ISYE 6740 Homework 2

Q1-1

lets consider a graph with $A_1 \dots A_m$ components.
we can consider the Adjancy matrix and degree matrix as below:

$$\begin{bmatrix} \boxed{D_1} & & & \\ & \boxed{D_2} & & \\ & & \ddots & \\ & & & \boxed{D_m} \end{bmatrix} \quad \begin{bmatrix} \boxed{A_1} & & & \\ & \boxed{A_2} & & \\ & & \ddots & \\ & & & \boxed{A_m} \end{bmatrix}$$

because each component is connected with minimum one edge then we can consider each A_1 to A_m and D_1 to D_m as block of semmetric matrix. $\boxed{\ddots}$

hence $L = \begin{bmatrix} L_1 & & & \\ & L_2 & & \\ & & \ddots & \\ & & & L_m \end{bmatrix}$. ①

Now lets consider an eigendecomposition of L_x for eigenvalue zero and eigen vector $\underline{V}$ $(V \in R^n)$

$$\lambda = 0 = V^T L_x V = V^T D_x V - V^T A_x V$$

$$= \sum_{i=1}^{n} d_i V_i^2 - \sum_{ij} a_{ij} V_i V_j$$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} d_i V_i^2 - 2\sum_{ij} a_{ij} V_i V_j + \sum_{j=1}^{n} d_j V_j^2\right)$$

$$= \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}(V_i - V_j)^2 \quad ②$$

$$\Rightarrow \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}(v_i - v_j)^2 = 0 \Rightarrow \sum_{i,j=1}^{n} \overbrace{a_{ij}}^{\text{positive}}(v_i - v_j)^2 = 0$$

the equation above can only equals to zero if $v_i = v_j$ for each i & j.

this means for component x, the eigent vector is scaled on identity vector.

if $v_i = v_j = v_x \Rightarrow$ eigen vector $= v_x I_x$ (I_x the indicator vector)

we can generalize this concep to all components

$\boxed{(v_1 \cdots v_m) \text{ are values}}$

$$V^T L V = \begin{bmatrix} v_1 I_{A_1} & & & \\ & v_2 I_{A_2} & & \\ & & \ddots & \\ & & & v_m I_{A_m} \end{bmatrix}$$

this shows that the eigenvectors to eigenvalue zero are linear combination of $I_{A_1}, \ldots, I_{A_m}$.

Q1-1

I used both my Kmeans and the Kmeans from scikit-learn package. In both cases the false classification rate was close to 5.2%. For details please check the Jupyter file attached to this report.

Q2-1

we have done the PCA in several Steps.
we also cleaned up the data and we got to 13 countries and 20 types of foods. using same model we used in the class (slides)

$m = 13 \quad d = 20$

$x^1, \ldots, x^{13} \in R^{d=20} \Rightarrow$ matrix $X \in R^{13 \times 20}$

next step: calculate the mean an covariance of the matrix

$$\mu = \frac{1}{m}\sum_{i=1} x^i \quad \& \quad C = \frac{1}{m}\sum_{i=1}^{m}(x^i-\mu)(x^i-\mu)^T$$

next step: find the eigenvectors $w_1, w_2, \ldots$ of the covariance matrix and pick the largest eigen value(λ), then the next one, then the next, ... Depend on the dimension we want to reduce to.

next step: computed the principle components

$$z_i = \begin{pmatrix} w_1^T (x^i-\mu)/\sqrt{\lambda_1} \\ w_2^T (x^i-\mu)/\sqrt{\lambda_2} \\ \vdots \end{pmatrix}$$

Q2-2

lets assume we have m data points $x^1, \dots, x^m \in R^d$
and the mean is $\mu = \frac{1}{m} \sum_{i=1}^{m} x^i$.

we will find direction w in the space R^d in which
$||w|| \leq 1$ and it captures the maximum
variance. therefore:

$$\max_{w: ||w|| \leq 1} \underbrace{\frac{1}{m} \sum_{i=1}^{m} (w^T x^i - w^T \mu)^2}_{\text{maximize it.}}$$

$$\frac{1}{m} \sum_{i=1}^{m} (w^T x^i - w^T \mu)^2 = \frac{1}{m} \sum_{i=1}^{m} (w^T (x^i - \mu))^2$$

$$= \frac{1}{m} \sum_{i=1}^{m} w^T (x^i - \mu)(x^i - \mu)^T w$$

$$= w^T \left(\frac{1}{m} \sum_{i=1}^{m} (x^i - \mu)(x^i - \mu)^T \right) w = w^T \underset{\text{covariance}}{C} w$$

the optimization

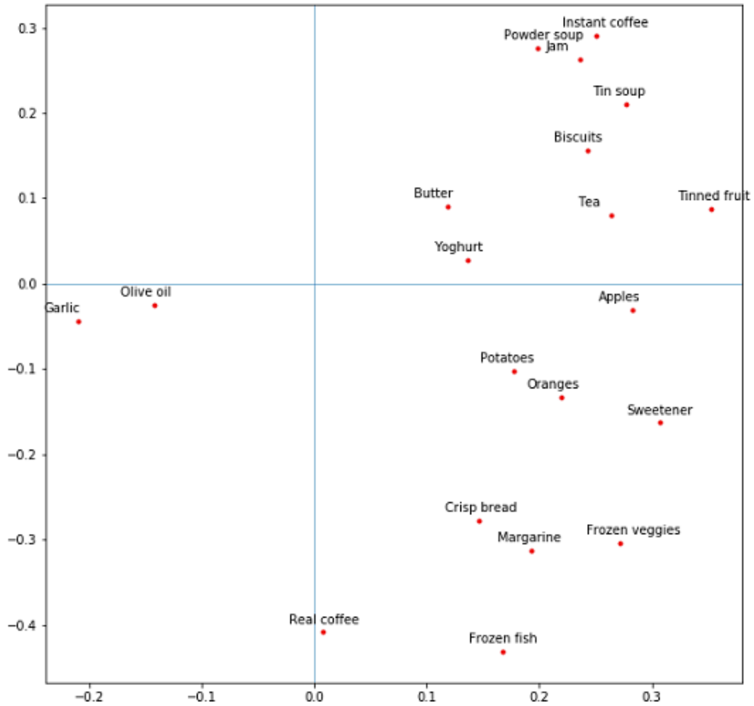
$$\boxed{\max_{w: ||w|| \leq 1} w^T C w}$$

if λ is the eigenvalue for matrix covariance C then:

$$CW = \lambda W$$

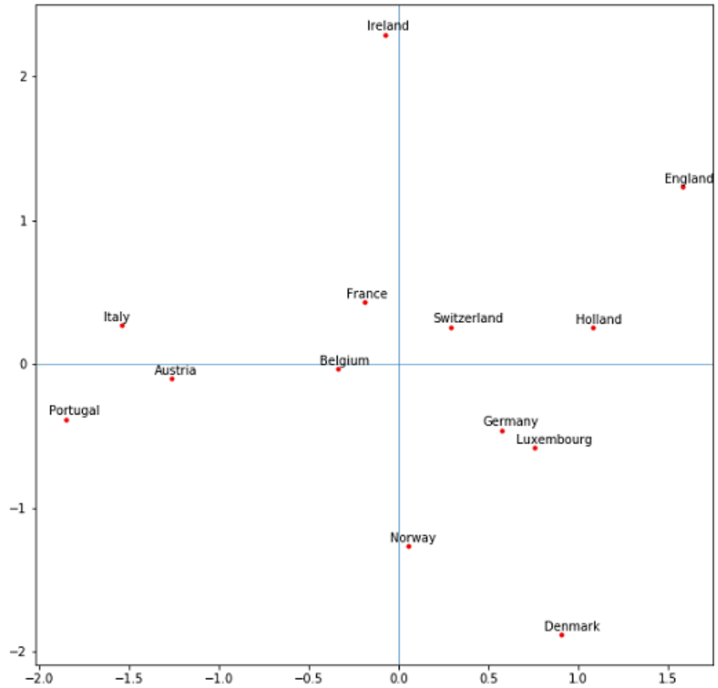
Therfore to get top k principle components all we need to do is to find the eigenvalues and its related eigenvectors. Pick the largest eigenvalue λ_1 and related eigenvector as first component, the second λ_2 and its eigenvector as second principle, ...

Q2-3



This figure shows the correlation between various foods in the reduced data components.
Tea and biscuits are positively correlated therefore, countries (such as England) uses biscuits and tea, the more biscuits are consumed, the more team is also consumed.
It also shows that places that consumes foods such as olive oil which is the opposite to for example apple, they consume less apple. Similarly, if the food are opposite each other from Y axis.

Q2-4



This figure shows the reduced data to two components. Countries close to each other, has similar daily diet. For example, Norway Denmark or Switzerland and Holland.
Countries far from each other has different diets. For example, Italy compared to Denmark. Italy consumes more garlic and olive oil while Denmark consumes more frozen fish and frozen veggie.