Inferência Bayesiana e algoritmos MCMC em Modelos Hierárquicos

Prof. Caio Azevedo

Visão geral: inferência Bayesiana (IB)

- Diversos conteúdos sobre inferência Bayesiana podem ser encontrados aqui, aqui, aqui, aqui e aqui.
- Essencialmente, toda a inferência Bayesiana (diferentemente da frequentista) baseia-se na chamada distribuição a posteriori ("estimador geral").
- Em IB, toda quantidade desconhecida não obversável (incluindo os parâmetros) é tratada como uma variável aleatória.
- Seja θ (possívelmente um vetor) um parâmetro associado a um determinado modelo (de regressão) $p(\boldsymbol{x}|\theta)$ (verossimilhança) para o qual assumimos uma priori $p(\theta)$.

Visão geral: inferência Bayesiana (IB)

- Posteriori: $p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int_\Theta p(\boldsymbol{x}|\theta)p(\theta)d\theta}$ (se Θ for um conjunto discreto, a integral é substituida por um somatório).
- Problema (nos modelos hierárquicos): é muito complicado obter as posterioris marginais de interesse, de forma analítica. Se $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ elas seriam dadas por $p(\theta_1|\boldsymbol{x}) = \int_{\theta_1} p(\boldsymbol{\theta}|\boldsymbol{x})d\theta_1$ e $p(\theta_2|\boldsymbol{x}) = \int_{\theta_2} p(\boldsymbol{\theta}|\boldsymbol{x})d\theta_2$, respectivamente.
- Distribuições condicionais completas: $p(\theta_1|\theta_2, \boldsymbol{x})$ e $p(\theta_2|\theta_1, \boldsymbol{x})$.
- Uma alternativa é maximizar a posteriori (marginal) dos parâmetros.

Aplicação de IB nos modelos estudados (até o momento)

- MLH (dos níveis) : parâmetros de interesse $(\boldsymbol{\gamma}^t, \sigma^2, \boldsymbol{\theta}^t, \boldsymbol{u}^t)^t$.
- MLGH : parâmetros de interesse $(\boldsymbol{\gamma}^t, \boldsymbol{\theta}^t, \phi, \boldsymbol{u}^t)^t$.
- Usualmente, em IB, trabalha-se com a verossimilhança completa (original) ao invés da marginal, a qual considera, também, os efeitos aleatórios ($\boldsymbol{u}$).
- Para outros modelos, como por exemplo, os modelos não lineares (normais/normais) hierárquicos e os modelos não lineares generalizados hierárquicos, outros parâmetros tem de ser considerados.

Aplicação de IB nos modelos estudados

- Mesmo sob modelos mais simples (aqui) não é possível obter as distribuições a posteriori marginais de interesse (por conta da estruturação das matrizes de covariância).
- Nem mesmo as algumas das chamadas distribuições condicionais completas, possuem forma conhecida.
- Uma forma de obter condicionais completas com forma conhecida pode ser vista em: aqui.

Posteriores: MLH de dois níveis

Seja $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}^t, \sigma^2, \boldsymbol{\theta}^t, \boldsymbol{u}^t)^t$ (mais geralmente, é o conjunto com todas as quantidades a serem estimadas). Assim:

$$
\begin{aligned}
p(\boldsymbol{\gamma}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\vartheta})p(\boldsymbol{u}|\boldsymbol{\Psi})p(\boldsymbol{\gamma})p(\sigma^2)p(\boldsymbol{\theta}) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{J}(\boldsymbol{y}_j - \boldsymbol{Z}_j\boldsymbol{\gamma} - \boldsymbol{X}_j\boldsymbol{u}_j)'(\boldsymbol{y}_j - \boldsymbol{Z}_j\boldsymbol{\gamma} - \boldsymbol{X}_j\boldsymbol{u}_j)\right\} \\
&\times (\sigma^2)^{-\sum_{j=1}^{J} n_j/2} p(\boldsymbol{u}|\boldsymbol{\Psi})p(\boldsymbol{\gamma})p(\sigma^2)p(\boldsymbol{\theta})
\end{aligned}
$$

Posteriores: MLH de dois níveis

Em que $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\theta})$. Posteriores marginais: calcular a integral para cada parâmetro (quantidade a ser estimada) em $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}^t, \sigma^2, \boldsymbol{\theta}^t, \boldsymbol{u}^t)^t$, com relação as demais componentes, ou seja:

$$p(\boldsymbol{\vartheta}_k|\boldsymbol{y}) = \int_{\boldsymbol{\vartheta}_{(-k)}} p(\boldsymbol{\gamma}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{\vartheta}_{(-k)}$$

em que $\boldsymbol{\vartheta}_{(-k)}$ corresponde ao vetor $\boldsymbol{\vartheta}$ sem o parâmetros $k \in \{\boldsymbol{\gamma}^t, \sigma^2, \boldsymbol{\theta}^t, \boldsymbol{u}^t\}$

Posteriores: MLGH de dois níveis

$$
\begin{aligned}
p(\boldsymbol{\gamma}, \boldsymbol{\theta}, \phi, \boldsymbol{u}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\vartheta}) p(\boldsymbol{u}|\boldsymbol{\Psi}) p(\boldsymbol{\gamma}) p(\boldsymbol{\theta}) p(\phi) \\
&\propto \exp\left\{ \phi \sum_{j=1}^{J} \left[\sum_{i=1}^{n_j} y_{ji}\theta_{ji} - b(\theta_{ji}) \right] + \sum_{j=1}^{J} \sum_{i=1}^{n_j} c(y_{ji}, \phi) \right\} \\
&\times p(\boldsymbol{u}|\boldsymbol{\Psi}) p(\boldsymbol{\gamma}) p(\boldsymbol{\theta}) p(\phi)
\end{aligned}
$$

Posteriores: MLGH de dois níveis

Em que $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\theta})$ e $\theta_{ji} = h(\mu_{ji}) = \mu_{ji} = g^{-1}(\boldsymbol{Z}'_{ji}\gamma + \boldsymbol{X}'_{ji}\boldsymbol{u}_j)$. Posteriores marginais: calcular a integral para cada parâmetro em $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}^t, \sigma^2, \boldsymbol{\theta}^t, \boldsymbol{u}^t, \phi)^t$, com relação as demais componentes, ou seja:

$$p(\boldsymbol{\vartheta}_k|\boldsymbol{y}) = \int_{\boldsymbol{\vartheta}_{(-k)}} p(\gamma, \boldsymbol{\theta}, \boldsymbol{u}, \phi|\boldsymbol{y}) d\boldsymbol{\vartheta}_{(-k)}$$

Prioris

- $\gamma \sim N_p(\boldsymbol{\mu}_\gamma, \boldsymbol{\Psi}_\gamma)$.
- $\sigma^2 \sim IG(\alpha_{\sigma^2}, \beta_{\sigma^2})$.
- $\phi \sim IG(\alpha_\phi, \beta_\phi)$.
- $\boldsymbol{u}_j \sim N_p(0, \boldsymbol{\Psi})$ (outras opções como t de Student multivadairada, normal assimétrica multivariada, mistura de escala (skew) normal).
- $\boldsymbol{\Psi} \sim IW_q(\nu_\Psi, \boldsymbol{W}_\Psi)$ (inversa-Wishart). Caso essa matriz seja estruturada (por exeplo Uniforme, Hankel etc veja aqui) podemos usar prioris específicas para cada componente:AR(1) - $\sigma^2 \sim$ gama$(\alpha_{\sigma^2}, \beta_{\sigma^2})$, $\rho \sim U(0,1)$; ARH(1) $\sigma_i^2 \sim$ gama$(\alpha_{\sigma_i^2}, \beta_{\sigma_i^2})$, $i = 1, 2, .., p$, mesma priori para ρ.

Obtenção das posterioris

- Como visto, não é possível obter, analiticamente, as distribuições marginais.
- Alternativa: emprego de algum método numérico.
  - Algoritmos MCMC: aqui, aqui, aqui, aqui, aqui e aqui.
  - INLA (Integrated Nested Laplace Approximation): aqui, aqui, aqui, aqui e aqui.
  - ABC (Approximate Bayesian Computation): aqui, aqui e aqui.
  - Maximização da posteriori (obtenção da moda a posteriori): aqui, aqui e aqui e aqui.
  - Expectation and Propagation/Variational Bayes: aqui e aqui.

Exemplo da normal bivariada (algoritmos MCMC)

- Seja $\boldsymbol{Y} = (Y_1, Y_2)' \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Desejamos estimar as fdp's marginais de Y_1, Y_2, com $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, conhecidos. Sabemos que $Y_i \sim N(\mu_i, \sigma_i^2), i = 1, 2$.
- Suponha, no entanto, que saibamos apenas que:

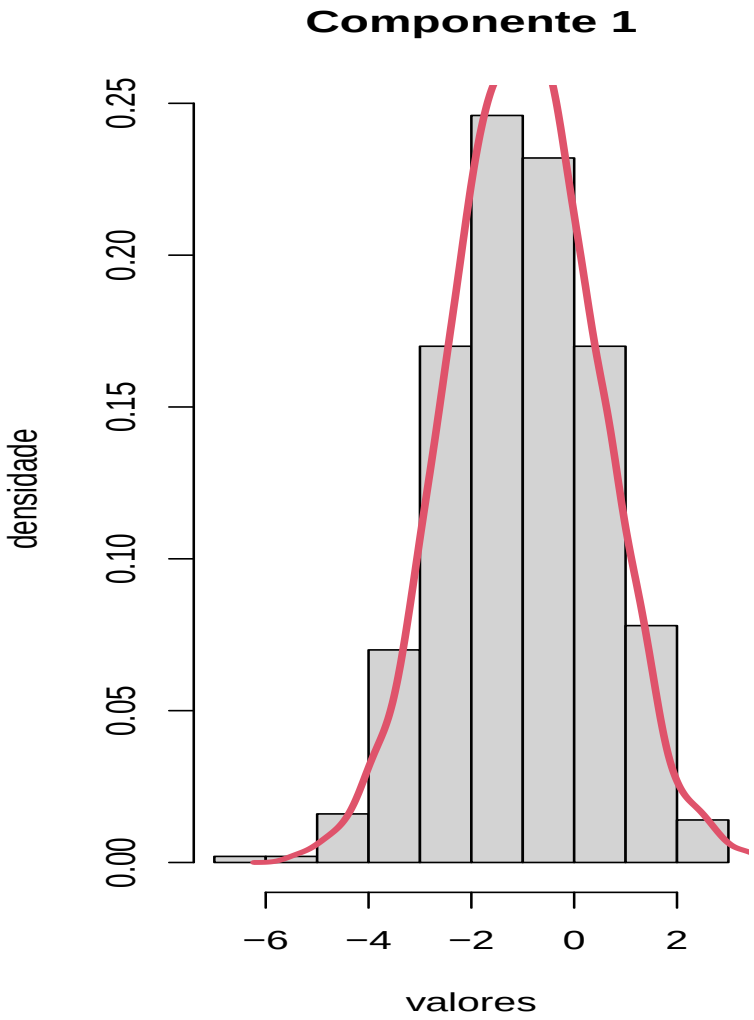
  $Y_1|y_2, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_1(\overline{\mu}_1, \overline{\sigma}_1^2)$, em que $\overline{\mu}_1 = \mu_1 + \sigma_{12} \left(\sigma_2^2\right)^{-1} (y_2 - \mu_2)$ e $\overline{\sigma}_1^2 = \sigma_1^2 - (\sigma_{12})^2 \left(\sigma_2^2\right)^{-1}$.

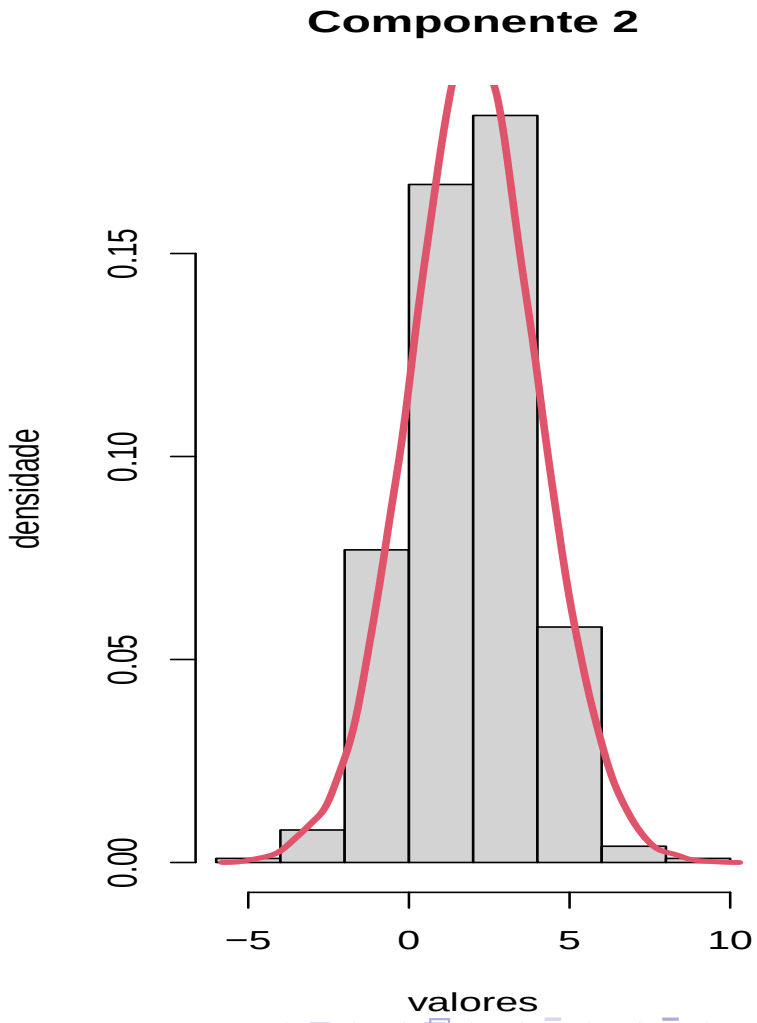
  $Y_2|y_1, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_1(\overline{\mu}_2, \overline{\sigma}_2^2)$, em que $\overline{\mu}_2 = \mu_2 + \sigma_{12} \left(\sigma_1^2\right)^{-1} (y_1 - \mu_1)$ e $\overline{\sigma}_2^2 = \sigma_2^2 - (\sigma_{12})^2 \left(\sigma_1^2\right)^{-1}$.
- As distribuições condicionais podem ser utilizadas para estimar as distribuições marginais? Veja aqui uma revisão sobre a distribuição normal multivariada.

Simulação iterativa

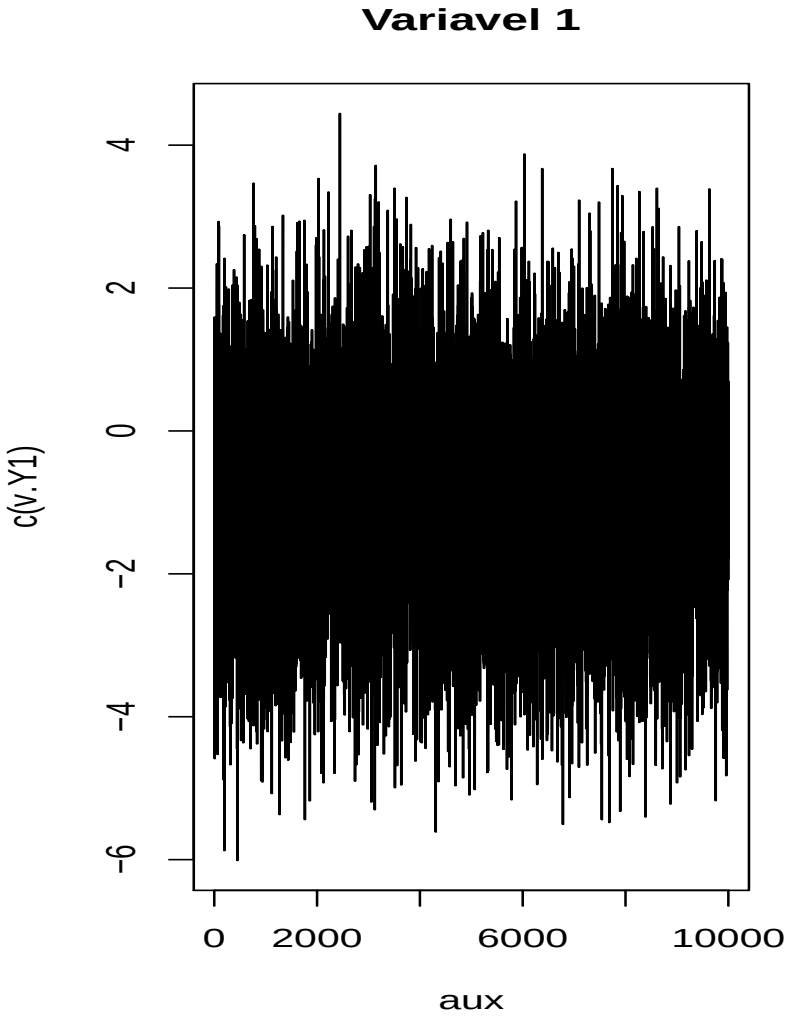
- Simule $y_1^{(1)} \sim N(0,1)$, por exemplo. Com este valor, simule $y_2^{(1)} | \left(y_1^{(1)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$. Agora, com $y_2^{(1)}$ simule $y_1^{(2)} | \left(y_2^{(2)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$.
- Repita o passo acima R vezes, obtendo-se $(y_1^{(1)}, y_2^{(1)}), (y_1^{(2)}, y_2^{(2)}), ..., (y_1^{(R)}, y_2^{(R)})$

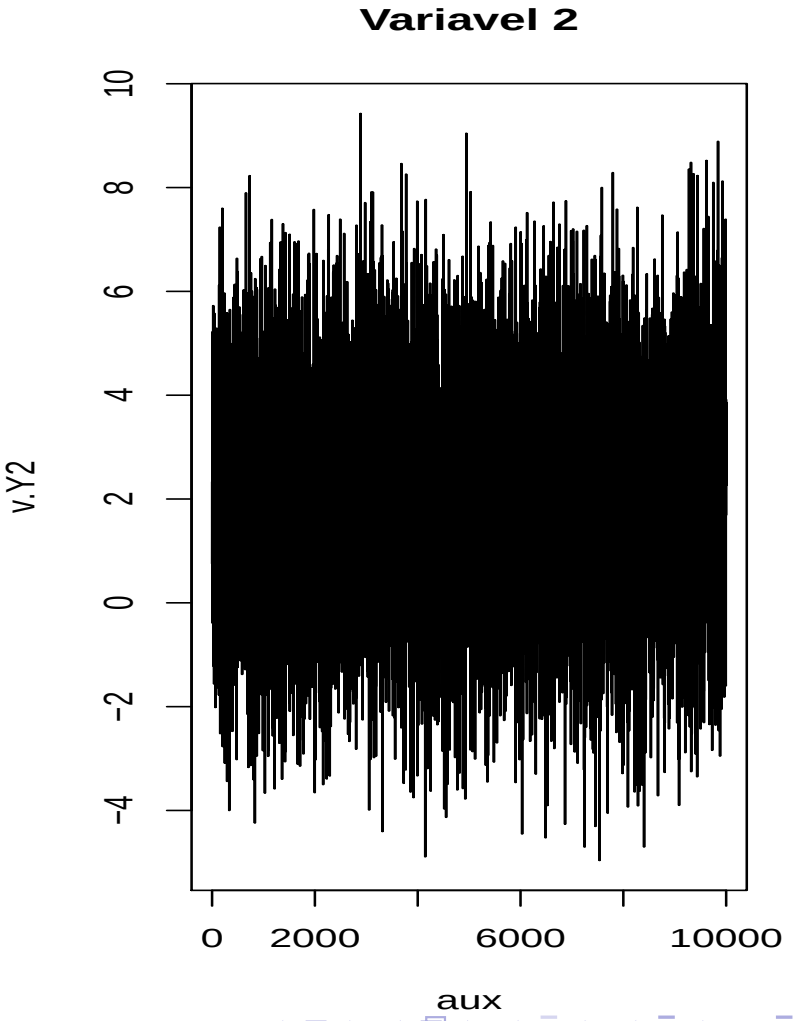
Histograma das amostras MCMC



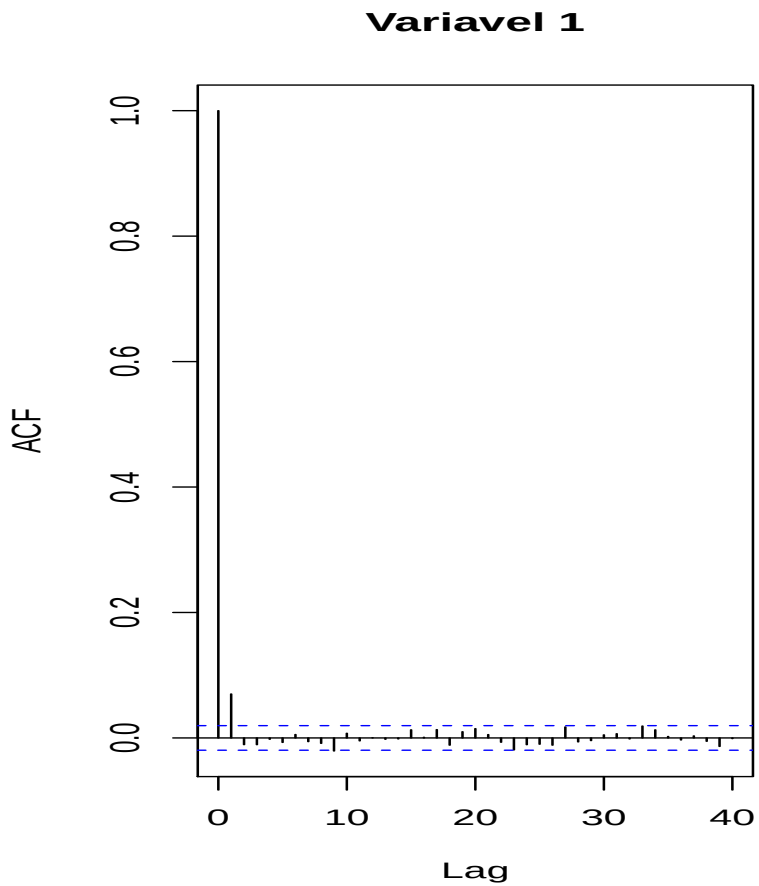


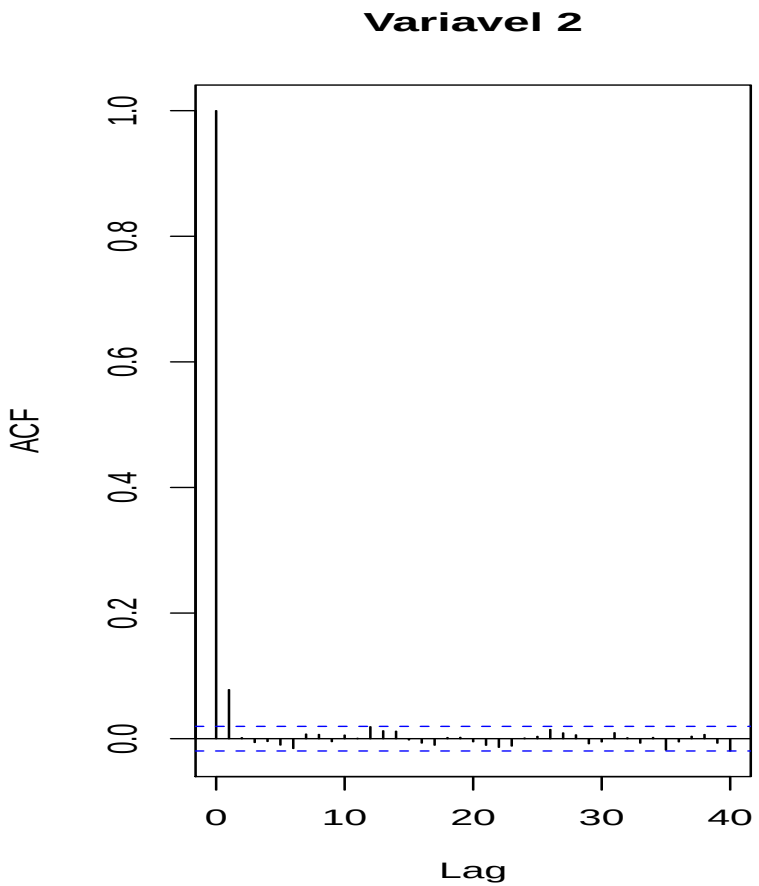
Gráficos das trajetórias (Trace plots) das amostras MCMC





ACF das amostras MCMC





Cadeias de Markov (link 1, link 2)

- Estudo de sequências (conjuntos) de variáveis aleatórias que guardam alguma estrutura de dependência entre si.
- Exemplo: seja $X_1, X_2, ...$ um sequência de Bernoullis que representam resultados de lançamentos de duas moedas (0, cara ; 1, coroa).
- Caso $X_i = 0$, lança-se uma moeda com probabilidade de cara igual a 0,55, caso contrário lança-se uma moeda com probabilidade de cara igual à 0,35.

Exemplo: lançamento de moedas

- Matriz de transição

$$\boldsymbol{P} = \begin{bmatrix} P(X_i = 0|X_{i-1} = 0) & P(X_i = 1|X_{i-1} = 0) \\ P(X_i = 0|X_{i-1} = 1) & P(X_i = 1|X_{i-1} = 1) \end{bmatrix}$$

$$\boldsymbol{P} = \begin{bmatrix} 0,55 & 0,45 \\ 0,35 & 0,65 \end{bmatrix}$$

- Distribuição estacionária ($\boldsymbol{x}$): $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x}$.

Exemplo: lançamento de moedas

- Resultado: Se a Cadeia de Markov (definida pela matriz de transição), for reversível, a distribuição estacionária existe e é única.
- Distribuição estacionária: Obtida através de observações da cadeia de Markov, $\boldsymbol{x} = \boldsymbol{P}^n$, para n suficiente grande.
- No exemplo em questão, temos que $\boldsymbol{x} = [0,45; 0,55]$

- No exemplo da normal bivariada, a matriz de transição (que possui um número infinito não enumerável de elementos) é definida pelas densidades de transição (distribuições condidionais completas):

$$Y_1|y_2, \boldsymbol{\mu}, \boldsymbol{\Sigma}; Y_2|y_1, \boldsymbol{\mu}, \boldsymbol{\Sigma}$$

- Neste caso, faz sentido calcular $P(Y_1 \in a|y_2, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ e $P(Y_2 \in b|y_1, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, em que a, b são intervalos da reta.
- Resultado: se a Cadeia de Markov, com matriz de transição definida pelas condicionais completas, for reversível, sua distribuição estacionária será única e convergirá para as distribuições marginais de interesse.

Algoritmos de Monte Carlo via Cadeias de Markov

- A família de algoritmos que permitem simular variáveis aleatórias (iterativamente) a partir das distribuições condicionais completas é conhecido como algoritmos de Monte Carlo via Cadeias de Markov (MCMC).
- Monte Carlo (resolver integrais) Cadeias de Markov (gera cadeias de Markov).
- Aplicações: obter distribuições (marginais, condicionais, conjuntas) de interesse. Em particular: Inferência Bayesiana.

Aplicações em Inferência Bayesiana

- Distribuição inversa gama: $X \sim IG(r, \gamma)$:

$$p(x) = \frac{\gamma^r}{\Gamma(r)} e^{-x/\gamma} x^{-(r+1)} \mathbb{1}_{(0,\infty)}(x)$$

- Distribuição normal-inversa gama: $(X, Y) \sim NIG(\mu, \nu, r, \gamma)$

$$\begin{aligned} X|y &\sim N(\mu, y/\nu) \\ Y &\sim IG(r, \gamma) \end{aligned}$$

Continuação

- Densidade conjunta

$$\begin{aligned} p(x,y) &= \frac{\sqrt{\nu}}{\sqrt{2\pi}} y^{-1/2} y^{-(r+1)} \exp\left(-\frac{\gamma}{y}\right) \exp\left\{-\frac{1}{2}\frac{\nu(x-\mu)^2}{\sigma^2}\right\} \\ &\times \quad \mathbb{1}_{(0,\infty)}(y)\mathbb{1}_{(-\infty,\infty)}(x) \end{aligned}$$

- Distribuição marginal de $X, X \sim t_{(2r)}\left(\mu, \sqrt{\frac{\gamma}{r\nu}}\right)$,

$$p(x) = \frac{\Gamma\left(\frac{2r+1}{2}\right)}{\Gamma\left(\frac{2r}{2}\right)\Gamma\left(\frac{1}{2}\right)} \left(\sqrt{2r}\delta\right)^{-1} \left[1 + \frac{(x-\mu)^2}{2r\delta^2}\right]^{-\frac{2r+1}{2}} \mathbb{1}_{(-\infty,\infty)}(x)$$

$$\delta^2 = \sqrt{\frac{\gamma}{r\nu}}$$

Aplicações em Inferência Bayesiana

- Seja $X_1|\boldsymbol{\theta}, ..., X_n|\boldsymbol{\theta}, \boldsymbol{\theta} = (\mu, \sigma^2)$ uma amostra aleatória de $X|\boldsymbol{\theta} \sim N(\mu, \sigma^2)$.
- Família conjugada (normal inversa gama)

$$\begin{aligned} \mu|\sigma^2 &\sim N(\alpha, \sigma^2/\kappa) \\ \sigma^2 &\sim \text{IG}(\gamma, \beta) \end{aligned}$$

Continuação

- Posteriori conjunta

$$\begin{aligned} \mu|\boldsymbol{x}, \sigma^2 &\sim N(\alpha^*, \sigma^2/\nu^*) \\ \sigma^2|\boldsymbol{x} &\sim \text{IG}(r^*, \gamma^*) \end{aligned}$$

em que $\nu^* = \nu + n$, $\gamma^* = \frac{1}{2}\left[\frac{n\nu}{n+\nu}(\overline{x} - \alpha)^2 + (n-1)s^2 + \gamma\right]$, $r^* = \frac{n}{2} + r$.

- Além disso, $\mu|\boldsymbol{x} \sim t_{(2r^*)}\left(\alpha^*, \sqrt{\frac{\gamma^*}{r^*\nu^*}}\right)$.
- Assim, as posterioris marginais são: $\mu|\boldsymbol{x} \sim t_{(2r^*)}\left(\alpha^*, \sqrt{\frac{\gamma^*}{r^*\nu^*}}\right)$ e $\sigma^2|\boldsymbol{x} \sim \text{IG}(r^*, \gamma^*)$.

Continuação

- Pode-se provar que (distribuições condicionais completas):

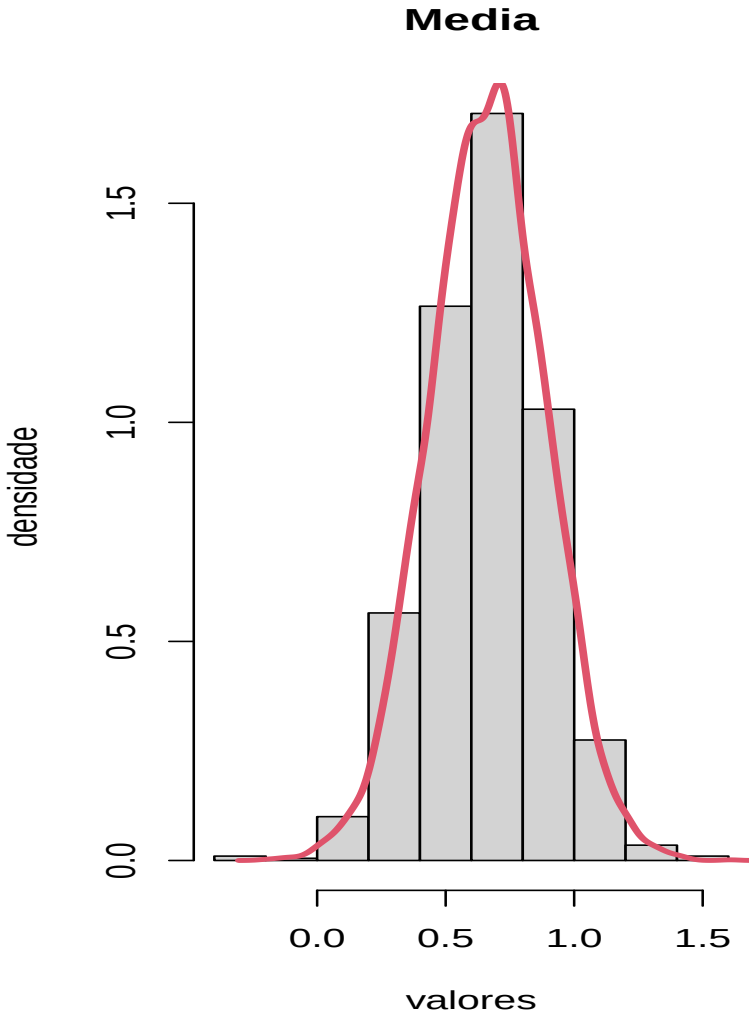
$$\begin{aligned} \mu|\boldsymbol{x}, \sigma^2 &\sim N(\alpha^*, \sigma^2/\nu^*) \\ \sigma^2|\boldsymbol{x}, \mu &\sim \text{IG}(r^{**}, \gamma^{**}) \end{aligned}$$

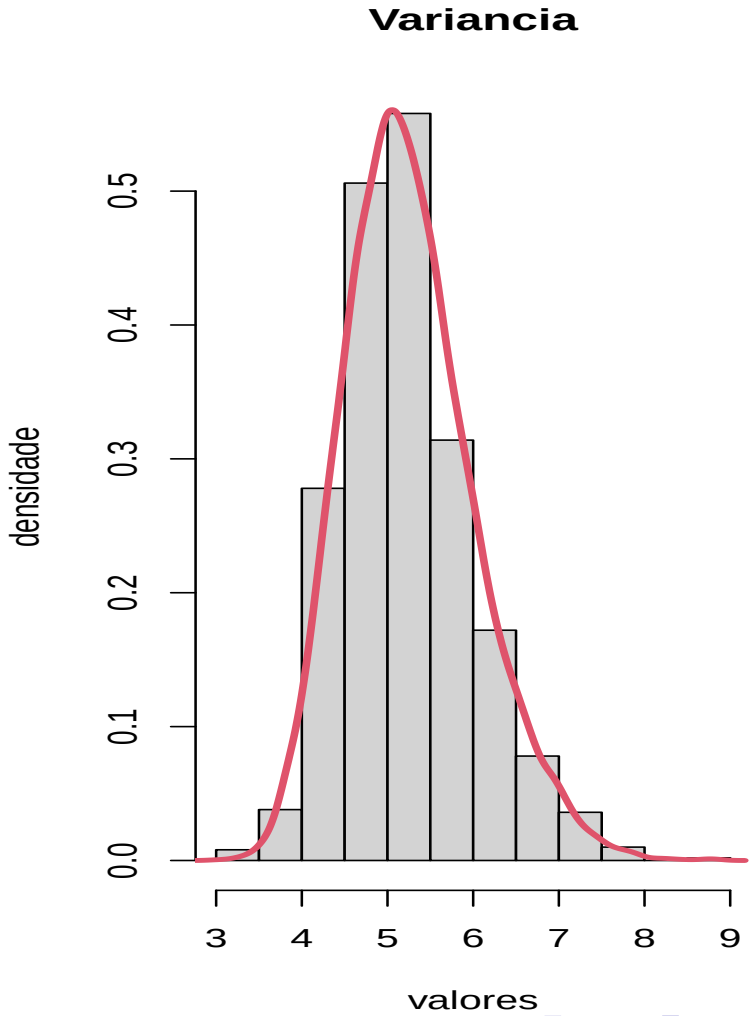
em que

$r^{**} = \frac{n+1}{2} + r; \gamma^{**} = \frac{n}{2}(\overline{x} - \mu)^2 + \frac{\nu}{2}(\mu - \alpha)^2 + \frac{1}{2}(n-1)s^2 + \gamma$

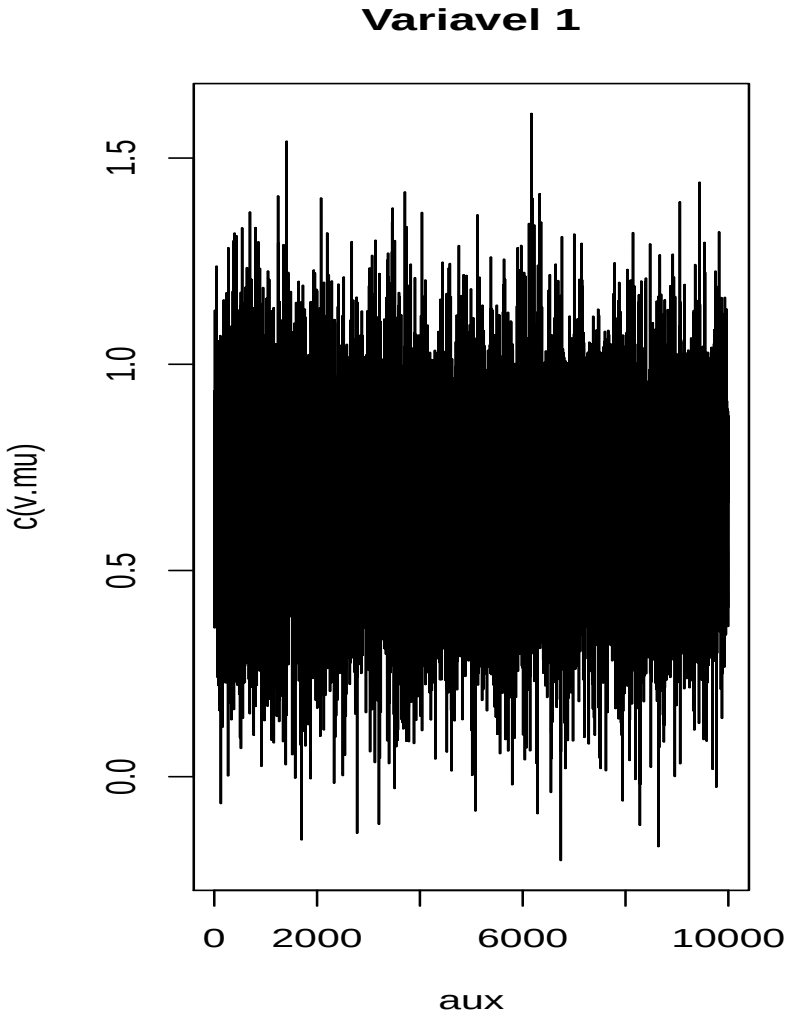
$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; s^2 = \frac{1}{n-2}\sum_{i=1}(x_i - \overline{x})^2.$

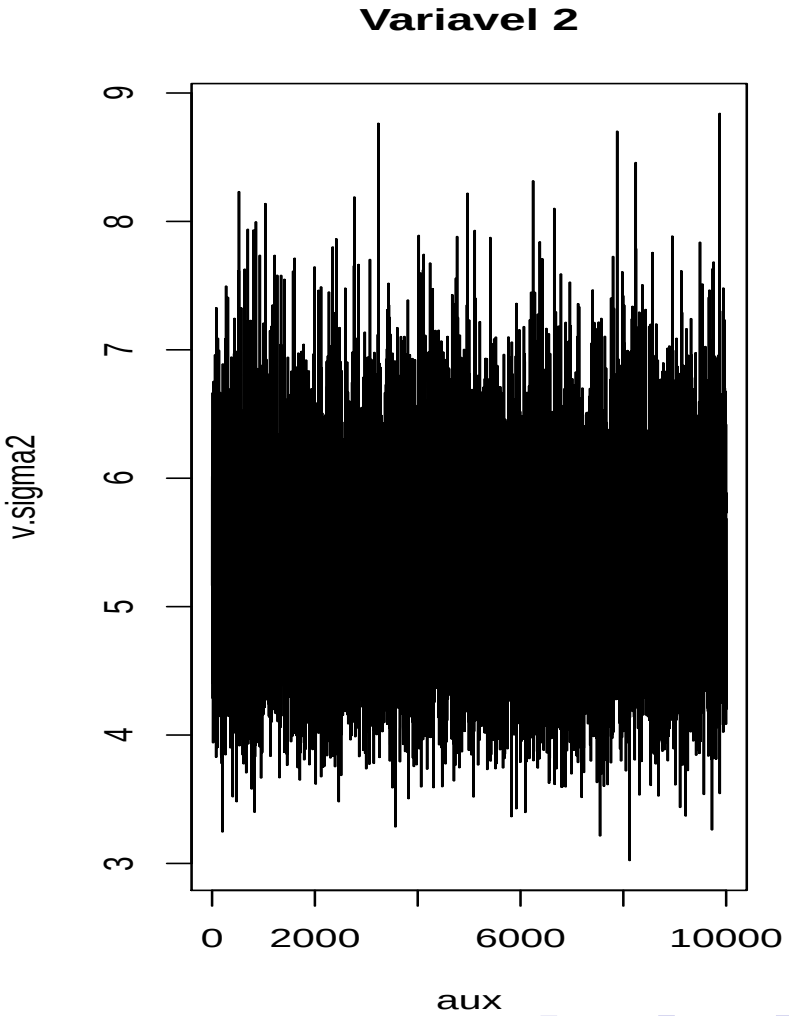
Histograma das amostras MCMC



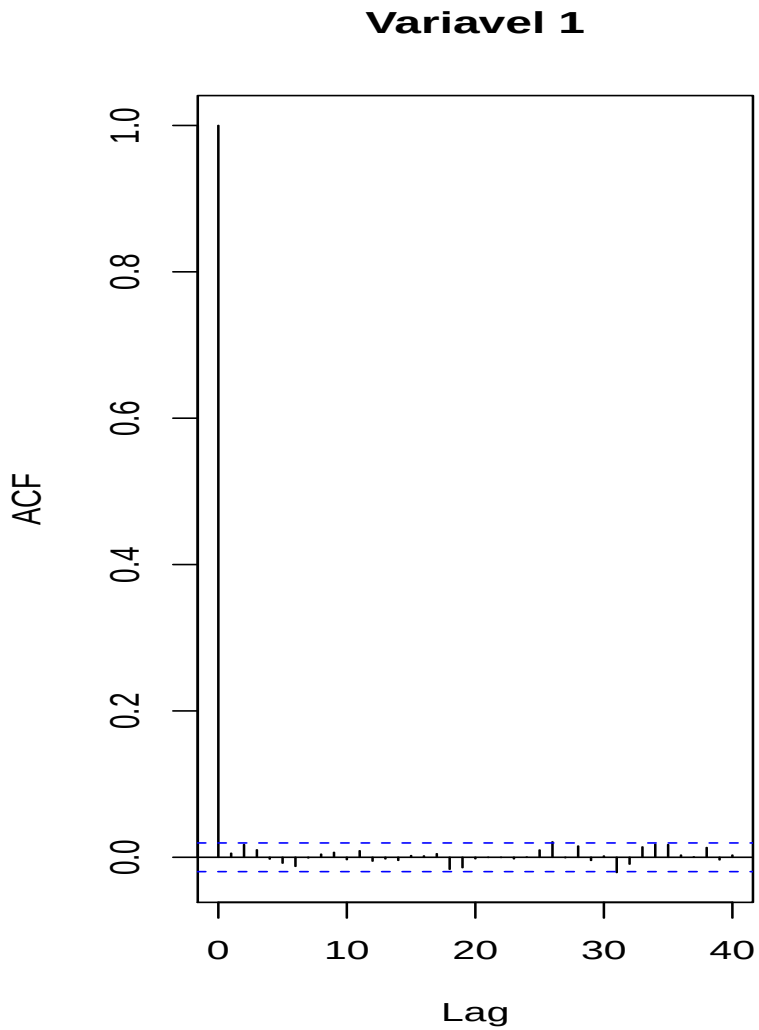


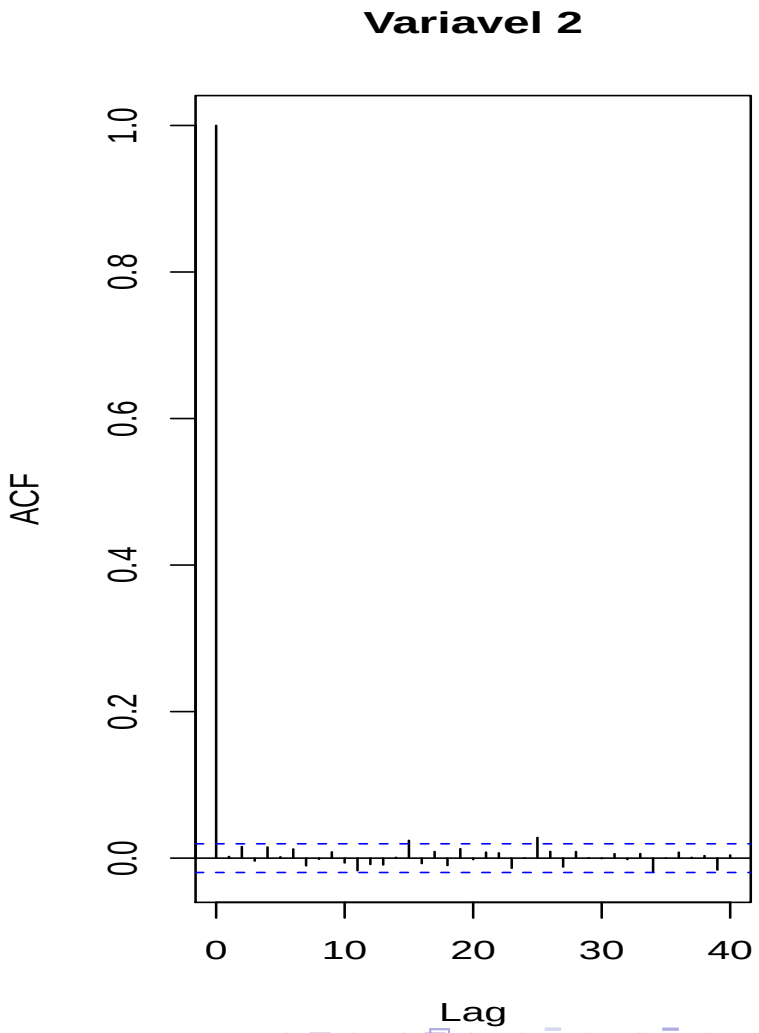
Gráficos das trajetórias (Trace plots) MCMC





ACF das amostras MCMC





Definição geral dos algoritmos MCMC

- Seja $p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ a distribuição a posteriori de interesse (com forma analítica intratável).
- Considere a seguinte partição para o (vetor) p-paramétrico $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_k), k \leq p$.
- Sejam $\boldsymbol{\theta}_j|\boldsymbol{\theta}_{(-j)}, \boldsymbol{x}, i = 1, 2, ..., k$ em que $\boldsymbol{\theta}_{(-j)}$ denota o vetor paramétrico $\boldsymbol{\theta}$ menos a componente $\boldsymbol{\theta}_j$, as **distribuições condicionais completas**.

Definição geral dos algoritmos MCMC

- Inicie a cadeia com um conjunto de valores apropriados (respeitando o espaço paramétrico).
- Para $r = 1, 2, ..., R$, $j = 1, 2, ..., k$ simule $\boldsymbol{\theta}_j^{(r)}$ de $\boldsymbol{\theta}_j | \boldsymbol{\theta}_{(-j)}^{(r-1)}, \boldsymbol{x}$.
- A partir de algum $B > 1$ (burn-in), retenha os valores a cada t (espaçamento) iterações (para evitar a presença de autocorrelação).
- Os valores retidos, se a cadeia tiver alcançado a convergência, corresponderá à uma amostra aleatória das posteriories (distribuições) de interesse.

Questões de interesse (verificação da convergência)

1. Como simular de θ_j caso a distribuição condicional completa não seja conhecida ou possível de simular diretamente?
2. Como verificar a convergência da cadeia gerada?
3. Como determinar B, t, R?
4. Veja: aqui, aqui, aqui.

Ferramentas para verificação de convergência

- Simular diferentes cadeias, a partir de valores iniciais diferentes.
- Colocar, num mesmo gráfico, as trajetórias (trace plots) de todas as cadeias, para cada parâmetros (B,t,R).
- Gráficos de autocorrelação (t).
- Acompanhar, visualmente, a evolução de momentos e/ou quantis, calculados desde o primeiro valor gerado até a iteração r, para cada valor de r (B,R).
- Estatística de Geweke: teste de igualdade de médias para sub-amostras disjuntas de uma única cadeia (B,R).
- Estatística de Gelman-Rubin: análise de variância entre as cadeias geradas (B,R).

Estatística de Geweke (EG)

- Considere uma única cadeia (para cada parâmetro).
- A primeira metade da cadeia é subdividade em b intervalos disjuntos (com aproximadamente o mesmo tamanho).
- Selecionam-se amostras de tamanho n_1 e n_2, da primeira e segunda metades respectivamente.
- Seja $\overline{\theta}_1 = \frac{1}{n_1}\sum_{r=1}^{n_1}\theta_1^{(r)}$ e $\overline{\theta}_2 = \frac{1}{n_2}\sum_{r=1}^{n_2}\theta_2^{(r)}$ (valores simulados).
- $EG = \frac{\overline{\theta}_1 - \overline{\theta}_2}{\sqrt{\frac{\widehat{s}_1(0)}{n_1} + \frac{\widehat{s}_2(0)}{n_2}}}$, em que $\widehat{s}_i^2(0)$ é a variância da amostra calculada pela estimativa consistente da densidade espectral na frequência zero (por causa de possível existência de autocorrelação), da amostra i.

Estatística de Geweke (EG)

- Repete-se o processo acima, descartando-se o primeiro subintervalo da primeira metade da cadeia.
- Repete-se, uma segunda vez, descartando-se os dois primeiros subintervalos da primeira metada da cadeia.
- Repete-se o processo até ficar-se com somente o último subintervalo.
- Espera-se, se a cadeia tiver convergido, que a partir de um determinado "descarte", as estatísticas não sejam mais significativas (estejam entre (-2,2)).

Estatística de Gelman Rubin (EGR)

- Consiste em uma análise de variância considerando M diferentes cadeias para cada parâmetro.
- Defina

$$W = \frac{1}{M}\sum_{j=1}^{M} s_j^2; s_j^2 = \frac{1}{R-1}\sum_{r=1}^{R}\left(\theta_j^{(r)} - \overline{\theta}_j\right)^2; \overline{\theta}_j = \frac{1}{R}\sum_{r=1}^{R}\theta_j^{(r)}$$

$$V = \frac{R}{M-1}\sum_{j=1}^{M}\left(\overline{\theta}_j - \overline{\overline{\theta}}\right)^2; \overline{\overline{\theta}} = \frac{1}{M}\sum_{j=1}^{M}\overline{\theta}_j$$

$$\widehat{Var}(\theta) = \left(1 - \frac{1}{R}\right)W + \frac{M+1}{MR}V; R_c = \sqrt{\frac{\widehat{d}+3}{\widehat{d}+1}\frac{\widehat{Var}(\theta)}{W}}$$

Estatística de Gelman Rubin (EGR)

- É possível provar que $R_c \geq 1$. Em geral, se $R_c \leq 1,2$, então a cadeia convergiu (cuidado, depende de vários fatores).
- Na prática: cada cadeia (para cada parâmetro) é subdividida em intervalos. Calcula-se a estatística R_c utilizado o primeiro intervalo (descartando-se a primeira metade) de todas as M cadeias. Depois, calcula-se R_c utilizando-se os dois primeiros intervalos (descartando-se a primeira metade do intervalo resultante), depois os três primeiros (descartando-se a primeira metade do intervalo resultante), assim por diante.
- Analisa-se o comportamento da estatística (calculando-se um intervalo de confiança) ao longo desse processo.

Modelo gama

- Exemplo: posteriori associada ao modelo gama(r, λ).
- Suponha uma amostra aleatória de tamanho n de $X|r, \lambda \sim$ gama(r, λ), ou seja:

$$p(x|r,\lambda) = \frac{1}{\lambda^r \Gamma(r)} e^{-x/\lambda} x^{r-1} \mathbb{1}_{(0,\infty)}(x)$$

- Verossimilhança:

$$p(\boldsymbol{x}|r) = \frac{1}{\lambda^{nr}\Gamma(r)^n} e^{-n\bar{x}/\lambda} \prod_{i=1}^{n} x_i^{r-1}$$

Priori e posteriori

- Priori ($r \sim \text{gama}(\alpha, \gamma)$ e $\lambda \sim \text{IG}(\delta, \phi)$).

$$\begin{aligned} p(r,\lambda) = p(r)p(\lambda) &= \frac{1}{\gamma^{\alpha}\Gamma(\alpha)} e^{-r/\gamma} r^{\alpha-1} \mathbb{1}_{(0,\infty)}(r) \\ &\times \frac{\phi^{\delta}}{\Gamma(\phi)} e^{-\phi/\lambda} \lambda^{-\delta-1} \mathbb{1}_{(0,\infty)}(\lambda) \end{aligned}$$

- Posteriori

$$\begin{aligned} p(r,\lambda|\boldsymbol{x}) &\propto \frac{1}{\lambda^{nr}\Gamma(r)^n} e^{-n\overline{x}/\lambda} \prod_{i=1}^{n} x_i^{r-1} \times e^{-r/\gamma} r^{\alpha-1} \\ &\times e^{-\phi/\lambda} \lambda^{-\delta-1} \mathbb{1}_{(0,\infty)}(r) \mathbb{1}_{(0,\infty)}(\lambda) \end{aligned}$$

Distribuições condicionais completas

- Condicionais completas

$$p(r|\lambda, \boldsymbol{x}) \propto \frac{1}{\lambda^{nr}\Gamma(r)^n} \prod_{i=1}^{n} x_i^{r-1} \times e^{-r/\gamma} r^{\alpha-1}$$

$$p(\lambda|r, \boldsymbol{x}) \propto e^{-(n\overline{x}+\phi)/\lambda} \lambda^{-nr-\delta-1}$$

- Note que: $\lambda|r, \boldsymbol{x} \sim \text{IG}(nr+\delta, n\overline{x}+\phi)$ mas $r|\lambda, \boldsymbol{x}$ não corresponde à nenhuma distribuição catalogada.
- Algoritmo: simular, iterativamente, valores para (r, λ) através das duas distribuições acima.

Simulação das distribuições condicionais completas

- Se todas as distribuições condicionais completas forem conhecidas e fáceis de simular, teremos o algoritmo do amostrador de Gibbs ("Gibbs sampling").
- Como no exemplo anterior da distribuição normal.
- Ponto importante, identificada a distribuição condicional completa, eg, normal, inversa gama, t de Student, devemos escolher algoritmos apropriados para simular delas (eg, transformada inversa, rejeição, rejeição adaptativa, amostragem por importância).

Simulação das distribuições condicionais completas

- Caso alguma(a) distribuição(ões) condicional(is) completa(s) não seja(m) "conhecida(s)", algum algoritmo auxiliar para simular valores dela tem de ser usado: Metropolis-Hastings, rejeição adaptativa, amostragem por corte etc.

Amostrador de Gibbs

- Inicie as cadeias com valores iniciais conveniente.
- Simule, para r=1,2,...,R:
  - $\theta_1^{(r)}$ de $\theta_1^{(r)}|\theta_2^{(r-1)}, \theta_3^{(r-1)}, ..., \theta_k^{(r-1)}, \boldsymbol{x}$.
  - $\theta_2^{(r)}$ de $\theta_2^{(r)}|\theta_1^{(r)}, \theta_3^{(r-1)}, ..., \theta_k^{(r-1)}, \boldsymbol{x}$.
  - $\theta_3^{(r)}$ de $\theta_3^{(r)}|\theta_1^{(r)}, \theta_2^{(r)}, ..., \theta_k^{(r-1)}, \boldsymbol{x}$.

    $\vdots$
  - $\theta_k^{(r)}$ de $\theta_k^{(r)}|\theta_1^{(r)}, \theta_2^{(r)}, ..., \theta_{k-1}^{(r)}, \boldsymbol{x}$.

WinBUGS: página, artigo, livro

- Programa WinBugs: permite ajustar modelos complexos usando diversos algoritmos do tipo MCMC.
- Em geral, basta apenas fornecer o modelo (verossimilhança) e prioris.
- Limitações no uso de prioris impróprias (Jeffreys).
- Dispõe de mecanismos para inserir verossimilhanças e prioris que não são padrão (não constam em seu banco de dados).
- Pode ser utilizado de modo mais simples através do pacote R2WinBUGS do R.
- O diagnóstico de convergência pode ser facilmente realizado através do pacote coda, disponível no programa R.
- Não está sendo mais atualizado (projeto desativado).

Algoritmos para simular das cond.(tradução do manual)

- Condicional completa contínua
  - Conjugada (conhecida): amostragem direta usando algoritmos padrão.
  - Log-côncava: Rejeição adaptativa de derivação livre (Gilks, 1992).
  - Espaço paramétrico restrito: Amostragem por corte ("slice-sampling"), Neal, 1997.
  - Espaço paramétrico irrestrito: Metropolis-Hastings (adaptativo).
- Condicional completa discreta
  - Limite superior finito: transformada inversa.
  - Poisson deslocada: amostragem direta usando algoritmos padrão.

Comparação de modelos

- Seja $h(\boldsymbol{\theta}) = -2 \ln p(\boldsymbol{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ (quanto menor, melhor o ajuste do modelo).
- Seja $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(s)}$, uma amostra MCMC válidade de tamanho s. Defina: $\overline{\boldsymbol{\theta}} = (\overline{\theta}_1, ..., \overline{\theta}_k)$, em que $\overline{\boldsymbol{\theta}}_j = \frac{1}{s} \sum_{r=1}^{s} \boldsymbol{\theta}_j^{(r)}, j = 1, 2, ...k$.
- Defina ainda: Deviance $= \overline{D} = \frac{1}{s} \sum_{r=1}^{s} h(\boldsymbol{\theta}^{(s)})$ e $\widehat{D} = h(\overline{\boldsymbol{\theta}})$.
- Estatística de comparação de ajuste de modelos: Deviance, $p_D = \overline{D} - \widehat{D}$ e $DIC = \overline{D} + p_D$.
- Em princípio, outras estatísticas podem ser utilizadas como aquelas vistas aqui, utilizando-se $\widehat{D}$ no lugar de $-2 \ln p(\boldsymbol{x}|\boldsymbol{\theta})$.
- Veja também as estatísticas "loo" e "WAIC".

Comparação de modelos

- Quanto maior o valor de p_D e menor os valores do *Deviance* e do *DIC* melhor o ajuste do modelo. O WinBUGS calcula ambos.
- Atenção: mesmo o modelo que apresenta o melhor ajuste (segundo os critérios acima) pode não estar bem ajustado aos dados. Faz-se necessário, sempre, a verificação de qualidade de ajuste do modelo eleito por análise de resíduos, distribuição preditiva, etc.
- Além de que a análise residual pode ser feita de forma parecida ao caso frequentistas (como visto anteriormente), novas abordagens podem ser consideradas veja: aqui, aqui, aqui, aqui, aqui, aqui, aqui.

Comentários

- Exemplos de cálculos de posterioris e condicionais completas podem ser encontrados aqui e aqui.
- Nos concentramos na análise de dados, ao invés das contas para obtenção das posteriores e condicionais completas.
- Mostraremos alguns exemplos em que modelos de um nível (probabilísticos e de regressão) são empregados, antes de apresentarmos modelagens de dados hierárquicos.

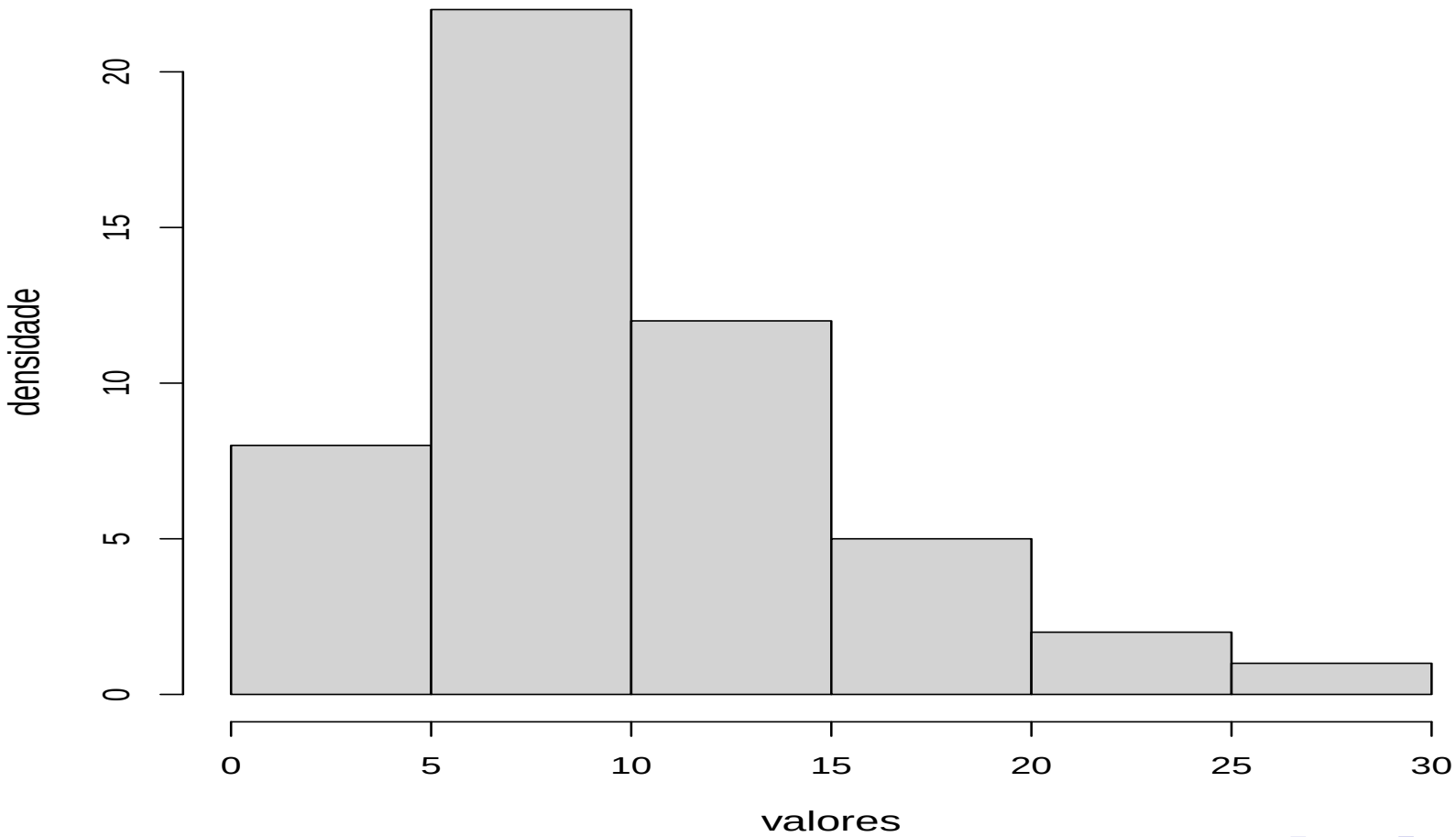
Comentários

- Nos concentraremos no WinBUGS/OpenBUGS e no brms. Os códigos feitos no WinBUGS/OpenBUGS podem ser adaptados, facilmente, para o outros programas genéricos de algoritmos MCMC.
- Além disso, outros pacotes no R podem ser utilizados, com algum esforço adicional.

Exemplo da potência das turbinas de aviões

- Cinco tipos de turbinas foram comparadas em relação aos respectivos tempos de vida (em milhões de ciclos).
- Quanto maior o número médio de ciclos, melhor o desempenho da turbina.
- Foram consideradas 10 turbinas de cada um dos cinco tipos.
- Vamos desconsiderar os tipos de turbina e modelar o tempo de vida como se fossem de uma única população.
- Seja $Y_i \overset{iid}{\sim} \text{gama}(r, \lambda)$, $\mathcal{E}(Y_j|r,\lambda) = r\lambda$ e $\mathcal{V}(Y_j|r,\lambda) = r\lambda^2$ o tempo de vida da i-ésima turbina, $i = 1, ..., 50$

Histograma dos dados



Cont.

- Prioris: $r \sim \text{gama}(0,01; 1/0,01), \mathcal{E}(R) = 1, \mathcal{V}(R) = 100$ e $\lambda \sim \text{IG}(0,01; 0,01)$ (vaga).
- Algoritmo WinBUGS para o ajuste do modelo.

gamamodel< −function(){

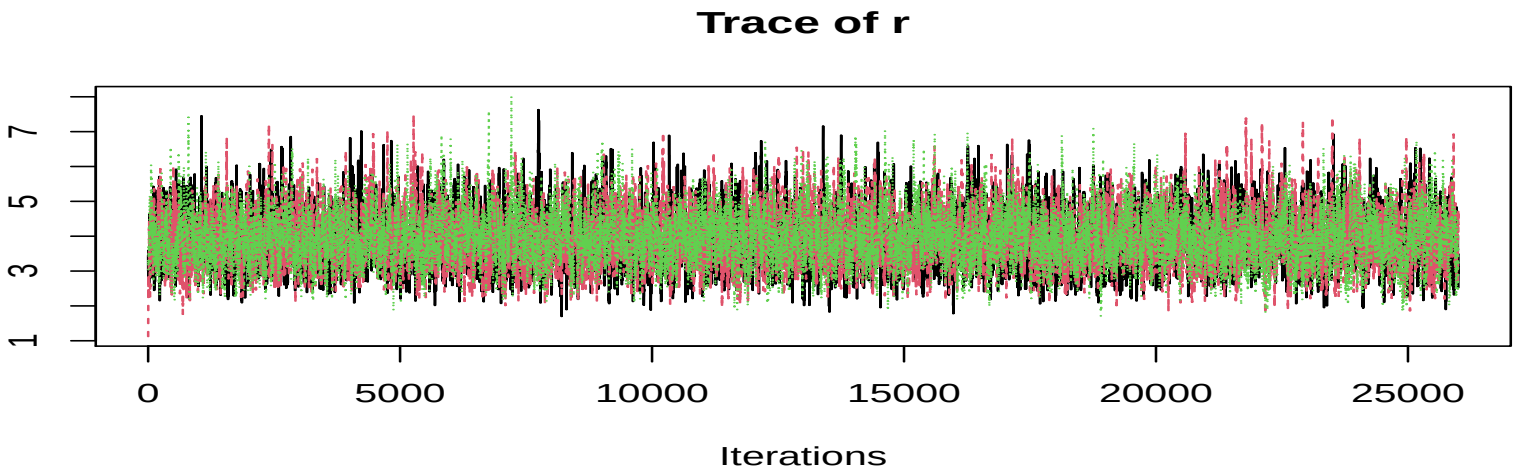
for (i in 1 : N){ y[i]∼dgamma(r,lambdaa)}

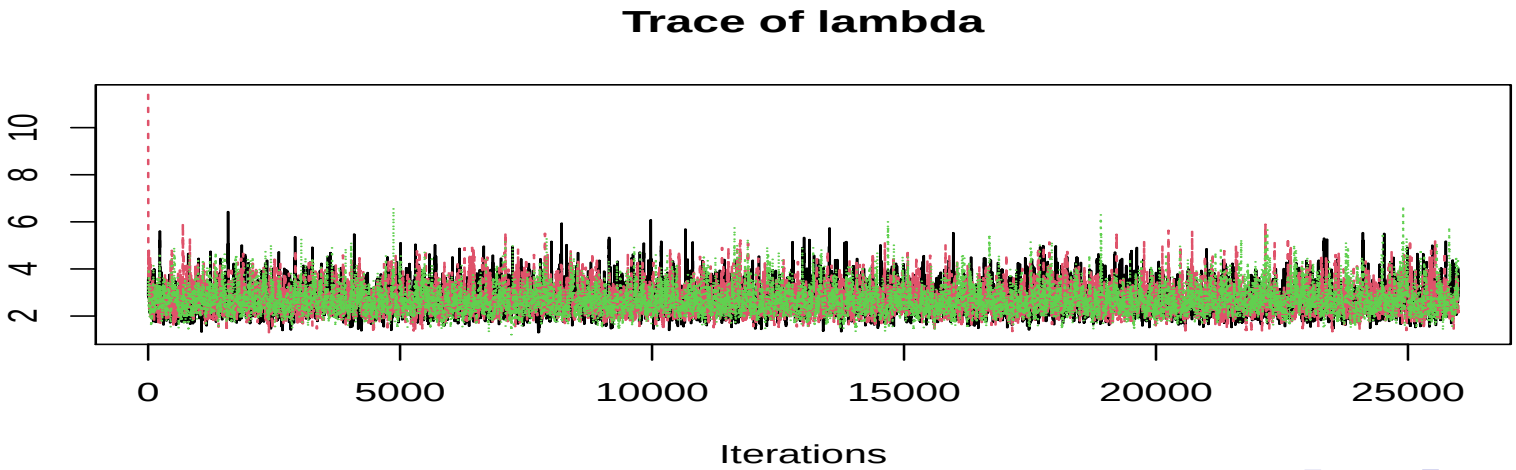
r∼dgamma(0.01,0.01)

lambdaa∼dgamma(0.01,0.01)

lambda< −pow(lambdaa,-1)}

Traceplots para os três conjuntos de cadeias geradas





Autocorrelações para um dos conjuntos de cadeias

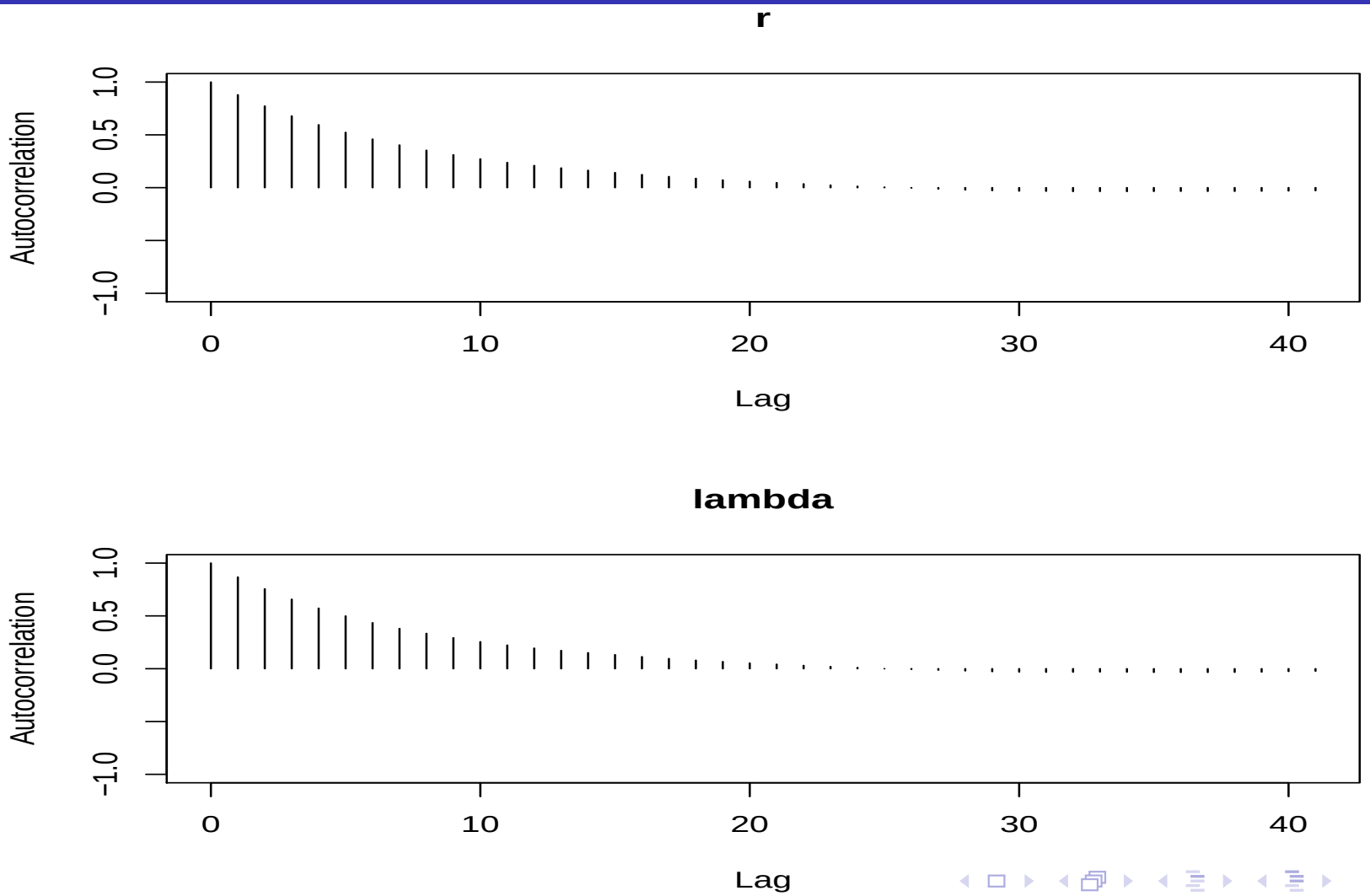


Gráfico das med. acumul. para um dos conj. de cadeias

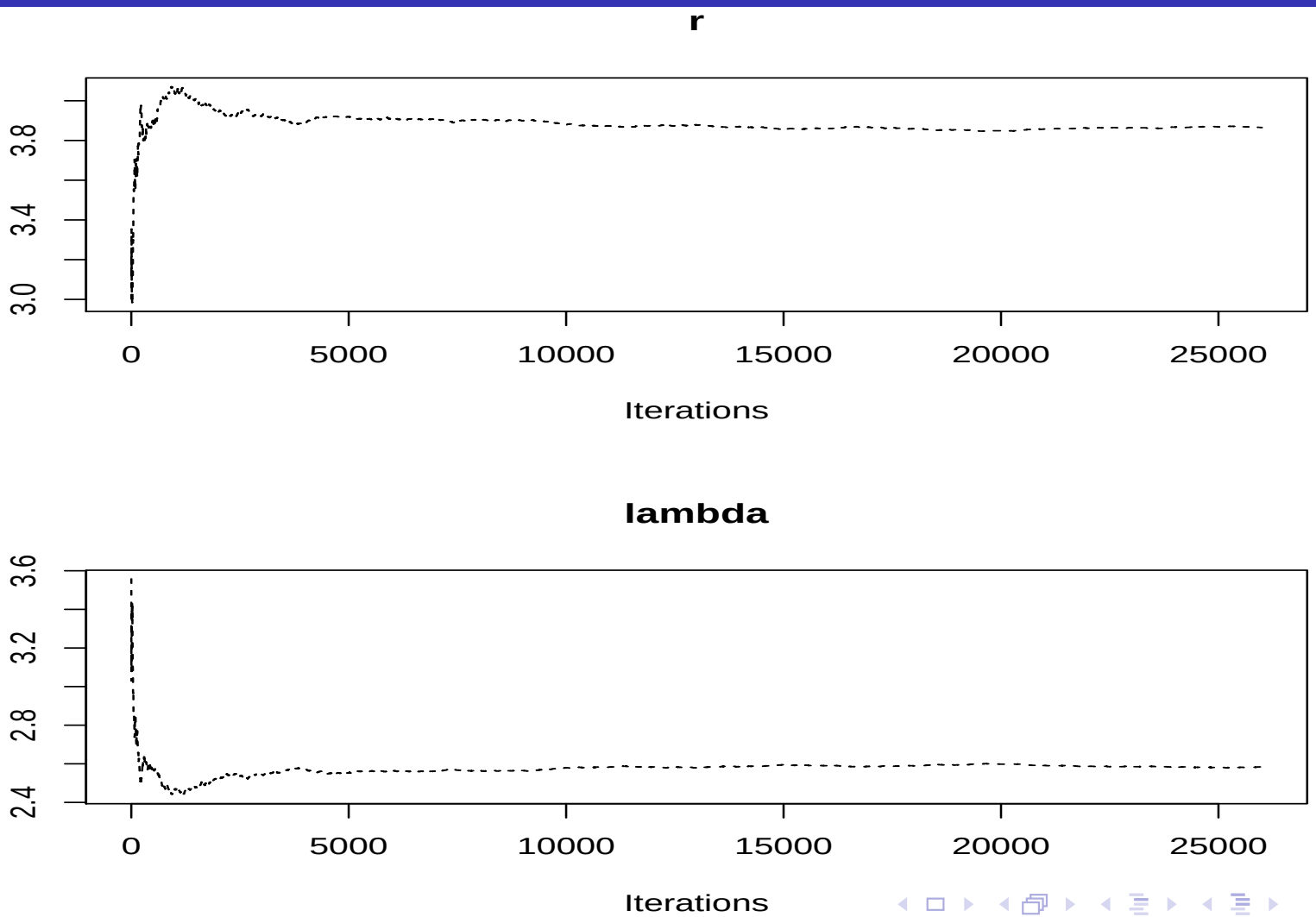
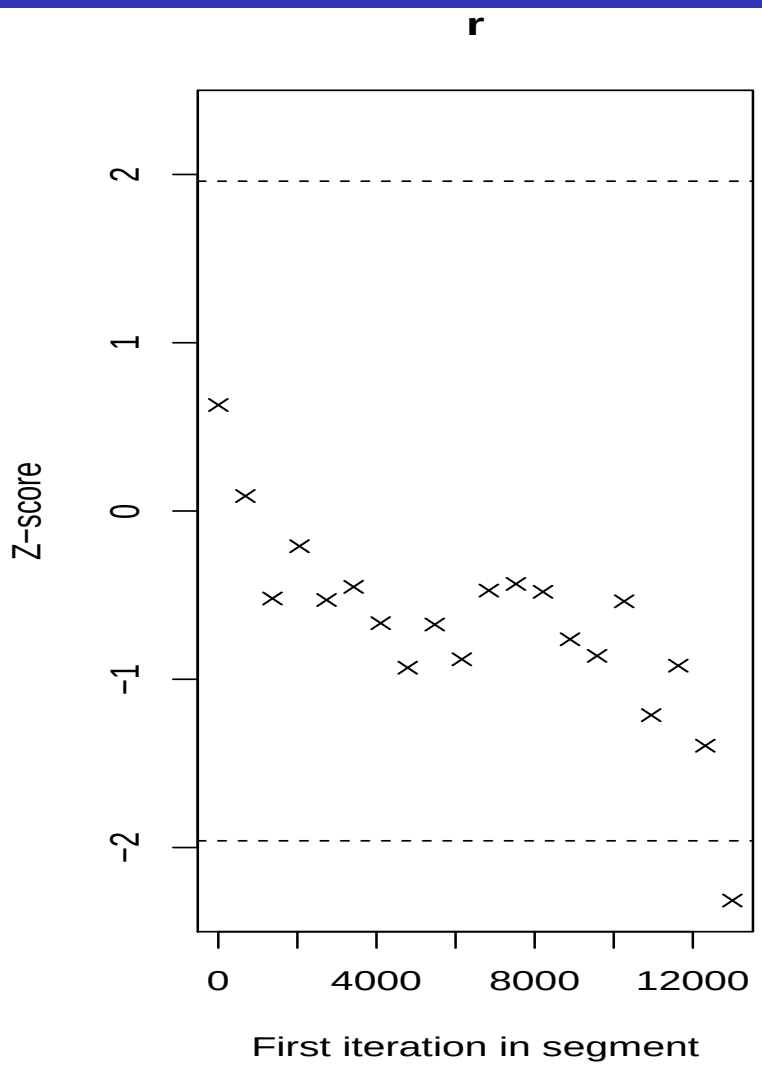


Gráfico da estat. de Geweke para um dos conj. de cadeias



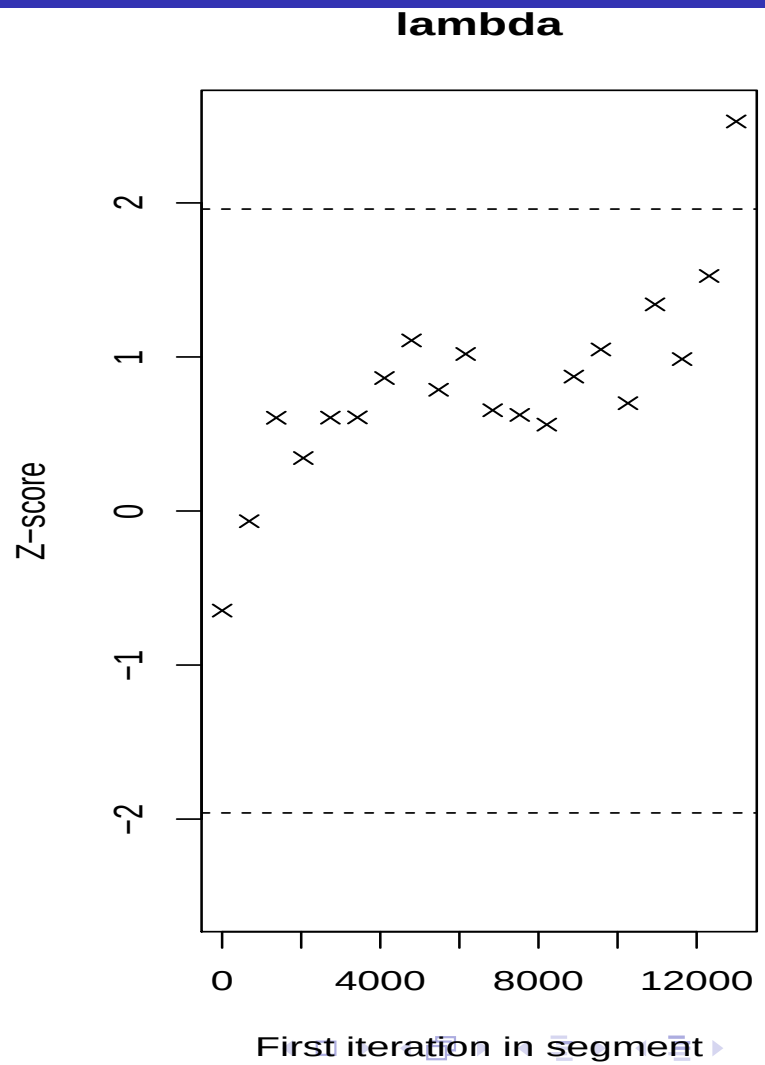
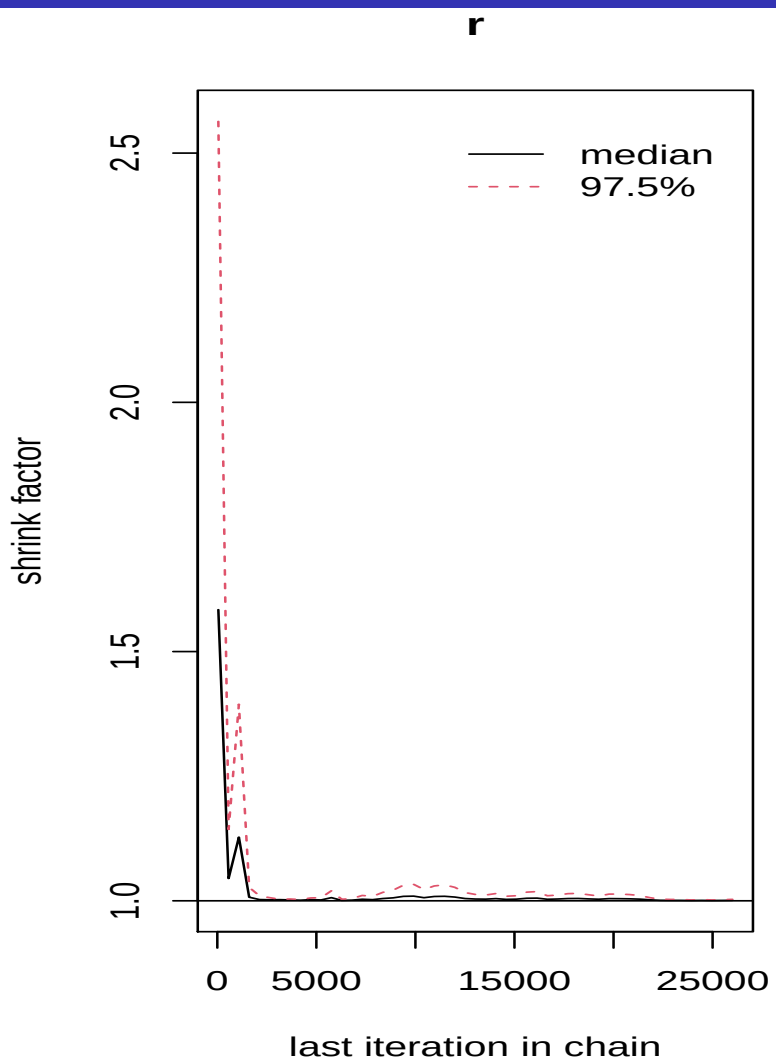
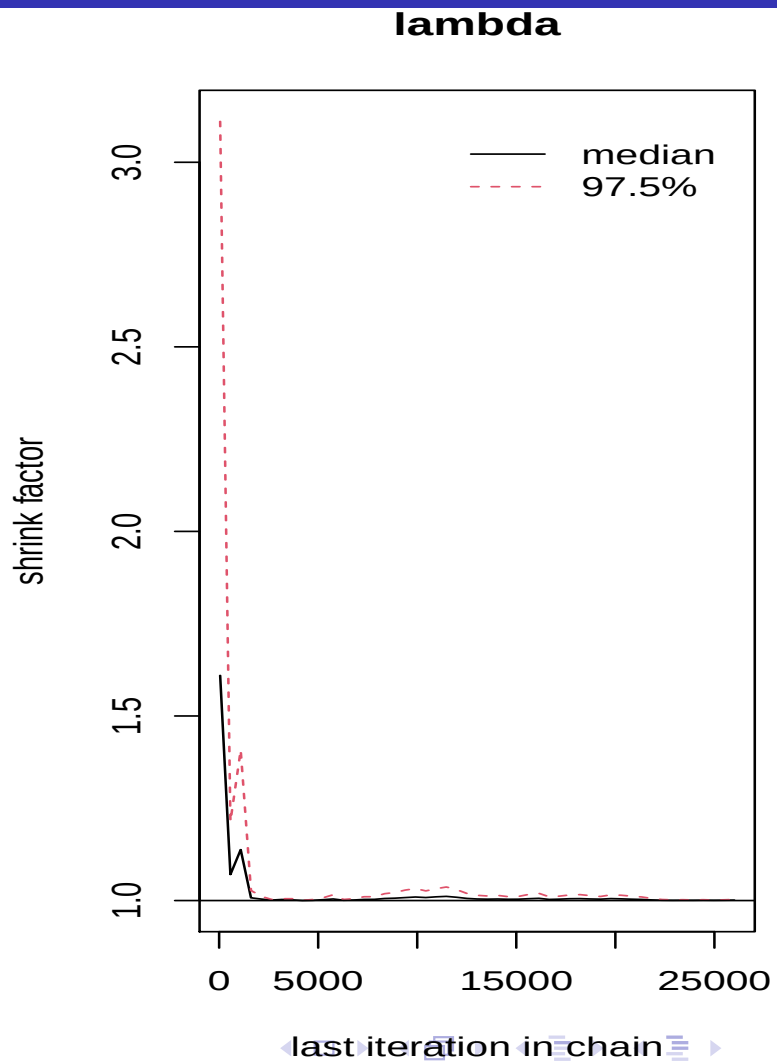
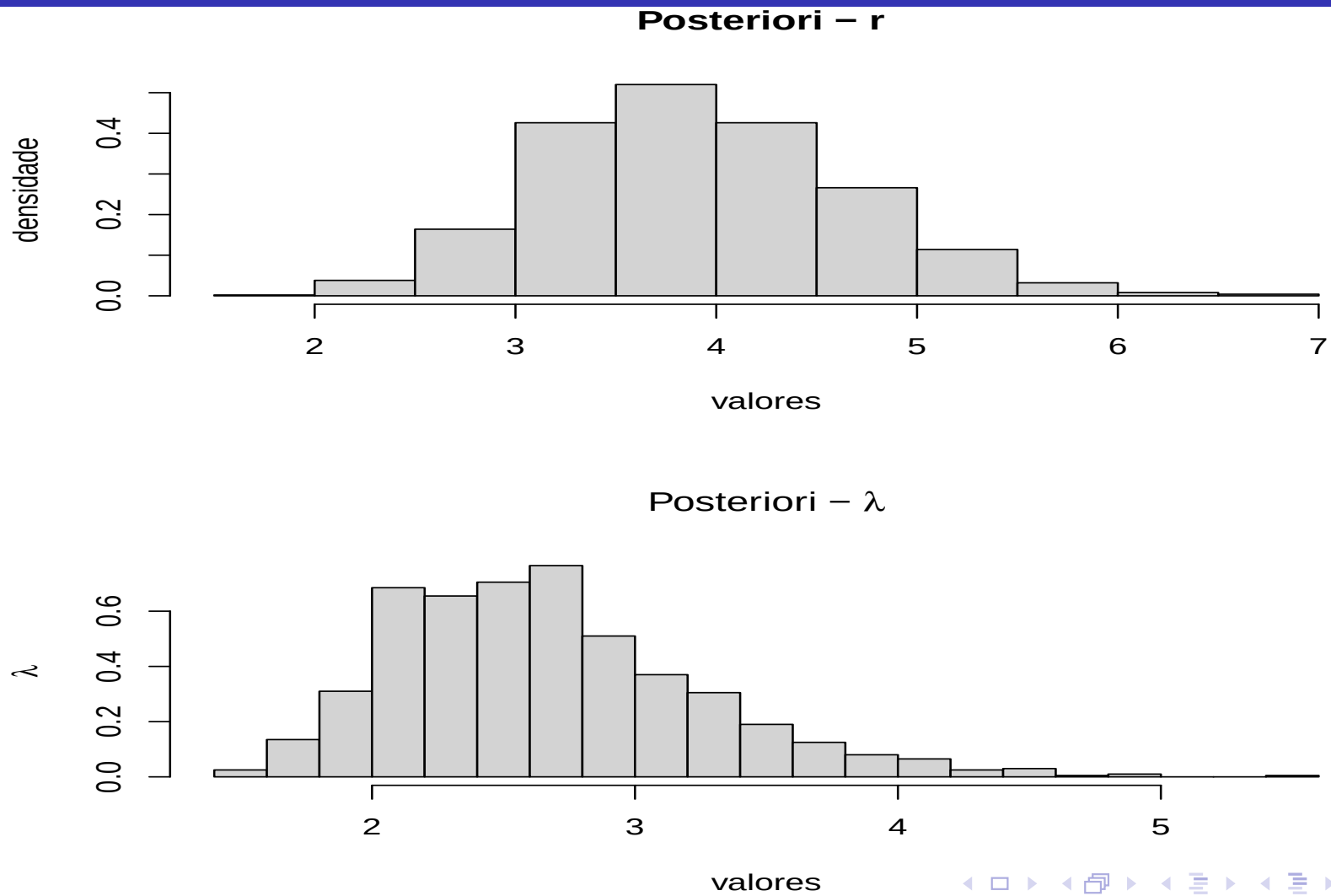


Gráfico da estat. de GR utilizando os três conj. de cadeias





Hist. da amostra válida para um dos conjuntos de cadeias



Estimativas Bayesianas para um dos conjuntos de cadeias

- Resumo: B = 6000, t = 20, R = 26000.

Estatística	Parâmetro	
	r	λ
EAP	3,92	2,66
EPAP	0,76	0,59
$IC_B(95\%)$	[2,55; 5,47]	[1,75; 4,15]
$HPD(95\%)$	[2,47; 5,39]	[1,62;3,87]

- Modelo gama: Deviance $= 295,7$, $p_D = 2,1$; $DIC = 297,7$. Modelo exponencial Deviance $= 331,0$; $p_D = 1,0$; $DIC = 332,1$.

Exemplo do número acidentes

- Descrição: número de acidentes (com algum tipo de trauma para as pessoas envolvidas) em 92 dias (correspondentes) em dois anos distintos (1961 e 1962), medidos em algumas regiões da Suécia.
- Considerou-se apenas 43 dias, correspondendo a dias de 1961 em que não havia limite de velocidade e de 1962 em que havia limites de velocidade (90 ou 100 km/h).

Exemplo do número acidentes

- Modelo proposto para analisar os dados: Considere (i =1, ano de 1961, i=2, ano de 1962). Lembrando que: 1961 (sem limite de velocidade) 1962 (com limite de velocidade), temos

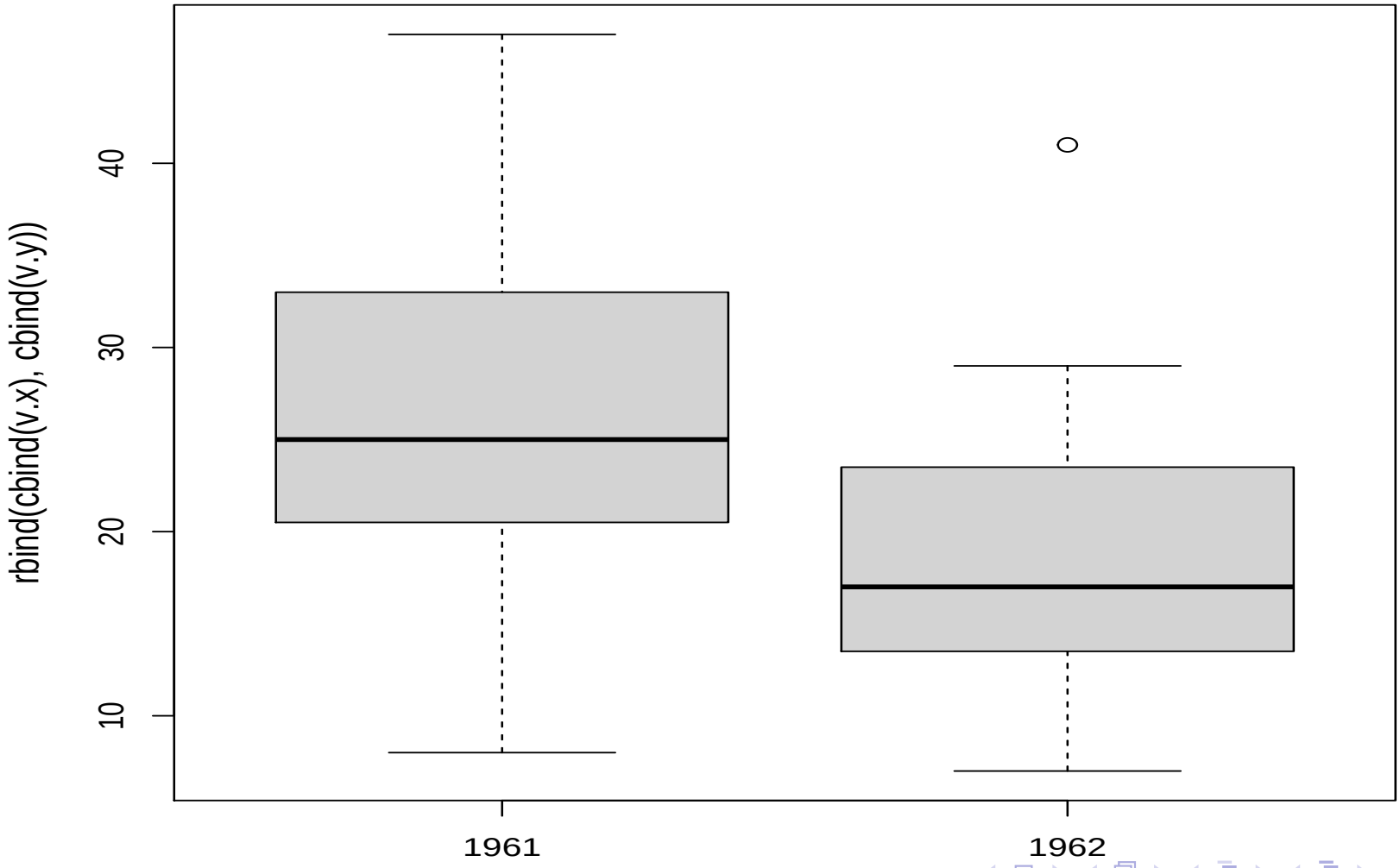
$$\begin{aligned} Y_{ji}|\boldsymbol{\beta} &\overset{ind.}{\sim} \text{Poisson}(\mu_i), i = 1, 2, j = 1, ..., 43 \\ \ln \mu_i &= \mu + \alpha_i, \alpha_1 = 0 \end{aligned}$$

em que $\boldsymbol{\beta} = (\mu, \alpha_2)'$. Assim, concluí-se que $\mathcal{E}(Y_{ji}|\boldsymbol{\beta}) = e^{\mu+\alpha_i}$.

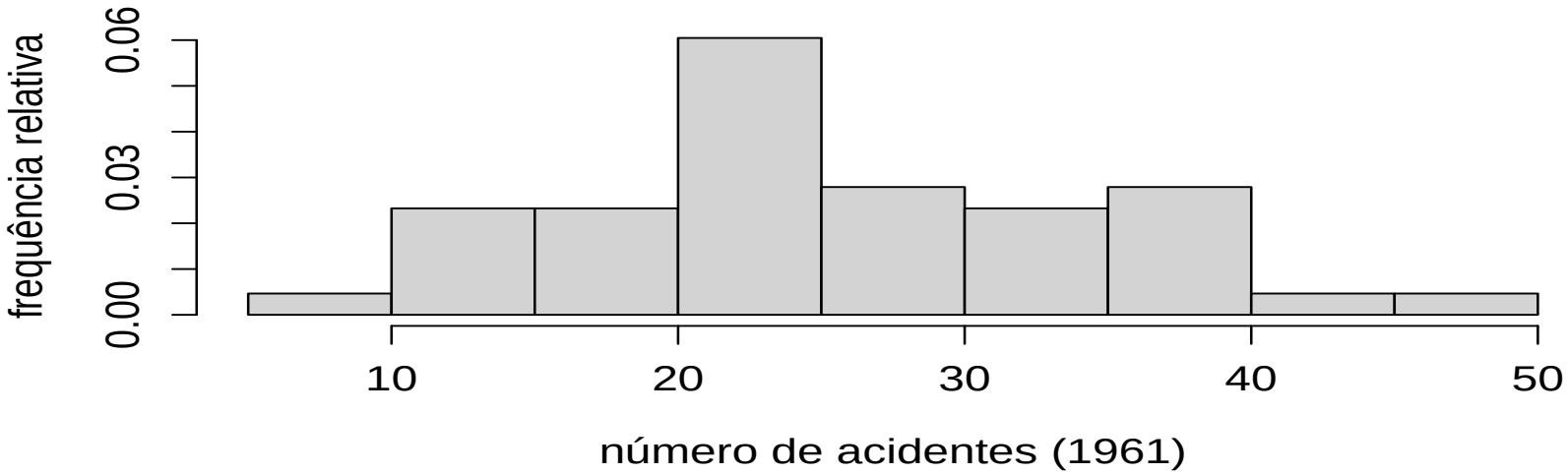
Medidas Resumo

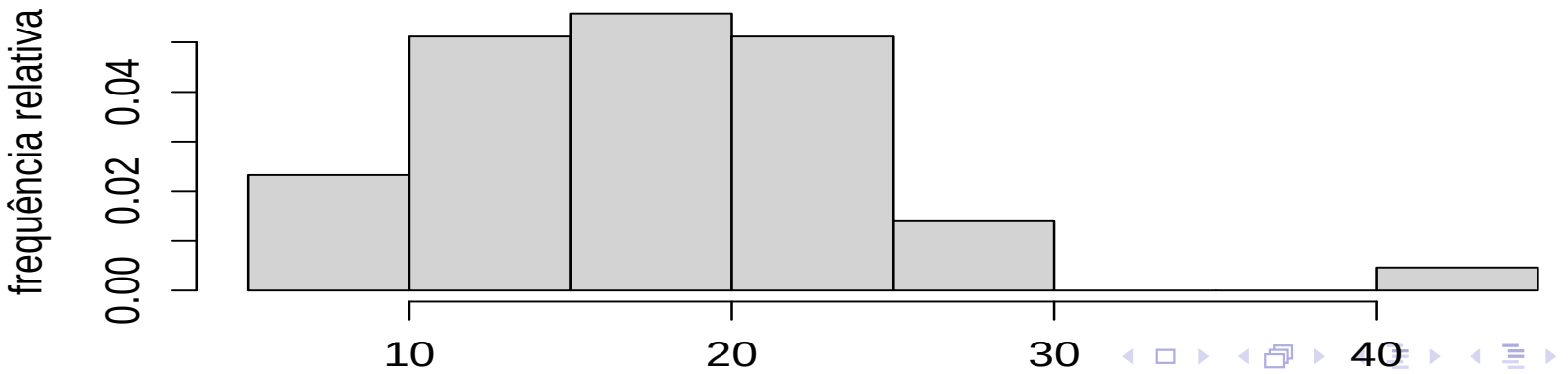
Ano	Média	Var.	DP	CV(%)	Mín.	Med.	Máx.
1961	26,05	82,66	9,09	34,91	8,00	25,00	47,00
1962	18,05	44,71	6,69	37,05	7,00	17,00	41,00

Boxplots do número de acidentes por ano

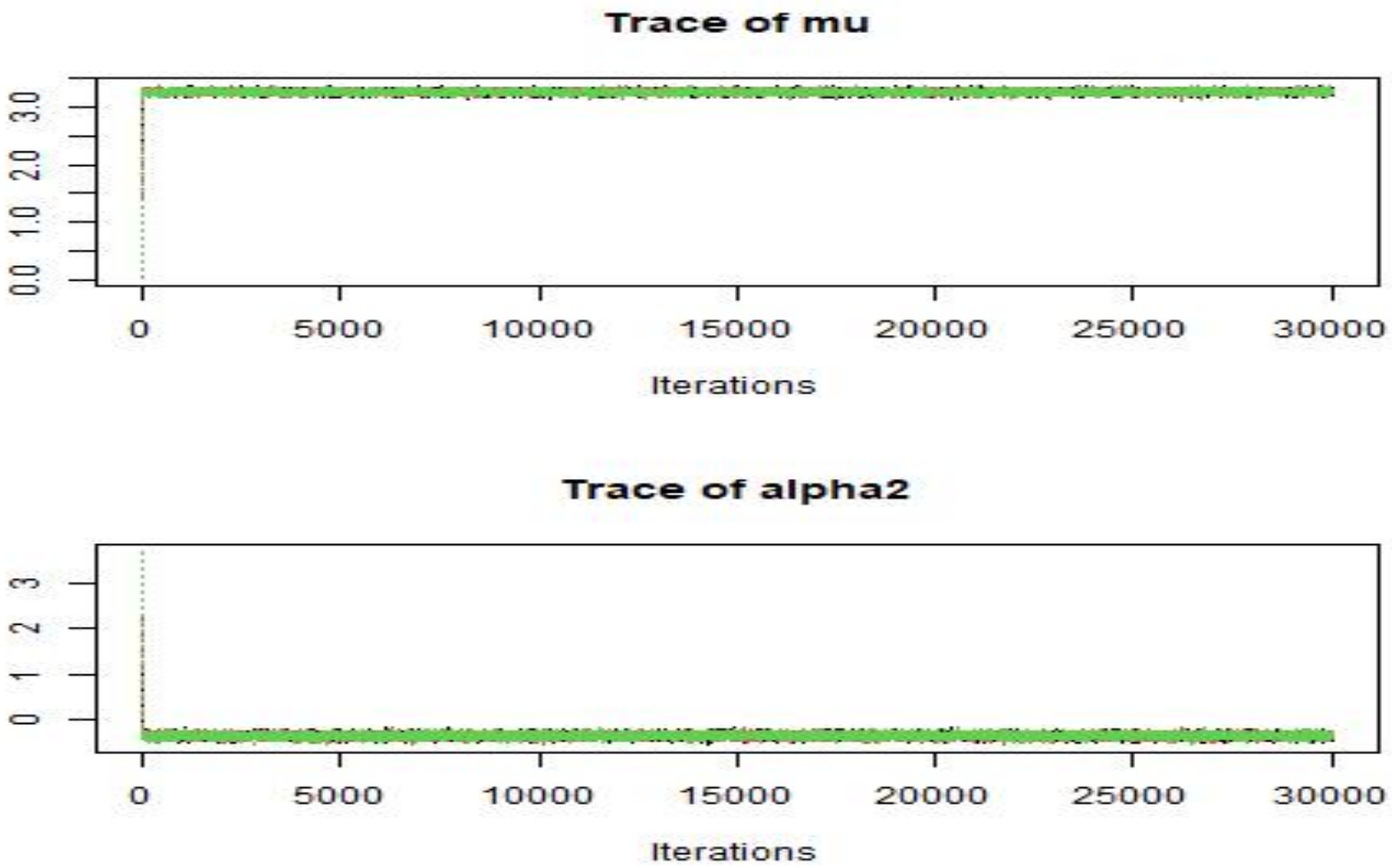


Histogramas do número de acidentes por ano

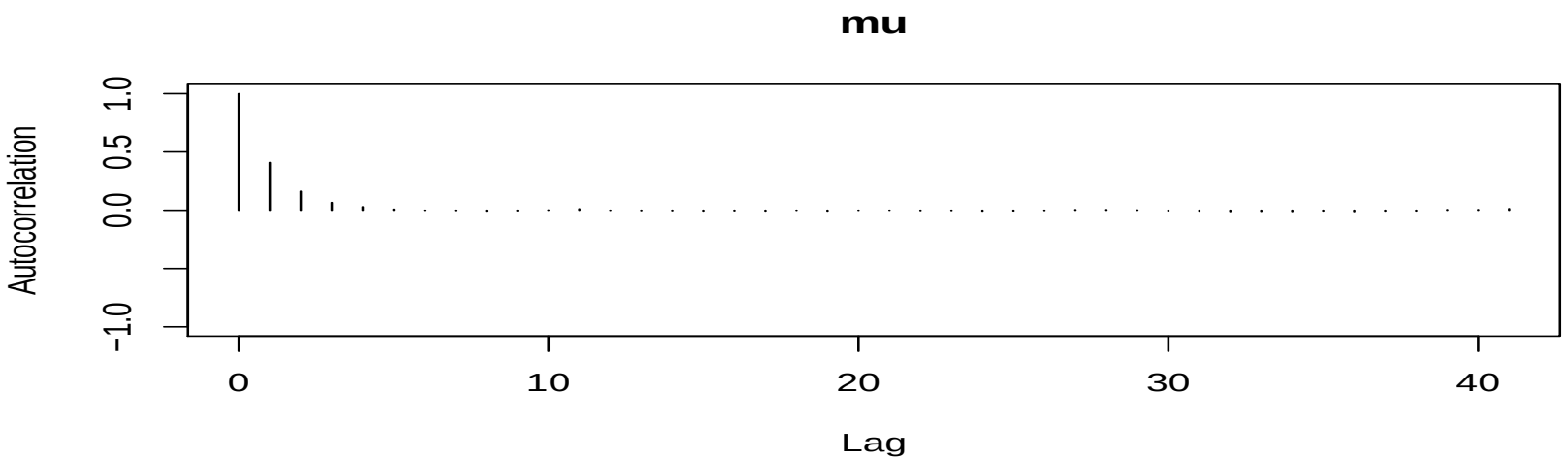




Traceplots para os três conjuntos de cadeias geradas



Autocorrelações para um dos conjuntos de cadeias



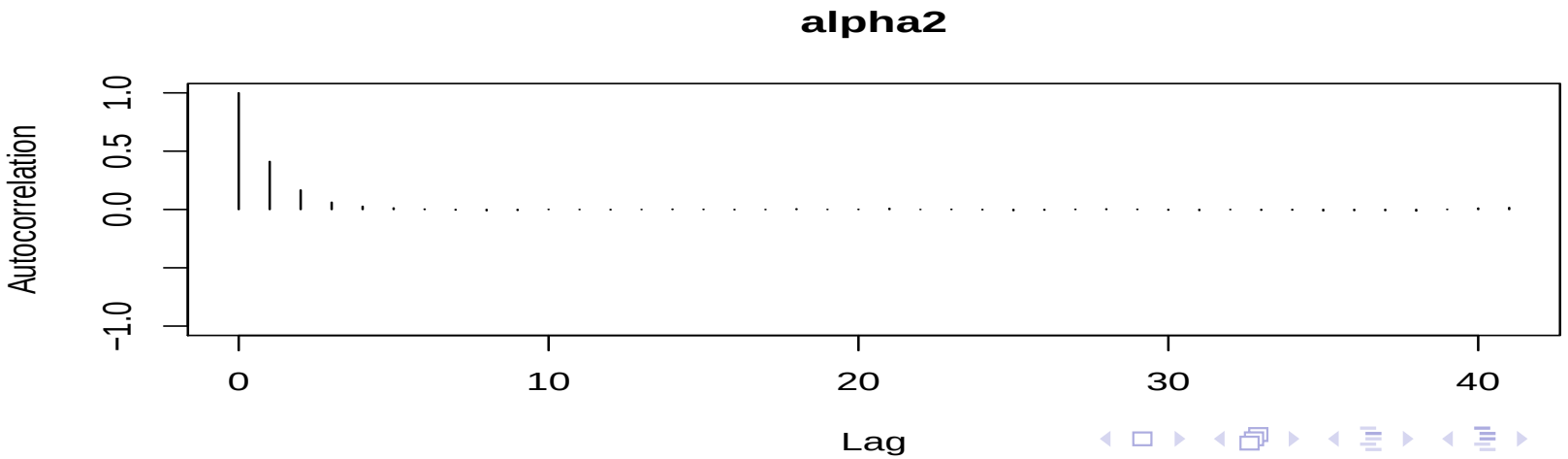
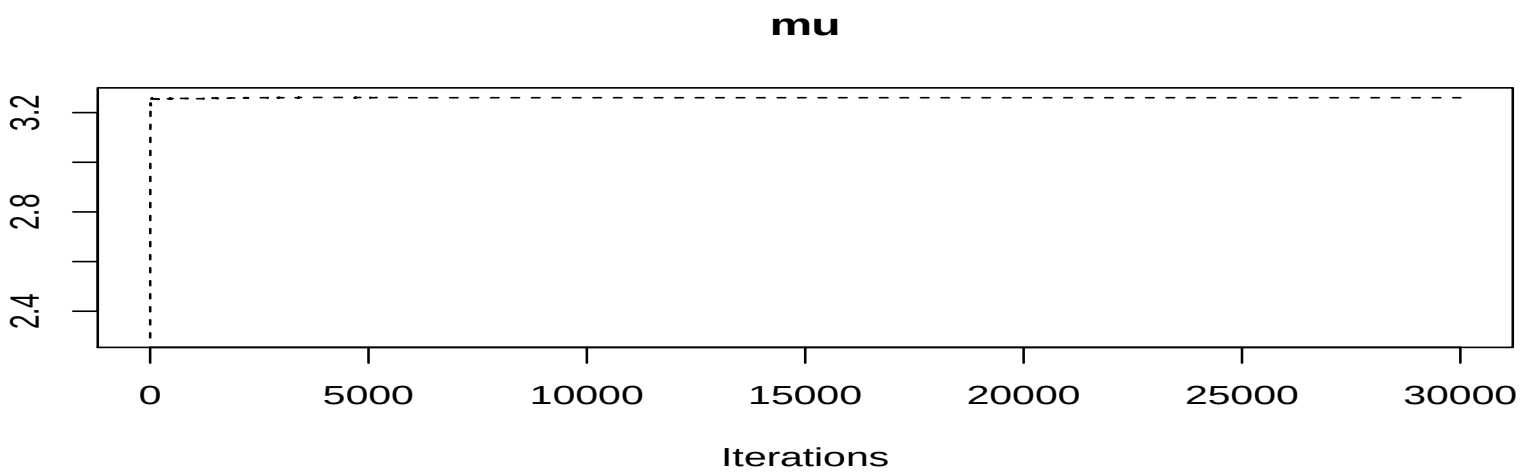


Gráfico das med. acumul. para um dos conj. de cadeias



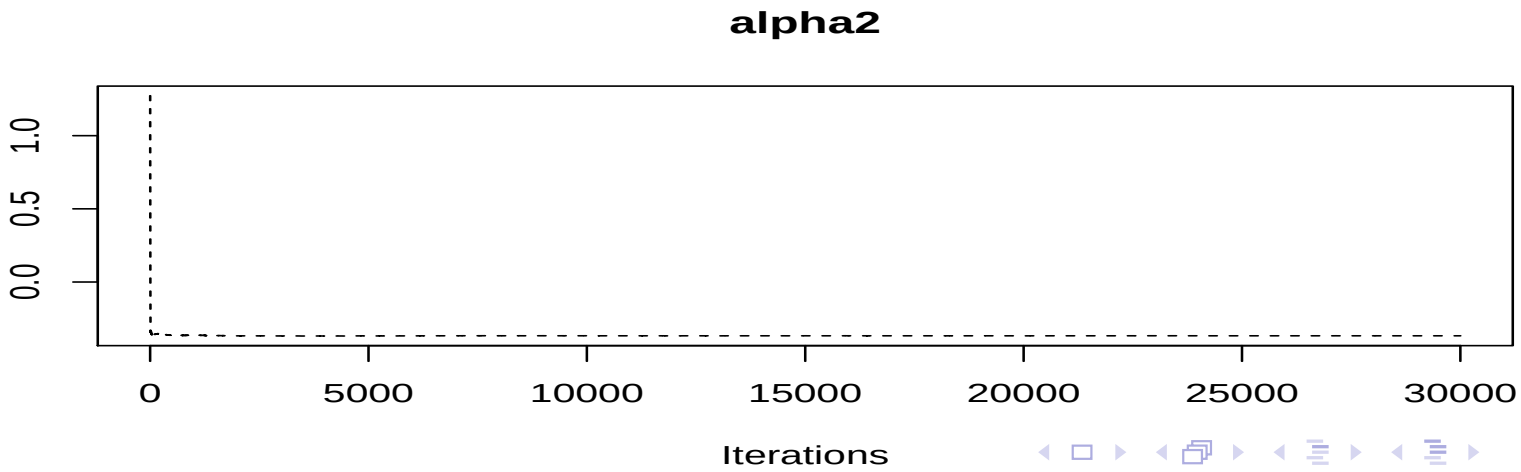
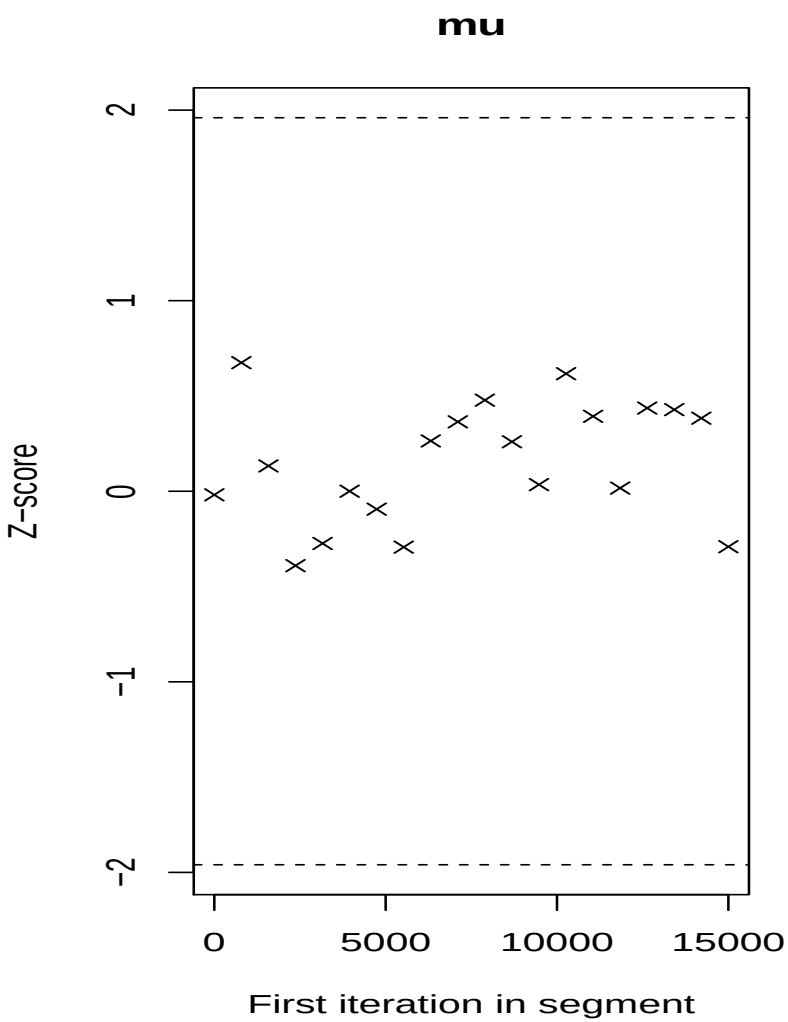


Gráfico da estat. de Geweke para um dos conj. de cadeias



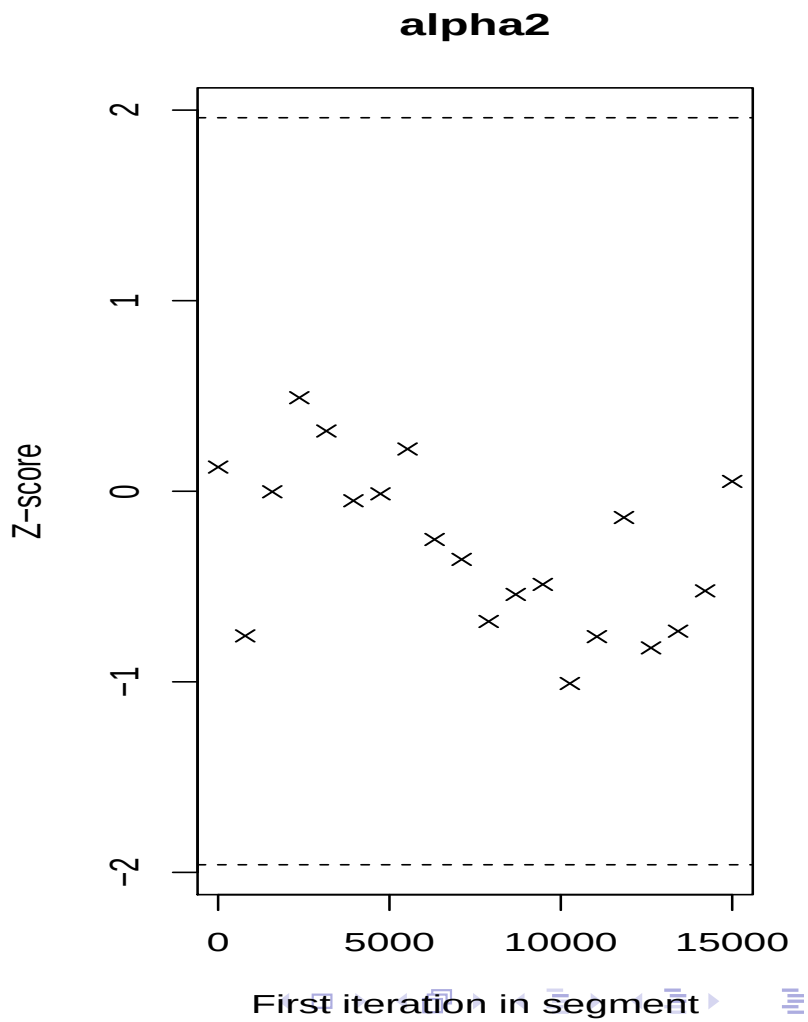
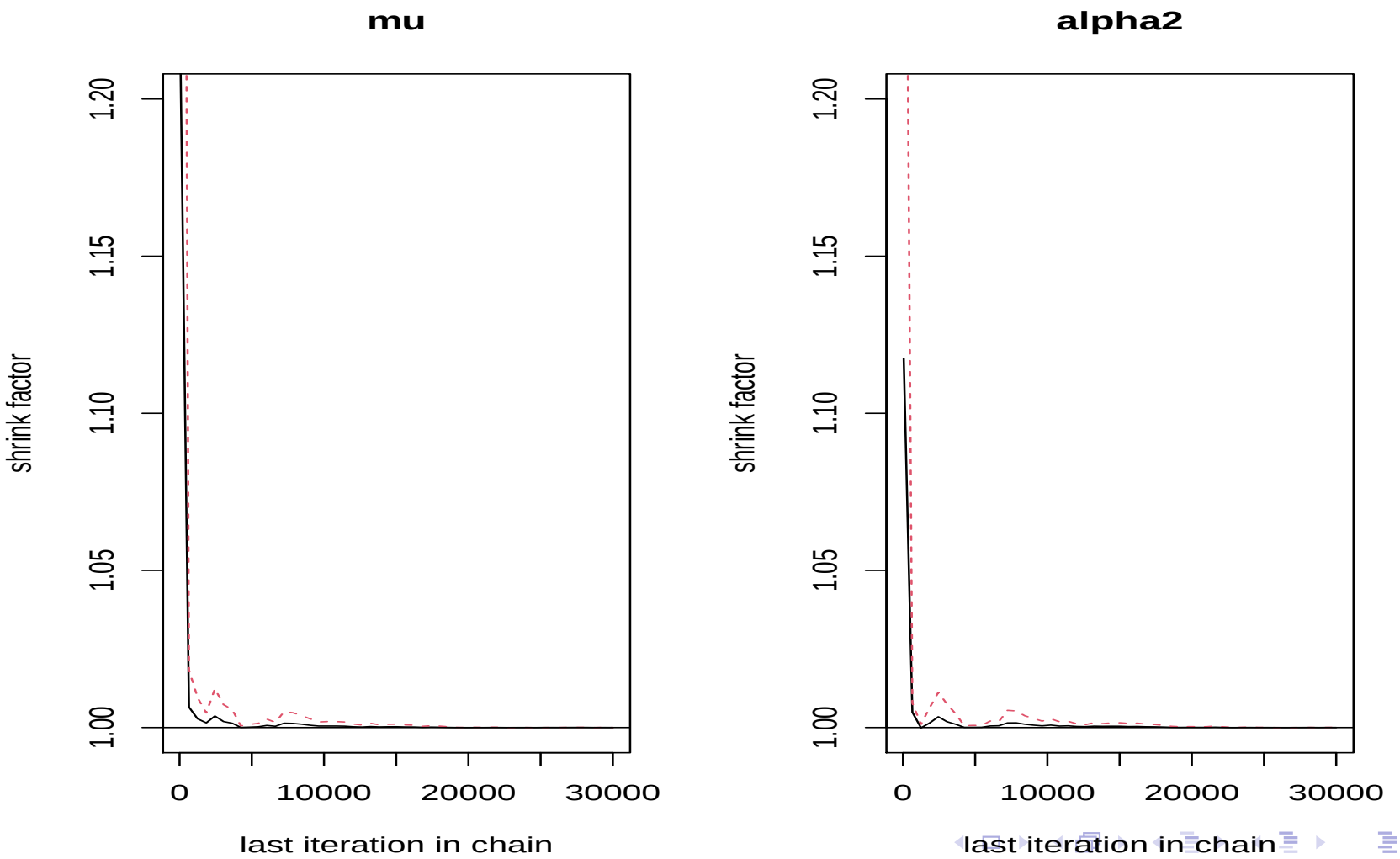
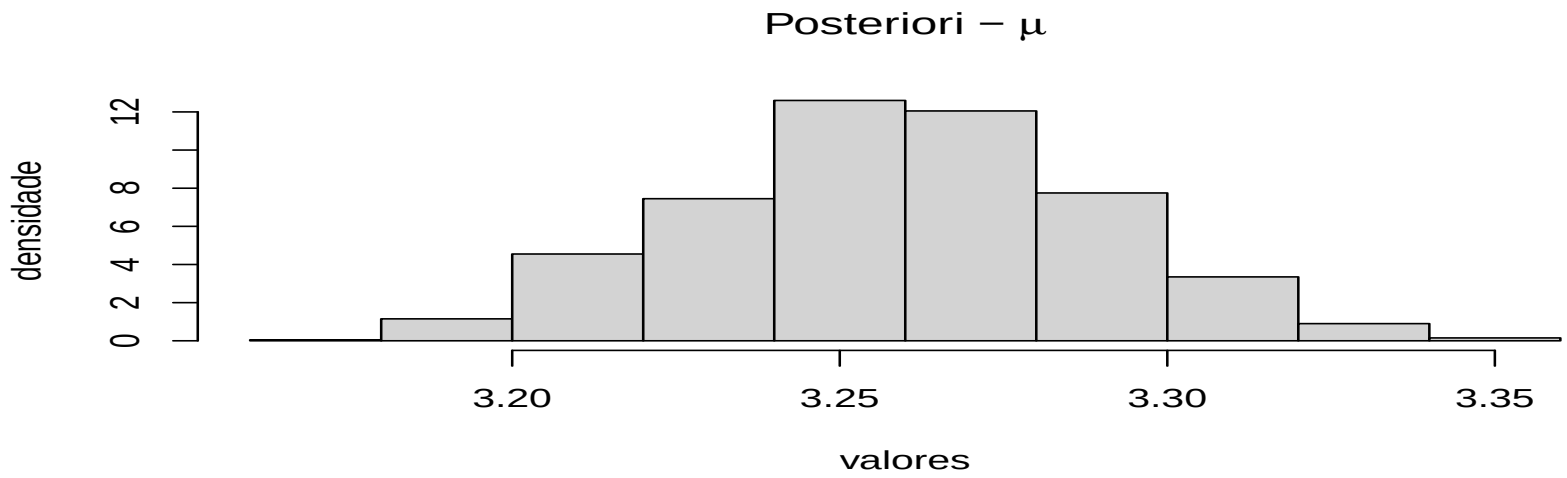
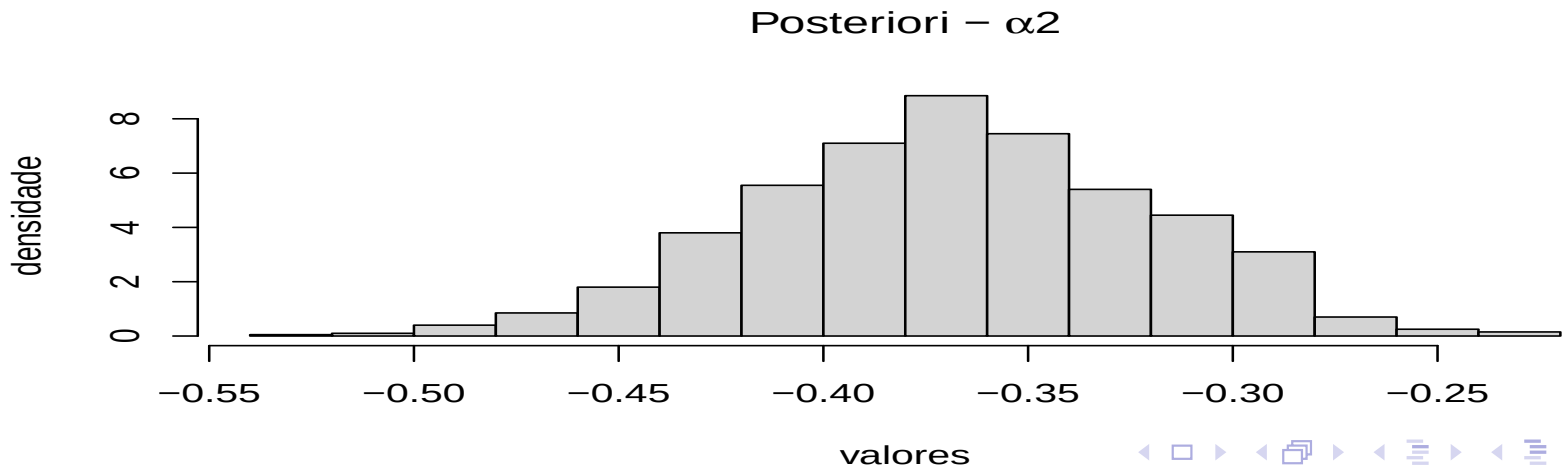


Gráfico da estat. de GR utilizando os três conj. de cadeias



Hist. da amostra válida para um dos conj. de cadeias





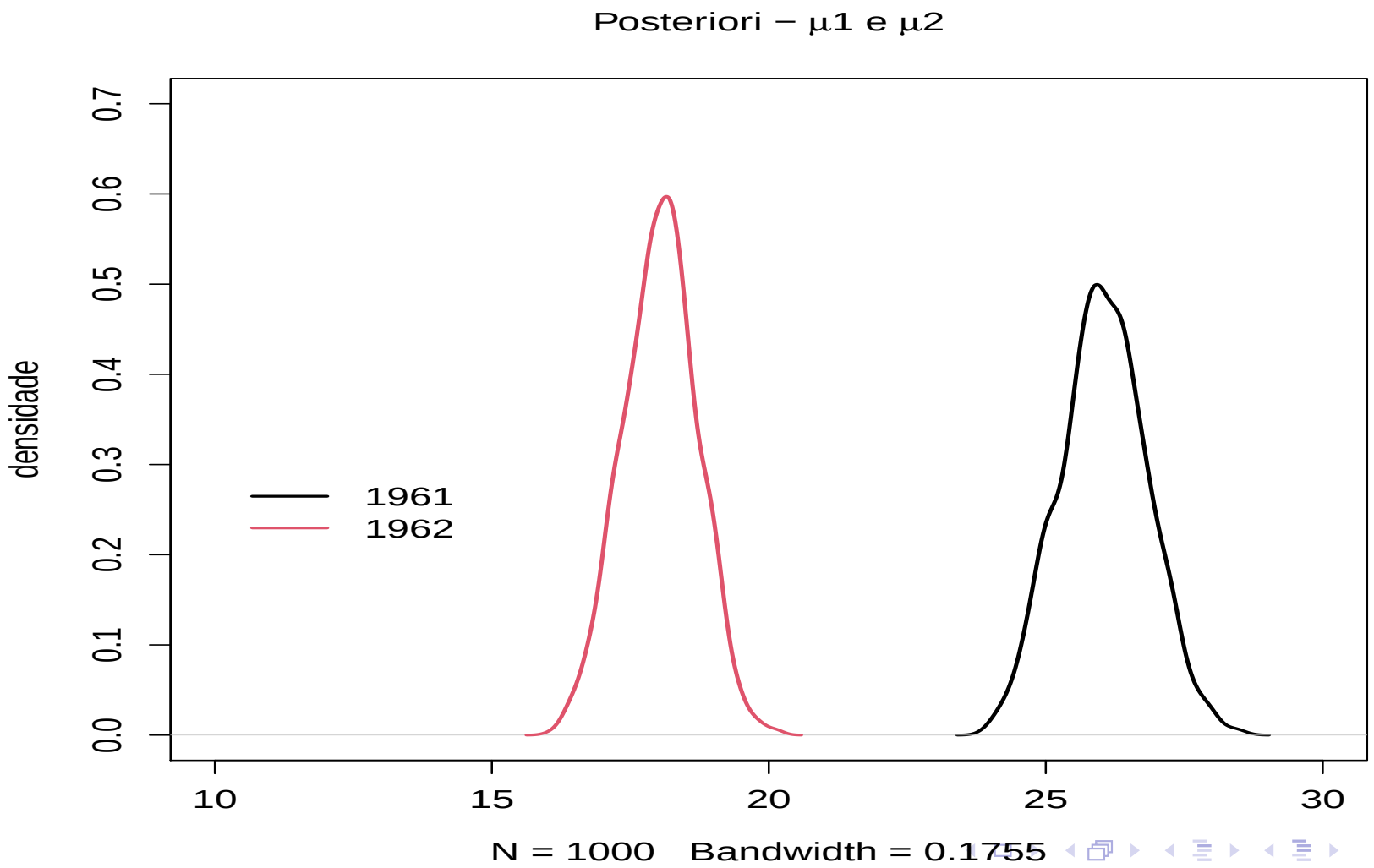
Estimativas Bayesianas para um dos conjuntos de cadeias

- Resumo

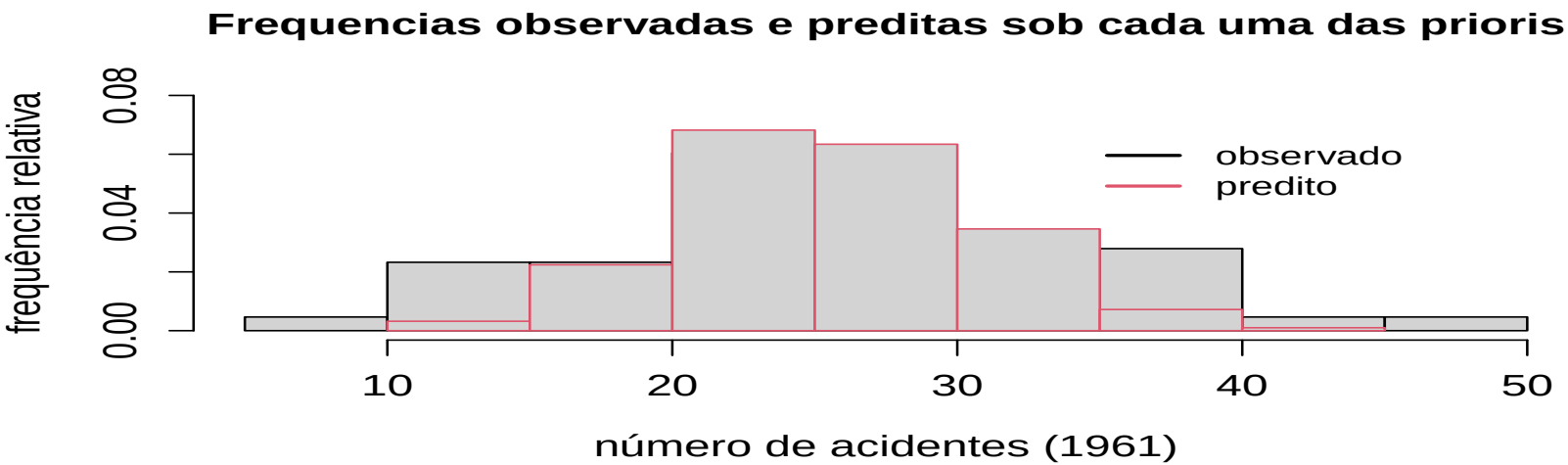
Parâmetro	EAP	DPAP	$IC_B(95\%)$	$HPD(95\%)$
μ	3,25	0,03	[3,20 ; 3,32]	[3,20 ; 3,32]
α_2	-0,37	0,05	[-0,46 ; -0,28]	[-0,46 ; -0,28]
$\mu_1 = exp(\mu)$	26,05	0,79	[24,58 ; 27,58]	[24,58 ; 27,58]
$\mu_2 = exp(\mu + \alpha_2)$	18,04	0,67	[16,73 ; 19,30]	[16,68 ; 19,22]

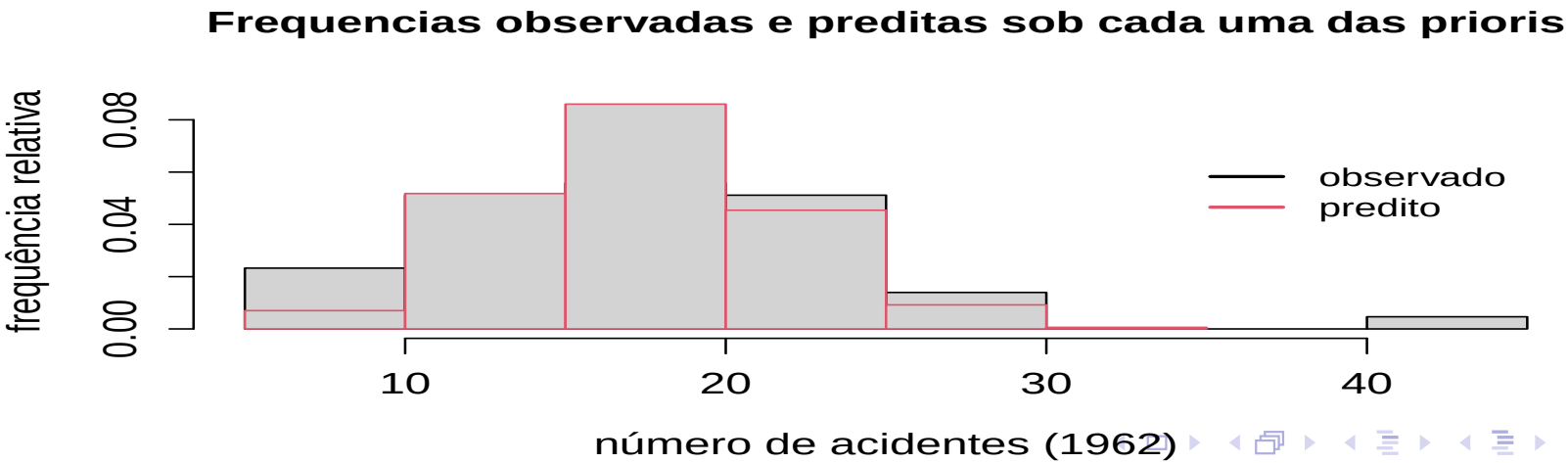
- Modelo com α_2: Deviance $= 654{,}8, p_D = 1,8; DIC = 656,6$. Modelo sem α_2: Deviance$= 716{,}7$ $p_D = 1,0; DIC = 717,8$.

Posterioris das médias de cada grupo



Distribuições observadas e preditas





Outro modelo para analisar o exemplo do número acidentes

- Superdispersão induzida pela introdução de efeitos aleatórios (veja várias opões de modelagem para dados de contagem aqui).

$$
\begin{aligned}
Y_{ji}|\boldsymbol{\beta}, b_j &\overset{ind.}{\sim} \text{Poisson}(\mu_{ji}), i = 1, 2, j = 1, ..., 43 \\
\ln \mu_{ji} &= \mu + \alpha_i + b_j, \alpha_1 = 0 \\
b_j &\overset{i.i.d.}{\sim} N(0, \sigma^2)
\end{aligned}
$$

em que $\boldsymbol{\beta} = (\mu, \alpha_2, \sigma^2)'$.

Outro modelo para analisar o exemplo do número acidentes

- Se $b_j \sim N(0, \sigma^2)$, então $e^{b_j} \sim$ log-normal$(0, \sigma^2)$
- Neste caso,

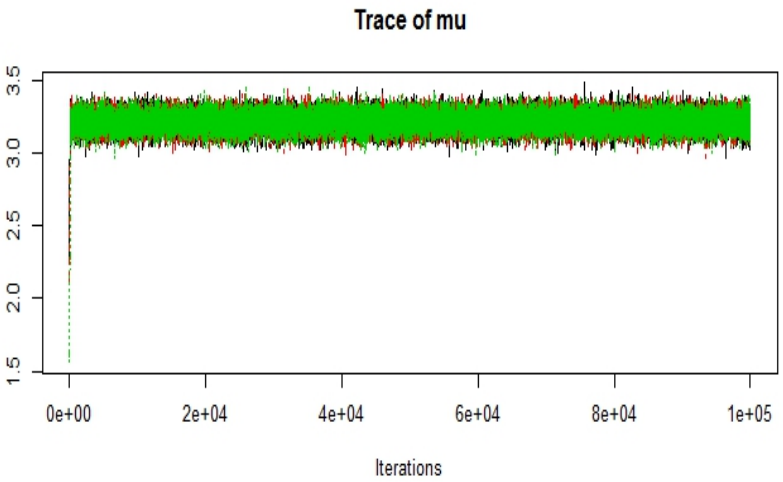
$$\mathcal{E}(Y_{ji}|\beta, \sigma^2) \equiv \mathcal{E}(Y_{ji}) = \mathcal{E}(\mathcal{E}(Y_{ji}|b_j)) = e^{\mu+\alpha_i}\mathcal{E}(e^{b_j}) = e^{\mu+\alpha_i}e^{\sigma^2/2} > e^{\mu+\alpha_i}$$

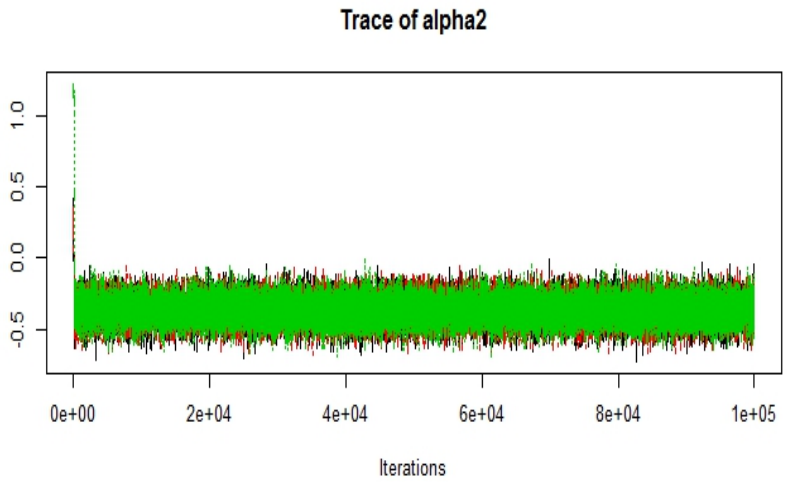
e

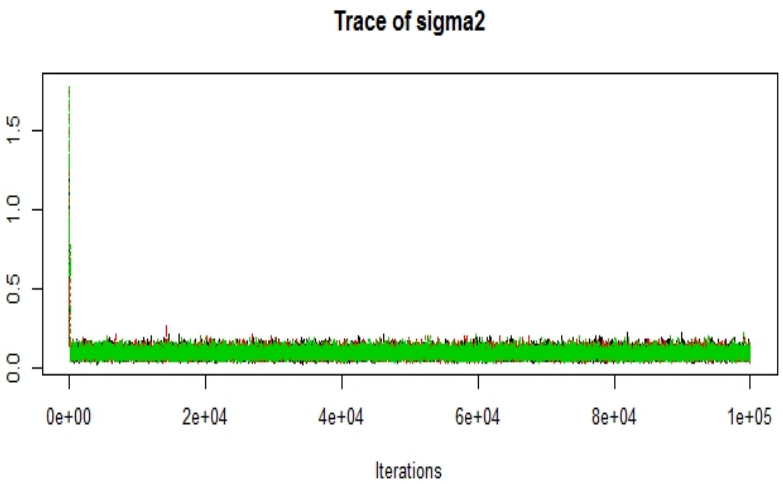
$$\begin{aligned}\mathcal{V}(Y_{ji}) &= \mathcal{V}(\mathcal{E}(Y_{ji}|b_j)) + \mathcal{E}(\mathcal{V}(Y_{ji}|b_j)) = e^{2(\mu+\alpha_i)}\mathcal{V}(e^{b_j}) + e^{\mu+\alpha_i}\mathcal{E}(e^{b_j}) \\ &= e^{2(\mu+\alpha_i)}(e^{\sigma^2} - 1)e^{\sigma^2} + e^{\mu+\alpha_i}e^{\sigma^2/2} > e^{\mu+\alpha_i}\end{aligned}$$

- Assim, o modelo em questão consegue contemplar uma variância maior do que aquela apresentada pelo modelo de regressão de Poisson.

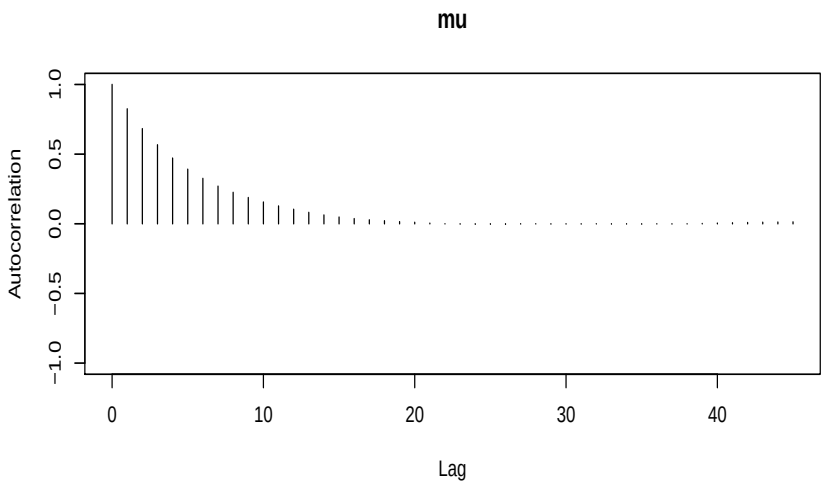
Traceplots para os três conjuntos de cadeias geradas

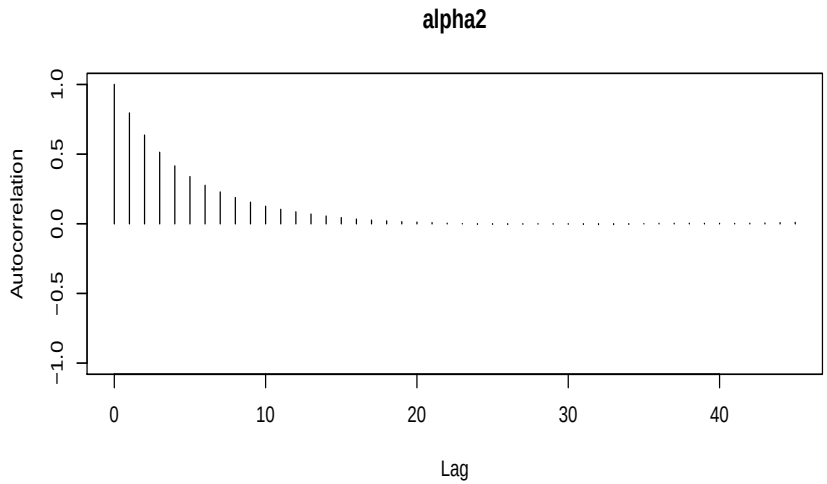






Autocorrelações para um dos conjuntos de cadeias





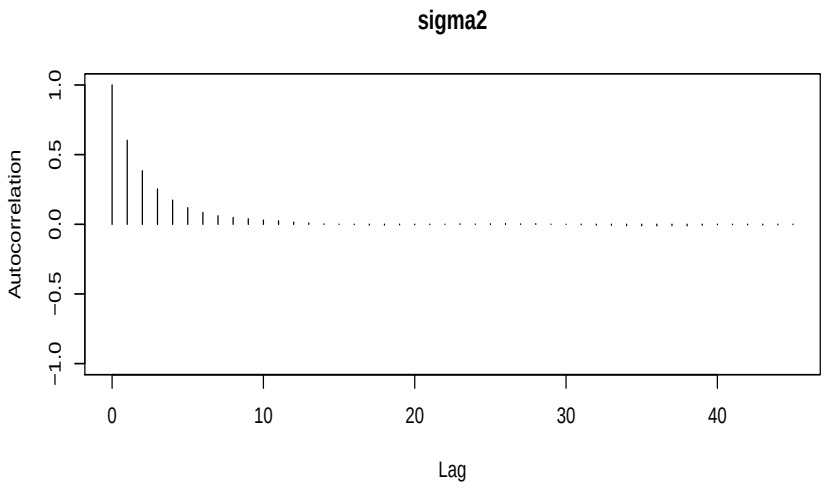
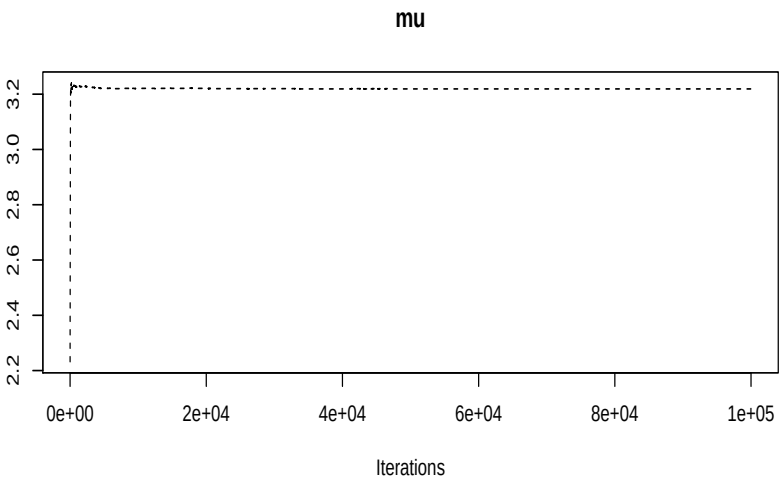
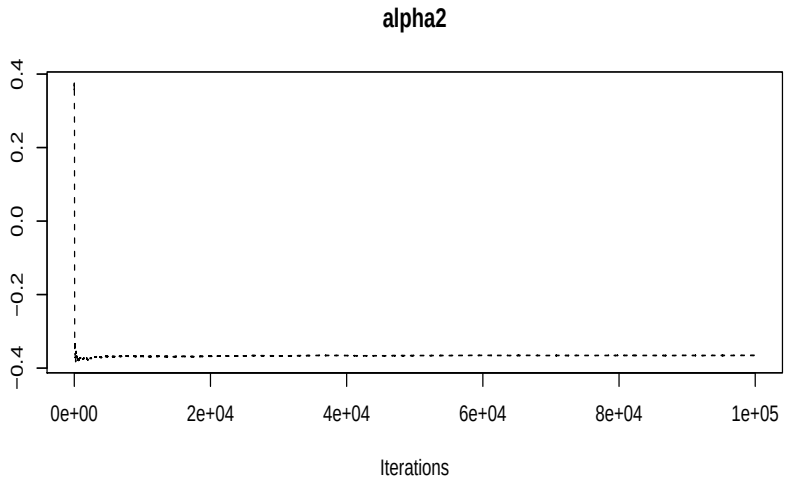


Gráfico das med. acum. para um dos conj. de cadeias





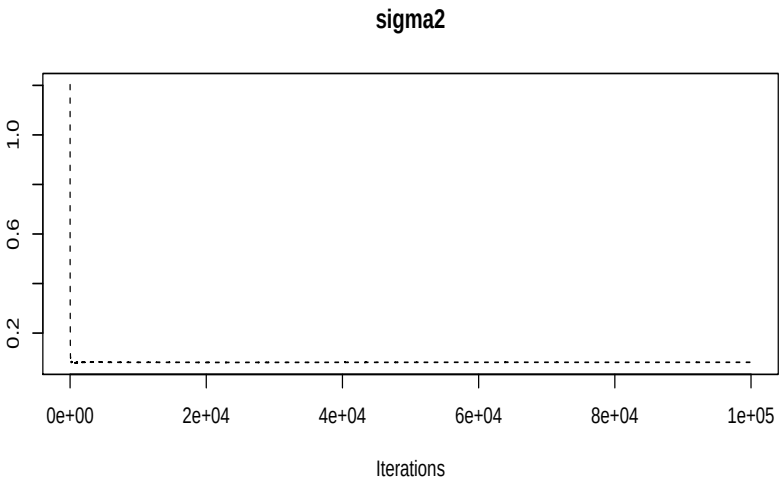
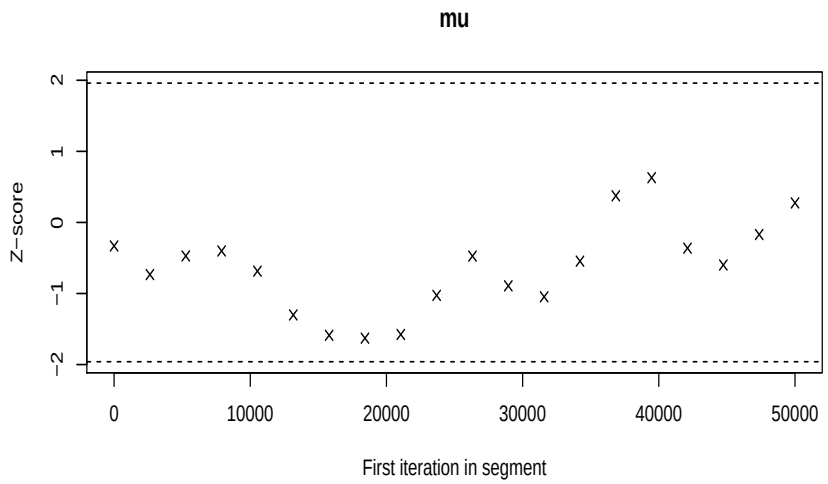
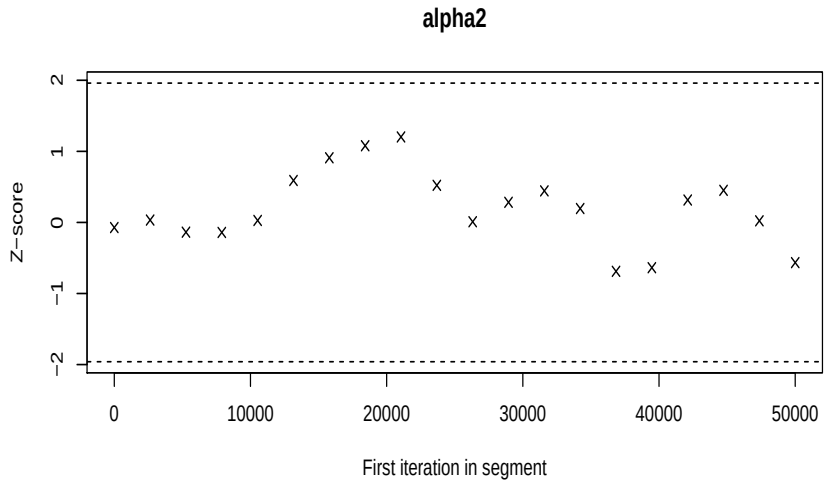


Gráfico da est. de Geweke para um dos conj. de cadeias





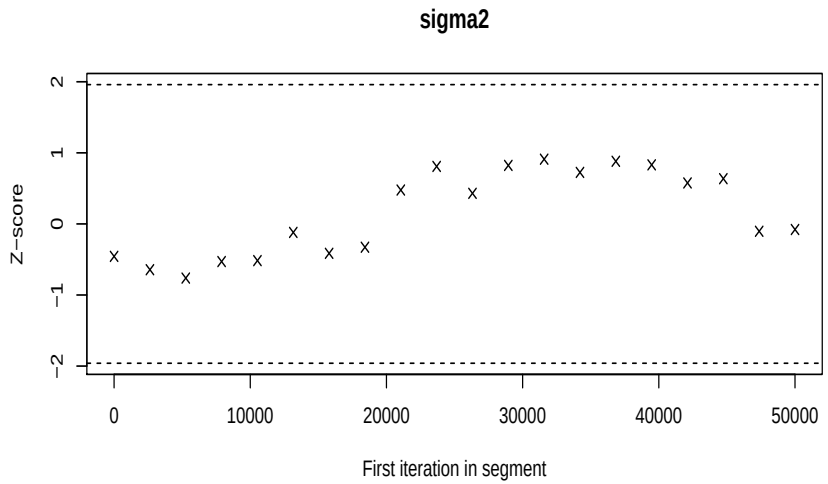
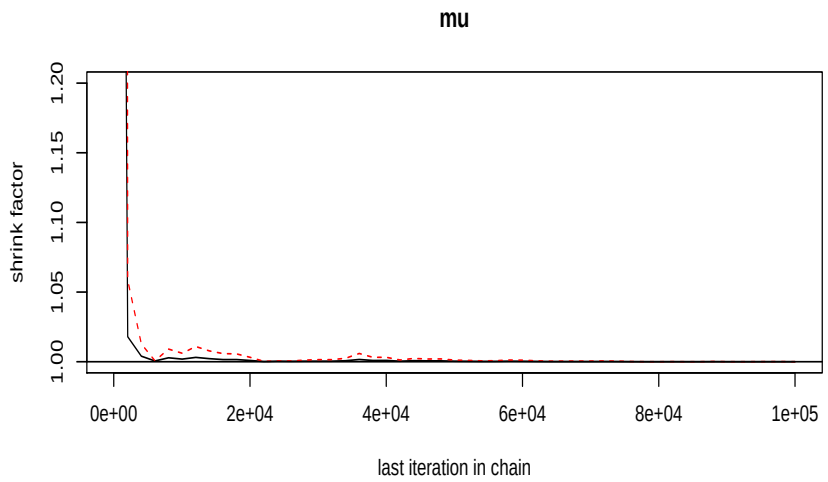
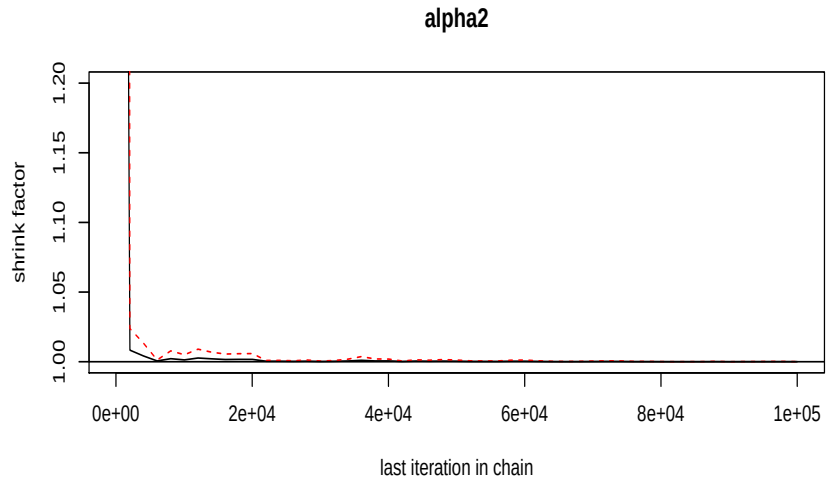
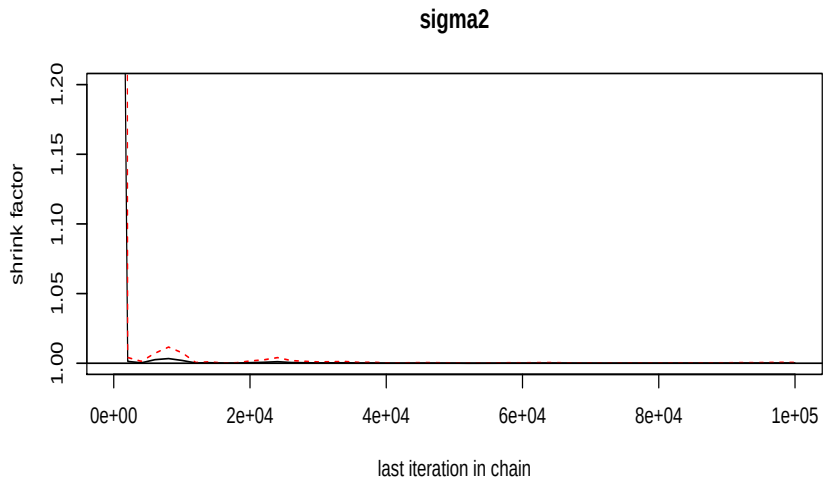


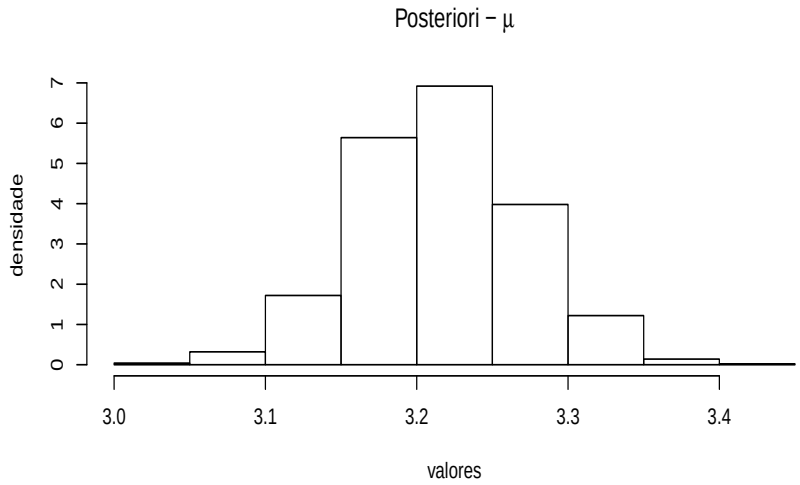
Gráfico da estat. de GR utilizando os três conj. de cadeias

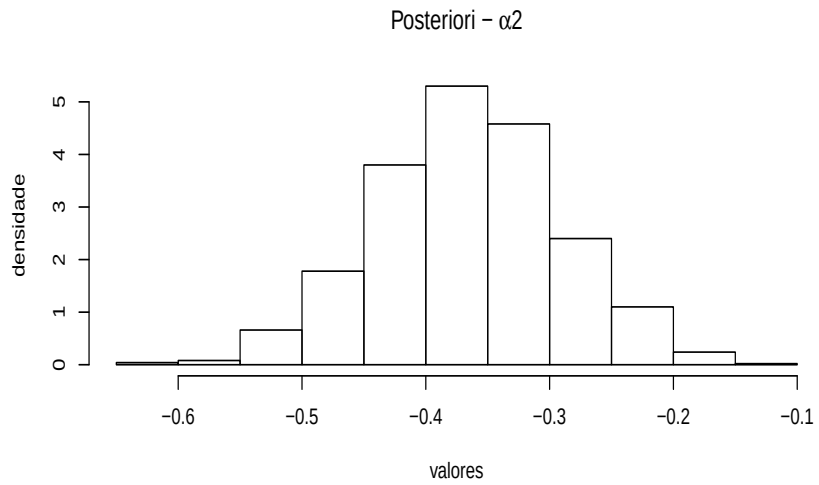


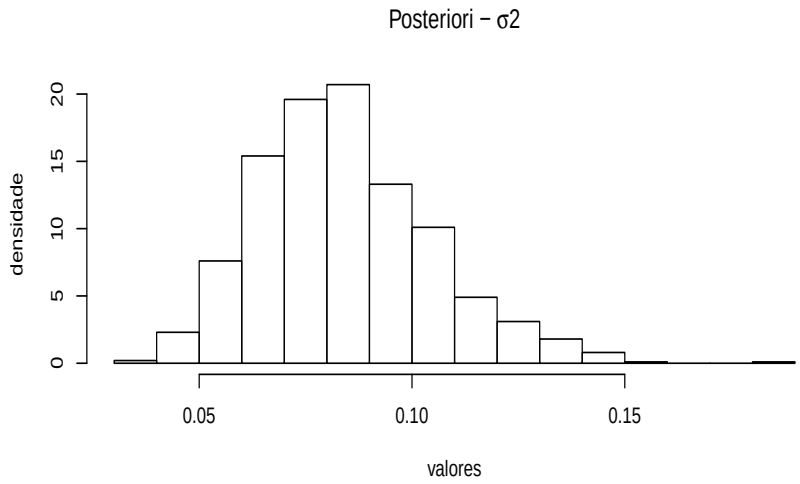




Hist. da amostra válida para um dos conj. de cadeias







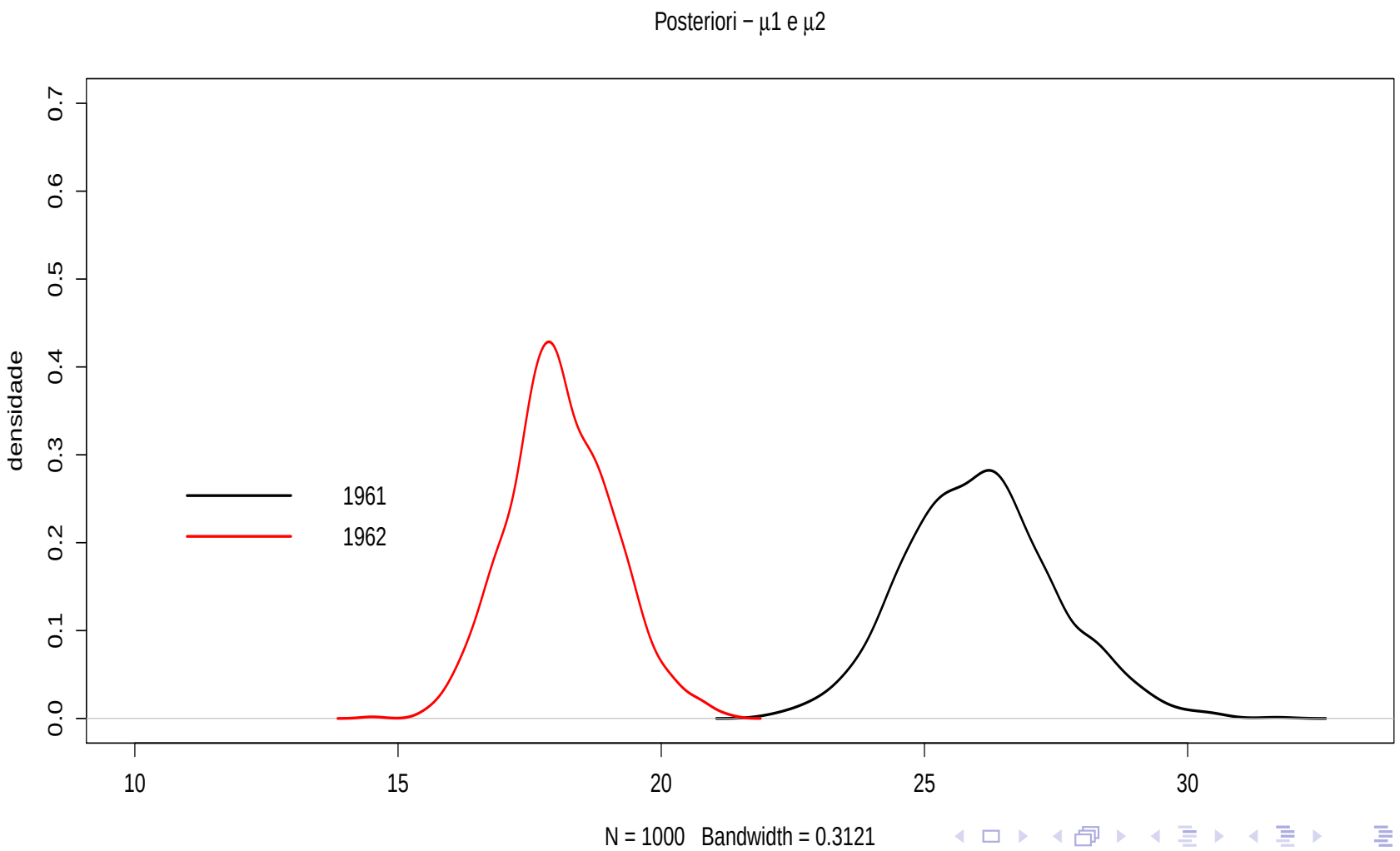
Estimativas Bayesianas para um dos conjuntos de cadeias

- Resumo

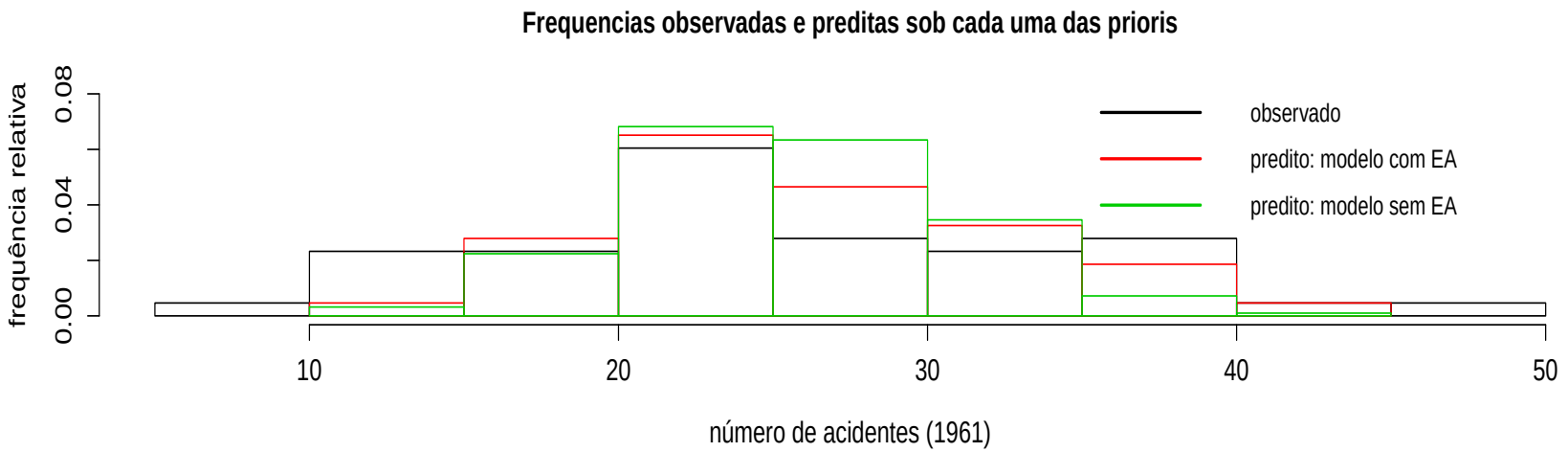
Parâmetro	EAP	DPAP	$IC_B(95\%)$	$HPD(95\%)$
μ	3,22	0,05	[3,12 ; 3,33]	[3,11 ; 3,31]
α_2	-0,37	0,08	[-0,51 ; -0,22]	[-0,51 ; -0,23]
σ^2	0,08	0,02	[0,05 ; 0,13]	[0,05 ; 0,13]
$\mu_1 = e^{\mu} e^{\sigma^2/2}$	26,07	1,43	[23,46 ; 29,10]	[23,50 ; 29,10]
$\mu_2 = e^{\mu+\alpha_2} e^{\sigma^2/2}$	18,10	0,99	[16,26 ; 20,16]	[16,24 ; 20,00]

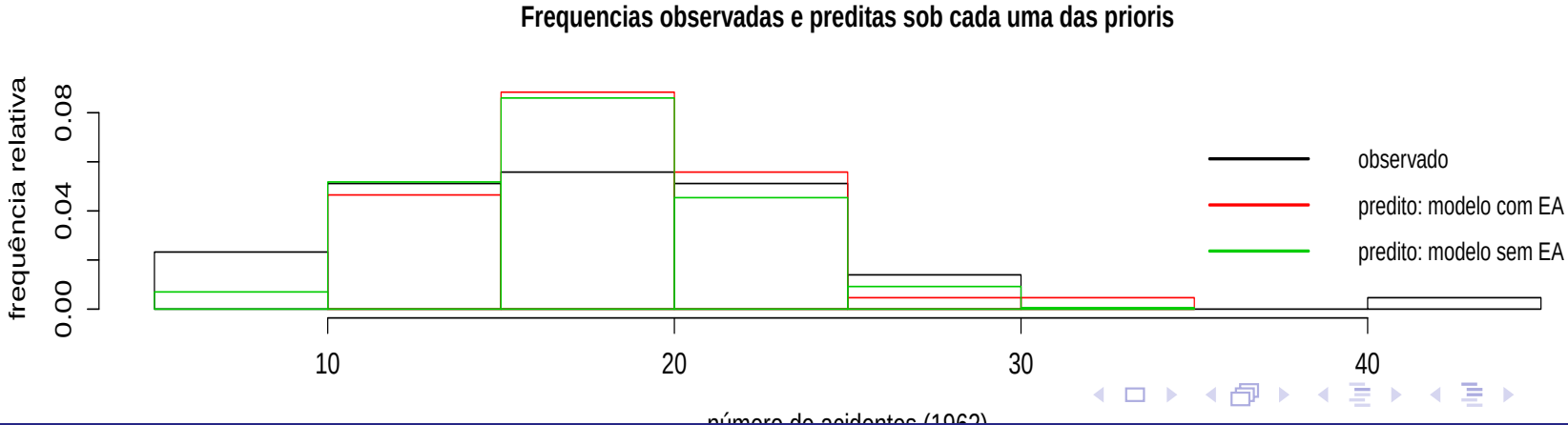
- Deviance $= 505{,}3, p_D = 55,30$; $DIC = 560,6$. Modelo sem efeito aleatório: Deviance $= 654{,}8, p_D = 1,8$; $DIC = 656,6$.

Posterioris das médias de cada grupo



Valores observados e distribuições preditivas





Recursos Computacionais

- Implementar o modelo diretamente em R (eventualmente utilizando funções/pacotes que permitem simular das condicionais completas diretamente ou indiretamente): rejeição adaptativa, slice sampling, Metropolis-Hastings adaptativo.
- Implementar o modelo diretamente nos programas: WinBUGS, OpenBUGS, JAGS, Stan; ou fazê-lo através dos respectivos pacotes que permitem usá-los diretamente no R, ou seja, o R2WinBUGS/R2OpenBUGS (WinBUGS/OpenBUGS), rjags, R2jags (JAGS), rstan (Stan).

Recursos Computacionais

- Pacotes no R: blme, MCMCglmm, MCMCpack, brms (modelos previamente implementados).
- Veja também: aqui, aqui, aqui.
- Compração entre inferência frequentista e bayesiana para o ajuste de modelos hierárquicos (link).

OpenBUGS: página, artigo, livro

- Programa OpenBUGS: permite ajustar modelos complexos usando diversos algoritmos do tipo MCMC. Trata-se da continuação do projeto BUGS (inicado com o WinBUGS). Várias modificações/adições foram implementadas (aqui), incluindo distribuições, funções, algoritmos de simulação e pacotes.
- Em geral, basta apenas fornecer o modelo (verossimilhança) e prioris.
- Limitações no uso de prioris impróprias (Jeffreys).
- Dispõe de mecanismos para inserir verossimilhanças e prioris que não são padrão (não constam em seu banco de dados).

OpenBUGS: página, artigo, livro

- Pode ser utilizado de modo mais simples através do pacote R2OpenBUGS do R.
- O diagnóstico de convergência pode ser facilmente realizado através do pacote coda, disponível no programa R.
- Continua sendo atualizado.

JAGS: página, artigo, livro

- Programa JAGS: permite ajustar modelos complexos usando diversos algoritmos do tipo MCMC. Semelhante, em diversos aspectos, ao OpenBUGS, embora possua funções que não estão no OpenBUGS/WinBUGS (aqui).
- Em geral, basta apenas fornecer o modelo (verossimilhança) e prioris.
- Limitações no uso de prioris impróprias (Jeffreys).
- Dispõe de mecanismos para inserir verossimilhanças e prioris que não são padrão (não constam em seu banco de dados).

JAGS: página, artigo, livro

- Planejado para:
  - Para ter um motor multiplataforma para a linguagem BUGS.
  - Para ser extensível, permitindo que os usuários escrevam suas próprias funções, distribuições e amostradores. Com efeito, é possível criar extensões (aqui, aqui).
  - Ser uma plataforma de experimentação de ideias em modelagem bayesiana.
- Continua sendo atualizado.

Stan: página, artigo, livro

- Diferente dos outros três programas. Além de realizar inferência Bayesiana com outros tipos de algoritmos de simulação, permite realizar inferência frequentista:
  - Inferência estatística bayesiana plena com amostragem MCMC (NUTS, HMC).
  - Inferência bayesiana aproximada com inferência variacional (ADVI).
  - Estimativas de máxima verossimilhança penalizada com otimização (L-BFGS).
- Apresenta diversas funções e distribuições de probabilidade que não estão disponíveis nos outros três programas (aqui).

Stan: página, artigo, livro

- Pode ser utilizado de modo mais simples através do pacote rstan do R.
- O diagnóstico de convergência pode ser facilmente realizado através do pacote coda, disponível no programa R, mas também há opções no "rstan".
- Continua sendo atualizado.

Recursos computacionais

- Existem versões dos quatro programas para Windows, Linux and OS.
- "Hints:"
  - Não usar WinBUGS.
  - OpenBUGS e JAGS tem sintaxes mais simples.
  - Em geral, o stan tende a ser mais lento para cada iteração mas, necessita de menos iterações para convergir (tradeoff).
  - Embora "não seja esperado", já obtive resultados melhores usando o stan.
- Veja também os projetos Nimble e MultiBUGS.