

# Regularização, GAM e Bayes

Elias T Krainski

Curso de Especialização em  
Data Science & Big Data  
Universidade Federal do Paraná

29 de Setembro

Regularização

**Generalized Additive Models - GAM**

Efeitos aleatórios

Modelo de passeio aleatório

Modelos hierárquicos

**Integrated Nested Laplace Aproximations - INLA**

Possibilidades

Implementação

# Introdução

## Modelos de regressão para média de uma variável resposta

$$\begin{aligned} E(y) &= g^{-1}(\eta) \\ \eta &= f(x_1, x_2, \dots, x_p) \end{aligned} \tag{1}$$

# Introdução

## Modelos de regressão para média de uma variável resposta

$$\begin{aligned} E(y) &= g^{-1}(\eta) \\ \eta &= f(x_1, x_2, \dots, x_p) \end{aligned} \tag{1}$$

- Regressão linear (múltipla): efeito de  $X_j$  é constante ( $\beta_j$ )

$$\eta = \alpha + \beta_1 X_1 + \dots + \beta_p X_p$$

- $\eta$  é um hiper-plano  $p$  dimensional

# Introdução

## Modelos de regressão para média de uma variável resposta

$$\begin{aligned} E(y) &= g^{-1}(\eta) \\ \eta &= f(x_1, x_2, \dots, x_p) \end{aligned} \tag{1}$$

- Regressão linear (múltipla): efeito de  $X_j$  é constante ( $\beta_j$ )

$$\eta = \alpha + \beta_1 X_1 + \dots + \beta_p X_p$$

- $\eta$  é um hiper-plano  $p$  dimensional
- **Generalized Additive Models - GAM:** adição de termos suaves que dependem de uma ou mais variáveis
  - Um exemplo com  $p = 4$  covariáveis disponíveis:

$$f(x_1, \dots, x_4) = \beta_0 + \beta_1 X_1 + s_1(x_2) + s_2(x_3, x_4)$$

# Introdução

## Modelos de regressão para média de uma variável resposta

$$\begin{aligned} E(y) &= g^{-1}(\eta) \\ \eta &= f(x_1, x_2, \dots, x_p) \end{aligned} \tag{1}$$

- ▶ Regressão linear (múltipla): efeito de  $X_j$  é constante ( $\beta_j$ )

$$\eta = \alpha + \beta_1 X_1 + \dots + \beta_p X_p$$

- ▶  $\eta$  é um hiper-plano  $p$  dimensional
- ▶ **Generalized Additive Models - GAM:** adição de termos suaves que dependem de uma ou mais variáveis
  - ▶ Um exemplo com  $p = 4$  covariáveis disponíveis:

$$f(x_1, \dots, x_4) = \beta_0 + \beta_1 X_1 + s_1(x_2) + s_2(x_3, x_4)$$

- ▶ Podemos ainda ter efeitos aleatórios
  - ▶ **Generalized Additive Mixed Models - GAMM**

# Regularização

# Regularização em regressão

Penaliza a adição de termos

- ▶ **Ridge:** Penalização de segunda ordem

$$\sum_i (y_i - \mathbf{x}_i \beta)^2 + \lambda \sum_j \beta_j^2$$

# Regularização em regressão

Penaliza a adição de termos

- ▶ **Ridge:** Penalização de segunda ordem

$$\sum_i (y_i - \mathbf{x}_i \beta)^2 + \lambda \sum_j \beta_j^2$$

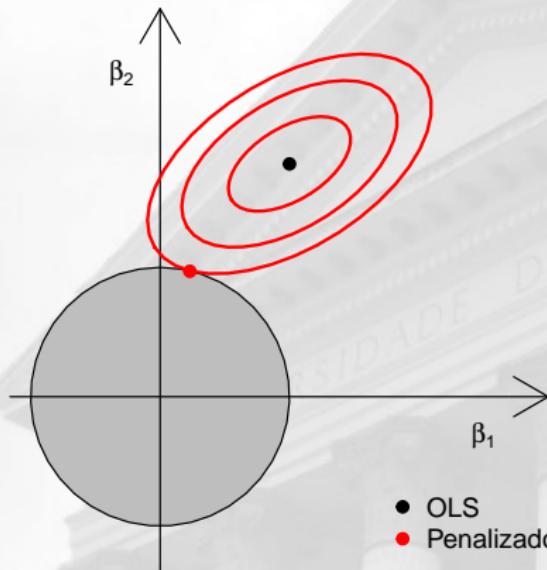
- ▶ **LASSO:** Penalização de primeira ordem

$$\sum_i (y_i - \mathbf{x}_i \beta)^2 + \lambda \sum_j |\beta_j|$$

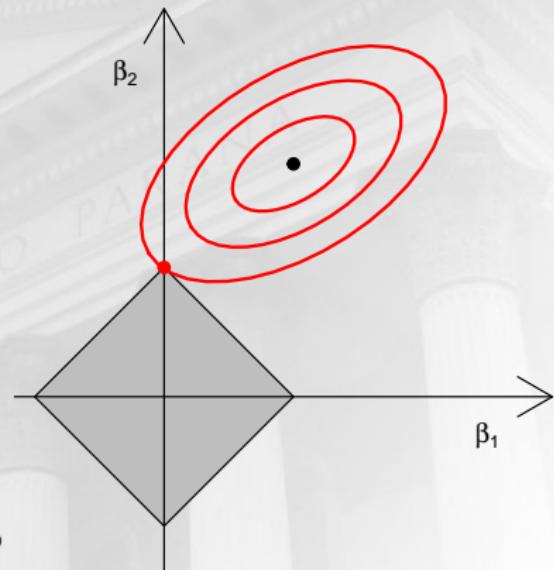
# Efeito da regularização

## Contração no feito das covariáveis

Ridge



LASSO



- OLS
- Penalizado

# Regularização: generalizações

- ▶ Ordem  $q$ ,  $q > 0$

$$\lambda \sum_j |\beta_j|^q$$

# Regularização: generalizações

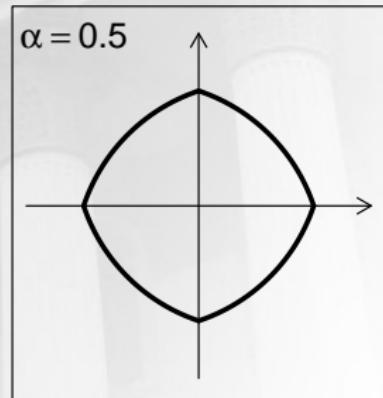
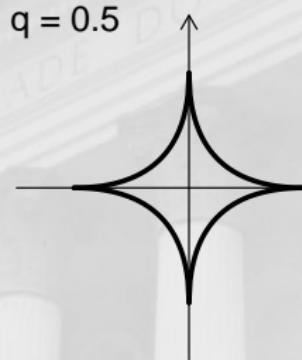
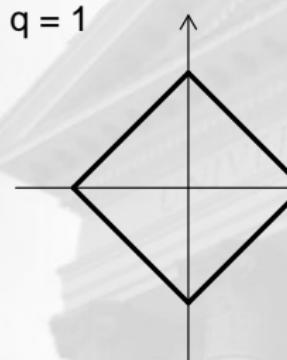
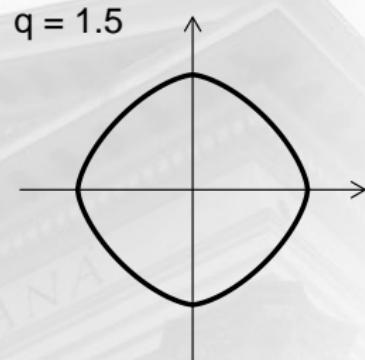
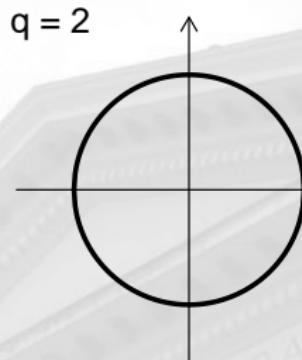
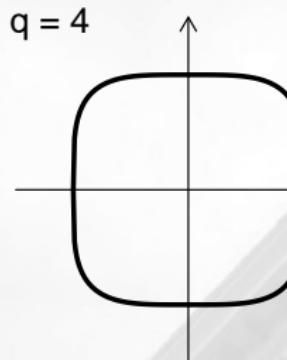
- ▶ Ordem  $q$ ,  $q > 0$

$$\lambda \sum_j |\beta_j|^q$$

- ▶ Rede elástica: pondera 1<sup>a</sup> e 2<sup>a</sup> ordem

$$\lambda \sum_j (\alpha |\beta_j| + (1 - \alpha) \beta_j^2)$$

# Efeito de Regularização: generalizações



# Exemplo: simulado

## Simula dados de um modelo linear

```
set.seed(1); n <- 1000; p <- 5    ## define n e p
X <- matrix(runif(n*p), ncol=p)      ## simula X
X <- scale(X)                      ## padroniza X
beta <- 0:(p-1)/p                  ## define coeficientes
mu <- 1 + X%*%beta                ## calcula E(Y)
y <- rnorm(n, m = mu, s = 0.5)     ## simula y
```

# Exemplo: simulado

## Simula dados de um modelo linear

```
set.seed(1); n <- 1000; p <- 5 ## define n e p
X <- matrix(runif(n*p), ncol=p) ## simula X
X <- scale(X) ## padroniza X
beta <- 0:(p-1)/p ## define coeficientes
mu <- 1 + X%*%beta ## calcula E(Y)
y <- rnorm(n, m = mu, s = 0.5) ## simula y
```

## Estima o modelo e visualiza sumário

```
result <- lm(y ~ X)
round(coef(summary(result)), 4)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.0278    0.0165  62.198  0.000
## X1          0.0096    0.0166   0.581  0.561
## X2          0.1736    0.0166  10.488  0.000
## X3          0.3887    0.0166  23.464  0.000
## X4          0.5808    0.0166  35.086  0.000
## X5          0.8046    0.0166  48.589  0.000
```

## Exemplo: estima manual

```
naive.lm.fun <- function(params, fmin=c("ols", "lik")) {  
  eta <- params[1] + X%*%params[1 + 1:p]  
  if (any(fmin=='ols'))  
    return(sum((y - eta)^2))  
  else  
    return(-sum(dnorm(y, eta, exp(params[p+2])), log=TRUE)))  
}
```

## Exemplo: estima manual

```
naive.lm.fun <- function(params, fmin=c("ols", "lik")) {  
  eta <- params[1] + X%*%params[1 + 1:p]  
  if (any(fmin=='ols'))  
    return(sum((y - eta)^2))  
  else  
    return(-sum(dnorm(y, eta, exp(params[p+2])), log=TRUE)))  
}
```

```
ols.res <- optim(rep(0, p+1), naive.lm.fun)  
lik.res <- optim(rep(0, p+2), naive.lm.fun, fmin='lik')  
rbind(true=c(beta0=1, beta=beta, sd=0.5),  
      lm=c(coef(result), NA), ols=c(ols.res$par, NA),  
      lik=c(lik.res$par[1:(p+1)], exp(lik.res$par[p+2])))  
##          beta0   beta1   beta2   beta3   beta4   beta5     sd  
## true  1.00  0.00000  0.200  0.400  0.600  0.800  0.500  
## lm    1.03  0.00963  0.174  0.389  0.581  0.805    NA  
## ols   1.03  0.00851  0.173  0.389  0.581  0.804    NA  
## lik   1.03  0.00523  0.175  0.398  0.579  0.802  0.521
```

# Exemplo: Ridge manual vs MASS

Estima  $\lambda$  e os coeficientes

```
naive.ridge.fun <- function(params) {  
  beta <- params[1 + 1:p]  
  eta <- params[1] + X%*%beta  
  lambda <- exp(params[p + 2])  
  penal <- lambda * sum(beta^2)  
  sum((y - eta)^2) + penal  
}  
(r.res <- optim(rep(0, p+2), naive.ridge.fun))$par  
## [1] 1.0280 0.0099 0.1729 0.3879 0.5793 0.8027 0.3663
```

# Exemplo: Ridge manual vs MASS

Estima  $\lambda$  e os coeficientes

```
naive.ridge.fun <- function(params) {  
  beta <- params[1 + 1:p]  
  eta <- params[1] + X%*%beta  
  lambda <- exp(params[p + 2])  
  penal <- lambda * sum(beta^2)  
  sum((y - eta)^2) + penal  
}  
(r.res <- optim(rep(0, p+2), naive.ridge.fun))$par  
## [1] 1.0280 0.0099 0.1729 0.3879 0.5793 0.8027 0.3663
```

Usa o  $\lambda$  estimado na função lm.ridge() do pacote MASS

```
lm.ridge(y ~ X, lambda=exp(r.res$par[p+2]))$coef  
##      X1      X2      X3      X4      X5  
## 0.00965 0.17329 0.38795 0.57965 0.80300
```

## Exemplo: Rede elástica manual vs glmnet

Estima  $\lambda$  e coeficientes fixando  $\alpha = 0.5$

```
naive.rel.fun <- function(params, alpha=0) {  
  beta <- params[1 + 1:p]  
  eta <- params[1] + X%*%beta  
  penal <- (1-alpha) * sum(beta^2)/2 +  
    alpha * sum(abs(beta))  
  lambda <- exp(params[p + 2])  
  mean((y-eta)^2)/2 + lambda * penal  
}  
r2.res <- optim(rep(0, p+2), naive.rel.fun, alpha=0.5)
```

## Exemplo: Rede elástica manual vs glmnet

Estima  $\lambda$  e coeficientes fixando  $\alpha = 0.5$

```
naive.rel.fun <- function(params, alpha=0) {  
  beta <- params[1 + 1:p]  
  eta <- params[1] + X%*%beta  
  penal <- (1-alpha) * sum(beta^2)/2 +  
    alpha * sum(abs(beta))  
  lambda <- exp(params[p + 2])  
  mean((y-eta)^2)/2 + lambda * penal  
}  
r2.res <- optim(rep(0, p+2), naive.rel.fun, alpha=0.5)
```

Usa o  $\lambda$  estimado na função `glmnet()` do pacote `glmnet`

```
g <- glmnet(X, y, standardize = FALSE,  
             alpha=0.5, lambda=exp(r2.res$par[p+2]))  
rbind(r2.res$par[1:(p+1)], c(g$a0, drop(g$beta)))  
##          s0      V1      V2      V3      V4      V5  
## [1,] 0.995 0.32 0.0941 0.334 0.617 0.804  
## [2,] 1.028 0.00 0.1593 0.372 0.563 0.785
```

# **Generalized Additive Models - GAM**

# GAM: Funções bases e penalização

- ▶ Ajusta modelos de regressão com termos suaves, Wood (2017)
  - ▶ Exemplo:

$$\eta = \beta_0 + s(x)$$

# GAM: Funções bases e penalização

- ▶ Ajusta modelos de regressão com termos suaves, Wood (2017)

- ▶ Exemplo:

$$\eta = \beta_0 + s(x)$$

- ▶ Cada termo suave,  $s(x)$  pode ser representado por

$$s(x) = \sum_k \beta_k B_k(x)$$

- ▶  $B_k(x)$  é a  $k$ -ésima função base em  $x$
- ▶ recomenda-se que  $k << n$

# GAM: Penalização

- ▶ Penalização quadrática

- ▶ evita sobre-estimação ( *overfit* )
- ▶ Função à minimizar:

$$\sum_i (y_i - \beta_0 - s(x))^2 + \lambda \int s''(x) dx$$

- ▶ Representação da segunda derivada de cada termo suave

$$\int s_j''(x)^2 dx = \beta_j^T \mathbf{S}_j \beta_j$$

- A matriz **S** é conhecida - Depende das funções bases consideradas

## Estimação no pacote mgcv

- ▶ O processo de estimativa é feito em dois estágios
  - ▶ Optimiza em relação a  $\lambda$
  - ▶ Para cada  $\lambda$ , estima um GLM

## Estimação no pacote mgcv

- ▶ O processo de estimação é feito em dois estágios
  - ▶ Optimiza em relação a  $\lambda$
  - ▶ Para cada  $\lambda$ , estima um GLM
- ▶ O **Generalized Cross Validation - GCV** é usado na optimização:

$$nD/(n + k)^2$$

- ▶  $D$ : deviance
- ▶  $n$ : número de observações
- ▶  $k$ : graus de liberdade do modelo

## Estimação no pacote mgcv

- ▶ O processo de estimação é feito em dois estágios
  - ▶ Optimiza em relação a  $\lambda$
  - ▶ Para cada  $\lambda$ , estima um GLM
- ▶ O **Generalized Cross Validation - GCV** é usado na optimização:

$$nD/(n + k)^2$$

- ▶  $D$ : deviance
- ▶  $n$ : número de observações
- ▶  $k$ : graus de liberdade do modelo
- ▶ Procedimento anàlogo à inferêncià Bayesiana empírica

# Efeitos aleatórios

# Efeitos aleatórios

- ▶ Efeito fixo
  - ▶ O efeito de uma covariável é assumido como fixo
  - ▶ Espera-se que o efeito seja o mesmo para uma amostra diferente

# Efeitos aleatórios

- ▶ Efeito fixo

- ▶ O efeito de uma covariável é assumido como fixo
- ▶ Espera-se que o efeito seja o mesmo para uma amostra diferente

- ▶ Efeito aleatório

- ▶ Há situações em que o efeito não é fixo, varia para um novo conjunto de dados
  - ▶ variações entre lotes fabricados,
  - ▶ variações temporais
  - ▶ variações de indivíduos

# Efeito aleatório no modelo linear

- ▶ Seja o preditor linear especificado como

$$\eta = \mathbf{X}\beta + \mathbf{Z}\mathbf{b}$$

- ▶  $\mathbf{X}$  é a matriz covariáveis,
- ▶  $\beta$  são os coeficientes de regressão (efeitos fixos),
- ▶  $\mathbf{Z}$  é uma matriz de efeitos aleatórios,
- ▶  $\mathbf{b}$  são os efeitos aleatórios.
  - ▶ exemplo: efeito de indivíduo, quando há mais de uma observação de cada indivíduo

# Efeito aleatório no modelo linear

- ▶ Seja o preditor linear especificado como

$$\eta = \mathbf{X}\beta + \mathbf{Z}\mathbf{b}$$

- ▶  $\mathbf{X}$  é a matriz covariáveis,
- ▶  $\beta$  são os coeficientes de regressão (efeitos fixos),
- ▶  $\mathbf{Z}$  é uma matriz de efeitos aleatórios,
- ▶  $\mathbf{b}$  são os efeitos aleatórios.
  - ▶ exemplo: efeito de indivíduo, quando há mais de uma observação de cada indivíduo
- ▶ Assume-se uma distribuição para os efeitos aleatórios  $\mathbf{b}$ 
  - ▶ geralmente:  $\mathbf{b} \sim N(\mathbf{0}, \Sigma)$

# Efeito aleatório no modelo linear

- ▶ Seja o preditor linear especificado como

$$\eta = \mathbf{X}\beta + \mathbf{Z}\mathbf{b}$$

- ▶  $\mathbf{X}$  é a matriz covariáveis,
- ▶  $\beta$  são os coeficientes de regressão (efeitos fixos),
- ▶  $\mathbf{Z}$  é uma matriz de efeitos aleatórios,
- ▶  $\mathbf{b}$  são os efeitos aleatórios.
  - ▶ exemplo: efeito de indivíduo, quando há mais de uma observação de cada indivíduo
- ▶ Assume-se uma distribuição para os efeitos aleatórios  $\mathbf{b}$ 
  - ▶ geralmente:  $\mathbf{b} \sim N(\mathbf{0}, \Sigma)$
- ▶ É comum considerar algum modelo para suavizar  $\mathbf{b}$ 
  - ▶ Exemplo: suavizar:  $b_i$  similar a  $b_{i+1}$

Exemplo: tempo discretizado em  $k$  períodos:

$$\mathbf{z} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ \ddots \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times k}$$

Exemplo: tempo discretizado em  $k$  períodos:

$$\mathbf{Z} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ \ddots \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times k}$$

- se  $k = n$  temos  $\mathbf{Z} = \mathbf{I}_{n \times n}$



## Modelo de passeio aleatório

# Modelo para diferenças sucessivas

- ▶ modelagem das **diferenças sucessivas**
  - ▶ Seja  $b_1, \dots, b_n$  variáveis aleatórias
  - ▶ Seja as diferenças sucessivas:  $b_i - b_{i+1}$
  - ▶ Considere  $b_i - b_{i+1} \sim N(0, \sigma^2)$

# Modelo para diferenças sucessivas

- ▶ modelagem das **diferenças sucessivas**
  - ▶ Seja  $b_1, \dots, b_n$  variáveis aleatórias
  - ▶ Seja as diferenças sucessivas:  $b_i - b_{i+1}$
  - ▶ Considere  $b_i - b_{i+1} \sim N(0, \sigma^2)$
- ▶  $\sigma^2$  é o parâmetro desse modelo
- ▶ quanto menor  $\sigma^2$ , menor a variação em  $b_i - b_{i+1}$
- ▶ sem variação =  $0 \leftarrow \sigma^2 \rightarrow \infty$  = sem suavização

# Distribuição conjunta

- distribuição conjunta do vetor  $\mathbf{b} = b_1, \dots, b_n$

$$\begin{aligned} p(\mathbf{b}) &= \prod_{i=1}^{n-1} p(b_i | b_{i+1}) \\ &= \prod_{i=1}^{n-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(b_i - b_{i+1})^2}{2\sigma^2}} \end{aligned}$$

# Distribuição conjunta

- distribuição conjunta do vetor  $\mathbf{b} = b_1, \dots, b_n$

$$\begin{aligned} p(\mathbf{b}) &= \prod_{i=1}^{n-1} p(b_i | b_{i+1}) \\ &= \prod_{i=1}^{n-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(b_i - b_{i+1})^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{(n-1)/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n-1} (b_i - b_{i+1})^2} \end{aligned}$$

## Distribuição conjunta (cont.)

- ▶ Notar que

$$\begin{aligned}\sum_{i=1}^{n-1} (b_i - b_{i+1})^2 &= \sum_{i=1}^{n-1} (b_i^2 - 2b_i b_{i+1} + b_{i+1}^2) \\ &= \sum_{i=1}^{n-1} b_i^2 - 2 \sum_{i=1}^{n-1} b_i b_{i+1} + \sum_{i=2}^n b_i^2\end{aligned}$$

## Distribuição conjunta (cont.)

- ▶ Notar que

$$\begin{aligned}\sum_{i=1}^{n-1} (b_i - b_{i+1})^2 &= \sum_{i=1}^{n-1} (b_i^2 - 2b_i b_{i+1} + b_{i+1}^2) \\ &= \sum_{i=1}^{n-1} b_i^2 - 2 \sum_{i=1}^{n-1} b_i b_{i+1} + \sum_{i=2}^n b_i^2 = \mathbf{b}^T \mathbf{R} \mathbf{b}\end{aligned}$$

onde

$$\mathbf{R} = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ -1 & -1 & 2 & -1 & & \\ & & & \ddots & & \\ & & & & -1 & 2 & -1 \\ & & & & -1 & -1 & 2 & -1 \\ & & & & & & 1 \end{pmatrix}$$

## Distribuição conjunta (cont.)

- ▶ Notar que

$$\begin{aligned}\sum_{i=1}^{n-1} (b_i - b_{i+1})^2 &= \sum_{i=1}^{n-1} (b_i^2 - 2b_i b_{i+1} + b_{i+1}^2) \\ &= \sum_{i=1}^{n-1} b_i^2 - 2 \sum_{i=1}^{n-1} b_i b_{i+1} + \sum_{i=2}^n b_i^2 = \mathbf{b}^T \mathbf{R} \mathbf{b}\end{aligned}$$

onde

$$\mathbf{R} = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & & \ddots & & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 & -1 \\ & & & & & & 1 & \end{pmatrix}$$

- ▶  $p(\mathbf{b}) = (2\pi\sigma^2)^{(n-1)/2} e^{-\frac{1}{2\sigma^2} \mathbf{b}^T \mathbf{R} \mathbf{b}}$

## Distribuição de $b$ para o exemplo de Tóquio

- ▶ é razoável considerar  $b_1$  similar a  $b_n$
- ▶ passeio aleatório cíclico
- ▶ multiplicando a distribuição conjunta anterior por

$$(2\pi\sigma^2)^{1/2} e^{-\frac{(b_n - b_1)^2}{2\sigma^2}}$$

## Distribuição de $\mathbf{b}$ para o exemplo de Tóquio

- ▶ é razoável considerar  $b_1$  similar a  $b_n$
- ▶ passeio aleatório cíclico
- ▶ multiplicando a distribuição conjunta anterior por

$$(2\pi\sigma^2)^{1/2} e^{-\frac{(b_n - b_1)^2}{2\sigma^2}}$$

temos

$$p(\mathbf{b}_c) = (2\pi\sigma^2)^{n/2} e^{-\frac{1}{2\sigma^2} \mathbf{b}_c^T \mathbf{R}_c \mathbf{b}_c}$$

onde

$$\mathbf{R}_c = \begin{pmatrix} 2 & -1 & & & & & -1 \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & & \ddots & & & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 & -1 \\ -1 & & & & & & -1 & 2 \end{pmatrix}$$



# Modelos hierárquicos

# Modelos hierárquicos

- Nível 1 - variável resposta,  $y$

$$y|\boldsymbol{b}, \theta_1 \sim \pi(y|\boldsymbol{b}, \theta_1) = \prod_{i=1}^n \pi(y_i|x, \theta_1) \text{(ind. cond.)}$$

$\theta_1$  é um hyper-parâmetro na verossimilhança

# Modelos hierárquicos

- ▶ Nível 1 - variável resposta,  $y$

$$y|\boldsymbol{b}, \theta_1 \sim \pi(y|\boldsymbol{b}, \theta_1) = \prod_{i=1}^n \pi(y_i|x, \theta_1) \text{(ind. cond.)}$$

$\theta_1$  é um hyper-parâmetro na verossimilhança

- ▶ Nível 2 - efeito/campo latente/não-observado,  $\boldsymbol{b}$ 
  - ▶ Geralmente assume distribuição Gaussiana → INLA
  - ▶ Essa distribuição, por sua vez, possui hyper-parâmetro(s)

# Modelos hierárquicos

- Nível 1 - variável resposta,  $y$

$$y|\mathbf{b}, \theta_1 \sim \pi(y|\mathbf{b}, \theta_1) = \prod_{i=1}^n \pi(y_i|x, \theta_1) \text{(ind. cond.)}$$

$\theta_1$  é um hyper-parâmetro na verossimilhança

- Nível 2 - efeito/campo latente/não-observado,  $\mathbf{b}$ 
  - Geralmente assume distribuição Gaussiana → INLA
  - Essa distribuição, por sua vez, possui hyper-parâmetro(s)

Se Bayesiano:

- Nível 3 - distribuição para os hyper-parâmetros,  $\theta_2$

$$\mathbf{b}|\theta_2 \sim \pi(\mathbf{b}|\theta_2) = N(\mathbf{0}, \mathbf{Q}(\theta_2)^{-1})$$

# Modelos hierárquicos

- ▶ Nível 1 - variável resposta,  $y$

$$y|\mathbf{b}, \theta_1 \sim \pi(y|\mathbf{b}, \theta_1) = \prod_{i=1}^n \pi(y_i|x, \theta_1) \text{(ind. cond.)}$$

$\theta_1$  é um hyper-parâmetro na verossimilhança

- ▶ Nível 2 - efeito/campo latente/não-observado,  $\mathbf{b}$ 
  - ▶ Geralmente assume distribuição Gaussiana → INLA
  - ▶ Essa distribuição, por sua vez, possui hyper-parâmetro(s)

Se Bayesiano:

- ▶ Nível 3 - distribuição para os hyper-parâmetros,  $\theta_2$

$$\mathbf{b}|\theta_2 \sim \pi(\mathbf{b}|\theta_2) = N(\mathbf{0}, \mathbf{Q}(\theta_2)^{-1})$$

- ▶ Temos:  $\theta = \{\theta_1, \theta_2\}$  (hyper-parâmetros)

$$\theta \sim \pi(\theta) \rightarrow \text{se Bayesiano}$$

# $\pi(\mathbf{y}|\mathbf{b}, \theta)$ : verossimilhança

Depende de

- ▶ tipo dos valores da variável resposta
  - ▶ binaria (sim/não), contagens, contínua, censurada

# $\pi(\mathbf{y}|\mathbf{b}, \theta)$ : verossimilhança

Depende de

- ▶ tipo dos valores da variável resposta
  - ▶ binaria (sim/não), contagens, contínua, censurada
- ▶ como é coletada
  - ▶ **usual:** cada indivíduo tem uma única observação
  - ▶ cada indivíduo pode ter mais de uma observação
  - ▶ Processo pontual: tem-se localizações (no tempo e/ou espaço) de eventos

## $\pi(\mathbf{b}|Q(\theta))$ : Campo latente

► É

- Não observável
- Chamado de campo aleatório **Gaussiano** latente
- Para o INLA: Gaussiano e Markoviano

## $\pi(\mathbf{b}|Q(\theta))$ : Campo latente

- ▶ É
  - ▶ Não observável
  - ▶ Chamado de campo aleatório **Gaussiano** latente
  - ▶ Para o INLA: Gaussiano e Markoviano
  
- ▶ Representa
  - ▶ Efeitos de covariáveis (coeficientes ou termo suave)
  - ▶ Efeitos aleatórios (de indivíduos, séries temporais)

## $\pi(\mathbf{b}|Q(\theta))$ : Campo latente

- ▶ É
  - ▶ Não observável
  - ▶ Chamado de campo aleatório **Gaussiano** latente
  - ▶ Para o INLA: Gaussiano e Markoviano
- ▶ Representa
  - ▶ Efeitos de covariáveis (coeficientes ou termo suave)
  - ▶ Efeitos aleatórios (de indivíduos, séries temporais)
- ▶ Pode ser
  - ▶ não estruturado (indivíduos independentes)
  - ▶ estruturado (efeito em tempos vizinhos são similares)
  - ▶ mais de uma estrutura combinada (covariaveis, não ou estruturados)

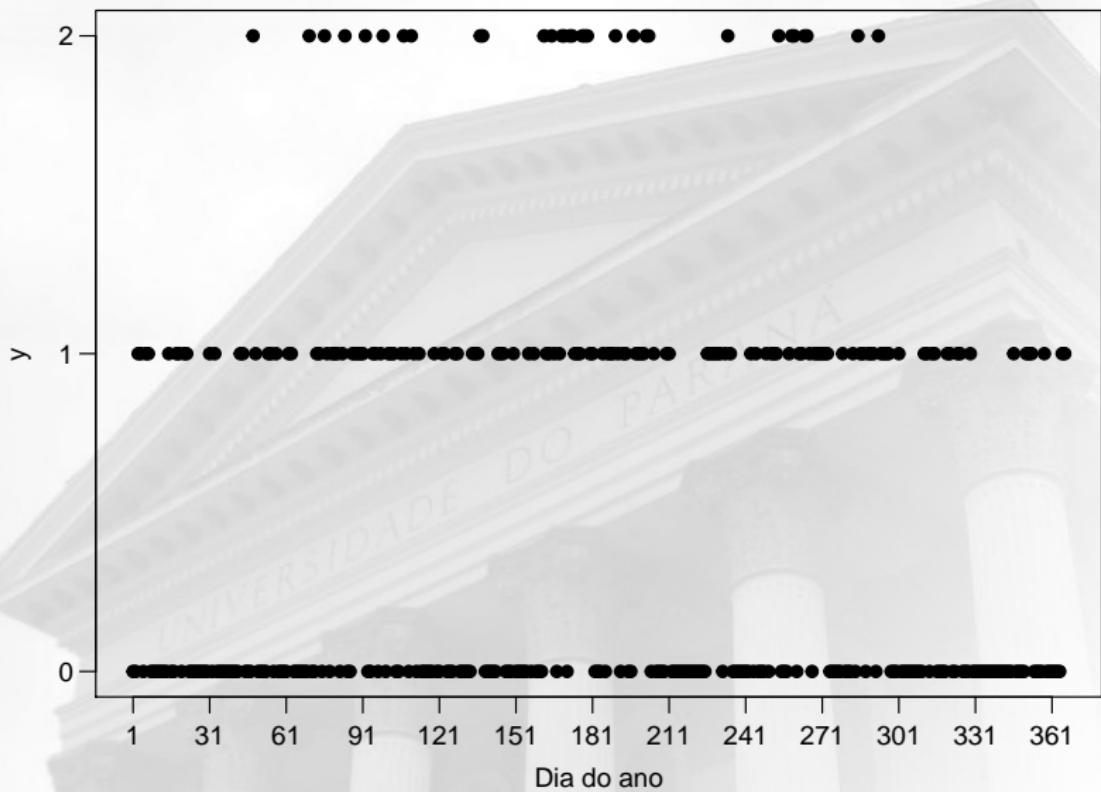
## Distribuição *a priori* de $\theta$ : $\pi(\theta)$

- ▶ Parâmetros da verossimilhança e da distribuição de  $b$
- ▶ Exemplos (verossimilhança):
  - ▶ parâmetro de precisão (Normal, gamma, beta, binomial negativa)
  - ▶ probabilidade de adicional de zero
- ▶ Exemplos (campo latente)
  - ▶ parâmetro de precisão/variância do efeito aleatório
  - ▶ parâmetro de correlação (no modelo AR1 por exemplo)
  - ▶ parâmetro de alcance (em alguns modelos espaciais)



# *Integrated Nested Laplace Aproximations -* INLA

## Exemplo 1: Dias chuvosos por dia do ano (dois anos)



Problema: modelar a probabilidade de chuva por dia do ano

# Modelo

- ▶  $y_i$  assume valores 0, 1 ou 2, para  $i = 1, \dots, n$ 
  - ▶ assumindo independência condicional, temos

$$y_i | p_i \sim \text{Binomial}(2, p_i)$$

- ▶ função de ligação (logito):

$$p_i = 1 / (1 + \exp(-x_i))$$

# Modelo

- ▶  $y_i$  assume valores 0, 1 ou 2, para  $i = 1, \dots, n$ 
  - ▶ assumindo independência condicional, temos

$$y_i | p_i \sim \text{Binomial}(2, p_i)$$

- ▶ função de ligação (logito):

$$p_i = 1 / (1 + \exp(-x_i))$$

- ▶ **b**: suavização ao longo do tempo
  - ▶ *Randon Walk* - RW de primeira ordem: rw1
  - ▶ distribuição Gaussiana para diferenças sucessivas (**R** esparsa)

$$x_i - x_{i-1} \sim N(0, \tau^{-1})$$

$$\log(\pi(\mathbf{b}|\tau)) \propto -\frac{\tau}{2} [(x_1 - x_n)^2 + \sum_{i=2}^n (x_i - x_{i-1})^2] = -\frac{\tau}{2} \mathbf{b}' \mathbf{R} \mathbf{b},$$

- ▶ **b** é um *Gaussian Markov Random Field* - GMRF, Rue and Held (2005)

# Modelo

- ▶  $y_i$  assume valores 0, 1 ou 2, para  $i = 1, \dots, n$ 
  - ▶ assumindo independência condicional, temos

$$y_i | p_i \sim \text{Binomial}(2, p_i)$$

- ▶ função de ligação (logito):

$$p_i = 1 / (1 + \exp(-x_i))$$

- ▶ **b**: suavização ao longo do tempo
  - ▶ *Randon Walk* - RW de primeira ordem: rw1
  - ▶ distribuição Gaussiana para diferenças sucessivas (**R** esparsa)

$$x_i - x_{i-1} \sim N(0, \tau^{-1})$$

$$\log(\pi(\mathbf{b}|\tau)) \propto -\frac{\tau}{2} [(x_1 - x_n)^2 + \sum_{i=2}^n (x_i - x_{i-1})^2] = -\frac{\tau}{2} \mathbf{b}' \mathbf{R} \mathbf{b},$$

- ▶ **b** é um *Gaussian Random Field* - GMRF, Rue and Held (2005)
- ▶  $\tau$ : parâmetro de precisão local

# Modelo, INLA

## ► Modelo

$y_i|x_i \sim \text{Binomial}(2, p_i)$  → verossimilhança

$\mathbf{b}|\tau \sim N(\mathbf{0}, (\tau \mathbf{R})^-)$  → campo latente GMRF

$\tau \sim p(\tau)$  → distribuição à priori

# Modelo, INLA

## ► Modelo

$$\begin{aligned}y_i | x_i &\sim \text{Binomial}(2, p_i) \rightarrow \text{verossimilhança} \\ \mathbf{b} | \tau &\sim N(\mathbf{0}, (\tau \mathbf{R})^-) \rightarrow \text{campo latente GMRF} \\ \tau &\sim p(\tau) \rightarrow \text{distribuição à priori}\end{aligned}$$

► INLA: Sigla de *Integrated Nested Laplace Approximation*, Rue, Martino, and Chopin (2009)

► Primeria *Laplace Approximation* - LA para calcular

$$\pi(\tau | \mathbf{y})$$

► Segunda (aninhada à primeira) LA para calcular

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

# Modelo, INLA

## ► Modelo

$$\begin{aligned}y_i | x_i &\sim \text{Binomial}(2, p_i) \rightarrow \text{verossimilhança} \\ \mathbf{b} | \tau &\sim N(\mathbf{0}, (\tau \mathbf{R})^-) \rightarrow \text{campo latente GMRF} \\ \tau &\sim p(\tau) \rightarrow \text{distribuição à priori}\end{aligned}$$

- INLA: Sigla de *Integrated Nested Laplace Approximation*, Rue, Martino, and Chopin (2009)

- Primeira *Laplace Approximation* - LA para calcular

$$\pi(\tau | \mathbf{y})$$

- Segunda (aninhada à primeira) LA para calcular

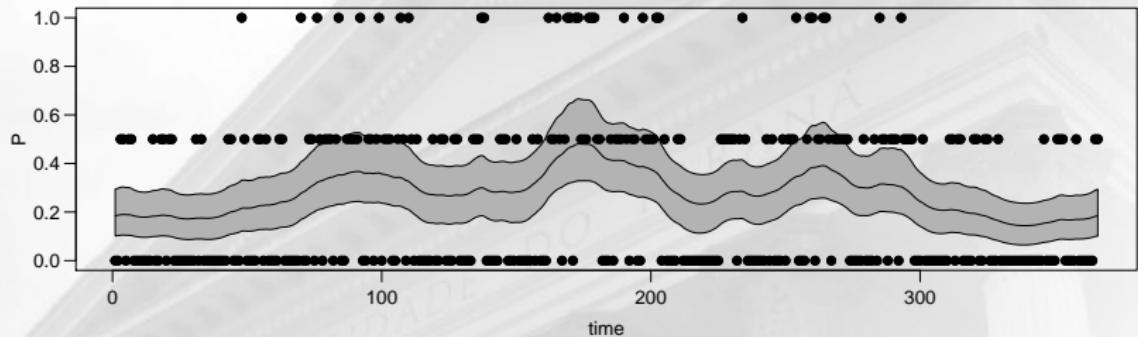
$$\pi(x_i | \mathbf{y})$$

- Se a verossimilhança é Gaussiana a “aproximação” é exata

## Exemplo de série temporal com rw1

```
head(Tokyo, 5)
##   y n time
## 1 0 2    1
## 2 0 2    2
## 3 1 2    3
## 4 1 2    4
## 5 0 2    5
modelo <- y ~ f(time, model='rw1', cyclic=TRUE)
resultado <- inla(modelo, family='binomial',
                    data=Tokyo, Ntrials=n,
                    control.compute=list(cpo=TRUE))
```

# Resultado da série temporal



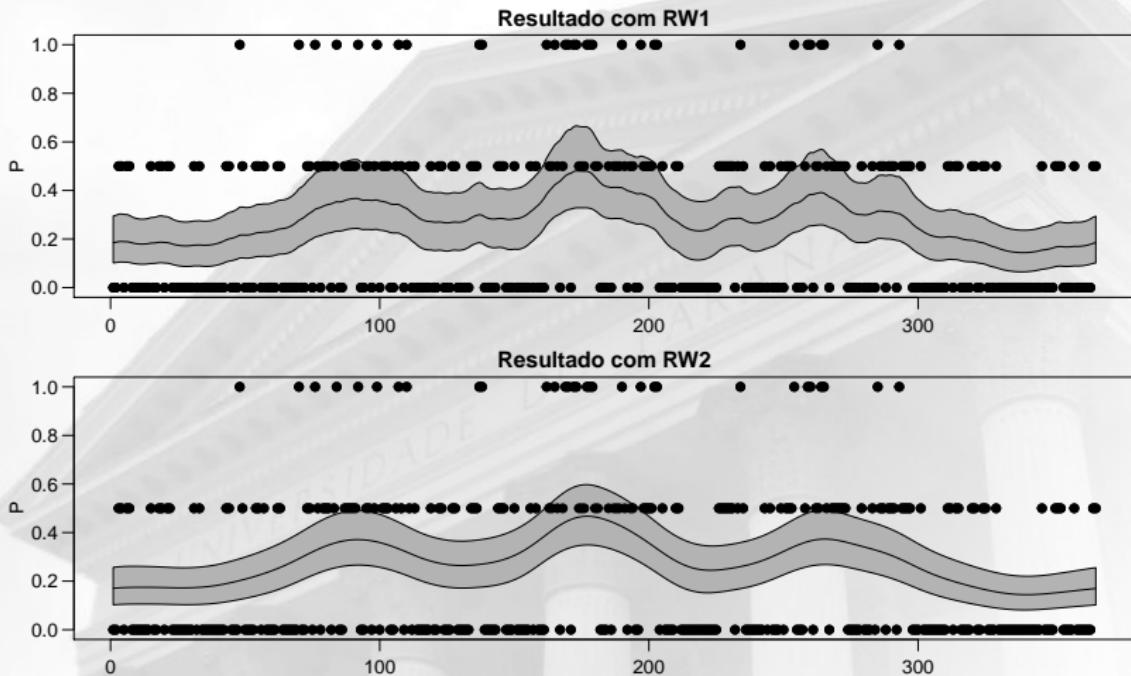
## Suavizando mais

Distribuição Gaussiana para diferenças de segunda ordem (rw2)

$$x_i - 2x_{i-1} + x_{i-2} \sim N(0, \tau^{-1})$$

```
modelo2 <- y ~ f(time, model='rw2', cyclic=TRUE)
resultado2 <- inla(modelo2, family='binomial',
                     data=Tokyo, Ntrials=n,
                     control.compute=list(cpo=TRUE))
```

# Ambos os resultados para a série temporal



# Medidas de ajuste

- ▶ *Conditional Predictive Ordinate - CPO:*

$$P(y_i^{\text{obs}} | \mathbf{y}_{-i})$$

$\mathbf{y}_{-i}$  é o vetor  $y$  sem a observação  $i$ .

```
c(-sum(log(resultado$cpo$cpo)),  
  -sum(log(resultado2$cpo$cpo)))  
## [1] 313 314
```

# Medidas de ajuste

- ▶ *Conditional Predictive Ordinate* - CPO:

$$P(y_i^{\text{obs}} | \mathbf{y}_{-i})$$

$\mathbf{y}_{-i}$  é o vetor  $y$  sem a observação  $i$ .

```
c(-sum(log(resultado$cpo$cpo)),  
  -sum(log(resultado2$cpo$cpo)))  
## [1] 313 314
```

- ▶ *Probability Integral Transform* - PIT:

$$P(Y_i \leq y_i^{\text{obs}} | \mathbf{y}_{-i}).$$

# Medidas de ajuste

- ▶ *Conditional Predictive Ordinate* - CPO:

$$P(y_i^{\text{obs}} | \mathbf{y}_{-i})$$

$\mathbf{y}_{-i}$  é o vetor  $y$  sem a observação  $i$ .

```
c(-sum(log(resultado$cpo$cpo)),  
  -sum(log(resultado2$cpo$cpo)))  
## [1] 313 314
```

- ▶ *Probability Integral Transform* - PIT:

$$P(Y_i \leq y_i^{\text{obs}} | \mathbf{y}_{-i}).$$

- ▶ úteis para detectar falta de ajuste ou *outliers*



Possibilidades

# Possibilidades de modelos

- ▶ Modelos (mistos) generalizados
- ▶ Modelos (mistos) aditivos generalizados
- ▶ Modelos de sobrevivência
- ▶ Modelos dinâmicos
- ▶ Modelos de volatilidade estocástica
- ▶ Suavização por *spline*
- ▶ Regressão semiparamétrica
- ▶ Mapeamento de doenças
- ▶ Geoestatística baseada em modelos\*
- ▶ Processos de Cox log-Gaussianos
- ▶ Modelos espaço-temporais
- ▶ Regressão (semiparamétrica) com coeficientes variando no espaço / espaço-temporal
- ▶ +++

# Possibilidades de modelos

- ▶ Modelos (mistos) generalizados
- ▶ Modelos (mistos) aditivos generalizados
- ▶ Modelos de sobrevivência
- ▶ Modelos dinâmicos
- ▶ Modelos de volatilidade estocástica
- ▶ Suavização por *spline*
- ▶ Regressão semiparamétrica
- ▶ Mapeamento de doenças
- ▶ Geoestatística baseada em modelos\*
- ▶ Processos de Cox log-Gaussianos
- ▶ Modelos espaço-temporais
- ▶ Regressão (semiparamétrica) com coeficientes variando no espaço / espaço-temporal
- ▶ +++

→ GLMM, GAM, GAMM, ... são diferentes nomes para a mesma coisa

## Flexibilidade e responsabilidade

- ▶ No exemplo ocorrência de chuva, suponha que adicionamos mais termos (covariáveis, efeito sazonal)

# Flexibilidade e responsabilidade

- ▶ No exemplo ocorrência de chuva, suponha que adicionamos mais termos (covariáveis, efeito sazonal)
- ▶ Flexibilidade deve ser acompanhada de responsabilidade

# Flexibilidade e responsabilidade

- ▶ No exemplo ocorrência de chuva, suponha que adicionamos mais termos (covariáveis, efeito sazonal)
- ▶ Flexibilidade deve ser acompanhada de responsabilidade
- ▶ PC-prior ***Penalized Complexity*** prior, Simpson et al. (2017)

# Implementação

# Implementação

## O código

- ▶ gmrflib, inlaprog, fmesh  
  - ▶ ~6000 linhas em Fortran
  - ▶ +100000 linhas em C++
- ▶ bibliotecas externas: muparser, taucs (matrizes esparsas),  
**PARDISO (experimental)**
- ▶ ~60000 linhas de código+help em R (R-INLA)

# Implementação

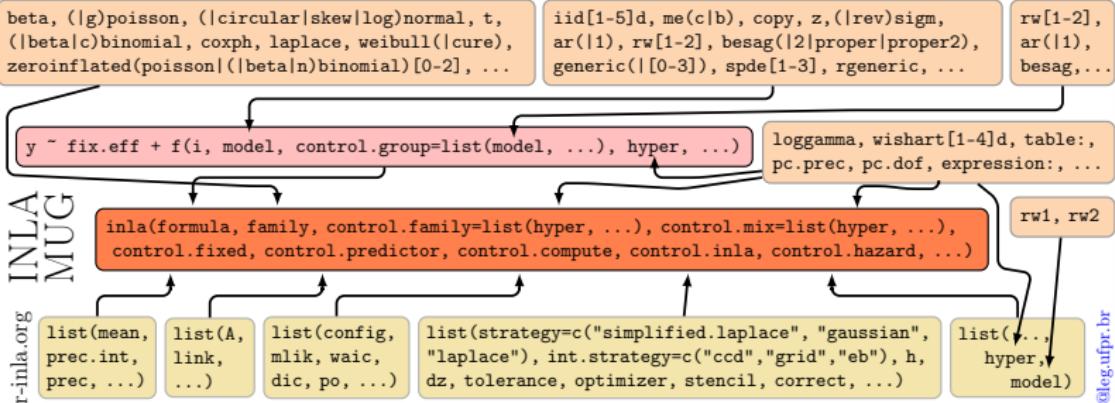
## O código

- ▶ gmrflib, inlaprog, fmesh  
  - ▶ ~6000 linhas em Fortran
  - ▶ +100000 linhas em C++
- ▶ bibliotecas externas: muparser, taucs (matrizes esparsas),  
**PARDISO (experimental)**
- ▶ ~60000 linhas de código+help em R (R-INLA)

## O trabalho

- ▶ +15 anos Håvard Rue (~80% do código)
- ▶ +10 anos Finn Lindgren (~20% do código)
- ▶ código fonte no bitbucket (open source), web page:  
[r-inla.org](http://r-inla.org)

# Função principal no R-INLA



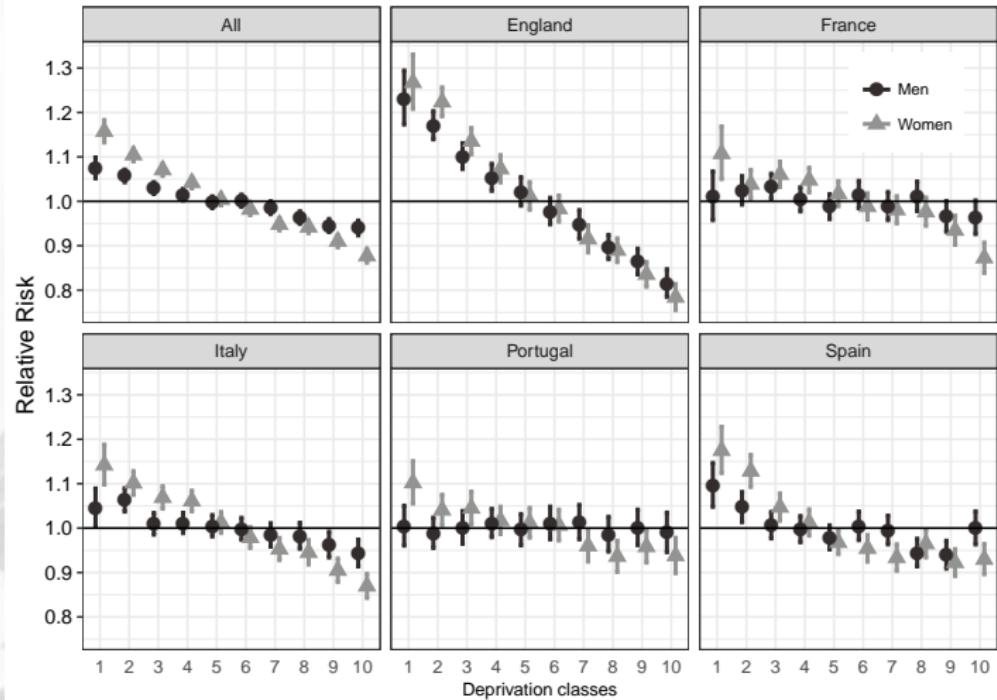
A faint, grayscale watermark-like image of the facade of the Paraná State University (Universidade do Paraná) is visible in the background. The building features classical architectural details like columns and a triangular pediment. The text "UNIVERSIDADE DO PARANÁ" is partially visible on the facade.

# Aplicações

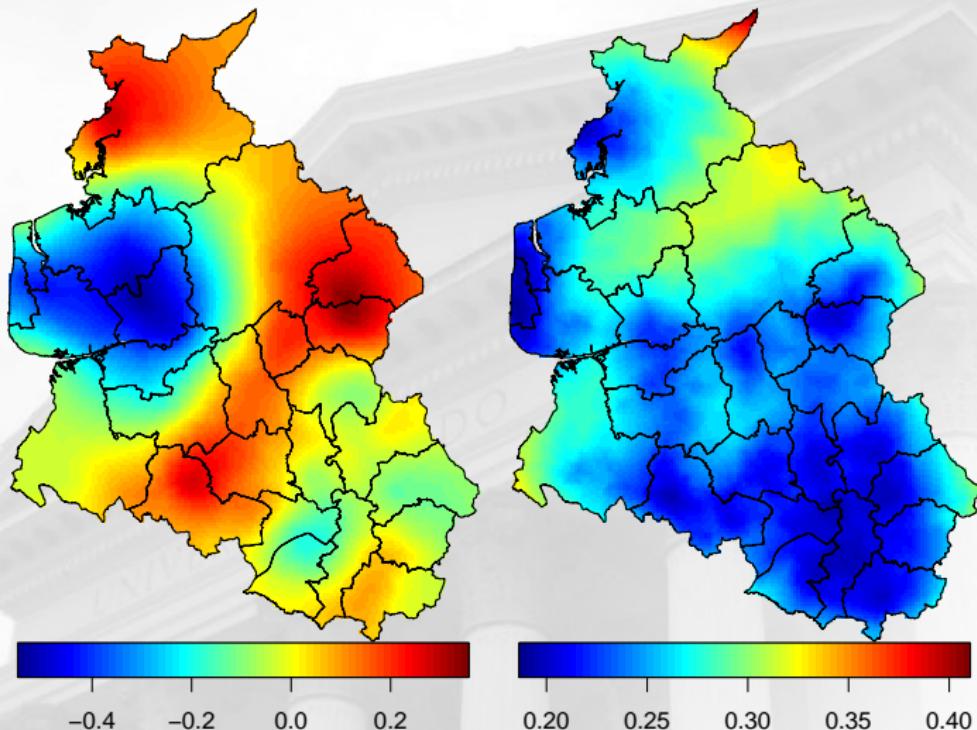
# Algumas aplicações citadas em Rue et al. (2017)

Recent examples of applications using the R-INLA package for statistical analysis include disease mapping ( Schrödle & Held 2011a , b ; Ugarte et al. 2014 , 2016 ; Papoila et al. 2014 ; Goicoa et al. 2016 ; Riebler et al. 2016 ); age-period-cohort models ( Riebler & Held 2016 ); a study of the evolution of the Ebola virus ( Santermans et al. 2016 ); the relationships between access to housing, health, and well-being in cities ( Kandt et al. 2016 ); the prevalence and correlates of intimate partner violence against men in Africa ( Tsiko 2016 ); a search for evidence of gene expression heterosis ( Niemi et al. 2015 ); analysis of traffic pollution and hospital admissions in London ( Halonen et al. 2016 ); early transcriptome changes in maize primary root tissues in response to moderate water deficit conditions by RNA sequencing ( Opitz et al. 2016 ); performance of inbred and hybrid genotypes in plant breeding and genetics ( Lithio & Nettleton 2015 ); a study of Norwegian emergency wards ( Goth et al. 2014 ); effects of measurement errors ( Muff et al. 2015 , Muff & Keller 2015 , Kröger et al. 2016 ); network meta-analysis ( Sauter & Held 2015 ); time-series analysis of genotyped human campylobacteriosis cases from the Manawatu region of New Zealand ( Friedrich et al. 2016 ); modeling of parrotfish habitats ( NC Roos et al. 2015 ); Bayesian outbreak detection ( Salmon et al. 2015 ); long-term trends in the number of Monarch butterflies ( Crewe & Mccracken 2015 ); long-term effects on hospital admission and mortality of road traffic noise ( Halonen et al. 2015 ); spatio-temporal dynamics of brain tumors ( Julian et al. 2015 ); ovarian cancer mortality ( García-Pérez et al. 2015 ); the effect of preferential sampling on phylodynamic inference ( Karcher et al. 2016 ); analysis of the impact of climate change on abundance trends in central Europe ( Bowler et al. 2015 ); investigation of drinking patterns in US counties from 2002 to 2012 ( Dwyer-Lindgren et al. 2015 ); resistance and resilience of terrestrial birds in drying climates ( Selwood et al. 2015 ); cluster analysis of population amyotrophic lateral sclerosis risk ( Rooney et al. 2015 ); malaria infection in Africa ( Noor et al. 2014 ); effects of fragmentation on infectious disease dynamics ( Jousimo et al. 2014 ); soil-transmitted helminth infection in sub-Saharan Africa ( Karagiannis-Voules et al. 2015 ); analysis of the effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015 ( Bhatt et al. 2015 ); adaptive prior weighting in generalized regression ( Held & Sauter 2016 ); analysis of hand, foot, and mouth disease surveillance data in China ( Bauer et al. 2016 ); estimation of the biomass of anchovies in the coast of Perú ( Quiroz et al. 2015 ); and many others.

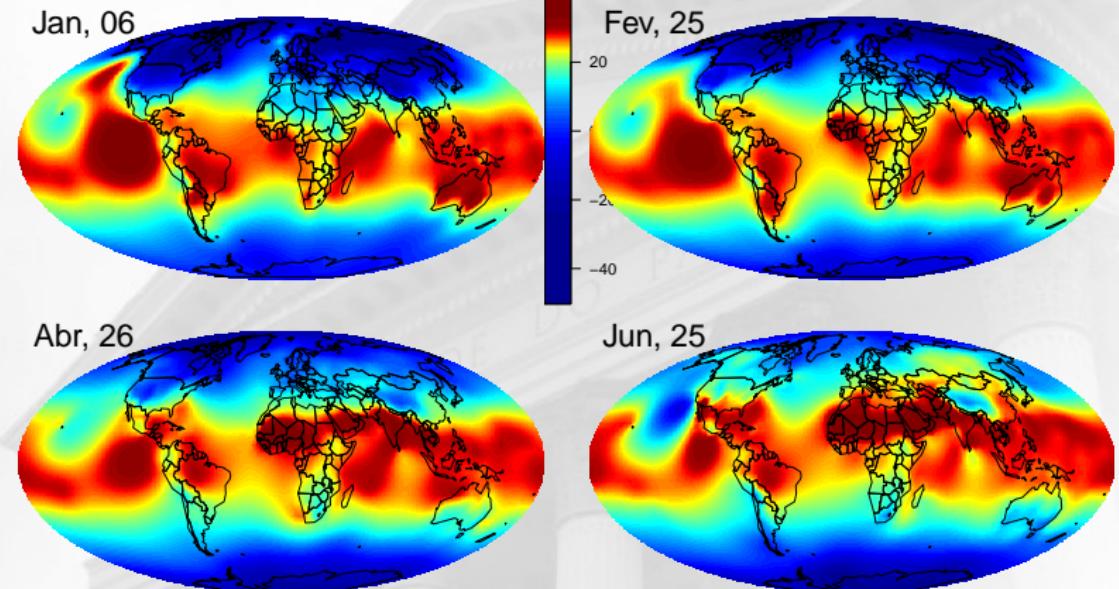
# Efeito de deprivação, Ribeiro et al. (2018)



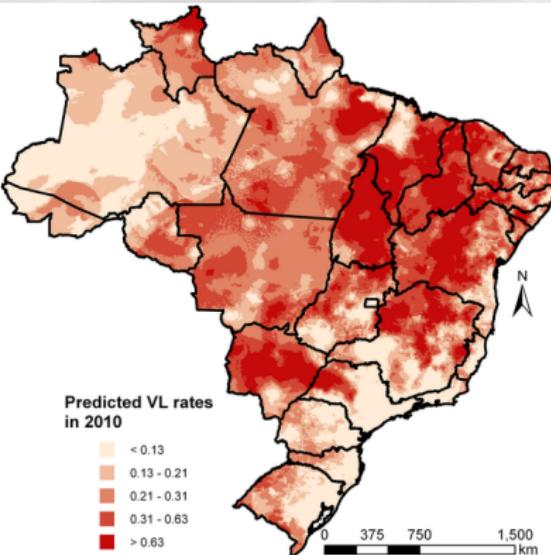
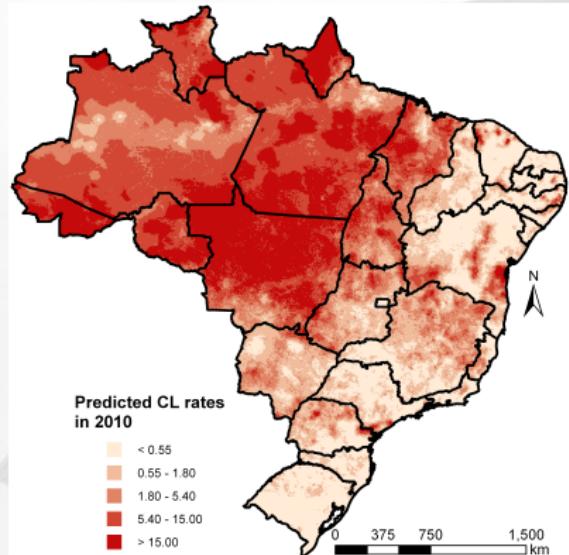
# Sobrevida: mapa de fragilidade



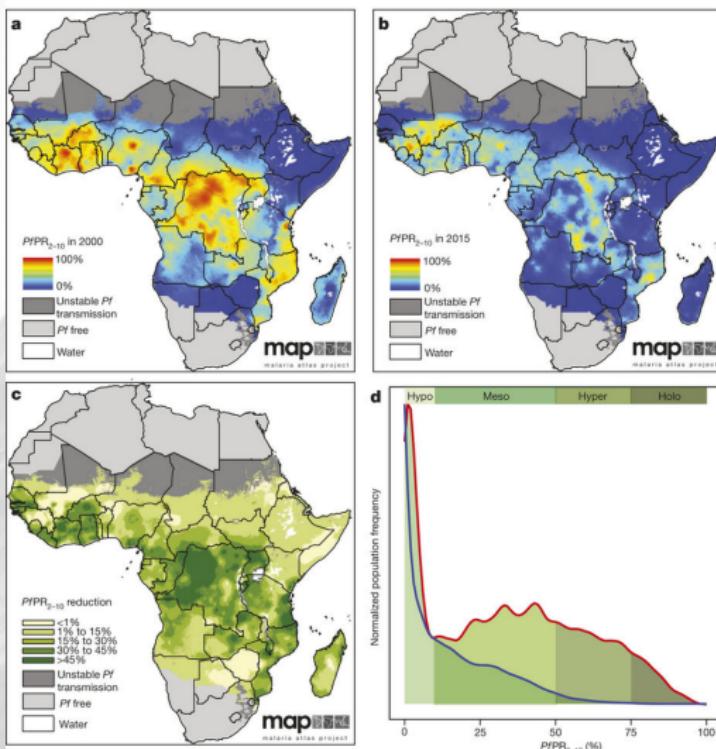
# Modelagem espaço temporal não separável no globo, E. T. Krainski (2018)



# Leishmaniasis in Brazil, Karagiannis-Voules et al. (2013)



# Malaria in Africa, Gething (2015)



## Referências

# Referências

- Gething, S. Bhatt AND D. J. Weiss AND E. Cameron AND D. Bisanzio AND B. Mappin AND U. Dalrymple AND K. E. Battle AND C. L. Moyes AND A. Henry AND P. A. Eckhoff AND E. A. Wenger AND O. Briët AND M. A. Penny AND T. A. Smith AND A. Bennett AND J. Yukich AND T. P. Eisele AND J. T. Griffin AND C. A. Fergus AND M. Lynch AND F. Lindgren AND J. M. Cohen AND C. L. J. Murray AND D. L. Smith AND S. I. Hay AND R. E. Cibulskis AND P. W. 2015. "The Effect of Malaria Control on Plasmodium Falciparum in Africa Between 2000 and 2015." *Nature*, no. 526 (October): 207–11.
- Karagiannis-Voules, D-A, R. G. C. Scholte, L. H. Guimarães, J. Utzinger, and P. Vounatsou. 2013. "Bayesian Geostatistical Modeling of Leishmaniasis Incidence in Brazil." *PLOS Neglected Tropical Diseases*, no. 5.
- Krainski, Elias T. 2018. "Statistical Analysis of Space-Time Data: New Models and Applications." PhD thesis, Norwegian University of Science; Technology.
- Ribeiro, A. I., E. T. Krainski, M. S. Carvalho, G. Launoy, C. Pernet, and M. F. de Pina. 2018. "Does Community Deprivation Determine Longevity After the Age of 75? A Cross-National Analysis." *International Journal of Public Health*, 1–11.
- Rue, H., and L. Held. 2005. *Gaussian Markov Random Fields: Theory and Applications*. Monographs on Statistics & Applied Probability. Boca Raton: Chapman; Hall.
- Rue, H., S. Martino, and N. Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with Discussion)." *Journal of the Royal Statistical Society, Series B* 71 (2): 319–92.
- Rue, H., A. I. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. 2017. "Bayesian Computing with Inla: A Review." *Annual Review of Statistics and Its Application* 4: 395–421.
- Simpson, D. P., H. Rue, T. G. Martins, A. Riebler, and S. H. Sørbye. 2017. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science* 32 (1): 1–28.
- Wood, S.N. 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman; Hall/CRC.