Assignment2Q3

Semester 2, 2016

Benjamin G. Moran; c3076448@uon.edu.au

22nd August 2016

Question 3 [Total: 20 marks]

Consider the example on corn yields but with an extended data set, with sample sizes, sample means and sample variances as given below. We have five different corn growers giving individual yields for a particular new type of genetically engineered corn that has three seasons per year. The corn is distributed by a research station to the growers. Growers give yields for the most recent seasons in tons/hectare. Summary statistics for the data are given below.

Table 1: Corn Yield Data by Grower

	1	2	3	4	5
n	16.0	19.0	14.0	12.0	8.0
y	15.3	16.2	16.4	13.2	13.5
s^2	8.2	12.3	7.9	5.2	6.2

(a) By adapting code from the week 7 lecture, fit a hierarchical Normal model to this data, using the Empirical Bayes approach. Provide estimates and inference for each unknown true grower mean yield, including a plot showing all estimated posterior densities, one for each unknown mean. [7 marks]

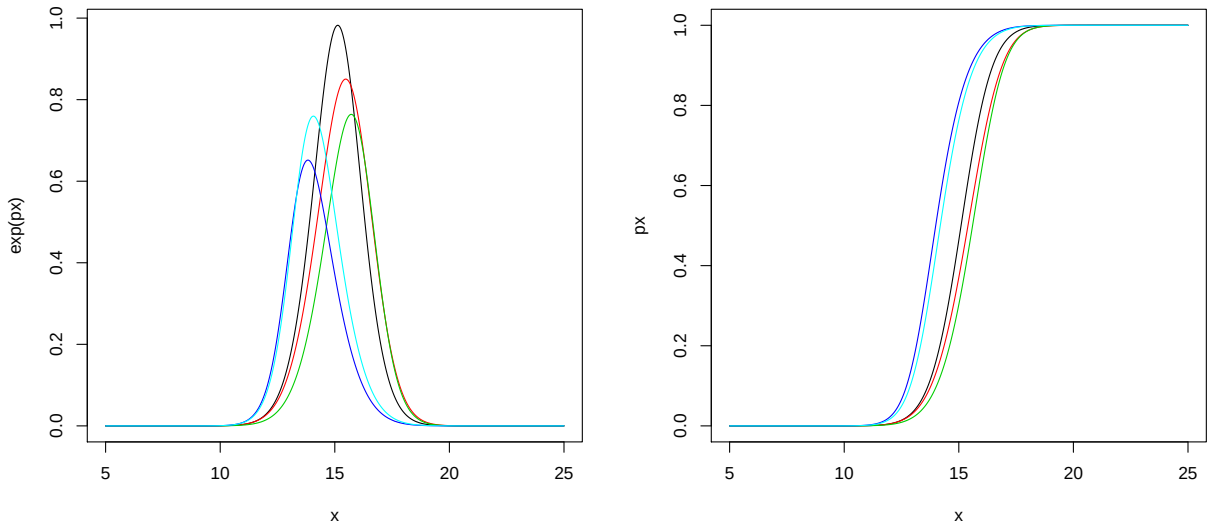
Answer: The lecture material presents us with 3 possible models, however only one is represented by the code in those same notes: the model that makes no assumptions about individual grower variances (3). So, the model we must fit to our data is a non-standard posterior distribution that is constructed by multiplying a Normal prior with the likelihood for μ_j in the form of a t-density.

$$\begin{aligned} p(\mu_j \mid y_j, \mu, \tau^2) &\propto p(y_j \mid \mu_j, \mu, \tau^2) p(\mu_j \mid \mu, \tau^2) \\ &= p(\mu_j \mid \mu, \tau^2) \int p(\mu_j \mid y_j, \sigma_j^2, \mu, \tau^2)(\sigma_j^2 \mid \mu, \tau^2) \\ &\propto p(N(\mu, \tau^2)) p(t_{n_j - 1}(\bar{y}_j, s_j^2)/n_j)) \end{aligned}$$

For the choice of hyperparameters μ and τ, we again proceed in line with the notes and calculate the average yield across all growers, along with the standard deviation of that average yield. This gives us our empirical priors, which we can use to perform the estimation. Plot

```
x <- seq(5, 25, by=0.001)
mu <- mean(y.bar) # 14.92
tau <- sqrt(var(y.bar)) # 1.49566...
par(mfrow = c(1,2))
pdf.data <- matrix(nrow=5,ncol=20001)
for(k in 1:5){
  px <- -(x-mu)^2/(2*tau^2)-3/2*log(1+(3*(x-y.bar[k])^2)/(2*sig2[k]))
  pdf.data[k,] <- px
```

```
  if(k==1){
    plot(x,exp(px),type='l',col=k)
  } else {
    lines(x,exp(px),type='l', col = k)
  }
}
cdf.data <- matrix(nrow=5,ncol=20001)
for(k in 1:5) {
  px <- -(x-mu)^2/(2*tau^2) -3/2*log(1+(3*(x-y.bar[k])^2)/(2*sig2[k]))
  cdf.data[k,] <- exp(px)/sum(exp(px))
}

for(k in 1:5) {
  px <- cumsum(cdf.data[k,])
  if(k==1) {
    plot(x,px,type='l', col=k)
  } else {
    lines(x,px,type='l', col = k)
  }
}
```



We can generate MCMC samples using the following code (Note: I have not included the code for each grower to save space):

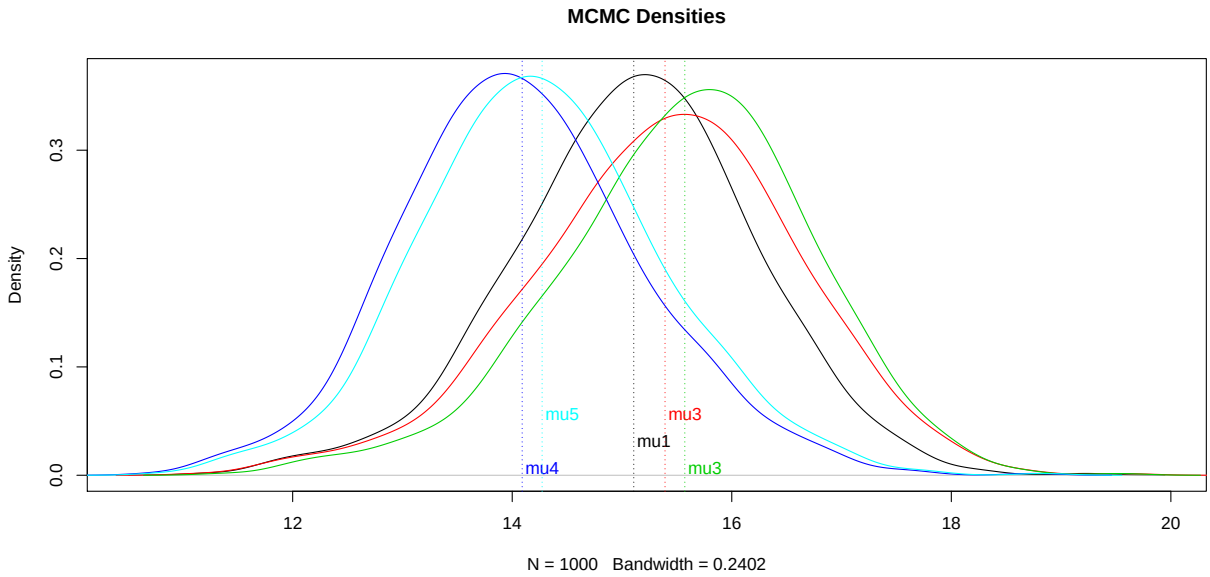
```
cdf1 <- cumsum(cdf.data[1,])
u=runif(1000,0,1)
mu1=0

for(j in 1:1000){
 st=0;k=1
 while(st==0) {
```

```
  if(cdf1[k]<u[j]&cdf1[k+1]>u[j]) {
    mu1[j] <- x[k]
    st <- 1
  } else {
    k <- k+1
    }
  }
}
```

Now, using our MCMC samples, we can calculate the posterior mean and CrI for each grower:

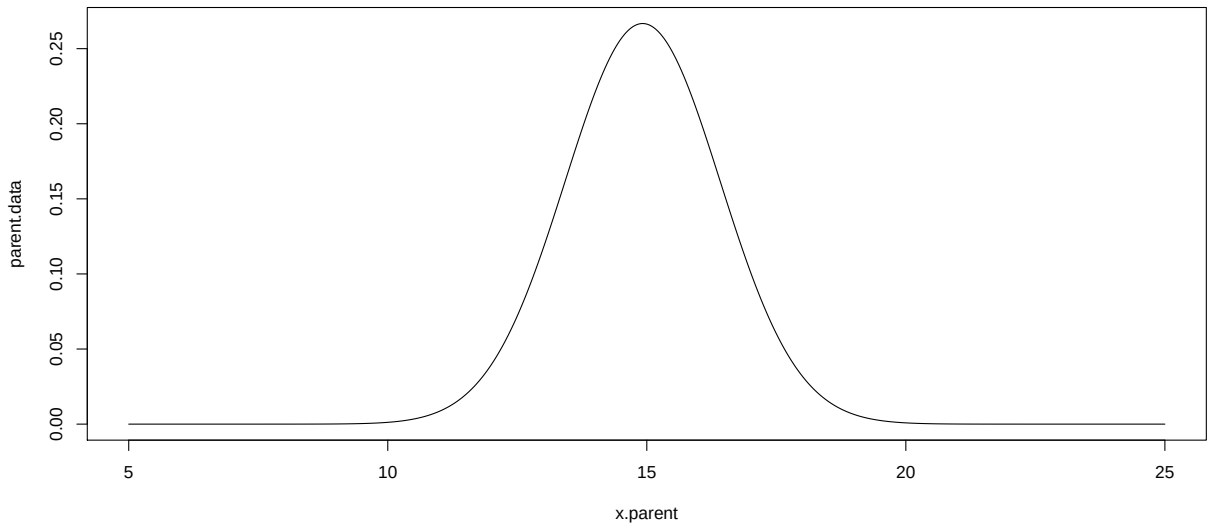
Grower	Mean	2.5%	97.5%
1	15.10747	12.75100	17.22105
2	15.39207	12.79385	17.63405
3	15.57251	13.03083	17.66505
4	14.09053	11.99640	16.41408
5	14.27223	12.12133	16.56005

This suggests an order (from the grower with the largest mean output to the lowest) of $3, 2, 1, 5, 4$. Plotting the MCMC sample densities reaffirms this idea:



(b) Plot the estimated parent density for the five grower means. At what percentile of this distribution does of each grower's estimated posterior mean yield lie? Do they all fit in with the prior distribution? [2 marks]

```
x.parent <- seq(5,25,by=0.001)
parent.data <- dnorm(x.parent,mu,tau)
plot(x.parent,parent.data, type = 'l')
```



We can easily calculate the corresponding percentile of the parent distribution for each posterior mean by:

```
# Grower 1
mu1.p <- round(mean(x.parent<mean(mu1))*100,2)
# Grower 2
mu2.p <- round(mean(x.parent<mean(mu2))*100,2)
# Grower 3
mu3.p <- round(mean(x.parent<mean(mu3))*100,2)
# Grower 4
mu4.p <- round(mean(x.parent<mean(mu4))*100,2)
# Grower 5
mu5.p <- round(mean(x.parent<mean(mu5))*100,2)
```

This tells us that the posterior means for Growers 1 through 5 are:

Grower	1	2	3	4	5
Percentile	50.54	51.96	52.86	45.45	46.36

(c) What is the posterior probability that each of the five growers has the largest mean yield? What is your conclusion in this case? [2 marks]

We can calculate these probabilities fairly simply, using the MCMC estimates for the posterior means of each grower and the following code:

```
# Grower 1
length(mu1[mu1>mu2&mu1>mu3&mu1>mu4&mu1>mu5])/n
## [1] 0
# Grower 2
length(mu2[mu2>mu1&mu2>mu3&mu2>mu4&mu2>mu5])/n
## [1] 0.005
```

```
# Grower 3
length(mu3[mu3>mu1&mu3>mu2&mu3>mu4&mu3>mu5])/n
## [1] 0.995
# Grower 4
length(mu4[mu4>mu1&mu4>mu2&mu4>mu3&mu4>mu5])/n
## [1] 0
# Grower 5
length(mu5[mu5>mu1&mu5>mu2&mu5>mu3&mu5>mu4])/n
## [1] 0
```

This tells us that, using our estimates, Grower 3 has the largest mean yield with a probability of $\approx 99.5\%$.

(d) What is the posterior probability that, for each pair of growers (grower i and i+k), the mean yield for grower(i) exceeds the mean yield for grower (i+k)? [2 marks]

```
# Probability that mean of Grower A (Row) is greater than mean of Grower B (Column)
colnames(probs) <- c("1","2","3","4","5")
rownames(probs) <- c("1","2","3","4","5")
probs.m <- melt(probs)

probs.m <- as.data.frame(probs.m)
names(probs.m) <- c("GrowerA", "GrowerB", "Pr")
probs.m <- probs.m %>%
  arrange(GrowerA,desc(Pr))

kable(probs.m, row.names = FALSE, col.names = c("Grower A", "Grower B", "Pr(A>B)"))
```

Grower A	Grower B	Pr(A>B)
1	4	1.000
1	5	1.000
1	2	0.009
1	1	0.000
1	3	0.000
2	4	1.000
2	5	1.000
2	1	0.991
2	3	0.005
2	2	0.000
3	1	1.000
3	4	1.000
3	5	1.000
3	2	0.995
3	3	0.000
4	1	0.000
4	2	0.000
4	3	0.000
4	4	0.000
4	5	0.000
5	4	1.000
5	1	0.000
5	2	0.000
5	3	0.000
5	5	0.000

```
length(mu3[mu3<mu2&mu3>mu1&mu3>mu5&mu3>mu4])/n
## [1] 0.005
length(mu2[mu2<mu3&mu2<mu1&mu2>mu5&mu2>mu4])/n
## [1] 0.009
length(mu1[mu1<mu3&mu1>mu2&mu1>mu5&mu1>mu4])/n
## [1] 0.009
```

```
length(mu3[mu3>mu2&mu3>mu1&mu3>mu5&mu3>mu4])/n
## [1] 0.995
length(mu2[mu2<mu3&mu2>mu1&mu2>mu5&mu2>mu4])/n
## [1] 0.986
length(mu1[mu1<mu3&mu1<mu2&mu1>mu5&mu1>mu4])/n
## [1] 0.991
length(mu5[mu5<mu3&mu5<mu2&mu5<mu1&mu5>mu4])/n
## [1] 1
length(mu4[mu4<mu3&mu4<mu2&mu4<mu1&mu4<mu5])/n
## [1] 1
length(mu3[mu3<mu2&mu3>mu1&mu3>mu5&mu3>mu4])/n
## [1] 0.005
length(mu2[mu2<mu3&mu2<mu1&mu2>mu5&mu2>mu4])/n
## [1] 0.009
length(mu1[mu1<mu3&mu1>mu2&mu1>mu5&mu1>mu4])/n
## [1] 0.009
```

(e) What is the most likely ordering for the 5 mean grower yields in your MC sample? What are the 2nd, 3rd and 4th most likely orderings? [4 marks]

```
length(mu3[mu3>mu2&mu3>mu1&mu3>mu5&mu3>mu4])/n
## [1] 0.995
length(mu2[mu2<mu3&mu2>mu1&mu2>mu5&mu2>mu4])/n
## [1] 0.986
length(mu1[mu1<mu3&mu1<mu2&mu1>mu5&mu1>mu4])/n
## [1] 0.991
length(mu5[mu5<mu3&mu5<mu2&mu5<mu1&mu5>mu4])/n
## [1] 1
length(mu4[mu4<mu3&mu4<mu2&mu4<mu1&mu4<mu5])/n
## [1] 1
```

(f) Based on (a)-(d), and any other evidence you can provide, where do you believe significant differences exist among the 5 grower means (if any)? [3 marks]

End of Assignment 2