HW6

Andrew Liang

11/4/2020

```r
library(MASS)
library(matlib)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```r
library(mvtnorm)
```

7.4

a

Intuitively, most married couples are probably between the ages of 25 to 80. I will set $\theta_0 = (\frac{25+80}{2}, \frac{25+80}{2}) = (52.5, 52.5)^T$.

Consequently, I will pick a semiconjugate prior distribution. It seems more likely that there will be more married couples around age 50, justifying a bell curve centered around 52.5 with variance $13.75^2 = 189$ such that 95% of my prior is within $(25, 80)$

Additionally, I believe that ages of couples are tightly correlated, and thus aim for a prior correlation of around 0.75. So the covariance between the two variables would be:

$$0.75 = \frac{\sigma_{1,2}}{189}$$
$$\sigma_{1,2} = 141.75$$

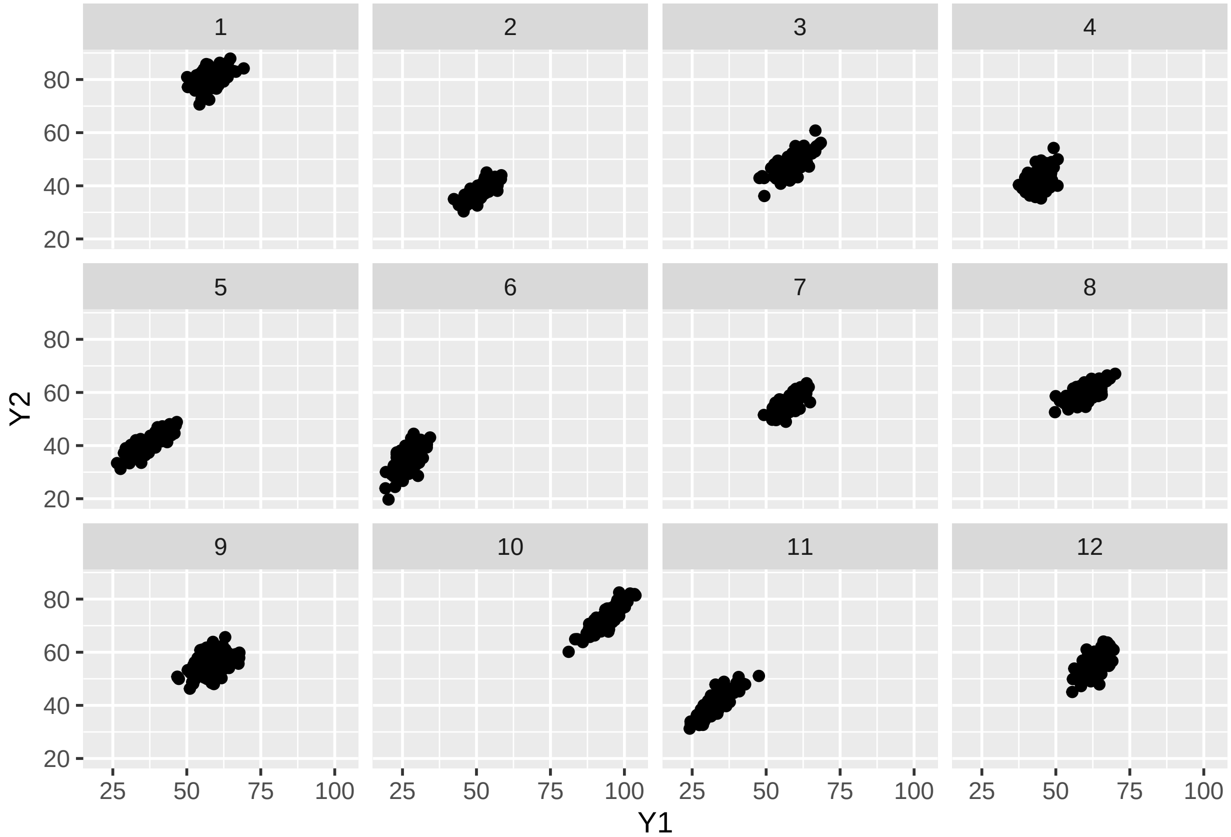
Thus the covariance matrix will be:

$$\Sigma_0 = \begin{bmatrix} 189 & 141.75 \\ 141.75 & 189 \end{bmatrix}$$

```r
agehw <- as.data.frame(read.csv('agehw.dat', sep = ""))

Y <- agehw
n <- nrow(agehw)
p <- ncol(agehw)

mu0 <- rep(52.5, p)
lambda0 <- s0 <- rbind(c(189, 141.75), c(141.75, 189))

#nu0 = p+2
nu0 <- p + 2 + 10
```

b

```r
N <- 100
S <- 12

set.seed(1651)

Y_preds <- lapply(1:S, function(s) {

  # Sample THETA according using prior
  theta <- mvrnorm(n = 1, mu0, lambda0)
  sigma <- inv(rWishart(1, nu0, inv(s0))[, , 1])
  Y_s <- mvrnorm(n = 100, theta, sigma)
  data.frame(Y1 = Y_s[, 1], Y2 = Y_s[, 2], dataset = s)

})

Y_comb <- do.call(rbind, Y_preds)

ggplot(Y_comb, aes(x = Y1, y = Y2)) +
  geom_point() +
  facet_wrap(~ dataset)
```



c

```
S <- 10000

set.seed(1651)
#monte carlo priors
do_mcmc <- function(Y, mu0, lambda0, s0, nu0) {
  ybar <- colMeans(Y)
  p <- ncol(Y)
  n <- nrow(Y)

  THETA <- matrix(nrow = S, ncol = p)
  SIGMA <- array(dim = c(p, p, S))

  #start with sigma sample
  sigma <- cov(Y)

  #gibbs sampling

  # more readable
  inv <- solve

  for (s in 1:S) {
    #update theta
    lambdan <- inv(inv(lambda0) + n * inv(sigma))
    mun <- lambdan %*% (inv(lambda0) %*% mu0 + n * inv(sigma) %*% ybar)
    theta <- mvrnorm(n = 1, mun, lambdan)
    #update sigma
    resid <- t(Y) - c(theta)
    stheta <- resid %*% t(resid)
    sn <- s0 + stheta
    sigma <- inv(rWishart(1, nu0 + n, inv(sn))[, , 1])

    THETA[s, ] <- theta
    SIGMA[, , s] <- sigma
  }

  list(theta = THETA, sigma = SIGMA)
}

my_prior_mcmc <- do_mcmc(agehw, mu0, lambda0, s0, nu0)
THETA <- my_prior_mcmc$theta
SIGMA <- my_prior_mcmc$sigma

#for reuse later
print_quantiles <- function(THETA, SIGMA) {

  #husband
  print("Husband")
  print(quantile(THETA[, 1], probs = c(0.025, 0.5, 0.975)))

  #wife
  print("Wife")
  print(quantile(THETA[, 2], probs = c(0.025, 0.5, 0.975)))

  cors <- apply(SIGMA, MARGIN = 3, FUN = function(covmat) {
    covmat[1, 2] / (sqrt(covmat[1, 1] * covmat[2, 2]))
  })
  print("Correlation")
  print(quantile(cors, probs = c(0.025, 0.5, 0.975)))
}

print_quantiles(THETA, SIGMA)
```

```
## [1] "Husband"
##     2.5%      50%    97.5%
## 41.95711 44.52705 47.10299
## [1] "Wife"
##     2.5%      50%    97.5%
## 38.60235 40.99225 43.39706
## [1] "Correlation"
##      2.5%       50%     97.5%
## 0.8617723 0.9024292 0.9313893
```

d part iii

```
set.seed(1651)
```

```
mu0 <- rep(0, p)
lambda0 <- 10^5 * diag(p)
s0 <- 1000 * diag(p)
nu0 <- 3
diffuse_mcmc <- do_mcmc(agehw, mu0, lambda0, s0, nu0)
print_quantiles(diffuse_mcmc$theta, diffuse_mcmc$sigma)
```

```
## [1] "Husband"
##     2.5%      50%    97.5%
## 41.67009 44.45815 47.24395
## [1] "Wife"
##     2.5%      50%    97.5%
## 38.31861 40.91042 43.51476
## [1] "Correlation"
##      2.5%       50%     97.5%
## 0.7927651 0.8548633 0.8995693
```

e

Comparing the confidence intervals, it doesn't seem that prior info matters since the sample size is large. The diffuse prior is slightly different, but not significantly at all. Regardless of the prior info, the quantiles and correlations being quite similar. A smaller sample size may lead to different results.

```
set.seed(1651)
# my prior

mu0 <- rep(52.5, p)
lambda0 <- s0 <- rbind(c(189, 141.75), c(141.75, 189))

nu0 <- p + 2 + 10
my_prior_mcmc_short <- do_mcmc(agehw[1:25, ], mu0, lambda0, s0, nu0)
print_quantiles(my_prior_mcmc_short$theta, my_prior_mcmc_short$sigma)
```

```
## [1] "Husband"
##     2.5%      50%    97.5%
## 41.01928 45.45946 49.99407
## [1] "Wife"
##     2.5%      50%    97.5%
## 38.42809 43.12462 47.75715
## [1] "Correlation"
##      2.5%       50%     97.5%
## 0.8388558 0.9136948 0.9545245
```

```
# diffuse prior
mu0 <- rep(0, p)
lambda0 <- 10^5 * diag(p)
s0 <- 1000 * diag(p)
nu0 <- 3
diffuse_mcmc_short <- do_mcmc(agehw[1:25, ], mu0, lambda0, s0, nu0)
print_quantiles(diffuse_mcmc_short$theta, diffuse_mcmc_short$sigma)
```

```
## [1] "Husband"
##     2.5%      50%    97.5%
## 39.28346 45.19486 51.26876
## [1] "Wife"
##     2.5%      50%    97.5%
## 36.69468 42.82984 49.01497
## [1] "Correlation"
##      2.5%       50%     97.5%
## 0.5453413 0.7618878 0.8808638
```

By reducing the sample size, we can see a noticeable effect on the intervals.

7.6

```
diab <- read.csv('azdiabetes.dat', sep = "")
diaby <- filter(diab, diabetes == "Yes"); diaby <- diaby[,-8] #yes diabetes
diabn <- filter(diab, diabetes == "No"); diabn <- diabn[,-8] # no diabetes

ny <- nrow(diaby)
nn <- nrow(diabn)

#prior info
mu0y <- colMeans(diaby)
mu0n <- colMeans(diabn)
p <- ncol(diaby)
v0 <- p + 2
S0y <- L0y <- covary <- var(diaby)
S0n <- L0n <-covarn <- var(diabn)
```

a

```
S <- 10000

set.seed(1651)

#for diabetes
prior_mcmcy <- do_mcmc(diaby, mu0y, L0y, S0y, v0)
THETAy <- prior_mcmcy$theta
SIGMAy <- prior_mcmcy$sigma

print_quantiles2 <- function(THETA, SIGMA) {
  print("npreg")
  print(quantile(THETAy[, 1], probs = c(0.025, 0.5, 0.975)))
  print("glu")
  print(quantile(THETAy[, 2], probs = c(0.025, 0.5, 0.975)))
  print("bp")
  print(quantile(THETAy[, 3], probs = c(0.025, 0.5, 0.975)))
  print("skin")
  print(quantile(THETAy[, 4], probs = c(0.025, 0.5, 0.975)))
  print("bmi")
  print(quantile(THETAy[, 5], probs = c(0.025, 0.5, 0.975)))
  print("ped")
  print(quantile(THETAy[, 6], probs = c(0.025, 0.5, 0.975)))
  print("age")
  print(quantile(THETAy[, 7], probs = c(0.025, 0.5, 0.975)))

}

print_quantiles2(THETAy, SIGMAy)
```

```
## [1] "npreg"
##     2.5%      50%    97.5% 
## 4.135005 4.699816 5.280675 
## [1] "glu"
##     2.5%      50%    97.5% 
## 138.4848 143.1119 147.7056 
## [1] "bp"
##     2.5%      50%    97.5% 
## 72.88016 74.70069 76.52159 
## [1] "skin"
##     2.5%      50%    97.5% 
## 31.47044 32.98546 34.49689 
## [1] "bmi"
##     2.5%      50%    97.5% 
## 34.86065 35.82489 36.81748 
## [1] "ped"
##      2.5%       50%     97.5% 
## 0.5600743 0.6167227 0.6759749 
## [1] "age"
##     2.5%      50%    97.5% 
## 34.80586 36.40912 38.02962
```

```
set.seed(1651)

# for non diabetes
prior_mcmcn <- do_mcmc(diabn, mu0n, L0n, S0n, v0)
THETAn <- prior_mcmcn$theta
SIGMAn <- prior_mcmcn$sigma

print_quantiles3 <- function(THETA, SIGMA) {
  print("npreg")
  print(quantile(THETAn[, 1], probs = c(0.025, 0.5, 0.975)))
  print("glu")
  print(quantile(THETAn[, 2], probs = c(0.025, 0.5, 0.975)))
  print("bp")
  print(quantile(THETAn[, 3], probs = c(0.025, 0.5, 0.975)))
  print("skin")
  print(quantile(THETAn[, 4], probs = c(0.025, 0.5, 0.975)))
  print("bmi")
  print(quantile(THETAn[, 5], probs = c(0.025, 0.5, 0.975)))
  print("ped")
  print(quantile(THETAn[, 6], probs = c(0.025, 0.5, 0.975)))
  print("age")
  print(quantile(THETAn[, 7], probs = c(0.025, 0.5, 0.975)))

}

print_quantiles3(THETAn, SIGMAn)
```

```
## [1] "npreg"
##     2.5%      50%    97.5%
## 2.647236 2.928532 3.222492
## [1] "glu"
##     2.5%      50%    97.5%
## 107.5246 110.0127 112.4528
## [1] "bp"
##     2.5%      50%    97.5%
## 68.69906 69.91445 71.16527
## [1] "skin"
##     2.5%      50%    97.5%
## 26.24637 27.29072 28.36996
## [1] "bmi"
##     2.5%      50%    97.5%
## 30.75585 31.42616 32.11361
## [1] "ped"
##      2.5%       50%     97.5%
## 0.4150773 0.4461704 0.4769973
## [1] "age"
##     2.5%      50%    97.5%
## 28.18156 29.22563 30.25816
```

Looking at the intervals above, it seems that the variables that differ significantly are the number of pregnancies, glucose level, diabetes pedigree, and age, assuming multivariate normal for both groups

```
#probability that means of thetay > thetan
for(j in 1:7){
  mean(THETAy[,j] > THETAn[,j])
}
```

It seems that for each variable j, the $P(\theta_{d,j} > \theta_{n,j}|Y)$ is 1, which indicates that the diabetes does in fact have a significant positive effect in all of these variables

b)

```
#posterior means of Sigmas
mean(SIGMAy)
```

```
## [1] 37.17405
```

```
mean(SIGMAn)
```

```
## [1] 34.29703
```