Name: Ethan Tan

Admin: p2012085

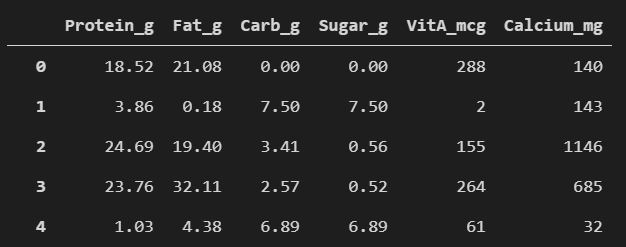
Class: DAAA/2A/03

Subject: Mathematics for A.I. Assignment 1

# Question 1

## Part (a)

### Should PCA be carried out on covariance or correlation matrix? Explain.

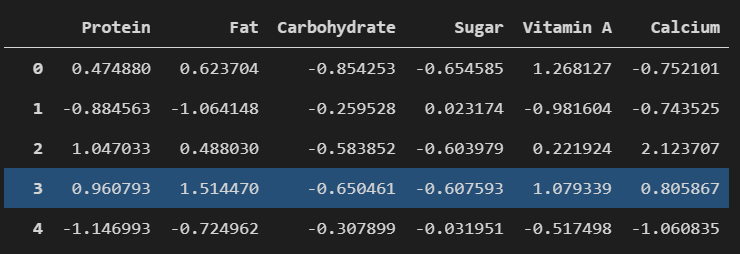


PCA should be carried out on the correlation matrix. All the variables (nutrients) measure mass. Although they currently have different scales (g/mg/mcg), it won't matter when they are standardized.

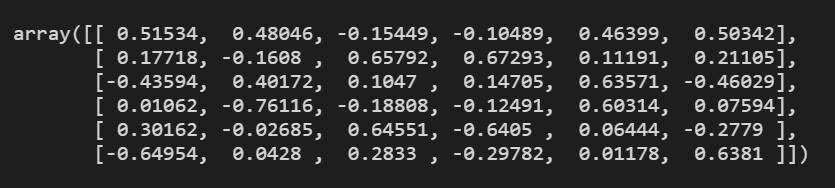
## Part (b)

### Extract the principal components. Justify your decision and interpret the principal components. You should include the necessary tables, outputs and graphs.

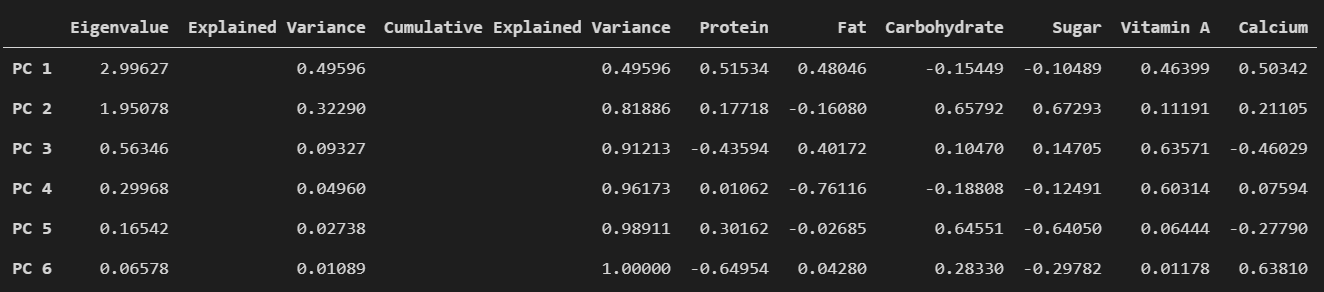
### Step 1: Scale the data



### Step 2: Perform Eigendecomposition



### Step 3: Summarize the Principal Components



### Step 4: Select the Principal Components

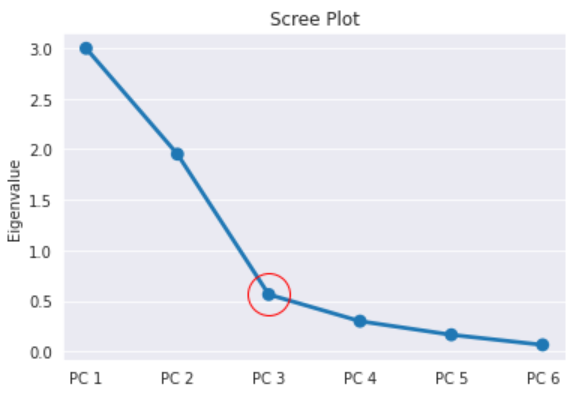
#### Method 1 - Kaiser's Rule

PCA was carried out on the correlation matrix, so Kaiser's Rule can be used in this case. By Kaiser's Rule, only the top 2 PCs should be kept.

#### Method 2 - Cumulative Explained Variance

The top 2 PCs explain more than 80% (81%) of the total variance; hence, only the top 2 PCs should be retained.

#### Method 3 - Scree Plot



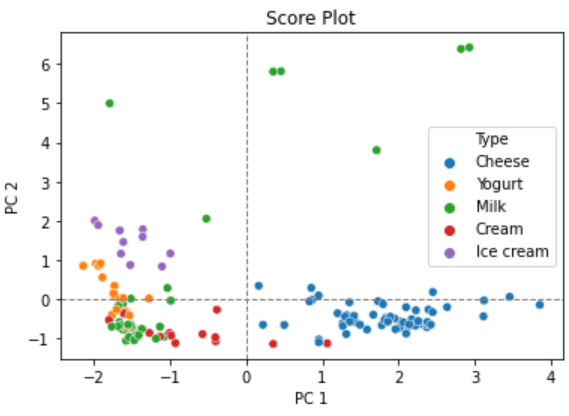
By the scree plot, there is an elbow at the 3rd PC. Therefore, only the top 2 PCs should be kept.

#### Summary of Principal Component Selection

All 3 methods agree that the top 2 Principal Components should be kept.

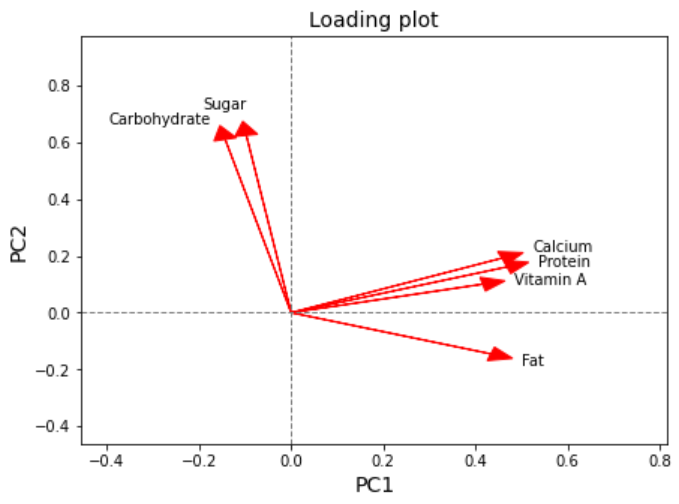
### Step 5: Plot the Transformed Data

#### Score Plot



### Step 6: Interpret the Principal Components

#### Loading Plot



From the loading plot, the interpretations of the PCs are as follows.

1. PC 1:
   * PC 1 seems to measure the contrast between the concentrations of sugar and non-sugar nutrients
   * Carbohydrate and Sugar are sugar nutrients
   * Carbohydrate and Sugar have negative loading values for PC 1, while
   * Calcium, Protein, Vitamin A and Fat have positive loading values for PC 1
   * A higher score for PC 1 means that the dairy product has a greater concentration of non-sugar nutrients as compared to sugar nutrients
   * A lower score for PC 1 means that the dairy product has a greater concentration of sugar nutrients as compared to non-sugar nutrients
2. PC 2:
   * PC 2 seems to measure the contrast between the concentrations of fats and the other nutrients
   * Fat is the only variable with a negative loading value
   * A higher score for PC 2 means that the dairy product has a lower concentration of fats as compared to other nutrients
   * A lower score for PC 2 means that the dairy product has a greater concentration of fats as compared to other nutrients

## Part (c)

### Which type(s) of dairy product has/have the following attributes? Explain your answer with the aid of a suitable graph with colour or marker to display "Type" information.

### Low carbohydrates and sugar but high in other nutrients.

Cheese. The cheese (blue) products have generally high PC 1 scores. Recalling the interpretations of the PCs, "a higher score for PC 1 means that the food has a greater concentration of non-sugar nutrients as compared to sugar nutrients". Therefore, the dairy product(s) with low carbohydrates and sugar but high in other nutrients should have a high PC 1 score.

### High carbohydrates and sugar but low in other nutrients.

Yogurt and ice cream. The yogurt (orange) and ice cream (purple) products have generally low PC 1 scores. Recalling the interpretations of the PCs, "a lower score for PC 1 means that the food has a greater concentration of sugar nutrients as compared to non-sugar nutrients". Therefore, the dairy product(s) with high carbohydrates and sugar but low in other nutrients should have low PC 1 scores. The milk (green) and cream (red) products have high variances for PC 1 scores, so they cannot be considered to have high carbohydrates and sugar but low in other nutrients.

## Part (d)

### A dairy product has its nutritional value listed below. Which type of dairy product is it likely to be? Show your working and explain.

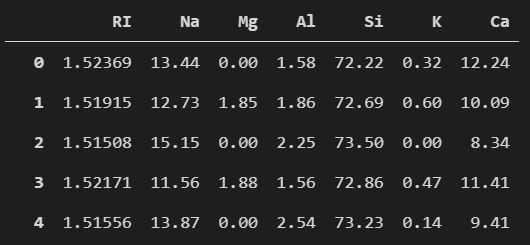
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Protein | 4.8 g | Sugar | 19.1 g |  |
| Fat | 23 g | Vitamin A | 17 mcg |
| Carbohydrate | 26.2 g | Calcium | 121 mg |

The dairy product is likely to be ice cream. The given (above) data was scaled and transformed by PCs 1 and 2, and plotted on a labelled score plot, together with the rest of the data points. As the target datum point (blue) was situated in the `Ice cream` (orange) cluster, it is likely that the dairy product is an ice cream.

# Question 2

## Part (a)

### Should PCA be carried out on covariance or correlation matrix? Explain.

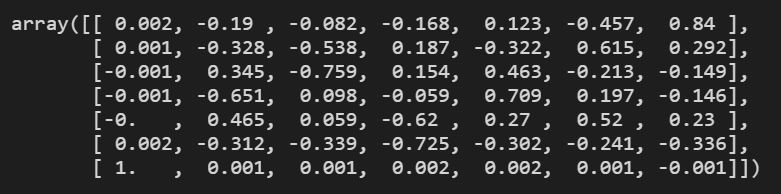


PCA should be carried out on the covariance matrix. RI measures the ratio of the velocity of light in a vacuum to its velocity in the objects. On the other hand, Na, Mg, Al, Si, K and Ca measure the proportions of their corresponding oxides in the objects by weight. As RI and the 6 other quantitative variables have different scales and measure different things, it doesn't make sense to carry out PCA on the correlation matrix. Therefore, PCA should be carried out on the covariance matrix.

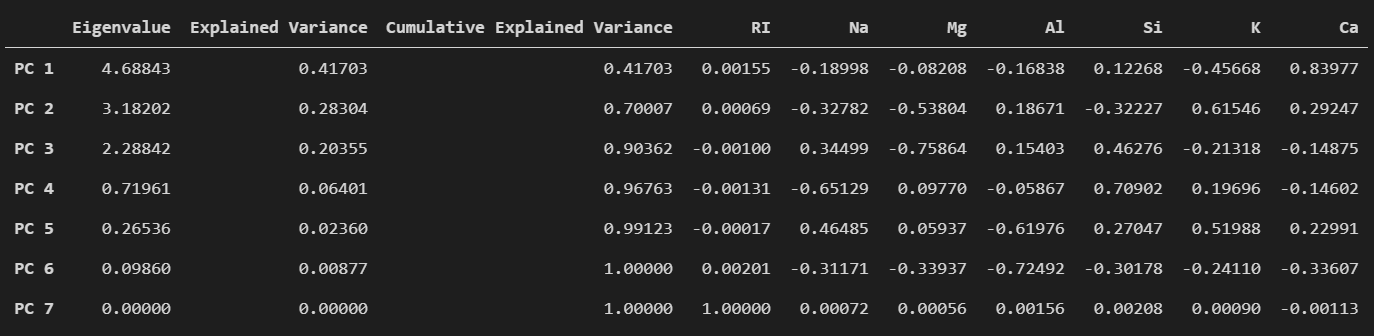
## Part (b)

### Extract the principal components. Justify your decision and interpret the principal components. You should include the necessary tables, outputs and graphs.

### Step 1: Perform Eigendecomposition



### Step 2: Summarize the Principal Components



### Step 3: Select the Principal Components

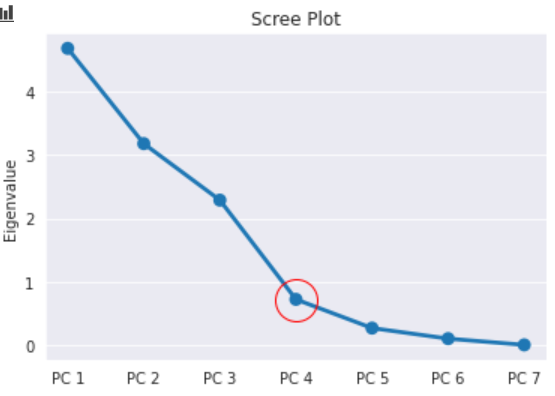
#### Method 1 - Kaiser's Rule

PCA was not carried out on a covariance matrix, so Kaiser's Rule is not applicable in this case.

#### Method 2 - Cumulative Explained Variance

The top 3 PCs explain more than 80% (90%) of the total variance; hence, only the top 3 PCs should be retained.

#### Method 3 - Scree Plot



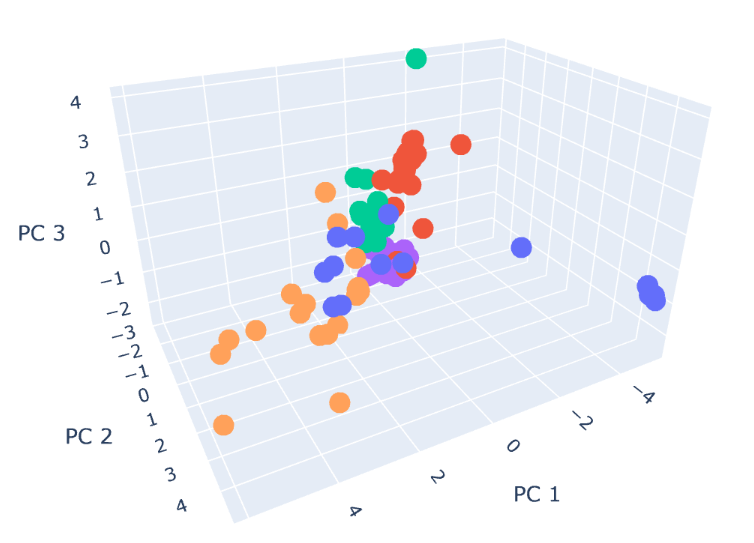
By the scree plot, there is an elbow at the 4th PC. Therefore, only the top 3 PCs should be kept.

#### Summary of Principal Component Selection

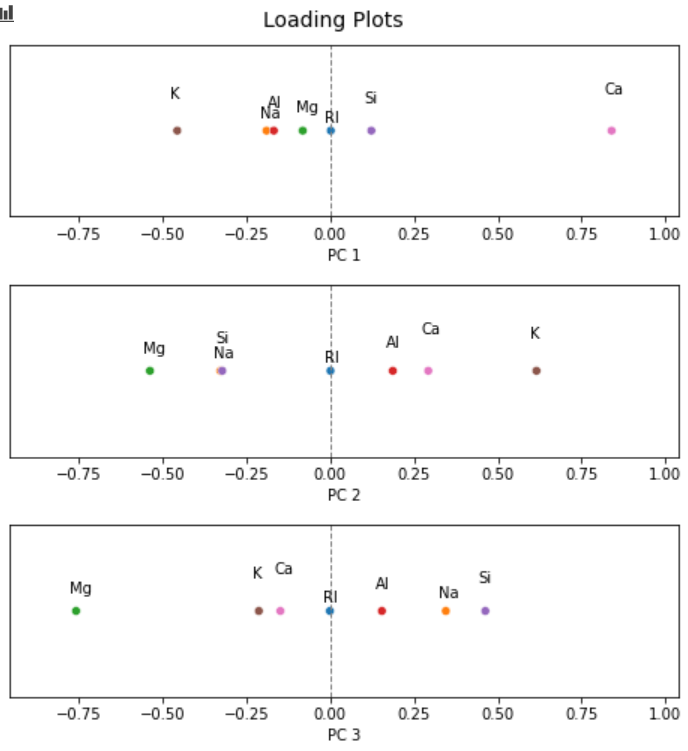
Both methods 2 and 3 agree that the top 3 Principal Components should be kept.

### Step 4: Plot the Transformed Data

#### Score Plot



#### Loading Plot



From the loading plots, the interpretations of the PCs are as follows.

1. PC 1:
   * PC 1 seems to measure the contrast between the concentrations of silicon oxide (SiO2) and calcium oxide (CaO) against the concentrations of potassium oxide (K2O), sodium oxide (Na2O), aluminium oxide (Al2O3) and magnesium oxide (MgO)
   * `Ca` and `Si` have positive loading values while
   * `K`, `Na`, `Al` and `Mg` have negative loading values
   * A higher PC 1 score means that the glass object has a higher concentration of silicon oxide and calcium oxide as compared to the rest of the oxides involved
   * A lower PC 1 score means that the glass object has a lower concentration of silicon oxide and calcium oxide as compared to the rest of the oxides involved
2. PC 2:
   * PC 2 seems to measure the contrast between the concentrations of potassium oxide, calcium oxide and aluminium oxide against the concentrations of magnesium oxide, sodium oxide and silicon oxide
   * `K`, `Ca` and `Al` have positive loading values while
   * `Na`, `Si` and `Mg` have negative loading values
   * A higher PC 2 score means that the glass object has a higher concentration of potassium oxide, calcium oxide and aluminium oxide as compared to the rest of the oxides involved
   * A lower PC 2 score means that the glass object has a lower concentration of potassium oxide, calcium oxide and aluminium oxide as compared to the rest of the oxides involved
3. PC 3:
   * PC 3 seems to measure the contrast between the concentrations of aluminium oxide, sodium oxide and silicon oxide against the concentrations of magnesium oxide, calcium oxide and potassium oxide
   * `Al`, `Na` and `Si` have positive loading values while
   * `Mg`, `Ca` and `K` have negative loading values
   * A higher PC 3 score means that the glass object has a higher concentration of aluminium oxide, sodium oxide and silicon oxide as compared to the rest of the oxides involved
   * A lower PC 3 score means that the glass object has a lower concentration of aluminium oxide, sodium oxide and silicon oxide as compared to the rest of the oxides involved

`RI` has a very small absolute loading value for all 3 PCs. This means that `RI` adds little to no information. Thus, it can be considered a redundant variable in this case study.

## Part (c)

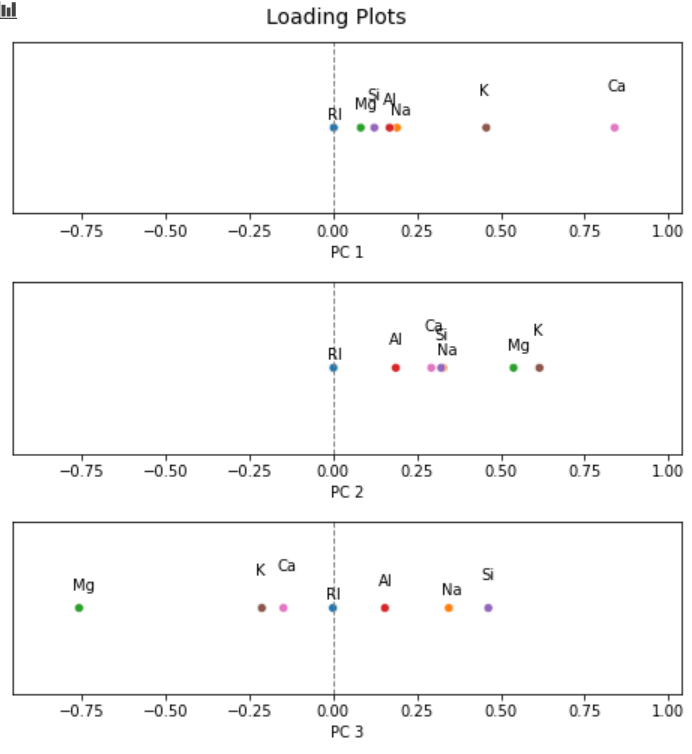
### The following shows the attributes of a glass object. Which class does it likely belong to? Explain your answer with the aid of a suitable graph with colour or marker to display "Class" information.

|  |  |  |  |
| --- | --- | --- | --- |
| RI | 1.51641 | Si | 73.05 |
| Na | 13.04 | K | 0.53 |
| Mg | 3.5 | Ca | 8.6 |
| Al | 1.28 |  |  |

The glass object likely belongs to the `VehWin` (Vehicle Window) class. The given (above) data was transformed by PCs 1 and 2, and plotted on a 3D labelled score plot, together with the rest of the data points. As the target datum point (light blue) was situated in the `VehWin` (purple) cluster, it is likely that the glass object belongs to the `VehWin` class.

## Part (d)

### Explain how PC3 is advantageous over the first two principal components.



PC3 captures the contrast between the more important variables and less important variables of PC1 and PC2.

* For PC1, `Ca` and `K` have the 2 highest absolute loading values
* For PC2, `K` and `Mg` have the 2 highest absolute loading values
* These 3 variables (`Mg`, `K` and `Ca`) have negative loading values for PC3
* The rest of the variables (`Si`, `Al` and `Na`) have positive loading values for PC3
* A higher PC3 score means that the glass object scored lower for PCs 1 and 2 ???????