Computation in Data Science HW1

R09946006 | 何青儒 | HO, Ching-Ru | Oct. 31, 2020

- Multiply all the numbers listed in the sixth column of Table 8.2 by 0.1.

```
$ data <- read.csv("table8.2.csv", header = TRUE, fill = FALSE)
$ data_mod <- data[2:6]
$ data_mod[5] <- data_mod[5]*0.1
```

1. Construct the sample covariance matrix, S
   - Covariance Matrix, $S = \begin{bmatrix} 4.308 & 1.684 & 1.803 & 2.155 & -0.025 \\ 1.684 & 1.767 & 0.588 & 0.178 & 0.018 \\ 1.803 & 0.588 & 0.801 & 1.065 & -0.016 \\ 2.155 & 0.178 & 1.065 & 1.969 & -0.036 \\ -0.025 & 0.018 & -0.016 & -0.036 & 0.005 \end{bmatrix}$

```
$ S <- round(var(data_mod), 3)
$ View(S)
```

2. Obtain the eigenvalue-eigenvector pairs and the first two sample principal components for the covariance matrix in Part a.
   - Eigenvalue: $\hat{\lambda_1} = 6.913$, $\hat{\lambda_2} = 1.689$, $\hat{\lambda_3} = 0.230$, $\hat{\lambda_4} = 0.015$, $\hat{\lambda_5} = 0.004$
   - Eigenvector:

     $\hat{e_1} = \begin{bmatrix} 0.783 \\ 0.309 \\ 0.335 \\ 0.424 \\ -0.005 \end{bmatrix}$, $\hat{e_2} = \begin{bmatrix} -0.060 \\ -0.785 \\ 0.091 \\ 0.610 \\ -0.021 \end{bmatrix}$, $\hat{e_3} = \begin{bmatrix} 0.540 \\ -0.537 \\ 0.051 \\ -0.646 \\ -0.003 \end{bmatrix}$, $\hat{e_4} = \begin{bmatrix} 0.302 \\ 0.003 \\ -0.930 \\ 0.176 \\ 0.110 \end{bmatrix}$, $\hat{e_5} = \begin{bmatrix} -0.029 \\ -0.017 \\ 0.107 \\ -0.006 \\ 0.994 \end{bmatrix}$
   - First two sample principle components:
     - $\hat{y_1} = 0.783x_1 + 0.309x_2 + 0.335x_3 + 0.424x_4 - 0.005x_5$
     - $\hat{y_2} = -0.060x_1 - 0.785x_2 + 0.091x_3 + 0.610x_4 - 0.021x_5$

```
$ eigen_original <- eigen(S)
$ eigen_value <- round(eigen$values, 3)
$ eigen_vector <- round(eigen_original$vectors, 3)

$ View(eigne_value)
>> [1] 6.913 1.689 0.230 0.015 0.004

$ View(eigne_vector)
>>      [,1]   [,2]   [,3]   [,4]   [,5]
[1,]  0.783 -0.060  0.540  0.302 -0.029
[2,]  0.309 -0.785 -0.537  0.003 -0.017
[3,]  0.335  0.091  0.051 -0.930  0.107
[4,]  0.424  0.610 -0.646  0.176 -0.006
[5,] -0.005 -0.021 -0.003  0.110  0.994
```

3. Compute the proportion of total variance explained by the first two principal components obtained in Part b. Calculate the correlation coefficients, $r_{\hat{y}_i, x_k}$ and interpret these components if possible. Compare your results with the results in Example 8.3. What can you say about the effects of this change in scale on the principal components?

   - Correlation Matrix, $R = \begin{bmatrix} 0.992 & -0.038 & 0.125 & 0.018 & -0.001 \\ 0.611 & -0.767 & -0.194 & 0.000 & -0.001 \\ 0.984 & 0.132 & 0.027 & -0.126 & 0.007 \\ 0.794 & 0.565 & -0.221 & 0.015 & -0.000 \\ -0.186 & -0.386 & -0.020 & 0.188 & 0.880 \end{bmatrix}$
   - The proportion of total sample variance explained by the first two principle Components is $0.781 + 0.1909 = 0.9719$ as below.
   - Compare to the result in Example 8.3 (lecture slide page 30), value of PC1 become bigger, from 74.1% up to 78.1%, however, PC2 does not have change, from 19.1 to 19.09 (usually round to 19.1). After scaling, variable x_5 (median value of home) has much more influence in the first principle component, making the percentage value become bigger.

```
$ prcomp(data_mod)
>> Standard deviations (1, .., p=5):
[1] 2.62912099 1.29978551 0.47956535 0.12036221 0.06344938

Rotation (n x k) = (5 x 5):
                                            PC1          PC2
Total_population_.thousands.          -0.782561519  0.05945889
Median_school_years                   -0.309002431  0.78507128
Total_employment_.thousands.          -0.334535931 -0.09057822
Health_services_employment_.hundreds. -0.424470512 -0.60949178
Median_value_home_.10000s.             0.005045987  0.02104967
                                            PC3          PC4
Total_population_.thousands.           0.540387761 -0.302462133
Median_school_years                   -0.536530093 -0.003400582
Total_employment_.thousands.           0.051244623  0.932131302
Health_services_employment_.hundreds. -0.646123390 -0.175650143
Median_value_home_.10000s.            -0.003883256 -0.093719568
                                            PC5
Total_population_.thousands.          -0.023660839
Median_school_years                   -0.017449491
Total_employment_.thousands.           0.091577958
Health_services_employment_.hundreds. -0.004018108
Median_value_home_.10000s.             0.995355722

$ summary(data.pca)
>> Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     2.629 1.2998 0.47957 0.12036 0.06345
Proportion of Variance 0.781 0.1909 0.02599 0.00164 0.00045
Cumulative Proportion  0.781 0.9719 0.99791 0.99955 1.00000
```