Trabajo práctico 3

IECD323: Estadística Bayesiana

Bastián Barraza Morales

December 15, 2022

Introducción

Al realizar un análisis estadístico desde un enfoque Bayesiano, se busca incorporar las creencias y la información a priori que uno tiene acerca de un evento aleatorio en una muestra (aleatoria), y de este modo, producir una distribución a posteriori con la información actualizada. El cálculo de una distribución a posteriori no siempre es fácil de obtener analíticamente, y en muchos casos es imposible, por lo que muchos autores han buscado métodos alternativos (principalmente computacionales) para definir o aproximar esta distribución.

El muestreo Gibbs (Gibbs sampling) es un método MCMC (Markov Chain Monte Carlo), el cual permite aproximar la distribución a posteriori conjunta $\pi(\theta_1, \theta_2, ..., \theta_n|\overrightarrow{X})$, cuando cada una de las distribuciones condicionales $\pi(\theta_i|\theta_j, ..., \theta_n|\overrightarrow{X})$ con $i \neq j$ son especificadas o simples de encontrar. Esto se realiza asumiendo que existe una distribución estacionaria, la cual se aproximará a la distribucion a posterior de los parámetros dada una muestra aleatoria .

Este documento tiene como objetivo aproximar la distribución a posteriori mediante el método de Gibbs Sampling con los datos de una muestra aleatoria simulada y la distribución a priori de los parámetros.

Para el presente informe se utilizará el software R para realizar las simulaciones, gráficos y cálculos necesarios.

1 Ejercicio 1:

Simular $x_1, x_2, ..., x_{50}|(\theta = 5, \sigma^2 = \phi^{-1} = 2) \sim N(5, 2)$.

Respuesta:
Al realizar la simulación de 50 valores con los valores dados (Apéndice A.1) se obtienen las siguientes estadísticas descriptivas de X.

Mínimo	Cuartil 1	Mediana	Media	Cuartil 4	Máximo	Varianza
2.75	4	4.887	5.032	5.806	7.956	1.68

Se observan que el cuartil 1 y cuartil 4 están casia la misma distancia de la media al igual que el valor mínimo y máximo. Por otro lado, en la figura 1 se observa el histograma de la distribución de probabilidad de X en donde la media está dibujada con una recta azul (Apéndice B.1).

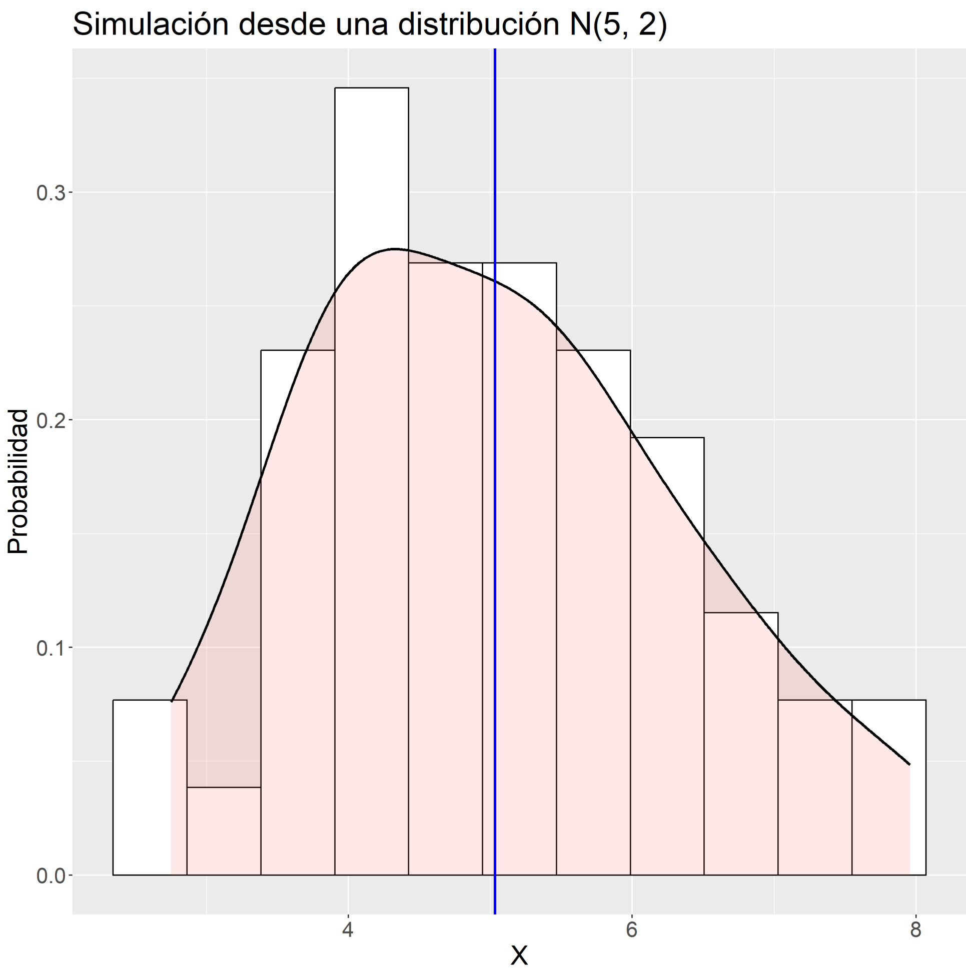

Figure 1: Distribución de probabilidad de X.

2 Ejercicio 2

Considerar $\pi(\theta) \sim N(0, \tau_0^2)$ y $\pi(\phi) \sim Gamma(\frac{a_0}{2}, \frac{b_0}{2})$, donde $a_0 = n_0 = 10$ y $b_0 = n_0\sigma_0^2 = 25$

Respuesta:

Dado que $X \sim N(5, 2)$ y con las distribuciones a priori de θ y ϕ se puede especificar completamente $\pi(\theta, \phi|X)$ mediante una constante de normalización de la expresión:

$$\pi(\theta, \phi|X) \propto ln(\theta, \phi|X)\pi(\theta)\pi(\phi)$$

Sin embargo, esta constante de normalización resulta imposible de calcular analíticamente. Ahora bien, si se puede obtener las densidades condicionales a posteriori $\pi(\theta|\phi, X)$ y $\pi(\phi|\theta, X)$. En efecto:

$$\pi(\theta|\phi, X) \sim N\left(\frac{\tau_0^{-2}\mu_0 + n\phi\bar{x}_n}{\tau_0^{-2} + n\phi}, (\tau_0^{-2} + n\phi)^{-1}\right)$$

$$\pi(\phi|\theta, X) \sim Gamma\left(\frac{n + n_0}{2}, (b_0 + (n-1)S^2 + n(\theta - \bar{X}_n)^2)^{-2}\right)$$

Con las densidades condicionales a posteriori se pueden aplicar el método de Gibbs sampling para aproximar la distribución a posteriori conjunta.

3 Ejercicio 3

Simule 1000 iteraciones usando 'Gibbs Sampling'.

Respuesta:

Para realizar el método Gibbs Sampling se consideran los valores iniciales de $\theta^{(0)}$ y $\phi^{(0)}$ desde $\pi(\theta|\phi^{(0)}, X)$ y $\pi(\phi|\theta^{(0)}, X)$ respectivamente.

Luego, a partir de estas cantidades iniciales, se simulan los valores de $\theta^{(i)}$ y $\phi^{(i)}$ con $i = 1, 2, ..., n$ en donde se considera la sucesión $\{((\theta^n, \phi^n)\})$ como una cadena de Markov (Apéndice A.2).

En la figura 2 y 3 se observan los valores de θ y ϕ respectivamente según el número de iteraciones (Apéndice B.2 y B.3).

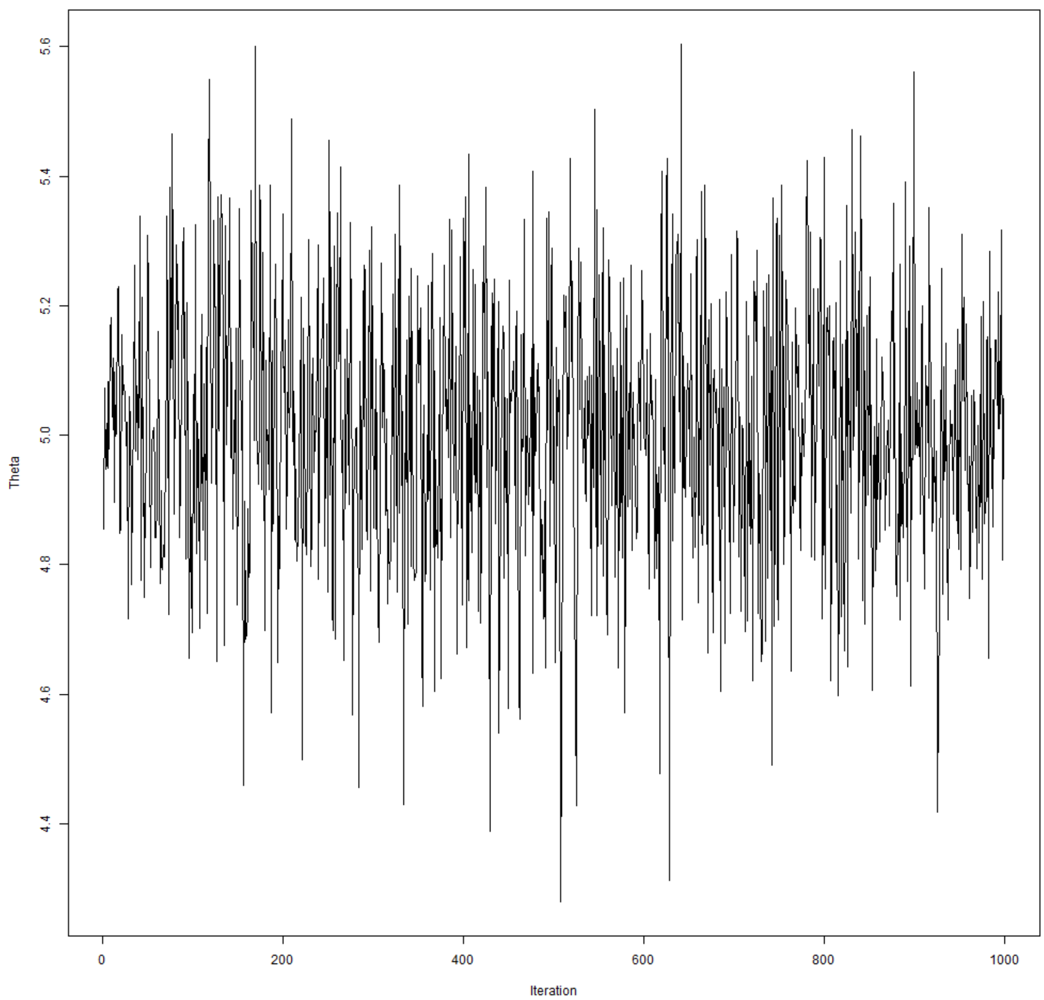

Figure 2: Valores de θ según su número de iteración.

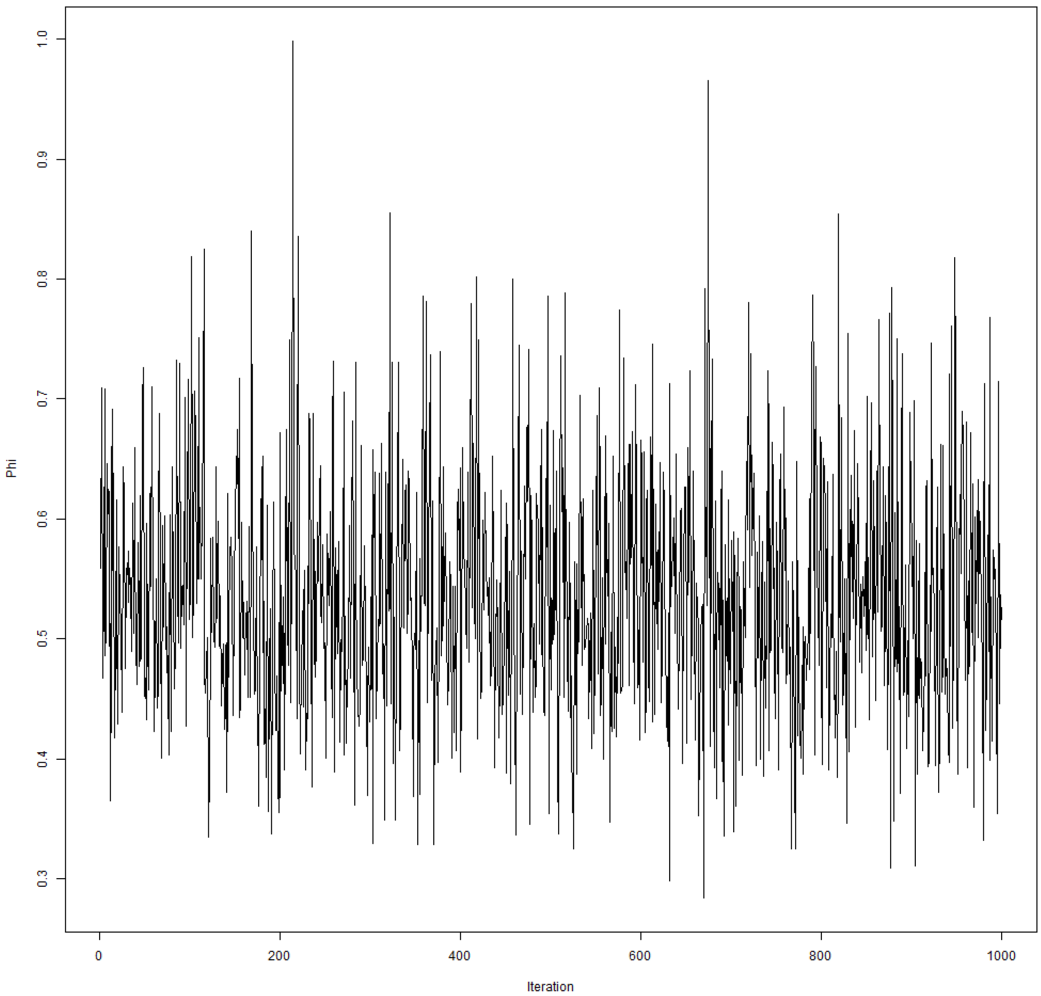

Figure 3: Valores de ϕ según su número de iteración.

4 Ejercicio 4

Removiendo las 500 primeras iteraciones calcule las medias y las vairanzas muestrales de $\theta|\overrightarrow{x}$ y $\sigma^2|\overrightarrow{x}$.

Respuesta:
Luego de eliminar las 500 primeras iteraciones (Apéndice A.3), se obtienen los siguientes valores para la media y la varianza de θ y ϕ (Apéndice A.3).

Variable	Media	Varianza
θ	5.005	0.039
ϕ	0.536	0.011

Los valores esperados de θ y ϕ son 5 y 0.5 respectivamente , dado que $X \sim N(\theta = 5, \phi^{-1} = 2)$. En nuestro caso, notamos que la media de θ y ϕ están muy cercanas a las cantidades esperadas. Además, las variables no poseen mucha variabilidad.

5 Ejercicio 5

Construir los histogramas para $\theta| + \overrightarrow{x}$ y $\sigma^2| + \overrightarrow{x}$ usando los últimos 500 datos.

Respuesta:

En la figura 4 y 5 se tienen los histogramas de θ y ϕ respectivamente (Apéndice B.4 y B.5), en donde se marca con una recta azul la media de la variable. Notamos que θ se acerca bastante a una normal, mientras que ϕ tiene forma de distribución normal con una leve asimetría positiva.

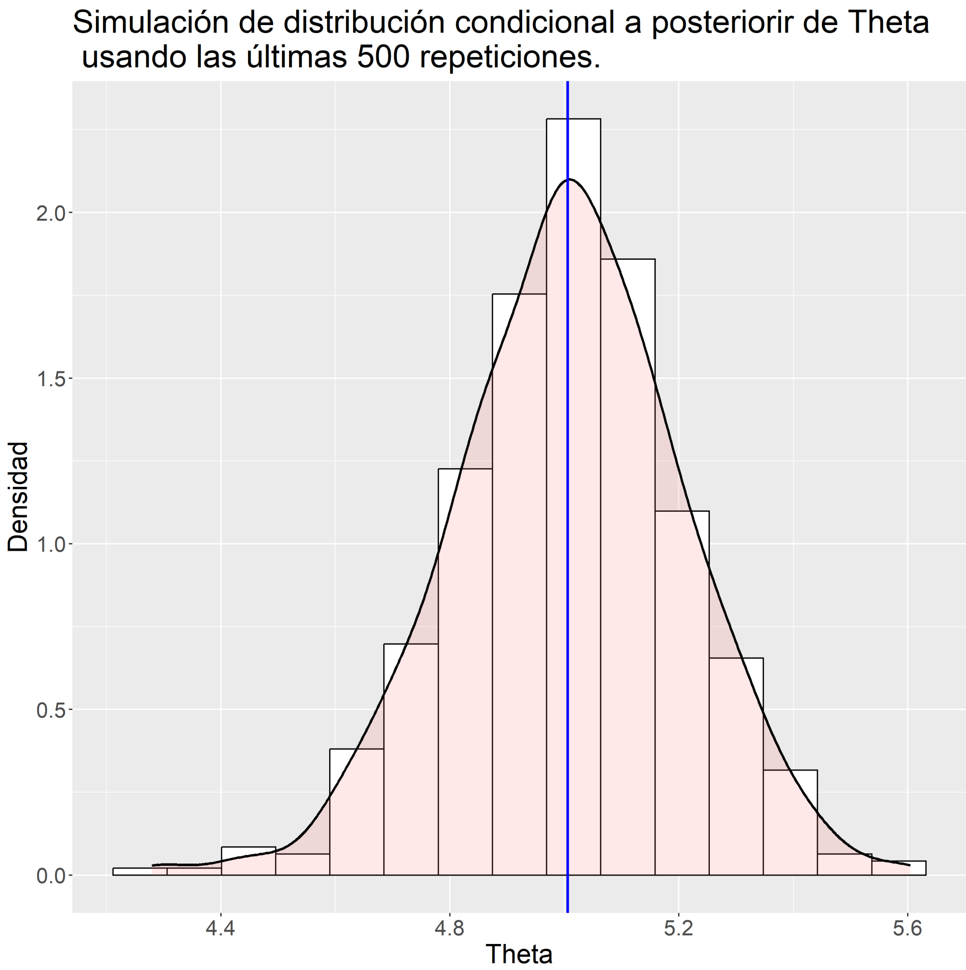

Figure 4: Distribución a posteriori de Theta.

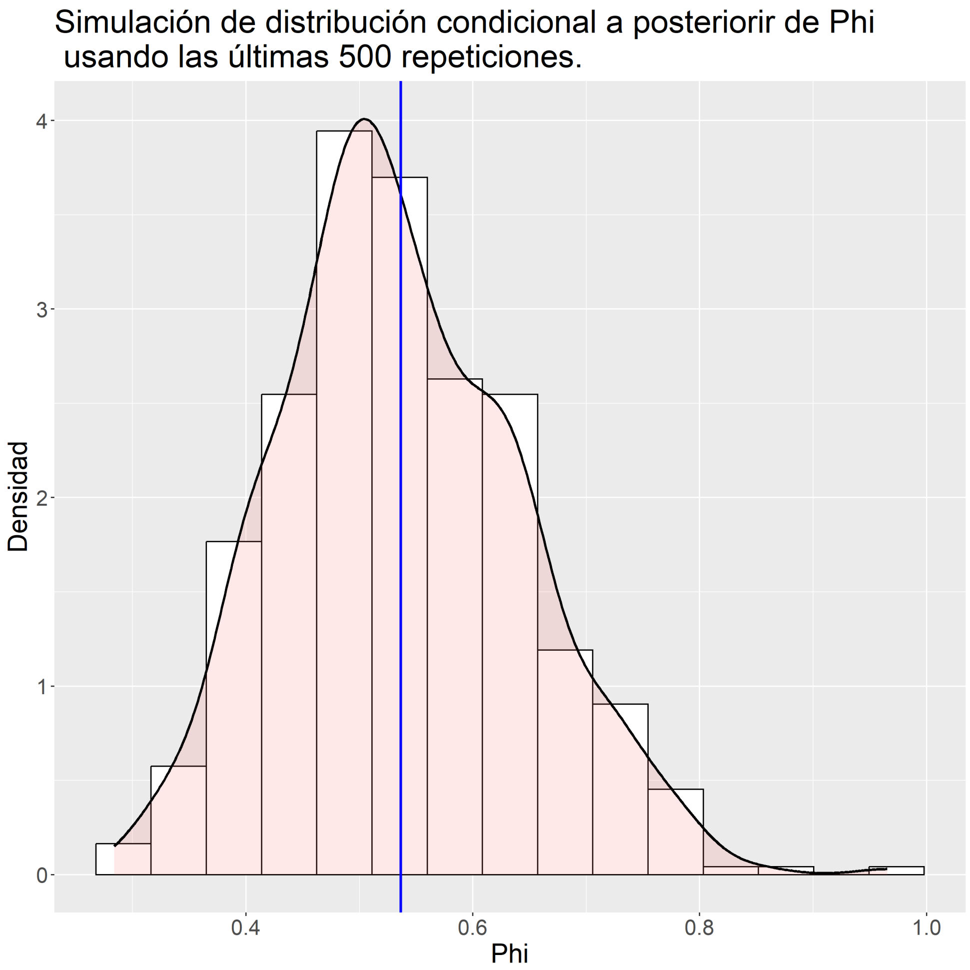

Figure 5: Distribución a posteriori de Phi.

6 Ejercicio 6

Generar el histograma de la distriución conjunta $\pi(\theta, \phi|\overrightarrow{x})$ (3D). Compare esto con un histograma generado por el método aceptación/rechazo.

Respuesta:

En la figura 8 se realiza el histograma de la distirbución conjunta (Apéndice B.6), notamos una gran concentración de datos cercano a la media de θ (graficado en el eje x), específicamente entre 4.80 y 5.28. Asimismo, los valores de ϕ (graficados en el eje Y) se concentran cercano a 0.5. El par con mayor cantidad de datos se encuentra en los intérvalos $\theta = [4.93, 4.98]$ y $\phi = [0.5, 0.53]$. Estos resultados son los esperados luego de visualizar el histograma de θ y ϕ calculadas con el método de Gibbs sampling.

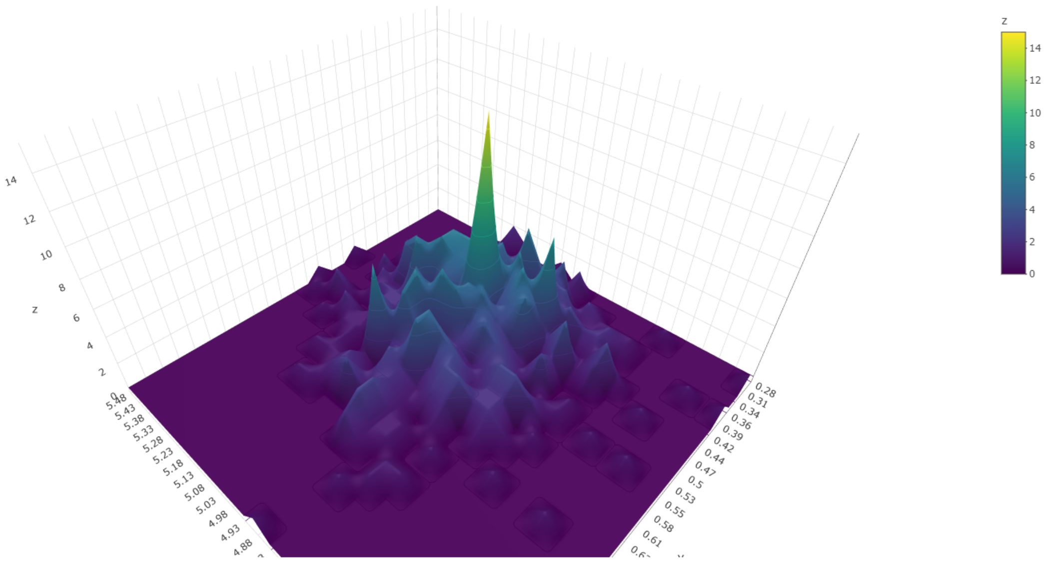

Figure 6: Distribución conjunta mediante el método Gibbs sampling

Luego, en la figura 9 se realiza el histograma mediante el método aceptación/rechazo (Apéndice A.4 y B.7). Notamos que θ (representado en el eje x) tiene valores en todo su dominio, concentrando un gran cantidad cercano a su media. Por otro lado, ϕ (representado en el eje Y) no posee datos en algunos intérvalos, como [0.51, 0.58] y la frecuencia de datos es mayor en valores mayor a 0.61, lo cual indica que el método de aceptación/rechazo genera mas cantidad de datos en valores mas grandes que el método Gibbs sampling.

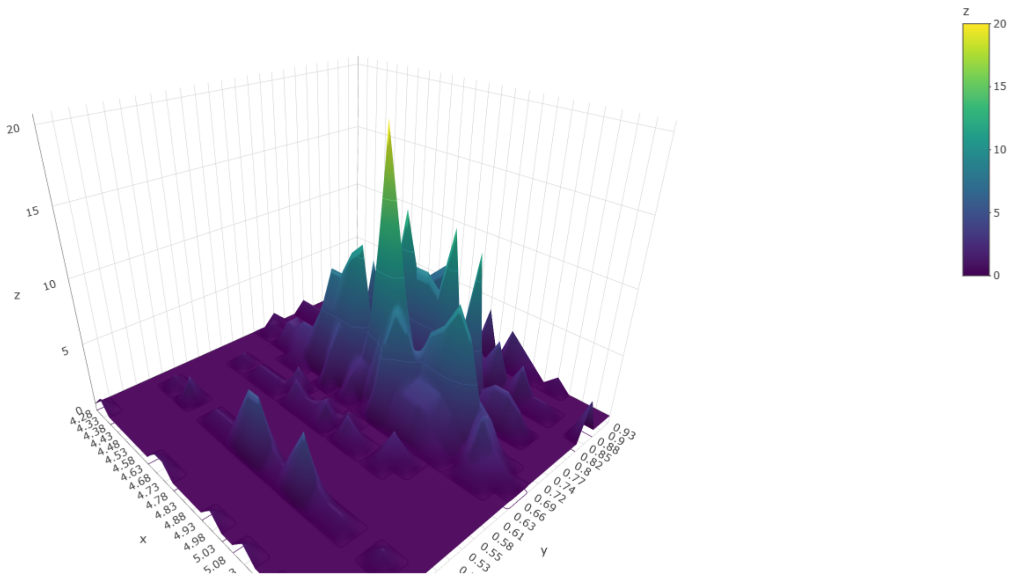

Figure 7: Distribución conjunta mediante el método aceptación/rechazo.

7 Apéndice A

Simulación de distribuciones

Apéndice A.1

```
set.seed(2)
n=50
theta = 5
phi1 = 2
SIM50 = data.frame('simulation' = rnorm(n,
                                         theta,
                                         sd = sqrt(phi1)))
```

Apéndice A.2

```
# numero de muestras
S <- 1000

# matriz para almacenar las muestras
PHI <- matrix(data = NA, nrow = S, ncol = 2)
# ALGORITMO (muestreador de Gibbs)
# 1. inicializar la cadena
#    valor inicial: simular de la previa
#    solo es necesario alguno de los valores
set.seed(1)
phi <- c( rnorm(1, mu0, sd = sqrt(t20)), rgamma(1, nu0/2, s20/2) )
PHI[1,] <- phi

# 2. simular iterativamente de las distribuciones condicionales completas
set.seed(1)
for(s in 2:S) {
        # 2.1 actualizar el valor de theta
        t2n    <- 1/( (t20**(-1)) + n*phi[2] )
        mun    <- (n*phi[2]*mean(SIM50$simulation)) / (t20**(-1) + n*phi[
        phi[1] <- rnorm( 1, mun, sd = sqrt(t2n) )
        # 2.2 actualizar el valor de sigma^2
        nun    <- nu0+n
        s2n    <- s20 + (n-1)*var(SIM50$simulation) + n*(phi[1] - mean(SIM
        phi[2] <- rgamma(1, nun/2, s2n/2)
        # 2.3 almacenar
        PHI[s,] <- phi
}
```

Apéndice A.3

```
sub_phi = data.frame('post_theta' = PHI[501:1000,1],
                     'post_sigma' = PHI[501:1000,2])
sub_phi

# Theta
mean(sub_phi$post_theta)
var(sub_phi$post_theta)

# PHI
mean(sub_phi$post_sigma)
var(sub_phi$post_sigma)
```

Apéndice A.4

```
set.seed(2)
rep = 500
PHI <- matrix(data = NA, nrow = rep, ncol = 2)
phi <- c( rnorm(1, mu0, sd = sqrt(t20)), rgamma(1, nu0/2, s20/2) )
PHI[1,] <- phi

for(s in 2:rep) {
        # 2.1 actualizar el valor de theta
        t2n    <- 1/( (t20**(-1)) + n*phi[2] )
        mun    <- (n*phi[2]*mean(SIM50$simulation)) / (t20**(-1) + n*phi[
        phi[1] <- rnorm( 1, mun, sd = sqrt(t2n) )
        alpha = dnorm(phi[1], mean=5, sd=sqrt(2)) / dnorm(PHI[s-1,1], mea

        if (runif(1) < alpha) {
          PHI[s, 1] = phi[1]}

        else{
          PHI[s] <- PHI[s - 1, 1]}

        #2.2 actualizar el valor de sigma^2

        nun    <- nu0+n
        s2n    <- s20 + (n-1)*var(SIM50$simulation) + n*(phi[1] - mean(SIM
        phi[2] <- rgamma(1, nun/2, s2n/2)
        alpha = dgamma(phi[2], shape = nun/2, scale = s2n/2) / dgamma(PHI
        # 2.3 almacenar
        if(runif(1) < alpha){
            PHI[s, 2] <- phi[2]}
        else{
            PHI[s, 2] <- PHI[s - 1, 2]}
}
```

8 Apéndice B

Gráficos

Apéndice B.1

```
ggplot(SIM50, aes(x=simulation)) +
 geom_histogram(aes(y=..density..),
                     colour=1,
                     fill="white",
                     bins = 11)+
 geom_density(lwd = 1,
                     alpha=.15,
                     fill="#FF6666") +
 geom_vline(aes(xintercept=mean(simulation)),
                color="blue", size=1) +
    xlab("X") +
    ylab("Probabilidad")+
    ggtitle('Simulaci n desde una distribuci n N(5, 2)') +
    theme(text = element_text(size = 20))
```

Apéndice B.2

```
plot(PHI[2:1000,1],type="l",ylab="Theta",xlab="Iteration")
```

Apéndice B.3

```
plot(PHI[,2],type="l",ylab="Phi",xlab="Iteration")
```

Apéndice B.4

```
ggplot(sub_phi, aes(x=post_theta)) +
 geom_histogram(aes(y=..density..),
                     colour=1,
                     fill="white",
                     bins = 15)+
 geom_density(lwd = 1,
                     alpha=.15,
                     fill="#FF6666") +
 geom_vline(aes(xintercept=mean(post_theta)),
                color="blue", size=1) +
    xlab("Theta") +
    ylab("Densidad")+
    ggtitle('Simulaci n de distribuci n condicional a posteriorir de Th
    theme(text = element_text(size = 20))
```

Apéndice B.5

```
ggplot(sub_phi, aes(x=post_sigma)) +
 geom_histogram(aes(y=..density..),
                  colour=1,
                  fill="white",
                  bins = 15)+
 geom_density(lwd = 1,
                  alpha=.15,
                  fill="#FF6666")  +
 geom_vline(aes(xintercept=mean(post_sigma)),
              color="blue", size=1)  +
    xlab("Phi") +
    ylab("Densidad")+
    ggtitle('Simulaci n de distribuci n condicional a posteriorir de Ph
    theme(text = element_text(size = 20))
```

Apéndice B.6

```
fig <- plot_ly(z = ~z) %>% add_surface(
  contours = list(
    z = list(
      show=TRUE,
      usecolormap = TRUE,
      highlightcolor="#ff0000",
      project=list(z=TRUE)
      )
    )
  )
fig <- fig %>% layout(
    scene = list(camera=list(eye = list(x=1.87, y=0.88, z=-0.64)),
               xaxis = list(ticketmode = 'array',
                            ticktext = seq(4.28, 5.61, 0.05),
                            tickvals = 0:24,
                            range = c(0,24)),

               yaxis = list(ticketmode = 'array',
                            ticktext = round(seq(0.283,0.966, 0.027), 2),
                              tickvals = 0:24,
                              range = c(0,24))
      )
  )

fig


##Generar las muestras mediante aceptacion / rechazo
```

Apéndice B.7

```
fig <- plot_ly(z = ~z) %>% add_surface(
  contours = list(
    z = list(
      show=TRUE,
      usecolormap = TRUE,
      highlightcolor="#ff0000",
      project=list(z=TRUE)
      )
    )
  )
fig <- fig %>% layout(
    scene = list(camera=list(eye = list(x=1.87, y=0.88, z=-0.64)),
               xaxis = list(ticketmode = 'array',
                            ticktext = seq(4.28, 5.61, 0.05),
                            tickvals = 0:24,
                            range = c(0,24)),

               yaxis = list(ticketmode = 'array',
                            ticktext = round(seq(0.283,0.966, 0.027), 2),
                              tickvals = 0:24,
                              range = c(0,24))
      )
  )
```