Math 475: Final Exam

Hannah Zmuda

01/08/2021

Problem 1

part a

Show the outputs of the EM algorithm are consistent with the given parameter equations To find the updated parameters (i.e. the maximized value) we fist need to find the Q-function (E step) then maximize the Q function by taking the derivative in regard to each of the parameters (M step). **E step:** Given the observed likelihood, we can compute the complete likelihood to be:

$$L(\theta|n_{k,i}) = \Pi_{i=0}^{16} \frac{[z(\theta)^{(n_0)} t(\theta)^{n_i} p(\theta)^{n_i}]}{i!}$$

Given the likelihood equation, we can work out the log likelihood to be:

$$log[L(\theta|n_{k,i})] = \Sigma_{i=0}^{16} z_0 log(\alpha 1_{i=0}) +$$

$$(t_i)[log(\alpha 1_{i=0}) + log(\beta \mu^i e^{-\mu}) + log((1-\alpha-\beta)\lambda^i e^{-\lambda})] +$$

$$(p_i)[log(\alpha 1_{i=0}) + log(\beta \mu^i e^{-\mu}) + log((1-\alpha-\beta)\lambda^i e^{-\lambda})] - log(i!)$$

where y is the complete data set and z_0, t_i, p_i represent three different groups. These are further broken down into the zero, typical, and promiscuous groups. To find your Q-function, take the expectation of the log likelihood function:

$$Q(\theta|\theta^{(t)}) = n_{z,0}^{(t)} \Sigma_{i=0}^{16} log(\alpha 1_{i=0}) + n_{t,i}^{(t)} \Sigma_{i=0}^{16} log(\beta \mu^i exp(-\mu)) + n_{p,i}^{(t)} \Sigma_{i=0}^{16} log((1-\alpha-\beta)\lambda^i exp(\lambda))$$

M step: For the M step of the EM algorithm, we need to maximize the Q function in regard to each parameter then set it equal to zero.

When the derivative is set equal to zero, we find that the updated parameters equal to what we expected:

$$\alpha^{(t+1)} = \frac{n_0 z_0(\theta^{(t)} t)}{N}$$

$$\beta^{(t+1)} = \Sigma_{i=0}^{16} \frac{n_i t_i(\theta^{(t)})}{N}$$

$$\mu^{(t+1)} = \frac{\Sigma_{i=0}^{16} i n_i t_i(\theta^{(t)})}{\Sigma_{i=0}^{16} n_i t_i(\theta^{(t)})}$$

$$\lambda^{(t+1)} = \frac{\Sigma_{i=0}^{16} i n_i p_i(\theta^{(t)})}{\Sigma_{i=0}^{16} n_i p_i(\theta^{(t)})}$$

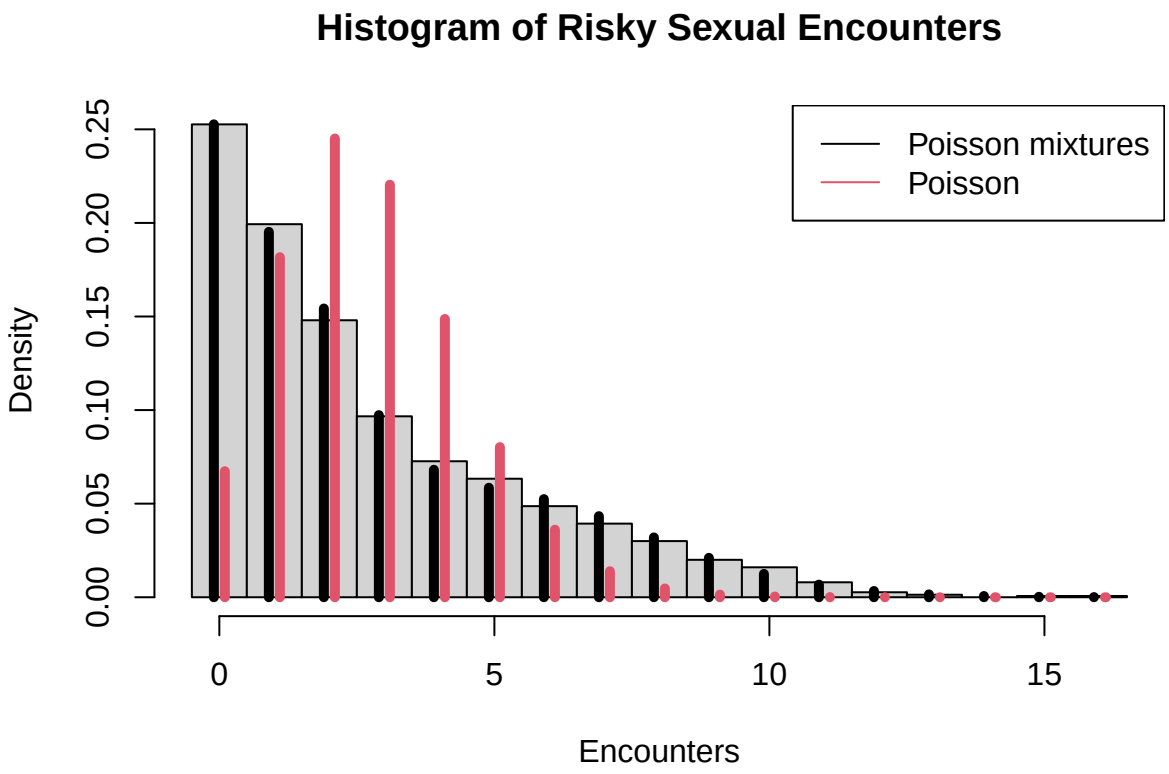
part b and c

```r
set.seed(475)
# initialize variables
data = data.frame(enc = 0:16, freq = c(379, 299, 222, 145, 109,
    95, 73, 59, 45, 30, 24, 12, 4, 2, 0, 1, 1))
N = sum(data$freq)
y = rep(data$enc, data$freq)
alpha = 0.5
beta = 0.8
mu = 2
lambda = 15
param = c(alpha, beta, mu, lambda)
param.guess = c(0.1, 0.2, 3, 4)
tol = 1e-10
tol.cur = 100
time = 0
i = 0:16
# functions
log.likelihood <- function(alpha, beta, mu, lambda, x) {
    l = 0
    alpha = exp(alpha)/(1 + exp(alpha))
    beta = exp(beta)/(1 + exp(beta))
    mu = exp(mu)/(1 + exp(mu))
    lambda = exp(lambda)/(1 + exp(lambda))
    for (i in 1:length(x$enc)) {
        e = x$enc[i]
        n = x$freq[i]
        if (e == 0) {
            l = l + n * log(alpha + beta * exp(-mu) + (1 - alpha -
                beta) * exp(-lambda))
            print(l)
        } else {
            l = l + n * log(beta * (mu^e) * exp(-mu) + (1 - alpha -
                beta) * exp(-lambda) * lambda^e) - log(factorial(e))
            print(l)
        }
    }
    return(l)
}

# EM Algorithm
while (tol.cur > tol) {
    pi = (beta * exp(-mu) * mu^i) + ((1 - alpha - beta) * exp(-lambda) *
        lambda^i)
    pi[1] = pi[1] + alpha
    z.stat = alpha/(pi[1])
    t.stat = (beta * (mu^i) * exp(-mu))/pi
    p.stat = ((1 - alpha - beta) * exp(-lambda) * lambda^i)/pi
    alpha = (data$freq[1] * z.stat)/N
    beta = sum(data$freq * t.stat)/N
    mu = sum(i * data$freq * t.stat)/sum(data$freq * t.stat)
    lambda = sum(i * data$freq * p.stat)/sum(data$freq * p.stat)
    new.param = c(alpha, beta, mu, lambda)
    tol.cur = sum(abs(new.param - param))
```

```
    param = new.param
    time = time + 1
}
param
```

```
## [1] 0.1221661 0.5625419 1.4674746 5.9388889
```

```
hist(rep(data$enc, data$freq), breaks = -0.5 + c(0:17), freq = F,
    main = "Histogram of Risky Sexual Encounters", xlab = "Encounters")
# hist(y,freq=F)
z = 0:16
prob = (beta * exp(-mu) * mu^z + (1 - alpha - beta) * exp(-lambda) *
    lambda^z)/(factorial(z))
prob[1] = prob[1] + alpha
for (i in 1:length(z)) {
    lines(c(z[i] - 0.1, z[i] - 0.1), c(0, prob[i]), lwd = 5,
        col = 1)
}

pois.hat = mean(y)

for (i in 1:length(z)) {
    lines(c(z[i] + 0.1, z[i] + 0.1), c(0, dpois(z[i], pois.hat)),
        lwd = 5, col = 2)
}

legend("topright", c("Poisson mixtures", "Poisson"), lty = 1,
    col = 1:2)
```




```
# standard error Use log likelihood at theta.hat values (i.e.
# new parameter values)
set.seed(1234)
data = data.frame(enc = 0:16, freq = c(379, 299, 222, 145, 109,
    95, 73, 59, 45, 30, 24, 12, 4, 2, 0, 1, 1))
tol = 1e-10
B = 1000
result.boot <- NULL
alpha <- 1
beta <- 2
mu <- 4
lambda <- 10
param <- c(alpha, beta, mu, lambda)

for (j in 1:B) {
    # randomize samples
    data.boot <- rmultinom(1, sum(data$freq), prob = data$freq/length(data$freq))
    # set initial values
    tol.cur <- 100
    N <- sum(data.boot)
    i = c(0:16)
    # loop
    while (tol.cur > tol) {
        pi = (beta * exp(-mu) * mu^i) + ((1 - alpha - beta) *
            exp(-lambda) * lambda^i)
        pi[1] = pi[1] + alpha
```

```
        z.stat = alpha/(pi[1])
        t.stat = (beta * (mu^i) * exp(-mu))/pi
        p.stat = ((1 - alpha - beta) * exp(-lambda) * (lambda^i))/pi

        alpha = (data.boot[1] * z.stat)/N
        beta = sum(data.boot * t.stat)/N
        mu = sum(i * data.boot * t.stat)/sum(data.boot * t.stat)
        lambda = sum(i * data.boot * p.stat)/sum(data.boot *
            p.stat)

        new.param = c(alpha, beta, mu, lambda)
        tol.cur = sum(abs(new.param - param))
        param = new.param
    }
    result.boot <- rbind(result.boot, param)
}
result.boot[B, ]
```

```
## [1] 0.1365674 0.5705522 1.6564927 6.2467511
```

```
cov(result.boot)  #covariance matrix to show standard error
```

```
##               [,1]          [,2]         [,3]        [,4]
## [1,]  0.0003955050 -0.0001790008 0.0015697232 0.001484594
## [2,] -0.0001790008  0.0004392937 0.0001291111 0.001400945
## [3,]  0.0015697232  0.0001291111 0.0121965277 0.013091738
## [4,]  0.0014845938  0.0014009455 0.0130917380 0.038778808
```

```
# pairwise correlation
cor(result.boot)
```

```
##            [,1]        [,2]       [,3]      [,4]
## [1,]  1.0000000 -0.42943889 0.71470853 0.3790831
## [2,] -0.4294389  1.00000000 0.05577864 0.3394271
## [3,]  0.7147085  0.05577864 1.00000000 0.6019799
## [4,]  0.3790831  0.33942709 0.60197988 1.0000000
```

Problem 2

part a: Metropolis and M-H Algorithm

We are seeking to estimate $\alpha, \beta, \mu, \lambda$ using the Metropolis Algorithm.

Because the proposal distribution is symmetric, it can be canceled out in the ratio calculation, and we can then focus on the ratio of the target distribution with theta star and the previous theta value.

part 2: MCMH

```
# Metropolis Samples (Symmetric Proposal Distribution)
set.seed(575)
# initialize variables and constants
data = data.frame(enc = 0:16, freq = c(379, 299, 222, 145, 109,
    95, 73, 59, 45, 30, 24, 12, 4, 2, 0, 1, 1))
N = sum(data$freq)
y = rep(data$enc, data$freq)
reject <- 0
alpha = rep(0, N)
beta = rep(0, N)
mu = rep(0, N)
lambda = rep(0, N)
minn = -1
maxx = 1

# functions
log.likelihood <- function(alpha, beta, mu, lambda, x) {
    l = 0
    # reparameterization
    alpha = exp(alpha)/(1 + exp(alpha))
    beta = exp(beta)/(1 + exp(beta))
    mu = exp(mu)/(1 + exp(mu))
    lambda = exp(lambda)/(1 + exp(lambda))
    for (i in 1:length(x$enc)) {
        e = x$enc[i]
        n = x$freq[i]
        if (e == 0) {
            l = l + n * log(alpha + beta * exp(-mu) + (1 - alpha -
                beta) * exp(-lambda))
            print(l)
        } else {
            l = l + n * log(beta * (mu^e) * exp(-mu) + (1 - alpha -
                beta) * exp(-lambda) * lambda^e) - log(factorial(e))
            print(l)
        }
    }
    return(l)
}

# initialize sample values i.e chains
alpha[1] = rnorm(1, 0, 1)
```

```
beta[1] = rnorm(1, 0, 1)
mu[1] = rnorm(1, 0, 1)
lambda[1] = rnorm(1, 0, 1)
sigma = 1
i = 1

# for(i in 2:N){ #sample from proposal distribution
# (symmetrical) alpha.star <- rnorm(1,mean = alpha[i-1], sd =
# sigma) beta.star <- rnorm(1,mean = beta[i-1], sd = sigma)
# mu.star <- rnorm(1,mean = mu[i-1], sd = sigma) lambda.star
# <- rnorm(1,mean = lambda[i-1], sd = sigma)
# print(c(alpha.star,beta.star,mu.star,lambda.star)) num <-
# log.likelihood(alpha.star,beta.star,mu.star,lambda.star,data)
# dem <-
# log.likelihood(alpha[i-1],beta[i-1],mu[i-1],lambda[i-1],data)
# ratio <- exp(num) - exp(dem) accept.prob <- min(1,ratio) U
# = runif(1) if(U <= accept.prob){ alpha[i] <- alpha.star
# beta[i] <- beta.star mu[i] <- mu.star lambda[i] <-
# lambda.star } else{ alpha[i] <- alpha[i-1] beta[i] <-
# beta[i-1] mu[i] <- mu[i-1] lambda[i] <- lambda[i-1] reject
# = reject + 1 } } # alpha = exp(alpha)/(1+exp(alpha)) # beta
# = exp(beta)/(1+exp(beta)) # mu = exp(mu) # lambda =
# exp(lambda) c(mean(alpha),mean(beta),mean(mu),mean(lambda))
# print('Rejection Rate:') 100*(reject/N)
```

Problem 3

part a:

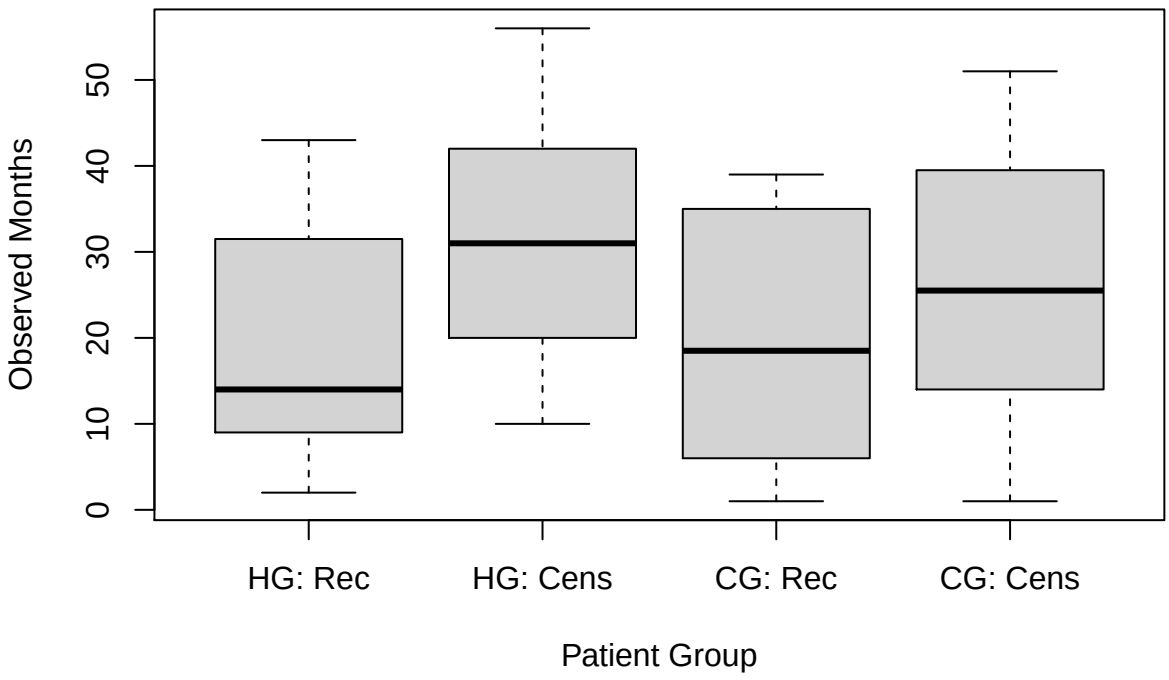
From the given data for the clinical trial, we can see from the box-and-whisker plot that the Hormone group with the censored time has the most patients out of the four groups. Not only that, but the means are more spread apart within the hormone group (difference of 13.68) compared to the control group (difference of 7.2). Based on the combined Normal QQ plot, the Control Group with Censored time most closely follows a normal distribution ("3" from list.id legend). The two from the hormone group of Recurrence and Censor times (0 and 1, respectively), do not closely align with a normal distribution. ## part b: Given the likelihood and prior, we can use those to calculate the conditional distributions of θ and τ. To do this we will first need to calculate the joint probability density function:

$$P(y|\theta,\tau) = L(\theta,\tau|y)f(\theta,\tau)$$

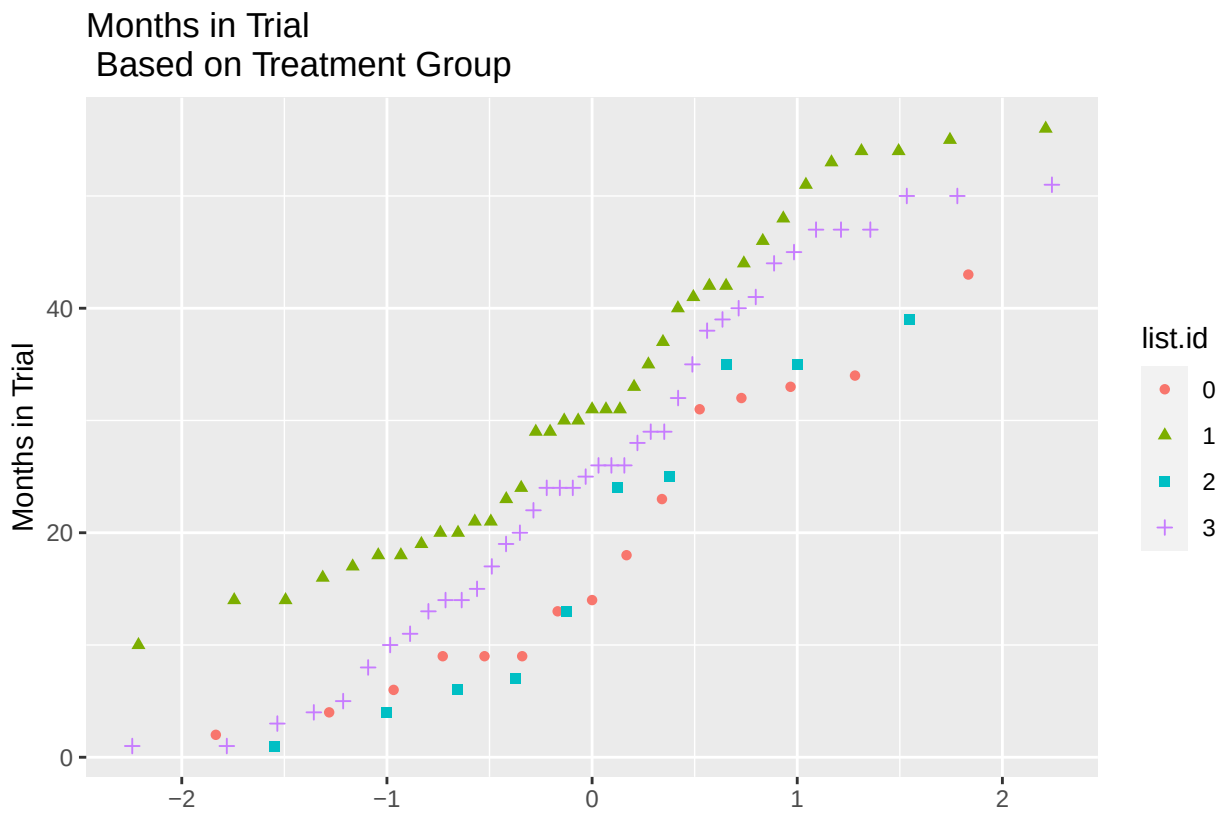
$$P(\theta,\tau|y) = P(y|\theta,\tau)f(\theta,\tau) = L(\theta,\tau|y)f(\theta,\tau)f(\theta,\tau)$$

$$P(\theta,\tau|y) = \theta^{\Sigma\delta_i^c+\Sigma\delta_i^H+2a}\tau^{\Sigma\delta_i^H+2b}exp(-\theta(\Sigma x_i^C+2c)-\tau\theta(\Sigma x_i^H+2d))$$

```
set.seed(12345)
library(ggplot2)
# initialize data sets
hormone.rec <- c(2, 4, 6, 9, 9, 9, 13, 14, 18, 23, 31, 32, 33,
    34, 43)
lab.h.r <- rep(0, length(hormone.rec))
hormone.cens <- c(10, 14, 14, 16, 17, 18, 18, 19, 20, 20, 21,
    21, 23, 24, 29, 29, 30, 30, 31, 31, 31, 33, 35, 37, 40, 41,
    42, 42, 44, 46, 48, 51, 53, 54, 54, 55, 56)
lab.h.c <- rep(1, length(hormone.cens))
control.rec = c(1, 4, 6, 7, 13, 24, 25, 35, 35, 39)
lab.c.r <- rep(2, length(control.rec))
control.cens = c(1, 1, 3, 4, 5, 8, 10, 11, 13, 14, 14, 15, 17,
    19, 20, 22, 24, 24, 24, 25, 26, 26, 26, 28, 29, 29, 32, 35,
    38, 39, 40, 41, 44, 45, 47, 47, 47, 50, 50, 51)
lab.c.c <- rep(3, length(control.cens))
list.id <- c(lab.h.r, lab.h.c, lab.c.r, lab.c.c)
dat.df <- c(hormone.rec, hormone.cens, control.rec, control.cens)

cancer.dat <- data.frame(dat.df, list.id)
cancer.dat$list.id <- as.factor(cancer.dat$list.id)

# part a: Plots and quantiles
dat.list <- list(hormone.rec, hormone.cens, control.rec, control.cens)
boxplot(dat.list, range = 0, names = c("HG: Rec", "HG: Cens",
    "CG: Rec", "CG: Cens"), xlab = "Patient Group", ylab = "Observed Months")
```

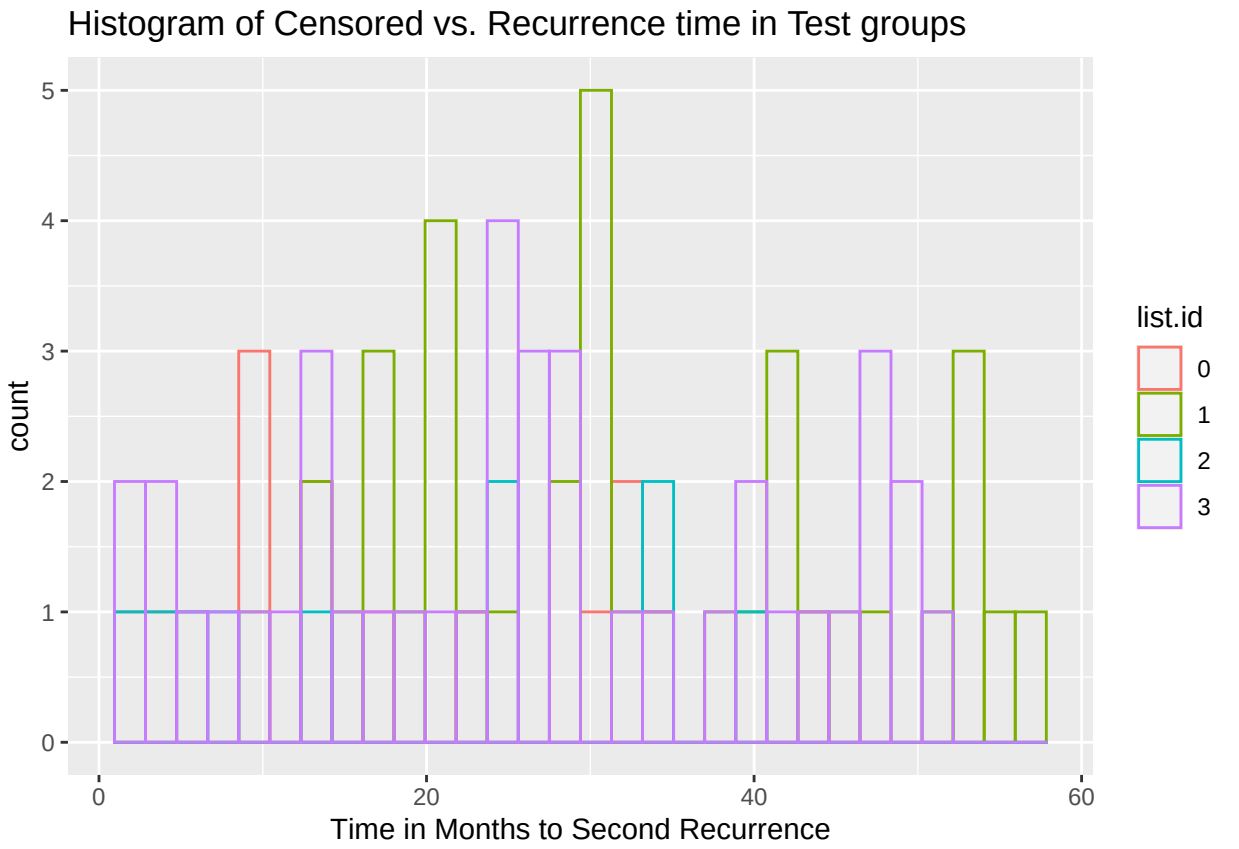



```
p <- qplot(sample = dat.df, data = cancer.dat, color = list.id,
    shape = list.id)
p + labs(title = "Months in Trial \n Based on Treatment Group",
    y = "Months in Trial")
```




```
ggplot(cancer.dat, aes(x = dat.df, fill = list.id, color = list.id)) +
    geom_histogram(position = "identity", fill = "transparent") +
    labs(title = "Histogram of Censored vs. Recurrence time in Test groups",
        x = "Time in Months to Second Recurrence")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

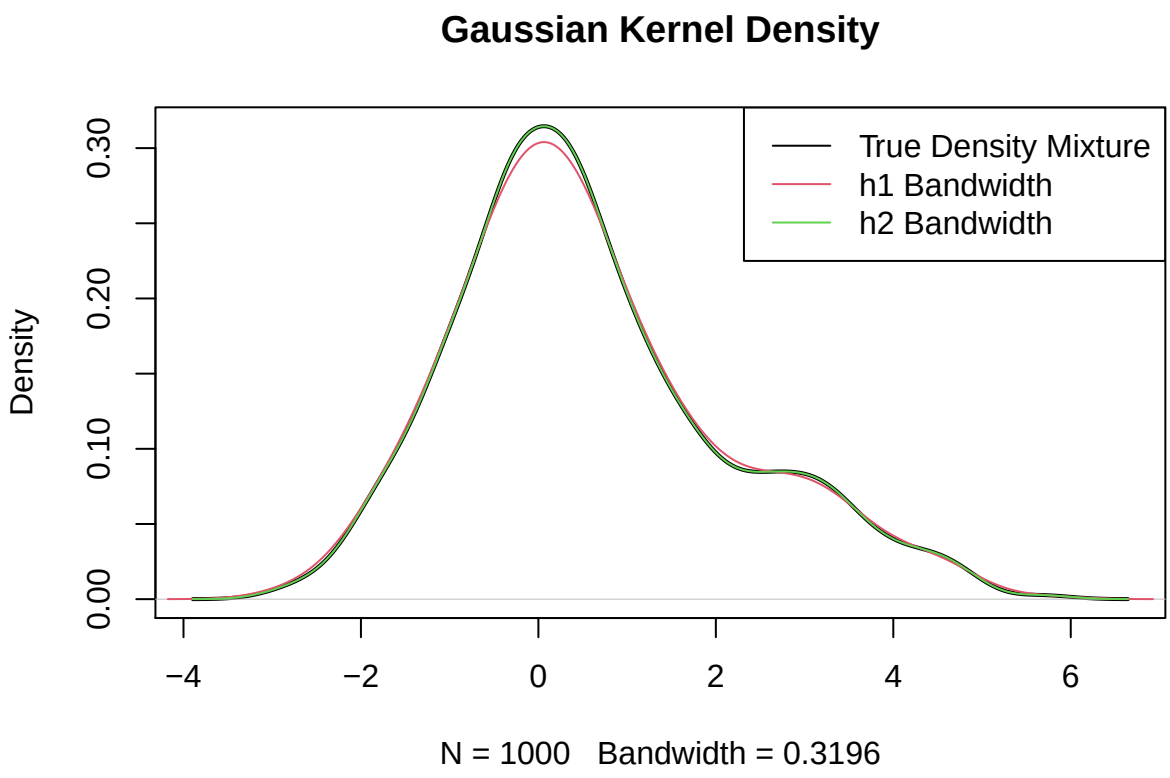



```
# part c: Gibbs Sampler Initalize and prepare recurrence and
# censored data based on patient group
key.hormone <- c(rep(1, length(hormone.rec)), rep(0, length(hormone.cens)))
key.control <- c(rep(1, length(control.rec)), rep(0, length(control.cens)))
dat.hormone <- data.frame(dat = c(hormone.rec, hormone.cens),
    key.hormone)
dat.control <- data.frame(dat = c(control.rec, control.cens),
    key.control)
```

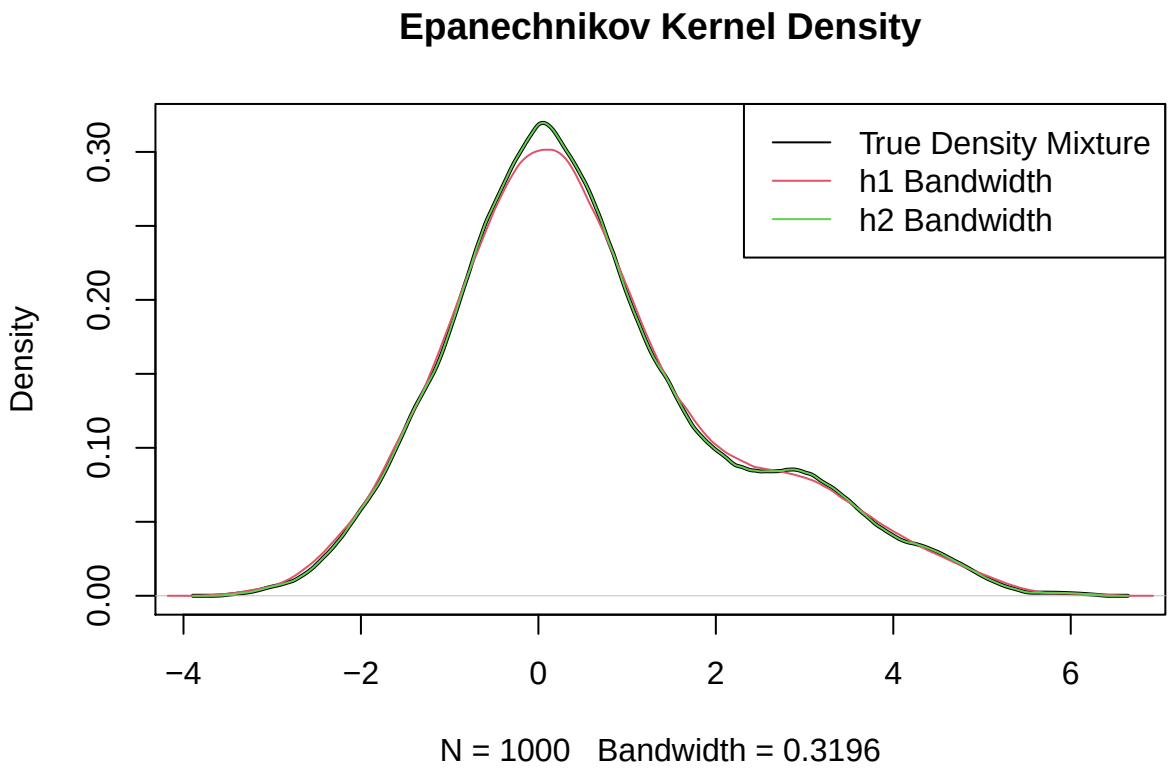
Problem 4

Problem 5

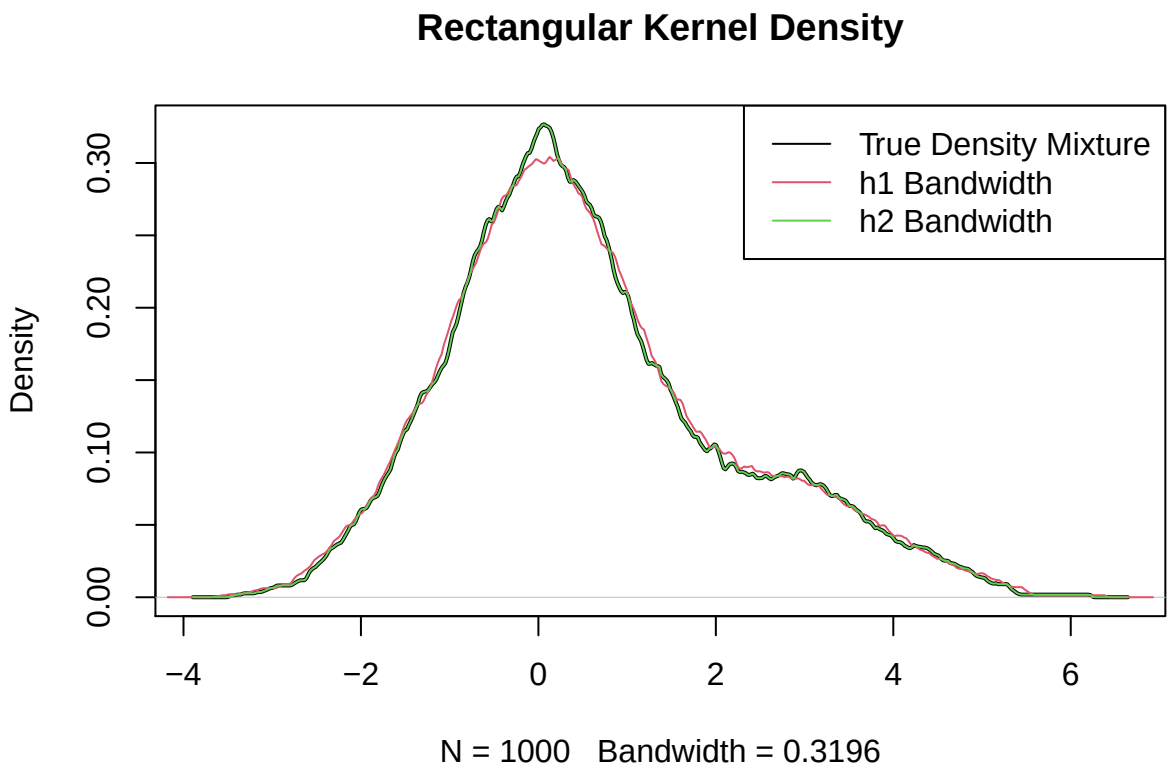
```r
set.seed(775)
n <- 1000  #sample size
comp <- sample(1:2, prob = c(0.8, 0.2), size = n, replace = TRUE)
mix <- 0.8 * rnorm(n, 0, 1) + 0.2 * rnorm(n, 3, 1)
mu = c(0, 3)
stan.dev = c(1, 1)
samp <- rnorm(n, mean = mu[comp], sd = stan.dev[comp])

h1 <- 1.06 * (n^(-1/5)) * sd(samp)
h2 <- 0.9 * (n^(-1/5)) * min(sd(samp), (IQR(samp))/1.34)
k.gauss <- function(t) {
    return((1/sqrt(2 * pi)) * -0.5 * t^2)
}

plot(density(samp), col = 1, main = "Gaussian Kernel Density",
    lwd = 2)  #true density estimate
lines(density(samp, bw = h1, kernel = "gaussian"), col = 2)
lines(density(samp, bw = h2, kernel = "gaussian"), col = 3)
legend("topright", c("True Density Mixture", "h1 Bandwidth",
    "h2 Bandwidth"), lty = 1, col = 1:3)
```



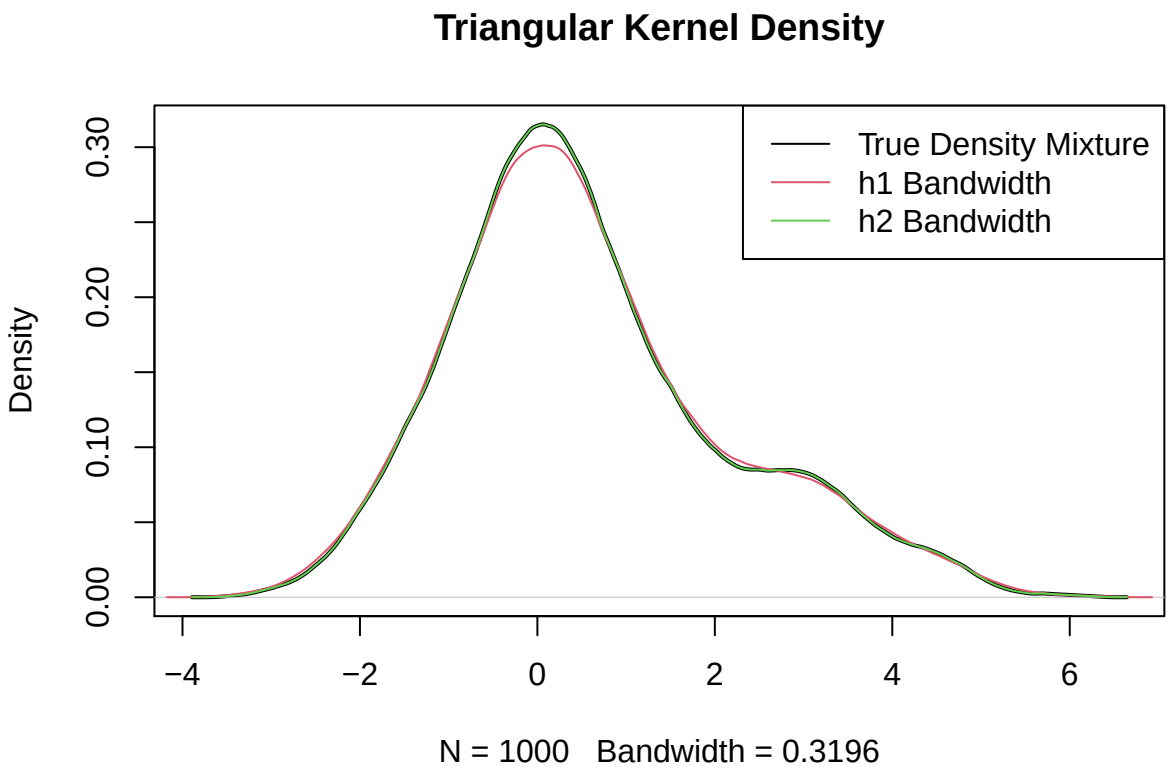
```r
# part b
plot(density(samp, kernel = "epanechnikov"), col = 1, main = "Epanechnikov Kernel Density",
    lwd = 2)  #true density estimate
lines(density(samp, bw = h1, kernel = "epanechnikov"), col = 2)
lines(density(samp, bw = h2, kernel = "epanechnikov"), col = 3)
legend("topright", c("True Density Mixture", "h1 Bandwidth",
    "h2 Bandwidth"), lty = 1, col = 1:3)
```




```r
plot(density(samp, kernel = "rectangular"), col = 1, main = "Rectangular Kernel Density",
    lwd = 2)  #true density estimate
lines(density(samp, bw = h1, kernel = "rectangular"), col = 2)
lines(density(samp, bw = h2, kernel = "rectangular"), col = 3)
legend("topright", c("True Density Mixture", "h1 Bandwidth",
    "h2 Bandwidth"), lty = 1, col = 1:3)
```




```
plot(density(samp, kernel = "triangular"), col = 1, main = "Triangular Kernel Density",
    lwd = 2)  #true density estimate
lines(density(samp, bw = h1, kernel = "triangular"), col = 2)
lines(density(samp, bw = h2, kernel = "triangular"), col = 3)
legend("topright", c("True Density Mixture", "h1 Bandwidth",
    "h2 Bandwidth"), lty = 1, col = 1:3)
```



Based on the plots from above. The h2 bandwidth is a better smoothing parameter compared to using the h1 bandwidth.