Practicals-Sessional 1

Principal Component Analysis-Answers

Your Name (Reg.No)

Follow the instructions in the chunks and answer the questions.

Consider the dataset nutritian.xlsx' which contains nutitional data of the food taken by 8618 indivituals. The variables "Protein_g" "Fat_g" "Carb_g" "Sugar_g" "VitB6_mg" "VitB12_mcg" "VitE_mg" are self explanatory and the units of the variables are also given with variable name splitted by_(For example_g` is the variable is mesured in grams).

```r
library(readxl)
#Load the data
data = read_excel('nutritian.xlsx')
head(data)
```

Protein_g	Fat_g	Carb_g	Sugar_g	VitB6_mg	VitB12_mcg	VitE_mg
0.85	81.1	0.06	0.06	0.003	0.17	2.32
0.85	81.1	0.06	0.06	0.003	0.13	2.32
0.28	99.5	0.00	0.00	0.001	0.01	2.80
21.40	28.7	2.34	0.50	0.166	1.22	0.25
23.24	29.7	2.79	0.51	0.065	1.26	0.26
20.75	27.7	0.45	0.45	0.235	1.65	0.24

Question 1: (5 marks)

(a) What is the mean vector of the dataset? (2marks)

```r
#Code here
X = data
n = nrow(X)
meanVec = 1/n * t(X) %*% rep(1, n)
meanVec
```

Protein_g	11.524
Fat_g	10.647
Carb_g	21.819
Sugar_g	6.560
VitB6_mg	0.264
VitB12_mcg	1.225
VitE_mg	0.872

(b) Find the **Sample covariance matrix** of the dataset using matrix algebra. (Note that, you are not allowed to use the command `cov(data)` to find the covariance) (3marks)

```
#Code here
Y = as.matrix(X - rep(1, n) %*% t(meanVec))
Sample.Cov = 1/(n-1) *t(Y) %*% Y
Sample.Cov
```

	Protein_g	Fat_g	Carb_g	Sugar_g	VitB6_mg	VitB12_mcg	VitE_mg
Protein_g	111.31	9.159	-86.78	-38.198	1.154	11.184	-1.197
Fat_g	9.16	251.741	-23.39	-0.423	-0.357	-1.424	20.642
Carb_g	-86.78	-23.388	741.96	227.990	2.550	-11.134	7.370
Sugar_g	-38.20	-0.423	227.99	185.017	0.578	-2.964	3.458
VitB6_mg	1.15	-0.357	2.55	0.578	0.229	0.546	0.527
VitB12_mcg	11.18	-1.424	-11.13	-2.964	0.546	18.655	1.008
VitE_mg	-1.20	20.642	7.37	3.458	0.527	1.008	14.815

Question 2. (15 marks)

Perform principal component analysis. (Use the Covariance matrix obtained in the Quesiton 1b) Note that the following are to be done to answer this question:

(a) Find the eigenvalues and eigenvectors of the covariance matrix (3marks)

(b) Tabularise the eigenvalues, proportion of total variance explained by each variable and the cumulative propotion. (6marks)

(c) Decide how many principal components to be retained if the researcher wants 98% or more of the total sample variance to be explained? (2mark)

(d) Conform the same by plotting screeplot. (2marks)

(e) What are the principal components found in (c) and (d)? (Write the principal components in the sheet provided or type here.)(2marks)

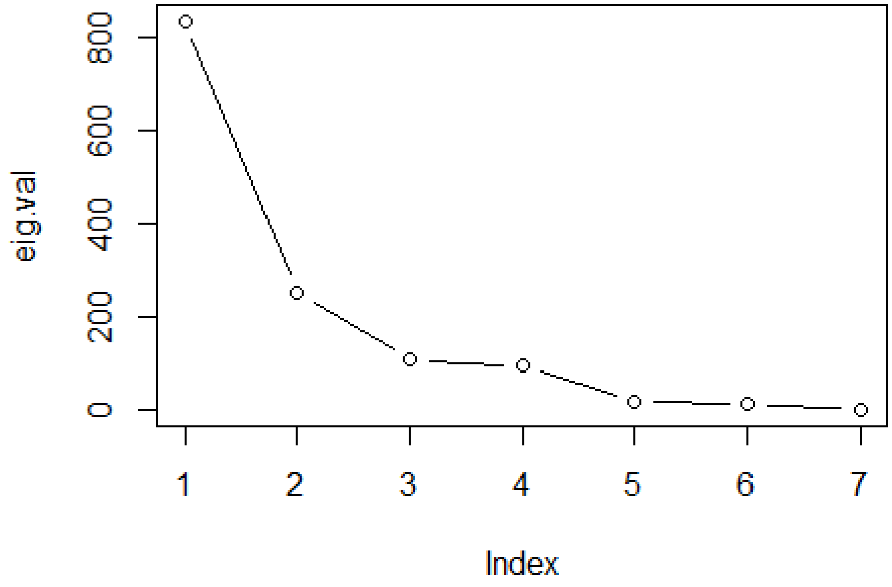
```
#Find the eigenvalues and eigenvectors. Note that the function which you use
to find the eigenvales and eigenvectors already order the eigenvalues in the
desending order.
eigen = eigen(Sample.Cov)
eigen

## eigen() decomposition
## $values
## [1] 836.925 253.277 108.640  94.533  17.666  12.527   0.168
##
## $vectors
##           [,1]      [,2]     [,3]     [,4]     [,5]      [,6]      [,7]
## [1,] -0.12986 -0.032180  0.60951  0.77193  0.11393  0.039946 -0.012511
## [2,] -0.03920 -0.993537  0.01576 -0.05792  0.01373 -0.086814  0.004529
## [3,]  0.93261 -0.023807  0.34171 -0.11274 -0.00412 -0.012266 -0.004932
## [4,]  0.33392 -0.060029 -0.71284  0.61346  0.02059  0.000842  0.000767
## [5,]  0.00291  0.000707  0.01094  0.01095 -0.03090  0.032507  0.998869
## [6,] -0.01559  0.006012  0.05674  0.10740 -0.95331 -0.275066 -0.022299
## [7,]  0.00895 -0.087422 -0.00307 -0.00897 -0.27682  0.956033 -0.039509
```

Table of Proportion of Total Variance Explained.

```
eig.val = eigen$values
prop.eig = eig.val/sum(diag(Sample.Cov))
cumSum = cumsum(prop.eig)
Tab.prop = data.frame('EigenValues' = eig.val, "Proportion" = prop.eig,
"Cumulative Proportion" = cumSum)
Tab.prop
```

EigenValues	Proportion	Cumulative.Proportion
836.925	0.632	0.632
253.277	0.191	0.824
108.640	0.082	0.906
94.533	0.071	0.977
17.666	0.013	0.990
12.527	0.009	1.000
0.168	0.000	1.000

```
#Plot the Scree plot here
plot(eig.val, type = 'b')
```



Since the elbow bend is at 5th principal component, the first 4 principal components itself will explain the data better.

Answer for (e):

List the four principal components.