STAT401\_HW3

김정현

# Q1.

options(digits=4)  
mu <- c(2, -1)  
sigma <- matrix(c(2, 2, 2, 5), byrow = TRUE, nrow = 2)  
eigen\_result <- eigen(sigma)

## (a)

a11 = round(eigen\_result$vectors[1, 1],4)  
a12 = round(eigen\_result$vectors[2, 1],4)  
a21 = round(eigen\_result$vectors[1, 2],4)  
a22 = round(eigen\_result$vectors[2, 2],4)  
  
cat(paste("PC 1:",a11, "x\_1 + ", a12, "x\_2"))

## PC 1: 0.4472 x\_1 + 0.8944 x\_2

cat("\n")

cat(paste("PC 2:", a21, "x\_1 + ", a22, "x\_2"))

## PC 2: -0.8944 x\_1 + 0.4472 x\_2

First PC: Second PC:

## (b)

round(eigen\_result$values[1]/sum(eigen\_result$values),4)

## [1] 0.8571

The proportion of total population variance explained by first PC is 0.8571.

## (c)

i = 2  
j = 1  
cat("Correlation between X\_2 and Y\_1 : ",sqrt(eigen\_result$values[j])\*a12 / sqrt(sigma[i,i]))

## Correlation between X\_2 and Y\_1 : 0.9798

Correlation between and is 0.9798.

## (d)

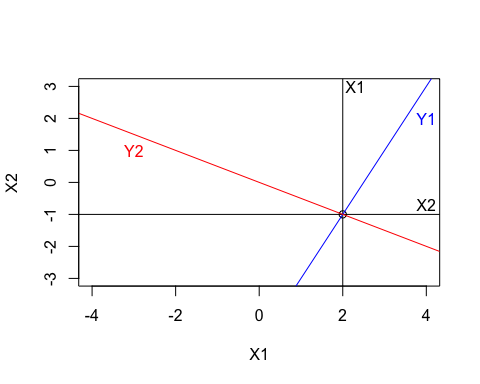
obs <- matrix(c(3,1), byrow = T, nrow = 2)  
a1 <- matrix(c(a11, a12), byrow = F, nrow = 1)  
a1 %\*% (obs-mu)

## [,1]  
## [1,] 2.236

The PC score for first principal component is 2.236 (same with ).

## (e)

pc1 <- mu[2] - a12/ a11 \* mu[1]  
pc2 <- mu[2] - a22/ a21 \* mu[1]  
plot(mu[1],mu[2], xlim=c(-4,4), ylim=c(-3,3), xlab='X1', ylab='X2')  
abline(h=-1); abline(v=2)  
abline(pc1, a12 / a11, col='blue')  
abline(pc2, a22 / a21, col='red')  
text(mu[1]+0.3, 3, "X1") ;text(4, mu[2]+0.3, "X2")  
text(4, 2, "Y1", col='blue'); text(-3, 1, "Y2", col='red')

 Since is the largest eigenvalue, the major axis lie on direction and the minor axis lie on direction .

# Q2.

# Mean vector  
x\_bar <- c(95.5, 164.4, 55.7, 93.4, 18.0, 31.1)  
  
# Covariance matrix  
S <- matrix(c(3266, 1344, 732, 1176, 163, 238,  
 1344, 722, 324, 537, 80, 118,  
 732, 324, 179, 281, 39, 57,  
 1176, 537, 281, 475, 64, 95,  
 163, 80, 39, 64, 10, 14,  
 238, 118, 57, 95, 14, 21), nrow = 6, byrow = TRUE)

## (a)

q2.pca.cov = princomp(covmat = S)  
summary(q2.pca.cov)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6  
## Standard deviation 66.9242 12.33604 5.679053 2.813162 1.1730752 6.545e-01  
## Proportion of Variance 0.9585 0.03257 0.006902 0.001694 0.0002945 9.166e-05  
## Cumulative Proportion 0.9585 0.99102 0.997920 0.999614 0.9999083 1.000e+00

The data can be summarized by 1 dimension (with cumulative proportion approximate 95.8%), which is smaller than 6 dimensions.

## (b)

std\_devs <- sqrt(diag(S))  
  
R <- matrix(nrow = 6, ncol = 6)  
  
for (i in 1:6) {  
 for (j in 1:6) {  
 R[i, j] <- S[i, j] / (std\_devs[i] \* std\_devs[j])  
 }  
}  
  
rownames(R) <- colnames(R) <- c("X1", "X2", "X3", "X4", "X5", "X6")  
  
print(R)

## X1 X2 X3 X4 X5 X6  
## X1 1.0000 0.8752 0.9574 0.9442 0.9019 0.9088  
## X2 0.8752 1.0000 0.9013 0.9170 0.9415 0.9583  
## X3 0.9574 0.9013 1.0000 0.9637 0.9218 0.9297  
## X4 0.9442 0.9170 0.9637 1.0000 0.9286 0.9512  
## X5 0.9019 0.9415 0.9218 0.9286 1.0000 0.9661  
## X6 0.9088 0.9583 0.9297 0.9512 0.9661 1.0000

summary(princomp(covmat = R))

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6  
## Standard deviation 2.3782 0.42023 0.240086 0.224189 0.190853 0.152003  
## Proportion of Variance 0.9427 0.02943 0.009607 0.008377 0.006071 0.003851  
## Cumulative Proportion 0.9427 0.97209 0.981702 0.990078 0.996149 1.000000

The data can be summarized by 1 dimension (with cumulative proportion approximate 94.3%), which is smaller than 6 dimensions.

## (c)

The proportion of total variance has similar value. However, the eigen values and eigen vector are different since scaling is made in correlation matrix.

# Q3.

## (a)

radio <- read.table('radiotherapy.dat', header = TRUE, sep = "")[,-6]  
sapply(radio, var)

## X1 X2 X3 X4 X5   
## 4.6548 0.6128 0.5714 0.1104 0.8622

Since the variance has difference in scale (especially in Symptoms), it is better to use correlation matrix R.

## (b)

q3.pca = prcomp(radio, scale = T)  
round(q3.pca$rotation, 3)

## PC1 PC2 PC3 PC4 PC5  
## X1 0.445 0.231 0.608 0.603 0.127  
## X2 0.432 0.572 0.117 -0.679 0.105  
## X3 0.356 -0.779 0.333 -0.342 0.196  
## X4 0.463 -0.039 -0.665 0.231 0.537  
## X5 0.523 -0.105 -0.252 0.077 -0.804

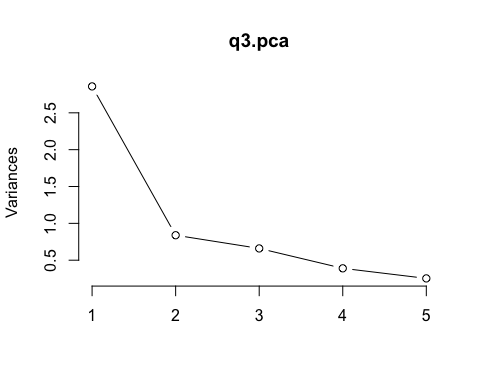
## (c)

summary(q3.pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 1.691 0.916 0.812 0.624 0.5030  
## Proportion of Variance 0.572 0.168 0.132 0.078 0.0506  
## Cumulative Proportion 0.572 0.739 0.871 0.949 1.0000

1. By total proportion of variance: Since adding second proportion explains over 70% variance, choosing 2 principal components is appropriate.
2. By using scree plot: By drawing the scree plot, choosing 2 principal components is appropriate.

screeplot(q3.pca, type = "l")



1. By choosing variance larger than 1: Choosing 1 principal component is appropriate.

Therefore, using 2 principal component is appropriate.

## (d)

1. First principal component explains the overall effect of 5 variables.
2. Second principal component: By considering the absolute value of coefficients larger than 0.2, exclude Food-consumed and appetite. Then, the second principal component can be interpreted as active reason (Symptom, Activity) against non-active reason (Sleep).

## (e)

Since 2 PC explain 73.95% of total sample variance, the data is summarized with 2 PC given in this data.