# HW 1


Daniel Brigham Évora 99473
Natacha Sousa 107413


**I. Pen-and-paper** [9v]

Consider the bivariate observations $\{\mathbf{x}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix}0\\2\end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix}3\\-1\end{bmatrix}\}$ and the multivariate

Gaussian mixture given by

$$\mathbf{u}_1 = \begin{bmatrix}2\\-1\end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix}1\\1\end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix}4 & 1\\1 & 4\end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}, \quad \pi_1 = 0.5, \quad \pi_2 = 0.5$$

Answer the following questions by presenting all intermediary steps, and use 3 decimal places in each.

1) [6v] Perform two epochs of the EM clustering algorithm and determine the new parameters.

## ① E step:

$$p(c_k, \mathbf{x}_i) = p(\mathbf{x}_i | c_k) p(c_k) = N(\mathbf{x}_i | \mathbf{u}_k, \Sigma_k) \bullet \pi_k$$

and then normalize, $\gamma_{ki} = p(c_k \mid \mathbf{x}_i) = \dfrac{p(c_k, \mathbf{x}_i)}{\sum_j p(c_j, \mathbf{x}_i)}$

$$p(\mathbf{x} | c_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{u}_k)^T \Sigma_k^{-1} (\mathbf{x}-\mathbf{u}_k)}$$

### Cluster 1:

$$P(C_1|x_1) = P(x_1|C_1)P(C_k) = N(x_1 | \mu_1, \Sigma_1) \cdot \pi_1$$

$x_1 = \begin{bmatrix}1\\0\end{bmatrix}$, $\mu_1 = \begin{bmatrix}2\\-1\end{bmatrix}$, $\pi_1 = 0{,}5$

$$= \left( \frac{1}{2\pi \quad |\Sigma_1|^{1/2}} e^{-\frac{1}{2}(x_1-\mu_1)^T \Sigma_1^{-1} (x_1-\mu_1)} \right) \pi_1$$

$\Sigma_1 = \begin{bmatrix}4 & 1\\1 & 4\end{bmatrix}$

$$= \left[ \frac{1}{2\pi} \frac{1}{(16-1)^{1/2}} e^{-\frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix}^T \frac{1}{15}\begin{bmatrix}4 & -1\\-1 & 4\end{bmatrix}\begin{bmatrix}-1\\1\end{bmatrix}\right)} \right] \pi_1$$

$$= \left( \frac{1}{2\pi} \frac{1}{(15)^{1/2}} e^{-\frac{10}{30}} \right) \pi_1$$

$P(c_1|x_1) = 1{,}472 \times 10^{-2}$

$x_2 = \begin{bmatrix}0\\2\end{bmatrix}$

$P(C_1|x_2) = N(x_2 | \mu_1, \Sigma_1) \cdot \pi_1 = 2{,}4336 \times 10^{-3}$

$x_3 = \begin{bmatrix}3\\-1\end{bmatrix}$

$P(C_1|x_3) = N(x_3 | \mu_1, \Sigma_1) \cdot \pi_1 = 1{,}798 \times 10^{-2}$

### Cluster 2:

$P(c_2|x_1) = N(x_1 | \mu_2, \Sigma_2) \cdot \pi_2 = 3{,}0987 \times 10^{-2}$

$P(C_2|x_2) = N(x_2 | \mu_2, \Sigma_2) \cdot \pi_2 = 2{,}413 \times 10^{-2}$

$P(C_2|x_2) = N(x_3 | \mu_2, \Sigma_2) \cdot \pi_2 = 5{,}3948 \times 10^{-3}$

### Normalization:

and then normalize, $\gamma_{ki} = p(c_k \mid \mathbf{x}_i) = \dfrac{p(c_k, \mathbf{x}_i)}{\sum_j p(c_j, \mathbf{x}_i)}$

$$\gamma_{11} = P(C_1|x_1) = \frac{P(C_1|x_1)}{P(C_1|x_1) + P(C_2|x_1)} = \frac{1{,}472 \times 10^{-2}}{1{,}472 \times 10^{-1} + 3{,}0987 \times 10^{-2}} = 3{,}221 \times 10^{-1}$$

$$\gamma_{21} = P(C_2|x_1) = \frac{P(C_2|x_1)}{P(C_1|x_1) + P(C_2|x_1)} = 6{,}779 \times 10^{-1}$$

$$\gamma_{12} = P(C_1|x_2) = \dots = 9{,}160 \times 10^{-2}$$

$$\gamma_{22} = P(C_2|x_2) = \dots = 9{,}084 \times 10^{-1}$$

$$\gamma_{13} = P(C_1 \mid x_3) = \dots = 7{,}6955 \times 10^{-1}$$

$$\gamma_{23} \quad P(C_2 \mid x_3) = \dots = 2{,}304 \times 10^{-1}$$

## M step:

$$N_k = \sum_{i=1}^{n} \gamma_{ki}$$

$$\mathbf{u}_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ki} \bullet \mathbf{x}_i$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ki} \bullet (\mathbf{x}_i - \mathbf{u}_k) \bullet (\mathbf{x}_i - \mathbf{u}_k)^T$$

$$\pi_k = p(c_k) = \frac{N_k}{N}$$

$$N_1 = \sum_{i=1}^{3} \gamma_{1i} = \gamma_{11} + \gamma_{12} + \gamma_{13} = 3{,}221 \times 10^{-1} + 9{,}160 \times 10^{-2} + 7{,}6955 \times 10^{-1}$$

$$N_1 = 1{,}183$$

$$N_2 = \sum_{i=1}^{2} \gamma_{2i} = \gamma_{21} + \gamma_{22} + \gamma_{23} = 6{,}779 \times 10^{-1} + 9{,}084 \times 10^{-1} + 2{,}304 \times 10^{-1}$$

$$N_2 = 1{,}8166$$

$$\mu_1 = \frac{1}{N_1} \sum_{i=1}^{3} \gamma_{1i} \cdot x_i = \frac{1}{N_1} \left( \gamma_{11} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \gamma_{12} \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \gamma_{13} \begin{bmatrix} 3 \\ -1 \end{bmatrix} \right)$$

$$\Rightarrow \mu_1 = \begin{bmatrix} 2{,}223 \\ -4{,}955 \times 10^{-1} \end{bmatrix}$$

$$\mu_2 = \frac{1}{N_2} \sum_{i=1}^{3} \gamma_{2i} \cdot x_i = \frac{1}{N_2} \left( \gamma_{21} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \gamma_{22} \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \gamma_{23} \begin{bmatrix} 3 \\ -1 \end{bmatrix} \right)$$

$$\Rightarrow \mu_2 = \begin{bmatrix} 7{,}5368 \times 10^{-1} \\ 8{,}7317 \times 10^{-1} \end{bmatrix}$$

new mean calculated

$$\Sigma_1 = \frac{1}{N_1} \left( \sum_{i=1}^{3} \gamma_{1i} \cdot (x_i - \mu_1) \cdot (x_i - \mu_1)^T \right)$$

$$= \frac{1}{N_1} \Big( \gamma_{11} \cdot (x_1 - \mu_1) \cdot (x_1 - \mu_1)^T + \gamma_{12} \cdot (x_2 - \mu_1) \cdot (x_2 - \mu_1)^T + \gamma_{13} \cdot (x_3 - \mu_1) \cdot (x_3 - \mu_1)^T \Big)$$

$$= \frac{1}{N_1} \Bigg( \gamma_{11} \cdot \left( \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 2{,}233 \\ -4{,}955 \times 10^{-1} \end{bmatrix} \right) \cdot \left( \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 2{,}233 \\ -4{,}955 \times 10^{-1} \end{bmatrix} \right)^T + \gamma_{12} \cdot \left( \begin{bmatrix} 0 \\ 2 \end{bmatrix} - \begin{bmatrix} 2{,}233 \\ -4{,}955 \times 10^{-1} \end{bmatrix} \right) \cdot \left( \begin{bmatrix} 0 \\ 2 \end{bmatrix} - \begin{bmatrix} 2{,}233 \\ -4{,}955 \times 10^{-1} \end{bmatrix} \right)^T$$

$$+ \gamma_{13} \cdot \left( \begin{bmatrix} 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 2{,}233 \\ -4{,}955 \times 10^{-1} \end{bmatrix} \right) \cdot \left( \begin{bmatrix} 3 \\ -1 \end{bmatrix} - \begin{bmatrix} 2{,}233 \\ -4{,}955 \times 10^{-1} \end{bmatrix} \right)^T \Bigg)$$

$$\Rightarrow \Sigma_1 = \begin{bmatrix} 1{,}182 & -8{,}4937 \times 10^{-1} \\ -8{,}4937 \times 10^{-1} & 7{,}1448 \times 10^{-1} \end{bmatrix}$$

$$\Sigma_2 = \frac{1}{N_2} \left( \sum_{i=1}^{3} \gamma_{2i} \cdot (x_i - \mu_2) \cdot (x_i - \mu_2)^T \right)$$

$$\Rightarrow \Sigma_2 = \begin{bmatrix} 9{,}467 \times 10^{-3} & -1{,}0386 \\ -1{,}0386 & 1{,}364 \end{bmatrix}$$

$$\pi_1 = p(C_1) = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} \Rightarrow \pi_1 = 3{,}9944 \times 10^{-1}$$

$$\pi_2 = P(C_2) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} \Rightarrow \pi_2 = 6{,}056 \times 10^{-1}$$

# E-step

(The same with new mean ($\mu$), new cov matrix ($\Sigma$) and new priors ($\pi$))

## Cluster 1:

$P(c_1|x_1) = 4,695 \times 10^{-2}$

$P(C_1|x_2) = 4,618 \times 10^{-4}$

$P(C_1|x_3) = 1,366 \times 10^{-1}$

## Cluster 2:

$P(c_2|x_1) = 9,0161 \times 10^{-2}$

$P(C_2|x_2) = 1,267 \times 10^{-1}$

$P(C_2|x_2) = 6,684 \times 10^{-3}$

## normalization:

and then normalize, $\gamma_{ki} = p(c_k \mid \mathbf{x}_i) = \dfrac{p(c_k, \mathbf{x}_i)}{\sum_j p(c_j, \mathbf{x}_i)}$

$\gamma_{11} = P(c_1|x_1) = 3,429 \times 10^{-1}$

$\gamma_{21} = P(C_2|x_1) = 6,576 \times 10^{-1}$

$\gamma_{12} = P(C_1|x_2) = 3,6325 \times 10^{-3}$

$\gamma_{22} = P(c_2|x_2) = 9,964 \times 10^{-1}$

$\gamma_{13} = P(C_1|x_3) = 9,5335 \times 10^{-1}$

$\gamma_{23} = P(C_2|x_3) = 4,665 \times 10^{-2}$

# M step:

$$N_k = \sum_{i=1}^{n} \gamma_{ki}$$

$$\mathbf{u}_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ki} \bullet \mathbf{x}_i$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ki} \bullet (\mathbf{x}_i - \mathbf{u}_k) \bullet (\mathbf{x}_i - \mathbf{u}_k)^T$$

$$\pi_k = p(c_k) = \frac{N_k}{N}$$

$N_1 = \sum_{i=1}^{3} \gamma_{1i}$

$N_1 = 1,299$

$N_2 = \sum_{i=1}^{2} \gamma_{2i}$

$N_2 = 1,701$

$\mu_1 = \frac{1}{N_1} \sum_{i=1}^{3} \gamma_{1i} \cdot x_i$

$\Rightarrow \mu_1 = \begin{bmatrix} 2,465 \\ -7,281 \times 10^{-1} \end{bmatrix}$

$\mu_2 = \frac{1}{N_2} \sum_{i=1}^{3} \gamma_{2i} \cdot x_i$

$\Rightarrow \mu_2 = \begin{bmatrix} 4,6897 \times 10^{-1} \\ 1,144 \end{bmatrix}$

$\Sigma_1 = \frac{1}{N_1} \left( \sum_{i=1}^{3} \gamma_{1i} \cdot (x_i - \mu_1) \cdot (x_i - \mu_1)^T \right)$

$\Rightarrow \Sigma_1 = \begin{bmatrix} 7,925 \times 10^{-1} & -4,066 \times 10^{-1} \\ -4,066 \times 10^{-1} & 2,157 \times 10^{-1} \end{bmatrix}$

$\Sigma_2 = \frac{1}{N_2} \left( \sum_{i=1}^{3} \gamma_{2i} \cdot (x_i - \mu_2) \cdot (x_i - \mu_2)^T \right)$

$\Rightarrow \Sigma_2 = \begin{bmatrix} 4,136 \times 10^{-1} & -6,189 \times 10^{-1} \\ -6,189 \times 10^{-1} & 1,061 \end{bmatrix}$

$\pi_1 = P(C_1) = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} \Rightarrow \pi_1 = 4,331 \times 10^{-1}$

$\pi_2 = P(C_2) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} \Rightarrow \pi_2 = 5,6687 \times 10^{-1}$

2) Using the final parameters computed in previous question:

a. [1v] perform a hard assignment of observations to clusters under a MAP assumption.

## 2 a)

and then normalize, $\gamma_{ki} = p(c_k \mid \mathbf{x}_i) = \dfrac{p(c_k, \mathbf{x}_i)}{\sum_j p(c_j, \mathbf{x}_i)}$

$$\begin{cases} P(x_1 \mid C=1) = 5{,}860 \times 10^{-1} \\ P(x_1 \mid C=2) = 4{,}1396 \times 10^{-1} \end{cases}$$

$x_1 \longrightarrow$ cluster 1

$$\begin{cases} P(x_2 \mid C=1) = 0 \\ P(x_2 \mid C=2) = 1 \end{cases}$$

$x_2 \longrightarrow$ cluster 2

$x_3 \longrightarrow$ cluster 1

$$\begin{cases} P(x_3 \mid C=1) = 9{,}9999 \times 10^{-1} \simeq 1 \\ P(x_3 \mid C=2) = 9{,}251 \times 10^{-9} \simeq 0 \end{cases}$$

clusters $= \{ C_1 = \{x_1, x_3\},\ C_2 = \{x_2\} \}$

2) Using the final parameters computed in previous question:

a. [1v] perform a hard assignment of observations to clusters under a MAP assumption.

b. [2v] compute the silhouette of the larger cluster (the one that has more observations assigned to it) using the Euclidean distance.

- Silhouette combines both **cohesion** and **separation**
- Calculated for a specific object $\mathbf{x}_i$
  - $a(\mathbf{x}_i)$ = average distance of $\mathbf{x}_i$ to the points in its cluster
  - $b(\mathbf{x}_i)$ = min (average distance of $\mathbf{x}_i$ to points in another cluster)
  - the silhouette coefficient for a point is then given by

$$s(\mathbf{x}_i) = 1 - \frac{a(\mathbf{x}_i)}{b(\mathbf{x}_i)} \text{ if } a < b, \quad (\text{or } s = b/a - 1 \text{ if } a \geq b, \text{ not the usual case})$$

between -1 and 1 (the closer to 1 the better)

- Silhouette of **cluster**: average of observation silhouettes
- Silhouette of **clustering solution**: average of cluster silhouettes

## 2 b)

larger cluster $\Longrightarrow C_1$

**CA:**

$$a(x_1) = \|x_1 - x_3\|_2 = \left\| \begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}3\\-1\end{bmatrix} \right\|_2 = \left\| \begin{bmatrix}-2\\1\end{bmatrix} \right\|_2 = \sqrt{(-2)^2 + 1^2} = \sqrt{5}$$

$$b(x_1) = \|x_1 - x_2\|_2 = \left\| \begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0\\2\end{bmatrix} \right\|_2 = \left\| \begin{bmatrix}1\\-2\end{bmatrix} \right\|_2 = \sqrt{5}$$

$$a(x_3) = \|x_3 - x_1\|_2 = \left\| \begin{bmatrix}3\\-1\end{bmatrix} - \begin{bmatrix}1\\0\end{bmatrix} \right\|_2 = \left\| \begin{bmatrix}2\\-1\end{bmatrix} \right\|_2 = \sqrt{5}$$

$$b(x_3) = \|x_3 - x_2\|_2 = \left\| \begin{bmatrix}3\\-1\end{bmatrix} - \begin{bmatrix}0\\2\end{bmatrix} \right\|_2 = \left\| \begin{bmatrix}3\\-3\end{bmatrix} \right\|_2 = \sqrt{18}$$

$$S(x_1) \overset{a=b}{=} \frac{b(x_1)}{a(x_1)} - 1 = \frac{\sqrt{5}}{\sqrt{5}} - 1 = 0$$

$$S(x_3) \overset{b>a}{=} 1 - \frac{a(x_3)}{b(x_3)} = 1 - \frac{\sqrt{5}}{\sqrt{18}} = 4{,}7295 \times 10^{-1}$$

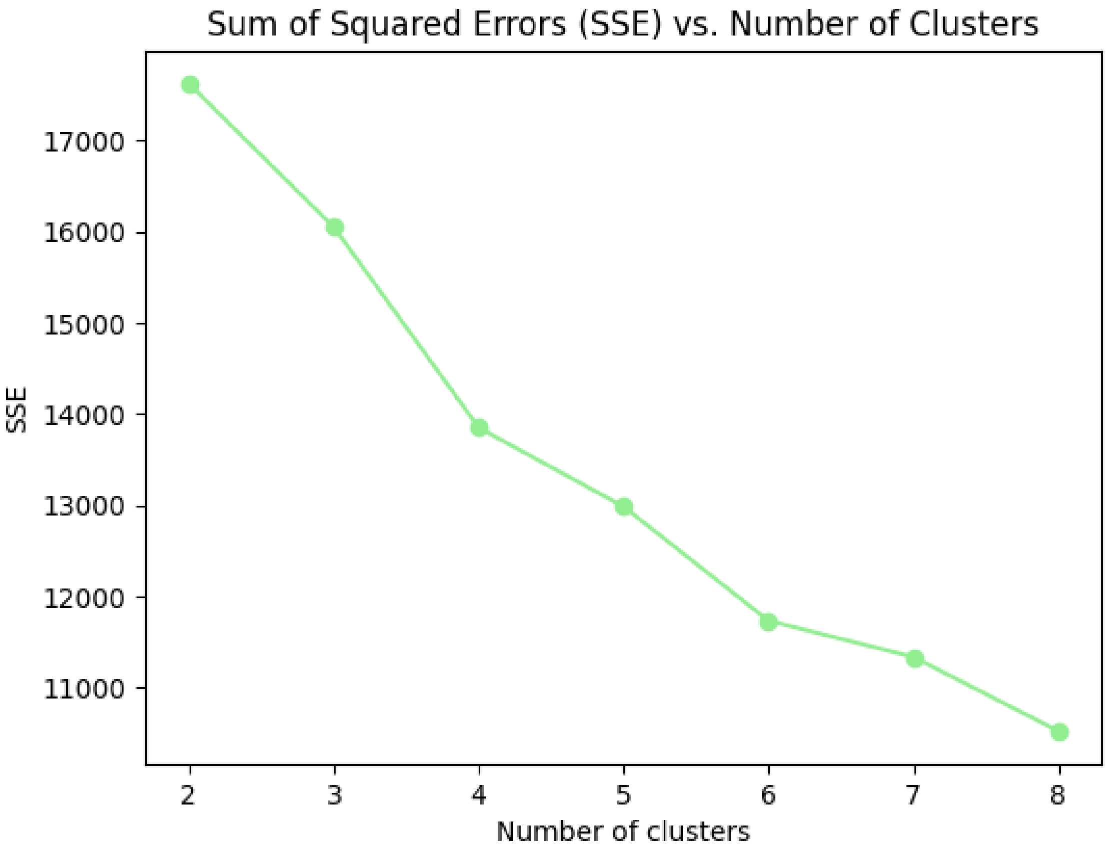
$$S(C_1) = \frac{S(x_1) + S(x_3)}{2} = \frac{0 + 4{,}7255 \times 10^{-1}}{2} = 2{,}3648 \times 10^{-1}$$

$\Leftrightarrow S(C_1) = 2{,}3648 \times 10^{-1}$

silhouette of cluster //

# Parte II

## 1. a)




## b)

The purpose is to identify the ‘elbow’, where the decrease in SSE (Sum of Square Errors) begins to be less significant as the number of clusters increases. This point marks a balance between the quantity of clusters and the quality of the segmentation, since adding more clusters beyond this point generally brings little improvement in terms of reducing inertia.

Based on the graph, the ‘elbow’ appers to be around 4 clusters. Up to this point the SSE (Sum of Square Errors) decreases significantly when we increase the number of clusters, but from 4 clusters onwards, the improvements in SSE reduction become smaller. Therefore 4 clusters seems to be the ideal number of segments, as it balances the number of groups with a good reduction in inertia. If we opted for more than 4 clusters, the gain in terms of SSE reduction would increase without much benefit. The graphic suggests that the optimum segmentation, based on the trade-off between clusters and inertia, would be 4 clusters.

## c)

K-means is a suitable algorithm for numerical data, as it calculates Euclidean distances to form clusters based on the average of the points, this method is ideal for continuous variables.

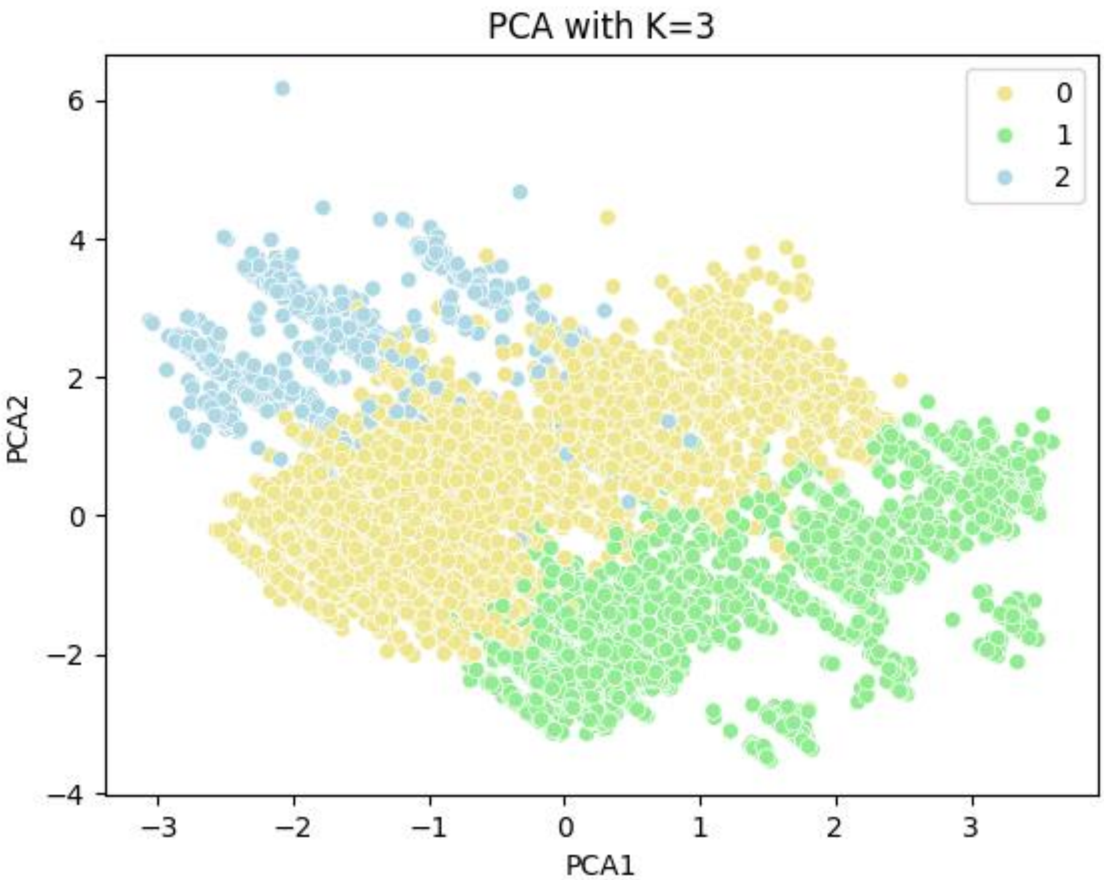
K-modes, on the other hand, is a variation of K-means adapted for categorical data. Instead of using the mean to define the centre of the cluster, it uses the mode, witch makes it more appropriate for qualitative variables.

In the case of the accounts.csv dataset, if many features are categorical (for example, 'job', 'education'), K-modes may be a better choice since K-means does not handle categorical data well without first encoding them.

## 2.

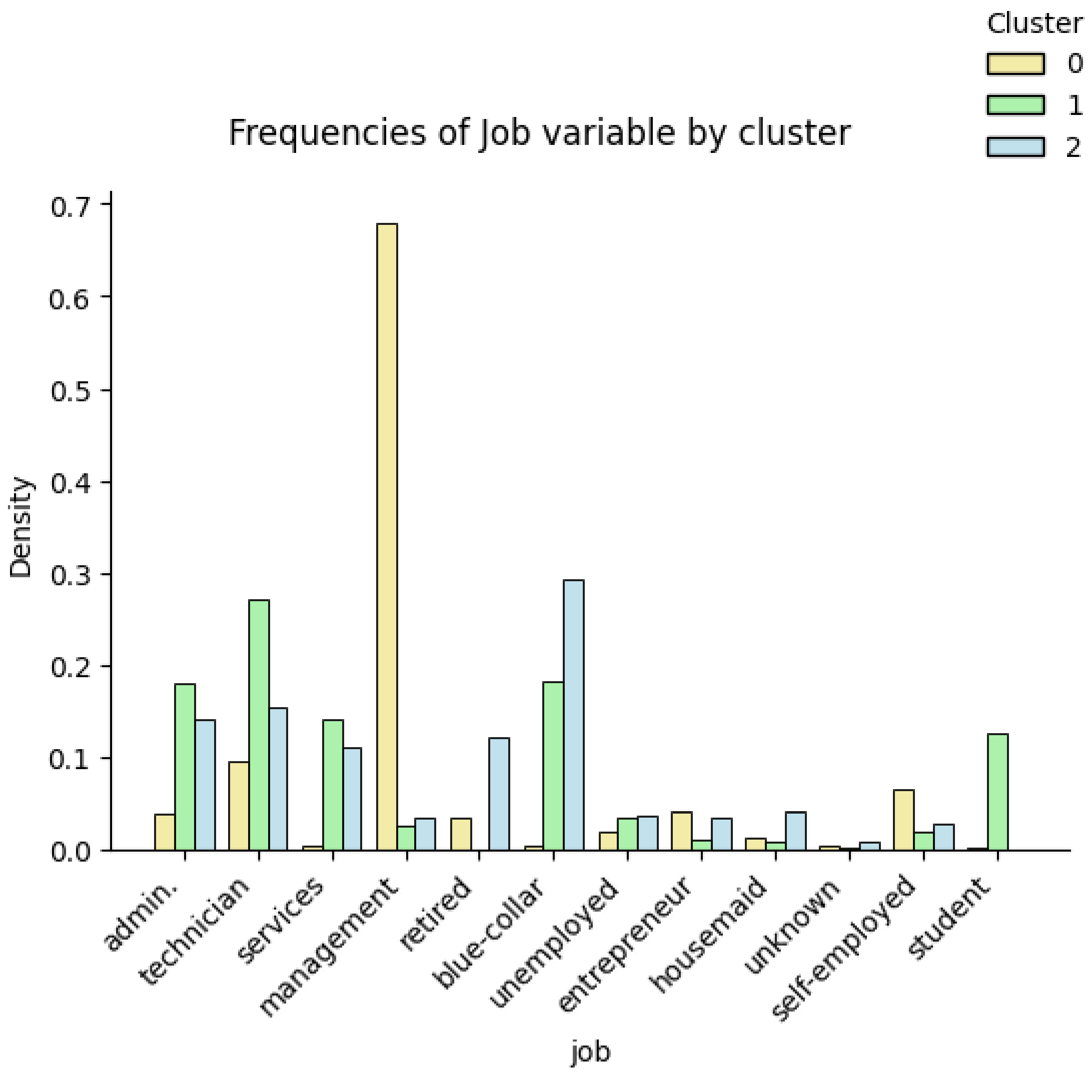
**a)** Top 2 components: 0.22755012102819194

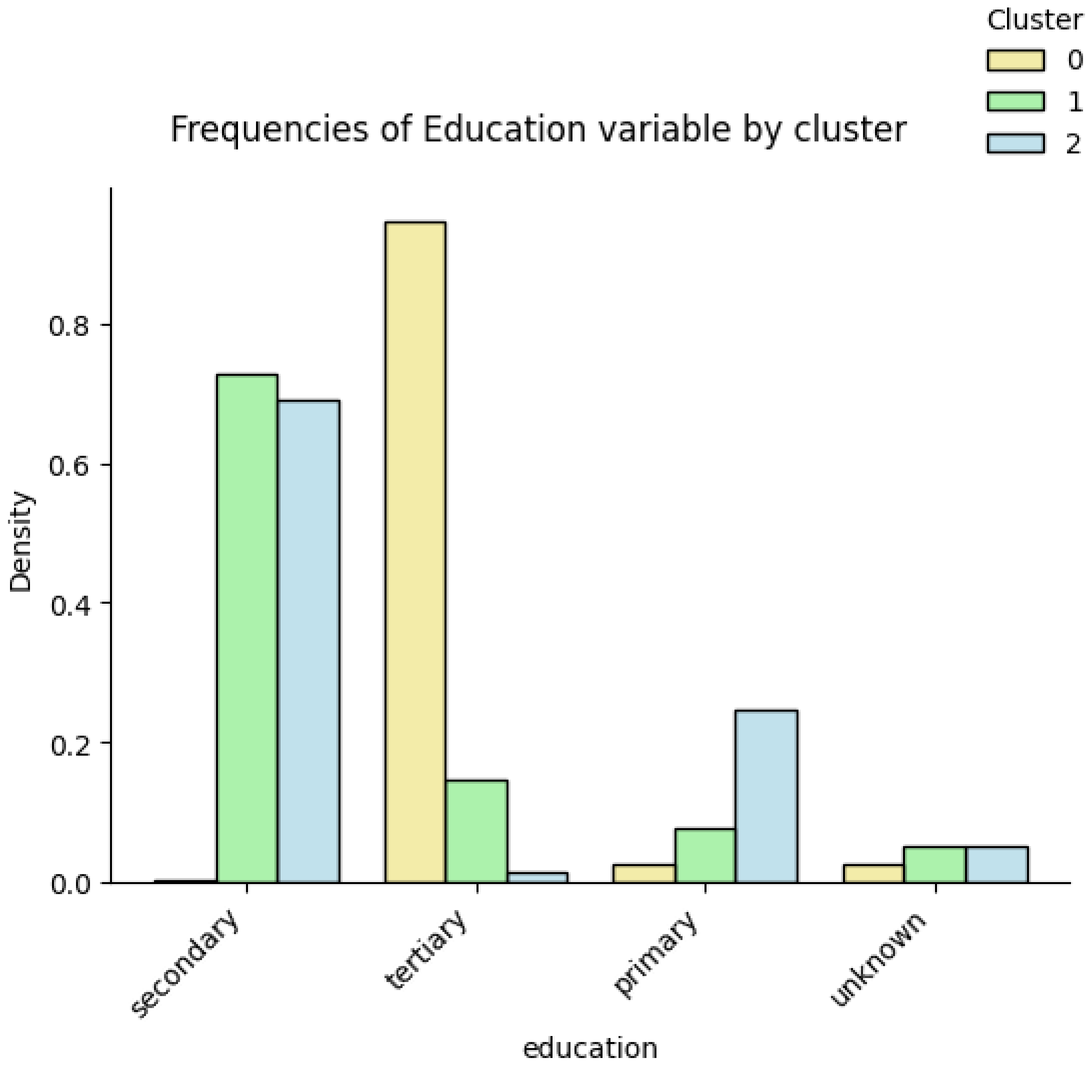
**b)**



The three clusters are visually distinct, though there is some overlap, with a few elements from Cluster 2 appearing within Cluster 0 and vice versa. Overall, however, the clusters remain well-defined and separate.

**c)**






The clusters reveal distinct profiles based on education levels, employment roles, and job types. Cluster 0 predominantly consists of individuals with tertiary education who occupy management positions or are self-employed, including a high proportion of entrepreneurs. This cluster has relatively few people in other job areas and almost no individuals with only primary education, highlighting a concentration of highly educated professionals in higher-paying roles.

Cluster 1 is characterized by a significant number of individuals with secondary education, followed by a smaller yet notable group with tertiary education. This cluster includes a high proportion of people in administrative, technical, service, and student roles, suggesting a mix of mid-level positions aligned with moderate educational attainment.

Cluster 2 primarily comprises individuals with primary education and almost no tertiary-educated individuals. People in this cluster are largely found in roles like blue-collar work, housekeeping, and retired or unemployed status. These roles are generally lower-paying and reflect the limited job opportunities available to those with less education.

These clusters illustrate a trend: people with higher education levels, such as those in Cluster 0, are more likely to hold management or self-employed positions, which tend to offer higher earnings. Conversely, individuals with lower education levels, as seen in Cluster 2, are more likely to be in lower-paying roles. This trend underscores the link between education level and access to higher-paying job opportunities.