Métodos Computacionais para Inferência Estatística

Capítulo 6 - Tópicos Adicionais - *Clonagem de Dados*

Paulo Justiniano Ribeiro Jr.
Wagner Hugo Bonat
Elias Teixeira Krainski
Walmes Marques Zeviani

LEG: Laboratório de Estatística e Geoinformação
Universidade Federal do Paraná

20º *SINAPE*, 30-31/07/2012

Laboratório de Estatística e Geoinformação

Motivação

Clonagem de dados (data cloning)

- Lele (2007) e Lele (2010)
- (Outras referências)
- Implementação: algoritmos MCMC
- Pacote **R**: *clone* (Sólymos, 2010)
- utiliza JAGS, BUGS ou OpenBUGS

Intuição

Clonagem de dados (data cloning)

- Busca conciliar obtenção da verossimilhança com flexibilidade dos algoritmos de Inf. Bayesiana (MCMC)
- Dados "abundantes" *dominam* a priori
- *Posteriori* reflete a verossimilhança
- Proposta: **Clonar dados** (K vezes): preserva informação essencial
  - Média da *posteriori*: converge para EMV
  - K vezes a variância de *posteriori*: converge para variância assintótica do MLE
- Especificacão de modelos e identificabilidade

Laboratório de Estatística e Geoinformação

Construção - Genérica

- Modelo hierárquico (Bayesiano)

  1. $[Y|\underline{b}, X] \sim f(\underline{\mu}, \phi)$
  2. $g(\underline{\mu}) = X\underline{\beta} + Z\underline{b}$
  3. $\underline{b} \sim NMV(\underline{0}, \Sigma)$.
  4. *priori*: $[\beta, \Sigma, \phi]$.

Posteriori informa sobre verossimilhança clonada $L^K(\boldsymbol{\beta}, \Sigma, \phi)$

$$\pi^K(\boldsymbol{\beta}, \Sigma, \phi|y_{ij}) = \frac{[\int f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i]^K \pi(\boldsymbol{\beta})\pi(\Sigma)\pi(\phi)}{C(K; y_{ij})}$$

$$C(K; y_{ij}) = \int [\int f_i(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma, \phi) f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i]^K \pi(\boldsymbol{\beta})\pi(\Sigma)\pi(\phi) d\boldsymbol{\beta} d\Sigma d\phi$$

Passos

- Especificar modelo completo (Bayesiano)
- Clonar dados (K vezes)
- MCMC em dados clonados
- Repetir para diferentes K
- Verificar comportamento
- Resumos da *posteriori* informam sobre:
  - $L^K(\cdot)$
  - $L(\cdot)$ (assintoticamente)

Algoritmo

1. Dados K-clonado $\underline{Y}^k = (\underline{Y}, \underline{Y}, \ldots, \underline{Y})$
2. Gerar amostras (MCMC) da *posteriori* $[\beta, \Sigma, \phi]$ utilizando dados clonados $\underline{Y}^k$
   - Gere estado atual $(\beta, \Sigma, \phi)*$ de $[\beta, \Sigma, \phi]$
   - Gere K valores dos efeitos aleatórios $\underline{b}$, digamos $\underline{b}^1, \underline{b}^2, \ldots, \underline{b}^K$ de $[\underline{b}|\theta*]$.
   - Calcule $q^* = f(\underline{y}|\underline{b}^1, \phi*) f(\underline{y}|\underline{b}^2, \phi*), \ldots, f(\underline{y}|\underline{b}^K, \phi*)$ e faça $q_1 = q^*$
   - Repita (a) e (b) obtendo novos valores $(\beta, \Sigma, \phi)^{@}$ e $q^{@}$.
   - Gere uma $U(0,1)$ e calcule $p = \min(1, \frac{q^{@}}{q_1})$. Se $U > p$ $(\beta, \Sigma, \phi)_{j+1} = (\beta, \Sigma, \phi)_j$ caso contrário $(\beta, \Sigma, \phi)_{j+1} = (\beta, \Sigma, \phi)^{@}$.
   - Repita (d) e (e) **muitas** vezes.
3. Calcule as médias e as variâncias amostrais para $(\theta, \phi)_j$.

Laboratório de Estatística e Geoinformação

Identificabilidade

1. Estudar a identificabilidade de modelos em geral é não trivial.
2. Por vezes modelos são ajustados como se fossem identificáveis.
3. A atribuição de *priori's* pode tornar um modelo 'identificável'.

Sob clonagem dos dados:

1. Parâmetros são não-estimáveis: *posteriori* converge para a *priori* truncada no espaço de não-identificabilidade dos parâmetros quando aumenta-se o número de clones
2. Maior autovalor da matriz de variância-covariância a *posteriori* não converge para 0.
3. Se a variância à *posteriori* de um parâmetro converge para 0 quando aumentamos o número de clones, ele é estimável.

Exemplo I: Poisson com efeito aleatório

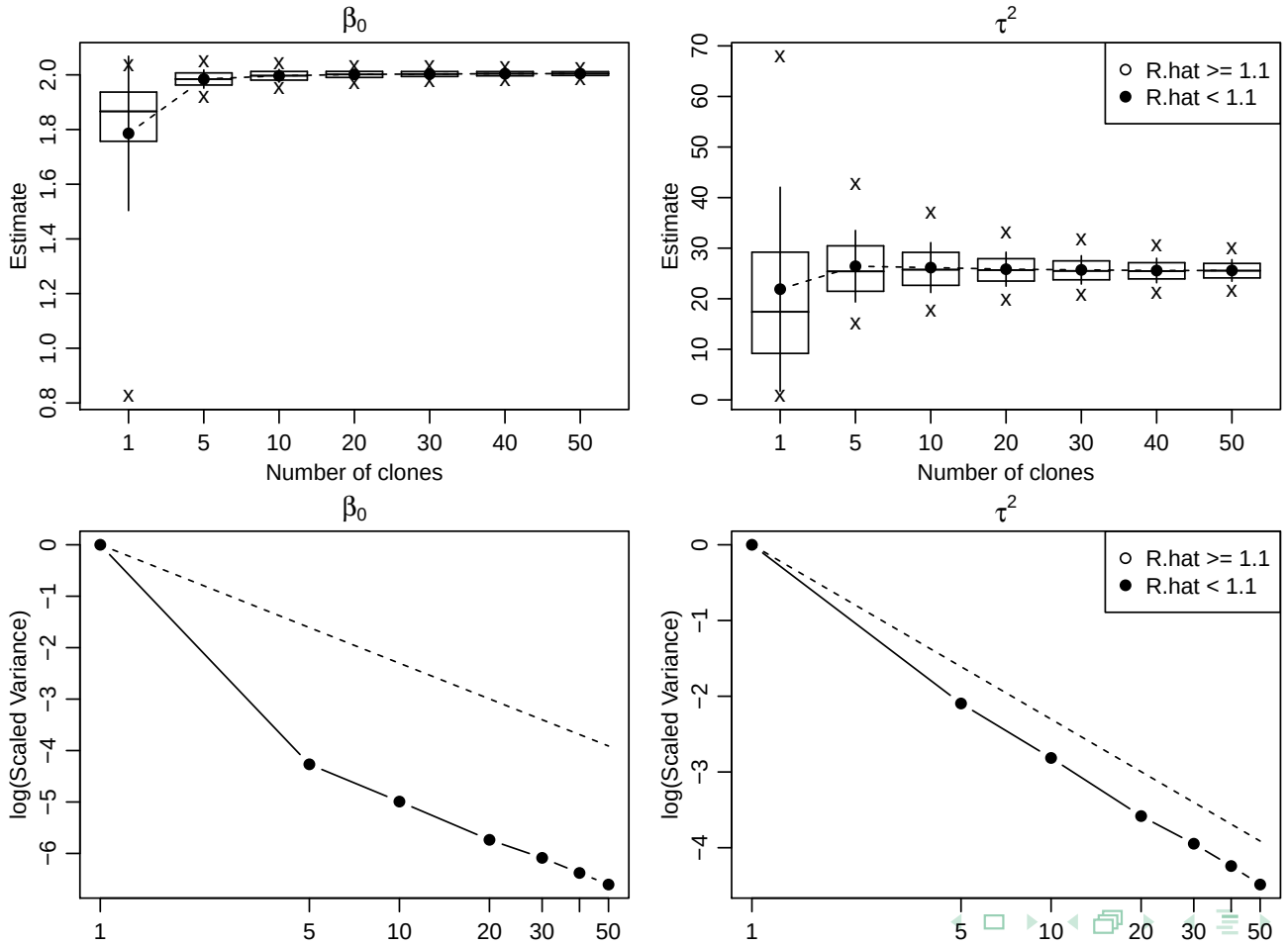
1. Modelo:
   - $Y_{ij}|b_i \sim P(\lambda_i)$
   - $log(\lambda_i) = \beta_0 + b_i$
   - $b_i \sim N(0, 1/\tau^2)$
   - $\tau^2 \sim G(1;0{,}1)$

2. Código

```
mod.poisson <- function(){
  for(j in 1:n.ua){
        for(i in 1:n.rep){
                Y[j,i] ~ dpois(lambda[j,i])
        log(lambda[j,i]) <- beta0 + b[j]
    }
    b[j] ~ dnorm(0,tau)
  }
  beta0 ~ dnorm(0, 0.001)
  tau   ~ dgamma(1, 0.01)
}
```

Exemplo I: Poisson com efeito aleatório (cont)

```
dados.list <- list(Y = t(matrix(dados$y, 10, 10)), n.ua = 10, n.rep = 10)
clone<-dc.fit(data=dados.list, model=mod.poisson, params=c("beta0","tau"),
      n.clones=c(1,5,10,20,30,40,50), multiply="n.ua", unchanged="n.rep",
            n.iter= 10000, n.adapt = 500, n.update = 500, thin =5)
```



Exemplo II: Normal sem replicações

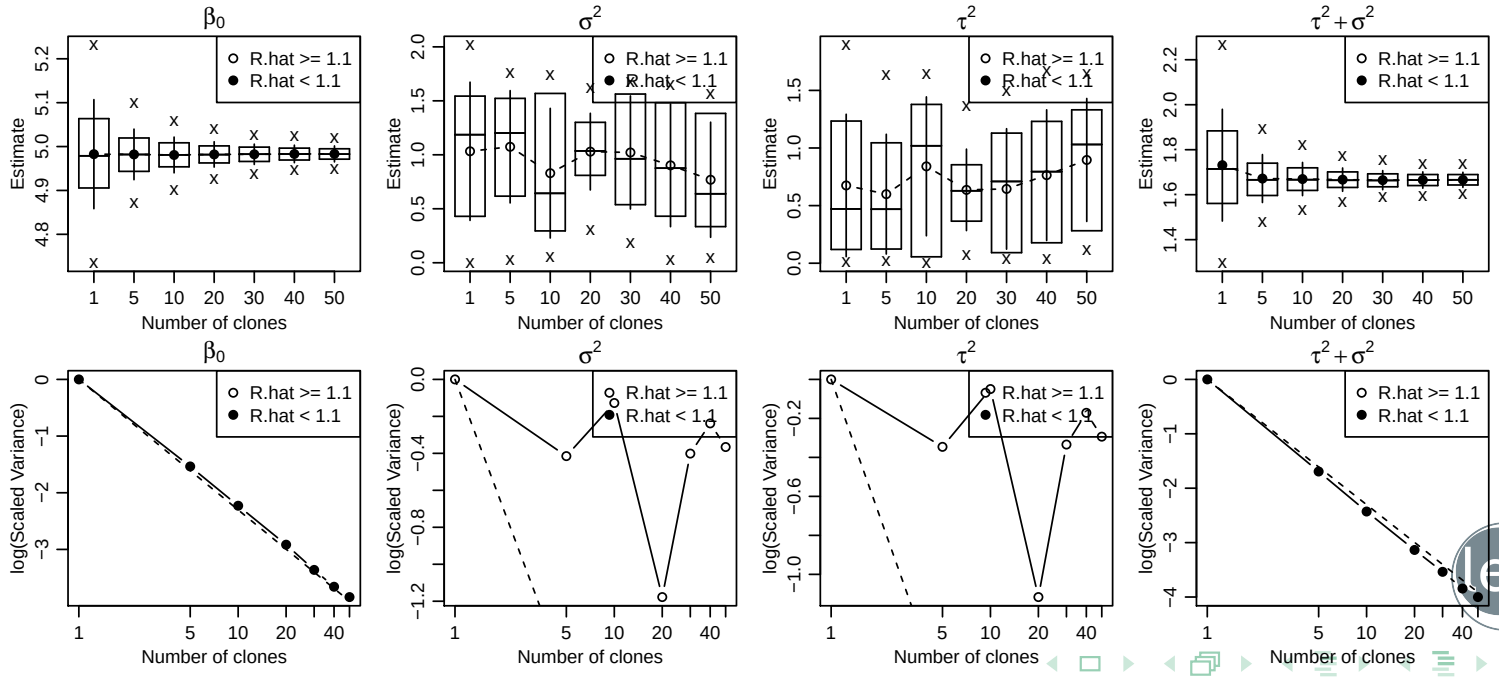
1. Modelo:
   - $Y_i \sim N(\mu_i, 1/\sigma^2)$
   - $log(\mu_i) = \beta_0 + b_i$
   - $b_i \sim N(0, 1/\tau^2)$
   - $\sigma^2 \sim G(0{,}5; 0{,}5)$; $\tau^2 \sim G(0{,}5; 0{,}5)$
2. Código

```
model.normal <- function(){
  for(i in 1:n){
    Y[i] ~ dnorm(mu[i], 1/sigma2)
    mu[i] <- b0 + b[i]
    b[i] ~dnorm(0, 1/tau2)
  }
  b0 ~ dnorm(0, 0.01)
  sigma2 ~ dgamma(0.5, 0.5)
  tau2 ~ dgamma(0.5, 0.5)
  soma <- sigma2 + tau2
}
```

Exemplo II: Normal sem replicações (cont)

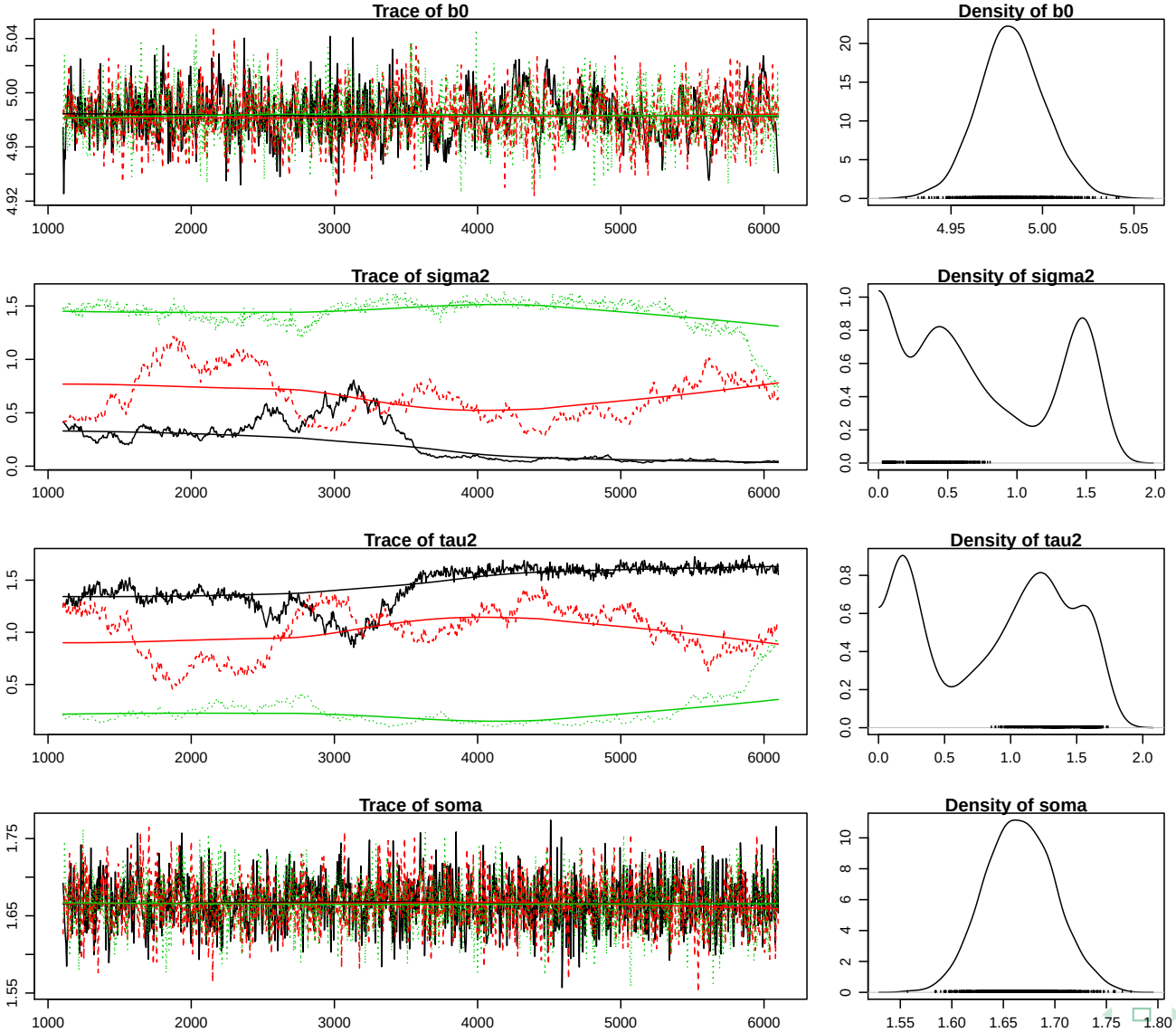
```
k <- c(1,5,10,20,30,40,50)
Gclone1 <- dc.fit(data = dat.gauss, params= c("b0","tau2", "sigma2"),
                  model = model.normal, n.clones=k, multiply="n",
                  n.iter= 5000, n.adapt = 1000, n.update = 100, thin = 5)
Gclone2 <- dc.fit(data = dat.gauss, params= c("soma"),
                  model = model.normal, n.clones=k, multiply="n",
                  n.iter= 5000, n.adapt = 1000, n.update = 100, thin = 5)
```

$$\phi = \sigma^2 + \tau^2$$



Exemplo II: Normal sem replicações (cont)

Cadeias MCMC:



Exemplo III: Normal com replicações

1. Modelo:
   - $Y_{ij} \sim N(\mu_i, 1/\sigma^2)$
   - $log(\mu_i) = \beta_0 + b_i$
   - $b_i \sim N(0, 1/\tau^2)$
   - $\sigma^2 \sim G(1; 0{,}01)$; $\tau^2 \sim G(1; 0.01)$

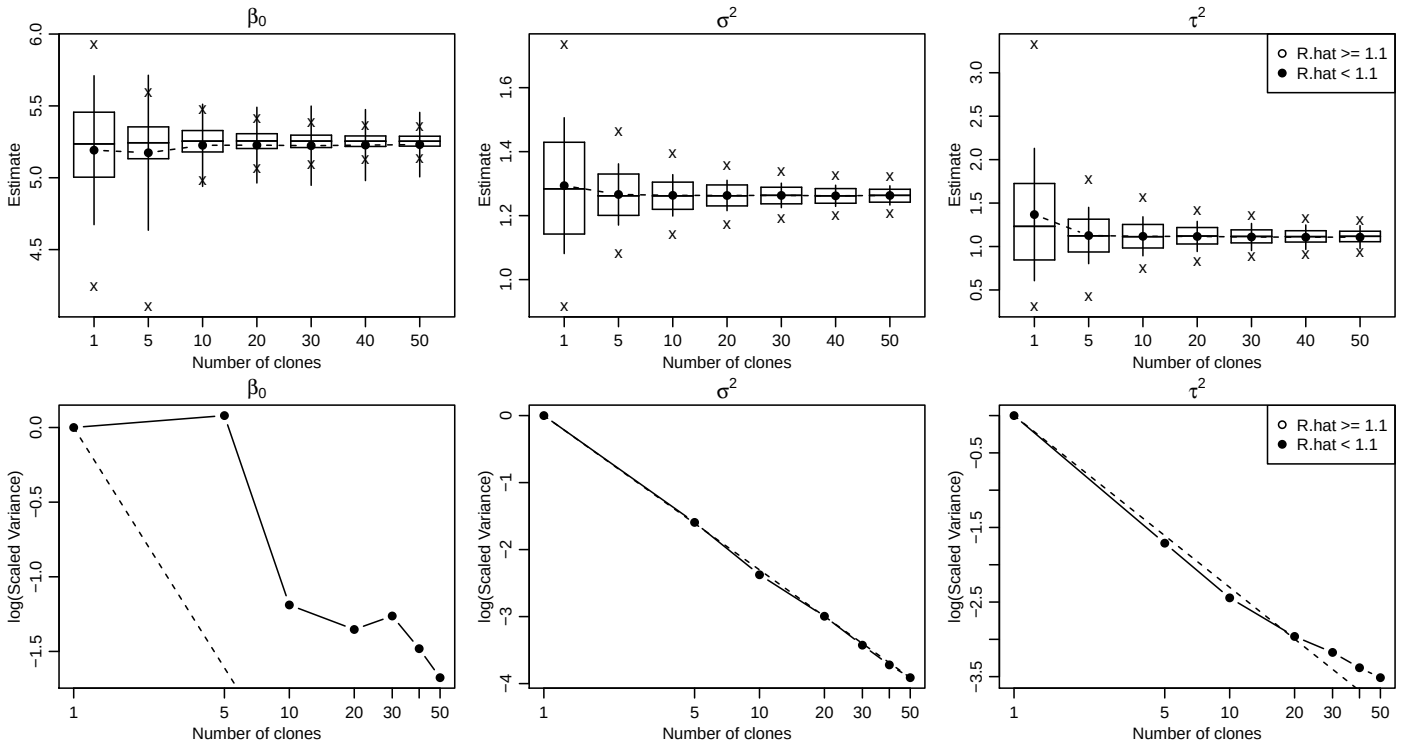
2. Código

```
mod.gauss.rep <- function(){
  for(j in 1:n.ua){
        for(i in 1:n.rep){
                Y[j,i] ~ dnorm(mu[j,i], sigma2)
                mu[j,i] <- beta0 + b[j]
     }
     b[j] ~ dnorm(0,tau2)
  }
  beta0 ~ dnorm(0,0.01)
  tau2 ~ dgamma(1,0.01)
  sigma2 ~ dgamma(1,0.01)
}
```

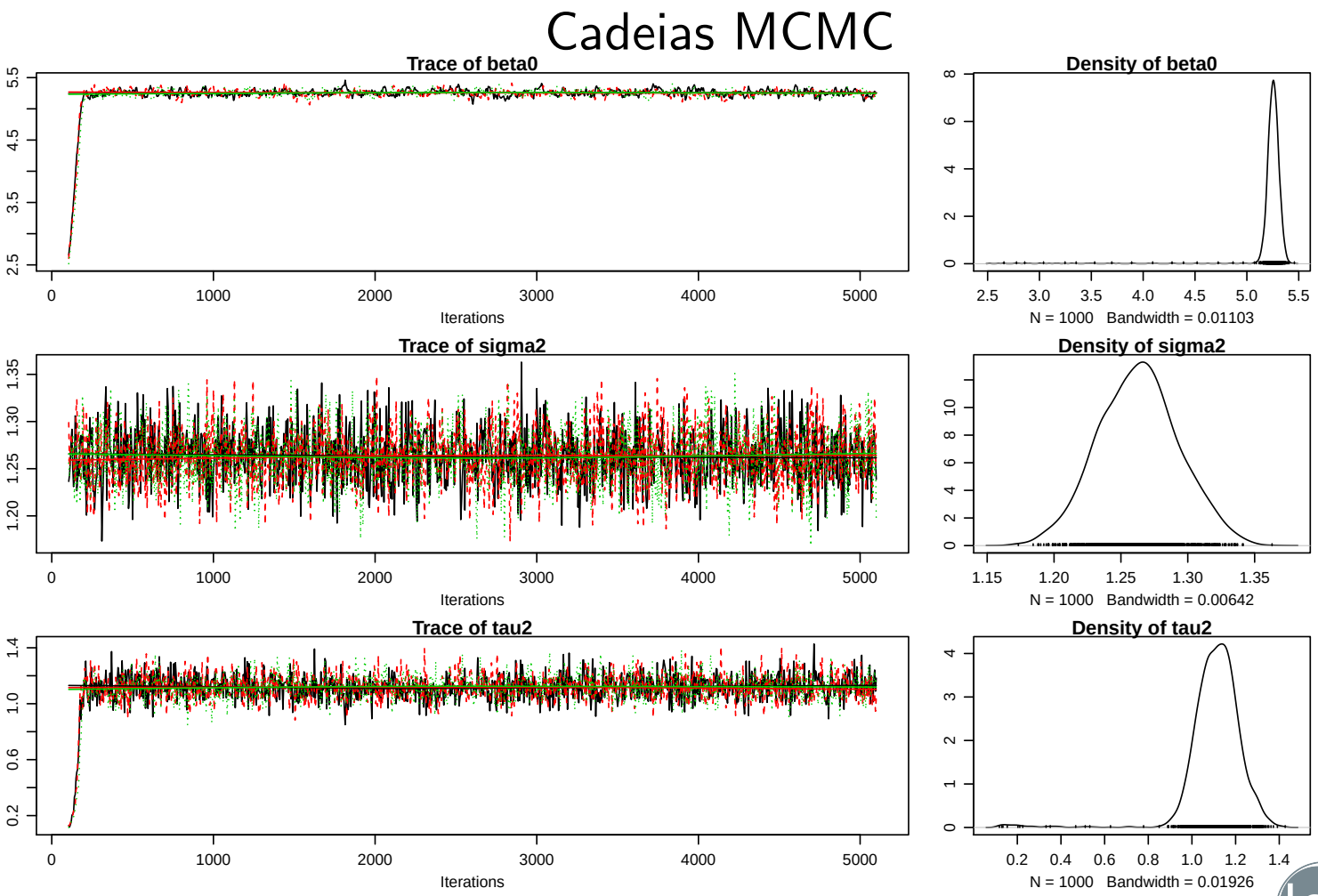
tório de Estatística
nformação

Exemplo III: Normal com replicações (cont)

```
GcloneR<- dc.fit(data=dados.list, params= c("beta0","tau2", "sigma2"),
                 model = mod.gauss.rep, n.clones=c(1,5,10,20,30,40,50),
                 multiply="n.ua", unchanged = "n.rep",
                 n.iter= 5000, n.adapt = 1000, n.update = 100, thin = 5)
```



Exemplo III: Normal com replicações (cont)



Laboratório de Estatística e Geoinformação

Exemplo III: Normal com replicação (cont)

Cuidado com as interpretações das saídas!

```
summary(GcloneR)

Iterations = 105:5100
Thinning interval = 5
Number of chains = 3
Sample size per chain = 1000
Number of clones = 50

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

        Mean      SD  DC SD  Naive SE Time-series SE R hat
beta0  5.231 0.22414 1.5849 0.0040922      0.0165402 1.007
sigma2 1.263 0.03004 0.2124 0.0005484      0.0005426 1.000
tau2   1.108 0.13144 0.9294 0.0023998      0.0075793 1.000

2. Quantiles for each variable:

         2.5%   25%   50%   75% 97.5%
beta0  5.1332 5.220 5.254 5.289 5.354
sigma2 1.2046 1.242 1.263 1.282 1.322
tau2   0.9265 1.055 1.117 1.176 1.301
```

Exemplo IV: Regressão Beta com efeitos aleatórios

Bibliografia

Lele, S.; Dennis, B. ; Lutscher, F.
Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods.
Ecology Letter 10: 551-563 (2007)

Solymos, P.
dclone: Data Cloning in R.
The R Journal 2: 29-37 (2010)

Lele, S. ; Nadeem, K. ; Schmuland, B.
Estimability and Likelihood Inference for Generalized Linear Mixed Models Using data Cloning.
Journal of the American Statistical Association 105:1617-1625 (2010)

Laboratório de Estatística e Geoinformação