hw8

2024-03-29

2a

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
Y <- read.table("http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/azdiabetes.dat", header=TRUE)
diabetics <- Y %>%
  filter(diabetes=="Yes")
Y_d <- diabetics[, 1:7]
nondiabetics <- Y %>%
  filter(diabetes=="No")
Y_n <- nondiabetics[, 1:7]
```

```r
mu0_d <- colMeans(diabetics[, 1:7])
samp_cov_d <- cov(diabetics[, 1:7])
lambda0_d <- cov(diabetics[, 1:7])
S0_d <- samp_cov_d
nu0_d <- 9

mu0_n <- colMeans(nondiabetics[, 1:7])
samp_cov_n <- cov(nondiabetics[, 1:7])
lambda0_n <- cov(nondiabetics[, 1:7])
S0_n <- samp_cov_n
nu0_n <- 9
```

```r
### Simulate multivariate normal vector
rmvnorm<-function(n,mu,Sigma)
{
  p<-length(mu)
```

```r
  res<-matrix(0,nrow=n,ncol=p)
  if(n>0 & p>0)
  {
    E<-matrix(rnorm(n*p),n,p)
    res<-t(  t(E%*%chol(Sigma)) +c(mu))
  }
  res
}

### Simulate from the Wishart distribution
rwish<-function(n,nu0,S0)
{
  sS0 <- chol(S0)
  S<-array( dim=c( dim(S0),n ) )
  for(i in 1:n)
  {
    Z <- matrix(rnorm(nu0 * dim(S0)[1]), nu0, dim(S0)[1]) %*% sS0
    S[,,i]<- t(Z)%*%Z
  }
  S[,,1:n]
}
```

```r
### Gibbs sampler

Sigma_d <- samp_cov_d
n_d<-dim(Y_d)[1]
S0_d <- samp_cov_d
THETA_d <- NULL
SIGMA_d <- NULL

Sigma_n <- samp_cov_n
n_n<-dim(Y_n)[1]
S0_n <- samp_cov_n
THETA_n <- NULL
SIGMA_n <- NULL

S <- 500

for (s in 1:S){
    ###update theta_d
  Ln_d<-solve( solve(lambda0_d) + n_d*solve(Sigma_d) )
  mun_d<-Ln_d%*%( solve(lambda0_d)%*%mu0_d + n_d*solve(Sigma_d)%*%mu0_d )
  theta_d<-rmvnorm(1,mun_d,Ln_d)
  ###

    ###update Sigma_d
  Sn_d<- S0_d + ( t(Y_d)-c(theta_d) )%*%t( t(Y_d)-c(theta_d) )
  Sigma_d<-solve( rwish(1, nu0_d+n_d, solve(Sn_d)) )
  ###

    ### save results
  THETA_d<-rbind(THETA_d,theta_d) ; SIGMA_d<-rbind(SIGMA_d,c(Sigma_d))
  ###
```

```
    ###update theta_n
  Ln_n<-solve( solve(lambda0_n) + n_n*solve(Sigma_n) )
  mun_n<-Ln_n%*%( solve(lambda0_n)%*%mu0_n + n_n*solve(Sigma_n)%*%mu0_n )
  theta_n<-rmvnorm(1,mun_n,Ln_n)
  ###

    ###update Sigma_n
  Sn_n<- S0_n + ( t(Y_n)-c(theta_n) )%*%t( t(Y_n)-c(theta_n) )
  Sigma_n<-solve( rwish(1, nu0_d+n_n, solve(Sn_n)) )
  ###

    ### save results
  THETA_n<-rbind(THETA_n,theta_n) ; SIGMA_n<-rbind(SIGMA_n,c(Sigma_n))
  ###

  if (s %% 100 == 0){
    print(s)
  }

}
```

```
## [1] 100
## [1] 200
## [1] 300
## [1] 400
## [1] 500
```

```
colMeans(THETA_n) - colMeans(THETA_d)
```

```
##        npreg         glu          bp        skin         bmi         ped
##   -1.7816555 -32.9302286  -4.6869680  -5.7121708  -4.3873174  -0.1689197
##          age
##   -7.1778575
```

```
apply(THETA_n, 2, var) - apply(THETA_d, 2, var)
```

```
##         npreg          glu           bp         skin          bmi
## -0.0625538205 -4.2937236289 -0.4560368370 -0.2792719934 -0.1101451668
##           ped          age
## -0.0006127821 -0.3416572219
```

The glucose of diabetics seems to be, on average, much higher than that of non-diabetics. Additionally, the variance of glucose is much greater for diabetics than it is for non-diabetics.

```
colMeans(THETA_d > THETA_n)
```

```
## npreg   glu    bp  skin   bmi   ped   age
##     1     1     1     1     1     1     1
```

$Pr(\theta_{d,j} > \theta_{n,j}|Y) = 1$ for all j $\in$ {1, 2, 3, 4, 5, 6, 7}.

2b

```r
colMeans(SIGMA_n) - colMeans(SIGMA_d)
```

```
##  [1]  -7.57570136   14.22913421    0.68366192    7.78210451    4.58786817
##  [6]   0.06127550   -4.86451002   14.22913421 -390.28944104   25.66920551
## [11]   2.35791231   15.49342050    0.44297895    7.71647816    0.68366192
## [16]  25.66920551  -13.27539500   16.86205911    5.43727934    0.07169206
## [21]   3.63355464    7.78210451    2.35791231   16.86205911   -6.26736091
## [26]   9.13864636   -0.47069163   24.32173631    4.58786817   15.49342050
## [31]   5.43727934    9.13864636   -0.64629268   -0.28028391   17.89175550
## [36]   0.06127550    0.44297895    0.07169206   -0.47069163   -0.28028391
## [41]  -0.06819314    0.22429395   -4.86451002    7.71647816    3.63355464
## [46]  24.32173631   17.89175550    0.22429395  -18.49115330
```

The 2,2 entry corresponds to glucose, so this supports our observation that the variability in glucose is especially high in diabetics.

```r
plot(colMeans(SIGMA_n), colMeans(SIGMA_d),
     xlab = "Posterior Mean of Covariance Matrix 1",
     ylab = "Posterior Mean of Covariance Matrix 2",
     main = "Posterior Means of Covariance Matrices",
     pch = 19, col = "blue")
```



```r
library(knitr)
```
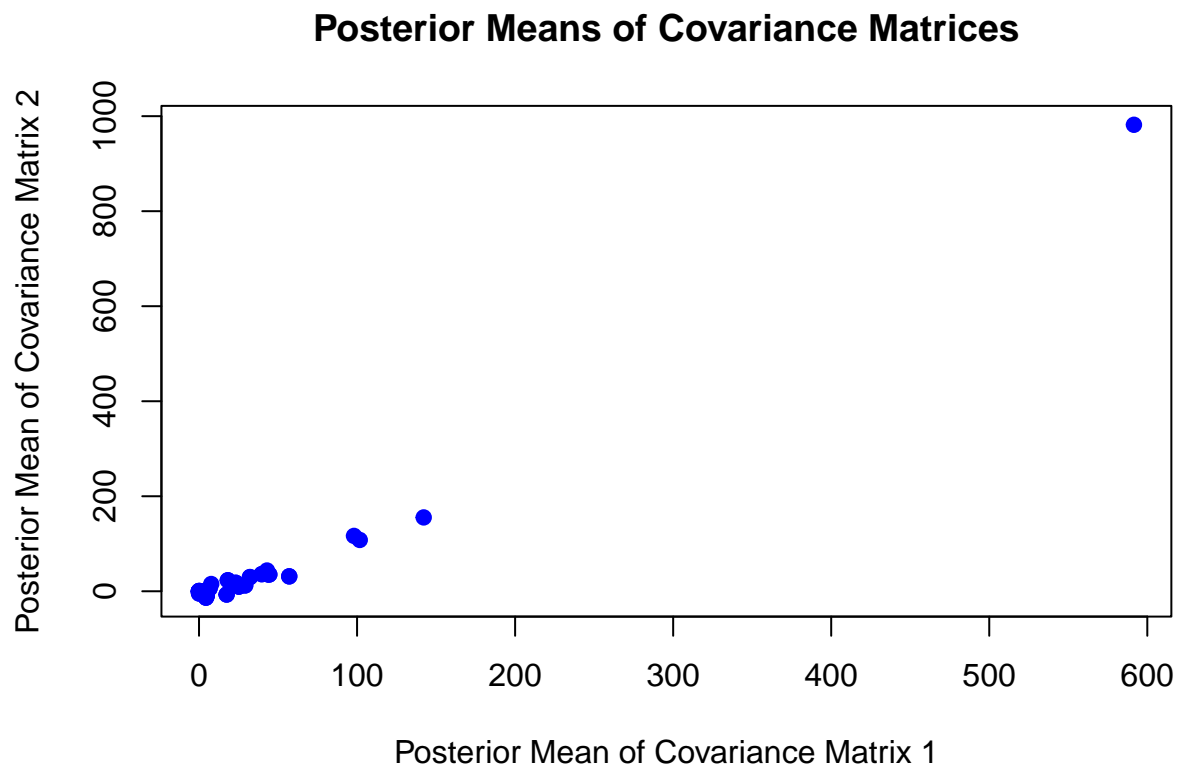
```r
#### Function for posterior quantile intervals for matrices
```

```r
#### From the `sbgcop` package
plotci.sA<-function(sA, ylabs = colnames(sA[, , 1]), mgp = c(1.75, 0.75,
    0))
{
    qA <- qM.sM(sA)
    p <- dim(qA)[1]
    tmp <- c(qA)
    tmp <- tmp[tmp != 1]
    par(mgp = mgp)
    for (j in 1:p) {
        plot(0, 0, type = "n", ylim = range(c(tmp), na.rm = TRUE),
            xlim = c(1, p), ylab = ylabs[j], xaxt = "n", xlab = "")
        points((1:p)[-j], qA[j, -j, 2], pch = 16, cex = 0.6)
        segments(x0 = (1:p)[-j], y0 = qA[j, -j, 1], x1 = (1:p)[-j],
            y1 = qA[j, -j, 3])
        abline(h = 0, col = "gray")
        abline(v = j, col = "gray")
    }
    axis(side = 1, at = 1:p, labels = colnames(qA[, , 1]), las = 2)
}


sR.sC<-function(sC)
{
    p <- dim(sC)[1]
    s <- dim(sC)[3]
    sR <- array(dim = c(p, p, s))
    dimnames(sR) <- dimnames(sC)
    for (l in 1:s) {
        C <- sC[, , l]
        R <- C * NA
        for (j in 1:p) {
            R[j, -j] <- C[j, -j] %*% solve(C[-j, -j])
        }
        sR[, , l] <- R
    }
    sR
}


qM.sM<-function (sM, quantiles = c(0.025, 0.5, 0.975))
{
    p1 <- dim(sM)[1]
    p2 <- dim(sM)[2]
    s <- dim(sM)[3]
    qM <- array(dim = c(p1, p2, length(quantiles)))
    dimnames(qM) <- list(dimnames(sM)[[1]], dimnames(sM)[[2]],
        paste(quantiles * 100, rep("% quantile", length(quantiles)),
            sep = ""))
    for (l in 1:length(quantiles)) {
        qM[, , l] <- apply(sM, c(1, 2), quantile, prob = quantiles[l],
            na.rm = TRUE)
    }
    qM
}
```

```
######


p <- 7

COR <- array( dim=c(p,p,S) )

for(s in 1:S){
Sig <- matrix( SIGMA_d[s,] ,nrow=p,ncol=p)
COR[ , ,s] <- Sig/sqrt( outer( diag(Sig),diag(Sig) ) )
}

apply(COR, c(1,2), quantile,prob=c(.025,.975))
```

```
## , , 1
##
##      [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
## 2.5%    1 -0.21603624 -0.02333336 -0.23897292 -0.30874507 -0.20513279
## 97.5%   1  0.06442201  0.24955069  0.04882643 -0.03512163  0.08022421
##          [,7]
## 2.5%  0.4303916
## 97.5% 0.6476300
##
## , , 2
##
##              [,1] [,2]       [,3]        [,4]        [,5]       [,6]
## 2.5%  -0.21603624    1 -0.0746227 -0.05773455 -0.09412677 -0.1197722
## 97.5%  0.06442201    1  0.2269402  0.22408940  0.19566058  0.1598039
##            [,7]
## 2.5%  -0.04348288
## 97.5%  0.24625907
##
## , , 3
##
##              [,1]       [,2] [,3]        [,4]      [,5]       [,6]      [,7]
## 2.5%  -0.02333336 -0.0746227    1 -0.06026897 0.0831292 -0.1750032 0.1405919
## 97.5%  0.24955069  0.2269402    1  0.22708395 0.3422572  0.1055482 0.3832435
##
## , , 4
##
##              [,1]        [,2]        [,3] [,4]      [,5]        [,6]
## 2.5%  -0.23897292 -0.05773455 -0.06026897    1 0.4025493 -0.01769833
## 97.5%  0.04882643  0.22408940  0.22708395    1 0.6124939  0.28493195
##            [,7]
## 2.5%  -0.19629059
## 97.5%  0.08193227
##
## , , 5
##
##              [,1]        [,2]       [,3]       [,4] [,5]         [,6]         [,7]
## 2.5%  -0.30874507 -0.09412677 0.0831292 0.4025493    1 -0.007774265 -0.34022597
## 97.5% -0.03512163  0.19566058 0.3422572 0.6124939    1  0.291489471 -0.03690773
```

```
##
## , , 6
##
##                [,1]        [,2]        [,3]         [,4]         [,5] [,6]
## 2.5%  -0.20513279 -0.1197722 -0.1750032 -0.01769833 -0.007774265    1
## 97.5%  0.08022421  0.1598039  0.1055482  0.28493195  0.291489471    1
##                [,7]
## 2.5%  -0.1900568
## 97.5%  0.1135090
##
## , , 7
##
##                [,1]         [,2]       [,3]         [,4]         [,5]        [,6] [,7]
## 2.5%   0.4303916 -0.04348288 0.1405919 -0.19629059 -0.34022597 -0.1900568    1
## 97.5%  0.6476300  0.24625907 0.3832435  0.08193227 -0.03690773  0.1135090    1
```

```r
pdf("cor_graph.pdf",height=6,width=6,family="Times")

par(mfcol=c(7,2),mar=c(1,2.75,1,1),mgp=c(1.75,.75,0),oma=c(1.5,0,0,0))


plotci.sA(COR)

REG<-sR.sC(COR)
plotci.sA(REG)
dev.off()
```

```
## pdf
##   2
```

```r
CQ<-apply(COR, c(1,2), quantile,prob=c(.025,.5,.975) )

round(CQ[1,,],2)
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
## [1,]  1.00 -0.22 -0.02 -0.24 -0.31 -0.21  0.43
## [2,] -0.22  1.00 -0.07 -0.06 -0.09 -0.12 -0.04
## [3,] -0.02 -0.07  1.00 -0.06  0.08 -0.18  0.14
## [4,] -0.24 -0.06 -0.06  1.00  0.40 -0.02 -0.20
## [5,] -0.31 -0.09  0.08  0.40  1.00 -0.01 -0.34
## [6,] -0.21 -0.12 -0.18 -0.02 -0.01  1.00 -0.19
## [7,]  0.43 -0.04  0.14 -0.20 -0.34 -0.19  1.00
```

```r
round(CQ[2,,],2)
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
## [1,]  1.00 -0.08  0.13 -0.10 -0.18 -0.06  0.55
## [2,] -0.08  1.00  0.08  0.09  0.05  0.02  0.11
## [3,]  0.13  0.08  1.00  0.10  0.22 -0.04  0.27
## [4,] -0.10  0.09  0.10  1.00  0.51  0.12 -0.06
## [5,] -0.18  0.05  0.22  0.51  1.00  0.15 -0.19
## [6,] -0.06  0.02 -0.04  0.12  0.15  1.00 -0.03
## [7,]  0.55  0.11  0.27 -0.06 -0.19 -0.03  1.00
```

```r
round(CQ[3,,],2)
```

```
##       [,1] [,2] [,3] [,4]  [,5] [,6]  [,7]
```

```
## [1,]   1.00 0.06 0.25 0.05 -0.04 0.08  0.65
## [2,]   0.06 1.00 0.23 0.22  0.20 0.16  0.25
## [3,]   0.25 0.23 1.00 0.23  0.34 0.11  0.38
## [4,]   0.05 0.22 0.23 1.00  0.61 0.28  0.08
## [5,]  -0.04 0.20 0.34 0.61  1.00 0.29 -0.04
## [6,]   0.08 0.16 0.11 0.28  0.29 1.00  0.11
## [7,]   0.65 0.25 0.38 0.08 -0.04 0.11  1.00
```

```r
round(apply(COR,c(1,2),mean),2)
```

```
##         [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
## [1,]   1.00 -0.08  0.12 -0.10 -0.18 -0.06  0.55
## [2,]  -0.08  1.00  0.08  0.09  0.05  0.02  0.11
## [3,]   0.12  0.08  1.00  0.10  0.22 -0.04  0.27
## [4,]  -0.10  0.09  0.10  1.00  0.51  0.13 -0.06
## [5,]  -0.18  0.05  0.22  0.51  1.00  0.14 -0.19
## [6,]  -0.06  0.02 -0.04  0.13  0.14  1.00 -0.03
## [7,]   0.55  0.11  0.27 -0.06 -0.19 -0.03  1.00
```

```r
knitr::include_graphics("cor_graph.pdf")
```
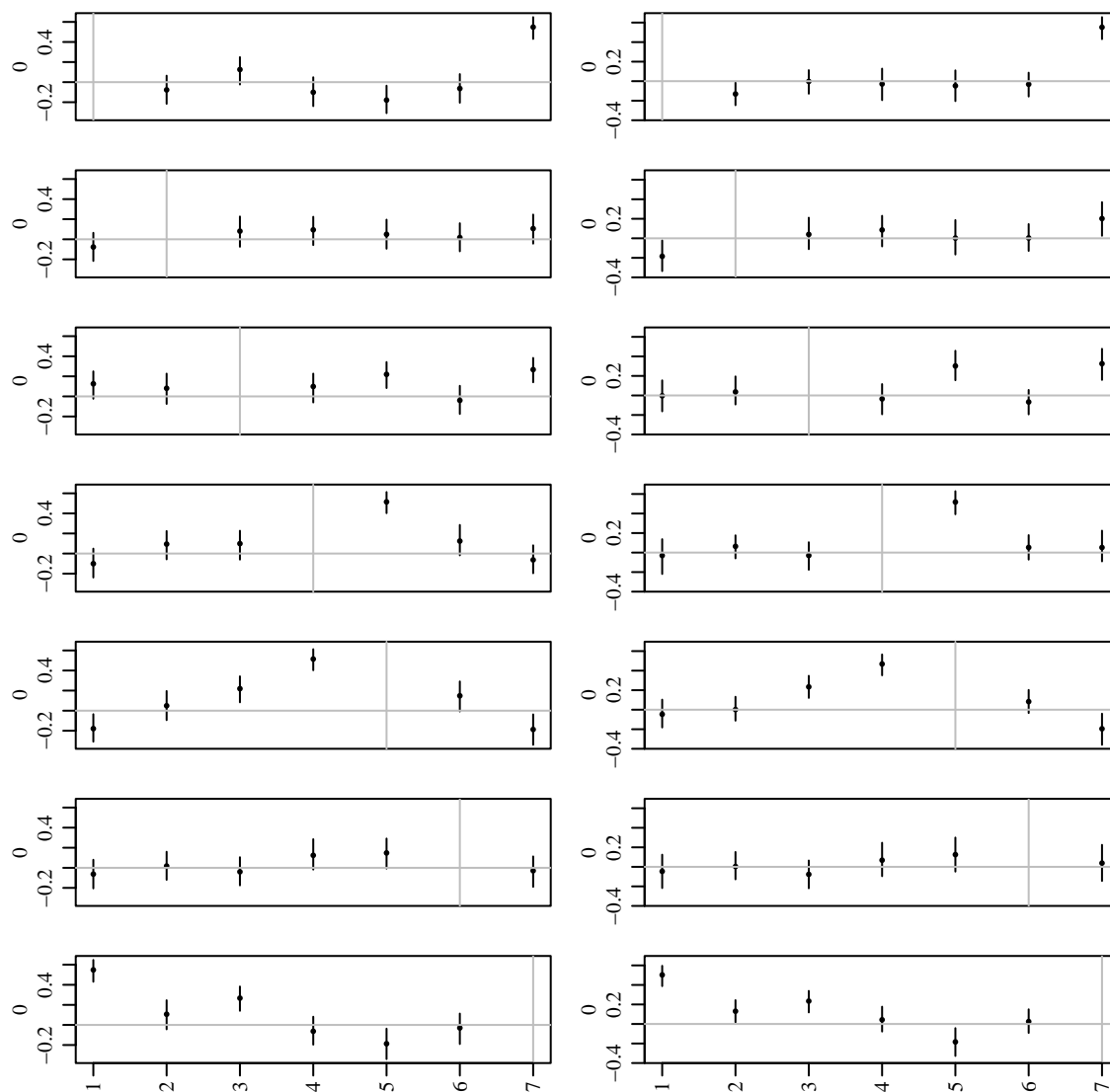
There do not seem to be many significant differences between the two groups. One difference we see is that

4

```r
Y<-dget(url("http://www2.stat.duke.edu/~pdh10/FCBS/Inline/Y.pima.miss"))

Y=Y%>%
  filter(!is.na(glu))%>%
  filter(!is.na(bp))%>%
  filter(!is.na(skin))%>%
  filter(!is.na(bmi))
colMeans(Y)
```

```
##       glu        bp      skin       bmi
## 121.85039  70.61417  28.83465  31.67874
```

```
###
Y <- readRDS("hw8train.rds")

### prior parameters
n<-dim(Y)[1] ; p<-dim(Y)[2]
mu0<-c(rep(0,14))
sd0<-(mu0/2)
L0<-matrix(.1,p,p) ; diag(L0)<-1 ; L0<-L0*outer(sd0,sd0)
nu0<-p+2 ; S0<-L0
###

### starting values
Sigma<-S0
Y.full<-Y
O<-1*(!is.na(Y))
```