4630 Assignment 3

Ravish Kamath: 213893664

06 December, 2022

Functions Created

Here are some functions I have created to save time when completing questions.

```
xbar = function(matrix){
  n = dim(matrix)[1]
  onevec = rep(1,n)
  xbar = 1/n*t(matrix)%*%onevec
  return(xbar)
}

Spool = function(dat){
  #Getting the n lengths and parameters
  nbar = rep(0,length(dat))
  for (i in 1: length(dat)){
    nbar[i] = dim(dat[[i]])[1]
  }
  n = sum(nbar)
  params = dim(dat[[1]])[2]
  g = length(dat)
  #Getting the variance matrices
  S = list()
  for (i in 1:g){
    S = append(S,list(var(dat[[i]])))
  }
  #Calculating the Within Matrix
  W = matrix(0,params,params)
  Wnew = matrix(0,params,params)
  for (i in 1:g){
    Wnew = as.matrix(lapply(S[i], '*', (nbar[i] -1)))
    Wnew = matrix(unlist(Wnew), ncol = params, byrow = T)
    W = W + Wnew
  }
  #s_pooled
  result = W/(n - g)
  return(result)
}
spectral = function(matrix){
  eig = eigen(matrix)
  D   = diag(eig$values)
  P   = eig$vectors
  matsqrt = P%*%sqrt(D)%*%t(P)
```

```
  output = list(D, P, matsqrt)
  names(output) = c('D', 'P', 'SqrtMat')
  return(output)
}
```

Question 1

Consider the data given in the EXCEL file tab "q1".

(a) State the multivariable linear regression model with all the necessary assumptions.

(b) Find the predicted model.

(c) Test the significance of the model.

(d) Regarless of your result in part (c), test if X_1 is significant? How about X_2?

(e) Find a 95% woking Hotelling confidence region for the mean response when $X_1 = 192$ and $X_2 = 152$.

(f) Find a 95% woking Hotelling prediction region for a new response when $X_1 = 192$ and $X_2 = 152$.

Solution

Part A

$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$

where $\underset{\sim}{Y} = [Y_{(1)}, Y_{(2)}]$, $X = [\underset{\sim}{1}, X_{(1)}, X_{(2)}]$, $\underset{\sim}{\beta} = [\beta_{(0)}, \beta_{(1)}, \beta_{(2)}]$ and $\underset{\sim}{\epsilon} = [\epsilon_{(1)}, \epsilon_{(2)}]$

Assumptions: $E(\underset{\sim}{\epsilon}_{(i)}) = \underset{\sim}{0}$ $V(\underset{\sim}{\epsilon}_{(i)}) = \sigma_{ii} I$ $cov(\underset{\sim}{\epsilon}_{(i)}, \underset{\sim}{\epsilon}_{(j)}) = \sigma_{ij} I$

Part B

For this section we will show both ways of getting the predicted model. One will be through the actual LS method, and the other will be using the lm function built in R.

Least Squares Method

```
Y = as.matrix(q1df[,1:2])
X = as.matrix(q1df[,3:4])
onevec = rep(1, dim(X)[1])
Xnew = cbind(onevec,X)
beta_coef = solve(crossprod(Xnew))%*%crossprod(Xnew, Y)
beta_coef
```

```
##                y1         y2
## onevec 28.0020858 35.8023808
## x1      0.3876325  0.2446617
## x2      0.5766253  0.4713147
```

Using the built in function in R

```
fit = lm(cbind(y1, y2) ~ x1 + x2, data = q1df)
summary(fit)
```

```
## Response y1 :
##
## Call:
## lm(formula = y1 ~ x1 + x2, data = q1df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.361  -3.794   0.001   4.041  17.925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.0021    29.8446   0.938    0.358
## x1            0.3876     0.2454   1.579    0.129
## x2            0.5766     0.3673   1.570    0.131
##
## Residual standard error: 6.564 on 22 degrees of freedom
## Multiple R-squared:  0.5837, Adjusted R-squared:  0.5459
## F-statistic: 15.43 on 2 and 22 DF,  p-value: 6.502e-05
##
##
## Response y2 :
##
## Call:
## lm(formula = y2 ~ x1 + x2, data = q1df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6192 -3.2907 -0.5662  2.5827 10.7487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.8024    23.8778   1.499    0.148
## x1            0.2447     0.1964   1.246    0.226
## x2            0.4713     0.2938   1.604    0.123
##
## Residual standard error: 5.252 on 22 degrees of freedom
## Multiple R-squared:  0.5349, Adjusted R-squared:  0.4926
## F-statistic: 12.65 on 2 and 22 DF,  p-value: 0.0002205
```

Part C

We have our null hypothesis to be $H_0 : X_1 = X_2 = 0$. Let $\hat{e} = Y - \hat{Y}$ and $\hat{\Sigma} = n^{-1}\hat{e}^T\hat{e}$. Below is the calculation for or sample variance.

```
Yhat = Xnew%*%beta_coef
ehat = Y - Yhat
n = dim(X)[1]
Sigmahat = 1/n*t(ehat)%*%ehat
Sigmahat
```

```
##         y1       y2
## y1 37.9206 10.70300
## y2 10.7030 24.27345
```

Now to calculate our test statistic (which will be using the Wilks Lambda), we need too solve for SSE and SST. We have shown the calculation below. Recall that $SSE = n\hat{\Sigma}$ and $SST = (Y - \hat{Y})^T(Y - \hat{Y})$. Below are the given outputs for SSE and SST.

```
SSE = n*Sigmahat
SSE
```

```
##         y1       y2
## y1 948.015 267.5750
## y2 267.575 606.8363
```

```
ybar = colMeans(Y)
n = nrow(Y)
m = ncol(Y)
r = ncol(X)
q = 1

Ybar = matrix(ybar, n ,m, byrow = T)
SST = crossprod(Y - Ybar)
SST
```

```
##         y1      y2
## y1 2277.44 1230.04
## y2 1230.04 1304.64
```

We now finally calculate the Wilks Lambda and complete the Bartlett method. Let it be shown that Wilks Lambda (Λ) is $\Lambda = \frac{|SSE|}{|SST|}$. Furthermore the formula for the Bartlett test statistic is as follows:

$$-(n - r - 1 - \frac{1}{2}(m - r + q + 1)) \log(\Lambda) \sim \chi^2_{m(r-q)}$$

The output is shown below.

```
WilksLam = det(SSE)/det(SST)

Bartlett = -(n - r - 1 - 1/2*(m - r + q + 1))*log(WilksLam)
df = m*(r-q)
1 - pchisq(Bartlett, df)
```

```
## [1] 1.420822e-05
```

Thus we get the Barlett observed value to be **22.32338** with the p-value being, **1.420822e-05**. Based of the very small p-value, we will **reject the null hypothesis**. This implies that there is evidence that the **model would be signficant**.

Part D

Let $H_0 : X_2 = 0$

Same procedure as part (c), we have our Λ statistic to be **0.8552134**, which when used through Bartlett's method, we get our test statistic to be 3.284489, with a p-value of **0.1935451**. Based of the **large** p-value, we cannot reject H_0. This implies that X_2 is **not significant**. Below is the output code.

```
#Testing if Beta(2) is significant
#H_0: Beta(2)  or X2 = 0
X1 = Xnew[,1:2]
Betahat1 = solve(crossprod(X1))%*%crossprod(X1,Y)
Sigmahat1 = 1/n*crossprod(Y - X1%*%Betahat1)
E = n*Sigmahat
H = n*(Sigmahat1 - Sigmahat)
n = nrow(Y)
m = ncol(Y)
r = ncol(X1)
q = 1

WilksLam = det(E)/det(E + H)
Bartlett = -(n - r - 1 - 1/2*(m - r + q + 1))*log(WilksLam)
df = m*(r-q)
1 - pchisq(Bartlett, df)
```

```
## [1] 0.1935451
```

Let $H_0 : X_1 = 0$

Same procedure as part (c), we have our Λ statistic to be **0.8787331**, which when used through Bartlett's method, we get our test statistic to be 3.284489, with a p-value of **0.2573347**. Based of the **large** p-value, we cannot reject H_0. This implies that X_1 is **not significant**. Below is the output code.

```
#Testing if Beta(1) is significant
#H_0: Beta(1)  or X1 = 0
X2 = cbind(Xnew[,1], Xnew[,3])
Betahat2 = solve(crossprod(X2))%*%crossprod(X2,Y)
Sigmahat2 = 1/n*crossprod(Y - X2%*%Betahat2)
E = n*Sigmahat
H = n*(Sigmahat2 - Sigmahat)
n = nrow(Y)
m = ncol(Y)
r = ncol(X2)
q = 1

WilksLam = det(E)/det(E + H)
Bartlett = -(n - r - 1 - 1/2*(m - r + q + 1))*log(WilksLam)
df = m*(r-q)
1 - pchisq(Bartlett, df)
```

```
## [1] 0.2573347
```

Part E

The formula for a $100(1-\alpha)\%$ Confidence Region is as follow:

$$\underset{\sim}{x}_0^T \hat{\beta}_{(i)} \pm \sqrt{[\frac{(m(n-r-1))}{n-r-m}] F_{m,n-r-m,\alpha}} \sqrt{\underset{\sim}{x}_0 (X^T X)^{-1} \underset{\sim}{x}_0 \hat{\Sigma}_{ii}}$$

Let $\underset{\sim}{x}_0 = [192, 152]^T$. Let us now compute $\underset{\sim}{x}_0$ into the formula, given above.

```
newobs = data.frame(x1 = 192, x2 = 152)
pred = predict(fit, newobs)
n = nrow(Y)
m = ncol(Y)
r = ncol(X)
table = sqrt( ((m*(n-r-1))/(n-r-m))*qf(0.95, df1 = m, df2 = n-r-m) )
x0 = c(1,192, 152)
sd1 = sqrt(t(x0)%*%solve(crossprod(Xnew))%*%x0*Sigmahat[1,1])
sd2 = sqrt(t(x0)%*%solve(crossprod(Xnew))%*%x0*Sigmahat[2,2])
sd = c(sd1, sd2)
CR_L = pred - table*sd
CR_U = pred + table*sd
CR = cbind(t(CR_L), t(CR_U))
colnames(CR) = c("Lower", "Upper")
CR
```

```
##       Lower    Upper
## y1 185.4438 194.7054
## y2 150.7123 158.1222
```

Hence our Confidence Region will be:

$$\begin{bmatrix} 140.0746 \\ 154.4173 \end{bmatrix} \pm 2.695139 \begin{bmatrix} 1.718209 \\ 1.374688 \end{bmatrix}$$

Part F

Very similar to our C.R. formula from part (e), our Hotelling Prediction Region is as follows:

$$x_0^T \hat{\beta}_{(i)} \pm \sqrt{[\frac{(m(n-r-1))}{n-r-m}] F_{m,n-r-m,\alpha}} \sqrt{(1 + x_0 (X^T X)^{-1} x_0) \hat{\Sigma}_{ii}}$$

Let us now use R,to calculate the P.R.

```
sd1 = sqrt( (1 + t(x0)%*%solve(crossprod(Xnew))%*%x0)*Sigmahat[1,1] )
sd2 = sqrt( (1 + t(x0)%*%solve(crossprod(Xnew))%*%x0)*Sigmahat[2,2] )
sd = c(sd1, sd2)
PR_L = pred - table*sd
PR_U = pred + table*sd
PR = cbind(t(PR_L), t(PR_U))
colnames(PR) = c("Lower", "Upper")
PR
```

```
##       Lower    Upper
## y1 172.8440 207.3051
## y2 140.6316 168.2029
```

Hence our Prediction Region will be:

$$\begin{bmatrix} 140.0746 \\ 154.4173 \end{bmatrix} \pm 2.695139 \begin{bmatrix} 6.3931987 \\ 5.115 \end{bmatrix}$$

Question 2

Timm (1975) reported the results of an experiment in which subjects' respond time to "probe words" at five position (Y_1 is at the beginning of the sentence, Y_2 is in the first quartile of the sentence, Y_3 is in the middle of the sentence, Y_4 is in the third quartile of the sentence, and Y_5 is at end of the sentence). The data are recorded in the EXCEL file tab "q2".

(a) Use the sample variance and obtain all the principle components.

(b) Timm specifically required the reduction in dimension should cover at least 90% of the total variance. How many principle components are needed? Why?

(c) Repeat parts (a) and (b) using the sample correlation matrix.

Solution

Part A

In this part, we will try to hard code the principal components, using the sample variance. Note that we could also use the prcomp function that is usually used to calculate P.C. First we need to calculate the sample variance. The output of the sample variance is calculated below.

```
X = data.matrix(q2df)
n = dim(X)[1]
params = dim(X)[2]
onevec = rep(1, n)
#Mean Vector
xbar = 1/n*t(X)%*%onevec
#Variance Matrix
S = 1/(n - 1)*(t(X)%*%X - n*xbar%*%t(xbar))
S
```

```
##          x1       x2       x3       x4       x5
## x1 65.09091 33.64545 47.59091 36.77273 25.42727
## x2 33.64545 46.07273 28.94545 40.33636 28.36364
## x3 47.59091 28.94545 60.69091 37.37273 41.12727
## x4 36.77273 40.33636 37.37273 62.81818 31.68182
## x5 25.42727 28.36364 41.12727 31.68182 58.21818
```

Now we get the eigen value and eigen vectors. The eigen vectors will represent the Principal Components

```
#Getting the eigenvalues and vectors
eig = eigen(S)
eig$vectors
```

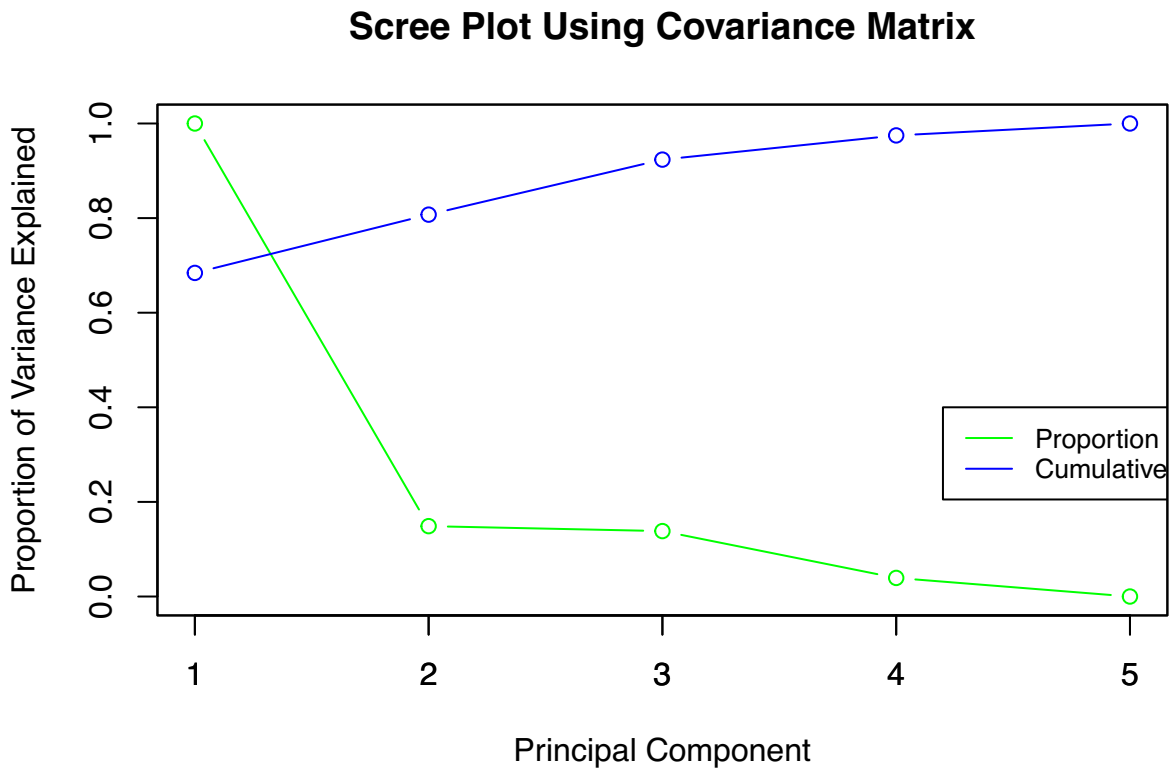
```
##            [,1]        [,2]        [,3]       [,4]       [,5]
## [1,] -0.4727831  0.57631826  0.41685181  0.2285789  0.4672469
## [2,] -0.3918187  0.10826396 -0.45239805  0.6558114 -0.4472185
## [3,] -0.4875471 -0.09600807  0.47945493 -0.3689607 -0.6221505
## [4,] -0.4677199  0.12056819 -0.61953744 -0.5803451  0.2146494
## [5,] -0.4080320 -0.79522446  0.08869553  0.2114962  0.3853964
```

Hence our Principal Components will be as follow:

$$
\begin{aligned}
Y_1 &= -0.4728X_1 - 0.3918X_2 - 0.4875X_3 - 0.4677X_4 - 0.4080X_5 \\
Y_2 &= 0.5763X_1 + 0.1082X_2 - 0.0960X_3 + 0.1206X_4 - 0.7952X_5 \\
Y_3 &= 0.4168X_1 - 0.4524X_2 + 0.4794X_3 - 0.6195X_4 + 0.0887X_5 \\
Y_4 &= 0.2286X_1 + 0.6558X_2 - 0.3690X_3 - 0.5803X_4 + 0.2115X_5 \\
Y_5 &= 0.4672X_1 - 0.4472X_2 - 0.6222X_3 + 0.2146X_4 + 0.3854X_5
\end{aligned}
$$

Part B

```
result = prcomp(X, scale = F)
var_exp = result$sdev^2/sum(result$sdev^2)
plot(var_exp, type = 'b', col = 'green', yaxt = 'n', ylab = '', xlab = '')
par(new = T)
plot(cumsum(var_exp), xlab = "Principal Component",
     ylab = "Proportion of Variance Explained",
     main = "Scree Plot Using Covariance Matrix",
     ylim = c(0, 1), type = "b", col = 'blue')
legend(4.2, 0.4, legend=c("Proportion", "Cumulative"),
       col=c("green", "blue"), lty=1:1, cex=0.8)
```



I would say that using the first 3 P.C. would be needed to cover 90% of the total variance. When comparing P.C. 3 and P.C. 4 we oly get an extra 5% of explained variance.

Part C

Same procedure for part (a), we first need to find the correlation matrix.

```
sd = sqrt(diag(diag(S),5))
invsd = solve(sd)
cor = invsd%*%S%*%t(invsd)
cor
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.6143902 0.7571850 0.5750730 0.4130573
## [2,] 0.6143902 1.0000000 0.5473897 0.7497770 0.5476595
## [3,] 0.7571850 0.5473897 1.0000000 0.6052716 0.6918927
## [4,] 0.5750730 0.7497770 0.6052716 1.0000000 0.5238876
## [5,] 0.4130573 0.5476595 0.6918927 0.5238876 1.0000000
```

Now let us compute the P.C for the correlation matrix. But before we do that, when using the correlation matrix for PCA, we need to standardize our independent variables. Let $Z_i = \frac{(X_i - \mu_i)}{\sqrt{\sigma_{ii}}}$ for $i = 1, ..., 5$.

```
result = prcomp(scale(X), scale = F )
print(result)
```

```
## Standard deviations (1, .., p=5):
## [1] 1.8483758 0.7838567 0.7564879 0.5207797 0.3543867
##
## Rotation (n x k) = (5 x 5):
##           PC1        PC2        PC3        PC4        PC5
## x1 -0.4418394  0.2006104 -0.6786078  0.2125365 -0.5087760
## x2 -0.4535595  0.4280646  0.3491277  0.6055405  0.3499642
## x3 -0.4727808 -0.3678765 -0.3754368 -0.2581448  0.6584479
## x4 -0.4536224  0.3934629  0.3345386 -0.7010073 -0.1899641
## x5 -0.4120276 -0.6974023  0.4058723  0.1734903 -0.3860467
```

Hence our principal components are as follows:

$$
\begin{aligned}
Y_1 &= -0.4418Z_1 - 0.4536Z_2 - 0.47288Z_3 - 0.4536Z_4 - 0.4120Z_5 \\
Y_2 &= 0.2006Z_1 - 0.4281Z_2 - 0.3679Z_3 + 0.3935Z_4 - 0.6974Z_5 \\
Y_3 &= 0.6786Z_1 - 0.3491Z_2 + 0.3754Z_3 - 0.3345Z_4 - 0.4059Z_5 \\
Y_4 &= 0.2125Z_1 + 0.6055Z_2 - 0.2581Z_3 - 0.7010Z_4 + 0.1735Z_5 \\
Y_5 &= 0.5088Z_1 - 0.3410Z_2 - 0.6584Z_3 + 0.1900Z_4 + 0.3860Z_5
\end{aligned}
$$

Question 3

Use the data in the EXCEL file tab "q2".

(a) Find the canonical correlation between (x_1, x_2, x_3) and (x_4, x_5).

(b) Test the significance canonical correlations.

(c) Regardless of your answer in part (b), is each canonical correlation individually significant?

Solution

```
X = data.matrix(q3df)
col_order = c('x4', 'x5', 'x1', 'x2', 'x3')
X = X[,col_order]
```

Part A

We first need to get the correlation matrix in order to calculate the canonical correlations.

```
cor(X)
```

```
##           x4        x5        x1        x2        x3
## x4 1.0000000 0.5238876 0.5750730 0.7497770 0.6052716
## x5 0.5238876 1.0000000 0.4130573 0.5476595 0.6918927
## x1 0.5750730 0.4130573 1.0000000 0.6143902 0.7571850
## x2 0.7497770 0.5476595 0.6143902 1.0000000 0.5473897
## x3 0.6052716 0.6918927 0.7571850 0.5473897 1.0000000
```

Now let $X^{(1)} = (\underset{\sim}{X}_4, \underset{\sim}{X}_5)$ and $X^{(2)} = (\underset{\sim}{X}_1, \underset{\sim}{X}_2, \underset{\sim}{X}_3)$

Furthermore, let the following Σ matrices be subsections of our correlation matrix.

$$\Sigma_{11} = \begin{bmatrix} 1 & 0.5239 \\ 0.5239 & 1 \end{bmatrix} \quad \Sigma_{12} = \begin{bmatrix} 0.5751 & 0.7498 & 0.6053 \\ 0.4131 & 0.5477 & 0.6919 \end{bmatrix}$$

$$\Sigma_{21} = \Sigma_{12} \qquad \Sigma_{12} = \begin{bmatrix} 1 & 0.6144 & 0.7572 \\ 0.6144 & 1 & 0.5474 \\ 0.7572 & 0.5474 & 1 \end{bmatrix}$$

Let $A = \Sigma_{11}^{-\frac{1}{2}} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}$. Let us use R to calculate $\Sigma_{11}^{-\frac{1}{2}}$ and Σ_{22}^{-1}.

```
cor11 = matrix(c(1.0000000, 0.5238876,
                 0.5238876, 1.0000000),
               ncol = 2, byrow = T)
cor11_sqrt = spectral(cor11)
cor11_invsqrt = solve(cor11_sqrt$SqrtMat)
cor12 = matrix(c(0.5750730, 0.7497770, 0.6052716,
                 0.4130573, 0.5476595, 0.6918927),
               ncol = 3, byrow = T)
cor21 = t(cor12)
cor22 = matrix(c(1.0000000, 0.6143902, 0.7571850,
                 0.6143902, 1.0000000, 0.5473897,
                 0.7571850, 0.5473897, 1.0000000),
               ncol = 3, byrow = T)
cor22_inv = solve(cor22)
```

Now we can finally calculate our A matrix. This is shown below.

```
A = cor11_invsqrt%*%cor12%*%cor22_inv%*%cor21%*%cor11_invsqrt
A
```

```
##           [,1]      [,2]
## [1,] 0.4705581 0.2831067
## [2,] 0.2831067 0.4371090
```

To find the canonical correlation, we now need to compute the eigen values of the A matrix. Using R once again we come with the following output.

```
canon_cor = sqrt((eigen(A)$values))
canon_cor
```

```
## [1] 0.8587396 0.4125933
```

By taking the square root of the eigen values of the matrix A, we get the following canonical correlations:

$$p_1^* = 0.8587396 \qquad p_2^* = 0.4125933$$

Part B

Let our null hypothesis be: $H_0 : p_1^* = p_2^* = \Sigma_{12} = 0$. We will be using what Bartlett suggests:

$$-[n - 1 - \frac{1}{2}(p + q + 1)] \log\Big(\prod_{i=1}^{p}(1 - (\hat{p}_i^*)^2)\Big) > \chi^2_{pq}(\alpha)$$

Let $p \leq q$ which will be the number of variables in each group. Hence, $p = 2$ and $q = 3$. Using the test statistic formula from above we get the p-value to be 0.09923, which is quite high. Hence we cannot reject our null hypothesis. This implies that our canonical correlations are not significant. Below is the R code to provide the evidence for our statement.

```
n = dim(X)[1]
p = min(3,2)
q = max(3,2)
Bartlett = -(n - 1 - 1/2*(p + q + 1))*log(prod((1-canon_cor^2)))
Bartlett
```

```
## [1] 10.66704
```

```
df = p*q
1-pchisq(Bartlett, df)
```

```
## [1] 0.09922844
```

Part C

Let $H^{(1)} : p_1^* \neq 0$, $p_2^* = 0$. Hence our alternative hypothesis will be $H_a^{(1)} : p_2^* \neq 0$.

The Bartlett formula to find the test statistic is as follows:

$$-[n - 1 - \frac{1}{2}(p + q + 1)] \log(1 - \ (\hat{p}_2^*)^2) > \chi^2_{(p-1)(q-1)}(\alpha)$$

Using the above test statistic, our Bartlett observed value is **1.306275** with a p-value of **0.5204105**.Due to the large p-value we **fail to reject our null hypothesis**. Hence we have do not have evidence that p_2^* is significant. Below is the R code for this problem.

```
#Testing p_2 not equal to  0
Bartlett = -(n - 1 - 1/2*(p + q + 1))*log((1-canon_cor[2]^2))
Bartlett
```

```
## [1] 1.306275
```

```
df = (p - 1)*(q-1)
1-pchisq(Bartlett, df)
```

```
## [1] 0.5204105
```

Let $H^{(2)} : p_2^* \neq 0$, $p_1^* = 0$. Hence our alternative hypothesis will be $H_a^{(2)} : p_1^* \neq 0$. Essentially we are using the same Bartlett formula except we replace p_2^*, with p_1^*.

Using the above test statistic, our Bartlett observed value is **9.360764** with a p-value of **0.009275471**.Due to the very small p-value we **reject our null hypothesis**. Hence we have evidence that p_1^* is significant. Below is the R code for this problem.

```
#Testing p_1 not equal to 0
Bartlett = -(n - 1 - 1/2*(p + q + 1))*log((1-canon_cor[1]^2))
Bartlett
```

```
## [1] 9.360764
```

```
df = (p - 1)*(q - 1)
1-pchisq(Bartlett, df)
```

```
## [1] 0.009275471
```

Question 4

(a) Show that

$$-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)+\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)=(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)$$

(b) Let

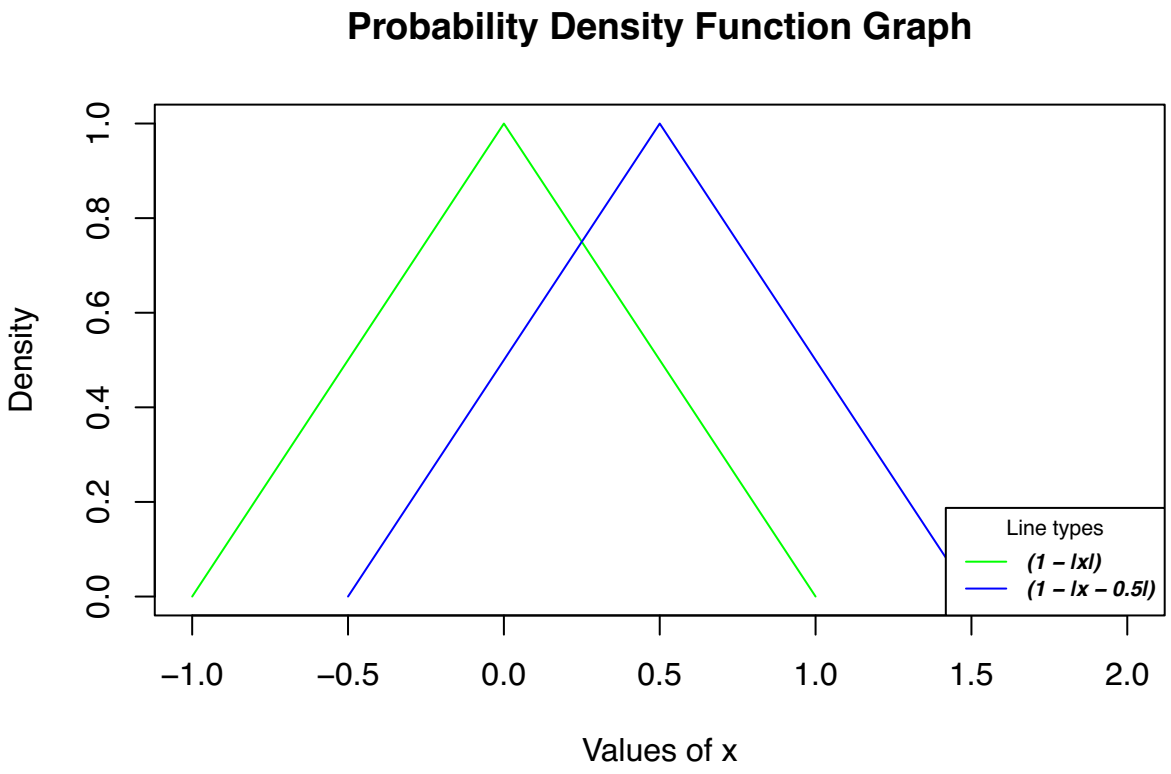
$$\begin{aligned} f_1(x) &= (1-|x|) \qquad \text{for} \quad |x|\leq 1 \\ f_2(x) &= (1-|x-0.5|) \qquad \text{for} \quad -0.5\leq x\leq 1 \end{aligned}$$

Solution

Part A

Handwritten notes.

Part B.1



Part B.2

Handwritten notes.

Part B.3

Handwritten notes.

Question 4

Question 4:

a. Show that

$$-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)+\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_x)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)$$

$$(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)$$

b. Let

$$f_1(x) = (1-|x|) \quad \text{for} \quad |x|\le 1$$
$$f_2(x) = (1-|x-0.5|) \quad \text{for} \quad -0.5\le x\le 1.5$$

1. Sketch the two densities in the same graph.
2. Identify the classification rule for the case $p_1=p_2$ and $c(1|2)=c(2|1)$.
3. Identify the classification rule for the case $p_1=0.2$ and $c(1|2)=c(2|1)$.

a)

$$\frac{-1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)^T\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)+\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)^T\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)$$

$$=\frac{-1}{2}\left[(\underset{\sim}{x}-\underset{\sim}{\mu}_1)^T\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)-(\underset{\sim}{x}-\underset{\sim}{\mu}_2)^T\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)\right]$$

$$=\frac{-1}{2}\left[\underset{\sim}{x}^T\Sigma^{-1}\underset{\sim}{x}-2\underset{\sim}{\mu}_1^T\Sigma^{-1}\underset{\sim}{x}+\underset{\sim}{\mu}_1^T\Sigma^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{x}^T\Sigma^{-1}\underset{\sim}{x}+2\underset{\sim}{\mu}_2^T\Sigma^{-1}\underset{\sim}{x}-\underset{\sim}{\mu}_2^T\Sigma^{-1}\underset{\sim}{\mu}_2\right]$$

$$=\frac{-1}{2}\left[2\underset{\sim}{\mu}_2^T\Sigma^{-1}\underset{\sim}{x}-2\underset{\sim}{\mu}_1^T\Sigma^{-1}\underset{\sim}{x}+\underset{\sim}{\mu}_1^T\Sigma^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2^T\Sigma^{-1}\underset{\sim}{\mu}_2\right]$$

$$=\frac{-1}{2}\left[(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)^T\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)-2(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)^T\Sigma^{-1}\underset{\sim}{x}\right]$$

$$=(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)^T\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)^T\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)$$

b.2) $p_1=p_2$ and $c(1|2)=c(2|1)=1$ Hence $R_1: f_1(x)\ge f_2(x)$ $R_2: f_1(x)<f_2(x)$

Based of the graph we sketched, we can see that the intersection point is at 0.25.

When $f_1(x)>f_2(x)$, R_1 is between $(-1, 0.25)$

" $f_1(x)>f_2(x)$, R_2 is between $(0.25, 1.5)$

b.3) $p_1=0.2$ $p_2=0.8$ and $c(1|2)=c(2|1)$. Hence

$$R_1: \frac{f_1(x)}{f_2(x)}\ge\frac{p_2}{p_1} \qquad R_2: \frac{f_1(x)}{f_2(x)}<\frac{p_2}{p_1}=4$$

$$\ge\frac{0.8}{0.2}=4$$

Question 5

Use the data in the EXCEL file tab "q5". Assume the data is a sample from two multivariate normal distributions.

(a) Identify the classification rule for the case $p1 = p2$ and $c(1|2) = c(2|1)$.

(b) Identify the classification rule for the case $p_1 = 0.25$ and $c(1|2)$ is half of $c(2|1)$.

(c) Identify the Bayesian rule using $p_1 = 0.6$.

(d) Based on the rules given in part (a), which population will a new data point $(50, 48, 47, 49)$ be classified into? How about using the rules in part (b)? Using the rule in part (c)?

Solution

```
A = as.matrix(q5df[q5df$Group == "A",1:4])
B =as.matrix(q5df[q5df$Group =="B",1:4])
dat = list(A, B)
```

Part A

All the formulation of the classification rule will be in the handwritten notes, however here is the R code for the calculation.

```
xbar = function(matrix){
  n = dim(matrix)[1]
  onevec = rep(1,n)
  xbar = 1/n*t(matrix)%*%onevec
  return(xbar)
}
# Mean Vectors
xbar1 = xbar(A)
xbar2 = xbar(B)
spooled = Spool(dat)

ahat = t(xbar1 - xbar2)%*%solve(spooled)

ybar1 = ahat%*%xbar1
ybar2 = ahat%*%xbar2

midpoint = 1/2*(ybar1 + ybar2)
midpoint
```

```
##           [,1]
## [1,] -7.062536
```

```
# Hence the classification is -7.062536
```

Part B

Handwritten Notes

Part C

Handwritten Notes.

Question 5

Question 5: Use the data in the EXCEL file tab "q5". Assume the data is a sample from two multivariate normal distributions.

a. Identify the classification rule for the case $p_1 = p_2$ and $c(1|2) = c(2|1)$.

b. Identify the classification rule for the case $p_1 = 0.25$ and $c(1|2)$ is half of $c(2|1)$.

c. Identify the Bayesian rule using $p_1 = 0.6$.

d. Based on the rules given in part (a), which population will a new data point $(50, 48, 47, 49)$ be classified into? How about using the rules in part (b)? Using the rule in part (c)?

a)

$$R_1: \frac{f_1(x)}{f_2(x)} = \exp\left\{\frac{-1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right\} \geq 1$$

$$= (\mu_1-\mu_2)^T\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2) \geq 0$$

$$R_2: \frac{f_1(x)}{f_2(x)} < 1$$

Since μ_1, μ_2 and Σ are unknown let us use $\bar{x}_1$, $\bar{x}_2$ and S_{pooled}

$$\bar{x}_1 = \begin{bmatrix} 48.21667 \\ 49.2750 \\ 50.1\bar{3} \\ 51.1750 \end{bmatrix} \quad \bar{x}_2 = \begin{bmatrix} 49.2\bar{0} \\ 50.1\bar{5}0 \\ 51.2250 \\ 51.8625 \end{bmatrix} \quad S_{pooled} =$$

	[,1]	[,2]	[,3]	[,4]
[1,]	6.352037	6.284722	5.825185	5.613611
[2,]	6.284722	6.603472	6.240556	6.092083
[3,]	5.825185	6.240556	6.984537	7.132083
[4,]	5.613611	6.092083	7.132083	7.753403

Let $\hat{a} = (\bar{x}_1 - \bar{x}_2)S_{pooled}^{-1}$

$\bar{y}_1 = \hat{a}\cdot\bar{x}_1$; $\bar{y}_2 = \hat{a}\bar{x}_2$

Then $R_1: \hat{y} \geq \frac{1}{2}(\bar{y}_1+\bar{y}_2) = -7.662536$

and $R_2: \hat{y} < -7.662536$

If our observation falls within R_1 we will classify it to A population, else we will classify it to B.

b) We $p_1 = 0.25$, $p_2 = 0.75$ and $c(1|2)$ is half of $c(2|1)$

$$R_1: \frac{f_1(x)}{f_2(x)} = (\mu_1-\mu_2)^T\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2) \geq \log\left(\frac{c(1|2)}{c(2|1)}\cdot\frac{p_2}{p_1}\right)$$

$$\geq \log\left(\frac{c(1|2)}{2c(1|2)}\cdot\frac{3/4}{1/4}\right)$$

$$\geq \log(1.5)$$

$$\geq 0.4055$$

Hence $R_2: \frac{f_1(x)}{f_2(x)} < 0.4055$

If our observation falls within R_1 we will classify it to A population, else we will classify it to B.

c) $P(A|\underset{\sim}{x}) = \dfrac{p_1 \cdot P(\underset{\sim}{x}|A)}{p_1 \cdot P(\underset{\sim}{x}|A) + p_2 \cdot P(\underset{\sim}{x}|B)}$

$= \dfrac{0.6 \cdot P(\underset{\sim}{x}|A)}{0.6 P(\underset{\sim}{x}|A) + 0.4 \cdot P(\underset{\sim}{x}|B)}$

$P(B|\underset{\sim}{x}) = \dfrac{p_2 \cdot P(\underset{\sim}{x}|B)}{p_1 \cdot P(\underset{\sim}{x}|A) + p_2 \cdot P(\underset{\sim}{x}|B)}$

$= \dfrac{0.4 \cdot P(\underset{\sim}{x}|B)}{0.6 P(\underset{\sim}{x}|A) + 0.4 \cdot P(\underset{\sim}{x}|B)}$

If $P(\underset{\sim}{x}|A) > P(\underset{\sim}{x}|B)$, then we would classify it to pop^{n} A, else we would classify it to pop^{n} B.

Part D.1

When we have equal cost and prior discriminants:

```
newobs = c(50, 48, 47, 49)
ahat%*%newobs
```

```
##           [,1]
## [1,] -5.481962
```

Since the observation is bigger than the cutoff point, which was -7.062536, we will classify this obs. to population A.

Part D.2

When p1 = 0.25, and c(1|2) is half of c(2|1)

```
ahat%*%newobs - midpoint
```

```
##          [,1]
## [1,] 1.580574
```

Since the cutoff value for this classification is 0.4055, we will also classify this obs. to population A.

Part D.3

Most of this will be shown in the handwritten, however to calculation certain probabilities for the Bayesian rule, we need to use R. This is shown below

```
#When we use Bayesian Rule
probA_newob = (2*pi)^(-2)*(det(spooled))^(-1/2)*
              exp( (-1/2)*t((newobs - xbar1))%*% solve(spooled)%*%(newobs -
                                                                   xbar1))
probA_newob
```

```
##              [,1]
## [1,] 1.475133e-09
```

```
probB_newob = (2*pi)^(-2)*(det(spooled))^(-1/2)*
  exp( (-1/2)*t((newobs - xbar2))%*% solve(spooled)%*%(newobs -
probB_newob
```

```
##              [,1]
## [1,] 3.036662e-10
```

```
A_newob = (0.6*probA_newob)/(0.6*probA_newob + 0.4*probB_newob)
A_newob
```

```
##           [,1]
## [1,] 0.8793235
```

```
B_newob = (0.4*probB_newob)/(0.6*probA_newob + 0.4*probB_newob)
B_newob
```

```
##           [,1]
## [1,] 0.1206765
```

D.3 let $\underset{\sim}{x}_0 = (50, 48, 47, 49)$

$$P(\underset{\sim}{x}_0 | A) = f_A(\underset{\sim}{x}_0)$$
$$= (2\pi)^{-2} |S_{pooled}|^{-1/2} \exp\left\{-\tfrac{1}{2} (\underset{\sim}{x}_0 - \bar{\underset{\sim}{x}}_1)^T S_{pooled}^{-1} (\underset{\sim}{x}_0 - \bar{\underset{\sim}{x}}_1)\right\}$$

Using R we have that the determinant of S_{pooled} is 1.003394. Furthermore we carry out the rest of the calculation in R and we get

$f_A(\underset{\sim}{x}_0) = 1.475133e-09$

Same procedure for $P(x_0|B)$ where we use $\bar{\underset{\sim}{x}}_2$ and S_{pooled} we have:

$f_B(x_0) = 3.036662e-10$

We have:

$$P(A|\underset{\sim}{x}_0) = \frac{0.6 \cdot P(\underset{\sim}{x}_0|A)}{0.6P(\underset{\sim}{x}|A) + 0.4 \cdot P(\underset{\sim}{x}|B)} \quad \text{and} \quad P(B|\underset{\sim}{x}_0) = \frac{0.4 \cdot P(\underset{\sim}{x}_0|B)}{0.6P(\underset{\sim}{x}_0|A) + 0.4 \cdot P(\underset{\sim}{x}_0|B)}$$

Subbing in for $P(\underset{\sim}{x}_0|A)$ and $P(\underset{\sim}{x}_0|B)$ for both probabilities, we get:

$P(A|\underset{\sim}{x}_0) = 0.8793235$ and $P(B|\underset{\sim}{x}_0) = 0.1206765$

Since $P(A|x_0)$ is greater than $P(B|\underset{\sim}{x}_0)$, we would classify it to pop $\simeq$ A.