Problemas de Dimensão Variável

Ricardo Ehlers
ehlers@icmc.usp.br

Departamento de Matemática Aplicada e Estatística
Universidade de São Paulo

Ch. 7 in Gamerman & Lopes

"One of the things we do not know is the number of things we do not know"
Peter Green

Em muitas aplicações práticas é razoável assumir que existe incerteza também em relação ao modelo que melhor se ajusta a um conjunto de dados.

- Crie uma variável aleatória discreta k (o indicador de modelo) e atribua probabilidades $p(k)$.
- Para cada k existe um vetor de parâmetros $\boldsymbol{\theta}^{(k)} \in \mathbb{R}^{n_k}$ com
  - uma função de verossimilhança: $p(\mathbf{y}|\boldsymbol{\theta}^{(k)}, k)$,
  - uma distribuição a priori: $p(\boldsymbol{\theta}^{(k)}|k)$.

A distribuição de interesse agora é dada por,

$$\pi(\boldsymbol{\theta}, k|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}, k)\ p(\boldsymbol{\theta}|k)\ p(k)$$

e temos que simular valores desta distribuição.

- A dimensão de $\boldsymbol{\theta}$ pode variar ao longo dos modelos. Precisamos construir uma cadeia com espaço de estados que muda de dimensão ao longo das iterações.
- Os algoritmos de Metropolis-Hastings e o amostrador de Gibbs não podem ser utilizados já que são definidos apenas para distribuições com dimensão fixa.
- Embora existam outras possibilidades iremos estudar os algoritmos MCMC com saltos reversíveis.

Example. Sejam $Y_1, \ldots, Y_n$ os tempos de vida de componentes eletrônicos sorteados ao acaso e existe incerteza em relação a distribuição dos dados. Sabe-se que

$$Y_i \sim Exp(\lambda) \text{ (Modelo 1), ou}$$

$$Y_i \sim Gamma(\alpha, \beta) \text{ (Modelo 2),}$$

$i = 1, \ldots, n$.

- $k = 1$ (Modelo 1), $\theta^{(1)} = \lambda$,

$$p(\mathbf{y}|\lambda, k = 1) = \lambda^n e^{-\lambda \sum y_i}$$

- $k = 2$ (Modelo 2), $\boldsymbol{\theta}^{(2)} = (\alpha, \beta)$,

$$p(\mathbf{y}|\alpha, \beta, k = 2) = \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \prod y_i^{\alpha-1} e^{-\beta \sum y_i}.$$

Seja M conjunto de todos os possíveis modelos.

- As probabilidades a posteriori de cada possível modelo são dadas por,

$$\pi(k|\mathbf{y}) = \frac{p(k)\ p(\mathbf{y}|k)}{\sum_{k \in M} p(k)\ p(\mathbf{y}|k)},\ k \in M$$

- $p(\mathbf{y}|k)$ é a *verossimilhança marginal* obtida como,

$$p(\mathbf{y}|k) = \int p(\mathbf{y}|\boldsymbol{\theta}, k) p(\boldsymbol{\theta}|k) d\boldsymbol{\theta}.$$

- Esta última integral só é analiticamente tratável em alguns casos restritos.
- Se o número de modelos candidatos for muito grande calcular (ou aproximar) $p(\mathbf{y}|k)$ pode ser inviável na prática.

- Se for especificada a distribuição de interesse como,

$$\pi(\boldsymbol{\theta}, k|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}, k)\ p(\boldsymbol{\theta}|k)\ p(k)$$

  e conseguirmos simular valores desta distribuição então automaticamente teremos uma amostra aproximada de $\pi(k|\mathbf{y})$ e $\pi(\boldsymbol{\theta}|k, \mathbf{y})$.

MCMC com Saltos Reversiveis (RJMCMC)

- Proponha um novo valor para a cadeia e defina uma probabilidade de aceitação.
- Os movimentos podem ser entre espaços de dimensões diferentes.
- Em cada iteração atualize os parâmetros, dado o modelo, usando os métodos MCMC usuais.
- Atualize a dimensão.

Suponha que o estado atual da cadeia é $(k, \boldsymbol{\theta})$, i.e. estamos no modelo k com parâmetros $\boldsymbol{\theta}$

Um novo modelo k' com parâmetros $\boldsymbol{\theta}'$ é proposto com probabilidade $r_{k,k'}$. Em geral isto significa incluir ou retirar parâmetros do modelo atual.

Vamos assumir inicialmente que o modelo proposto tem dimensão maior, i.e. $n_{k'} > n_k$ e que $\boldsymbol{\theta}' = g(\boldsymbol{\theta}, \mathbf{u})$ para uma função deterministica g e um vetor aleatório $\mathbf{u} \sim q(\mathbf{u})$ com dimensão $n_{k'} - n_k$.

Então o seguinte algoritmo é utilizado,

1. proponha $(k, \boldsymbol{\theta}) \to (k', \boldsymbol{\theta}')$ com probabilidade $r_{k,k'}$
2. gere $\mathbf{u} \sim q(\mathbf{u})$ com dimensão $n_{k'} - n_k$
3. faça $\boldsymbol{\theta}' = g(\boldsymbol{\theta}, \mathbf{u})$,
4. aceite $(k', \boldsymbol{\theta}')$ com probabilidade $\min(1, A)$ sendo

$$A = \frac{\pi(k', \boldsymbol{\theta}')}{\pi(k, \boldsymbol{\theta})} \times \frac{r_{k',k}}{r_{k,k'}\ q(\mathbf{u})} \left| \frac{\partial g(\boldsymbol{\theta}, \mathbf{u})}{\partial(\boldsymbol{\theta}, \mathbf{u})} \right|.$$

Example. Sejam $Y_1, \ldots, Y_n$ os tempos de vida de componentes eletrônicos sorteados ao acaso e existe incerteza em relação a distribuição dos dados. Sabe-se que,

$$Y_i \sim Exp(\lambda) \text{ (Modelo 1)} \quad \text{ou} \quad Y_i \sim Gamma(\alpha, \beta) \text{ (Modelo 2)},$$

$i = 1, \ldots, n$. O objetivo é estimar qual modelo explica melhor os dados.

Distribuições a priori,

$$\begin{aligned} p(k) &= 1/2 \\ \lambda|k=1 &\sim Gamma(2,1) \\ \alpha|k=2 &\sim Gamma(4,2) \\ \beta|k=2 &\sim Gamma(4,2) \end{aligned}$$

Funções de verossimilhança,

$$p(\mathbf{y}|\lambda, k=1) = \lambda^n e^{-\lambda \sum y_i}$$

$$p(\mathbf{y}|\alpha, \beta, k=2) = \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \prod y_i^{\alpha-1} e^{-\beta \sum y_i}$$

Distribuições condicionais completas,

$$\begin{aligned} p(\lambda|\mathbf{y},\alpha,\beta,k=1) &\propto p(\mathbf{y}|\lambda,k=1)p(\lambda) \\ &\propto \lambda^n e^{-\lambda\sum y_i}\lambda e^{-\lambda} \\ &\propto \lambda^{n+1}e^{-\lambda(1+\sum y_i)} \end{aligned}$$

Portanto,

$$\lambda|\mathbf{y},\alpha,\beta,k=1 \sim Gamma(n+2, 1+\sum y_i)$$

$$\begin{aligned} p(\beta|\mathbf{y},\alpha,\lambda,k=2) &\propto p(\mathbf{y}|\alpha,\beta,k=2)p(\beta) \\ &\propto \beta^{n\alpha}e^{-\beta\sum y_i}\beta^3 e^{-2\beta} \\ &\propto \beta^{n\alpha+3}e^{-\beta(2+\sum y_i)} \end{aligned}$$

Portanto,

$$\beta|\mathbf{y},\alpha,\lambda,k=2 \sim Gamma(n\alpha+4, 2+\sum y_i)$$

$$p(\alpha|\mathbf{y},\beta,\lambda,k=2) \quad \propto \quad \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)}\prod y_i^{\alpha-1}\alpha^3 e^{-2\alpha}$$

A distribuição condicional completa de α não é conhecida então vamos usar o algoritmo de Metropolis-Hastings propondo valores $\alpha' \sim U[\alpha-\epsilon,\alpha+\epsilon]$.

A probabilidade de aceitação é,

$$\min\left\{1, \frac{p(\mathbf{y}|\alpha',\beta,k=2)}{p(\mathbf{y}|\alpha,\beta,k=2)}\frac{p(\alpha'|k=2)}{p(\alpha|k=2)}\right\}$$

já que $q(\alpha'|\alpha) = q(\alpha|\alpha') = 1/2\epsilon$.

```
> mh.alpha <- function(y,n,alpha,beta,eps) {
+ z = runif(1, alpha - eps, alpha + eps)
+ if (z <= 0){
+   acc=0
+ } else {
+   t1=prod(y)
+   num = beta^(n*z) * t1^(z-1)/(gamma(z)^n)
+   den = beta^(n*alpha) * t1^(alpha-1)/(gamma(alpha)^n)
+   num = num * exp(-2*z)*z^3
+   den = den * exp(-2*alpha)*alpha^3
+ }
+ aceita = min(1,num/den)
+ u = runif(1)
+ newalpha = ifelse(u < aceita, z, alpha)
+ return(newalpha)
+ }
```

Suponha que o modelo atual é $Exp(\lambda)$ e queremos propor o modelo $Gamma(\alpha, \beta)$. Um possivel esquema de atualização é o seguine,

1. gere $u \sim Gamma(a, b)$
2. defina $(\alpha, \beta) = g(\lambda, u) = (u, \lambda u)$
3. calcule o Jacobiano,

$$\begin{vmatrix} 0 & 1 \\ u & \lambda \end{vmatrix} = u$$

4. aceite o novo modelo com probabilidade $\min(1, A)$ sendo

$$A = \frac{p(\mathbf{y} \mid \alpha, \beta, k = 2)}{p(\mathbf{y} \mid \lambda, k = 1)} \frac{p(\alpha)p(\beta)}{p(\lambda)} \frac{u}{q(u)}$$

- A transformação no item (2) preserva a média, ou seja $E(Y) = 1/\lambda$ sob o modelo exponencial e $E(Y) = u/\lambda u = 1/\lambda$ sob o modelo gamma.

- Se o modelo atual for $Gamma(\alpha, \beta)$ e propomos o modelo $Exp(\lambda)$ o esquema reverso consiste em fazer

$$(\lambda, u) = g^{-1}(\alpha, \beta) = (\beta/\alpha, \alpha).$$

- A probabilidade de aceitação é simplesmente $\min(1, 1/A)$ substituindo $u = \alpha$,

$$A = \frac{p(\mathbf{y} \mid \lambda, k = 1)}{p(\mathbf{y} \mid \alpha, \beta, k = 2)} \frac{p(\lambda)}{p(\alpha)p(\beta)} \frac{q(\alpha)}{\alpha}.$$

```
> rj.modelo <- function(y,n,lambda,alpha,beta,model,a,b) {
+ if (model == 1) {
+   u = rgamma(1,a,b)
+   alpha1 = u
+   beta1 = lambda*u
+   lambda1 = lambda
+ } else {
+   lambda1 = beta/alpha
+   alpha1 = alpha
+   beta1  = beta
+   u = alpha
+ }
+ t1 = prod(y); t2 = sum(y)
+ num=beta1^(n*alpha1)*t1^(alpha1-1)*exp(-beta1*t2)/(gamma(alpha1)^n)
+ num=num * 2^4 * alpha1^3 * exp(-2*alpha1)/gamma(4)
+ num=num * 2^4 *  beta1^3 * exp(-2* beta1)/gamma(4) * alpha1
+ den=(lambda1^n) * exp(-lambda1*t2)
+ den=den * lambda1 * exp(-lambda1)/gamma(2)
+ den=den * b^a * u^(a-1) * exp(-b*u)/gamma(a)
+ u = runif(1,0,1)
```

```
+ if (model == 1) {
+   aceita = min(1,num/den)
+   if (u < aceita) {
+     model = 2
+     alpha = alpha1
+     beta  = beta1
+   }
+ } else {
+   aceita = min(1,den/num)
+   if (u < aceita) {
+     model  = 1
+     lambda = lambda1
+   }
+ }
+ if (model == 1) return(list(model=model, lambda=lambda))
+ else return(list(model=model, alpha=alpha, beta=beta))
+ }
```

```
> rjmcmc <- function(niter,nburn,y,n,a,b,eps=0.25){
+ x  = matrix(0, nrow=niter+1, ncol=3)
+ x1 = matrix(0, nrow=niter-nburn, ncol=3)
+ nv = nv1= array(0,2)
+ mod= array(0,niter-nburn)
+ x[1,] = c(1,1,1)
+ model  = 1
+ t1 = prod(y)
+ t2 = sum(y)
+ for (i in 1:niter){
+   if (model==1){
+     x[nv[1]+1,1] = rgamma(1, n + 2, t2 + 1)
+   } else {
+     x[nv[2]+1,3] = rgamma(1, 4 + n*x[nv[2],2], t2 + 2)
+     x[nv[2]+1,2] = mh.alpha(y,n,x[nv[2],2],x[nv[2]+1,3],eps)
+   }
+   new = rj.modelo(y,n,x[nv[1]+1,1],x[nv[2]+1,2],x[nv[2]+1,3],model,a,
+   model = new$model
+   if (i>nburn) mod[i-nburn]= model
+   if (model == 1) {
```

```
+     x[nv[1]+1,1] = new$lambda
+     nv[1] = nv[1] + 1
+     if (i > nburn) {
+       x1[nv1[1]+1,1] = new$lambda
+       nv1[1] = nv1[1] + 1
+     }
+   } else {
+     x[nv[2]+1,2] = new$alpha
+     x[nv[2]+1,3] = new$beta
+     nv[2] = nv[2] + 1
+     if (i > nburn) {
+       x1[nv1[2]+1,2] = new$alpha
+       x1[nv1[2]+1,3] = new$beta
+       nv1[2] = nv1[2] + 1
+     }
+   }
+ }
+ cat("Probabilidades a posteriori dos modelos","\n")
+ print(nv1/(niter-nburn))
+ cat("Medias a posteriori dos parametros","\n")
+ somas  = apply(x1,2,sum)
```

```
+ print(somas/c(nv1[1],nv1[2],nv1[2]))
+ return(list(x=x,nv=nv, x1=x1, nv1=nv1, model=mod))
+ }
```

Example. Testando o algoritmo com saltos reversiveis para o exemplo anterior. Os dados foram simulados como $Y_1, \ldots, Y_n \sim Exp(3)$, sendo $n = 10$.

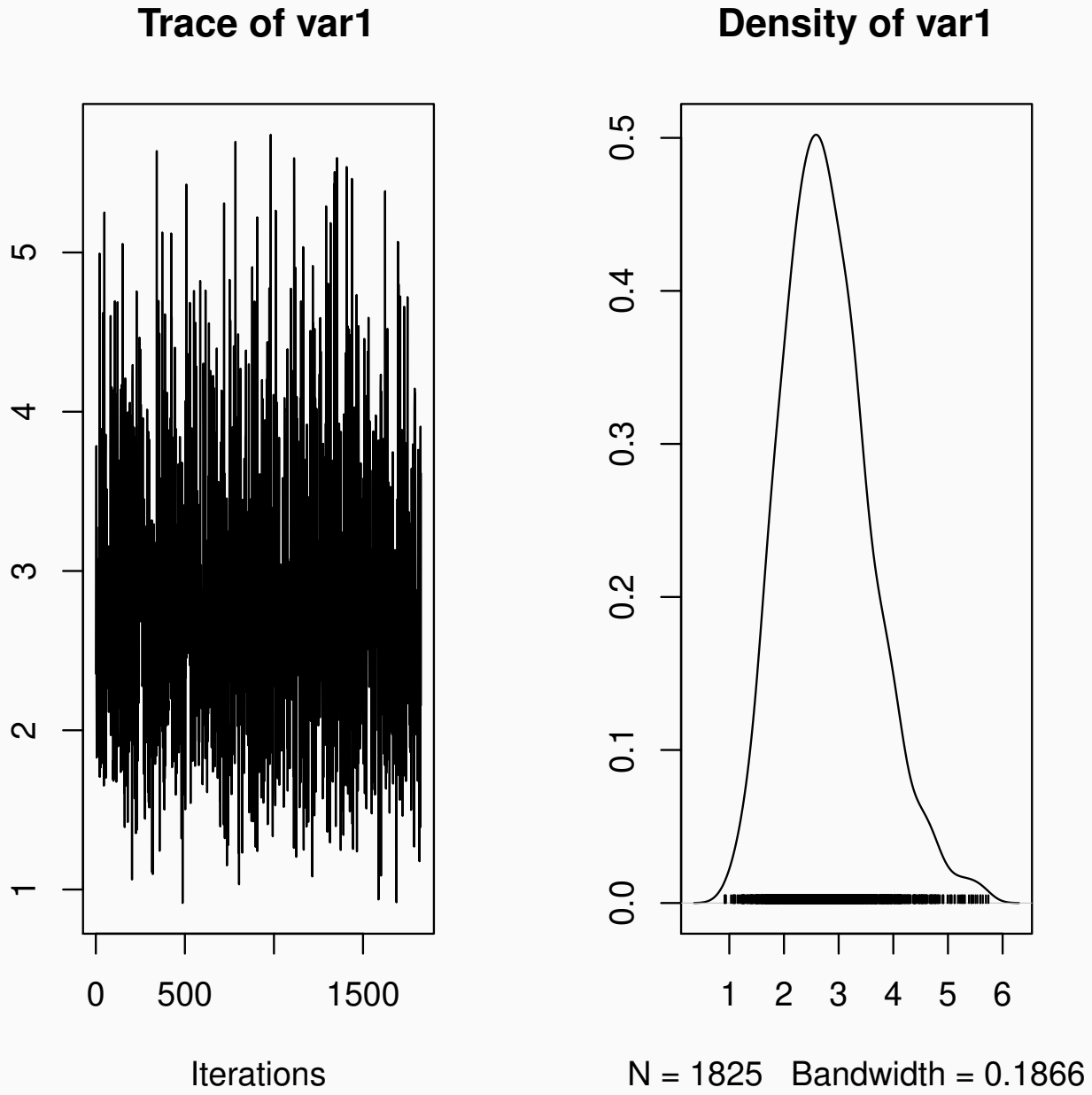
Total de 5000 iterações com 2500 de aquecimento. Distribuição proposta: $u \sim Gamma(1, 1)$.

```
Probabilidades a posteriori dos modelos
[1] 0.73 0.27
Medias a posteriori dos parametros
[1] 2.7955622 0.8739996 2.4279896
```

O modelo exponencial tem probabilidade a posteriori bem maior que o modelo Gamma.

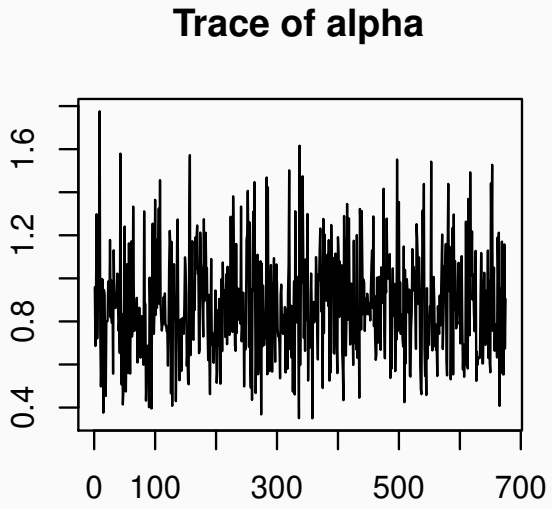
Análise sob o modelo Exponencial.

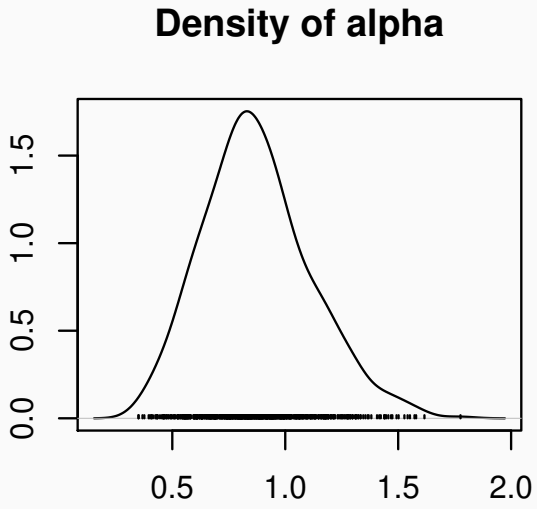
```
Iterations = 1:1825
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1825

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
       2.79556        0.81005         0.01896        0.01896

2. Quantiles for each variable:

 2.5%   25%   50%   75% 97.5%
1.455 2.218 2.718 3.277 4.636
```

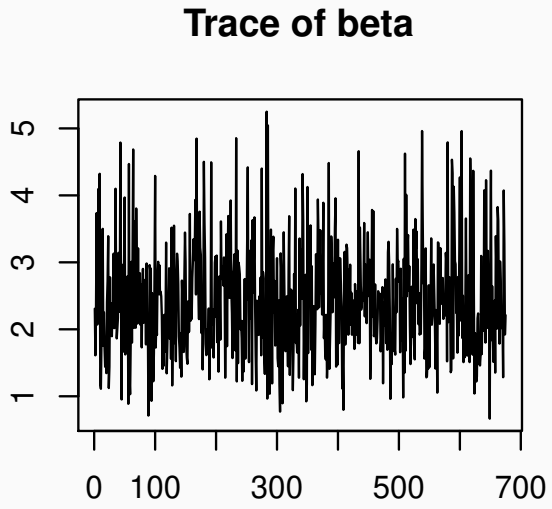

Trace of var1
Density of var1
1
2
3
4
5
0
500
1500
Iterations
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
N = 1825 Bandwidth = 0.1866

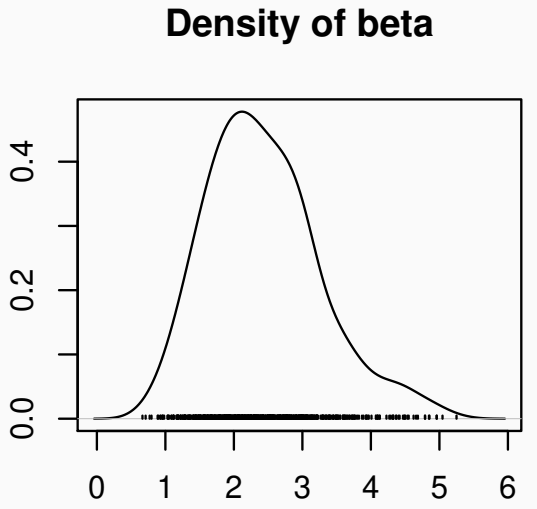
Análise sob o modelo Gamma.

```
Iterations = 1:675
Thinning interval = 1
Number of chains = 1
Sample size per chain = 675

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

       Mean     SD Naive SE Time-series SE
alpha 0.874 0.2387 0.009187       0.009187
beta  2.428 0.8303 0.031959       0.031959

2. Quantiles for each variable:

        2.5%    25%    50%   75% 97.5%
alpha 0.4601 0.7081 0.8493 1.013 1.417
beta  1.0898 1.8367 2.3270 2.932 4.426
```


Trace of alpha
0.4
0.8
1.2
1.6
0
100
300
500
700
Iterations


Density of alpha
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
N = 675 Bandwidth = 0.06548


Trace of beta
1
2
3
4
5
0
100
300
500
700
Iterations


Density of beta
0.0
0.2
0.4
0
1
2
3
4
5
6
N = 675 Bandwidth = 0.2353

Estimando os tempos médios de vida ($E(Y)$) sob o modelo 1 ($1/\lambda$) e sob o modelo 2 (α/β).

```
Iterations = 1:1825
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1825

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
      0.390317       0.123066       0.002881       0.002881

2. Quantiles for each variable:

  2.5%    25%    50%    75%  97.5%
0.2157 0.3052 0.3679 0.4510 0.6871
```

```
Iterations = 1:675
Thinning interval = 1
Number of chains = 1
Sample size per chain = 675

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
      0.385603       0.122357       0.004710       0.004981

2. Quantiles for each variable:

  2.5%    25%    50%    75%  97.5%
0.2098 0.2987 0.3610 0.4494 0.6609
```

Outros exemplos,

Example. Sejam $Y_1, \ldots, Y_n$ independentes tais que,

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(p_i) \\ p_i &= F(\alpha + \beta x_i),\ i = 1, \ldots, n. \end{aligned}$$

Poderiamos considerar diferentes funções de ligação $F(\cdot)$: logito, probito, Gumbel, etc.

Example. Sejam $Y_1, \ldots, Y_n$ independentes tais que,

$$Y_i \sim N(\mu_i, \sigma^2),\ i = 1, \ldots, n$$

sendo,

$$\mu_i = \begin{cases} \beta_0, \text{ (Modelo 0) ou,} \\ \beta_0 + \beta_1 x_i, \text{ (Modelo 1) ou,} \\ \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \text{ (Modelo 2)} \end{cases}$$

para uma covariável x.

Example. Sejam $Y_1, \ldots, Y_n$ independentes tais que,

$$Y_i \sim N(\mu_i, \sigma^2),\ i = 1, \ldots, n$$

sendo,

$$\mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Quais covariáveis devem entrar no modelo?

Para $k = 3$ e assumindo que $\beta_0 \neq 0$ temos $2^3 = 8$ possiveis modelos: M_0, M_1, M_2, M_3, M_{12}, M_{13}, M_{23} e M_{123}, sendo M_0 o modelo sem covariáveis e M_{123} o modelo completo.

- Remover uma covariável consiste em fazer seu coeficiente igual a zero.
- incluir uma covariável consiste em gerar um novo coeficiente. Por exemplo, podemos propor um salto de M_1 para M_{12} gerando $u \sim N(0, \gamma^2)$ e fazendo a transformação,

$$(\beta_1, \beta_2) = (\beta_1, u)$$

- Neste caso a função $g(\cdot)$ é a identidade e o Jacobiano é igual a 1.

Por exemplo, podemos assumir que em cada iteração cada covariável é escolhida ao acaso para entrar ou sair do modelo.

Assim, $r_{k,k'} = r_{k',k} = 1/3$ na probabilidade de aceitação.

Temos então que,

$$\alpha((\boldsymbol{\beta}, k), (\boldsymbol{\beta}', k')) = \min \left\{ 1, \frac{\pi(\boldsymbol{\beta}', k')}{\pi(\boldsymbol{\beta}, k)} \frac{1}{f_N(u|0, \gamma^2)} \right\}$$

Approximating the marginal likelihood

Denoting the competing models by $M_1, M_2, \ldots,$ the marginal likelihood of model M_i is,

$$p(\mathbf{y}|M_i) = \int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i) p(\boldsymbol{\theta}_i|M_i) d\boldsymbol{\theta}_i \qquad (1)$$

Computation of the marginal likelihood requires a proper prior and the integral (1) is in general difficult to calculate.

Approximating (1) from a MCMC output is not trivial because we integrate with respect to the prior.

Chib, S. and E. Jeliazkov (2001). Marginal likelihood from the Metropolis-Hastings output. Journal of the American Statistical Association 96, 270–281.

Chib, S. and I. Jeliazvok (2005). Accept-reject Metropolis- Hastings sampling and marginal likelihood estimation. Statistica Neerlandica 59, 30–34.

Rewrite the Bayes theorem for $\boldsymbol{\theta}$ as,

$$p(\mathbf{y}|M_i) = \frac{p(\mathbf{y}|\boldsymbol{\theta}_i^*, M_i)p(\boldsymbol{\theta}_i^*|M_i)}{\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, M_i)}$$

for a particular value $\boldsymbol{\theta}^*$.

- The denominator os this expression is unknown.
- If we can find an estimate of the posterior ordinate $\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, M_i)$ then the marginal likelihood can be calculated.
- For estimation efficiency we take the point $\boldsymbol{\theta}_i^*$ as the posterior mode given model M_i.

In a Random Walk Metropolis, suppose that the current value is $\boldsymbol{\theta}$ and a new value $\boldsymbol{\phi}$ is proposed. Then, by reversibility

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) q(\boldsymbol{\phi}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\mathbf{y}) = \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) \pi(\boldsymbol{\theta}^*|\mathbf{y}).$$

Integrating both sides with respect to $\boldsymbol{\theta}$ we obtain that,

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{\int \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}}{\int \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) d\boldsymbol{\theta}}.$$

This is estimated as,

$$\hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}) \approx \frac{N^{-1}\sum_{g=1}^{N}\alpha(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(g)})}{J^{-1}\sum_{j=1}^{J}\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{(j)})}$$

where

$\{\boldsymbol{\theta}^{(g)}\}_{g=1}^{N}$ are the sampled values from the posterior distribution and,

$\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{J}$ are draws from $q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$.