Lisa Over
Homework 12
April 21, 2015

CODE

#Gibbs function receives five parameters: "y" for the data values of interest, "p" for the probability that data value, yi, comes from distribution 1, "m1" and "m2" for the population means of distributions 1 and 2, "sigsq1" and "sigsq2" for the population variance of distributions 1 and 2, "N" for number of realizations, "lag" for determining how many realizations to skip between saves, and "burnin" for determining how many realizations to skip before starting to save. Gibbs generates N independent z vectors of indicator variables, N independent means from a normal distribution centered at the mean of the mixed data with variance 1/sum(z) (using the z just generated), and N independent probabilities that are each a probability of drawing from distribution 1.

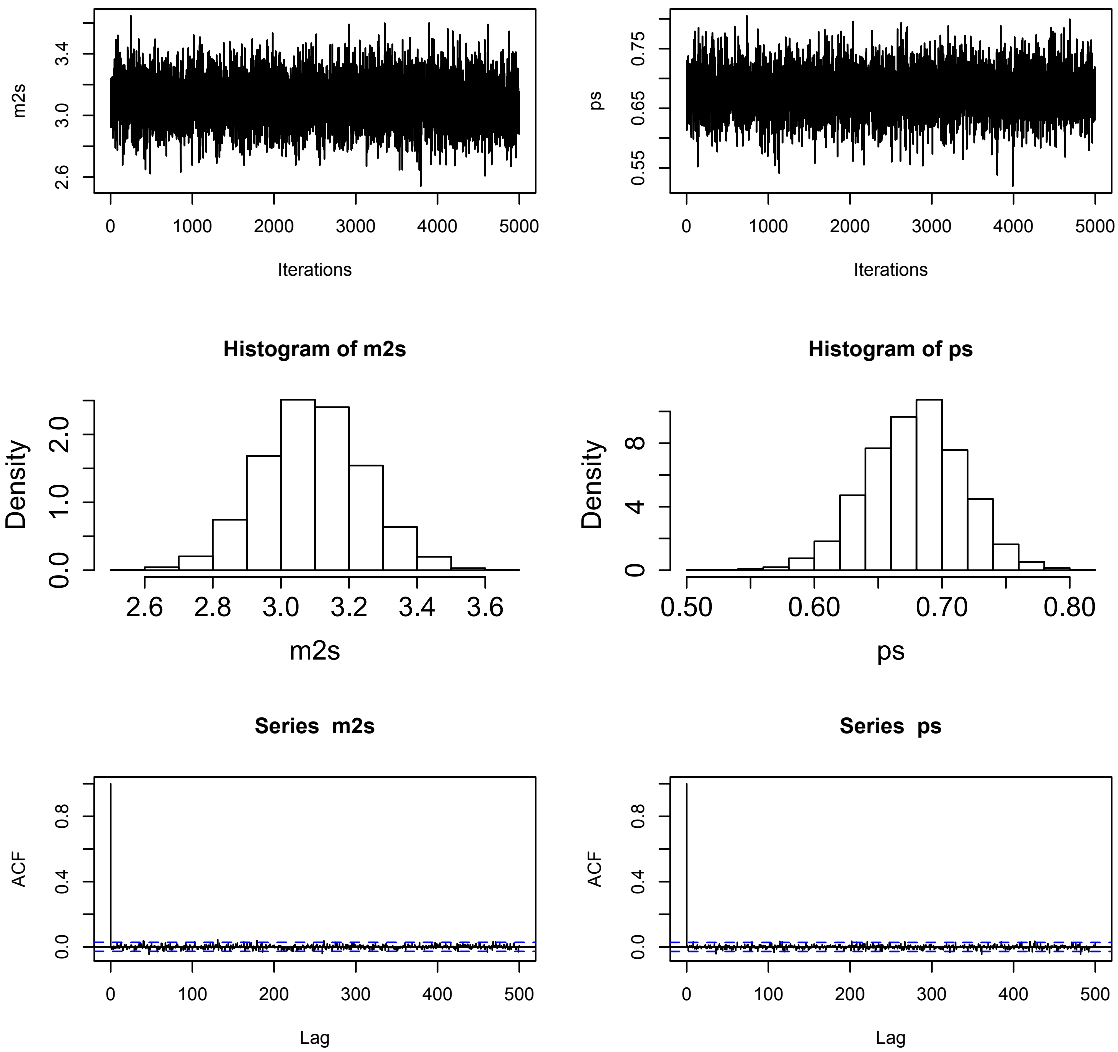
```
gibbs <- function(y,p,m1,m2,sigsq1,sigsq2,N,lag,burnin) {

        #obtain mean (y.bar)
        ybar = mean(y)
        n = length(y)

        #Set N to be N*lag+burnin
        N <- N*lag + burnin

        #Initialize vectors to hold the z vector realizations, m2 realizations, and p
realizations
        zs = NULL
        m2s = NULL
        ps = NULL

        for(i in 1:N) {

        #Calculate the vector of probabilities where each value represents the
probabiliy that the corresponding value in the data vector y was drawn from
distribution 2
        probz1 = (1-p)*exp(-(1/2)*(y-m2)^2)
        probz2 = p*exp(-(1/2)*(y)^2)
        probz = probz1 + probz2
        pvec = probz1/probz

        #Calculate the z vector of latent indicator values where each value is either 0
or 1 to indicate which distribution the corresponding value in the data vector y is
from - "0" indicates distribution 1 and "1" indicates distribution 2
        z = rbinom(n,1,pvec)
```

```
        #Compute a mean for the mixed model from a normal distribution N(y.bar,
1/sum(z)) Use the mean of the y values that were drawn from distribution 2 -
determine these y values by multiplying the y data vector by the z vector
        ybar2 = sum(y*z)/sum(z)
        m2 = rnorm(1,ybar2,sqrt(1/sum(z)))

        #Compute a p for the probability that the value comes from distribution 1
        p = rbeta(1,n-sum(z)+1,sum(z)+1)

        #if i is greater than burnin and if i is a multiple of the lag, store z, m2, and p
  if(i > burnin) {
         if(i %% lag == 0) {
                zs <- c(zs,z)
                m2s <- c(m2s,m2)
                ps <- c(ps,p)
   }
   }
}

  vectors <- list("zs" = zs, "m2s" = m2s, "ps" = ps)
  return(vectors)

}

#Set burnin=0 and leave lag=80 and run Gibbs with N=5000
N = 5000
lag = 80
burnin = 0

#Out of n trials with probability of success p, draw a random binomial variable that
represents the number of Yi values that are to be drawn from N(mu1,sigsq1).
n = 200
p = 0.7
k = rbinom(1, n, p)

m1 = 0
m2 = 3
sigsq1 = 1
sigsq2 = 1

#Generate the data with k values being drawn from distribution 1 and n - k values
being generated from distribution 2
y = NULL
for(i in 1:n) {
```

```
        if(i <= k) {
                y = c(y, rnorm(1,m1,sqrt(sigsq1)))
        }
        else y = c(y, rnorm(1,m2,sqrt(sigsq2)))
}


vectors = gibbs(y,p,m1,m2,sigsq1,sigsq2,N,lag,burnin)

#vectors$zbars represents Gibbs sampler realizations for the mean of the z vector
#vectors$m2s represents Gibbs sampler realizations for the mean of distribution 2
#vectors$ps represents Gibbs sampler realizations for the probability that a y value
comes from distribution 1
zs = vectors$zs
m2s = vectors$m2s
ps = vectors$ps

par(mfrow=c(3,2)) #split plotting window into 2 rows and 2 columns
ts.plot(m2s,xlab="Iterations")
ts.plot(ps,xlab="Iterations")
hist(m2s,probability=T, cex.lab=1.5, cex.axis=1.5)
hist(ps,probability=T, cex.lab=1.5, cex.axis=1.5)
acf(m2s,lag.max=500)
acf(ps, lag.max=500)

#Convert z to an n column matrix
zmat = matrix(zs, ncol=n, byrow=TRUE)
#Get column means of matrix zmat
zimeans = colMeans(zmat)

plot(y,zimeans)

mean(m2s)
#[1] 0.8351638
mean(ps)
#[1] 0.3077043
```

Lisa Over
Homework 12
April 21, 2015

RESULTS



In Figure 1 below, the proportion of 1s for each y_i is plotted against the corresponding y_i from the 5000 draws. The proportion of 1s indicates the probability that the corresponding value was drawn from distribution 2. The y values that could only be drawn from distribution 1 are those forming a horizontal line at z=0. The y values that could only be drawn from distribution 2 are those forming a horizontal line at z=1. The y values that correspond to $0 < z < 1$ are those that could have been drawn from either distribution 1 or 2. There is a positive relationship between the proportion of 1s and the y values. As the y values increase, they become more likely to have come from distribution 2.

Lisa Over
Homework 12
April 21, 2015

The mean of the distribution of m2 values is 3.09. The mean of the distribution of p values is 0.68. This is consistent with the mean of distribution 2, which was 3, and with the probability that a data value comes from distribution 1, which is 0.7.

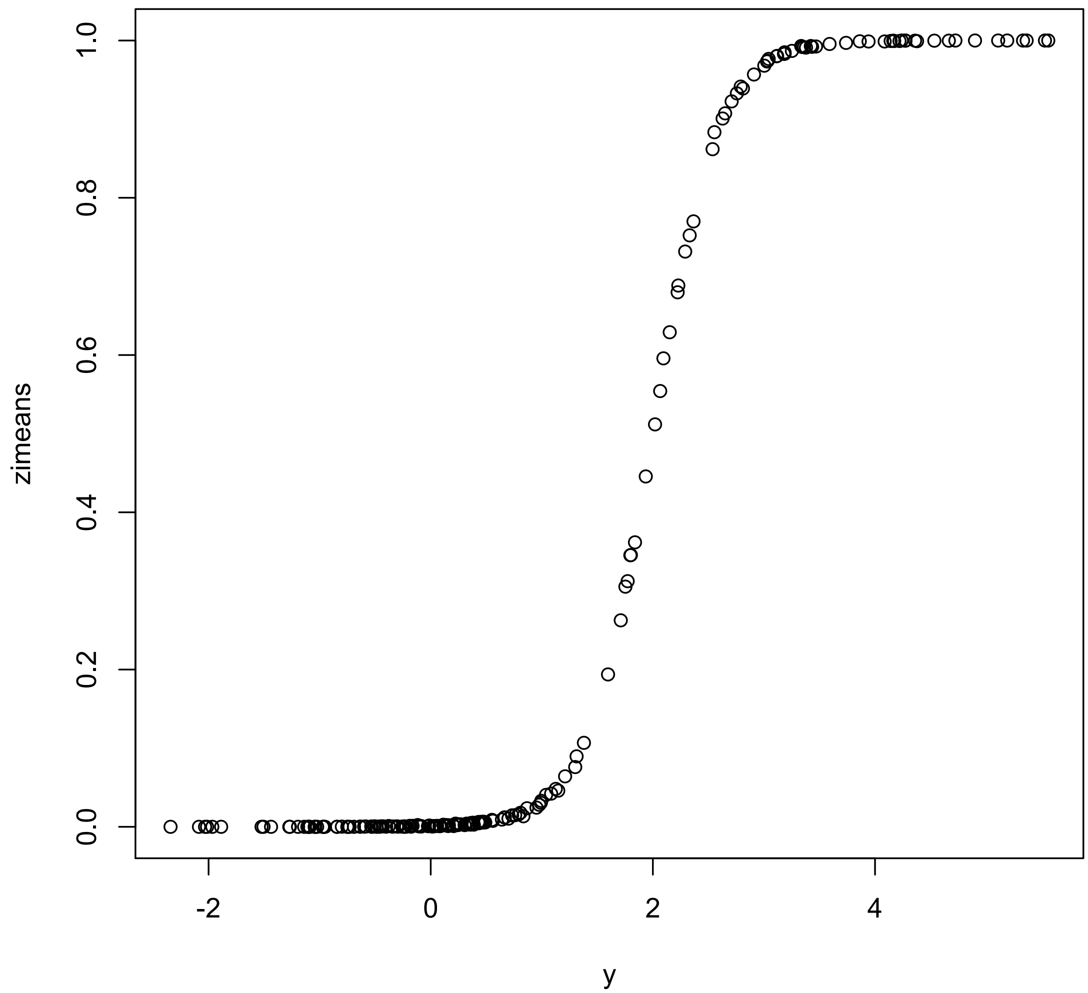


Figure 1 Scatterplot of proportion of 1s plotted against corresponding y_i.