de mortalidad p_i

$$Y_i \sim Binomial(n_i, p_i),\ i = 1, \ldots, 12.$$

2. El modelo también supone que las tasas de mortalidad entre hospitales son similares y especifica un modelo de efectos aleatorios para las verdaderas tasas de mortalidad p_i

$$logit(p_i) = \alpha + u_i,\ u_i \sim N(0, \sigma^2).$$

3. Se especifica una prior no informativa para α con una tasa de mortalidad logística poblacional

$$\alpha \sim N(0, 1/\tau),\ \tau = 0.$$

```
Surg
```

```
##      n  r hospital
## 1   47  0        A
## 2  148 18        B
## 3  119  8        C
## 4  810 46        D
## 5  211  8        E
## 6  196 13        F
## 7  148  9        G
## 8  215 31        H
## 9  207 14        I
## 10  97  8        J
## 11 256 29        K
## 12 360 24        L
```

R-INLA - Ejemplo de ajuste de un Modelo

Ejemplo: teniendo como información el número de operaciones y muertes en 12 hospitales, el objetivo es evaluar el desempeño de cada hospital con respecto a las tasas de mortalidad quirúrgica.

4. El prior predeterminado de la precisión de los efectos aleatorios u_i es $1/\sigma^2 \sim Gamma(1, 5 \times 10^{-5})$
5. Este se puede cambiar configurando un Prior de Complejidad Penalizada en la desviación estándar σ. Por ejemplo, podemos especificar que la probabilidad de que σ sea mayor que 1 es menor o igual a 0.01. Este prior se especifica como

```
prior.prec <- list(prec = list(prior = "pc.prec", param = c(1, 0.01)))
```

6. El modelo sería escrito en R como

```
formula <- r ~ f(hospital, model = "iid", hyper = prior.prec)
```

7. Luego, llamamos a inla()

```
res <- inla(formula, data = Surg,  family = "binomial", Ntrials = n,
control.predictor = list(compute = TRUE), control.compute = list(dic = TRUE))
```

```
Surg
```

```
##      n  r hospital
## 1   47  0        A
## 2  148 18        B
## 3  119  8        C
## 4  810 46        D
## 5  211  8        E
## 6  196 13        F
## 7  148  9        G
## 8  215 31        H
## 9  207 14        I
## 10  97  8        J
## 11 256 29        K
## 12 360 24        L
```

R-INLA – Visualización de resultados

Ejemplo: teniendo como información el número de operaciones y muertes en 12 hospitales, el objetivo es evaluar el desempeño de cada hospital con respecto a las tasas de mortalidad quirúrgica.

Para obtener resumen del objeto, el DIC del modelo, resúmenes de objetos fijos, efectos aleatorios e hiperparámetros se obtienen, respectivamente, con:

summary(res)

resdicdic

res$summary.fixed

res$summary.random

res$summary.hyperpar

DIC=Devianza del modelo + 2×Desviación estándar de la devianza posterior

```
## Fixed effects:
##                mean   sd 0.025quant 0.5quant 0.975quant
## (Intercept) -2.545 0.14     -2.838   -2.539     -2.281
##               mode kld
## (Intercept) -2.53   0
##
## Random effects:
## Name   Model
##  hospital   IID model
##
## Model hyperparameters:
##                           mean    sd 0.025quant 0.5quant
## Precision for hospital 12.04 18.30      2.366    8.292
##                          0.975quant  mode
## Precision for hospital      41.86 5.337
##
## Expected number of effective parameters(std dev): 7.257(1.703)
## Number of equivalent replicates : 1.654
##
## Deviance Information Criterion (DIC) ...............: 74.93
## Deviance Information Criterion (DIC, saturated) ....: 25.16
## Effective number of parameters .....................: 8.173
##
## Marginal log-Likelihood:  -41.16
```

R-INLA – Visualización de resultados

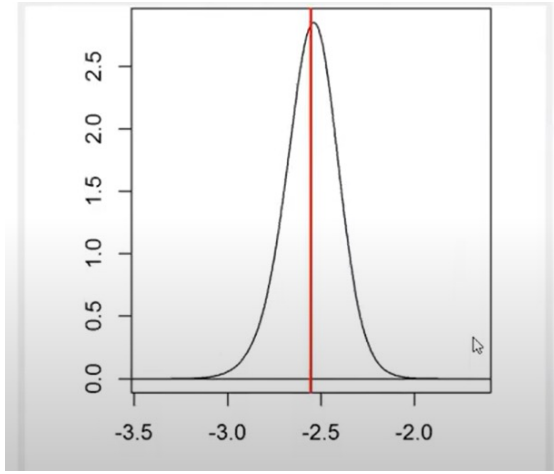
```r
res$summary.fitted.values
```

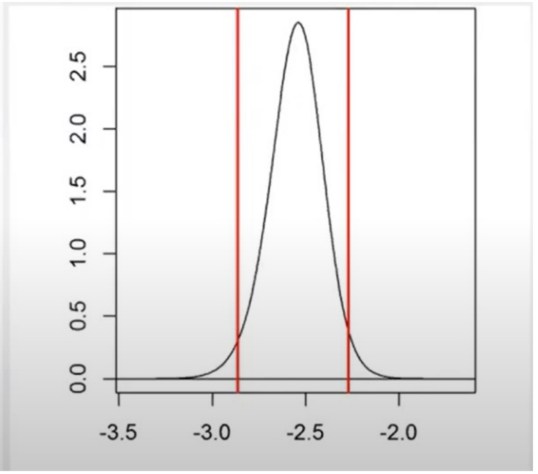
```
##                      mean      sd 0.025quant
## fitted.Predictor.01 0.05667 0.01872    0.02284
## fitted.Predictor.02 0.10224 0.02132    0.06685
## fitted.Predictor.03 0.07220 0.01695    0.04229
## fitted.Predictor.04 0.06011 0.00787    0.04540
## fitted.Predictor.05 0.05410 0.01298    0.03041
## fitted.Predictor.06 0.07057 0.01438    0.04465
## fitted.Predictor.07 0.06838 0.01545    0.04061
## fitted.Predictor.08 0.12140 0.02205    0.08256
## fitted.Predictor.09 0.07123 0.01420    0.04563
## fitted.Predictor.10 0.07941 0.01918    0.04674
## fitted.Predictor.11 0.10160 0.01723    0.07156
## fitted.Predictor.12 0.06951 0.01152    0.04836
##                     0.5quant 0.975quant    mode
## fitted.Predictor.01  0.05595    0.09579 0.05534
## fitted.Predictor.02  0.10007    0.14968 0.09535
## fitted.Predictor.03  0.07103    0.10923 0.06920
## fitted.Predictor.04  0.05986    0.07621 0.05936
## fitted.Predictor.05  0.05357    0.08088 0.05244
## fitted.Predictor.06  0.06978    0.10128 0.06844
## fitted.Predictor.07  0.06752    0.10147 0.06620
## fitted.Predictor.08  0.12004    0.16822 0.11740
## fitted.Predictor.09  0.07044    0.10153 0.06909
## fitted.Predictor.10  0.07759    0.12256 0.07448
## fitted.Predictor.11  0.10036    0.13865 0.09777
## fitted.Predictor.12  0.06901    0.09360 0.06811
```

Se obtienen las tasas de mortalidad predichas para cada hospital con **res$summary.fitted.values**.

La columna **“mean"** muestra que los hospitales 2, 8 y 11 son los que tienen las medias posteriores más altas de las tasas de mortalidad.

Las columnas "Cuantil 2.5%" y "Cuantil 97.5%" contienen los límites inferior y superior de los intervalos creíbles del 95% de las tasas de mortalidad y proporcionan medidas de incertidumbre.





R-INLA – Transformación marginal

Si deseamos obtener una transformación de la marginal, podemos usar **inla.tmarginal().** Por ejemplo, si queremos obtener la varianza del efecto aleatorio u_i, podemos obtener la marginal de la precisión τ y luego aplicar la función inversa.

```
marg.variance <- inla.tmarginal(function(x) 1/x, res$marginals.hyperpar$"Precision for hospital")
```

También podemos utilizar **inla.zmarginal()** para obtener estadísticas resumidas de la marginal.

```
inla.zmarginal(marg.variance)
```

```
## Mean                0.146459
## Stdev               0.106093
## Quantile  0.025 0.0236223
## Quantile  0.25   0.075129
## Quantile  0.5    0.120269
## Quantile  0.75   0.1868
## Quantile  0.975 0.421436
```

R-INLA – Tasas de mortalidad posteriores

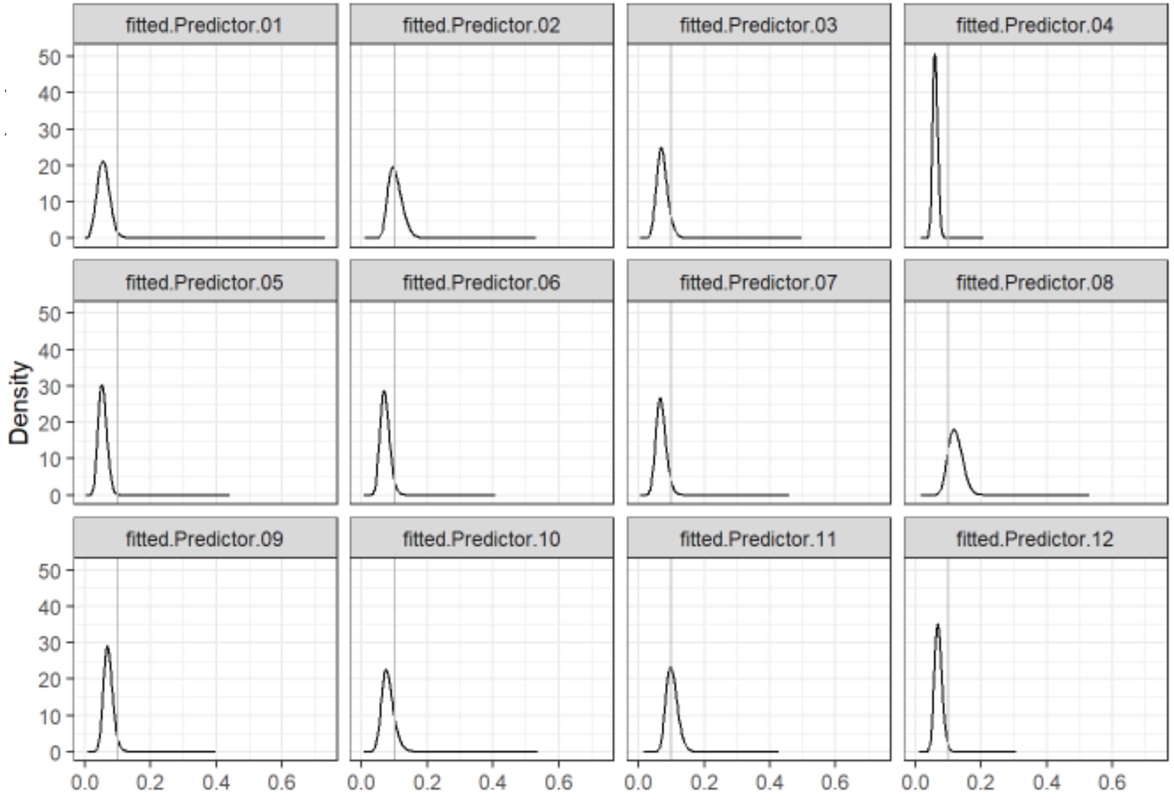
En este ejemplo, deseamos evaluar el rendimiento de los hospitales examinando las tasas de mortalidad. **res$marginals.fitted.values** es una lista que contiene las tasas de mortalidad posteriores de cada uno de los hospitales.

Podemos representar gráficamente estas posteriores construyendo un marco de datos **(marginals)** a partir de la lista **res$marginals.fitted.values** y agregando una columna hospital que denote el hospital.

```
list_marginals <- res$marginals.fitted.values

marginals <- data.frame(do.call(rbind, list_marginals)

marginals$hospital <- rep(names(list_marginals),
          times = sapply(list_marginals, nrow))
```



R-INLA – Probabilidad de excedencia

También podemos calcular las probabilidades de que las tasas de mortalidad sean **mayores que un valor umbral dado.** Estas probabilidades se conocen **como probabilidades de excedencia** y se expresan como $P(p_i>c)$. Se puede calcular con la función **inla.pmarginal()**, pasando como argumentos la distribución marginal de p_1 y el valor umbral c:

```
marg <- res$marginals.fitted.values[[1]]
1 - inla.pmarginal(q = 0.1, marginal = marg)
## [1] 0.01644
```

Podemos calcular las probabilidades de que las tasas de mortalidad sean mayores que 0.1 para todos los hospitales usando la función **sapply(),** pasando como argumentos la lista con todas las marginales (**res$marginals.fitted.values**) y la función para calcular las probabilidades de **excedencia (1 - inla.pmarginal()).** **sapply()** devuelve un vector del mismo largo que la lista **res$marginals.fitted.values**, con valores iguales al resultado de aplicar la **función 1 - inla.pmarginal()** a cada uno de los elementos de la lista de marginales.

```
## fitted.Predictor.01 fitted.Predictor.02
##           1.644e-02           5.001e-01
## fitted.Predictor.03 fitted.Predictor.04
##           5.925e-02           4.353e-06
## fitted.Predictor.05 fitted.Predictor.06
##           7.952e-04           2.901e-02
## fitted.Predictor.07 fitted.Predictor.08
##           2.919e-02           8.301e-01
## fitted.Predictor.09 fitted.Predictor.10
##           2.999e-02           1.357e-01
## fitted.Predictor.11 fitted.Predictor.12
##           5.070e-01           7.982e-03
```

Ejemplos en epidemiología

- Shaddick, Thomas y Green (2018) producen estimaciones globales de la contaminación atmosférica por partículas finas. https://academic.oup.com/jrsssc/article/67/1/231/7058389?login=false#396910509

- Moraga et al. (2015) predicen la prevalencia de filariasis linfática en África subsahariana

https://parasitesandvectors.biomedcentral.com/articles/10.1186/s13071-015-1166-x

- Osgood-Zimmerman et al. (2018) mapean el fallo en el crecimiento infantil en África.

https://www.nature.com/articles/nature25760

Video: https://www.youtube.com/watch?v=Tdb5EPczE9E&t=4082s

Paquete R-INLA – Especificación de priors

Se puede ver escribiendo names(inla.models()$prior), y una lista con las opciones de cada una de las priors se puede ver con inla.models()$prior. La documentación sobre una prior específica se puede ver con inla.doc("nombredeprior"). Los valores de estas priors se pueden cambiar en el argumento control.fixed de inla() al asignar una lista con la media y la precisión de las distribuciones gaussianas.

```
prior.fixed <- list(mean.intercept = <>, prec.intercept = <>, mean = <>, prec = <>)
res <- inla(formula, data = d, control.fixed = prior.fixed)
prior.prec <- list(initial = <>, prior = <>, param = <>, fixed = <>)
prior <- list(prec = prior.prec)
```

hyper acepta una lista con nombres iguales a cada uno de los hiperparámetros, y valores iguales a una lista con la especificación de las priors. Específicamente, la lista contiene los siguientes valores:

initial: valor inicial del hiperparámetro (buenos valores iniciales pueden acelerar el proceso de inferencia).

prior: nombre de la distribución prior (por ejemplo, "iid", "bym2").

param: vector con los valores de los parámetros de la distribución prior.

fixed: variable booleana que indica si el hiperparámetro es un valor fijo.

INLA-package

```
install.packages("INLA",
repos = "https://inla.r-inla-download.org/R/stable",
dep = TRUE)

library(INLA)
```

Model

Observaciones siguen distribución normal

$$Y_i|\eta_i, \sigma^2 \sim N(\eta_i, \sigma^2), i = 1, \ldots, n$$

$$\eta_i = \beta_0 + \beta_1 \times x_{1i} + \beta_2 \times x_{2i} + u_i, \; u_i \sim N(0, \sigma_u^2)$$

covariables **efectos aleatorios**

1 Write the linear predictor as a formula object in R

resultado **cov2** **Modelo es similar a la regresión lineal lm()**

```
formula <- y ~ x1 + x2 + f(id, model = "iid")
```

cov1 **Índices y modelo de los efectos aleatorios**

Random effects specified with `f()`. First argument `id` is an index vector that specifies the element of the random effect that applies to each observation, second argument model name

2 Call the function `inla()`

```
res <- inla(formula, family = "gaussian", data = d)
```

Paquete R-INLA – Ejemplo de especificación y ajuste de un Modelo usando R-INLA

iid, es uno de los modelos a probar. Más info con **inla.doc("iid").** La fórmula del modelo se guardará como **formula.**

Con el comando **res** se guardarán los resúmenes y densidades marginales posteriores, usando el comando inla para especificar los datos, familia de la distribución de probabildiad y numero de intentos.

formula <- r ~ f(hospital, model = "iid", hyper = prior.prec)

res <- inla(formula, data = Surg, family = "binomial", Ntrials = n, control.predictor = list(compute = TRUE), control.compute = list(dic = TRUE))

Paquete R-INLA

- La aproximación laplaciana simplificada (que es la opción predeterminada en el paquete R-INLA) tiene un menor costo y remedia de manera satisfactoria las inexactitudes en la ubicación y asimetría de la aproximación gaussiana.

Cap 3: Inferencia Bayesiana e INLA

- Mejoran la predicción y estimación de los modelos.
- Recordemos probabilidades.

y=(y1,y2,..,yn) : Datos observados

θ : Vector de parámetros desconocidos

π(y|θ): Distribución de probabilidad de los datos dado el vector

(Función de verosimilitud)

π(θ|y): Distribución de probabilidad del vector dados los datos

(Distribución posterior)

Distribución de probabilidad de los datos y el vector

Distribución de probabilidad del vector

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}, \boldsymbol{\theta})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Verosimilitud marginal de los datos observados

Se asume constante de escala.

INLA

$$y_i|\boldsymbol{x}, \boldsymbol{\theta} \sim \pi(y_i|x_i, \boldsymbol{\theta}),\ i = 1, \ldots, n,$$

$$\boldsymbol{x}|\boldsymbol{\theta} \sim N(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{Q}(\boldsymbol{\theta})^{-1}),$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}),$$

$$\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}).$$

y=(y1,y2,..,yn) : Datos observados
θ : Vector de hiperparámetros
x : Campo gaussiano latente

Paquete R-INLA – Predictor lineal

install.packages("INLA" , repos = "https://inla.r-inla-download.org/R/stable", dep = TRUE)

inla(), donde especificamos la fórmula, la familia, los datos y otras opciones.

El formato del predictor lineal en R-INLA es similar al de otros modelos lineales como lm()

$Y_i \sim N(\eta_i, \sigma^2), \quad i = 1, \ldots, n$

$\eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_i$

donde Y_i es la variable de respuesta, η_i es el predictor lineal, x_1, x_2 son dos variables explicativas, y $u_i \sim N(0, \sigma_u^2)$, la fórmula se escribe como

$y \sim x_1 + x_2 + f(i, model = "iid")$

Ten en cuenta que por defecto, la fórmula incluye una intersección (intercepto). Si quisiéramos incluir explícitamente β_0 en la fórmula, necesitaríamos eliminar el intercepto (agregando 0) e incluirlo como un término de covariable (agregando b_0):

$y \sim 0 + b_0 + x_1 + x_2 + f(i, model = "iid")$