Inferência Estatística com Abordagem Bayesiana

Rosangela Helena Loschi [1]

[1]Departamento de Estatística
Universidade Federal de Minas Gerais

26 de outubro de 2023

Métodos Computationais para aproximar a distribuição *a posteriori*

- Métodos para aproximar a integral
  - métodos numéricos: quadraturas
  - método Monte Carlo
- Métodos de Reamostragem (SIR e Método da Rejeição)
- Métodos MCMC (*Gibbs sampler* e Metropolis-Hastings)

Métodos Computacionais

Obter as formas fechadas das distribuições *a posteriori* não é, em geral, fácil pois usualmente o computo das integrais no denominador da fórmula de Bayes é muito complexo e difícil.

Para solucionar este problema podemos:

- Aproximar a integral do denominador via
  - métodos numéricos de integração tipo quadraturas
  - método Monte Carlo
- Obter uma amostra da distribuição *a posteriori*
  - métodos de reamostragem (Rejeição, SIR)
  - métodos Monte Carlo via Cadeias de Markov (amostrador de Gibbs, Metropolis-Hastings)

Exemplo Motivador: Na análise de dados de trissomia cromossômica distribuição *a posteriori* para a proporção de não-disjunção na meiose I ϕ é

$$\pi(\phi|\mathbf{y}) = \frac{[\theta_1(\phi)]^{y_1}[\theta_2(\phi)]^{y_2}[\theta_3(\phi)]^{y_3}\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\int_0^1 [\theta_1(\phi)]^{y_1}[\theta_2(\phi)]^{y_2}[\theta_3(\phi)]^{y_3}\phi^{\alpha-1}(1-\phi)^{\beta-1}d\phi}, \quad \phi \in (0,1).$$

- Apesar de ser um modelo simples e depender de um único parâmetro ϕ a integral do denominador não pode se obtida analiticamente.
- Como fazer inferência *a posteriori* neste caso?
  - aproximando a integral: quadraturas, Monte Carlo
  - obtendo uma amostra da disttribuição *a posteriori*: SIR, Método da Rejeição, algoritmo de Metropolis-Hastings

Aproximando a Integral

- Suponha que tenhamos como meta determinar a distribuição

$$h(\theta) = \frac{f(\theta)}{\int_a^b f(\theta) d\theta}.$$

- Admita que a integral $I = \int_a^b f(\theta) d\theta$ no denominador seja de difícil solução analítica.
- Nossa meta é fornecer uma estimativa $\hat{I}$ para I.

Aproximando a Integral: Métodos de Quadraturas

- Os métodos de quadratura são métodos determinísticos, isto é, não tem erro Monte Carlo envolvido.
- Uma aproximação de I obtidas da seguinte forma
  - selecionamos pontos de avaliação θ_i, $i = 0, \ldots, n$ no intevalo de integração (a, b) tal que

$$a = \theta_0 < \theta_1 < \cdots < \theta_n = b$$

.

  - avaliamos f em cada ponto θ_i obtendo-se $f(\theta_i)$
  - a estimativa de I é

$$\hat{I} = \sum_{i=1}^{n} f(\theta_i)\omega_i$$

  onde ω_i é o peso de cada componente $f(\theta_i)$.
  - Note que estamos aproximando a integral pela soma dos retângulos de base ω_i e altura $f(\theta_i)$.
  - os métodos de quadratura se diferem pela forma como os pesos e os pontos de avaliação são escolhidos.

Aproximando a Integral: Métodos de Quadraturas

Regra de Newton-Cotes:

P1: o intervalo de integração (a, b) é dividido em n partes iguais de forma que $\theta_i,\ i = 0, \ldots, n$ estejam equi-espaçados. Assim.

$$\theta_0 = a,\ \theta_1 = a + (b-a)/n,\ \theta_2 = a + 2(b-a)/n, \ldots,\ \theta_n = b$$

P2: A função f é avaliada no ponto médio de cada itervalo que para o intervalo (θ_{i-1}, θ_i) é

$$(\theta_{i-1} + \theta_i)/2 = a + (2i-1)\frac{(b-a)}{2n}.$$

P3: Atribui-se peso igual para cada ponto de avaliação e este peso é $(b-a)/n$

P4: A integral é aproximada por

$$\hat{I}_{NC} = \frac{(b-a)}{n} \sum_{i=1}^{n} f\left(a + \frac{(2i-1)(b-a)}{2n}\right)$$

Aproximando a Integral: Métodos de Quadraturas

Exemplo: Suponha que queiramos resolver a integral

$$I = \int_0^{\infty} e^{-2\theta} d\theta.$$

- Sabemos que o resultado exato da integral é $I = 2$.
- Aproximando: Consideramos um intervalo em que a função a ser integrada tem massa probabilistica significativa: $(0, 6)$ $(P(0 < \theta < 6) = 0,999)$
- Tomemos $n = 2$. Neste caso, $\theta_1 = 3$. Os pontos médios seriam $1,5$ e $4,5$. A amplitude de cada intervalo seria 3.
- $\hat{I} = 3\left[exp\{-2 * 1,5\} + exp\{-2 * 4,5\}\right] = 0,149731$
- Tomemos $n = 3$. Neste caso, $\theta_1 = 2$ e $\theta_2 = 4$. Os pontos médios seriam 1, 3 e 5. A amplitude de cada intervalo seria 2.
- $\hat{I} = 2\left[exp\{-2 * 1\} + exp\{-2 * 3\} + exp\{-2 * 5\}\right] = 0,275719$

Aproximando a Integral: Métodos de Quadraturas

- Se particionamos este intevalo em $n = 10$ subintervalos de mesmo tamanho. A amplitude de cada intervalo seria $0,6$.

i	θ_i	Ponto médio(P_i)	Exp{-2* P_i}
1	0,6	0,3	0,548812
2	1,2	0,9	0,165299
3	1,8	1,5	0,049787
4	2,4	2,1	0,014996
5	3	2,7	0,004517
6	3,6	3,3	0,00136
7	4,2	3,9	0,00041
8	4,8	4,5	0,000123
9	5,4	5,1	3,72E-05
10	6	5,7	1,12E-05

- Se n=10 temos que Î =0,471211
- Se n=20 temos que Î =0,492575
- Se n=10 temos que Î =0,500243

Aproximando a Integral: Métodos de Quadraturas

Regra Trapezóidal: É uma modificação da Regra de Newton-Cotes

P1: O intervalo de integração (a, b) é dividido em n partes iguais de forma que θ_i, $i = 0, \ldots, n$ estejam equi-espaçados.

P2: Para pontos no interior do intervalo, isto é, $\theta_1, \ldots, \theta_{n-1}$

- A função f é avaliada no ponto médio de cada itervalo que para o intervalo (θ_{i-1}, θ_i) é

$$(\theta_{i-1} + \theta_i)/2 = a + (2i - 1)\frac{(b-a)}{2n}.$$

- Atribui-se peso igual para cada ponto de avaliação e este peso é $(b - a)/n$.

P3: Para os pontos a e b no extremo do intevalo calcula-se $f(a)$ e $f(b)$ e atribui-se metade do peso $((b-a)/2n)$ a estes componentes.

P4: A integral é aproximada por

$$\hat{I}_T = \frac{(b-a)}{n}\left\{f(a)/2 + \sum_{i=1}^{n} f\left(a + \frac{(2i-1)(b-a)}{2n}\right) + f(b)/2\right\}$$

Aproximando a Integral: Métodos de Quadraturas

Regra Simpson: É uma modificação da Regra de Newton-Cotes.

- Aproxima a área sob a curva situada entre θ_i e θ_{i+2} pela área sob uma parábola passando pelos pontos θ_i, θ_{i+1} e θ_{i+2}.

P1: O intervalo de integração (a, b) é dividido em n partes iguais de forma que θ_i, $i = 0, \ldots, n$ estejam equi-espaçados.
P2: Para pontos no interior do intervalo, isto é, $\theta_1, \ldots, \theta_{n-1}$ os pesos são alternados entre $4/3 * h$ e $2/3 * h$ onde $h = (b - a)/n$.
P3: Para os pontos a e b no extremo do intevalo calcula-se $f(a)$ e $f(b)$ e atribui-se peso$(h/3)$ a estes componentes.
P4: A integral é aproximada por

$$\begin{aligned} \hat{I}_T &= \frac{(b-a)}{3n}\left\{ f(a) + 4\sum_{i=1}^{n/2} f\left(a + \frac{(4i-1)(b-a)}{2}\right)\right. \\ &\left. +2\sum_{i=1}^{n/2} f\left(a + \frac{(4i-3)(b-a)}{2n}\right) + f(b)\right\} \end{aligned} \quad (1)$$

Aproximando a Integral: Métodos de Quadraturas

- Quando maior o valor de n melhor a aproximação.
- Para o caso unidimensional a escolha de n na ordem de 10^2 já fornece uma aproximação razoável.
- no caso multidimensional o raciocínio é analogo. Se queremos $I = \int_a^b \int_c^d f(x,y)dxdy$ devemos:
  - dividir o intervalo (a,b) em n subintervalos de mesmo tamanho $(b-a)/n$ e dividir o intervalo (c,d) em m intervalos de mesmo tamanho $(d-c)/m$.
  - isto divide a região de integração em $n * m$ areas. Avaliamos f em cada ponto médio da área.
  - Usando a Regra de Newton-cotes, aproximamos a integral por

$$\hat{I}_{NC} = \frac{(b-a)}{n}\frac{(d-c)}{m} \sum_{i=1}^{n}\sum_{j=1}^{m} f\left(a + \frac{(2i-1)(b-a)}{2n}, c + \frac{(2j-1)(d-c)}{2m}\right).$$

Aproximando a Integral: Métodos Monte Carlo

- A idéia básica é escrever a integral $I = \int_a^b f(\theta)d\theta$ como um valor esperado com respeito a alguma distribuição.
- A integral I pode ser re-escrita como

$$\begin{aligned} I &= \int_a^b f(\theta)d\theta = \int_a^b \frac{f(\theta)}{p(\theta)} p(\theta)d\theta \\ &= E_{p(\theta)}\left(\frac{f(\theta)}{p(\theta)}\right) \end{aligned}$$

- a aproximação para I é $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\theta_i)}{p(\theta_i)}$

Algoritmo:

P1 Gere uma amostra $\theta_1, \ldots, \theta_n$ da distibuição $p(\theta)$

P2 calcule $\frac{f(\theta_1)}{p(\theta_1)}, \ldots, \frac{f(\theta_n)}{p(\theta_n)}$

P3 calcule a média $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\theta_i)}{p(\theta_i)}$

Aproximando a Integral: Métodos Monte Carlo

- se $p(\theta)$ é uma $Uniforme(a, b)$, a integral I pode ser re-escrita como

$$\begin{aligned} I &= \int_a^b f(\theta) d\theta \\ &= \int_a^b (b-a) f(\theta) \frac{1}{(b-a)} d\theta \\ &= E_{U(a,b)} \left((b-a) f(\theta) \right) \end{aligned}$$

- a aproximação para I

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} (b-a) f(\theta_i)$$

Aproximando a Integral: Métodos Monte Carlo

Algoritmo:

P1 Gere uma amostra $\theta_1, \ldots, \theta_n$ da distibuição $U(a, b)$

P2 calcule $f(\theta_1), \ldots, f(\theta_n)$

P3 calcule a média $\bar{f} = 1/n \sum_{i=1}^{n} f(\theta_i)$

P4 calcule $\hat{I} = (b - a)\bar{f}$

- Como $\hat{I}$ é uma média, sua precisão para estimar I é dada por sua variância

$$Var(\bar{I}) = \frac{(b-a)^2}{n} Var(f(\theta))$$

Exemplo: se quisermos resolver a integral $\int_0^1 e^x dx$, pelo método Monter Carlo esta integral é aproximada por $\hat{I} = 1/n \sum_{i=1}^{n} e^{x_i}$ onde x_i são valores gerados da distribuição uniforme no intervalo $(0, 1)$.

Aproximando a Integral: Métodos Monte Carlo

Integral de exp(x) no intervalo (0,1) ➔ valor=1,71828

Theta_i	Exp(theta_i)	I^
0	1	1,732389
0,1	1,105171	
0,2	1,221403	
0,3	1,349859	
0,4	1,491825	
0,5	1,648721	
0,6	1,822119	
0,7	2,013753	
0,8	2,225541	
0,9	2,459603	
1	2,718282	
soma	19,05628	

Aproximando a Integral: Métodos Monte Carlo

No caso multivariado, e usando a distribuição de referência o sendo a uniforme

$$I = \int_a^b \int_c^d f(\theta, \alpha) d\theta d\alpha.$$

Algoritmo:

P1 Gere uma amostra $(\theta_1, \alpha_1), \ldots, (\theta_m, \alpha_m)$ da distibuição uniforme no retângulo $(a, b) \times (c, d)$.

P2 calcule $f(\theta_1, \alpha_1), \ldots, f(\theta_m, \alpha_m)$

P3 calcule a média $\bar{f} = 1/m \sum_{i=1}^{m} f(\theta_i, \alpha_i)$

P4 calcule $\hat{I} = (b - a)(d - c)\bar{f}$

Métodos Computacionais: Métodos de Reamostragem

Estes métodos fornecem uma amostra da distribuição *a posteriori* quando somente o kernel da distribuição é conhecido.

- Suponha que tenhamos como meta obter uma amostra da função

$$h(\theta) = \frac{f(\theta)}{\int_{\Theta} f(\theta) d\theta}.$$

- Suponha que uma amostra de $g(\theta)$ é facilmente obtida.
- Se pode utilizar as amostras de $g(\theta)$ para obter amostras de $h(\theta)$?

A resposta é sim e temos duas propostas para obter as amostras de $h(\theta)$: o método da rejeição e o método SIR.

Métodos Computacionais: Método da Rejeição

Para aplicar método da rejeição é necessário encontramos uma constante $M > 1$ tal que

$$\frac{f(\theta)}{g(\theta)} \leq M \;\; \forall \theta \in \Theta, \Leftrightarrow M = max \left\{ \frac{f(\theta)}{g(\theta)} : \theta \in \Theta \right\}$$

ou seja, necessitamos encapsular a função $f(\theta)$ de interesse.

* O algoritmo
  - $P1$ Gere θ^* da função $g(\theta)$.
  - $P2$ Gere $u \sim \text{Uniforme}(0, 1)$.
  - $P3$ Se $u \leq \frac{f(\theta^*)}{Mg(\theta^*)} \Rightarrow$ aceitamos θ^*.
  - $P4$ Volte a $P1$ até obter a amostra com o tamanho que deseja.

- O número esperado de valores aceitos em n rodadas do algoritmo é n/M.
- Se $M \approx 1$ temos um melhor desempenho do algoritmo.
- Se $M = 1$ aceitariamos todos os candidatos mas neste caso $f() = g()$ (não faz sentido!).

Por que?

Métodos Computacionais: Método da Rejeição

Por que $P(Aceitar\ \theta^*) = 1/M$! (Se f() é uma fdp)

$$\begin{aligned} P(Aceitar\ \theta^*) &= P\left(U \leq \frac{f(\theta^*)}{Mg(\theta^*)}\right) = E\left(1\left\{U \leq \frac{f(\theta^*)}{Mg(\theta^*)}\right\}\right) \\ &= E\left[E\left(1\left\{U \leq \frac{f(\theta^*)}{Mg(\theta^*)}\right\} \mid \theta^*\right)\right] \\ &= E\left[P\left(U \leq \frac{f(\theta^*)}{Mg(\theta^*)} \mid \theta^*\right)\right] \end{aligned}$$

$$U \sim U(0,1) \Rightarrow P\left(U \leq \frac{f(\theta^*)}{Mg(\theta^*)} \mid \theta^*\right) = \frac{f(\theta^*)}{Mg(\theta^*)}$$

$$P(Aceitar\ \theta^*) = \frac{1}{M} E_{\theta^*}\left[\frac{f(\theta^*)}{g(\theta^*)}\right] = \frac{1}{M}\int \frac{f(\theta^*)}{g(\theta^*)} g(\theta^*) d\theta^*$$

Métodos Computacionais: Método da Rejeição

* Se $h(\theta) = \pi(\theta \mid x)$ devemos encontrar M tal que

$$M = \max_{\theta} \left\{ \frac{f(x \mid \theta)\pi(\theta)}{g(\theta)} \right\}$$

* $g(\theta)$ é a distribuição de importância
  + deve ter caudas mais pesadas que $\pi(\theta \mid x)$.
  + devemos poder gerar amostras de $g(\theta)$ facilmente.
  + a amostra gerada de $g(\theta)$ deve ser suficientemente grande para cobrir todo o espaço paramétrico

Métodos Computacionais: Método da Rejeição

- **Exemplo:** A distribuição de importância $g(\theta)$ pode ser a distribuição *a priori* para θ.
- Neste caso

$$\frac{f(\theta)}{g(\theta)} = \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{\pi(\theta)} = f(\boldsymbol{x} \mid \theta) \tag{2}$$

- Devemos encontrar M tal que $\frac{f(\theta)}{g(\theta)} \leq M$ para todo θ.

$$\Rightarrow M = f(\boldsymbol{x} \mid \hat{\theta}_{MV}).$$

Neste caso a razão

$$\frac{f(\theta)}{Mg(\theta)} = \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{f(\boldsymbol{x} \mid \hat{\theta}_{MV})\pi(\theta)} = \frac{f(\boldsymbol{x} \mid \theta)}{f(\boldsymbol{x} \mid \hat{\theta}_{MV})} \tag{3}$$

Métodos Computacionais: Método da Rejeição

Neste caso o algoritmo torna-se:

O algoritmo

*P*1 Gere θ^* da função *a prior* $\pi(\theta)$.

*P*2 Gere $u \sim \text{Uniforme}(0, 1)$.

*P*3 Se $u \leq \frac{f(\boldsymbol{x}|\theta^*)}{f(\boldsymbol{x}|\hat{\theta}_{MV})} \Rightarrow$ aceitamos θ^*.

*P*4 Volte a *P*1 até obter a amostra com o tamanho que deseja.

- Esta escolha pode ser arriscada e gerar péssimos resultados se a distribuição *a priori* e a função de verossimilhança não forem razoavelmente concordantes.
- Neste caso, os valores de θ candidatos gerados de $\pi(\theta)$ podem estar muito distantes da região onde a distribuição *a posteriori* põe massa significativa.
- Isto levaria a um percentual grande de rejeições.

Métodos Computacionais: SIR

Se M não existe uma estratégia possível é utilizar o médoto SIR (*Sampling Importance Resampling*) ou *Bootstrap* Bayesiano (Rubin, 1979)

* O algoritmo

  P1 Gere uma amostra $\theta_1, \ldots, \theta_J$ da função de importância $g(\theta)$.

  P2 Para cada θ_i, $i = 1, \ldots, J$, calcule

$$\omega_i = \frac{f(\theta_i)}{g(\theta_i)} \text{ e os pesos } q_i = \frac{\omega_i}{\sum_i^J \omega_i}$$

  P3 Selecione uma amostra $\theta_1^*, \ldots, \theta_T^*$, com reposição, da amostra $\theta_1, \ldots, \theta_J$ assumindo que $P(\theta = \theta_i) = q_i$

    + gere $u \sim \text{Uniforme}(0, 1)$
    + Se $u \in (0, q_1)$ selecione θ_1
    + Se $u \in (q_1, q_1 + q_2)$ selecione θ_2 ....

A amostra $\theta_1^*, \ldots, \theta_T^*$ pode ser selecionada com reposição e se recomenda que $T = J/20$.

Métodos Computacionais: Dados Genéticos

- Implementamos o SIR e um método de quadratura (regra se Simpson) para obter uma aproximação da distribuição *a posteriori* no estudio para dados genéticos.
- a distribuição *a posteriori* para ϕ é

$$\pi(\phi|\mathrm{y}) = \frac{[\theta_1(\phi)]^{y_1}[\theta_2(\phi)]^{y_2}[\theta_3(\phi)]^{y_3}\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\int_0^1 [\theta_1(\phi)]^{y_1}[\theta_2(\phi)]^{y_2}[\theta_3(\phi)]^{y_3}\phi^{\alpha-1}(1-\phi)^{\beta-1}d\phi}, \quad \phi \in (0,1).$$

- Os resultados são muito similares.

Table: Estimadores de Bayes *a posteriori*

Especificação a Priori					*Método SIR*			*Regra de Simpson*		
α	β	Média	Variância	Moda	Média	Variância	Moda	Média	Variância	Moda
1.0	1.0	0.500	0.080	–	0.6571	0.0311	0.6690	0.6549	0.0305	0.6553
2.0	1.0	0.667	0.060	–	0.7043	0.0268	0.6842	0.7015	0.0272	0.7244
4.0	2.0	0.667	0.030	0.750	0.6498	0.0194	0.7147	0.6814	0.0195	0.7013
20.0	10.0	0.667	0.007	0.677	0.6671	0.0065	0.6765	0.6670	0.0063	0.6753

Métodos Computacionais: Métodos MCMC

Os métodos numéricos, SIR e de rejeição podem não ser boas alternativas para aproximar a distribuição *a posteriori* em modelos mais complexos. Nestes casos, os métodos MCMC são boas alternativas.

- Os métodos MCMC foram utilizados em Estatística Bayesiana pela primeira vez no artigo de Gelfand & Smith (1990).
- Produziram uma grande mudança na prática da Estadística Bayesiana fornecem uma ótima solução para aproximar a distribuição *a posteriori* em problemas muito complexo e de alta dimensão.
- Estos métodos também fornecem uma amostra da distribuição *a posteriori* quando somente o kernel da distribuição é conhecido. Mas, somente depois que a cadeia convergir.

Métodos Computacionais: Métodos MCMC

- A meta dos métodos MCMC é construir uma cadeia de Markov, invariante no tempo, que tenha como distribuição estacionária a distribuição *a posteriori*.
- Os métodos constroem uma cadeia ergódica, ou seja , uma cadeia que é irredutível, aperiódica e cujos estados são todos recorrentes positivos.
- Estos métodos fornecem uma amostra da distribuição *a posteriori* mas apenas após a cadeia gerada ter convergido (ter atingido o estado estacionário)
- A amostra gerada após a convergência ser atingida é correlacionada.

Métodos Computacionais: Métodos MCMC

O teorema que fundamenta os métodos MCMC.

Teorema: Se uma cadeia de Markov é ergódica, isto é, é irredutível, aperiódica e todos os estados são recorrentes positivos, então $\{\pi_j, J = 1, 2, \dots\}$ onde

$$\pi_j = \lim_{n \to \infty} p_{ij}^{(n)}$$

e $p_{ij} = P(X_t = j \mid X_{t-1} = i)$, é a distribuição estacionária da cadeia.

Métodos Computacionais: Métodos MCMC

- Um processos estocástico é uma CM se

  $$P(X_t = j \mid X_{t-1} = i, X_{t-2} = x_{i_{t-1}}, \ldots, X_1 = x_1) = P(X_t = x_j \mid X_{t-1} = i) = p_{ij}.$$

- CM é irredutível se todos os estados da cadeia se comunicaom entre si.
- Dois estados i e j se comunicam entre si, se para algum $n > 0$, $p_{ij}^{(n)} > 0$.
- CM é aperiódica se não possui estados absorventes.
- CM é recorrente se cada estado da cadeia for visitado infinitas vezes.

Métodos MCMC: Amostrador de Gibbs

Admita que nossa meta seja obter uma amostra da distribuição *a posteriori*

$$\pi(\theta_1, \ldots, \theta_n \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta_1, \ldots, \theta_n)\pi(\theta_1, \ldots, \theta_n)$$

No amostrador de Gibbs ou *Gibbs Sampler* a matriz de transição da cadeia de Markov é construída a partir das distribuições condicionais completas *a posteriori*.

* Seja $\theta = (\theta_1, \ldots, \theta_n)$ e $\theta_{(-i)}$ o vetor θ sem o componente i.
* a distribuição condicional completa *a posteriori* de θ_i é

$$\begin{aligned} f(\theta_i \mid \theta_{(-i)}, x) &= \frac{f(x \mid \theta_i, \theta_{(-i)})\pi(\theta_i, \theta_{(-i)})}{f(x, \theta_{(-i)})} \\ &\propto f(x \mid \theta_i, \theta_{(-i)})\pi(\theta_i, \theta_{(-i)}) \end{aligned}$$

* Se $f(\theta_i \mid \theta_{(-i)}, x)$ tem forma fechada e conhecida, podemos obter uma amostra da distribuição *a posteriori* via amostrador de Gibbs como segue:

Métodos MCMC: Amostrador de Gibbs

O algoritmo

P1 Inicie a cadeia atribuindo um valor inicial válido $\theta^0 \leftarrow (\theta_1^0, \ldots, \theta_n^0)$ a cada parâmetro.

P2 Para $i = 1, \ldots, T$ gere

$$\begin{aligned}
\theta_1^i &\leftarrow f(\theta_1 \mid \theta_2^{i-1}, \ldots, \theta_n^{i-1}, x) \\
\theta_2^i &\leftarrow f(\theta_2 \mid \theta_1^i, \theta_3^{i-1} \ldots, \theta_n^{i-1}, x) \\
&\vdots \qquad\qquad \vdots \\
\theta_n^i &\leftarrow f(\theta_n \mid \theta_1^i, \theta_2^i \ldots, \theta_{n-1}^i, x)
\end{aligned}$$

P3 Avalie a convergência e a auto-correlação da cadeia para definir o período de *burn-in* ou aquecimento B da cadeia e o *lag* l.

P4 θ^B, θ^{B+l},θ^{B+2l}, ... é a amostra de $\pi(\theta \mid x)$.

Amostrador de Gibbs: Ele funciona?

Seja a distribuição conjunta de (θ_1, θ_2) dada por

θ_2	θ_1		D.Marg.θ_2
	0	1	
0	0,1	0,4	0,5
1	0,3	0,2	0,5
D.Marg. θ_1	0,4	0,6	

Verifique se a partit das distribuições condicionais completas conseguimos obter a distribuição marginal de θ_1.

Cálculo das Distribuições condicionais completas: O cálculo da distribuição condicional de completa de θ_1 dado θ_2

$$P(\theta_1 = 0 \mid \theta_2 = 0) = 0,1/0,5 = 1/5 \text{ e } P(\theta_1 = 1 \mid \theta_2 = 0) = 0,4/0,5 = 4/5$$

$$P(\theta_1 = 0 \mid \theta_2 = 1) = 0,3/0,5 = 3/5 \text{ e } P(\theta_1 = 1 \mid \theta_2 = 1) = 0,2/0,5 = 2/5$$

Amostrador de Gibbs: Ele funciona?

- temos que tal distribuição condicional completa é

$$\pi(\theta_1 \mid \theta_2) = \begin{bmatrix} 0,2 & 0,8 \\ 0,6 & 0,4 \end{bmatrix} \tag{4}$$

- a primeira linha da matrix em (4) corresponde a $\pi(\theta_1 \mid \theta_2 = 0)$ e a segunda linha corresponde a $\pi(\theta_1 \mid \theta_2 = 1)$
- Analogamente, construímos a distribuição condicional de completa de θ_2 dado θ_1:

$$\pi(\theta_2 \mid \theta_1) = \begin{bmatrix} 1/4 & 3/4 \\ 4/6 & 2/6 \end{bmatrix} \tag{5}$$

- a primeira linha da matrix em (5) corresponde a $\pi(\theta_2 \mid \theta_1 = 0)$ e a segunda linha corresponde a $\pi(\theta_2 \mid \theta_1 = 1)$
- Para gerar da distribuição marginal de θ_1 usando o algoritmo de Gibbs devemos calcular as probabilidades de sairmos de cada estado 0 ou 1 e chegarmos a cada estado 0 ou 1 em T passos. Isto é dado pela matriz de transição a T passos $A^T_{\theta_1|\theta_1}$.

Amostrador de Gibbs: Ele funciona?

- matriz de transição a um passo de $\theta_1^{(0)} \rightarrow \theta_1^{(1)}$:

$$
\begin{aligned}
P(\theta_1^{(1)} = 1 \mid \theta_1^{(0)} = 0) &= P(\theta_1^{(1)} = 1 \mid \theta_2^{(0)} = 0) P(\theta_2^{(0)} = 0 \mid \theta_1^{(0)} = 0) \\
&+ P(\theta_1^{(1)} = 1 \mid \theta_2^{(0)} = 1) P(\theta_2^{(0)} = 1 \mid \theta_1^{(0)} = 0) \\
&= 0,8 * 1/4 + 0,4 * 3/4 = 0,5
\end{aligned}
$$

$$
\begin{aligned}
P(\theta_1^{(1)} = 1 \mid \theta_1^{(0)} = 1) &= P(\theta_1^{(1)} = 1 \mid \theta_2^{(0)} = 0) P(\theta_2^{(0)} = 0 \mid \theta_1^{(0)} = 1) \\
&+ P(\theta_1^{(1)} = 1 \mid \theta_2^{(0)} = 1) P(\theta_2^{(0)} = 1 \mid \theta_1^{(0)} = 1) \\
&= 0,8 * 4/6 + 0,4 * 2/6 = 0,6668
\end{aligned}
$$

$$
\begin{aligned}
P(\theta_1^{(1)} = 0 \mid \theta_1^{(0)} = 0) &= P(\theta_1^{(1)} = 0 \mid \theta_2^{(0)} = 0) P(\theta_2^{(0)} = 0 \mid \theta_1^{(0)} = 0) \\
&+ P(\theta_1^{(1)} = 0 \mid \theta_2^{(0)} = 1) P(\theta_2^{(0)} = 1 \mid \theta_1^{(0)} = 0) \\
&= 0,2 * 1/4 + 0,6 * 3/4 = 0,5
\end{aligned}
$$

$$
\begin{aligned}
P(\theta_1^{(1)} = 0 \mid \theta_1^{(0)} = 1) &= P(\theta_1^{(1)} = 0 \mid \theta_2^{(0)} = 0) P(\theta_2^{(0)} = 0 \mid \theta_1^{(0)} = 1) \\
&+ P(\theta_1^{(1)} = 0 \mid \theta_2^{(0)} = 1) P(\theta_2^{(0)} = 1 \mid \theta_1^{(0)} = 1) \\
&= 0,2 * 4/6 + 0,6 * 2/6 = 0,3314
\end{aligned}
$$

Amostrador de Gibbs: Ele funciona?

- A matriz de transição a 1 passo é

$$A^1_{\theta_1|\theta_1} = \begin{bmatrix} 0,5 & 0,5 \\ 0,333 & 0,667 \end{bmatrix}$$

- A matrix de transição a 2 passos é

$$(A^1_{\theta_1|\theta_1})^2 = \begin{bmatrix} 0,41667 & 0,58333 \\ 0,38889 & 0,61111 \end{bmatrix}$$

- A matrix de transição a 10 passos é

$$(A^1_{\theta_1|\theta_1})^{10} = \begin{bmatrix} 0,4 & 0,6 \\ 0,4 & 0,6 \end{bmatrix}.$$

- Ou seja, depois de 10 passos já atingimos a distribuição estacionária que é a distribuição marginal de θ_1.

Amostrador de Gibbs: Família Normal (μ, σ^2)

Se $X_i \mid \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$ e , *a priori* temos que $\mu \mid \sigma^2 \sim N(m, V\sigma^2)$ e $\sigma^2 \sim GI(a/2, d/2)$, as dist. condicionais completas para μ e σ^2 são, respectivamente,

$$\begin{aligned}
\pi(\mu \mid \sigma^2, x) &\propto f(x \mid \mu, \sigma^2)\pi(\mu \mid \sigma^2) \\
&\propto \exp\left\{-\frac{nV+1}{2V\sigma^2}\left[\mu^2 - 2\mu\left(\frac{nV\bar{x}+m}{nV+1}\right)\right]\right\} \qquad (6)
\end{aligned}$$

$$\begin{aligned}
\pi(\sigma^2 \mid \mu, x) &\propto f(x \mid \mu, \sigma^2)\pi(\mu \mid \sigma^2)\pi(\sigma^2) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{(d+n+3)}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(a + \sum_i (x_i - \mu)^2 + \frac{(\mu - m)^2}{V}\right)\right\}.
\end{aligned}$$

Ambas tem formas fechadas conhecidas

$$\mu \mid \sigma^2, x \sim N(m^*, V^*\sigma^2) \quad \text{e} \quad \sigma^2 \mid \mu, x \sim GI(a^*/2, d^*/2)$$

.

Amostrador de Gibbs: Família Normal (μ, σ^2)

Note que a variância da distribuição condicional completa de μ depende de σ^2 e que o parâmetro $a^* = a + \sum_i (x_i - \mu)^2 + \frac{(\mu - m)^2}{V}$ da distribuição condicional completa de σ^2 depende de μ.

Algoritmo:

P1 Inicialize: $(\mu^{(0)}, \sigma^{2(0)})$

P2 Para $i = 1, \ldots, T$ gere

$$\mu^{(i)} \quad | \quad (\sigma^2)^{(i-1)}, \boldsymbol{x} \sim N(m^*, V^* \sigma^{2(i-1)})$$

$$(\sigma^2)^{(i)} \quad | \quad \mu^{(i)}, \boldsymbol{x} \sim GI\left(\frac{1}{2}\left(a + \sum_i (x_i - \mu^{(i)})^2 + \frac{(\mu^{(i)} - m)^2}{V}\right), \frac{d + n + 1}{2}\right)$$

P3 Repita até obter a amostra de tamanho desejado.

Métodos MCMC: Metropolis-Hastings

Suponha que nossa distribuição alvo seja $\psi(\theta)$.

- A meta é construir uma matriz de transição tal que $\psi(\theta)$ seja a distribuição invariante relacionada a esta cadeia, ou seja, tal matriz tem que ser ergodica.
- O método começa gerando um candidato θ^* de uma densidade candidata $H(\theta^* \mid \theta)$ conhecida.
- Isto faz com que tenhamos amostras dependentes $\psi(\theta)$.
- parece com o método da rejeição só que agora o candidato gerado dependerá do estado inicial da Cadeia θ.
- Para ser ergodica a matrix de transição tem que ser reversível, ou seja, tem que valer

$$\psi(\theta)H(\theta^* \mid \theta) = \psi(\theta^*)H(\theta \mid \theta^*)$$

- Se isto ocorre, $H(\theta^* \mid \theta)$ será a probabilidade de transição.

Métodos MCMC: Metropolis-Hastings

- Se ocorre
$$\psi(\theta)H(\theta^* \mid \theta) > \psi(\theta^*)H(\theta \mid \theta^*)$$
então o processo move mais rapidamente de θ para θ^* do que de θ^* para θ.
- Neste caso,
  - Para reduzir o número de movimentos de θ para θ^* e garantir reversibilidade, acrescenta-se uma probabilidade $K(\theta^* \mid \theta)$ tal que
$$\psi(\theta)H(\theta^* \mid \theta)K(\theta^* \mid \theta) = \psi(\theta^*)H(\theta \mid \theta^*).$$
  - $K(\theta^* \mid \theta)$ é chamada de probabilidade de movimento e é dada por
$$K(\theta^* \mid \theta) = \frac{\psi(\theta^*)H(\theta \mid \theta^*)}{\psi(\theta)H(\theta^* \mid \theta)}.$$
  - Como não queremos reduzir o número de movimentos de θ^* para θ então tomamos $K(\theta \mid \theta^*) = 1$
- Teríamos algo similar se a desigualdade fosse ao contrário.

Métodos MCMC: Metropolis-Hastings

- A probabilidade de transição de θ para θ^* de tal forma que $\psi(\theta)$ seja a distribuição estacionária é

$$K(\theta^* \mid \theta) = \min\left\{\frac{\psi(\theta^*)H(\theta \mid \theta^*)}{\psi(\theta)H(\theta^* \mid \theta)}, 1\right\}.$$

É desejável que distribuição de referência (*proposal distribution*) $H(\theta \mid \theta^*)$ tenhas as seguintes características

  - é perfeitamente conhecida
  - se pode gerar fácilmente de H

Métodos MCMC: Metropolis-Hastings

Casos particulares:

- Algoritmo de Metropolis
  - $H(\theta \mid \theta^*)$ não tem que, necessariamente, ser simétrica mas simetria pode tornar mais fácil os cálculos pois $H(\theta^* \mid \theta) = H(\theta \mid \theta^*)$.
  - A probabilidade de transição de θ para θ^* é

$$K(\theta^* \mid \theta) = \min\left\{\frac{\psi(\theta^*)}{\psi(\theta)}, 1\right\}.$$

  - Neste caso, cadeia gerada seria um passeio aleatório.
- Amostrador independente
  - $H(\theta^* \mid \theta) = g(\theta^*)$.
  - A probabilidade de transição de θ para θ^* de tal forma que $\pi(\theta)$ seja a distribuição estacionária é

$$K(\theta^* \mid \theta) = \min\left\{\frac{\psi(\theta^*)g(\theta)}{\psi(\theta)g(\theta^*)}, 1\right\}.$$

  - Continuaríamos gera uma cadeia dependente pois a probabilidade de aceitação depende de θ, o estado atual da cadeia.

Métodos MCMC: Metropolis-Hastings

Se a distribuição estacionária é a diatribuição *a posteriori*, teríamos que $\psi(\theta) = \pi(\theta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta)\pi(\theta)$ e o algorítmo seria:

O algoritmo

*P*1 Inicie a cadeia θ^0

*P*2 Para cada $i = 1, \ldots, T$ gere um candidato $\theta^* \leftarrow H(\theta^i \mid \theta^{i-1})$

*P*3 Calcule os saltos

$$R = \frac{f(x \mid \theta^*)\pi(\theta^*)H(\theta^{i-1} \mid \theta^*)}{f(x \mid \theta^{i-1})\pi(\theta^{i-1})H(\theta^* \mid \theta^{i-1})}$$

*P*4 Calcule a probabilidade da cadeia mover-se do estado $i-1$ para o estado i

$$K = \min\{1, R\}$$

Métodos MCMC: Metropolis-Hastings

O algoritmo (cont.)

P5 Gere $u \sim U(0,1)$.

- Se $u \leq k \rightarrow \theta^i = \theta^*$ (o candidato é aceito)
- Se $u > k \rightarrow \theta^i = \theta^{i-1}$ (rejeitamos o candidato)

P6 Avalie a convergência e auto-correlação da cadeia para definir o período de *burn-in* B e o *lag* l.

P7 θ^B, θ^{B+l},θ^{B+2l}, ... é a amostra de $\pi(\theta \mid x)$.

Observações: É desejável que os saltos R sejam de fácil computo e não levem à rejeição muito frequentemente. Ainda mais, devem ter tamanho razoável para que a convergência seja rápidamente atingida.

Intuição do algoritmo Metropolis-Hastings

- Por simplicidade, consideremos um passeio aleatório, isto é, $H(\theta^* \mid \theta^{(i-1)}) = H(\theta^{(i-1)} \mid \theta^*)$, e assuma que a distribuição estacionária $\psi(\theta)$ seja a distribuição *a posteriori* $\pi(\theta \mid \boldsymbol{x})$.
- Neste caso, a probabilidade de movimento ou salto da cadeia é

$$R = \frac{f(x \mid \theta^*)\pi(\theta^*)}{f(x \mid \theta^{(i-1)})\pi(\theta^{(i-1)})}$$

- Se $R > 1$ temos que θ^* tem peso maior na distribuição *a posteriori* do que o valor $\theta^{(i-1)}$ já pertencente à amostra.
- A intuição nos manda aceitar θ^* como um elemento da amostra da distribuição *a posteriori* pois tem mais probabilidade que $\theta^{(i-1)}$.
- O que faz o algoritmo MH neste caso?
  - Ele considera $K(\theta^* \mid \theta^{(i-1)}) = min\{1, R\}$. Como $R > 1$ teremos que $K(\theta^* \mid \theta^{(i-1)}) = 1$
  - Faz $\theta^{(i)} = \theta^*$ como probabilidade 1

Intuição do algoritmo Metropolis-Hastings

- Se $R < 1$ temos que θ^* tem peso menor na distribuição *a posteriori* do que o valor $\theta^{(i-1)}$ já pertencente à amostra.
- Intuição: A frequência relativa dos valores de θ em nossa amostra igual à θ^* em relação à frequência de valores na amostra igual a $\theta^{(i-1)}$ deve ser

$$R = \frac{f(x \mid \theta^*)\pi(\theta^*)}{f(x \mid \theta^{(i-1)})\pi(\theta^{(i-1)})} \in (0,1)$$

.

- Isto significa que para cada estado $\theta^{(i-1)}$ presente na amostra, devemos ter apenas uma "fração"R de estados θ^* na amostra.
- O que o algoritmo faz neste caso?
  - Faz $\theta^{(i)} = \theta^*$ como probabilidade R
  - Faz $\theta^{(i)} = \theta^{(i-1)}$ como probabilidade $1 - R$

Métodos MCMC: Metropolis-Hastings

Exemplo: Se $X_1, \ldots, X_n \mid \mu \overset{iid}{\sim} N(\mu, 1)$ e *a priori* $\pi(\mu) \propto 1$ calcule a probabilidade de aceitação R do algoritmo Metropolis-Hastings usando como distribuição candidata a distribuição $\mu^{(i)} \mid \mu^{(i-1)} \sim N(\mu^{(i-1)}, V)$.

Solução:

$$
\begin{aligned}
R^{(i)} &= \frac{f(x \mid \mu^*)\pi(\mu^*)H(\mu^{(i-1)} \mid \mu^*)}{f(x \mid \mu^{(i-1)})\pi(\mu^{(i-1)})H(\mu^* \mid \mu^{(i-1)})} \\
&= \frac{\left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{\frac{-\sum(x_j-\mu^*)^2}{2}\right\}\left(\frac{1}{2V}\right)^{1/2} \exp\left\{\frac{-(\mu^{(i-1)}-\mu^*)^2}{2V}\right\}}{\left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{\frac{-\sum(x_j-\mu^{(i-1)})^2}{2}\right\}\left(\frac{1}{2V}\right)^{1/2} \exp\left\{\frac{-(\mu^*-\mu^{(i-1)})^2}{2V}\right\}} \\
&= \exp\left\{\frac{-1}{2}[(\mu^*)^2 - (\mu^{(i-1)})^2] - 2\sum_{j=1}^{n} x_j(\mu^* - \mu^{(i-1)})\right\} \qquad (7)
\end{aligned}
$$

Métodos MCMC: Metropolis-Hastings

Sobre esta estratégia

- Se V é pequeno tenderemos a gerar sempre valores próximos do valor gerado no passo anterior.
- Isto pode tornar lendo o processo de amostragem pois demoraremos a varrer todo o espaço paramétrico.
- no entanto, pode melhorar a taxa de aceitação do algoritmo.
- Dica 1: Gere as amostras considerando vários parametros de tunagem V. Escolha aquele que fornecer cadeias mais estáveis e taxa de aceitação razoável.
  - taxa de aceitação sugerida: entre 20% e 50%.
- Dica 2:Podemos ter instabilidades computacionais no cálculo da razao R em (7). Uma maneira de minimizar o problema é calculando $\ln R$ e depois exponencie este logaritmo.

Métodos MCMC:Construção da amostra da distribuição *a posteriori*

Métodos MCMC: construção da amostra

Suponha que quieramos uma amostra de tamanho n da distribuição *a posteriori*, independente.

- Estratégia 1: construa n cadeias paralelas, geradas a partir de n pontos iniciais distintos e independentes.
  - verifique a convergência de cada uma delas.
  - após convergirem no passo M_i , $i = 1, \ldots, n$, tome os elementos gerados nos passos $M_1 + 1, \ldots, M_n + 1$ como sendo os componente da amostra *a posteriori*.
  - este processo garante que a amostra gerada é indepenente mas é um processo computationalmente caro.
  - o período ou iterações $\{0, 1, \ldots, M_i\}$ é denominado período de *burn-in* ou aquecimento da cadeia i.
  - Observações relacionadas a estas iterações devem ser descartadas pois não são amostras da distribuição estacionária.

Métodos MCMC: construção da amostra

- Estratégia 2: Construa uma única cadeia.
  - Após convergir no passo M, descarte o período de *burn-in*.
  - A cadeia resultante é auto-correlacionada.
  - Estude a autocorrelação da cadeia e defina a ordem L dos saltos a partir do qual a cadeia resultante tenha auto-correlação não-significativa.
  - Forme a amostra considerando os valores gerados nos passos $M+1, M+L+1,\ M+2L+1, \ldots$.
  - a amostra resultande é aproximadamente independente.

Caso a auto-correlação seja muito alta pode ser necessário uma amostra muito grande para percorrer todo espaço paramétrico.

Métodos MCMC: Diagnóstico de convergência e auto-correlação

- Somente teremos uma amostra da distribuição *a posteriori* a partir do momento que a convergência da cadeia foi atingida, ou seja, quando a cadeia se "esqueceu"dos pontos iniciais.
- O monitoramento da convergência pode ser feito de forma gráfica considerando-se
  - + um gráfico de iteração versus valor gerado
  - + um gráfico da média ergódica

$$ME_i = \frac{\sum_{i=1}^{i} \theta_i}{i}$$

  - + Considerando-se muitos pontos iniciais e verificando quando as cadeias se tornam indistinguíveis.
- a correlação entre os elementos gerados pode ser avaliada através do gráfico de autocorrelação da cadeia
- Há formas mais formais para tal monitoramento que são baseadas em testes de estacionariedade da cadeia. Alguns destes testes estão no pacote CODA do R.

Métodos MCMC: Diagnóstico de convergência e auto-correlação

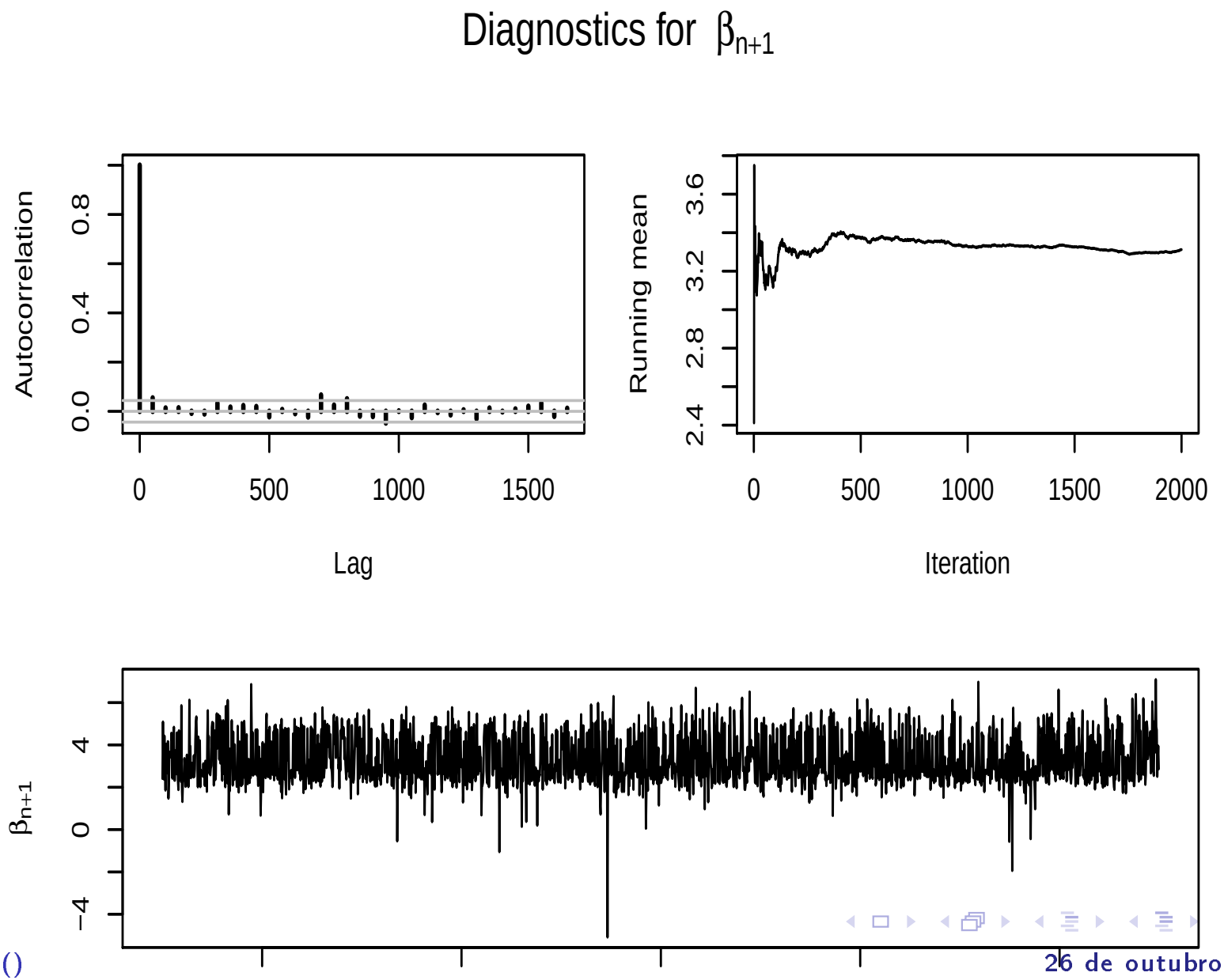

Figure 1: Time-series plots for the NAEP iterations each for the

Figure 2: Running mean plot

iterations each for the NAEP

Figure 5: Running mean plots

iterations each for the NAEP

Métodos MCMC: Diagnóstico de convergência e auto-correlação

Como resolver o problema de auto-correlação?

- Uma possibilidade é reparametrizar o modelo
- Se os parâmetros são fortemente correlacionados, uma outra estratégia para melhorar a eficiência do algoritmo é a bocagem.
  - Suponha que voce queira obter uma amostra da distribuição *a posteriori* de $(\alpha_1, \alpha_2, \theta)$. Suponha que α_1 e α_2 sejam autamente correlacionados.
  - Calcule as seguintes distribuições condicionais completas

$$\begin{aligned} \pi(\alpha_1, \alpha_2 \mid \theta, \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \alpha_1, \alpha_2, \theta)\pi(\alpha_1, \alpha_2) \\ \pi(\theta \mid \alpha_1, \alpha_2, \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \alpha_1, \alpha_2, \theta)\pi(\theta) \end{aligned}$$

  - Se tais distribuições condicionais completas tem formas fechadas conhecidas, use o amostrador de Gibbs para gerar amostras da distribuição *a posteriori*.
  - Esta estratégia aumenta a eficiência do algoritmo gerando a amostras de cada parâmetro menos auto-correlacionadas e aumentando a velocidade de convergência.