Temas en Estadística Contemporánea: Práctica N.1

Huertas, Anthony* and **Saenz, Miguel****

*Cod: 20173728
**Cod: 19921255

Compiled September 19, 2018

Maestría en Estadística, Escuela de Posgrado, Pontificia Universidad Católica del Perú, Lima, Perú

PROBLEMA 1

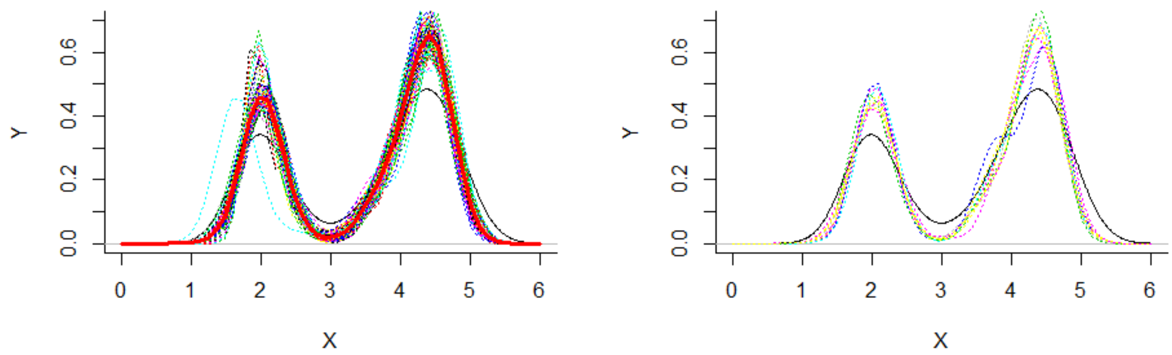
Listing 1. Modificación en funciones gibbs() y ex() de ex1a.R: Con la finalidad de eliminar las primeras distribuciones a posteriori que se encuentran correlacionadas con las demás y visualización de distribuciones.

```
gibbs <- function(n.iter=100,posteriori.draw = T, posteriori.mean = T,
                  draws = c(1:100), seed = 1963){
        set.seed(seed)
        .........
        ## now the Gibbs sampler
        for(iter in 1:n.iter){
                .........
                if(posteriori.draw==T){
                        if(is.element(iter,draws)){
                                lines(xgrid,f,col=iter,lty=3)
                        }
                }
                ........
        }
        .......
        if(posteriori.mean == T){
                lines(xgrid,fbar,lwd=3,col=2)
        }
        .......
}

ex <- function(n.it=100, post.d = T,post.m = T,d = c(1:100), set.s=1963){
        mcmc <- gibbs(n.iter=n.it,posteriori.draw = post.d, posteriori.mean = post.m,
        draws = d, seed = set.s)
        ........
}
```

Para ex1a.R, se visualizará (i) la densidad estimada media a posteriori y (i) 10 muestras de gráficas a posteriori para $G \sim p(G \mid data)$.

Listing 2. Plot de gráficas.

```
##### PREGUNTA 1
####(i)
ex()

####(ii)
it = sample(50:300,10,replace = T)
#gibbs.H(n.iter = 1000,posteriori.mean = F,posteriori.draw = T,draws = it)
ex(n.it = 300, post.d = T,post.m = F,d = it, set.s = 1963)
```



(a) Densidad estimada media a posteriori .

(b) 10 muestras de distribución a posteriori $G \sim p(G \mid data)$.

En (a) se realizó un muestreo de Gibbs con 100 iteraciones, obteniéndose la densidad estimada media a posteriori (curva gruesa roja) de 100 distribuciones. En (b), con el objetivo de tomar una muestra de 10 distribuciones a posteriori, se optó por desarrollar un muestreo de Gibbs con 1000 iteraciones eliminando los 100 primeros resultados, del conjunto restante se extrajo la muestra de 10 distribuciones.

PROBLEMA 2

Para ex1a, el modelo genérico es

$$y_i \sim F(y_i) = \int N(y_i; \theta_i, \sigma) dp(G)$$

El modelo jerárquico (hierarchical model) viene determinado por

$$\begin{aligned} y_i \mid \theta_i &\sim N(\theta_i, \sigma^2) \\ \theta_i &\sim G \\ G &\sim DP(1, G_0) \text{ con } G_0 \sim N(0,4) \end{aligned}$$

y un hiperapriori

$$\frac{1}{\sigma^2} \sim Gamma(1,1)$$

PROBLEMA 3

Para ex1a.R, la distribucion a posteriori condicional generada por la función sample.th() es de la forma siguiente

$$\theta_i \sim p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n, \sigma^2, y)$$

Primero tenemos un muestreo en los indicadores siguientes

$$s_i = j \sim p(s_i = j \mid \ldots, y_i) = \begin{cases} n_j^- N(y_i, \theta_j^*, \sigma^2) & 1, \ldots, i-1, i+1, \ldots, k \\ N(y_i, 0, 4+\sigma^2) & k+1 \end{cases}$$

donde se tienen previamente únicos valores iniciales $\{\theta_1^* \ldots, \theta_{10}^*\}$. Si $s_i = 11$, entonces

$$\theta_{11}^* \sim N\left(\left(\frac{1}{4} + \frac{1}{\sigma^2}\right)^{-1}\left(\frac{1}{\sigma^2} y_i\right), \left(\frac{1}{4} + \frac{1}{\sigma^2}\right)^{-1}\right) \quad \text{(S1)}$$

PROBLEMA 4

Para ex1a.R, la distribucion a posteriori condicional generada por la función sample.sig() es de la forma siguiente

$$\begin{aligned} \sigma^2 &\sim p(\sigma^2 \mid \theta_1, \ldots, \theta_n, y) \\ \frac{1}{\sigma^2} &\sim Gamma(1 + 0.5 * (n), 1 + 0.5 \sum_{i=1}^{n} (y_i - \theta_i)^2) \end{aligned}$$

donde $n = 272$ para los datos de la muestra.

PROBLEMA 5

Para ex1a.R, la distribucion a posteriori condicional generada por la función sample.ths() es de la forma siguiente

$$\begin{aligned}\theta_{s_j} &\sim p(\theta_{s_j} \mid \ldots) = N(\theta_{s_j}, 0, 4) * N(\bar{y}_j, \theta_{s_j}, \sigma^2/n_j) \\ \theta_{s_j} &\sim N\left(\left(\frac{1}{4} + \frac{n_j}{\sigma^2}\right)^{-1}\left(\frac{n_j}{\sigma^2}\bar{y}_j\right), \left(\frac{1}{4} + \frac{n_j}{\sigma^2}\right)^{-1}\right)\end{aligned}$$

donde $A_j = \{i : s_i = j\}$ y $n_j = \#\{i : s_i = j\}$, entonces $\bar{y} = (1/n_j)\sum_{i \in A_j} y_i$.

PROBLEMA 6

La función sample.th() muestrea los índices de un vector inicial $\{\theta_i^*\}_{i=1}^{k}$ y genera un nuevo elemento θ_{k+1} si al muestrear el índice, este toma el valor k+1. Sin embargo, sample.th() usa solo un dato (y_i) en caso se necesita determinar un nuevo valor para θ_{new}, a diferencia de la función sample.ths() que usa el promedio de los datos que cumplan ciertas condiciones y ello impacta en las varianzas de las distribuciones a posteriori siendo estas como siguen

i) Para sample.th(), tenemos una varianza:

$$\left(\frac{1}{4} + \frac{1}{\sigma^2}\right)^{-1} \tag{S2}$$

ii) Para sample.ths(), tenemos una varianza:

$$\left(\frac{1}{4} + \frac{n_j}{\sigma^2}\right)^{-1} \tag{S3}$$

Observamos que la varianza en la segunda es más baja, sabiéndose $n_j > 0$ (siendo especificado a detalle en Problema 3 y Problema 5), mejorando esto la simulación de las distribuciones a posteriori. Cabe indicar que la funcion sample.sig trabaja en conjunto con ambas, pero en la comparación entre sample.ths() y sample.th() es donde se manifiesta el efecto de mejora en la simulación.

PROBLEMA 7

Para ex1b, se realiza un muestreo de Gibbs por bloques:

$$\begin{aligned}G &\sim DP_H(\mathbf{a}, \mathbf{b}) \\ G &= \sum_{h}^{H} w_h \delta(\theta_h)\end{aligned}$$

Por lo que del modelo para la estimación determinado por

$$y_i \mid G, \sigma^2 \sim \int N(y_i, \theta_i, \sigma^2) dG(\theta_i)$$

tomariamos su equivalente al modelo

$$y_i \mid r_i = h \sim N(m_h, \sigma^2)$$

con a prioris

$$\begin{aligned}p(r_i = h) &= w_h \\ \frac{1}{\sigma^2} &\sim Gamma(1, 1)\end{aligned}$$

PROBLEMA 8

Las distribuciones a posteriori condicionales generadas en ex1b son

ii Para la función sample.r()

$$r_i \sim p(r_i \mid w_i, \sigma^2, m_h, y) = N(y_i, m_h, \sigma^2) * w_h$$

ii Para la función sample.mh()

$$m_j \sim p(m_j \mid r, \sigma^2, y) = \begin{cases} N\left(\left(\frac{1}{4} + \frac{n_j}{\sigma^2}\right)^{-1}\left(\frac{n_j}{\sigma^2}\bar{y}_j\right), \left(\frac{1}{4} + \frac{n_j}{\sigma^2}\right)^{-1}\right) & \text{si } r_i = j \text{ para algun } i \\ N(0, 4) & \text{otro caso.}\end{cases}$$

donde $A_j = \{i : r_i = j\}$ y $n_j = \#\{i : r_i = j\}$, entonces $\bar{y} = (1/n_j)\sum_{i \in A_j} y_i$.

iii Para la función sample.vh()

$$v_j \sim p(v_j \mid r) = Beta(1 + |A_j|, 1 + |B_j|)$$

donde $A_j = \{i : r_i = j\}$ y $B_j = \{i : r_i > j\}$.
Luego de ello se determina: $w_j = v_j \prod_{l<h}(1 - v_l)$.

iv Para la función sample.sig()

$$\sigma^2 \sim p(\sigma^2 \mid \theta_1, \ldots, \theta_n, y)$$
$$\frac{1}{\sigma^2} \sim Gamma(1 + 0.5 * (n), 1 + 0.5 \sum_{i=1}^{n} (y_i - \theta_i)^2)$$

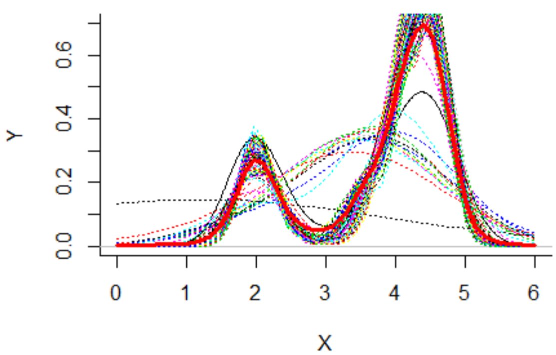
1. PROBLEMA 9

Se realizaron las mismas modificaciones que en el algoritmo gibbs de la pregunta 1, con la misma finalidad.

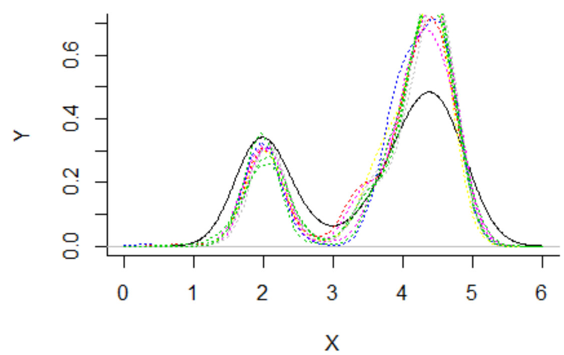
Listing 3. Plot de gráficas.

```
## PREGUNTA 9
# Item (i)
gibbs.H()

# Item (ii): 10 muestras de distribucion posteriori para G
it = sample(100:1000,10,replace = T)
gibbs.H(n.iter = 1000,posteriori.mean = F,posteriori.draw = T,draws = it)
```

Para ex1b.R, se visualizará (i) la densidad estimada media a posteriori y (i) 10 muestras de gráficas a posteriori para $G \sim p(G \mid data)$.



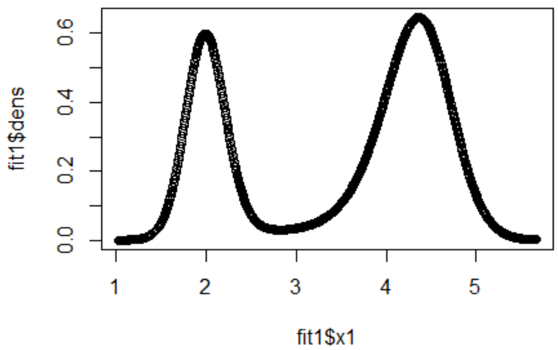
(c) Densidad estimada media a posteriori .



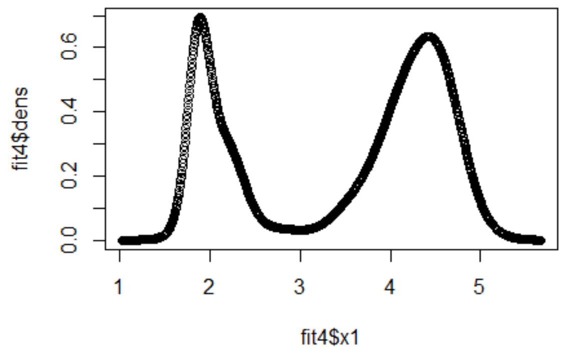
(d) 10 muestras de distribución a posteriori $G \sim p(G \mid data)$.

2. PROBLEMA 10

Para ex1c.R, se visualizará la densidad estimada para los dos modelos con distribuciones a priori distintas.



(e) Densidad estimada del Modelo 1.



(f) Densidad estimada del Modelo 2.

Se observa que en ambas se estiman densidades con picos muy altos a diferencias de los determinados en ex1a y ex1b, por lo que sería conveniente comparar con el uso de otras distribuciones a priori y más iteraciones.