ESTATÍSTICA COMPUTACIONAL E SIMULAÇÃO

Folha prática 3: Métodos MCMC

Ano letivo 2021/22

1. Considere a distribuição de Rayleigh com função densidade de probabilidade dada por

$$f(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, x \geq 0, \sigma > 0.$$

Utilize o algoritmo de Metropolis-Hastings para gerar valores desta distribuição, considerando para função proponente $q(y|x_t)$ a distribuição do qui-quadrado com x_t graus de liberdade. Considerando $\sigma = 4$ e 20000 réplicas, verifique se cadeia é eficiente.

2. Pretende-se gerar valores de uma distribuição (alvo) t-Student com n graus de liberdade, usando o algoritmo de Metropolis-Hastings com um passeio aleatório, considerando a distribuição normal $N(0, \sigma^2)$ para distribuição do erro aleatório.

   (a) Implemente o algoritmo de Metropolis referido considerando uma cadeia de dimensão $m = 10000$, $\sigma = 0.05$ e o número de graus de liberdade da t-Student igual a 4. Analise a eficiência.

   (b) Compare os resultados obtidos pelo algoritmo em termos do número de rejeições fazendo variar σ^2, considerando $\sigma^2 = 0.05; 0.5; 2; 16$.

   (c) Compare os quantis da distribuição teórica - distribuição alvo - t-Student com 4 g.l.,com os quantis da distribuição dos valores gerados, considerando cada um dos valores de σ^2 da alínea anterior e considerando um período de aquecimento igual a 1000. Tire conclusões.

3. Considere uma amostra aleatória de $(Z_1, \ldots, Z_n)$ da seguinte mistura de Normais:

$$pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2).$$

Supondo que não existe informação sobre p, o objetivo é estimar o valor de p usando uma distribuição $Be(1,1)$ para função proponente g e suponha que o ponto y gerado no algoritmo de Metropolis-Hastings é aceite com probabilidade $\alpha = min(1, \frac{f(Y)g(X_t)}{f(X_t)g(Y)})$ - corresponde ao amostrador independente.
Determine uma estimativa para p admitindo que os valores gerados provêm da seguinte mistura: $0.2N(0,1) + 0.8N(5,1)$.

4. Segundo um modelo genético, animais de uma determinada espécie estão distribuídos em 4 categorias, de acordo com as probabilidades

$$p_1 = \frac{2+\theta}{4}, p_2 = \frac{1-\theta}{4}, p_3 = \frac{1-\theta}{4}, p_4 = \frac{\theta}{4},$$

onde $0 \leq \theta \leq 1$ é um parâmetro desconhecido, relativamente ao qual se pretende fazer inferência.
Admita-se que
- se adopta para θ umas distribuição *a priori* beta de parâmetros (a, b).
-para uma amostra de dimensão N se observaram y_i animais na i-ésima categoria ($i = 1, \ldots, 4, \sum y_i = N$)
Pretende-se simular desta distribuição usando o algoritmo Metropolis-Hastings. Se usarmos, por exemplo, $q(\theta_i, \theta_j) = 1$ isso significa que simulamos de uniformes e aceitamos o valor simulado com probabilidade

$$\alpha(\theta_i, \theta_j) = min\{1, \frac{\pi(\theta_j|y)}{\pi(\theta_i|y)}\}.$$

Aplique este algoritmo para simular da distribuição

$$\pi(\theta|y) \propto (2+\theta)^{y_1}(1-\theta)^{y_2-y_3+b-1}\theta^{y_4+a-1}, 0 \leq \theta \leq 1,$$

considerando $a = b = 1$.

5. Num modelo binormal de vector valor médio (μ_1, μ_2), variâncias (σ_1^2, σ_2^2) e correlação ρ, considere as distribuições condicionais completas dadas por:

$$X_1|X_2 \sim N(\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2), (1-\rho^2)\sigma_1^2)$$

$$X_2|X_1 \sim N(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1 - \mu_1), (1-\rho^2)\sigma_2^2).$$

Utilize o amostrador Gibbs para gerar valores desta binormal, com valores gerados considerando para os parâmetros os valores $(\mu_1 = 0, \mu_2 = 2)$, $(\sigma_1^2 = 1, \sigma_2^2 = 0.25)$ e $\rho = -0.75$ e estime os parâmetros através das características amostrais correspondentes.

6. Suponhamos que $X = (X_1, \ldots, X_n)$ é uma sequência de variáveis aleatórias da forma $X_i = R_i Y_i$ onde Y_i são i.i.d. $Poi(\lambda)$, R_i são i.i.d. $Ber(p)$ e Y_i e R_i são independentes.
Dada uma amostra observada $x = (x_1, \ldots, x_n)$ de X, o objectivo é estimar os parâmetros (λ, p).
Do ponto de vista da metodologia Bayesiana considera-se o seguinte modelo hierárquico:
- $p \sim U(0,1)$
- $\lambda|p \sim Ga(a,b)$
- $(r_i|p,\lambda) \sim Ber(p)$ independentes
- $(x_i|r,p,\lambda) \sim Poi(\lambda r_i)$ independentes,
com $r = (r_1, \ldots, r_n)$ e a, b constantes conhecidas. Segue-se que a distribuição conjunta de (x, r, λ, p) é

$$f(x,r,\lambda,p) = \frac{b^a \lambda^{a-1} exp(-b\lambda)}{\Gamma(a)} \prod_{i=1}^{n} \frac{exp(-\lambda r_i)(\lambda r_i)^{x_i}}{x_i!} p^{r_i}(1-p)^{1-r_i}.$$

Podem ser feitas simulações sobre p e λ simulando da distribuição *a posteriori* - distribuição de (r, λ, p) condicional aos dados observados x - dada por

$$f(r,\lambda,p|x) = \frac{f(x,r,\lambda,p)}{\int f(x,r,\lambda,p)d(r,\lambda,p)},$$

usando, por exemplo, o algoritmo Gibbs.

(a) As condicionais completas são da forma
.$(\lambda|p,r,x) \sim Ga(a + \sum_i x_i, b + \sum_i r_i)$
.$(p|\lambda,r,x) \sim Be(1 + \sum_i r_i, n + 1 - \sum_i r_i)$
.$(r_i|\lambda,p,x) \sim Ber(\frac{pexp(-\lambda)}{pexp(-\lambda)+(1-p)I_{\{x_i=0\}}})$.

Prove as das duas primeiras expressões.

(b) Gere uma amostra de dimensão $n = 100$ do modelo ZIP usando parâmetros $p = 0.3$ e $\lambda = 2$.

(c) Implemente o algoritmo Gibbs para simular de $f(r,\lambda,p|x)$ para os dados da alínea anterior, determinar estimativas para os parâmetros $p = 0.3$ e $\lambda = 2$ e para caracterizar as distribuições de $p|x$ e de $\lambda|x$.

7. McCormick e Mathew(1983) examinaram questões de estimação relacionadas com o modelo

$$X_t = \gamma + \rho X_{t-1} + Y_t, t \geq 1$$

onde $\gamma \geq 0$, $0 \leq \rho < 1$ são parâmetros desconhecidos e Y_t são variáveis aleatórias i.i.d. A análise bayesiana deste modelo quando os Y_t são exponenciais com valor médio α^{-1} foi considerada por Pereira e Amaral Turkman(1983). Para uma distribuição *a priori* não informativa, a distribuição *a posteriori* baseada numa amostra de dimensão n é

$$h(\theta|x) \propto \alpha^{n-1} e^{-\alpha[S_0 - \rho S_1 - (n-1)\gamma]} I_{\Theta_N}(\theta)$$

onde $\theta = (\alpha, \gamma, \rho)$, $\Theta_n = \{\theta : \alpha \geq 0, 0 \leq \rho < 1, \gamma \geq 0, x_t - \rho x_{t-1} - \gamma \geq 0, \forall t = 1, \ldots, n\}$, $S_0 = \sum_{t=2}^{n} x_t$ e $S_1 = \sum_{t=2}^{n} x_{t-1}$. Mostre que:

(a) As distribuições condicionais completas são dadas por

i. $h(\alpha|x,\gamma,\rho) \sim Ga(n, S_0 - \rho S_1 - (n-1)\gamma)$.

ii. $h(\gamma|x,\alpha,\rho) \sim Exp_{esq}((n-1)\alpha, \gamma*)$, onde $\gamma* = min(x_t - \rho x_{t-1})$.

iii. $h(\rho|x,\alpha,\gamma) \sim Exp_{esq}(\alpha S_1, \rho*)$, onde $\rho* = min(1, \frac{x_t-\gamma}{x_{t-1}})$.

OBS: $f(x) \sim Exp_{esq}(\beta,\delta)$ significa $f(x) = \frac{\beta e^{-\beta(\delta-x)}}{1-e^{-\beta\delta}}$, para $0 < x < \delta$.

(b) Os seguintes dados representam valores dos teores de oxigénio dissolvido medidos na ponte de Angeja de Junho a Novembro de 1991:

4.0	4.1	3.9	4.4	3.2	4.0	3.7	4.2	4.5	4.3	3.6	1.9
3.3	1.9	2.9	2.7	2.4	2.9	3.8	3.5	2.7	3.9	2.8	3.3
2.9	3.8	4.4	5.1	5.2	7.2	6.2	4.8	4.0	2.7	4.4	3.4
4.2	4.8	5.3	4.5	4.1	4.0	2.9	0.8	5.2	7.3	5.1	5.3
7.1	8.1	7.8	6.9	7.5	6.0	5.0	5.3	4.8	4.3	5.8	4.6
4.5	4.1	4.6	6.4	6.3	6.2	6.2	6.8	7.5	7.4	7.0	6.7
7.5	6.1	5.7	5.4	5.3	4.0	3.7	2.5	0.8	1.3	3.3	4.1
5.7	4.3	3.5	3.8	2.0	3.8	4.1	1.8	3.0	4.7	6.2	6.0
5.3	4.4	3.4	4.7	4.5	3.7	4.3	1.6	2.9	3.6	3.7	3.9
4.6	5.0	5.3	4.7	6.5	5.7	5.8	8.0	7.4	6.1	7.6	

Usando esta amostra, implemente o algoritmo Gibbs, usando o pacote CODA para analisar a convergencia, com o objetivo de estimar os parâmetros e caracterizar aproximadamente as distribuições de $\alpha|x$, $\gamma|x$ e de $\rho|x$.

8. Considere o seguinte modelo de regressão (*one-way random effects model with k factors*)

$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j}$, $\alpha_i \sim N(0,\sigma_\alpha^2)$, $\epsilon_{i,j} \sim N(0,\sigma^2)$, $j = 1,\ldots,n_i; i = 1,\ldots,k$.

O modelo pode ser também escrito na forma hierárquica, i.e.,

$Y_{i,j}|\alpha_i \sim N(\mu+\alpha_i,\sigma^2)$ com $\alpha_i \sim N(0,\sigma_\alpha^2)$.

A densidade conjunta de

$$\mathbf{y}|\mu,\alpha_1,\ldots,\alpha_k,\sigma_\alpha^2,\sigma^2,$$

com $\mathbf{y} = (y_1,\ldots,y_n)$, é dada por:

$(\frac{1}{\sigma_\alpha^2})^{k/2} exp[\frac{1}{2\sigma_\alpha^2}\sum_{i=1}^k \alpha_i^2](\frac{1}{\sigma^2})^{Nk/2} exp[\frac{1}{2\sigma_\alpha^2}\sum_{i=1}^k\sum_{j=1}^{n_i}(y_{ij}-\mu-\alpha_i)^2]$, com $N = \sum_{i=1}^k n_i$

e as distribuições condicionais completas são dadas por:

$\alpha_i|\mathbf{y},\mu,\sigma_\alpha^2,\sigma^2 \sim N(\frac{n_i\sigma_\alpha^2}{n_i\sigma_\alpha^2+\sigma^2}(\bar{y}_i-\mu), \frac{\sigma_\alpha^2\sigma^2}{n_i\sigma_\alpha^2+\sigma^2}), i = 1,\ldots,k$

$\mu|\mathbf{y},\alpha_1,\ldots,\alpha_k,\sigma_\alpha^2,\sigma^2 \sim N(\bar{y}-\bar{\alpha}, \frac{\sigma^2}{Nk})$

$\sigma_\alpha^2|\mathbf{y},\alpha_1,\ldots,\alpha_k,\mu,\sigma^2 \sim GI(\frac{k}{2}-1, \frac{1}{2}\sum_{i=1}^k \alpha_i^2)$

$\sigma^2|\mathbf{y},\alpha_1,\ldots,\alpha_k,\mu,\sigma_\alpha^2 \sim GI(\frac{Nk}{2}-1, \frac{1}{2}\sum_{i,j}(y_{ij}-\mu-\alpha_i)^2)$
onde IG representa a distribuição gama invertida.

Considere os seguintes dados para avaliar a precisão da estimativa da composição química das folhas de nabo, onde as folhas representam um efeito aleatório

$factores$				
1	3.28	3.09	3.03	3.03
2	3.52	3.48	3.38	3.38
3	2.88	2.80	2.81	2.76
4	3.34	3.38	3.24	3.26

Implemente o algoritmo de Gibbs para obter estimativas para os parâmetros $\alpha_i, i = 1, \ldots, 4$, μ, σ_α^2 e σ^2 (depois de analisada a convergência) algumas das suas características e apresente graficamente as distribuições condicionais completas correspondentes.