Mistura de Normais com Variância Contaminada

Caio Gabriel Barreto Balieiro [1]

Taiguara Melo Tupinambás [1]

Walmir dos Reis Miranda Filho[1]

Programa de Pós-Graduação em Estatística

Departamento de Estatística - UFMG

02 de dezembro de 2019

Conteúdo

- 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 $\boldsymbol{\theta} = (\mu, \sigma^2, \nu)$, e identicamente distribuídas com função densidade dada por

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

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

$$p(\boldsymbol{\theta}) = p(\mu|\sigma^2)p(\sigma^2)p(\nu). \tag{2}$$

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

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

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 defina da seguinte forma

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

onde

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

- 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.

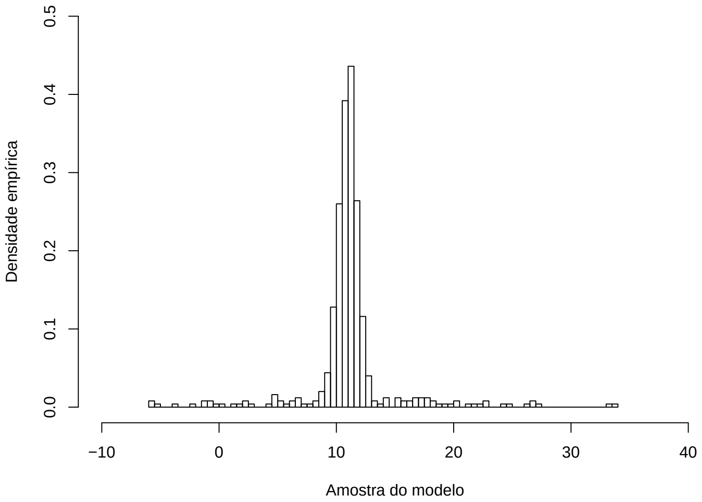
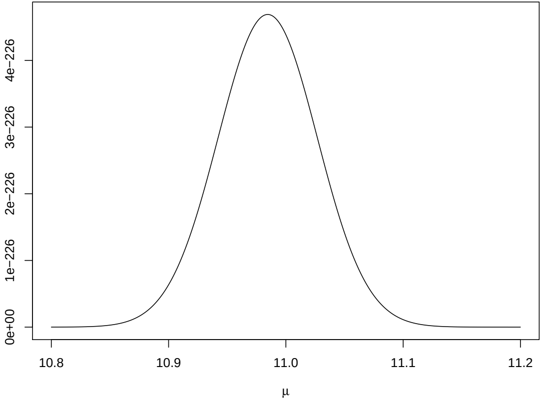
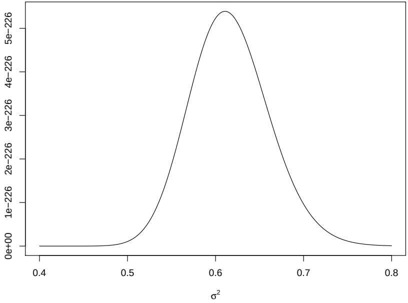


Figura: Histograma da amostra gerada do modelo

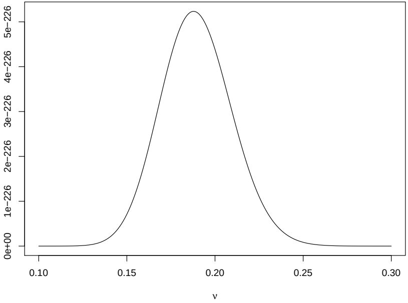


(a) $I_\mu = (10.85, 11.13)$, dados $\sigma^2 = 0.64, \nu = 0.2$

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



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

- 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, $i, j, k \in \{1, \ldots, L\}$. Temos pela quadratura de Riemann que

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

- Como c, a constante de proporcionalidade, é dada pelo inverso da densidade *a priori* preditiva $f(\boldsymbol{y})$, a qual é obtida integrando-se em todo o espaço paramétrico o produto entre a função de verossimilhança $f(\boldsymbol{y}|\alpha_1, \alpha_2, \alpha_3)$ 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, $c^{-1} \approx \sum_{i=1}^{L} \sum_{j=1}^{L} \sum_{k=1}^{L} h(r_i, s_j, t_k|\boldsymbol{y})\Delta_i\Delta_j\Delta_k$. 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

Tabela: 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 $\theta_1, \ldots, \theta_k$ uma amostra aleatória de g e $\boldsymbol{y} = (y_1, \ldots, y_n)$ uma amostra do modelo para os dados observados. Para cada ponto θ_j, $j = 1, \ldots, k$, os pesos são dados por
$$w_j(\theta_j) = \frac{p(\theta_j|\boldsymbol{y})/q(\theta_j)}{\sum_{j=1}^{k} p(\theta_j|\boldsymbol{y})/q(\theta_j)}. \tag{6}$$
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

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

- 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.

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

Cenário	Parâmetro	Média	Variância	Assimetria	Curtose
$k = 500$	μ	10.9956	0.0024	0.1394	3.0639
	σ^2	0.6423	0.0030	0.4046	4.0116
	ν	0.2010	0.0005	0.3050	3.5598
$k = 5000$	μ	11.0001	0.0023	-0.0244	2.9656
	σ^2	0.6428	0.0027	0.2869	3.1554
	ν	0.2010	0.0006	0.3157	3.0594
$k = 50000$	μ	11.0001	0.0022	0.0266	3.0106
	σ^2	0.6422	0.0027	0.2529	3.1077
	ν	0.2012	0.0005	0.2654	3.1680