574_HW3

Matthew Stoebe

2025-04-12

Question 1

a&b)

```
oil <- read.table("Data/oil.dat", header = FALSE)
colnames(oil) <- c("X1", "X2", "X3", "X4", "X5", "zone")


groups <- unique(oil$zone)
g <- length(groups)
n_total <- nrow(oil)

xbar <- colMeans(oil[, c("X1","X2","X3","X4","X5")])

p <- 5
B <- matrix(0, nrow = p, ncol = p)
W <- matrix(0, nrow = p, ncol = p)


for (grp in groups) {

  oil.grp <- subset(oil, zone == grp)
  n_grp <- nrow(oil.grp)


  xbar_grp <- colMeans(oil.grp[, c("X1", "X2", "X3", "X4", "X5")])

  diff_grp <- as.numeric(xbar_grp - xbar)
  B <- B + n_grp * (diff_grp %*% t(diff_grp))

  for (i in 1:n_grp) {
    diff_i <- as.numeric(oil.grp[i, c("X1", "X2", "X3", "X4", "X5")] - xbar_grp)

    W <- W + outer(diff_i, diff_i)
  }
}


A <- solve(W) %*% B
```

```
eigenA <- eigen(A)$values
print("Eigenvalues of A:")
```

```
## [1] "Eigenvalues of A:"
```

```
print(eigenA)
```

```
## [1]  4.178414e+00+0.000000e+00i  6.660138e-01+0.000000e+00i
## [3]  6.929707e-16+0.000000e+00i -1.085708e-16+4.672184e-17i
## [5] -1.085708e-16-4.672184e-17i
```

```
Lambda <- Re(prod(1 / (1 + eigenA)))
print("Wilks' Lambda:")
```

```
## [1] "Wilks' Lambda:"
```

```
print(Lambda)
```

```
## [1] 0.115911
```

```
T_stat <- - (n_total - 1 - p + g/2) * log(Lambda)
df <- p * (g - 1)
p_value <- 1 - pchisq(T_stat, df = df)
cat("Test Statistic T =", T_stat)
```

```
## Test Statistic T = 110.979
```

c)

```
fit <- manova(cbind(X1, X2, X3, X4, X5) ~ zone, data=oil)
summary_fit <- summary(fit, test="Wilks")
print(summary_fit)
```

```
##           Df   Wilks approx F num Df den Df    Pr(>F)
## zone       2 0.11591   18.985     10     98 < 2.2e-16 ***
## Residuals 53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d)

```
summary_aov <- summary.aov(fit)
print(summary_aov)
```

```
##  Response X1 :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## zone          2 135.67  67.837  19.167 5.451e-07 ***
## Residuals    53 187.57   3.539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response X2 :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## zone          2 3186.7 1593.34  20.006 3.366e-07 ***
## Residuals    53 4221.2   79.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response X3 :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## zone          2 0.9844 0.49221  5.8812 0.004935 **
## Residuals    53 4.4357 0.08369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response X4 :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## zone          2 48.803 24.4017  22.673 7.677e-08 ***
## Residuals    53 57.040  1.0762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response X5 :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## zone          2 209.29 104.647  16.408 2.843e-06 ***
## Residuals    53 338.02   6.378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e)

```r
fit12 <- manova(cbind(X1, X2, X3, X4, X5) ~ I(zone %in% c("Wilhelm","Sub-Muhinina")), data=oil)
summary(fit12, test="Wilks")
```

```
##                                              Df   Wilks approx F num Df den Df
## I(zone %in% c("Wilhelm", "Sub-Muhinina"))  1 0.33748   19.632      5     50
## Residuals                                    54
##                                                 Pr(>F)
## I(zone %in% c("Wilhelm", "Sub-Muhinina")) 9.027e-11 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 2

```r
stock <- read.table("Data/stock.dat", header = FALSE)

n <- nrow(stock)
p <- ncol(stock)


pca_stock <- prcomp(stock, scale. = TRUE)
summary(pca_stock)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.5612 1.1862 0.7075 0.63248 0.50514
## Proportion of Variance 0.4874 0.2814 0.1001 0.08001 0.05103
## Cumulative Proportion  0.4874 0.7689 0.8690 0.94897 1.00000
```

```r
pca_stock$rotation
```

```
##           PC1        PC2         PC3        PC4         PC5
## V1 -0.4690832 -0.3680070 -0.60431522 -0.3630228  0.38412160
## V2 -0.5324055 -0.2364624 -0.13610618  0.6292079 -0.49618794
## V3 -0.4651633 -0.3151795  0.77182810 -0.2889658  0.07116948
## V4 -0.3873459  0.5850373  0.09336192  0.3812515  0.59466408
## V5 -0.3606821  0.6058463 -0.10882629 -0.4934145 -0.49755167
```

```r
head(pca_stock$x)
```

```
##             PC1         PC2        PC3         PC4          PC5
## [1,]  0.7840702 -1.05108989 -0.7148399 -1.08451570 -0.701848643
## [2,] -0.5683963  0.23178924 -0.7335404  0.44711016 -0.275089301
## [3,]  0.5936081 -0.04770498  1.0710833  0.01350533  0.003451496
## [4,] -0.2343607 -1.59966660  0.1770439 -1.05408155  0.027038115
## [5,] -0.7759649  1.39648497 -0.9677149 -0.15321340 -0.056295703
## [6,]  0.3068333  0.38199461 -0.6591256 -0.19611501  0.657563969
```

a)

```r
variance_Y2 <- pca_stock$sdev[2]^2
cat("Estimated variance of Y2 =", variance_Y2, "\n\n")
```

```
## Estimated variance of Y2 = 1.407013
```

b)

```r
xi <- c(0.58, -0.41, -0.32, -1.82, 0.04)

loading1 <- pca_stock$rotation[, 1]
Y1_value <- sum(xi * loading1)
cat("The first principal component (Y1) for xi is", Y1_value, "\n\n")
```

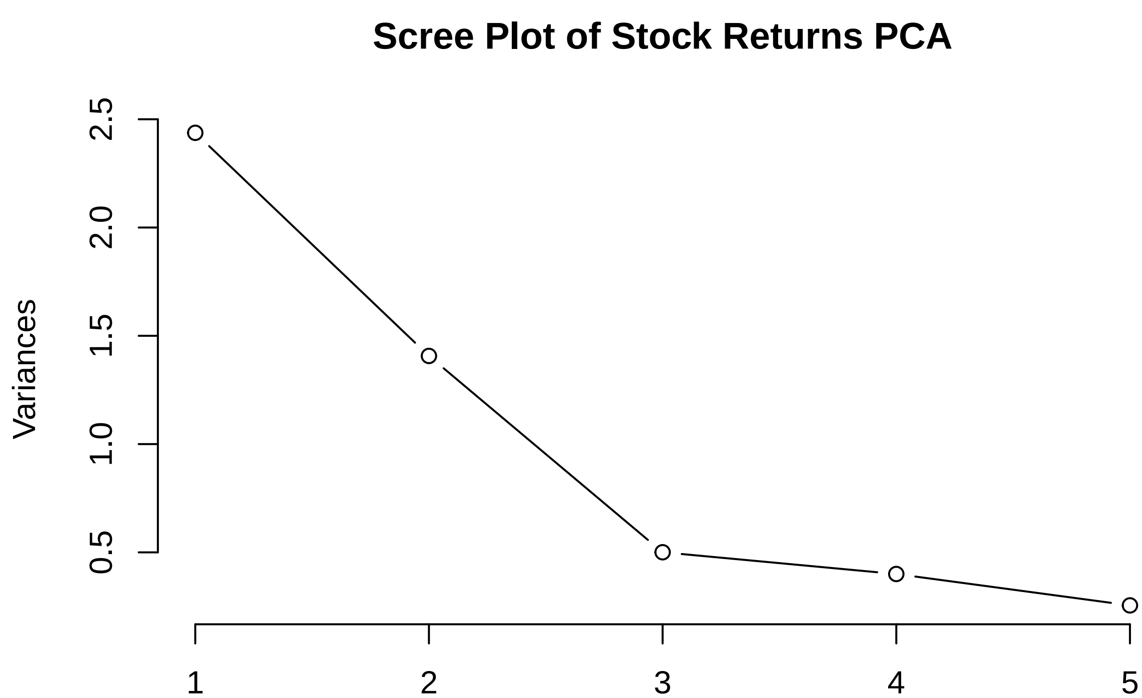
```
## The first principal component (Y1) for xi is 0.7856126
```

c)

```r
summary(pca_stock)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.5612 1.1862 0.7075 0.63248 0.50514
## Proportion of Variance 0.4874 0.2814 0.1001 0.08001 0.05103
## Cumulative Proportion  0.4874 0.7689 0.8690 0.94897 1.00000
```

```r
# Create a scree plot:
plot(pca_stock, type = "l", main = "Scree Plot of Stock Returns PCA")
```



d)

```r
loadings <- pca_stock$rotation
cat("Loadings for PC1:\n")
```

```
## Loadings for PC1:
```

```r
print(loadings[, 1])
```

```
##         V1         V2         V3         V4         V5
## -0.4690832 -0.5324055 -0.4651633 -0.3873459 -0.3606821
```

```r
cat("\nLoadings for PC2:\n")
```

```
##
## Loadings for PC2:
```

```r
print(loadings[, 2])
```

```
##         V1         V2         V3         V4         V5
## -0.3680070 -0.2364624 -0.3151795  0.5850373  0.6058463
```

```r
cat("\n")
```

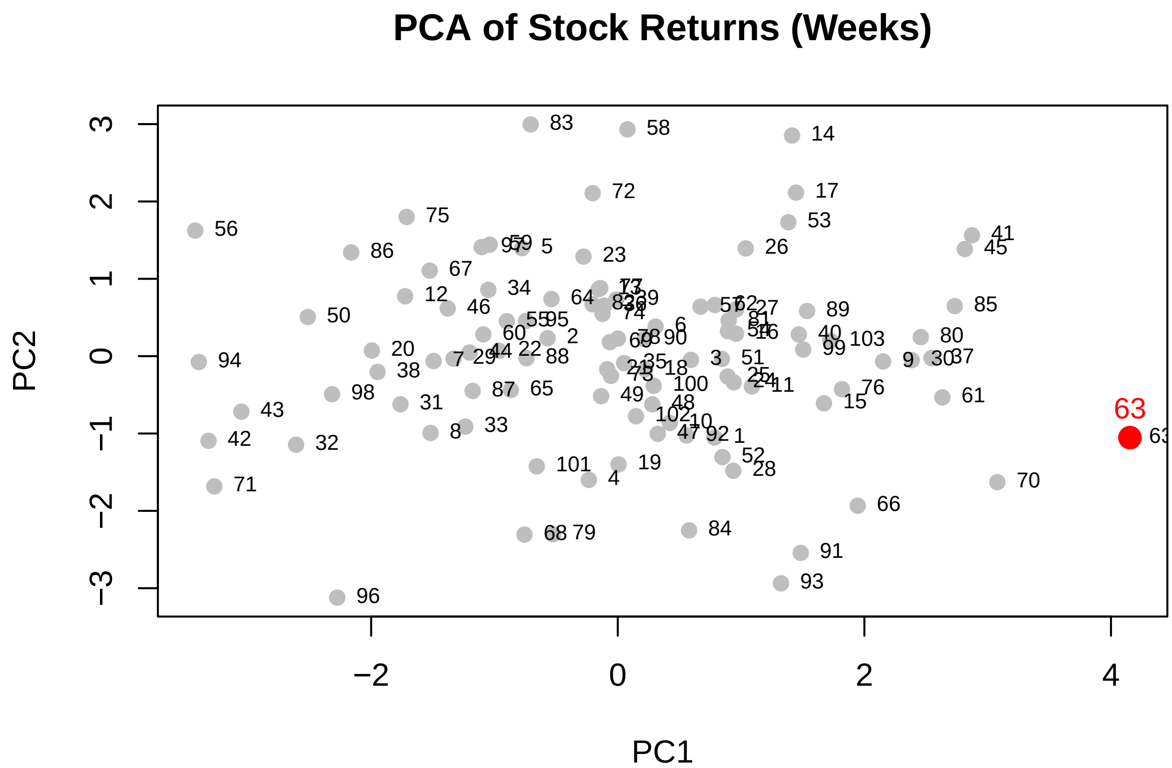
e)

```r
pc_scores <- pca_stock$x

plot(pc_scores[, 1], pc_scores[, 2],
     xlab = "PC1", ylab = "PC2",
     main = "PCA of Stock Returns (Weeks)",
     pch = 19, col = "gray")

text(pc_scores[, 1], pc_scores[, 2], labels = 1:n, cex = 0.7, pos = 4)

points(pc_scores[63, 1], pc_scores[63, 2], col = "red", pch = 19, cex = 1.5)
text(pc_scores[63, 1], pc_scores[63, 2], labels = "63", col = "red", cex = 0.9, pos = 3)
```



It appears that week 63 is an extreme outlier for PC1 and about average for PC2.

Question 3

```r
S <- matrix(c(
  6.59, 6.87, 7.01, 6.56, 6.67, 6.83, 5.97, 7.12,
  6.87, 9.79, 10.79, 10.72, 10.42, 11.01, 11.36, 13.44,
  7.01, 10.79, 13.97, 14.63, 14.13, 14.57, 15.77, 18.58,
  6.56, 10.72, 14.63, 16.72, 16.41, 16.99, 18.82, 21.76,
  6.67, 10.42, 14.13, 16.41, 17.37, 18.17, 19.62, 22.33,
  6.83, 11.01, 14.57, 16.99, 18.17, 21.52, 25.79, 29.68,
  5.97, 11.36, 15.77, 18.82, 19.62, 25.79, 39.47, 48.76,
  7.12, 13.44, 18.58, 21.76, 22.33, 29.68, 48.76, 75.14),
  nrow = 8, ncol = 8, byrow = TRUE)
print("Covariance matrix S:")
```

```
## [1] "Covariance matrix S:"
```

```r
print(S)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,] 6.59 6.87 7.01 6.56 6.67 6.83 5.97 7.12
```

```
## [2,] 6.87  9.79 10.79 10.72 10.42 11.01 11.36 13.44
## [3,] 7.01 10.79 13.97 14.63 14.13 14.57 15.77 18.58
## [4,] 6.56 10.72 14.63 16.72 16.41 16.99 18.82 21.76
## [5,] 6.67 10.42 14.13 16.41 17.37 18.17 19.62 22.33
## [6,] 6.83 11.01 14.57 16.99 18.17 21.52 25.79 29.68
## [7,] 5.97 11.36 15.77 18.82 19.62 25.79 39.47 48.76
## [8,] 7.12 13.44 18.58 21.76 22.33 29.68 48.76 75.14
```

```r
mean_vec <- c(5.96, 9.07, 12.18, 14.89, 18.26, 20.47, 20.13, 14.59)

lambda1 <- 159.28
lambda2 <- 28.54
lambda3 <- 6.59

e1 <- c(-0.11, -0.18, -0.24, -0.28, -0.28, -0.34, -0.47, -0.63)
e2 <- c(-0.27, -0.32, -0.37, -0.36, -0.36, -0.22,  0.16,  0.59)
e3 <- c( 0.46,  0.39,  0.21, -0.05, -0.19, -0.36, -0.50,  0.41)
```

a)

We can use the covariance matrix here becuase all fields are measured in the same units (snow depth) so the variance values are meaningful relative to one another.

b)

159.28

c)

0

d)

10.79

e)

```r
cov_Y1_X1 <- lambda1 * e1[1]
cov_Y1_X1
```

```
## [1] -17.5208
```

f)

```r
total_variability <- sum(diag(S))
explained_variability_PC1_PC2 <- lambda1 + lambda2
prop_explained <- explained_variability_PC1_PC2 / total_variability
round(prop_explained * 100, 2)
```

```
## [1] 93.64
```

g)

PC1 has consistent direction and magnitude. This indicates that PC1 reflects overall level of the snow pack where end of season periods contribute heavily. Because it is all negative, negative

h)

PC2 has negative loading for early periods and positive for later periods which contrasts early snow accumulation with late season accumulation

i)

1991s overall snowpack was average, but the high PC2 indicates that the majority of the snow was late season

j)

It looks like 2011 was the best year for rafting as there is a negative pc1 (negative pc1 means large snow pack i think) and a high PC2 which means a lot of the snow came later in the year

k)

```r
y8 <- c(-7.50, -2.25, -4.25)
x8_centered <- y8[1]*e1 + y8[2]*e2 + y8[3]*e3
x8_reconstructed <- mean_vec + x8_centered
round(x8_reconstructed, 2)
```

```
## [1]  5.44  9.48 13.92 18.01 21.98 25.04 25.42 16.25
```

l)

```r
x1 <- c(1.53, 2.47, 2.76, 4.76, 7.32, 11.20, 14.55, 3.11)
x1_centered <- x1 - mean_vec
y1_2 <- sum(e2 * x1_centered)
round(y1_2, 2)
```

```
## [1] 8.75
```