MULTIVARIATE ANALYSIS I
HUDM6122 – MIDTERM EXAM

Professor: Giovanni Motta
CA: Johnny Wang

March 8, 2023

Instructions

- This exam involves 3 questions, to be solved with R.
- You must upload ONE single file onto canvas/assignments/midterm. It can be a text file, a code or a markdown.
- You have 100 minutes to solve the questions and upload your file. If you submit late, we will attach a zero to your midterm score.
- You can use the pdf file of our textbook. No other material (notes, cheat-sheets, etc.) are allowed.
- Laptops, cellphones, and tablets are not allowed during the test.
- Using the internet is strictly forbidden.
- You must turn off your mobile device.

First Name	
Last Name	

Question	Points	Score
1	22	
2	16	
3	10	
Total:	48	

1. **PCA**

The data for this question are the deaths in London from Dec 1 to 15 1952, together with the levels of smoke and sulphur dioxide over the same period.

```
Col 1= Day
Col 2= No. Deaths
Col 3= Atmospheric Smoke mg/cu. m
Col 4= Atmospheric SO2 ppm
```

To load the data, you can use the following commands

```
library(epiDisplay)
data(SO2)
```

The predictors smoke and so2 are highly collinear, and could be used to create a synthetic variable ("pollution"?) that uses the information from both predictors. A PCA on the covariance matrix should produce one component aligned with the "sliver" connecting smoke and so2 .

(a) (2 points) Obtain a scatterplot of smoke and so2. Does it make sense to perform PCA?

Solution:
```
X=SO2[,-c(1,2)]
data <- as.data.frame(X)
library(lattice)
splom(~data) # question: scatterplot
```

(b) (3 points) Using the function princomp, obtain the loadings and the scores.

Solution:
```
PCA=princomp(X)
PCA$scores
PCA$loadings
```

(c) (2 points) Using the definition on page 13 of the textbook, compute the sample covariance matrix $\mathbf{S}$ of the 15×2 matrix $\mathbf{X}$ whose columns are smoke and so2. Do not use the function cov – or any other R function.

Solution:
```
n=dim(X)[1]
q=dim(X)[2]
dm.X=matrix(scale(X,center=TRUE,scale=FALSE),n,q)
S=t(dm.X)%*%dm.X/(n-1)
```

(d) (2 points) Using the spectral decomposition of $\mathbf{S}$, obtain the loadings and the scores – you should obtain the same as in (b), up-to-sign. Hint: use the R function eigen.

Solution:
```
A=eigen(S)$vectors
Y=dm.X%*%A
round(abs(Y)-abs(PCA$scores),5)
```

(e) (1 point) Obtain the screeplot from the outcome of the function princomp.

Solution:

```
screeplot(PCA)
```

(f) (1 point) Obtain the screeplot from the spectral decomposition of $\mathbf{S}$.

Solution:

```
barplot(eigen(S)$values,xlab="components",ylab="eigenvalues")
```

(g) (2 points) Compute the variances of the principal components, and compare them with the eigenvalues of $\mathbf{S}$.

Solution:

```
apply(Y,FUN=var,MARGIN=2)
eigen(S)$values
```

(h) (2 points) From the matrix $\mathbf{S}$ in (c) compute the matrix $\mathbf{D}$, and then obtain the correlation matrix $\mathbf{R}$ according to the formula on page 14 of the textbook. Do not use the function cor – or any other R function.

Solution:

```
D=diag(diag(S))
R=sqrt(solve(D))%*%S%*%sqrt(solve(D))
```

(i) (1 point) Apply the the function cor to $\mathbf{X}$, and obtain the correlation matrix of smoke and so2.

Solution:

```
cor(X)
```

(j) (2 points) Using the function scale, compute the standardized variables

$$\boldsymbol{Z}_j = \frac{\boldsymbol{X}_j - \bar{\boldsymbol{X}}_j}{sd(\boldsymbol{X}_j)}, \qquad j = smoke, so2.$$

Compute the sample covariance matrix of the 15×2 matrix $\mathbf{Z}$ according to the definition of $\mathbf{S}$ on page 13 of the textbook (with $\mathbf{Z}$ instead of $\mathbf{X}$). You should obtain the same as in (h) and (i).

Solution:

```
Z=dm.X%*%solve(sqrt(D))
Z=scale(X,center=TRUE,scale=TRUE)
t(Z)%*%Z/(n-1)
```

(k) (2 points) Using the spectral decomposition of $\mathbf{R}$, compute the principal components of $\mathbf{Z}$. Do not use the function princomp – or any other R function.

Solution:

```
B=eigen(R)$vectors
Y=Z%*%B
```

(l) (1 point) Using the function cor, compute the correlation between smoke and the second principal component of $\mathbf{Z}$.

Solution:

```
cor(X[,1],Y[,2])
```

(m) (1 point) Compute the correlation between `smoke` and the second principal component of $\mathbf{Z}$ according to page 70 of the textbook. You should obtain the same as in (l).

Solution:

```
B[1,2]*sqrt(eigen(R)$values[2])
```

2. **CCA**

The data set contains 3 classes (setosa, versicolor, and virginica) of 50 instances each ($n = 150$), where each class refers to a type of iris plant.

```
Columns Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. species
```

Remove column 5, the species labels

```
attach(iris)
data=iris[,-5]
```

Let $\boldsymbol{x}$ be the sepal variables (columns 1-2) and $\boldsymbol{y}$ be the petal variables (columns 3-4).

(a) (6 points) Compute the matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ according to page 96 of the textbook.

Solution:

```
data <- sweep(iris[,-5], 2, sqrt(apply(iris[,-5],2,var)), FUN="/")
x <- cbind(data[,1],data[,2])
y <- cbind(data[,3],data[,4])
R.11=cor(x)
R.22=cor(y)
R.12=cor(x,y)
R.21=cor(y,x)
E1=solve(R.11)%*%R.12%*%solve(R.22)%*%R.21
E2=solve(R.22)%*%R.21%*%solve(R.11)%*%R.12
```

(b) (4 points) Compute the eigenvectors of $\mathbf{E}_1$ and $\mathbf{E}_2$, and obtain the linear combinations u_i and v_i, $i = 1, \ldots, s = \min(q_1, q_2)$.

Solution:

```
A=eigen(E1)$vectors
B=eigen(E2)$vectors
u=x%*%A
v=y%*%B
```

(c) (1 point) Apply the function `cor` to $\boldsymbol{u}$ and $\boldsymbol{v}$ to obtain the canonical correlations R_i, $i = 1, \ldots, s = \min(q_1, q_2)$, according to page 96 of the textbook.

Solution:

```
cor(u[,1],v[,1])
cor(u[,2],v[,2])
```

(d) (1 point) Apply the function `cancor` to $\boldsymbol{x}$ and $\boldsymbol{y}$ to obtain the canonical correlations R_i, $i = 1, \ldots, s = min(q_1, q_2)$. You should obtain the same as in (c).

Solution:

```
cancor(x,y)$cor
```

Apply the test proposed by Bartlett (1947), in the same way as you did for Exercise 3.4 on page 103 of the textbook.

(e) (2 points) Compute the test statistic.

Solution:

```
n   = dim(data)[1]
q1 = dim(x)[2]
q2 = dim(y)[2]
test.stat = -(n-0.5*(q1+q2+1))*sum(log(1-eigen(E1)$values))
test.stat
```

(f) (1 point) What is the null hypothesis?

Solution: There is no significant canonical correlation.

(g) (1 point) What is the outcome of the test. Interpret.

Solution:

```
P.value <- pchisq(test.stat, df = q1*q2, lower.tail=F)
P.value
```

We reject H_0, that is, there is at least one significant canonical correlation.

3. **MDS**

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. It is a data frame with $n = 50$ observations on $q = 4$ variables.

```
data(USArrests)
data=USArrests
```

(a) (1 point) Compute the Euclidean proximity matrix.

Solution:

```
distance_matrix <- dist(data)
```

(b) (1 point) Using the R function cmdscale with $m = q$, compute the $n \times m$ coordinate values from the observed proximity matrix. Note that the argument k of the R function cmdscale corresponds to the parameter m on page 106.

Solution:

```
n=dim(data)[1]
q=dim(data)[2]
X.mds <- cmdscale(distance_matrix,k=q)
```

(c) (2 points) Plot the first $n \times 2$ coordinate values. Can you interpret the results?

Solution:

```
plot(X.mds[,1],X.mds[,2], type = "n")
text(X.mds[,1],X.mds[,2], labels = row.names(USArrests))
```

(d) (1 point) Compute $\mathbf{X}$, the $n \times q$ matrix containing the demeaned data.

Solution:

```
dm.X=scale(data,center=TRUE,scale=FALSE)
```

(e) (1 point) Compute $\mathbf{B} = \mathbf{X}\mathbf{X}^\top$, the $n \times n$ matrix in (4.1) on page 107 of the textbook.

Solution:

```
B=dm.X%*%t(dm.X)
```

(f) (2 points) Compute the spectral decomposition of $\mathbf{B}$, and obtain the $n \times q$ matrix $\mathbf{V}_1$ and the $q \times q$ diagonal matrix $\mathbf{\Lambda}_1$ according to the first two lines of page 109 of the textbook.

Solution:

```
V1=eigen(B)$vectors[,1:q]
L1=diag(eigen(B)$values[1:q])
```

(g) (2 points) Using $\mathbf{V}_1$ and $\mathbf{\Lambda}_1$ obtained in (f), compute the matrix $\mathbf{X}$ according to the first two lines of page 109 of the textbook. You should obtain the same matrix as in (d).

Solution:

```
X=V1%*%sqrt(L1)
round(abs(X-X.mds[,1:q]),5)
```