

Mistura de Normais com Variância Contaminada

Caio Gabriel Barreto Balieiro

Taiguara Melo Tupinambás

Walmir dos Reis Miranda Filho

Programa de Pós-Graduação em Estatística

Departamento de Estatística - UFMG

02 de dezembro de 2019

- 1 Introdução
- 2 Método de Quadratura de Riemann
- 3 Reamostragem por importância sequencial(SIR)
- 4 Integração via Monte Carlo em cadeias de Markov
- 5 Considerações finais

- O presente trabalho tem como objetivo obter, dada uma densidade a posteriori conjunta dos parâmetros de um modelo probabilístico para uma amostra previamente observada, as densidades a posteriori marginais de cada parâmetro.
- Obter as estatísticas de média; variância; assimetria e curtose associadas as densidades a posteriori marginais de cada parâmetro, a partir da implementação de três métodos numéricos.
- Sendo eles: (i) integração via quadratura de Riemann; (ii) reamostragem por importância sequencial (em inglês, *Sequential Importance Resampling*, ou SIR); e (iii) integração via Monte Carlo em cadeias de Markov (em inglês, *Markov Chain Monte Carlo*, ou MCMC) com inovações dadas pelo algoritmo de Metropolis-Hastings (MH).

- Sejam X_1, \dots, X_n amostras aleatórias independentes, condicionalmente a um vetor de parâmetros $\theta = (\mu, \sigma^2, \nu)$, e identicamente distribuídas com função densidade dada por

$$f(x|\mu, \sigma^2, \nu) = \nu \phi(x|\mu, 100\sigma^2) + (1 - \nu) \phi(x|\mu, \sigma^2), \quad x \in \mathbb{R}, \quad (1)$$

onde $\phi(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1} \exp[-(x - \mu)^2/(2\sigma^2)]$ denota a função densidade da distribuição normal com média μ e variância σ^2 avaliada no ponto x . Para o suporte de cada parâmetro, tem-se que $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$ e $\nu \in (0, 1)$.

- Para os parâmetros μ , σ^2 e ν , será pressuposto que cada um segue uma distribuição *a priori*: $\mu|\sigma^2 \sim N(m, V\sigma^2)$, onde $N(\cdot)$ denota a distribuição normal com média $m \in \mathbb{R}$ e variância $V\sigma^2$, $V > 0$; $\sigma^2 \sim GI(a, d)$, onde $GI(\cdot)$ denota a distribuição gama inversa com parâmetros de forma $a > 0$ e de taxa $d > 0$ (inverso da escala); e $\nu \sim U(0, 1)$, a distribuição uniforme contínua padrão.

- E então, propôs-se independência entre as prioris, portanto assumem a forma dada por

$$p(\theta) = p(\mu|\sigma^2)p(\sigma^2)p(\nu). \quad (2)$$

- Para gerar uma amostra aleatória do modelo em (1), foi utilizada uma representação hierárquica (Lachos *et al.*, 2013) tal que

$$X_i|\mu, \sigma^2, U_i = u_i \sim N(\mu, \sigma^2 u_i^{-1}), \quad U_i|\mu \sim p_d(1, 100) : P(U_i = 100) = \nu, \quad (3)$$

onde $p_d(a, b)$ denota uma função de probabilidade (discreta) que atribui massa probabilística apenas aos pontos a e b .

- Com isto, nossa posteriori será definida da seguinte forma

$$\begin{aligned}
 p(\mu, \sigma^2, \nu | \mathbf{x}) &= \frac{f(\mathbf{x} | \mu, \sigma^2, \nu) \times p(\mu, \sigma^2, \nu)}{f(\mathbf{x})} \propto \prod_{i=1}^n f(x_i) \times p(\mu | \sigma^2) \times p(\sigma^2) \times p(\nu) \\
 &\propto \prod_{i=1}^n \left[\nu \phi(x_i | \mu, 100\sigma^2) + (1 - \nu) \phi(x_i | \mu, \sigma^2) \right] \times \\
 &\times \phi(\mu | m, V\sigma^2) \times \frac{d^a}{\Gamma(a)} \left(\frac{1}{\sigma^2} \right)^{a+1} \exp \left(-\frac{d}{\sigma^2} \right) \\
 &\propto \left(\frac{1}{\sigma^2} \right)^{(n+1)/2+a+1} \exp \left\{ -\frac{[(\mu - m)^2/(2V) + d]}{\sigma^2} \right\} \times A(\mathbf{x} | \mu, \sigma^2, \nu),
 \end{aligned} \tag{4}$$

onde

$$A(\mathbf{x} | \mu, \sigma^2, \nu) = \prod_{i=1}^n \left\{ \frac{\nu}{10} \exp \left[-\frac{(x_i - \mu)^2}{200\sigma^2} \right] + (1 - \nu) \exp \left[-\frac{(x_i - \mu)^2}{2\sigma^2} \right] \right\}.$$

- Para o presente trabalho, foram considerados uma amostra de tamanho $n = 500$, da mistura finita de normais com variância contaminada parametrizada de tal forma que $\mu = 11$; $\sigma^2 = 0.64$ e $\nu = 0.2$.
- Os valores escolhidos para os hiperparâmetros são $m = 11$; $V = 1$; $a = 7$ e $d = 4$ nas distribuições *a priori*.
- Como não se tem uma expressão fechada para $p(\mu, \sigma^2, \nu | \mathbf{x})$, mas apenas de seu núcleo, para obter as densidades *a posteriori* marginais de μ , σ^2 e ν dado \mathbf{x} , bem como as estatísticas associadas a cada uma delas, é necessário aproximá-las por algum método numérico.

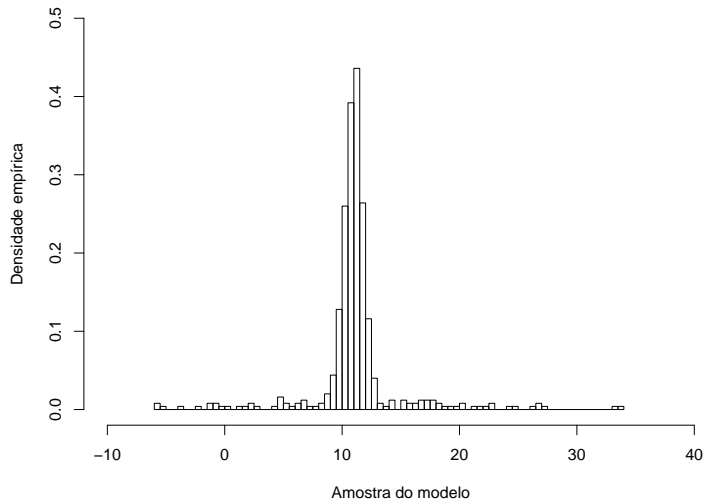
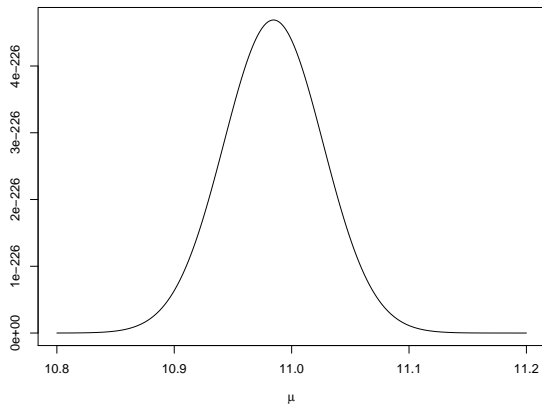
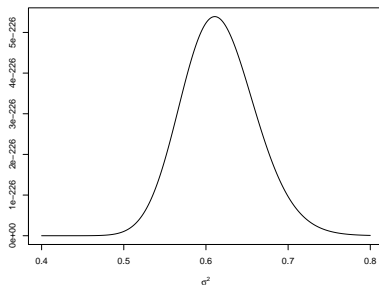


Figura 1: Histograma da amostra gerada do modelo

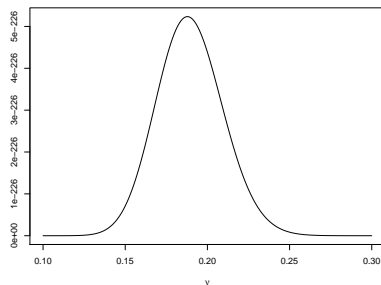


(a) $I_\mu = (10.85, 11.13)$, dados $\sigma^2 = 0.64, \nu = 0.2$

Figura 2: Intervalos de massa probabilística para cada parâmetro (variável aleatória) do núcleo de $p(\mu, \sigma^2, \nu | \mathbf{x})$



(a) $I_{\sigma^2} = (0.48, 0.78)$, dados $\mu = 11, \nu = 0.2$



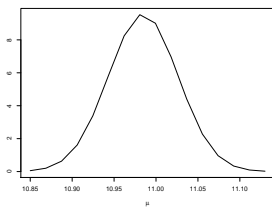
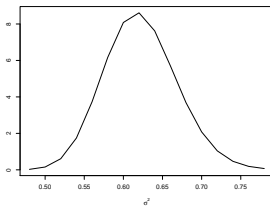
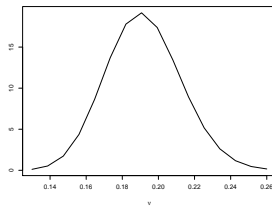
(b) $I_{\nu} = (0.13, 0.26)$ dados $\mu = 11, \sigma^2 = 0.64$

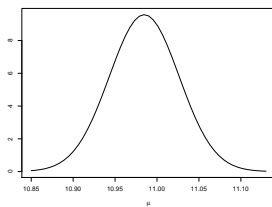
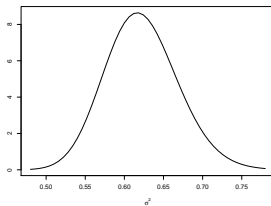
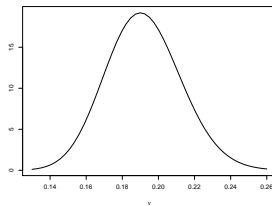
Figura 3: Intervalos de massa probabilística para cada parâmetro (variável aleatória) do núcleo de $p(\mu, \sigma^2, \nu | \mathbf{x})$

- Antes de aproximar as densidades *a posteriori* marginais de cada parâmetro, é necessário aproximar o inverso da constante de proporcionalidade.
- Dados três parâmetros $(\alpha_1, \alpha_2, \alpha_3)$ e uma amostra dos dados \mathbf{y} , quaisquer, suponha que se deseja aproximar a densidade *a posteriori* marginal de α_3 dados os pontos r_i, s_j, t_k da grade formada por todos os subintervalos de integração, $i, j, k \in \{1, \dots, L\}$. Temos pela quadratura de Riemann que

$$\begin{aligned}
 p(\alpha_3 | \mathbf{y}) &= \iint p(\alpha_1, \alpha_2, \alpha_3 | \mathbf{y}) d\alpha_1 d\alpha_2 \\
 \Rightarrow p(t_k | \mathbf{y}) &= \iint p(\alpha_1, \alpha_2, t_k | \mathbf{y}) d\alpha_1 d\alpha_2 \approx \sum_{i=1}^L \sum_{j=1}^L p(r_i, s_j, t_k | \mathbf{y}) \Delta_i \Delta_j \\
 &= \sum_{i=1}^L \sum_{j=1}^L c \cdot h(r_i, s_j, t_k | \mathbf{y}) \Delta_i \Delta_j.
 \end{aligned} \tag{5}$$

- Como c , a constante de proporcionalidade, é dada pelo inverso da densidade *a priori* preditiva $f(\mathbf{y})$, a qual é obtida integrando-se em todo o espaço paramétrico o produto entre a função de verossimilhança $f(\mathbf{y}|\alpha_1, \alpha_2, \alpha_3)$ e as densidades (ou funções de probabilidade) *a priori* para α_1 , α_2 e α_3 , também é possível aproximar c pela quadratura de Riemann
- Neste caso, $c^{-1} \approx \sum_{i=1}^L \sum_{j=1}^L \sum_{k=1}^L h(r_i, s_j, t_k|\mathbf{y})\Delta_i\Delta_j\Delta_k$. Com o valor aproximado para c , é possível calcular (5) nos limites superior e inferior de todos os subintervalos de um dado parâmetro e enfim obter uma aproximação da densidade *a posteriori* marginal deste mesmo parâmetro através de uma curva gráfica que liga todos os valores calculados

(a) Densidade *a posteriori* de μ (b) Densidade *a posteriori* de σ^2 (c) Densidade *a posteriori* de ν Figura 4: Densidades *a posteriori* marginais pela quadratura de Riemann com $L = 15$

(a) Densidade *a posteriori* de μ (b) Densidade *a posteriori* de σ^2 (c) Densidade *a posteriori* de ν Figura 5: Densidades *a posteriori* marginais pela quadratura de Riemann com $L = 50$

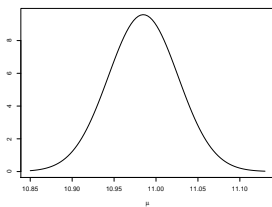
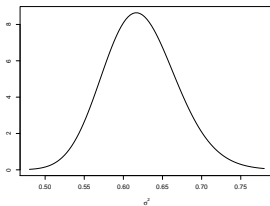
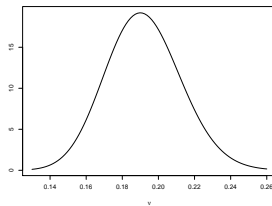
(a) Densidade *a posteriori* de μ (b) Densidade *a posteriori* de σ^2 (c) Densidade *a posteriori* de ν Figura 6: Densidades *a posteriori* marginais pela quadratura de Riemann com $L = 100$

Tabela 1: Estatísticas *a posteriori* para (μ, σ^2, ν) pela quadratura de Riemann

Cenário	Parâmetro	Média	Variância	Assimetria	Curtose
$L = 15$	μ	10.9847	0.0017	0.0022	2.9603
	σ^2	0.6222	0.0022	0.2230	2.9984
	ν	0.1918	0.0004	0.1572	2.9348
$L = 50$	μ	10.9848	0.0017	0.0056	2.9346
	σ^2	0.6222	0.0021	0.2149	2.9625
	ν	0.1918	0.0004	0.1523	2.8995
$L = 100$	μ	10.9848	0.0017	0.0064	2.9284
	σ^2	0.6222	0.0021	0.2131	2.9542
	ν	0.1918	0.0004	0.1512	2.8913

- Proposto por Gordon *et al.* (1993), o método SIR utiliza uma *função de amostragem por importância* g para aproximar (sem perda de generalidade) uma densidade de interesse p .
- Sejam $\theta_1, \dots, \theta_t$ uma amostra aleatória de g e $\mathbf{y} = (y_1, \dots, y_n)$ uma amostra do modelo para os dados observados. Para cada ponto θ_j , $j = 1, \dots, t$, os pesos são dados por

$$w_j(\theta_j) = \frac{p(\theta_j|\mathbf{y})/g(\theta_j)}{\sum_{j=1}^k p(\theta_j|\mathbf{y})/g(\theta_j)}. \quad (6)$$

em que g é uma densidade conhecida e da qual se sabe gerar uma amostra aleatória.

- Como feito em muitos trabalhos, para g será escolhida uma densidade normal trivariada $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, cujas componentes têm, cada uma, suporte em toda a reta real.

- Note que para 2 parâmetros, σ^2 e ν , o respectivo espaço paramétrico não é a reta real ($\Theta_{\sigma^2} = \mathbb{R}_+$ e $\Theta_{\nu} = [0, 1]$, respectivamente). logo será feita uma reparametrização.
- Para a reparametrização, consideram-se as transformações $\theta_1 = \mu$, $\theta_2 = \log(\sigma^2)$ e $\theta_3 = \log[\nu/(1 - \nu)]$. Logo, a expressão do núcleo reparametrizado é dada por

$$\begin{aligned}
 p(\theta_1, \theta_2, \theta_3 | \mathbf{x}) &= p(\theta_1 = \mu, \theta_2 = \log(\sigma^2), \theta_3 = \log[\nu/(1 - \nu)] | \mathbf{x}) \\
 &= p(\mu = \theta_1, \sigma^2 = \exp(\theta_2), \nu = 1/[1 + \exp(-\theta_3)] | \mathbf{x}) \times |J(\theta_1, \theta_2, \theta_3)| \\
 &\propto [\exp(\theta_2)]^{-[(n+1)/2+a+1]} \times \exp \left\{ -\frac{[(\theta_1 - m)^2/(2V) + d]}{\exp(\theta_2)} \right\} \\
 &\times A^*(\mathbf{x} | \theta_1, \theta_2, \theta_3) \times \frac{\exp(\theta_2) \exp(\theta_3)}{[1 + \exp(-\theta_3)]^{-2}} \\
 &\propto \left(\frac{1}{\sigma^2} \right)^{\frac{n+1}{2+a+1}} \times \exp \left\{ -\frac{[(\mu - m)^2/(2V) + d]}{\sigma^2} \right\} \times A(\mathbf{x} | \mu, \sigma^2, \nu) \\
 &\times \sigma^2 \nu^3 (1 - \nu)^{-1},
 \end{aligned} \tag{7}$$

- em que $|J(\theta_1, \theta_2, \theta_3)|$ é o determinante da matriz jacobiana das derivadas parciais de (μ, σ^2, ν) com respeito a $(\theta_1, \theta_2, \theta_3)$.
- Para a amostra de tamanho $n = 500$ da mistura finita de normais com variância contaminada tal que $\mu = 11$; $\sigma^2 = 0.64$; $\nu = 0.2$; $m = 11$; $V = 1$; $a = 7$ e $d = 4$ (Figura 1).
- Para as médias das componentes desta distribuição, será fixado $\boldsymbol{\mu} = (\theta_1, \theta_2, \theta_3) = (\mu, \log(\sigma^2), \log[\nu/(1 - \nu)]) = (11, \log(0.64), \log[0.2/0.8])$.
- Para a matriz de covariância $\boldsymbol{\Sigma}$, cada elemento da diagonal principal será dado pelo quadrado de $1/6$ do intervalo de massa probabilística do parâmetro correspondente.
- Esta escolha se justifica pelo fato de que as distribuições mostradas de 2a a 3b têm comportamento próximo à normalidade.

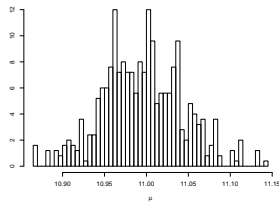
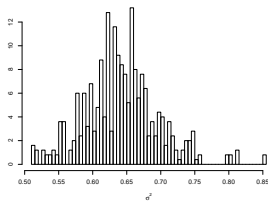
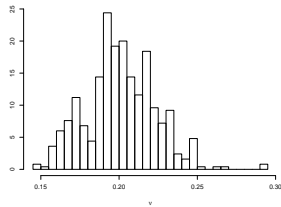
(a) Histograma de μ (b) Histograma de σ^2 (c) Histograma de ν

Figura 7: Histograma das densidades *a posteriori* marginais pela método SIR com $k = 500$

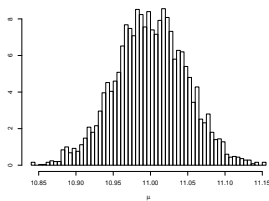
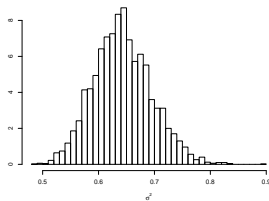
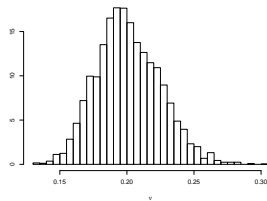
(a) Histograma de μ (b) Histograma de σ^2 (c) Histograma de ν

Figura 8: Histograma das densidades *a posteriori* marginais pela método SIR com $k = 5000$

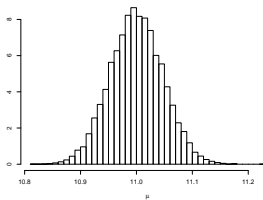
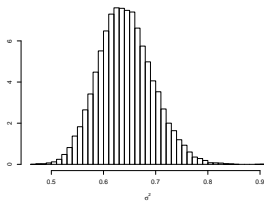
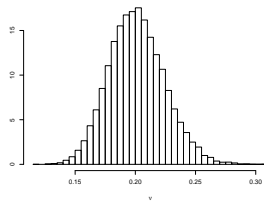
(a) Histograma de μ (b) Histograma de σ^2 (c) Histograma de ν

Figura 9: Histograma das densidades *a posteriori* marginais pela método SIR com $k = 50000$

Tabela 2: Estatísticas *a posteriori* para (μ, σ^2, ν) pelo método SIR

Cenário	Parâmetro	Média	Variância	Assimetria	Curtose
$k = 500$	μ	11.0003	0.0023	-0.0610	3.0351
	σ^2	0.6422	0.0026	0.1261	3.1453
	ν	0.2028	0.0006	0.3510	3.5475
$k = 5000$	μ	11.0006	0.0022	0.0216	3.0234
	σ^2	0.6418	0.0027	0.2349	3.0118
	ν	0.2013	0.0005	0.2485	3.1646
$k = 50000$	μ	11.0002	0.0022	-0.0081	3.0451
	σ^2	0.6420	0.0027	0.2353	3.1153
	ν	0.2009	0.0005	0.2608	3.0849

- Neste trabalho, será usado o algoritmo de Metropolis-Hastings (Metropolis *et al.*, 1953; Hastings, 1970), aqui abreviado por MH.
- Assim como o método SIR, o algoritmo MH também é baseado no uso de uma distribuição auxiliar ou proposta, aqui denotada por $q(\mathbf{y}, \mathbf{z})$. Assumindo-se que na interação j , $j = 1, \dots, k$ a cadeia está no estado $\mathbf{y}^{(j)}$, a posição da mesma na interação $j + 1$, denotada por $\mathbf{y}^{(j+1)}$, será dada após:
- Propor uma transição ou movimento para \mathbf{y}^* , onde \mathbf{y}^* é gerada de $q(\mathbf{y}^{(j)}, \cdot)$, a distribuição proposta, e, o valor inicial da cadeia $\mathbf{y}^{(1)}$;
- Aceitar a transição proposta com probabilidade

$$\rho(\mathbf{y}^{(j)}, \mathbf{y}^*) = \min \left(1, \frac{\rho(\mathbf{y}^*)/q(\mathbf{y}^{(j)}, \mathbf{y}^*)}{\rho(\mathbf{y}^{(j)})/q(\mathbf{y}^*, \mathbf{y}^{(j)})} \right) \quad (8)$$

e neste caso atribuir $\mathbf{y}^{(j+1)} = \mathbf{y}^*$ ou rejeitar a transição proposta e atribuir $\mathbf{y}^{(j+1)} = \mathbf{y}^{(j)}$, com probabilidade $1 - \rho(\mathbf{y}^{(j)}, \mathbf{y}^*)$.

- Para decidir sobre a aceitação ou não de \mathbf{y}^* quando amostrada a cada passo j , gere uma amostra u_1, \dots, u_k , onde k é o total de iterações prefixadas, da distribuição uniforme padrão $U(0, 1)$, independentemente de \mathbf{y}^* .
- Se a probabilidade de aceitação $\rho(\mathbf{y}^{(j)}, \mathbf{y}^*)$ for maior do que ou igual a u_j , então a transição proposta é aceita. Do contrário, ela é rejeitada.
- Para a distribuição proposta $q(\cdot)$ também será escolhida uma normal trivariada $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com valores do vetor de médias, e será reutilizado a reparametrização $\boldsymbol{\mu} = (\mu, \log(\sigma^2), \log[\nu/(1 - \nu)]) = (11, \log(0.64), \log[0.2/0.8])$ e da matriz de covariância $\boldsymbol{\Sigma} = \text{diag}\{0.0022, 0.0065, 0.0203\}$.
- É necessário definir um estado inicial da cadeia, em geral com densidade conjunta muito baixa, escolheu-se o ponto $\mathbf{y}^{(1)} = (10.86, \log(0.50), \log(0.14/0.86))$.

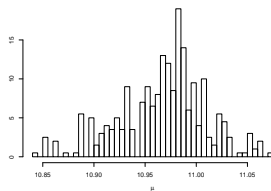
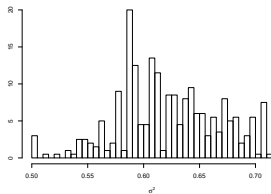
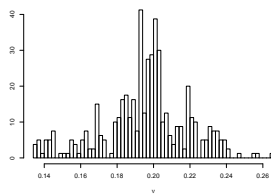
(a) Histograma de μ (b) Histograma de σ^2 (c) Histograma de ν

Figura 10: Histograma das densidades *a posteriori* marginais pelo método MCMC–MH, $k = 500$

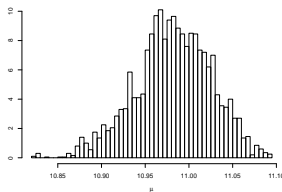
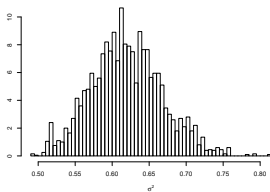
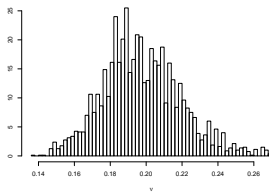
(a) Histograma de μ (b) Histograma de σ^2 (c) Histograma de ν

Figura 11: Histograma das densidades *a posteriori* marginais pela método MCMC–MH, $k = 5000$

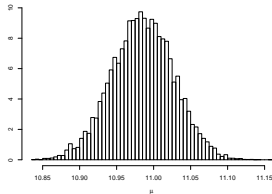
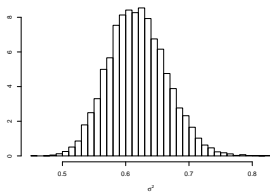
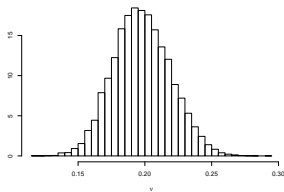
(a) Histograma de μ (b) Histograma de σ^2 (c) Histograma de ν

Figura 12: Histograma das densidades *a posteriori* marginais pela método MCMC–MH, $k = 50000$

Tabela 3: Estatísticas *a posteriori* para (μ, σ^2, ν) pelo método MCMC–MH

Cenário	Parâmetro	Média	Variância	Assimetria	Curtose
$k = 500$	μ	10.9672	0.0018	-0.3768	3.2193
	σ^2	0.6226	0.0020	-0.0684	2.6117
	ν	0.1957	0.0006	-0.3313	3.2991
$k = 5000$	μ	10.9818	0.0019	-0.2466	2.9894
	σ^2	0.6197	0.0023	0.2372	3.0057
	ν	0.1977	0.0005	0.3878	3.2038
$k = 50000$	μ	10.9852	0.0017	-0.0272	2.9451
	σ^2	0.6188	0.0021	0.2472	3.1088
	ν	0.1978	0.0005	0.1331	2.8971

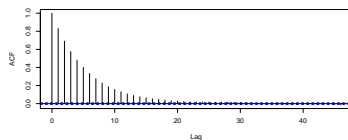
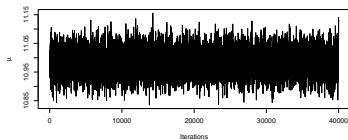
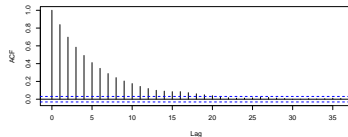
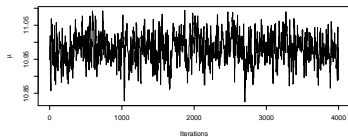
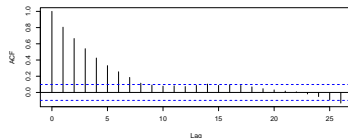
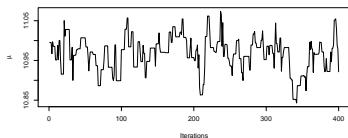
(a) Traço de μ (b) Autocorrelação de μ

Figura 13: Traço e autocorrelação da cadeia de μ , $k = \{500, 5000, 50000\}$, respectivamente

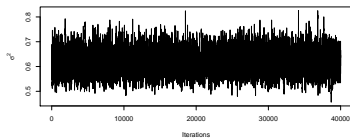
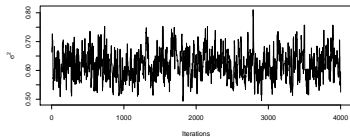
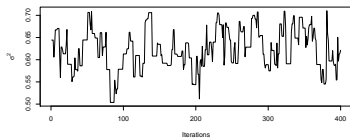
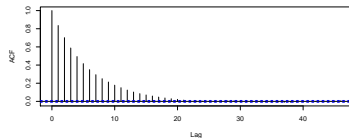
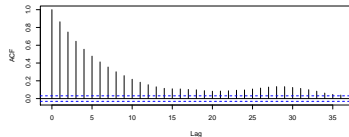
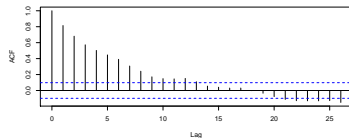
(a) Traço de σ^2 (b) Autocorrelação de σ^2

Figura 14: Traço e autocorrelação da cadeia de σ^2 , $k = \{500, 5000, 50000\}$, respectivamente

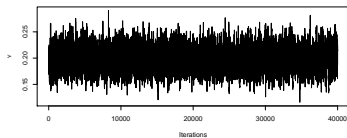
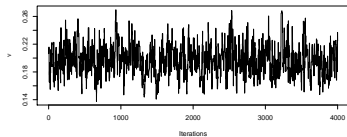
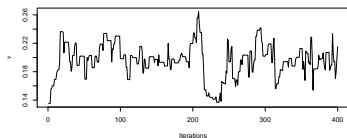
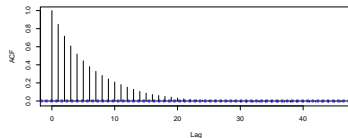
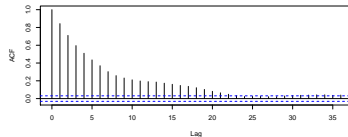
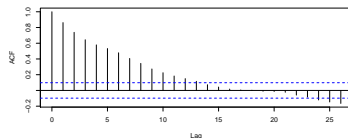
(a) Traço de ν (b) Autocorrelação de ν^2

Figura 15: Traço e autocorrelação da cadeia de ν^2 , $k = \{500, 5000, 50000\}$, respectivamente

- Neste trabalho, as densidades marginais *a posteriori* dos parâmetros de interesse (μ, σ^2, ν) foram aproximadas por três métodos numéricos, para uma amostra observada de tamanho 500.
- Para os três métodos implementados, foram desenvolvidos 3 cenários diferentes, para análise de convergência e complexidade computacional.
- No primeiro caso, do método de quadratura, três quantidades diferentes de intervalos de integração L foram escolhidos: 15, 50 e 100. O primeiro resultado não é muito satisfatório, apesar de aproximar bem os dois primeiros momentos.
- Para 50 e 100 subintervalos, a aproximação é bem melhor, às custas de um custo computacional maior, crescendo exponencialmente em relação a dimensão do espaço paramétrico ($O(L^n)$), com n igual a 3.

- Para os métodos estocásticos SIR e MCMC, foram considerados, respectivamente, quantidade de iterações k iguais a 500, 5000 e 50000.
- Em ambos os casos, para k igual a 500, as distribuições a posteriori aproximadas não foram muito satisfatórias, apresentando muitos outliers e modas locais.
- Para o SIR, a partir de 5000 amostras, as distribuições são bem mais suaves, sendo possível observar o comportamento de momentos de ordem superior.
- No MCMC, ainda se observa uma melhoria considerável de k igual a 5000 para k igual a 50000 passos de integração, quando o método parece convergir.
- Por fim, podemos concluir que, para o modelo probabilístico estudado, todos os três métodos desempenham bem a tarefa de aproximar a distribuição a posteriori, se escolhidos hiperparâmetros de convergência adequados.