Homework 10

Louis Bensard

May 7, 2018

Problem 1:

(a)

We want to use Gibbs Sampling to generate random values from the random vector $(X, Y) \sim N(\mu, \Sigma)$ with:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Moreover, we know that :

$$X|(Y = y) \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2} \cdot (y - \mu_2), (1 - \rho^2)\sigma_1^2)$$
$$Y|(X = x) \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1} \cdot (x - \mu_1), (1 - \rho^2)\sigma_2^2)$$

Thus, we get the following Gibbs Sampling Algorithm:

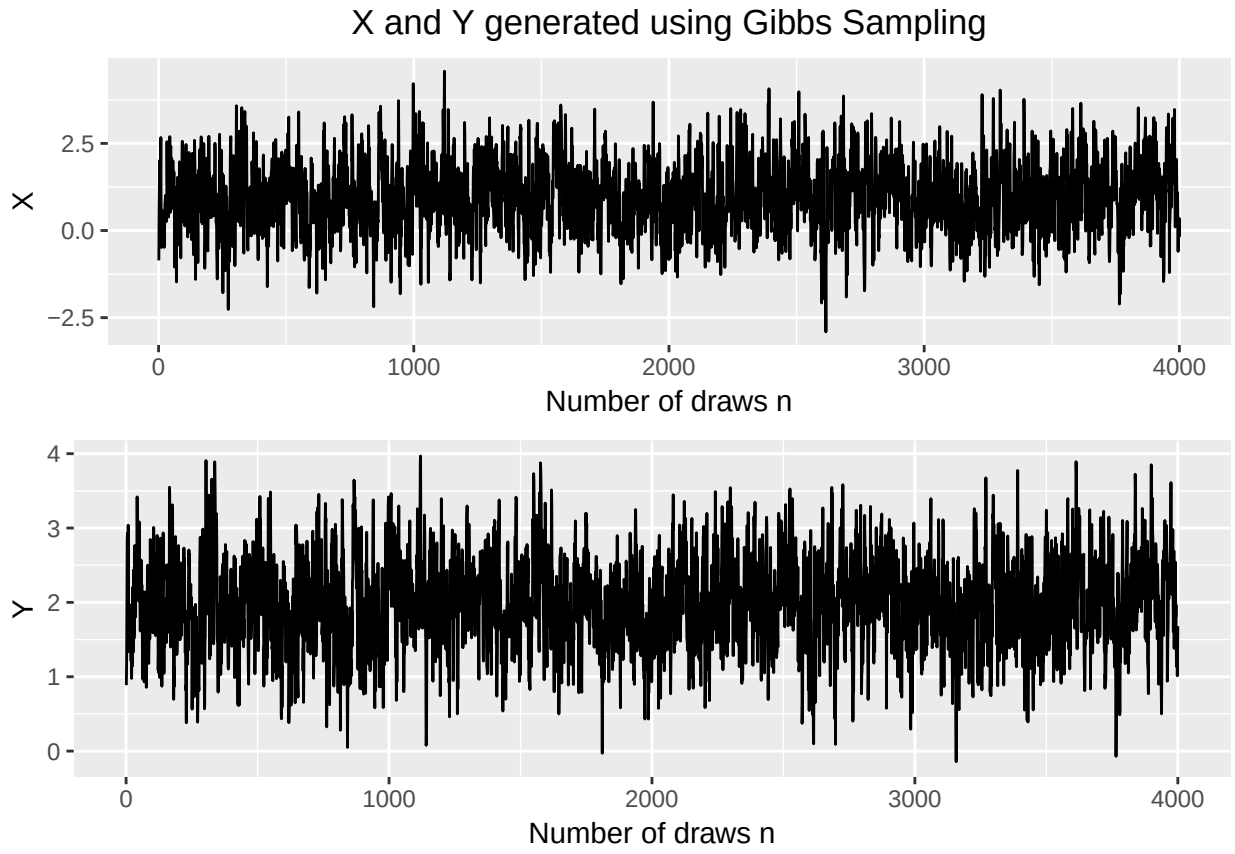
Step 1: Select a starting value $x^{(0)}$ and set $t = 0$

Step 2: Generate in turn: $Y^{(t+1)} \sim f(Y|x^{(t)})$, then $X^{(t+1)} \sim f(X|y^{(t+1)})$

Step 3: Increment t and go to step 2 until we reach n iterations

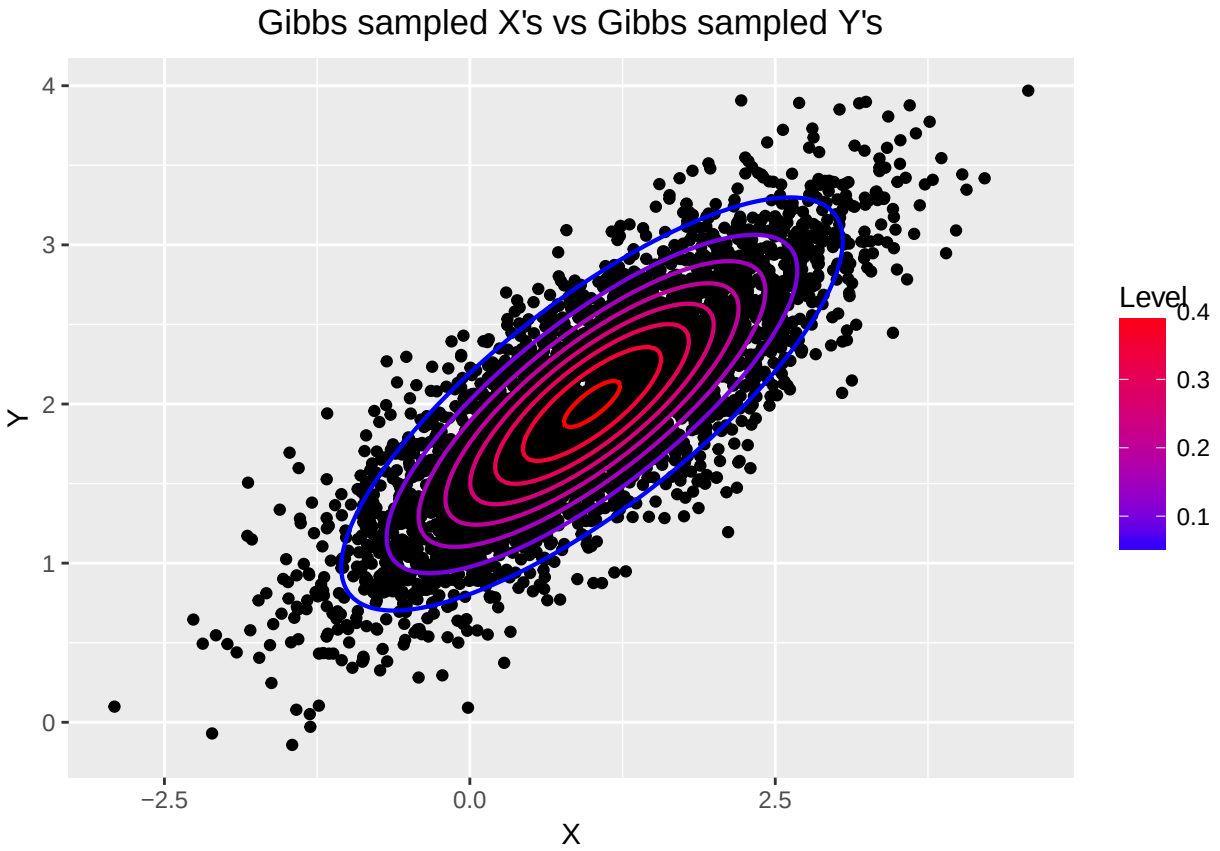
```
source('C:/Users/Louis/Documents/UPMC/M1/Spring 2018/MATH 534/Homework 10/functions_problem1.r')

Gibbs <- function(n, mu, Sigma, x0){

    X_t1 = x0
    x_vect = c(X_t1); y_vect = c()

    for(t in 1:n){

    Y_t1 = biv_norm_y_given_x(X_t1, mu, Sigma)
    X_t1 = biv_norm_x_given_y(Y_t1, mu, Sigma)

    x_vect= c(x_vect, X_t1)
    y_vect = c(y_vect, Y_t1)

    }

    return(list(X=x_vect, Y=y_vect))
}


mu = c(1,2)
Sigma = matrix(c(1,.5,.5,.4),2,2)
```

```r
n = 5000
x0 = 7 #inititial value

list = Gibbs(n, mu, Sigma, x0)
X = list$X
Y = list$Y

#burn-in
burn = 1000
X = X[(burn+1):n]
Y = Y[(burn+1):n]
```



After burn-in, the draws of the $X's$ and the $Y's$ looks good, around straight line evenly spread.

(b)

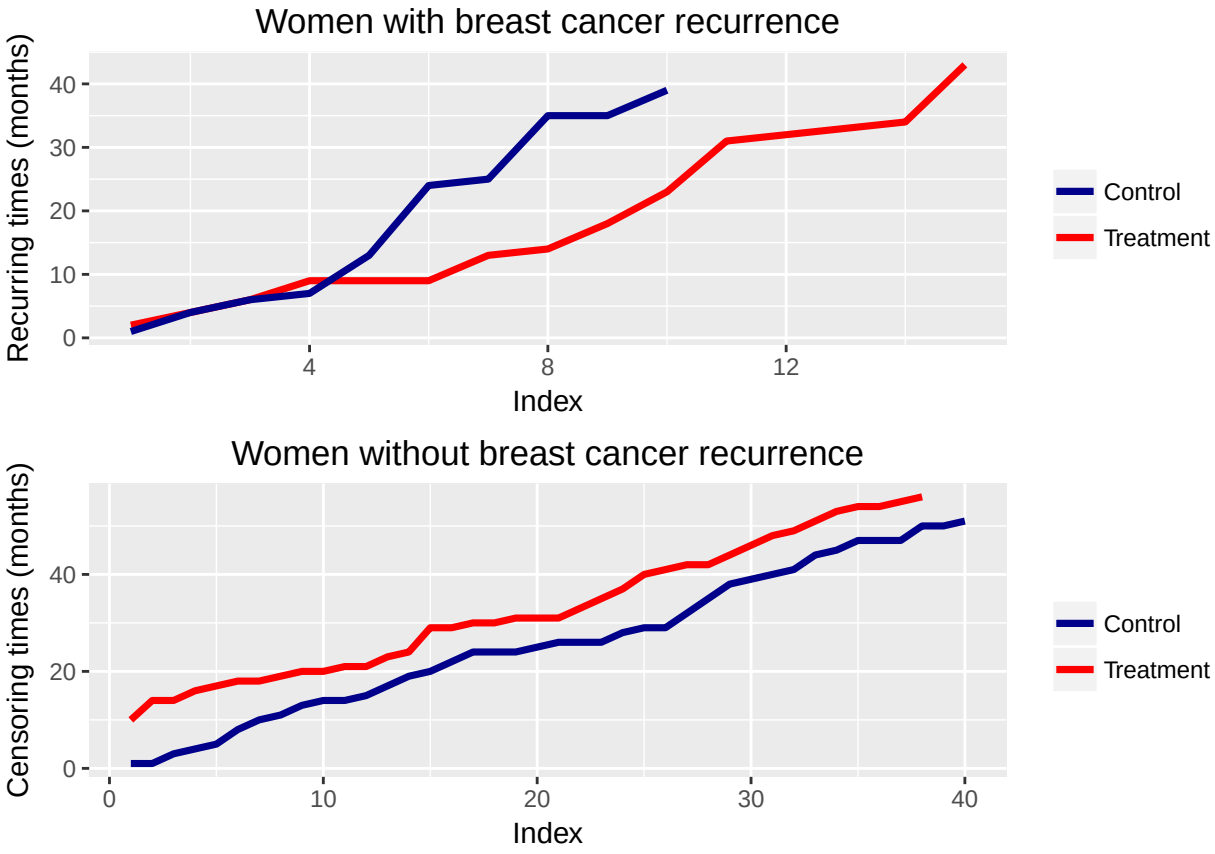


The plot is exactly as expected, a bivariate normal elipse, centered around $\mu = (1, 2)$. I overlayed the real contour of a bivariate normal with parameters μ and Σ and we can see that it fits very well the X vs Y scatterplot.

Problem 2:

(a)

The data is about women having recurrent breast cancer. Women having recurrent breast cancer are first treated by irradiation and then assigned to either a hormone therapy group or a control group. Then, in both groups, for each woman, we record how many months before the breast cancer comes back as "recurring time". If during the whole time of the clinical trial, the cancer does not come back, we keep the woman's time spent in the study as "censoring time" M and we say that her recurrence time is known to exceed M months.



(b)

We want to estimate the parameters τ and θ. The prior is:

$$f(\theta,\tau) \propto \theta^a \tau^b \cdot exp\{-c\theta - \theta\tau d\}, \text{ with } (a,b,c,d) = (3,60,1,120)$$

The likelihood of the data is:

$$L(\theta,\tau|y) \propto \theta^{(\sum \delta_i^C + \sum \delta_i^H)} \cdot \tau^{(\sum \delta_i^H)} \cdot exp\{-\theta \sum x_i^C - \tau\theta \sum x_i^H\}$$

As a result, the posterior $f(\theta,\tau|y) \propto L(\theta,\tau|y) \cdot f(\theta,\tau)$ is:

$$f(\theta,\tau|y) \propto \theta^{(\sum \delta_i^C + \sum \delta_i^H + a)} \cdot \tau^{(\sum \delta_i^H + b)} \cdot exp\{-\theta[\sum x_i^C + c] - \tau\theta[\sum x_i^H + d]\}$$

To implement Gibbs sampler, we need $f(\theta|\tau,y)$ and $f(\tau|\theta,y)$. From the above posterior distribution, we can conclude that:

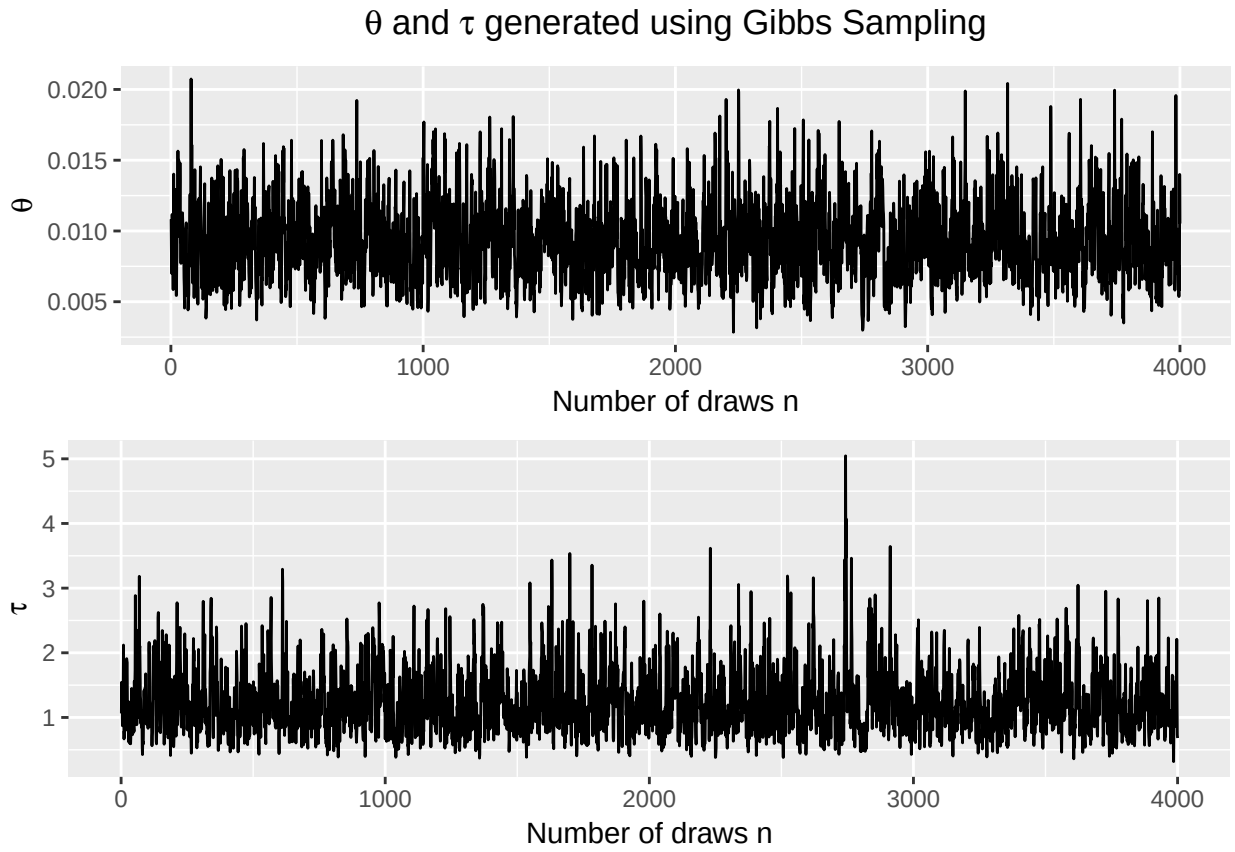
$$f(\theta|\tau,y) \propto \theta^{(\sum \delta_i^C + \sum \delta_i^H + a)} \cdot exp\{-\theta[\sum x_i^C + c + \tau(\sum x_i^H + d)]\}$$
$$\Rightarrow \theta|\tau \sim gamma(\sum \delta_i^C + \sum \delta_i^H + a + 1, \sum x_i^C + c + \tau(\sum x_i^H + d))$$

Similarly,

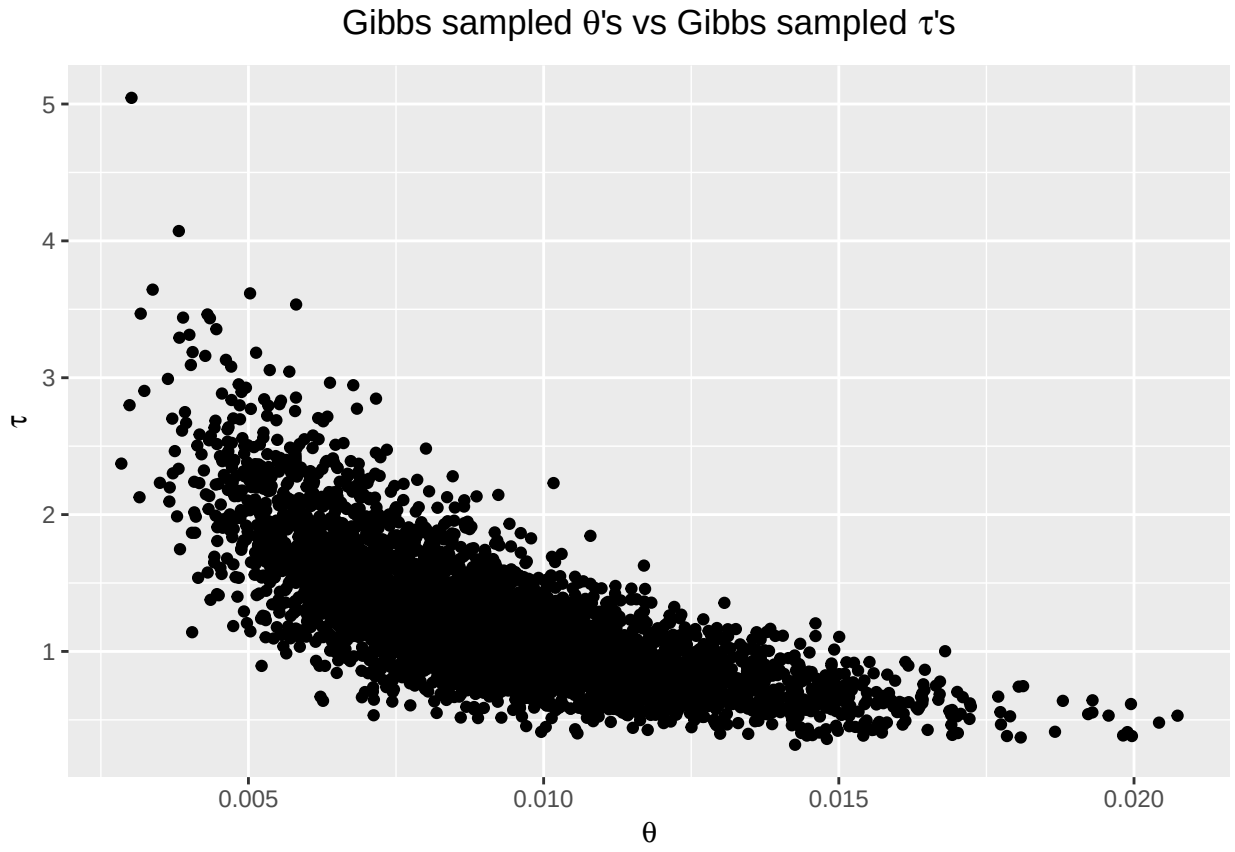
$$f(\tau|\theta, y) \propto \tau^{(\sum \delta_i^H + b)} \cdot exp\{-\tau\theta[\sum x_i^H + d]\}$$
$$\Rightarrow \tau|\theta \sim gamma(\sum \delta_i^H + b + 1, \theta[\sum x_i^H + d])$$

(c)

```
n=5000
sum_del_H = 15
sum_del_C = 10
sum_x_H = sum(data[data$treatment==1,1])
sum_x_C = sum(data[data$treatment==0,1])
a=3; b=1; c=60; d=120

tau_t1 = rgamma(1, 1, 1) #initial value
theta_vect = c(); tau_vect = c(tau_t1)

#Gibbs algo
for(t in 1:n){

    theta_t1 = rgamma(1, sum_del_H + sum_del_C + a + 1, sum_x_C + c + tau_t1*(d + sum_x_H))
    tau_t1 = rgamma(1, sum_del_H + b + 1, theta_t1*(d + sum_x_H))

    theta_vect = c(theta_vect, theta_t1)
    tau_vect = c(tau_vect, tau_t1)
}


#burn-in: after a 1000 draws, the distribution of the thetas generated are very close same for tau's
burn = 1000
theta_vect = theta_vect[(burn+1):n]
tau_vect = tau_vect[(burn+1):n]
```



The draws of the $\theta's$ and the $\tau's$ looks good, around straight line evenly spread.



(d)

```
m1 = mean(theta_vect); s1 = sd(theta_vect)
m2 = mean(tau_vect); s2 = sd(tau_vect)

L1 = sort(theta_vect)[0.025*(n-burn)]
U1 = sort(theta_vect)[0.975*(n-burn)]

L2 = sort(tau_vect)[0.025*(n-burn)]
U2 = sort(tau_vect)[0.975*(n-burn)]
```

	τ	θ
mean	1.22846	0.00922
sd	0.49069	0.0027
95% CI	[0.53994, 2.39334]	[0.00475, 0.01518]

Table 1: Summary Statistics

(e)

Let's find the exact prior for τ. Since the joint prior is $f(\theta, \tau) \propto \theta^a \tau^b \cdot exp\{-c\theta - \theta\tau d\}$ we have:

$$f(\tau) \propto \tau^b \int_0^\infty \theta^a \cdot exp\{-c\theta - \theta\tau d\} \cdot d\theta$$

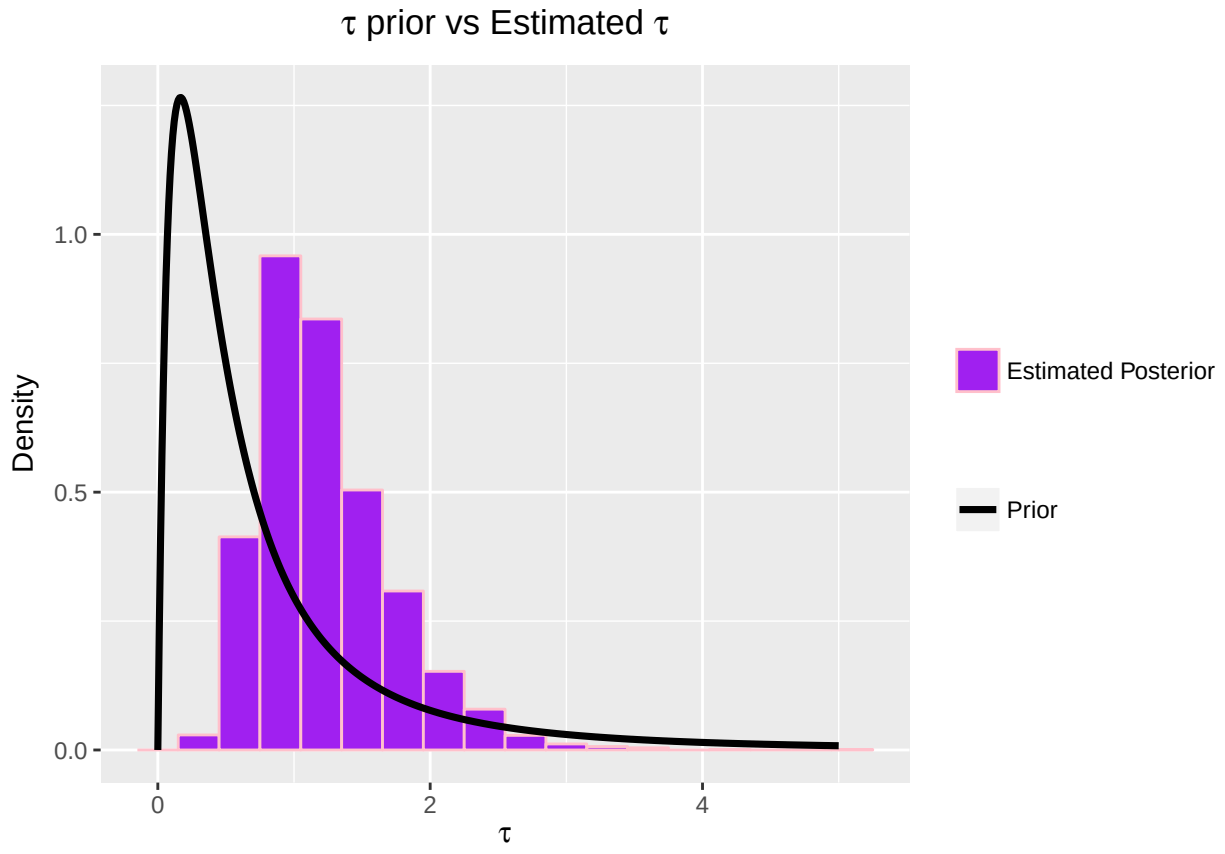
We recognize the kernel of a $gamma(a+1, c+\tau d)$ distribution, thus:

$$f(\tau) \propto \frac{\tau^b \cdot \Gamma(a+1)}{(c+\tau d)^{a+1}}$$

Since, $\int_0^\infty f(\tau)d\tau = 1/51840000$, then:

$$f(\tau) = (5.184 \cdot 10^7) \cdot \frac{\tau^b \cdot \Gamma(a+1)}{(c+\tau d)^{a+1}}$$

```
tau_temp = seq(0,5, length=500)
tau_prior = 51840000*(tau_temp^b)*gamma(a+1) / ((c + tau_temp*d)^(a+1))
```



(f)

We know that the time until second recurrence is assumed to have an exponential distribution with parameter $\tau\theta$ for the hormone therapy group and with parameter θ for the control group. Therefore, for the hormone group, the expected time until second recurrence is $\frac{1}{\tau\theta}$ and for the control group it is $\frac{1}{\theta}$. Therefore, if $\tau < 1$, then the time until second recurrence increases compared to the control group (which is what we want) and if $\tau > 1$ this time decreases.

In part (d), we showed that a 95% credible interval for τ is [0.53994, 2.39334]. This interval contains the value $\tau = 1$, therefore, even though it is centered at $\tau = 1.22846$, we cannot conclude that the hormone therapy increases (or decreases) the time until second recurrence. The Hormone and Control group don't have significantly different recurrence times.

(g)

Hyperparameters are halved:

```
a1=a/2; b1=b/2; c1=c/2; d1=d/2

tau_t1 = rgamma(1, 1, 1) #initial value
theta_vect = c(); tau_vect = c(tau_t1)

#Gibbs algo
for(t in 1:n){

    theta_t1 = rgamma(1, sum_del_H + sum_del_C + a1 + 1, sum_x_C + c1 + tau_t1*(d1 + sum_x_H))
    tau_t1 = rgamma(1, sum_del_H + b1 + 1, theta_t1*(d1 + sum_x_H))

    theta_vect = c(theta_vect, theta_t1)
    tau_vect = c(tau_vect, tau_t1)
}


#burn-in
burn = 1000 # after a 1000 draws, the distribution of the thetas generated are very close
#same for tau's
theta_vect = theta_vect[(burn+1):n]
tau_vect = tau_vect[(burn+1):n]

m1 = mean(theta_vect); s1 = sd(theta_vect)
m2 = mean(tau_vect); s2 = sd(tau_vect)

L1 = sort(theta_vect)[0.025*(n-burn)]
U1 = sort(theta_vect)[0.975*(n-burn)]

L2 = sort(tau_vect)[0.025*(n-burn)]
U2 = sort(tau_vect)[0.975*(n-burn)]
```

	τ	θ
mean	1.30499	0.00874
sd	0.53253	0.00259
95% CI	[0.55268, 2.62168]	[0.00449, 0.01458]

Table 2: Summary Statistics after halving hyperparameters

Hyperparameters are doubled:

```
a1=a*2; b1=b*2; c1=c*2; d1=d*2
```

```
tau_t1 = rgamma(1, 1, 1) #initial value
theta_vect = c(); tau_vect = c(tau_t1)

#Gibbs algo
for(t in 1:n){

    theta_t1 = rgamma(1, sum_del_H + sum_del_C + a1 + 1, sum_x_C + c1 + tau_t1*(d1 + sum_x_H))
    tau_t1 = rgamma(1, sum_del_H + b1 + 1, theta_t1*(d1 + sum_x_H))

    theta_vect = c(theta_vect, theta_t1)
    tau_vect = c(tau_vect, tau_t1)
}


#burn-in
burn = 1000 # after a 1000 draws, the distribution of the thetas generated are very close
#same for tau's
theta_vect = theta_vect[(burn+1):n]
tau_vect = tau_vect[(burn+1):n]

m1 = mean(theta_vect); s1 = sd(theta_vect)
m2 = mean(tau_vect); s2 = sd(tau_vect)

L1 = sort(theta_vect)[0.025*(n-burn)]
U1 = sort(theta_vect)[0.975*(n-burn)]

L2 = sort(tau_vect)[0.025*(n-burn)]
U2 = sort(tau_vect)[0.975*(n-burn)]
```

	τ	θ
mean	1.06259	0.01036
sd	0.4123	0.00281
95% CI	[0.48308, 2.08516]	[0.00548, 0.01649]

Table 3: Summary Statistics after doubling hyperparameters

We can see that in both cases, halving or doubling the hyperparameters only slighly changes the mean of τ and its 95% credible interval. In both cases, the credible interval still contains the value $\tau = 1$ thus we still cannot conclude that the time until second recurrence for the hormone therapy group is significantly different from the time of the control group. Thus, the results are not very sensitive to hyperparameter values. I would suggest the drug company to come up with hyperparameter values that make the results sensitive to them or change the hormone treatment.