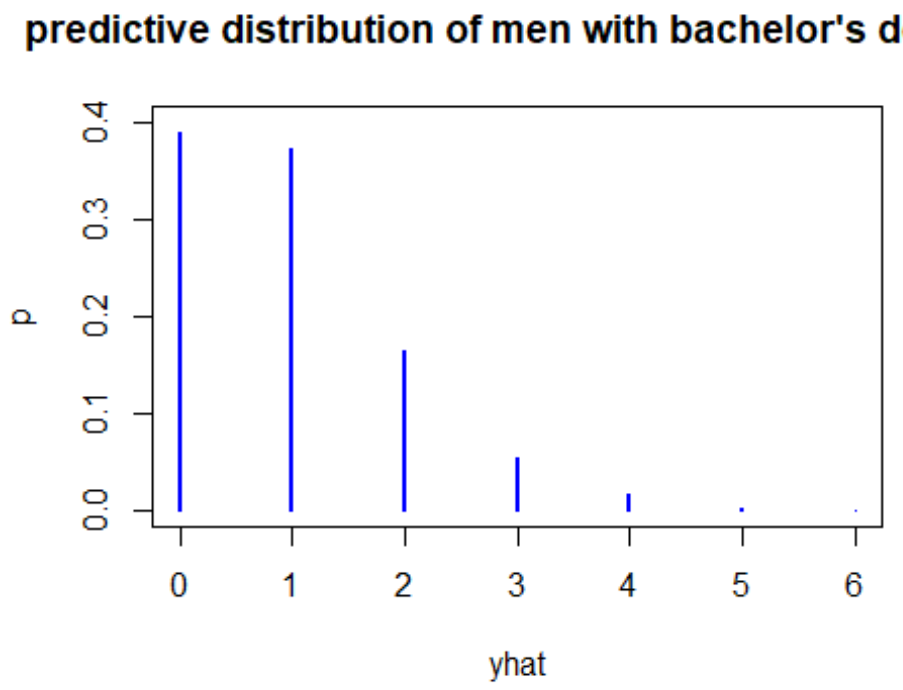
HW3 of Bayesian Statistics

Guanren Wang

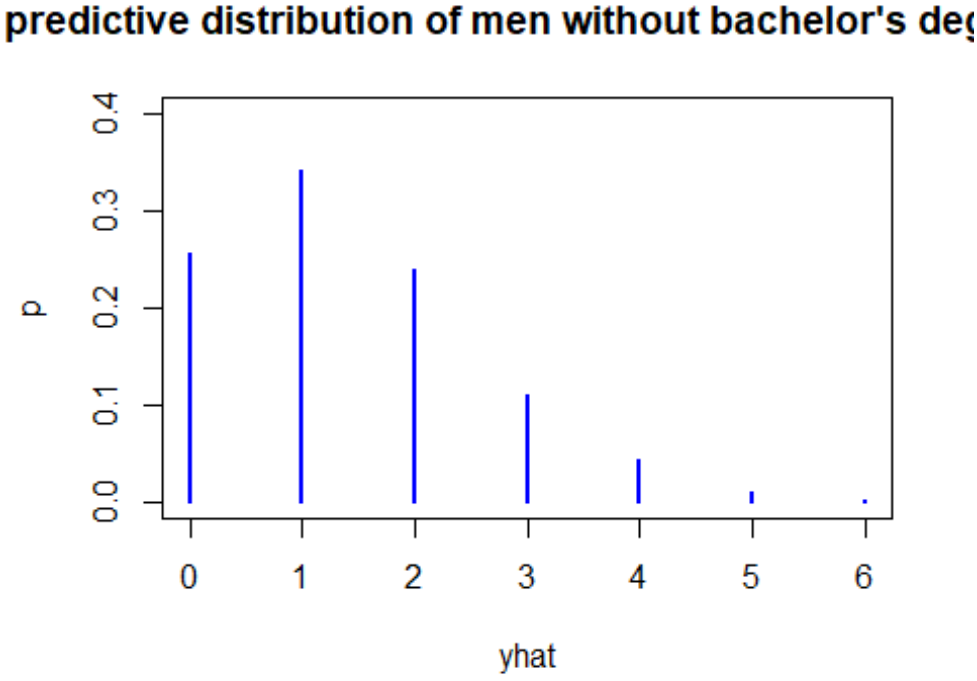
2019-3-22

1.(a).

```
library(tidyverse)

bach<-read.delim('./men30bach.txt',sep=' ',header = F)
bach<-t(bach)
nobach<-read.delim('./men30nobach.txt',sep = ' ',header = F)
nobach<-t(nobach)

n=5000
size.a=2+sum(bach)
size.b=2+sum(nobach)
mu.a=size.a/(1+length(bach))
mu.b=size.b/(1+length(nobach))
set.seed(123)
yhat.a<-rnbinom(n,size.a,mu=mu.a)%>%as.factor()
yhat.b<-rnbinom(n,size.b,mu=mu.b)%>%as.factor()

a<-data.frame(yhat.a)%>%group_by(yhat.a)%>%count(yhat.a)%>%mutate(p=n/5
000)%>%as.matrix()
b<-data.frame(yhat.b)%>%group_by(yhat.b)%>%count(yhat.b)%>%mutate(p=n/5
000)%>%as.matrix()

plot(a[,1],a[,3],type="h",ylim=c(0,0.4),lwd=2,col="blue",ylab="p",xlab
= 'yhat',main = "predictive distribution of men with bachelor's degree
")
```




```
plot(b[,1],b[,3],type = 'h',ylim =c(0,0.4),lwd=2,col='blue',ylab='p',xl
ab = 'yhat',main = "predictive distribution of men without bachelor's d
egree")
```



1.(b).

```
set.seed(123)
theta.a<-rgamma(n,size.a,rate = 1+length(bach))
theta.b<-rgamma(n,size.b,rate=1+length(nobach))
yhat.a<-as.numeric(yhat.a)
yhat.b<-as.numeric(yhat.b)
quantile(theta.b-theta.a,c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.1559684 0.7359099
```

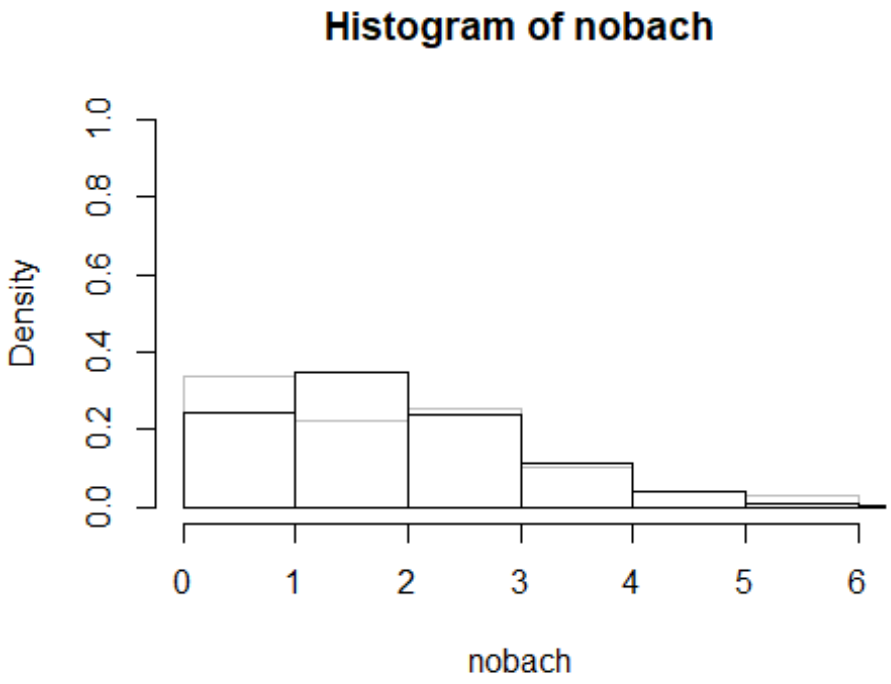
```
quantile(yhat.b-yhat.a,c(0.025,0.975))
```

```
##  2.5% 97.5%
##    -3     4
```

Apparently, men in their 30s with bachelor's degrees significantly have less childre, which can be proved by both histograms and quantiles. According to the histograms, preditive distribution of the children of men with bachelor's degree concentrate on 0 and 1, and the peak is at 0. In contrast, preditive distribution of the children of men without bachelor's degree reaches its maximum at 1, and dispersed more evenly. From the quantile we could find that lower bound of 95% CI of thetaB-thetaA is greater than 0, indicating thetaB is significantly greater than thetaA.

1.(c).

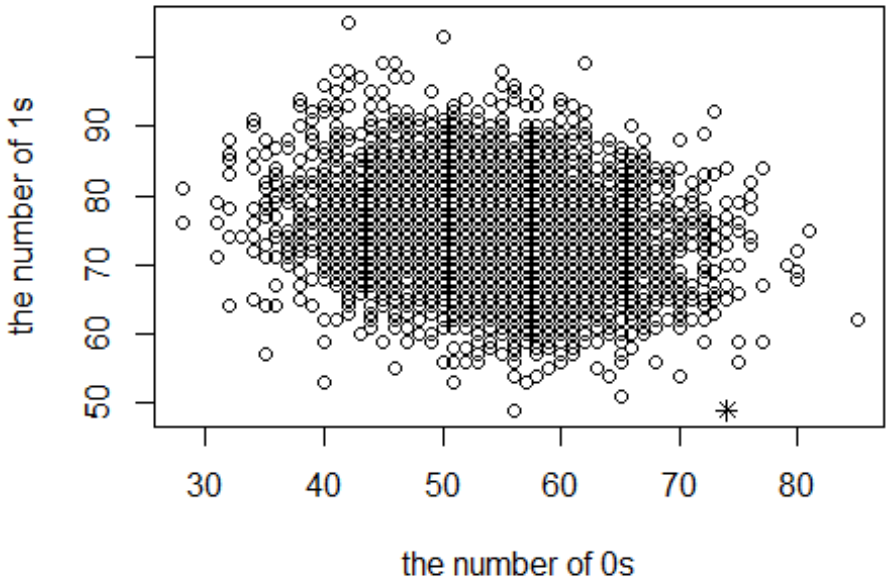
```
hist(nobach,breaks = 6,freq = F,ylim = 0:1,border = 'grey',right = F)
set.seed(123)
hist(rpois(100000,lambda = 1.4),breaks = 8,freq = F,ylim = 0:1,main = '
Histogram of theoretical poisson distribution',add=T,right = F)
```



The grey line is empirical one, the black line is theoretical one. Obviously, Poisson model fits not very well because the distributions of theoretical one reaches its peak at 1, while empirical one maximizes at 0.

1.(d).

```
zeros<-c()
ones<-c()
for (i in 1:length(theta.b)) {
  set.seed(i)
  sam<-rpois(218,theta.b[i])
  zeros[i]<-sum(sam==0)
  ones[i]<-sum(sam==1)
}
emp0<-sum(nobach==0)
emp1<-sum(nobach==1)
plot(zeros, ones, xlab="the number of 0s", ylab="the number of 1s")
points(emp0, emp1, pch=8)
```



According to the plot, the observation of data (marked as star sign) is far away from both population, meaning inadequacy of Poisson model.

2.(a).

The results are in the last page.

2.(b).

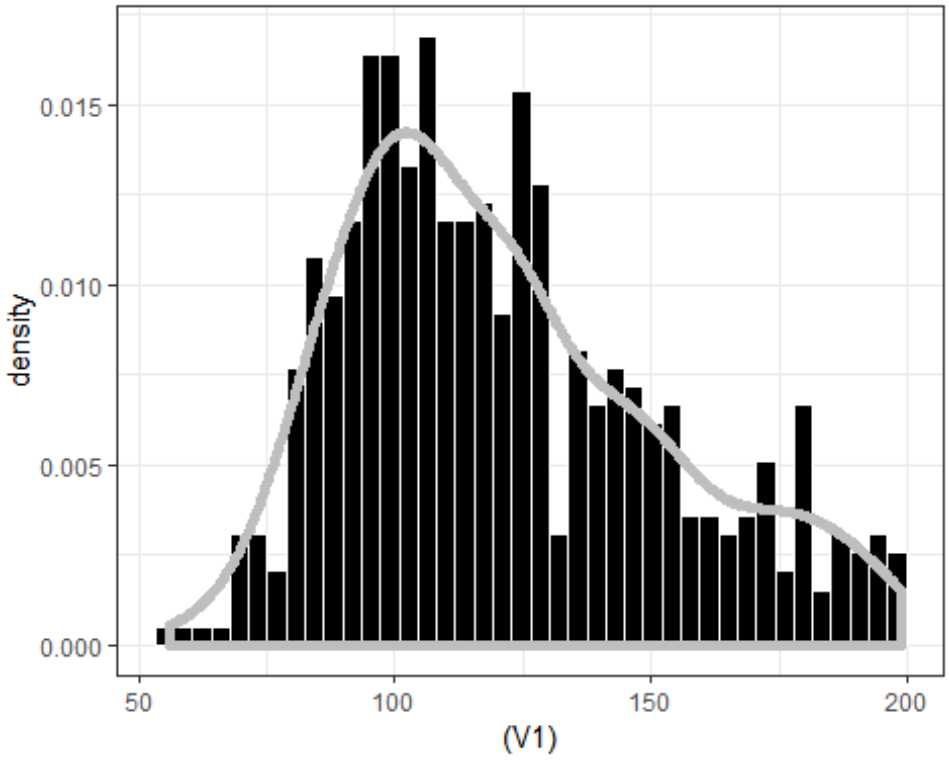
```
at=2
bt=1
ab=c(8, 16, 32, 64,128)
#starting values
theta<-mean(bach)
thetaB<-mean(nobach)
gamma<-thetaB/theta

s=10000
mtheta<-matrix(nrow = s,ncol = length(ab))
mgamma<-matrix(nrow = s,ncol = length(ab))
mtheta[1,]<-rep(theta,length(ab))
mgamma[1,]<-rep(gamma,length(ab))

set.seed(123)
for (i in ab) {
```

```
  for (j in 2:s) {
    mtheta[j,log2(i/4)]<-rgamma(n=1,shape = sum(bach)+sum(nobach)+at,ra
te = length(bach)+length(nobach)*mgamma[j-1,log2(i/4)]+bt)
    mgamma[j,log2(i/4)]<-rgamma(n=1,shape = sum(nobach)+i,rate = length
(nobach)*mtheta[j-1,log2(i/4)]+i)
  }
}
sim_thetaB<-mtheta*mgamma
sim_thetaA<-mtheta
apply(sim_thetaB-sim_thetaA,2,mean)

## [1] 0.3923780 0.3574256 0.2810234 0.2089362 0.1378038
```

Given a.gamma=b.gamma, as the parameter pair of prior distribution of gamma become greater, the expected difference between posterior thetaB and posterior thetaA become smaller and smaller.

3.(a).

```
dt<-read.delim('./glucose.txt',header = F)
ggplot(dt,aes((V1)))+geom_histogram(aes(y=..density..),bins = 40,fill='
black',color='white')+theme_bw()+geom_density(color='grey',size=2)
```



This empirical distribution is asymmetrical and positively skewed compared with norman distribution.

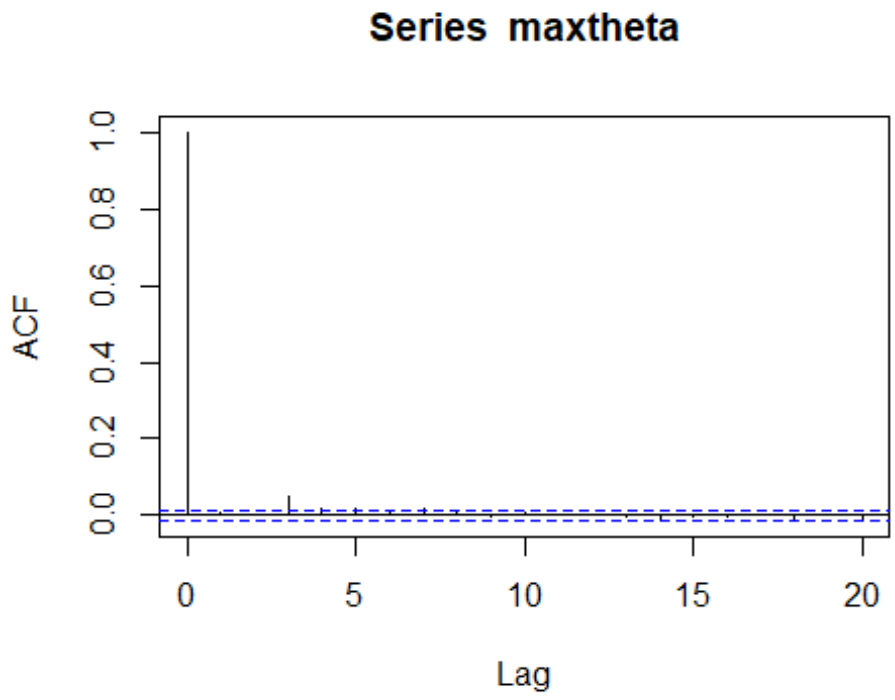
3.(b).

The results are in the last page.

3.(c).

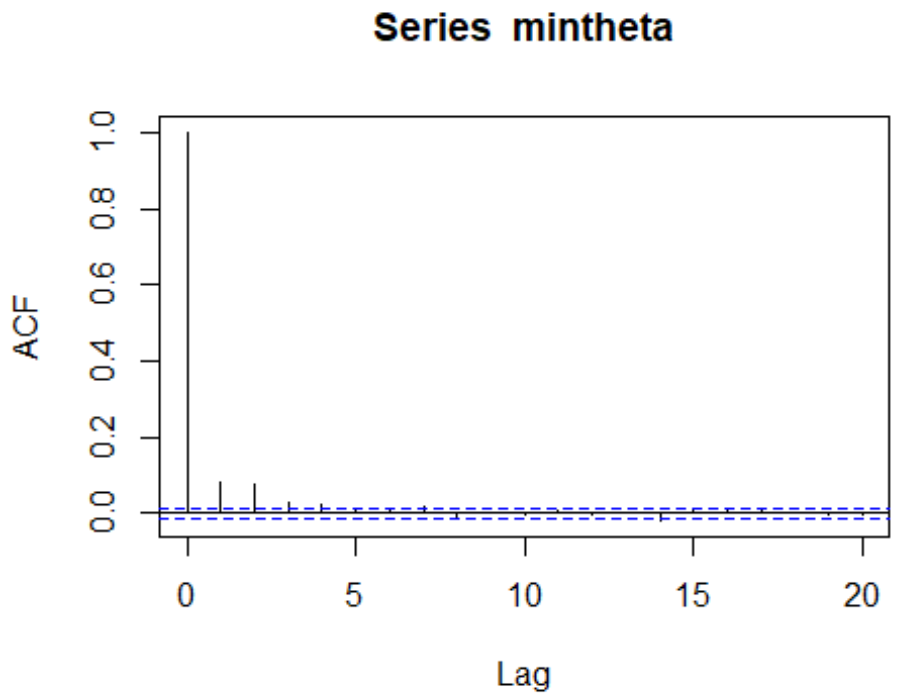
```
a=1
b=1
mu0=120
tau0_square=200
sigma0_square=1000
nu0=10
p=0.5 #when a=b=1,p~uniform(0,1), so I use expectation as initial value
n=20000 # number of iterations
obs=532

# restore original data
set.seed(123)
x<-rbinom(obs,size = 1,prob = p)+1

#gibbs sampler
theta<-matrix(ncol = 2,nrow = n)
sigma<-matrix(ncol = 2,nrow = n)

#set initial values
index1<-x==1
index2<-x==2
n1=sum(index1)
n2=obs-n1
y<-dt$V1
y1.bar<-mean(y[index1])
y2.bar<-mean(y[index2])
theta[1,1]<-mean(y[index1])
theta[1,2]<-mean(y[index2])
sigma[1,1]<-var(y[index1])
sigma[1,2]<-var(y[index2])

#start sampling
set.seed(123)
for (i in 1:(n-1)) {
  #sample p and x
  p<-c(p,rbeta(n=1,shape1 = a+n1,shape2 = b+n2))
  tempp1<-p[i+1]*dnorm(y,theta[i,1],sqrt(sigma[i,1]))
  tempp2<-(1-p[i+1])*dnorm(y,theta[i,2],sqrt(sigma[i,2]))
  px<-tempp1/(tempp1+tempp2)
  x<-rbinom(obs,1,px)+1

  #calculate n1, n2 and index
  index1<-x==1
```

```
  index2<-x==2
  n1=sum(index1)
  n2=obs-n1

  #xi=1
  mun1=(mu0/tau0_square+sum(y[index1])/sigma[i,1])/(1/tau0_square+n1/si
gma[i,1])
  taun1_square=1/(1/tau0_square+n1/sigma[i,1])
  theta[i+1,1]=rnorm(n=1,mean = mun1,sd= sqrt(taun1_square))

  nun1=nu0+n1
  sigman1_square=(nu0*sigma0_square+sum(y[index1]-theta[i,1])^2)/nun1
  sigma[i+1,1]=1/rgamma(n=1,shape = nun1/2,rate=(nu0*sigma0_square+nun1
*sigman1_square)/2)

  #xi=2
  mun2=(mu0/tau0_square+sum(y[index2])/sigma[i,2])/(1/tau0_square+n2/si
gma[i,2])
  taun2_square=1/(1/tau0_square+n2/sigma[i,2])
  theta[i+1,2]=rnorm(n=1,mean = mun2,sd= sqrt(taun2_square))

  nun2=nu0+n2
  sigman2_square=(nu0*sigma0_square+sum(y[index2]-theta[i,2])^2)/nun2
  sigma[i+1,2]=1/rgamma(n=1,shape = nun2/2,rate=(nu0*sigma0_square+nun2
*sigman2_square)/2)
}

maxtheta<-apply(theta,1,max)
mintheta<-apply(theta,1,min)
paste('The acf of maxtheta with lag',1:20,'is',acf(maxtheta,lag.max = 2
0)$acf[2:21])
```




```
##  [1] "The acf of maxtheta with lag 1 is 0.00714189902052996"
##  [2] "The acf of maxtheta with lag 2 is 0.00106108898598822"
##  [3] "The acf of maxtheta with lag 3 is 0.0512818253190321"
##  [4] "The acf of maxtheta with lag 4 is 0.0199075259701624"
##  [5] "The acf of maxtheta with lag 5 is 0.0150406643054137"
##  [6] "The acf of maxtheta with lag 6 is 0.00798117613277043"
##  [7] "The acf of maxtheta with lag 7 is 0.014733731885002"
##  [8] "The acf of maxtheta with lag 8 is 0.00545765425680194"
##  [9] "The acf of maxtheta with lag 9 is -0.00375015150509708"
## [10] "The acf of maxtheta with lag 10 is 0.00536132382417329"
## [11] "The acf of maxtheta with lag 11 is 2.85463513362642e-05"
## [12] "The acf of maxtheta with lag 12 is -0.000324125422246608"
## [13] "The acf of maxtheta with lag 13 is -0.00461453553466223"
## [14] "The acf of maxtheta with lag 14 is -0.00740303791475685"
## [15] "The acf of maxtheta with lag 15 is -0.00407982767612549"
## [16] "The acf of maxtheta with lag 16 is -0.0028294670528637"
## [17] "The acf of maxtheta with lag 17 is 0.000281424089367936"
## [18] "The acf of maxtheta with lag 18 is -0.0130905295358874"
## [19] "The acf of maxtheta with lag 19 is 0.0034160045629399"
## [20] "The acf of maxtheta with lag 20 is -0.00692608159061039"
```

```
paste('The acf of mintheta with lag',1:20,'is',acf(mintheta,lag.max = 2
0)$acf[2:21])
```




```
##  [1] "The acf of mintheta with lag 1 is 0.0785173122654554"
##  [2] "The acf of mintheta with lag 2 is 0.0753425418336816"
##  [3] "The acf of mintheta with lag 3 is 0.0250431347498966"
##  [4] "The acf of mintheta with lag 4 is 0.019795429120294"
##  [5] "The acf of mintheta with lag 5 is 0.00911023809433694"
##  [6] "The acf of mintheta with lag 6 is 0.0108424219159492"
##  [7] "The acf of mintheta with lag 7 is 0.0159325725024293"
##  [8] "The acf of mintheta with lag 8 is -0.0120964999294805"
##  [9] "The acf of mintheta with lag 9 is 0.00026227815132063"
## [10] "The acf of mintheta with lag 10 is -0.00336532699271723"
## [11] "The acf of mintheta with lag 11 is 0.00419508649756294"
## [12] "The acf of mintheta with lag 12 is -0.00406040245461257"
## [13] "The acf of mintheta with lag 13 is -0.00078380236202917"
## [14] "The acf of mintheta with lag 14 is -0.0197607016730424"
## [15] "The acf of mintheta with lag 15 is 0.00418380190592054"
## [16] "The acf of mintheta with lag 16 is 0.012986631995515"
## [17] "The acf of mintheta with lag 17 is 0.00615016253543692"
## [18] "The acf of mintheta with lag 18 is -0.00018225051620553"
## [19] "The acf of mintheta with lag 19 is -0.00696344015054308"
## [20] "The acf of mintheta with lag 20 is -0.00356427730463904"
```

```
library(coda)
paste('effective sample size of maxtheta is:',effectiveSize(maxtheta))
```

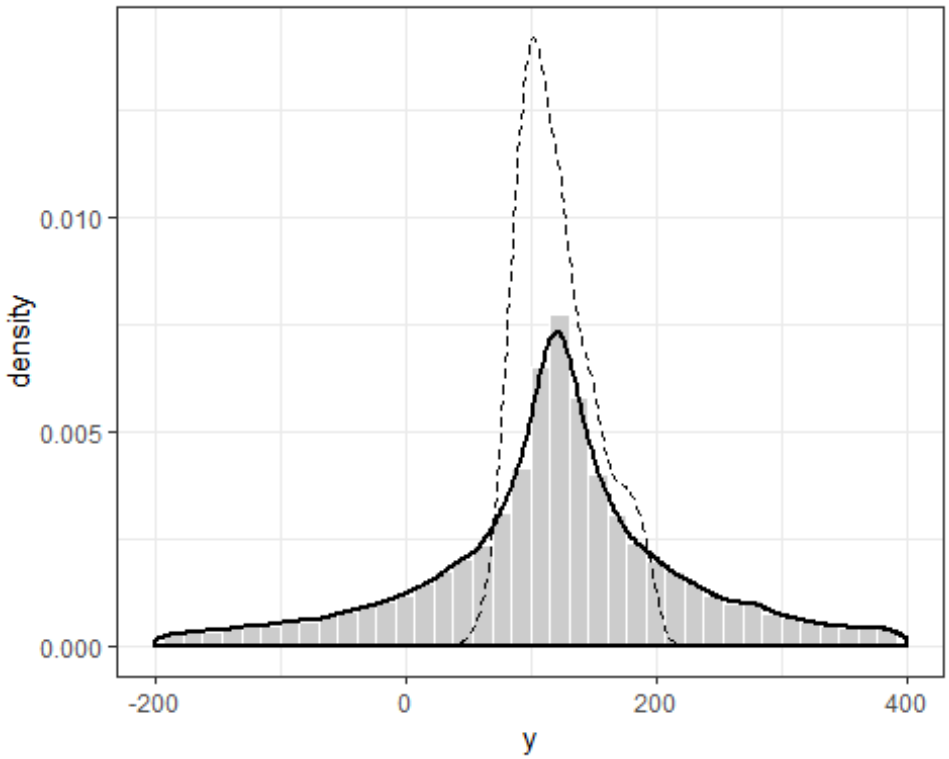
```
## [1] "effective sample size of maxtheta is: 16586.7331872733"
```

```
paste('effective sample size of mintheta is:',effectiveSize(mintheta))
```

```
## [1] "effective sample size of mintheta is: 13859.0231512964"
```

3.(d).

```
set.seed(125)
x<-rbinom(n,1,p)+1
y1<-rnorm(sum(x==1),mean = theta[x==1,1],sd=sqrt(sigma[x==1,1]))
y2<-rnorm(sum(x==2),mean = theta[x==2,2],sd=sqrt(sigma[x==2,2]))
y<-data.frame(V1=c(y1,y2))
ggplot(y,aes((V1)))+geom_histogram(aes(y=..density..),bins = 40,fill='g
rey80',color='white',show.legend = T)+theme_bw()+geom_density(color='bl
ack',size=1)+xlim(-200,400)+geom_density(data = dt,aes(V1),linetype=2,s
how.legend = T)+xlab('y')
```



The dotted line in the plot is density curve of data. The histogram and solid line are from predicted data. It seems the mixture model captures the centrality of empirical distribution but in a flattened shape. So the adequacy of mixture model isn't perfect.