

Inferência Bayesiana para a distribuição Power Lindley

Gustavo Durães¹

²Departamento de Estatística, Universidade de Brasília, Brasília, Brasil

Resumo. Este trabalho apresenta uma proposição bayesiana para a estimação dos parâmetros de interesse da distribuição power Lindley utilizando o software JAGS

Abstract. This work presents a Bayesian proposition for power Lindley distribution parameters using JAGS software

Keywords: Power Lindley, Bayesian Inference, JAGS: .

1 Introdução

A distribuição power Lindley foi proposta inicialmente por Ghitany et al. (2013) e sua função de densidade de probabilidade pode ser expressa como:

$$f(x|\alpha, \delta) = \begin{cases} \frac{\beta}{\beta+1} f_1(x|\alpha, \delta) + \left(1 - \frac{\beta}{\beta+1}\right) f_2(x|\alpha, \delta), & x > 0, \alpha, \beta > 0 \\ 0, & x < 0 \end{cases} \quad (1)$$

Onde,

$$f_1(x|\alpha, \delta) = \alpha \beta x^{\alpha-1} e^{-\beta x^\alpha}, \quad x > 0, \alpha, \beta > 0$$

$$f_2(x|\alpha, \delta) = \alpha \beta^2 x^{2\alpha-1} e^{-\beta x^\alpha}, \quad x > 0, \alpha, \beta > 0$$

A partir de (1), pode-se notar que a distribuição power Lindley é uma mistura das distribuições Weibull e Gamma Generalizada respectivamente. Para simplificar a interpretação da modelagem Bayesiana para a inferência dos parâmetros, pode-se realizar uma transformação no parâmetro de mistura $\beta = \frac{\delta}{1-\delta}$, que resulta na seguinte função de densidade de probabilidade:

$$g(x|\alpha, \delta) = \begin{cases} \delta g_1(x|\alpha, \delta) + (1 - \delta) g_2(x|\alpha, \delta), & x > 0, \alpha > 0, 0 < \delta < 1 \\ 0, & x < 0 \end{cases} \quad (2)$$

Onde,

$$g_1(x|\alpha, \delta) = \alpha \left(\frac{\delta}{1-\delta} \right) x^{\alpha-1} e^{-\left(\frac{\delta}{1-\delta}\right)x^\alpha}, \quad x > 0, \alpha > 0, 0 < \delta < 1$$

$$g_2(x|\alpha, \delta) = \alpha \left(\frac{\delta}{1-\delta} \right)^2 x^{2\alpha-1} e^{-\left(\frac{\delta}{1-\delta}\right)x^\alpha}, \quad x > 0, \alpha > 0, 0 < \delta < 1$$

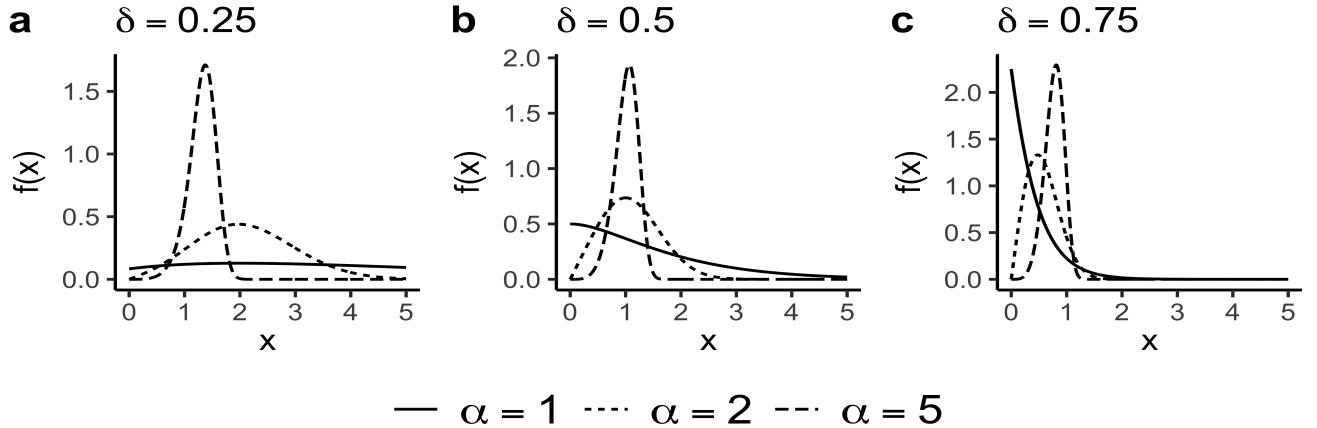


Figura 1: Gráficos da função de densidade de probabilidade power Lindley para diferentes valores do parâmetro de mistura δ e parâmetro de forma α . Valores de δ que se encontram próximos dos limites paramétricos resultam em densidades muito próximas aos componentes da mistura.

2 Método Bayesiano

A filosofia da inferência Bayesiana nos permite representar nossa incerteza acerca dos parâmetros através da atribuição de uma densidade de probabilidade adequada para, após a análise dos dados amostrais, analisar a distribuição posterior dos parâmetros afim de obter informações através do teorema de Bayes. Para o caso da distribuição power Lindley, é conveniente atribuirmos ao parâmetro α uma distribuição Gamma e ao parâmetro de mistura δ uma distribuição Beta. Dessa forma, o modelo se da por:

$$f(\alpha, \delta | \mathbf{x}) = \frac{f(\mathbf{x}|\alpha, \delta)f(\alpha)f(\delta)}{K} \propto L(\alpha, \delta | \mathbf{x})f(\alpha)f(\delta) \quad (3)$$

Podemos escrever a distribuição a posteriori para cada parâmetro separadamente da seguinte forma:

$$f(\alpha | \mathbf{x}, \delta) \propto L(\alpha, \delta | \mathbf{x})f(\alpha) \quad (4)$$

$$f(\delta | \mathbf{x}, \alpha) \propto L(\alpha, \delta | \mathbf{x})f(\delta) \quad (5)$$

Onde,

$$f(\alpha) = \frac{\theta^k}{\Gamma(k)} \alpha^{k-1} e^{-\theta\alpha} \quad \alpha > 0, k > 0, \theta > 0$$

$$f(\delta) = \frac{\delta^{\lambda-1} (1-\delta)^{\phi-1}}{B(\lambda, \phi)} \quad 0 < \delta < 1, \lambda > 0, \phi > 0$$

A obtenção das densidades a posteriori de forma analítica não é trivial. Uma solução para esse problema é a utilização de método de Monte Carlo utilizando Cadeias de Markov (MCMC), mais especificamente utilizando o método de Gibbs. Uma ferramenta possível para a execução desse procedimento é o software Jags Plummer (2003), que posse ser utilizado através do software R R Core Team (2019) utilizando o pacote rjags.

3 Implementação do Modelo

O software JAGS gera amostras aleatórias da distribuição a posteriori a partir das distribuições condicionais a priori e da verossimilhança. Inicialmente é necessário que seja estabelecido o modelo, entretanto, a distribuição power Lindley não está diretamente especificada no código base do JAGS, então, é necessário utilizar uma técnica conhecida como "Bernoulli ones trick" Kruschke (2015).

A partir dos dados acerca da força de tensão de 69 fibras de carbono utilizados em Ghitany et al. (2013), podemos realizar um procedimento análogo ao artigo e comparar resultados.

O modelo criado no R e aplicado no JAGS é o seguinte:

```
#Dados
y = c(1.312, 1.314, 1.479, 1.552, 1.700, 1.803, 1.861, 1.865, 1.944,
     , 1.958, 1.966, 1.997, 2.006, 2.021, 2.027, 2.055, 2.063, 2.098,
     2.14, 2.179, 2.224, 2.240, 2.253, 2.270, 2.272, 2.274, 2.301, 2.301,
     2.359, 2.382, 2.382, 2.426, 2.434, 2.435, 2.478, 2.490, 2.511,
     2.514, 2.535, 2.554, 2.566, 2.57, 2.586, 2.629, 2.633, 2.642, 2.648,
     2.684, 2.697, 2.726, 2.770, 2.773, 2.800, 2.809, 2.818, 2.821,
     2.848, 2.88, 2.954, 3.012, 3.067, 3.084, 3.090, 3.096, 3.128, 3.233,
     3.433, 3.585, 3.585)

#Tamanho da Amostra
N = length(y)
# Banco de dados
dataList = list(
  y=y, Ntotal=Ntotal  )

Modelo="
data {C <- 100000 ##Constante o para garantir py[i] <1
for(i in 1:N){
  ones[i] <- 1  }
model{
```

```

for (i in 1:Ntotal ) {
## Verossimilhanca criada utilizando "Bernoulli ones Trick"
py[ i ] <- (d*(a*(d/(1-d))*y[ i ]^(a-1)*exp(-d/(1-d)*y[ i ]^a))+  

(1-d)*((a*(d/(1-d))*2*y[ i ]^(2*a-1)*exp(-d/(1-d)*y[ i ]^a))))/C  

ones[ i ]~dbern(py[ i ])
}
## Priori para alpha
a~dgamma(39,10)
## Priori para delta
d~dbeta(4,4) } "

```

Uma forma possível para a elicitação das distribuições a priori, para o caso do parâmetro α , é a solução do seguinte sistema de equações:

$$\begin{cases} Moda = \frac{k-1}{\theta} \\ Variância = \frac{k}{\theta^2} \end{cases} \quad (6)$$

Por exemplo, para a priori utilizada para o parâmetro α no modelo acima, obtemos uma moda igual a 3.8 e variância de 0.39, valores diferentes dessa razão resultam em prioris mais concentradas, atribuindo maior impacto da priori nas estimativas da posteriori ou mais achatadas resultando em um menor impacto.

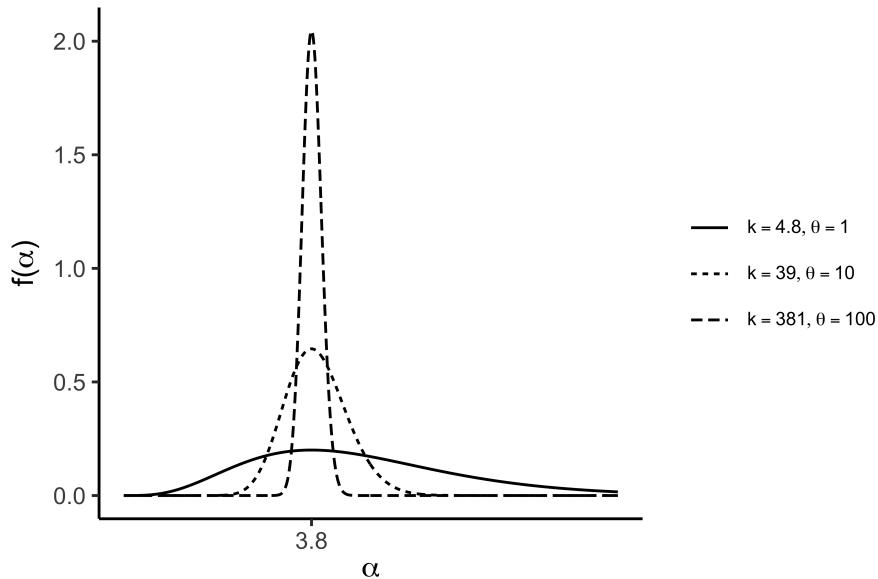


Figura 2: Função de densidade de probabilidade da distribuição Gamma

Para o parâmetro δ , a priori utilizada tem uma interpretação um pouco mais conveniente. Os parâmetros da distribuição beta podem ser interpretados como a proporção do total da amostra pertence a cada variável que compõe a mistura. A razão $\frac{\lambda}{\lambda+\phi}$ representa a "confiança" que essa proporção representa a proporção da mistura que segue distribuição Weibull, analogamente, a razão $\frac{\phi}{\lambda+\phi}$ representa a proporção da mistura que segue distribuição Gamma Generalizada. Além disso, quanto maior o valor

da soma $\lambda + \phi$ maior a confiança na priori estabelecida.

Também é necessário comentar que o parâmetro de escala das distribuições Weibull e Gamma Generalizada é determinado pelo valor de δ . O valor do parâmetro de escala das distribuições da mistura é maximizado a medida que a proporção da amostra é composta em sua maior parte pela distribuição Gamma Generalizada.

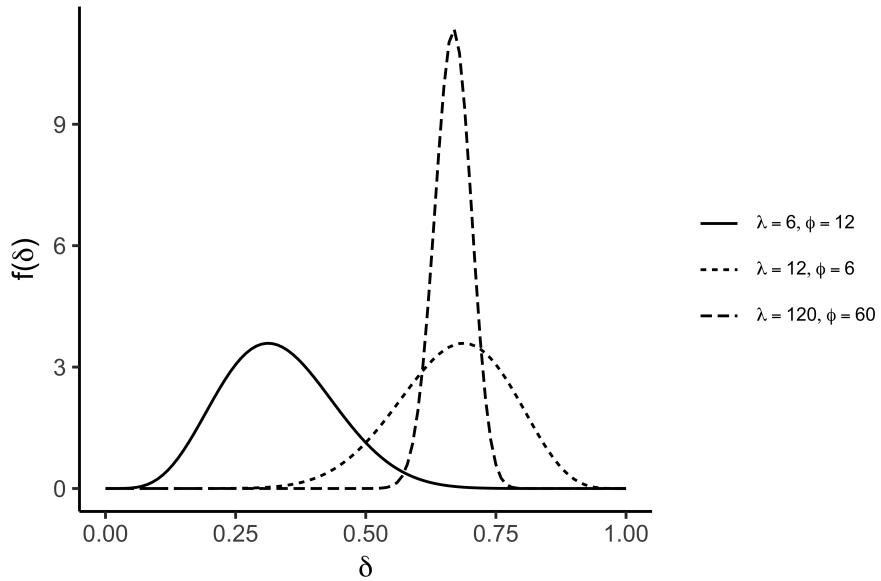


Figura 3: Função de densidade de probabilidade da Distribuição Beta.

Tendo essas informações em mente, podemos analisar as observações simuladas da distribuição á posteriori afim de obtermos informações razoáveis sobre as estimativas pontuais e intervalares do modelo. No caso proposto, foram geradas 10 mil amostras em 4 Cadeias de Markov paralelas. A análise de diagnóstico se dará pela análise visual dos Trace plots, Diagnóstico de Geweke e Diagnóstico de Gelman-Rubin. Também é interessante observação dos valores dos estimadores de máxima verossimilhança afim de comprar os métodos de estimação.

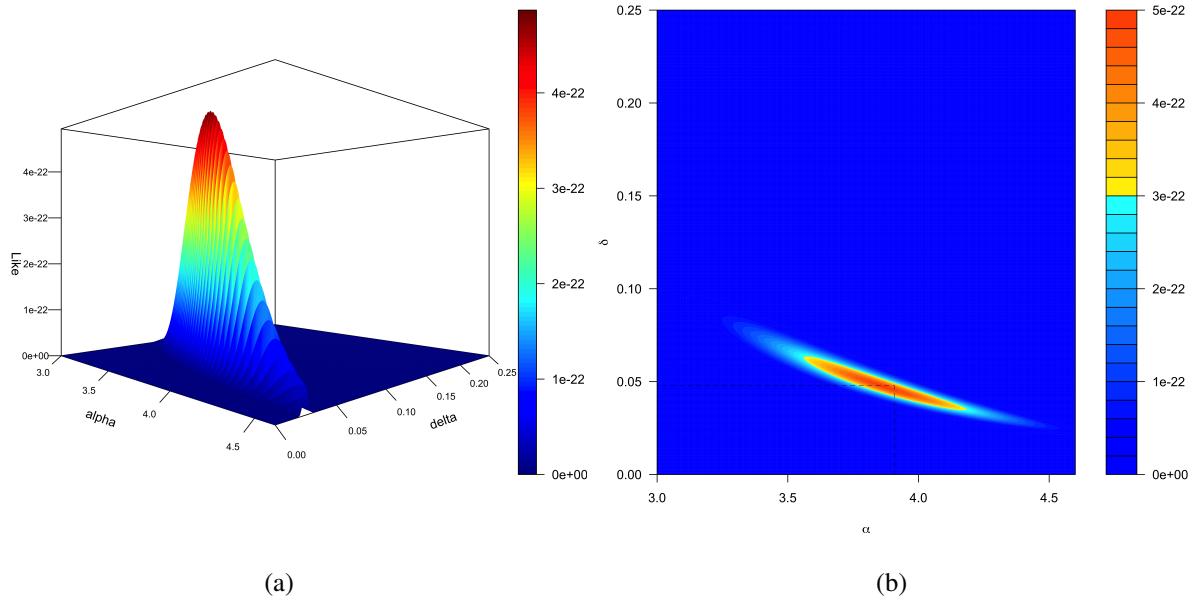


Figura 4: Podemos notar a partir dos gráficos da função de verossimilhança que os estimadores que a maximizam são, aproximadamente: $\alpha = 3.8$ e $\delta = 0.05$

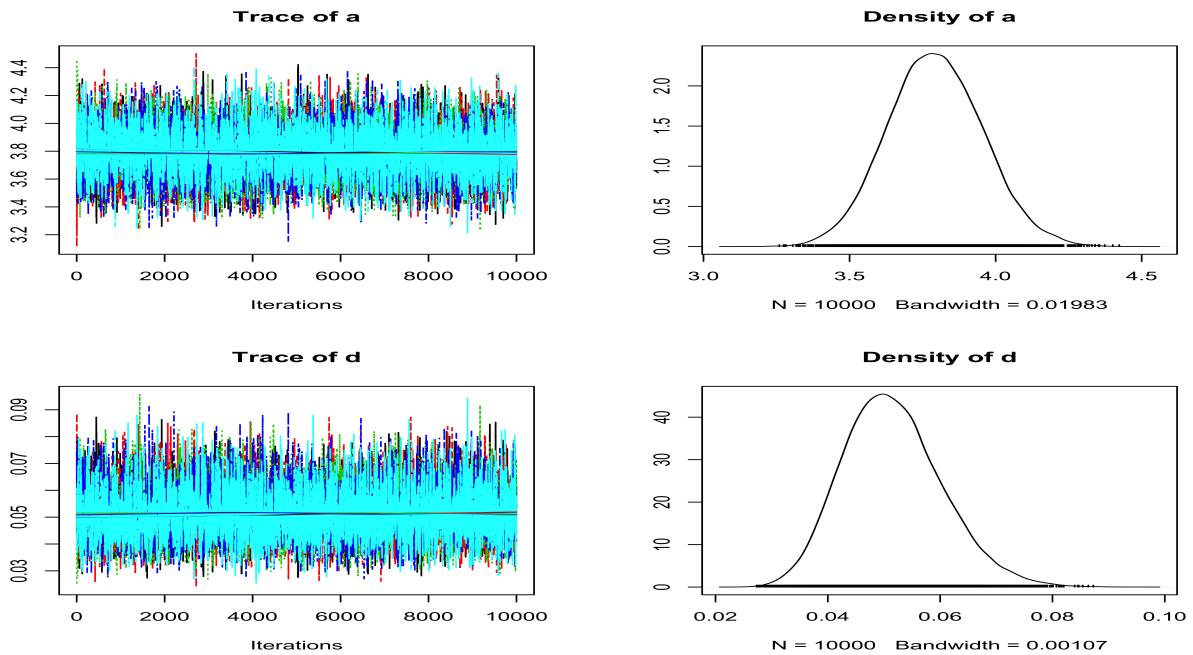


Figura 5: Trace plots e densidades empíricas da distribuição a posteriori.

Os Trace plots aparentam indicar convergência da simulação. As densidades simuladas nos forne-

cem as seguintes estimativas pontuais, através das medianas,

$$Md(f(\alpha|\boldsymbol{x}, \delta)) = 3.798157$$

$$Md(f(\delta|\boldsymbol{x}, \alpha)) = 0.05060089$$

Além disso, podemos calcular os intervalos de credibilidade á 95% para obter as estimativas inter-valares:

$$HDI(\alpha, 0.95) = (3.465; 4.100)$$

$$HDI(\delta, 0.95) = (0.0355; 0.07)$$

Os gráficos de diagnóstico de Geweke e Gelman aparentam indicar a convergência das simulações de Monte-Carlo, isso indica que as simulações geradas fornecem estimativas adequadas para as distribuições a posteriori.

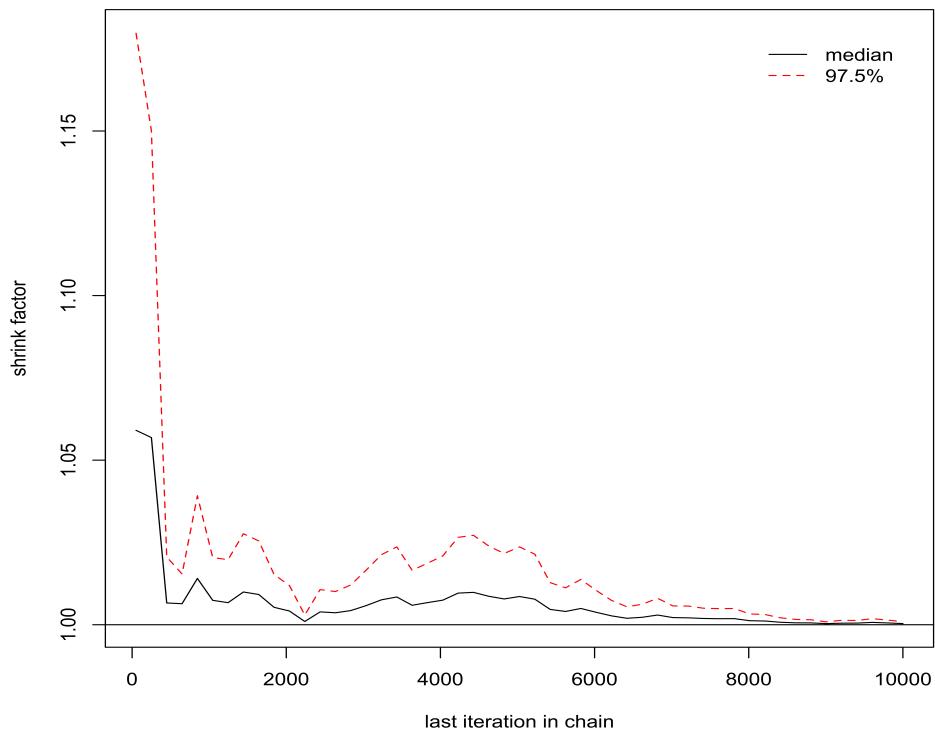


Figura 6: Diagnóstico de Geweke

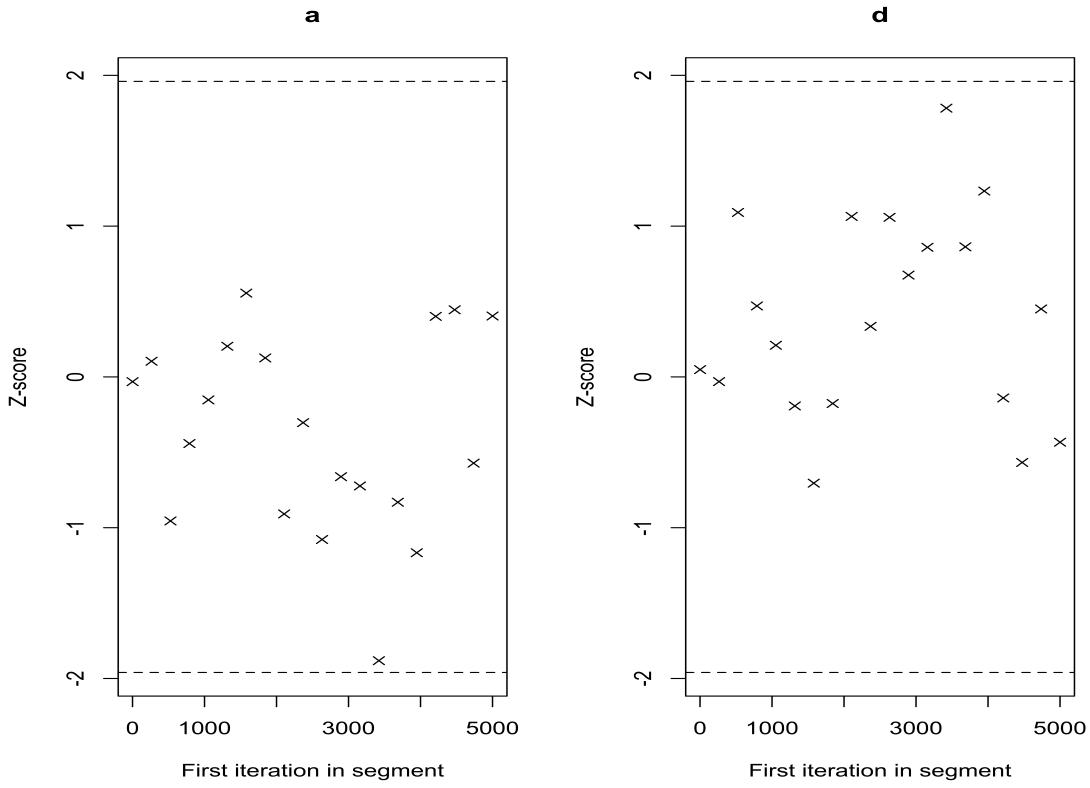


Figura 7: Diagnóstico de Gelman

4 Estudo do método de estimação para dados simulados via Monte Carlo

A partir do uso do método descrito nas seções anteriores, podemos avaliar a qualidade da aplicação através da estimação dos parâmetros em observações simuladas da distribuição power Lindley, variando o tamanho da amostra e analisando o percentual de intervalos de densidade máxima (HDI) contém o valor real. Espera-se que, ao se utilizar prioris razoavelmente apropriadas, o percentual se aproxime da credibilidade do intervalo a medida que o tamanho da amostra simulada aumente. Além disso, a medida que o tamanho da amostra cresça, é sabido que a priori estabelecida se torna menos importante.

Para o caso apresentado nas tabelas 1 e 2, foram utilizadas 1000 iterações do algoritmo de Gibbs e prioris fixas em relação ao tamanho da simulações da verossimilhança. Para obter o percentual de capturas do HDI gerado por essas amostras, foram geradas 100 simulações de cada iteração do procedimento.

A simulação das observações da verossimilhança foi feita de forma análoga ao algoritmo 2 proposto por Ghitany et al. (2013), utilizando funções nativas do software R para simular observações das densidades Weibull e Gama Generalizada que compõe a mistura.

α	δ	n	Captura para o HDI(50%)	$f(\alpha)$	$f(\delta)$	
0.2	0.8	25	0.57	$\alpha \sim Gamma(20, 100)$	$\delta \sim Beta(40, 10)$	
		50	0.58			
		75	0.46	$\alpha \sim Gamma(150, 100)$		
		100	0.53			
		200	0.49			
1.5	0.5	25	0.86	$\alpha \sim Gamma(150, 100)$	$\delta \sim Beta(25, 25)$	
		50	0.76			
		75	0.64			
		100	0.55			
		200	0.55			
0.9	0.26	25	0.74	$\alpha \sim Gamma(90, 100)$	$\delta \sim Beta(20, 60)$	
		50	0.64			
		75	0.68			
		100	0.61			
		200	0.51			

Tabela 1: Resultados para o parâmetro α

α	δ	n	Captura para o HDI(50%)	$f(\alpha)$	$f(\delta)$	
0.2	0.8	25	0.58	$\alpha \sim Gamma(20, 100)$	$\delta \sim Beta(40, 10)$	
		50	0.6			
		75	0.57	$\alpha \sim Gamma(150, 100)$		
		100	0.46			
		200	0.5			
1.5	0.5	25	0.51	$\alpha \sim Gamma(150, 100)$	$\delta \sim Beta(25, 25)$	
		50	0.47			
		75	0.59			
		100	0.37			
		200	0.49			
0.9	0.26	25	0.74	$\alpha \sim Gamma(90, 100)$	$\delta \sim Beta(20, 60)$	
		50	0.54			
		75	0.59			
		100	0.55			
		200	0.58			

Tabela 2: Resultados para o parâmetro δ

5 Considerações Finais

Os parâmetros estimado se mostram muito próximos dos encontrados em Ghitany et al. (2013). A modelagem aqui apresentada se mostra vantajosa por permitir a inclusão da subjetividade do pesquisador através da elicitação de prioris apropriadas, além disso, a interpretação dos parâmetros se dá de forma um pouco mais natural e mais tangível.

Referências

- Ghitany, M., Al-Mutairi, D., Balakrishnan, N., and Al-Enezi, L. (2013). Power lindley distribution and associated inference. *Computational Statistics Data Analysis*, 64:20 – 33.
- Kruschke, J. K. (2015). Chapter 8 - jags. In Kruschke, J. K., editor, *Doing Bayesian Data Analysis (Second Edition)*, pages 193 – 219. Academic Press, Boston, second edition edition.
- Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. *3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria*, 124.
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.