Aprendizagem 2024
Homework IV - Group 002
(ist1107095, ist1106078)

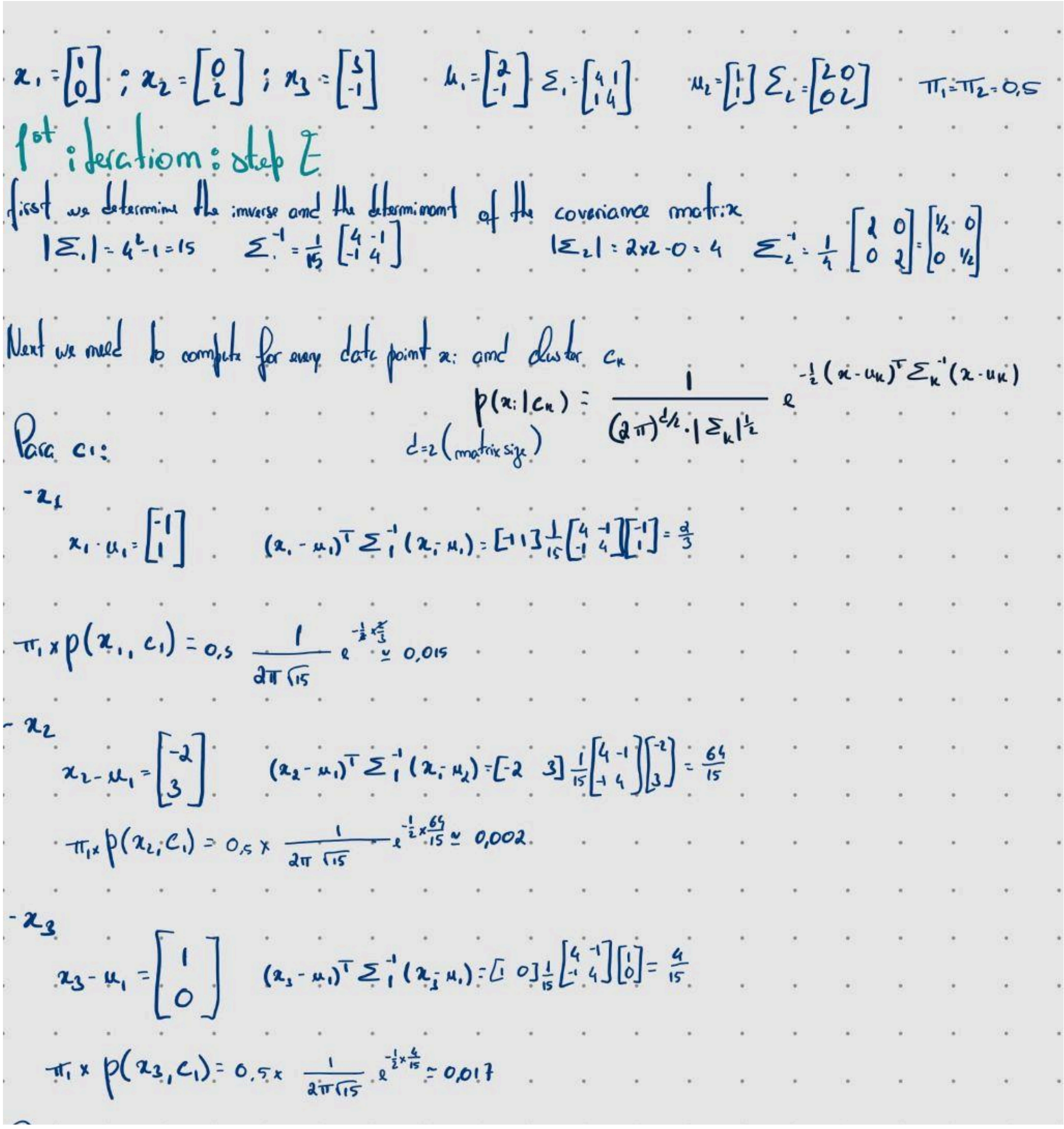
I. Pen-and-paper [9v]

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



$x_1 = \begin{bmatrix}1\\0\end{bmatrix}$; $x_2 = \begin{bmatrix}0\\2\end{bmatrix}$; $x_3 = \begin{bmatrix}3\\-1\end{bmatrix}$ $\quad u_1 = \begin{bmatrix}2\\-1\end{bmatrix}$ $\Sigma_1 = \begin{bmatrix}4 & 1\\1 & 4\end{bmatrix}$ $\quad u_2 = \begin{bmatrix}1\\1\end{bmatrix}$ $\Sigma_2 = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$ $\quad \pi_1 = \pi_2 = 0{,}5$

1st iteration: step E

first we determine the inverse and the determinant of the covariance matrix

$|\Sigma_1| = 4^2 - 1 = 15 \quad \Sigma_1^{-1} = \frac{1}{15}\begin{bmatrix}4 & -1\\-1 & 4\end{bmatrix} \qquad |\Sigma_2| = 2\times2 - 0 = 4 \quad \Sigma_2^{-1} = \frac{1}{4}\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix} = \begin{bmatrix}1/2 & 0\\0 & 1/2\end{bmatrix}$

Next we need to compute for every data point x_i and cluster c_k

$$p(x_i \mid c_k) = \frac{1}{(2\pi)^{d/2} \cdot |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x - u_k)^T \Sigma_k^{-1} (x - u_k)}$$

$d = 2$ (matrix size)

Para c_1:

- x_1

$x_1 - u_1 = \begin{bmatrix}-1\\1\end{bmatrix} \qquad (x_1 - \mu_1)^T \Sigma_1^{-1} (x_1 - \mu_1) = [-1\ 1]\frac{1}{15}\begin{bmatrix}4 & -1\\-1 & 4\end{bmatrix}\begin{bmatrix}-1\\1\end{bmatrix} = \frac{2}{3}$

$\pi_1 \times p(x_1, c_1) = 0{,}5 \ \frac{1}{2\pi\sqrt{15}} e^{-\frac{1}{2}\times\frac{2}{3}} \simeq 0{,}015$

- x_2

$x_2 - \mu_1 = \begin{bmatrix}-2\\3\end{bmatrix} \qquad (x_2 - \mu_1)^T \Sigma_1^{-1} (x_2 - \mu_2) = [-2\ \ 3]\frac{1}{15}\begin{bmatrix}4 & -1\\-1 & 4\end{bmatrix}\begin{bmatrix}-2\\3\end{bmatrix} = \frac{64}{15}$

$\pi_1 \times p(x_2, c_1) = 0{,}5 \times \frac{1}{2\pi\sqrt{15}} e^{-\frac{1}{2}\times\frac{64}{15}} \simeq 0{,}002$

- x_3

$x_3 - u_1 = \begin{bmatrix}1\\0\end{bmatrix} \qquad (x_3 - \mu_1)^T \Sigma_1^{-1} (x_3 - \mu_1) = [1\ 0]\frac{1}{15}\begin{bmatrix}4 & -1\\-1 & 4\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} = \frac{4}{15}$

$\pi_1 \times p(x_3, c_1) = 0{,}5 \times \frac{1}{2\pi\sqrt{15}} e^{-\frac{1}{2}\times\frac{4}{15}} \simeq 0{,}017$

Para c2:

x_1

$$x_1 - u_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \quad (x_1 - u_2)^T \Sigma_2^{-1} (x_1 - u_2) = [0 \; -1] \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = 1/2$$

$$\pi_2 \times p(x_1, c_2) = 0{,}5 \, \frac{1}{2\pi \times \sqrt{4}} e^{-\frac{1}{2} \times \frac{1}{2}} = 0{,}5 \, \frac{e^{-\frac{1}{4}}}{4\pi} \simeq 0{,}031$$

x_2

$$x_2 - u_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad (x_2 - u_2)^T \Sigma_2^{-1} (x_2 - u_2) = [-1 \;\; 1] \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = 1$$

$$\pi_2 \times p(x_2, c_2) = 0{,}5 \, \frac{1}{4\pi} e^{-\frac{1}{2}} \simeq 0{,}024$$

x_3

$$x_3 - u_2 = \begin{bmatrix} 2 \\ -2 \end{bmatrix} \quad (x_3 - u_2)^T \Sigma_2^{-1} (x_3 - u_2) = [2 \;\; -2] \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 2 \\ -2 \end{bmatrix} = 4$$

$$\pi_2 \times p(x_3, c_2) = 0{,}5 \, \frac{1}{4\pi} e^{-\frac{1}{2} \times 4} \simeq 0{,}005$$

Normalize the posteriors

$$\gamma_{11} = P(c_1 | x_1) = \frac{0{,}015}{0{,}031 + 0{,}015} \simeq 0{,}326 \qquad P(c_2|x_1) = 1 - P(c_1|x_1) = 0{,}674$$

$$p(c_1 | x_2) = \frac{0{,}002}{0{,}002 + 0{,}024} \simeq 0{,}077 \qquad P(c_2|x_1) = 0{,}923$$

$$p(c_1 | x_3) = \frac{0{,}017}{0{,}005 + 0{,}017} \simeq 0{,}773 \qquad P(c_2|x_1) = 0{,}227$$

1st iteration : step M

$$u_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_i \;\Big|\; \Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ki} \cdot (x_i - \mu_k)(x_i - \mu_k)^T \;\Big|\; \pi_k = p(c_k) = \frac{N_k}{N}$$

$$N_1 = 0{,}326 + 0{,}077 + 0{,}773 = 1{,}176 \quad ; \; N_2 = 0{,}674 + 0{,}923 + 0{,}227 = 1{,}824$$

$$u_1' = \frac{0{,}326 \times \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 0{,}077 \begin{bmatrix} 0 \\ 2 \end{bmatrix} + 0{,}773 \begin{bmatrix} 3 \\ -1 \end{bmatrix}}{1{,}176} = \frac{1}{1{,}176} \begin{bmatrix} 2{,}645 \\ -0{,}619 \end{bmatrix} = \begin{bmatrix} 2{,}249 \\ -0{,}526 \end{bmatrix}$$

$$u_2' = \frac{0{,}674 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 0{,}923 \begin{bmatrix} 0 \\ 2 \end{bmatrix} + 0{,}227 \begin{bmatrix} 3 \\ -1 \end{bmatrix}}{1{,}824} = \frac{1}{1{,}824} \begin{bmatrix} 1{,}355 \\ 1{,}619 \end{bmatrix} = \begin{bmatrix} 0{,}743 \\ 0{,}888 \end{bmatrix}$$

$$\Sigma_1' = \frac{1}{1{,}176} \left(0{,}326 \begin{bmatrix} -1 \\ 1 \end{bmatrix} [-1 \;\; 1] + 0{,}077 \begin{bmatrix} -2 \\ 3 \end{bmatrix} [-2 \;\; 3] + 0{,}773 \begin{bmatrix} 1 \\ 0 \end{bmatrix} [1 \; 0] \right) =$$

$$= \frac{1}{1{,}176} \left(0{,}326 \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + 0{,}077 \begin{bmatrix} 4 & -6 \\ -6 & 9 \end{bmatrix} + 0{,}773 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right) = \frac{1}{1{,}176} \begin{bmatrix} 1{,}407 & -0{,}788 \\ -0{,}788 & 1{,}019 \end{bmatrix} \simeq \begin{bmatrix} 1{,}196 & -0{,}670 \\ -0{,}670 & 0{,}866 \end{bmatrix}$$

$$\Sigma_2' = \frac{1}{1{,}824} \left(0{,}674 \begin{bmatrix} 0 \\ -1 \end{bmatrix} [0 \; -1] + 0{,}923 \begin{bmatrix} -1 \\ 1 \end{bmatrix} [-1 \;\; 1] + 0{,}227 \begin{bmatrix} 2 \\ -2 \end{bmatrix} [2 \;\; -2] \right) =$$

$$= \frac{1}{1{,}824} \left(0{,}674 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + 0{,}923 \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + 0{,}227 \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix} \right) = \frac{1}{1{,}824} \begin{bmatrix} 1{,}831 & -1{,}831 \\ -1{,}831 & 2{,}505 \end{bmatrix} \simeq \begin{bmatrix} 1{,}004 & -1{,}004 \\ -1{,}004 & 1{,}373 \end{bmatrix}$$

$$\pi'_1 = \frac{1,176}{1,176 + 1,824} = 0,392 \qquad \pi'_2 = 1 - \pi'_1 = 0,608$$

2nd iteration: step E

Again we will begin with the determinants and inverses of the covariance matrix

$$|\Sigma_1'| = 1,196 \times 0,866 - (-0,670)^2 \simeq 0,587 \qquad \Sigma_1'^{-1} = \frac{1}{0,587}\begin{bmatrix} 0,866 & 0,670 \\ 0,670 & 1,196 \end{bmatrix} \simeq \begin{bmatrix} 1,475 & 1,141 \\ 1,141 & 2,037 \end{bmatrix}$$

$$|\Sigma_2'| = 1,004 \times 1,373 - (-1,004)^2 \simeq 0,370 \qquad \Sigma_2'^{-1} = \frac{1}{0,370}\begin{bmatrix} 1,373 & 1,004 \\ 1,004 & 1,004 \end{bmatrix} \simeq \begin{bmatrix} 3,711 & 2,714 \\ 2,714 & 2,714 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix};\ x_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix};\ x_3 = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

$$\mu'_1 = \begin{bmatrix} 2,249 \\ -0,526 \end{bmatrix} \quad \mu'_2 = \begin{bmatrix} 0,743 \\ 0,888 \end{bmatrix} \qquad \sqrt{|\Sigma_1'|} \simeq 0,766 \qquad \sqrt{|\Sigma_2'|} \simeq 0,608$$

for cluster 1:

x_1:

$$x_1 - \mu'_1 = \begin{bmatrix} -1,249 \\ 0,526 \end{bmatrix} \qquad (x_1 - \mu'_1)^T \Sigma_1'^{-1}(x_1 - \mu'_1) = \begin{bmatrix} -1,249 & 0,526 \end{bmatrix}\begin{bmatrix} 1,475 & 1,141 \\ 1,141 & 2,037 \end{bmatrix}\begin{bmatrix} -1,249 \\ 0,526 \end{bmatrix} = 1,365$$

$$\pi'_1 \times p(c'_1 | x_1) = 0,392 \times \frac{1}{2\pi \times 0,776} e^{-\frac{1}{2} \times 1,365} \simeq 0,041$$

x_2:

$$x_2 - \mu'_1 = \begin{bmatrix} -2,249 \\ 2,526 \end{bmatrix} \qquad (x_2 - \mu'_1)^T \Sigma_1'^{-1}(x_2 - \mu'_1) = \begin{bmatrix} -2,249 & 2,526 \end{bmatrix}\begin{bmatrix} 1,475 & 1,141 \\ 1,141 & 2,037 \end{bmatrix}\begin{bmatrix} -2,249 \\ 2,526 \end{bmatrix} = 7,494$$

$$\pi'_1 \times p(c'_1 | x_2) = 0,392 \times \frac{1}{2\pi \times 0,776} e^{-\frac{1}{2} \times 7,494} \simeq 0,002$$

x_3

$$x_3 - \mu'_1 = \begin{bmatrix} 0,751 \\ 0,474 \end{bmatrix} \qquad (x - \mu'_1)^T \Sigma_1'^{-1}(x - \mu'_1) = \begin{bmatrix} 0,751 & 0,474 \end{bmatrix}\begin{bmatrix} 1,475 & 1,141 \\ 1,141 & 2,037 \end{bmatrix}\begin{bmatrix} 0,751 \\ 0,474 \end{bmatrix} = 2,102$$

$$\pi'_1 \times p(c'_1 | x_3) = 0,392 \times \frac{1}{2\pi \times 0,766} e^{-\frac{1}{2} \times 2,102} \simeq 0,028$$

for cluster 2:

x_1:

$$x_1 - \mu'_2 = \begin{bmatrix} 0,257 \\ -0,888 \end{bmatrix} \qquad (x_1 - \mu'_2)^T \Sigma_2'^{-1}(x_1 - \mu'_2) = \begin{bmatrix} 0,247 & -0,888 \end{bmatrix}\begin{bmatrix} 3,711 & 2,714 \\ 2,714 & 2,714 \end{bmatrix}\begin{bmatrix} 0,247 \\ -0,888 \end{bmatrix} \simeq 1,176$$

$$\pi'_2 \times p(c'_2 | x_1) = 0,608 \times \frac{1}{2\pi \times 0,608} \times e^{-\frac{1}{2} \times 1,176} \simeq 0,088$$

x_2:

$$x_2 - \mu'_2 = \begin{bmatrix} -0,743 \\ 1,112 \end{bmatrix} \qquad (x_2 - \mu'_2)^T \Sigma_2'^{-1}(x_2 - \mu'_2) = \begin{bmatrix} -0,743 & 1,112 \end{bmatrix}\begin{bmatrix} 3,711 & 2,714 \\ 2,714 & 2,714 \end{bmatrix}\begin{bmatrix} -0,743 \\ 1,112 \end{bmatrix} \simeq 0,920$$

$$\pi'_2 \times p(c'_2 | x_2) = 0,608 \times \frac{1}{2\pi \times 0,608} \times e^{-\frac{1}{2} \times 0,920} \simeq 0,100$$

x_3:

$$x_3 - \mu'_2 = \begin{bmatrix} 2,257 \\ -1,888 \end{bmatrix} \qquad (x_2 - \mu'_2)^T \Sigma_2'^{-1}(x_2 - \mu'_2) = \begin{bmatrix} 2,257 & -1,888 \end{bmatrix}\begin{bmatrix} 3,711 & 2,714 \\ 2,714 & 2,714 \end{bmatrix}\begin{bmatrix} 2,257 \\ -1,888 \end{bmatrix} \simeq 5,448$$

$$\pi'_2 \times p(c'_2 | x_3) = 0,608 \times \frac{1}{2\pi \times 0,608} \times e^{-\frac{1}{2} \times 5,448} \simeq 0,10$$

Normalize the posteriors

$$\gamma'_{11} = p(c'_1 | x_1) = \frac{0,041}{0,041 + 0,088} = 0,318 \qquad \gamma'_{21}\ p(c'_2 | x_1) = 0,682$$

$$p(c'_1 \mid x_2) = \frac{0{,}002}{0{,}002 + 0{,}100} = 0{,}020 \qquad p(c'_2 \mid x_2) = 0{,}980$$

$$p(c'_1 \mid x_3) = \frac{0{,}028}{0{,}028 + 0{,}010} = 0{,}737 \qquad p(c'_2 \mid x_2) = 0{,}263$$

2nd iteration: Step M

$$N_1 = 0{,}318 + 0{,}02 + 0{,}737 = 1{,}075 \qquad N_2 = 0{,}682 + 0{,}980 + 0{,}263 = 1{,}925$$

$$\mu''_1 = \frac{0{,}318 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 0{,}02 \begin{bmatrix} 0 \\ 2 \end{bmatrix} + 0{,}737 \begin{bmatrix} 3 \\ -1 \end{bmatrix}}{1{,}075} = \frac{1}{1{,}075} \begin{bmatrix} 2{,}529 \\ -0{,}697 \end{bmatrix} = \begin{bmatrix} 2{,}353 \\ -0{,}648 \end{bmatrix}$$

$$\mu''_2 = \frac{0{,}682 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 0{,}98 \begin{bmatrix} 0 \\ 2 \end{bmatrix} + 0{,}263 \begin{bmatrix} 3 \\ -1 \end{bmatrix}}{1{,}925} = \frac{1}{1{,}925} \begin{bmatrix} 1{,}471 \\ 1{,}697 \end{bmatrix} = \begin{bmatrix} 0{,}764 \\ 0{,}882 \end{bmatrix}$$

$$\Sigma''_1 = \frac{1}{1{,}075} \left(0{,}318 \times \begin{bmatrix} -1{,}249 \\ 0{,}526 \end{bmatrix} \begin{bmatrix} 1{,}249 & 0{,}526 \end{bmatrix} + 0{,}02 \times \begin{bmatrix} -2{,}249 \\ 2{,}526 \end{bmatrix} \begin{bmatrix} -2{,}249 & 2{,}526 \end{bmatrix} + 0{,}737 \begin{bmatrix} -0{,}743 \\ 1{,}112 \end{bmatrix} \begin{bmatrix} -0{,}743 & 1{,}112 \end{bmatrix} \right) \simeq$$

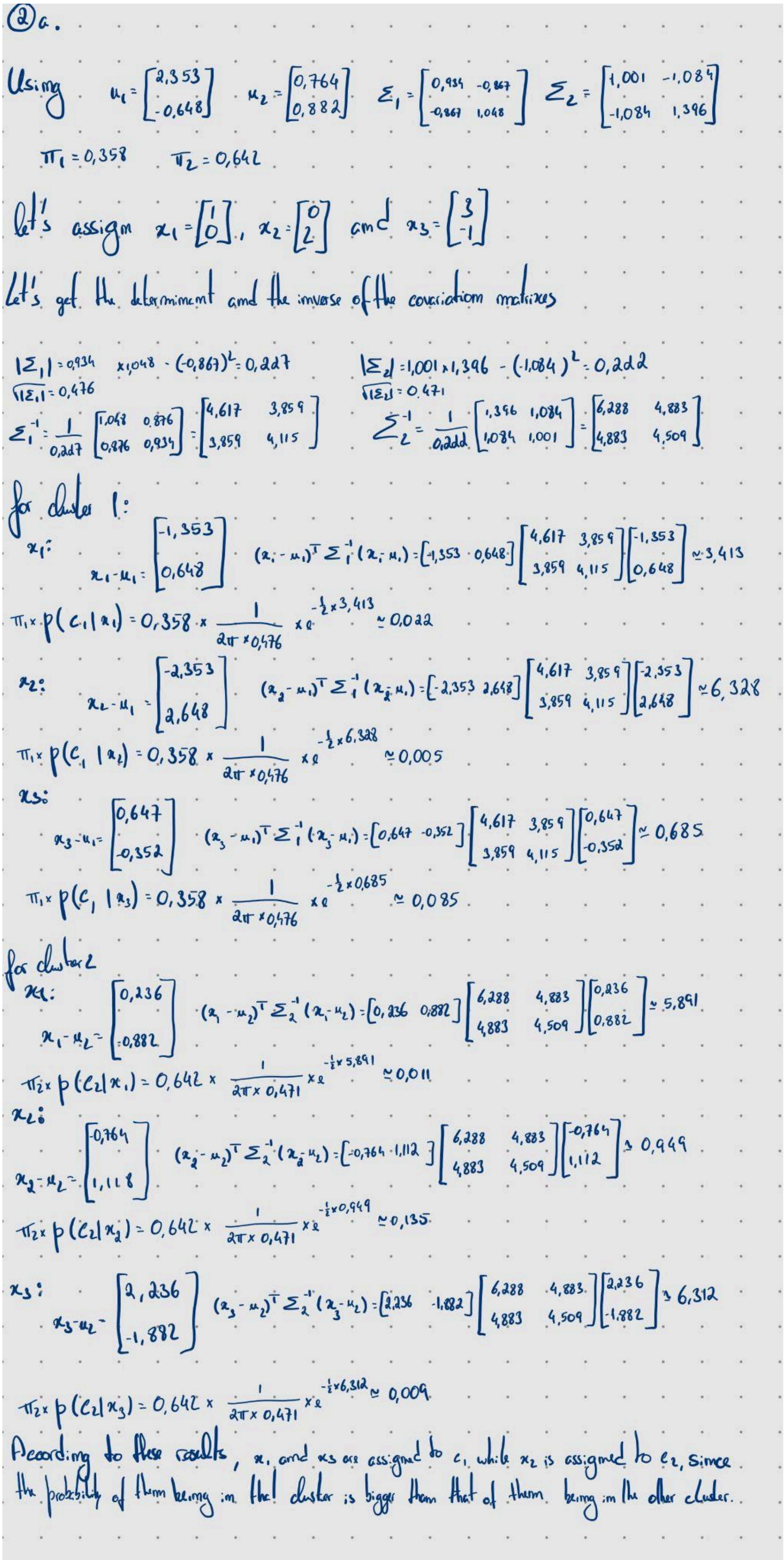
$$\simeq \frac{1}{1{,}075} \left(\begin{bmatrix} 0{,}496 & -0{,}209 \\ -0{,}209 & 0{,}088 \end{bmatrix} + \begin{bmatrix} 0{,}101 & -0{,}114 \\ -0{,}114 & 0{,}128 \end{bmatrix} + \begin{bmatrix} 0{,}407 & -0{,}609 \\ -0{,}609 & 0{,}911 \end{bmatrix} \right) = \begin{bmatrix} 0{,}934 & -0{,}867 \\ -0{,}867 & 1{,}048 \end{bmatrix}$$

$$\Sigma''_2 = \frac{1}{1{,}925} \left(0{,}682 \begin{bmatrix} 0{,}257 \\ -0{,}888 \end{bmatrix} \begin{bmatrix} 0{,}257 & -0{,}888 \end{bmatrix} + 0{,}98 \begin{bmatrix} -0{,}743 \\ 1{,}112 \end{bmatrix} \begin{bmatrix} -0{,}743 & 1{,}112 \end{bmatrix} + 0{,}263 \begin{bmatrix} 2{,}257 \\ -1{,}888 \end{bmatrix} \begin{bmatrix} 2{,}257 & -1{,}888 \end{bmatrix} \right) \simeq$$

$$\simeq \frac{1}{1{,}925} \left(\begin{bmatrix} 0{,}045 & -0{,}156 \\ -0{,}156 & 0{,}538 \end{bmatrix} + \begin{bmatrix} 0{,}541 & -0{,}810 \\ -0{,}810 & 1{,}212 \end{bmatrix} + \begin{bmatrix} 1{,}340 & -1{,}121 \\ -1{,}121 & 0{,}937 \end{bmatrix} \right) \simeq \begin{bmatrix} 1{,}001 & -1{,}084 \\ -1{,}084 & 1{,}396 \end{bmatrix}$$

$$\pi''_1 = \frac{1{,}075}{3} = 0{,}358 \qquad \pi''_2 = 0{,}642$$

2. Using the final parameters computed in previous question:
   a. [1v] Perform a hard assignment of observations to clusters under a MAP assumption.



②a.

Using $u_1 = \begin{bmatrix} 2{,}353 \\ -0{,}648 \end{bmatrix}$ $u_2 = \begin{bmatrix} 0{,}764 \\ 0{,}882 \end{bmatrix}$ $\Sigma_1 = \begin{bmatrix} 0{,}934 & -0{,}867 \\ -0{,}867 & 1{,}048 \end{bmatrix}$ $\Sigma_2 = \begin{bmatrix} 1{,}001 & -1{,}084 \\ -1{,}084 & 1{,}396 \end{bmatrix}$

$\pi_1 = 0{,}358$ $\pi_2 = 0{,}642$

let's assign $x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $x_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$ and $x_3 = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$

Let's get the determinant and the inverse of the covariation matrixes

$|\Sigma_1| = 0{,}934 \times 1{,}048 - (-0{,}867)^2 = 0{,}227$ $|\Sigma_2| = 1{,}001 \times 1{,}396 - (-1{,}084)^2 = 0{,}222$

$\sqrt{|\Sigma_1|} = 0{,}476$ $\sqrt{|\Sigma_2|} = 0.471$

$\Sigma_1^{-1} = \frac{1}{0{,}227} \begin{bmatrix} 1{,}048 & 0{,}876 \\ 0{,}876 & 0{,}934 \end{bmatrix} = \begin{bmatrix} 4{,}617 & 3{,}859 \\ 3{,}859 & 4{,}115 \end{bmatrix}$ $\Sigma_2^{-1} = \frac{1}{0{,}222} \begin{bmatrix} 1{,}396 & 1{,}084 \\ 1{,}084 & 1{,}001 \end{bmatrix} = \begin{bmatrix} 6{,}288 & 4{,}883 \\ 4{,}883 & 4{,}509 \end{bmatrix}$

for cluster 1:

x_1: $x_1 - u_1 = \begin{bmatrix} -1{,}353 \\ 0{,}648 \end{bmatrix}$ $(x_i - u_1)^T \Sigma_1^{-1} (x_i - u_1) = [-1{,}353 \;\; 0{,}648] \begin{bmatrix} 4{,}617 & 3{,}859 \\ 3{,}859 & 4{,}115 \end{bmatrix} \begin{bmatrix} -1{,}353 \\ 0{,}648 \end{bmatrix} \simeq 3{,}413$

$\pi_1 \times p(c_1 | x_1) = 0{,}358 \times \frac{1}{2\pi \times 0{,}476} \times e^{-\frac{1}{2} \times 3{,}413} \simeq 0{,}022$

x_2: $x_2 - u_1 = \begin{bmatrix} -2{,}353 \\ 2{,}648 \end{bmatrix}$ $(x_2 - u_1)^T \Sigma_1^{-1} (x_2 - u_1) = [-2{,}353 \;\; 2{,}648] \begin{bmatrix} 4{,}617 & 3{,}859 \\ 3{,}859 & 4{,}115 \end{bmatrix} \begin{bmatrix} -2{,}353 \\ 2{,}648 \end{bmatrix} \simeq 6{,}328$

$\pi_1 \times p(c_1 | x_2) = 0{,}358 \times \frac{1}{2\pi \times 0{,}476} \times e^{-\frac{1}{2} \times 6{,}328} \simeq 0{,}005$

x_3: $x_3 - u_1 = \begin{bmatrix} 0{,}647 \\ -0{,}352 \end{bmatrix}$ $(x_3 - u_1)^T \Sigma_1^{-1} (x_3 - u_1) = [0{,}647 \;\; -0{,}352] \begin{bmatrix} 4{,}617 & 3{,}859 \\ 3{,}859 & 4{,}115 \end{bmatrix} \begin{bmatrix} 0{,}647 \\ -0{,}352 \end{bmatrix} \simeq 0{,}685$

$\pi_1 \times p(c_1 | x_3) = 0{,}358 \times \frac{1}{2\pi \times 0{,}476} \times e^{-\frac{1}{2} \times 0{,}685} \simeq 0{,}085$

for cluster 2

x_1: $x_1 - u_2 = \begin{bmatrix} 0{,}236 \\ -0{,}882 \end{bmatrix}$ $(x_1 - u_2)^T \Sigma_2^{-1} (x_1 - u_2) = [0{,}236 \;\; 0{,}882] \begin{bmatrix} 6{,}288 & 4{,}883 \\ 4{,}883 & 4{,}509 \end{bmatrix} \begin{bmatrix} 0{,}236 \\ 0{,}882 \end{bmatrix} \simeq 5{,}891$

$\pi_2 \times p(c_2 | x_1) = 0{,}642 \times \frac{1}{2\pi \times 0{,}471} \times e^{-\frac{1}{2} \times 5{,}891} \simeq 0{,}011$

x_2: $x_2 - u_2 = \begin{bmatrix} -0{,}764 \\ 1{,}118 \end{bmatrix}$ $(x_2 - u_2)^T \Sigma_2^{-1} (x_2 - u_2) = [-0{,}764 \;\; 1{,}112] \begin{bmatrix} 6{,}288 & 4{,}883 \\ 4{,}883 & 4{,}509 \end{bmatrix} \begin{bmatrix} -0{,}764 \\ 1{,}112 \end{bmatrix} \simeq 0{,}949$

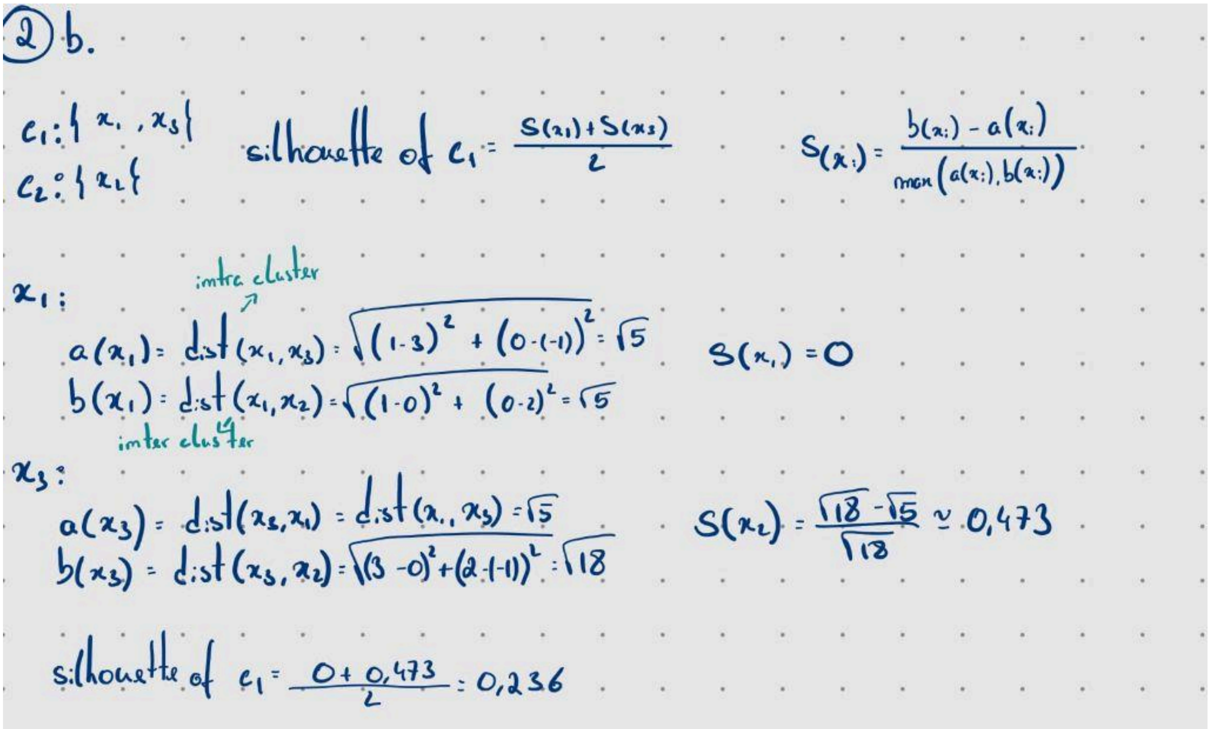
$\pi_2 \times p(c_2 | x_2) = 0{,}642 \times \frac{1}{2\pi \times 0{,}471} \times e^{-\frac{1}{2} \times 0{,}949} \simeq 0{,}135$

x_3: $x_3 - u_2 = \begin{bmatrix} 2{,}236 \\ -1{,}882 \end{bmatrix}$ $(x_3 - u_2)^T \Sigma_2^{-1} (x_3 - u_2) = [2{,}236 \;\; -1{,}882] \begin{bmatrix} 6{,}288 & 4{,}883 \\ 4{,}883 & 4{,}509 \end{bmatrix} \begin{bmatrix} 2{,}236 \\ -1{,}882 \end{bmatrix} \simeq 6{,}312$

$\pi_2 \times p(c_2 | x_3) = 0{,}642 \times \frac{1}{2\pi \times 0{,}471} \times e^{-\frac{1}{2} \times 6{,}312} \simeq 0{,}009$

According to these results, x_1 and x_3 are assigned to c_1 while x_2 is assigned to c_2, since the probability of them being in that cluster is bigger than that of them being in the other cluster.

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



② b.

$c_1: \{x_1, x_3\}$ silhouette of $c_1 = \frac{S(x_1)+S(x_3)}{2}$ $S(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$

$c_2: \{x_2\}$

x_1:

intra cluster

$a(x_1) = dist(x_1, x_3) = \sqrt{(1-3)^2 + (0-(-1))^2} = \sqrt{5}$ $S(x_1) = 0$

$b(x_1) = dist(x_1, x_2) = \sqrt{(1-0)^2 + (0-2)^2} = \sqrt{5}$

inter cluster

x_3:

$a(x_3) = dist(x_3, x_1) = dist(x_1, x_3) = \sqrt{5}$ $S(x_2) = \frac{\sqrt{18}-\sqrt{5}}{\sqrt{18}} \simeq 0{,}473$

$b(x_3) = dist(x_3, x_2) = \sqrt{(3-0)^2+(2-(-1))^2} = \sqrt{18}$

silhouette of $c_1 = \frac{0 + 0{,}473}{2} = 0{,}236$

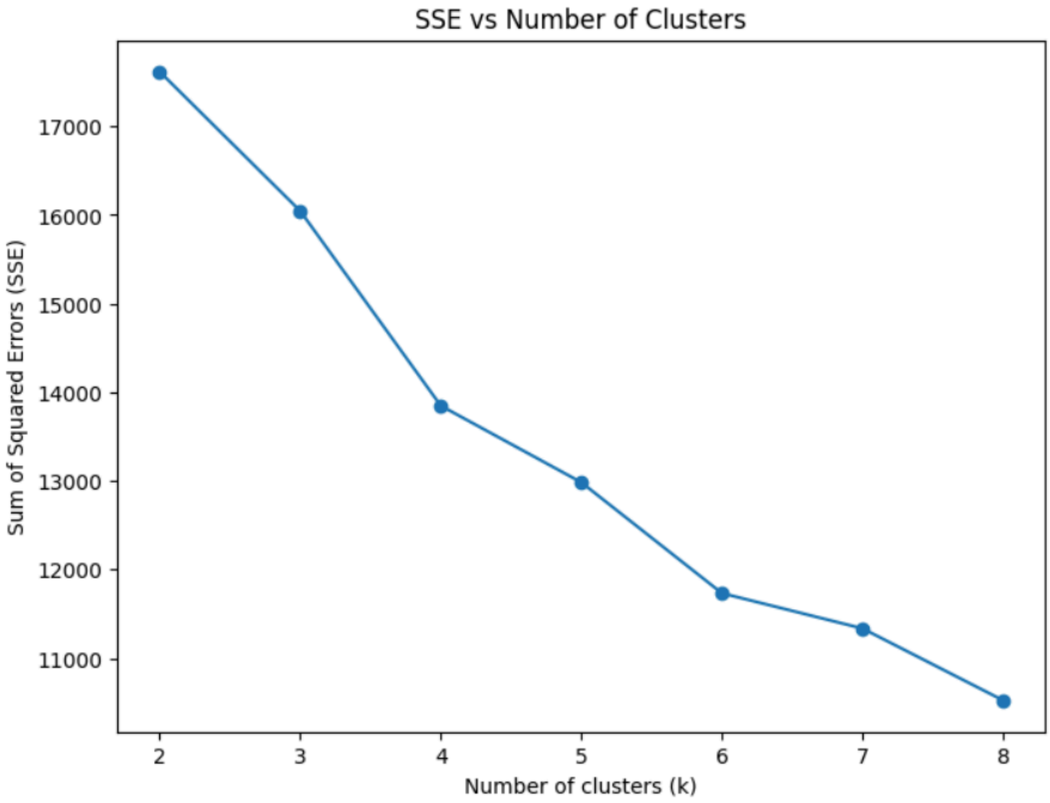
II. Programming and critical analysis [11v]

In the next exercise you will use the `accounts.csv` dataset. This dataset contains account details of bank clients, and the target variable *y* is binary ('has the client subscribed a term deposit?'). Select the first 8 features and remove duplicates and null values.

> ***Hint:*** You can use `get_dummies()` to change the feature type (e.g. `pd.get_dummies(data, drop_first=True)`).

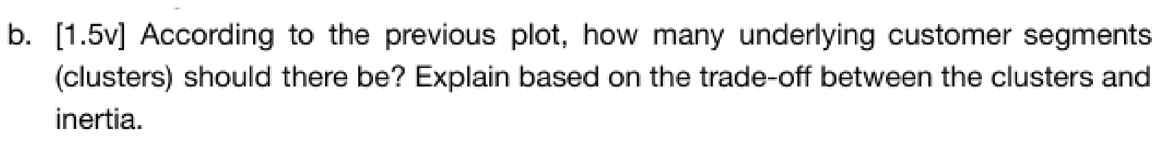
1. Normalize the data using MinMaxScaler:
   a. [4v] Using *sklearn*, apply *k*-means clustering (without targets) on the normalized data with `k={2,3,4,5,6,7,8}`, `max_iter=500` and `random_state=42`. Plot the different sum of squared errors (SSE) using the `_inertia` attribute of *k*-means according to the number of clusters.

```
import pandas as pd

from sklearn.preprocessing import MinMaxScaler

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt


file_path = 'accounts.csv'  # Adjust the path to your file

data = pd.read_csv(file_path)


data_selected = data.iloc[:, :8]


data_cleaned = data_selected.drop_duplicates().dropna()


data_encoded = pd.get_dummies(data_cleaned, drop_first=True)


scaler = MinMaxScaler()

data_normalized = scaler.fit_transform(data_encoded)


sse = []

k_value = range(2, 9)  # k values from 2 to 8


for k in k_value:

    kmeans = KMeans(n_clusters=k, max_iter=500, random_state=42)
    kmeans.fit(data_normalized)
    sse.append(kmeans.inertia_)  # Store the SSE for each k


plt.figure(figsize=(8, 6))
plt.plot(k_value, sse, marker='o')
plt.xlabel('Number of clusters (k)')
plt.ylabel('Sum of Squared Errors (SSE)')
plt.title('SSE vs Number of Clusters')
plt.show()
```



Using the sklearn library, I applied the k-means clustering algorithm to the normalized data with values of k k ranging from {2, 3, 4, 5, 6, 7, 8}, setting max_iter=500 and random_state=42 to ensure reproducibility. The plot above shows the sum of squared errors (SSE) for each value of k k.

This plot was generated using the "Elbow Method." The elbow method is a heuristic approach used to identify the optimal number of clusters. By plotting the SSE (which represents the sum of the squared distances between each data point and the centroid of its cluster) against the number of clusters, we can observe where the SSE starts to decrease at a slower rate. This point, where the slope of the curve begins to level off, is known as the "elbow." It represents the trade-off between the number of clusters and the model's accuracy, helping to identify an appropriate number of clusters to use.

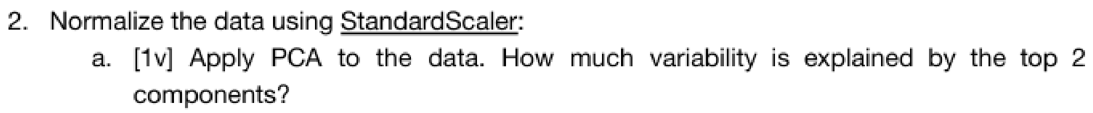


b. [1.5v] According to the previous plot, how many underlying customer segments (clusters) should there be? Explain based on the trade-off between the clusters and inertia.

According to the plot above, the ideal number of clusters appears to be around 4, where the "elbow" is observed. At this point, adding more clusters does not significantly reduce the SSE, indicating that the model does not benefit substantially from additional clusters. Thus, choosing 4 clusters strikes a good balance between model complexity and the reduction of error. This suggests that there are likely 4 distinct customer segments within this dataset. Increasing the number of clusters beyond this point would add to the model's complexity without providing a meaningful improvement in terms of segment differentiation.

c. [1.5v] Would *k*-modes be a better clustering approach? Explain why based on the dataset features.

Using k-modes might be a better clustering approach for this dataset because it contains a significant number of categorical variables, such as job, marital, education, default, and more. K-means is designed primarily for numerical data, where it minimizes SSE based on Euclidean distance. This method does not handle categorical variables well because it relies on continuous distance measures that are not meaningful for nominal data. In contrast, the k-modes algorithm is specifically designed for categorical data. It minimizes the dissimilarity between clusters by using matching distance (counting the number of mismatches between categorical attributes), which is more appropriate for nominal data. Therefore, using k-modes would likely result in clusters that more accurately reflect the true structure of customer segments in this dataset, as it would consider the categorical nature of most features.

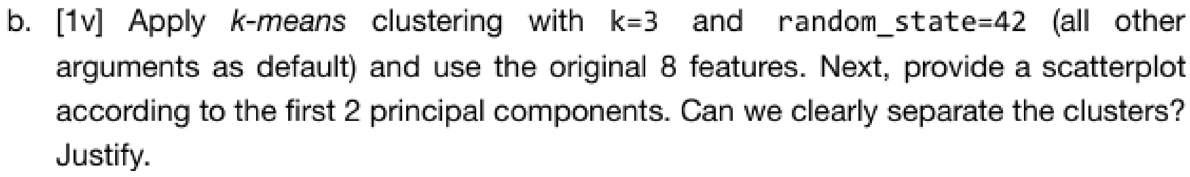


2. Normalize the data using StandardScaler:
   a. [1v] Apply PCA to the data. How much variability is explained by the top 2 components?

The explained variance ratio, which was printed in the output, shows the proportion of the dataset's total variance captured by the first two components. This value represents how much of the information (or variability) in the original eight features is retained in the two-dimensional representation created by PCA.

If the combined explained variance of the first two components is high (typically above 70-80%), it indicates that most of the information from the original features is preserved in these two dimensions, which makes this reduced representation a good summary of the data. On the other hand, if the explained variance is lower, it may suggest that more components are needed to retain an adequate amount of the original data's information.

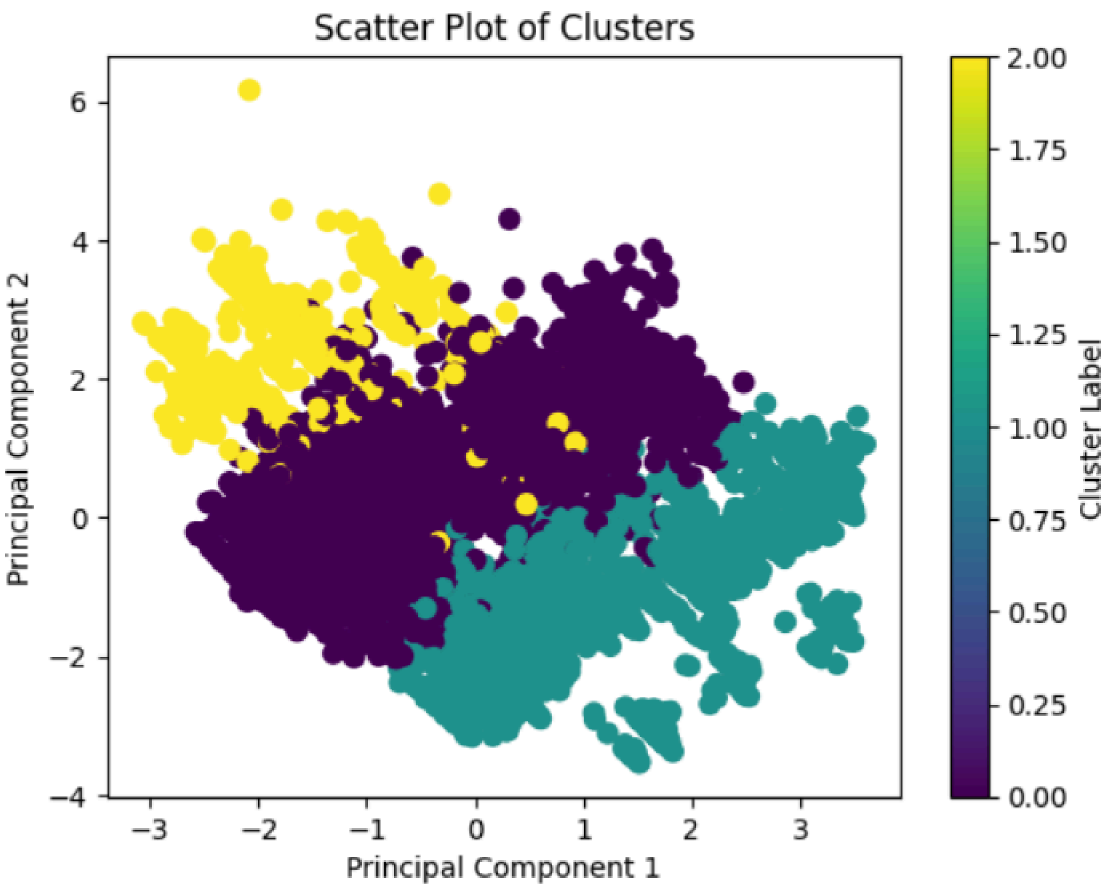
This explained variance metric is essential for assessing the effectiveness of PCA as a dimensionality reduction technique, as it helps determine how well we can represent the data with just two principal components.



b. [1v] Apply *k-means* clustering with `k=3` and `random_state=42` (all other arguments as default) and use the original 8 features. Next, provide a scatterplot according to the first 2 principal components. Can we clearly separate the clusters? Justify.

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
data = pd.read_csv('accounts.csv')
```

```
# Select the first 8 features and preprocess
data = data.iloc[:, :8]
data = data.drop_duplicates().dropna()
data_encoded = pd.get_dummies(data, drop_first=True)

# Normalize data using StandardScaler
scaler = StandardScaler()
data_standardized = scaler.fit_transform(data_encoded)

# Apply PCA
pca = PCA(n_components=2)
pca_components = pca.fit_transform(data_standardized)

# Explained variance ratio
explained_variance = pca.explained_variance_ratio_
print(f"Explained variance by the first 2 components:
{explained_variance[0] + explained_variance[1]:.2f}")



# Apply K-Means clustering with k=3
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(data_standardized)

# Scatter plot using the first two principal components
plt.scatter(pca_components[:, 0], pca_components[:, 1], c=clusters,
cmap='viridis', s=50)
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('Scatter Plot of Clusters')
plt.colorbar(label='Cluster Label')
plt.show()
```



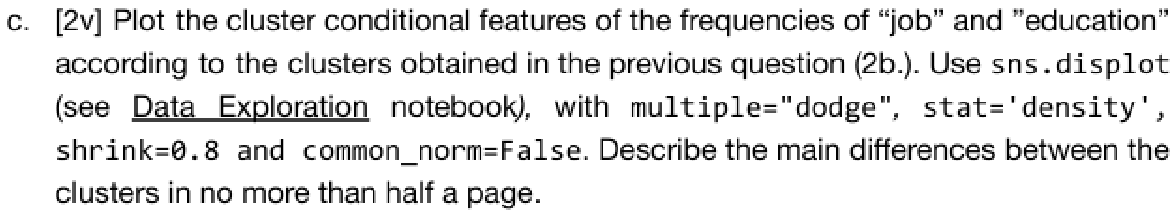
The explained variance calculated by the code shows how much of the original data's variability is retained when projecting it onto the two-dimensional space defined by the first two components. A higher explained variance indicates that the 2D representation is a more accurate reflection of the original dataset. The printed value reveals how much of the data's total variance is captured by these components.

The scatter plot shows clusters in the space of the two principal components, where each point represents an observation, and each color indicates a cluster. If the points from different clusters are well separated in the plot, it suggests that the K-means algorithm effectively identified distinct groups within the data. In contrast, if there is significant overlap between clusters, it may indicate that the groups are less distinct and that the number of clusters might need to be adjusted.

Observing the shape and dispersion of each cluster can also reveal insights. For instance, clusters that appear more spread out indicate greater internal variability, while more compact clusters suggest more homogeneous groupings.

If the clusters are clearly separated, then selecting three clusters (with $k = 3$ k=3) appears to be appropriate for this dataset, suggesting that the identified customer segments are meaningfully distinct. If the separation is less clear, it could be worth trying different values for k k or exploring alternative clustering methods, such as k-modes, especially if the dataset includes many categorical variables, as suggested in the previous question.

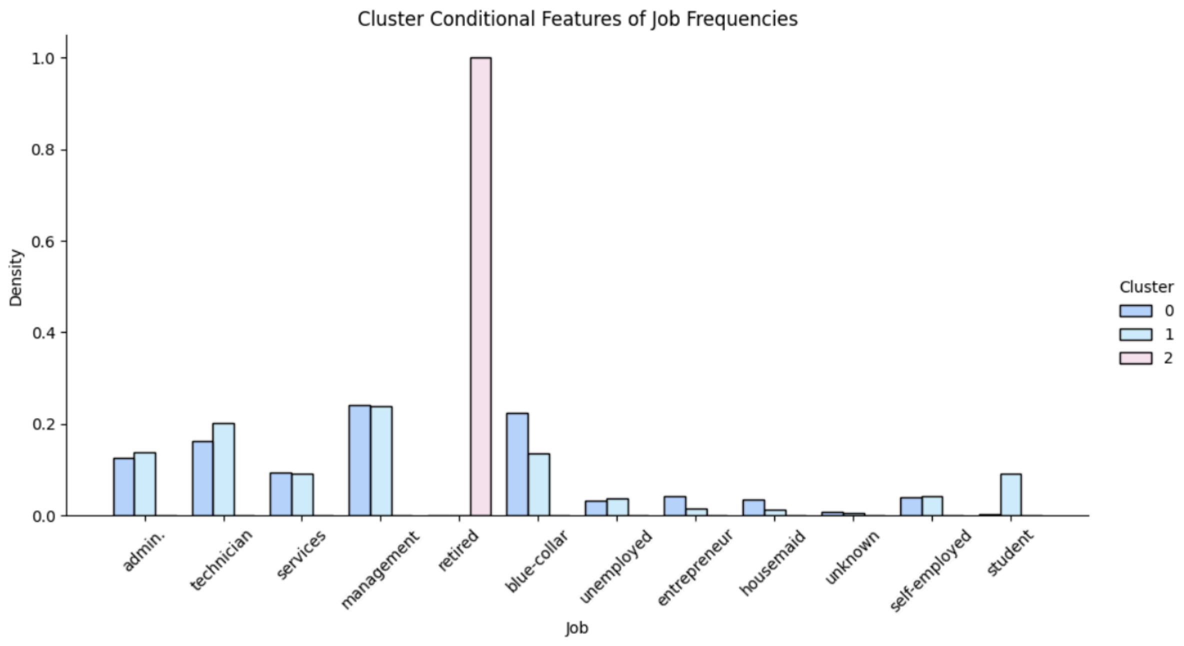
This visual analysis helps evaluate whether the chosen number of clusters is suitable and provides an initial understanding of each customer segment's characteristics.

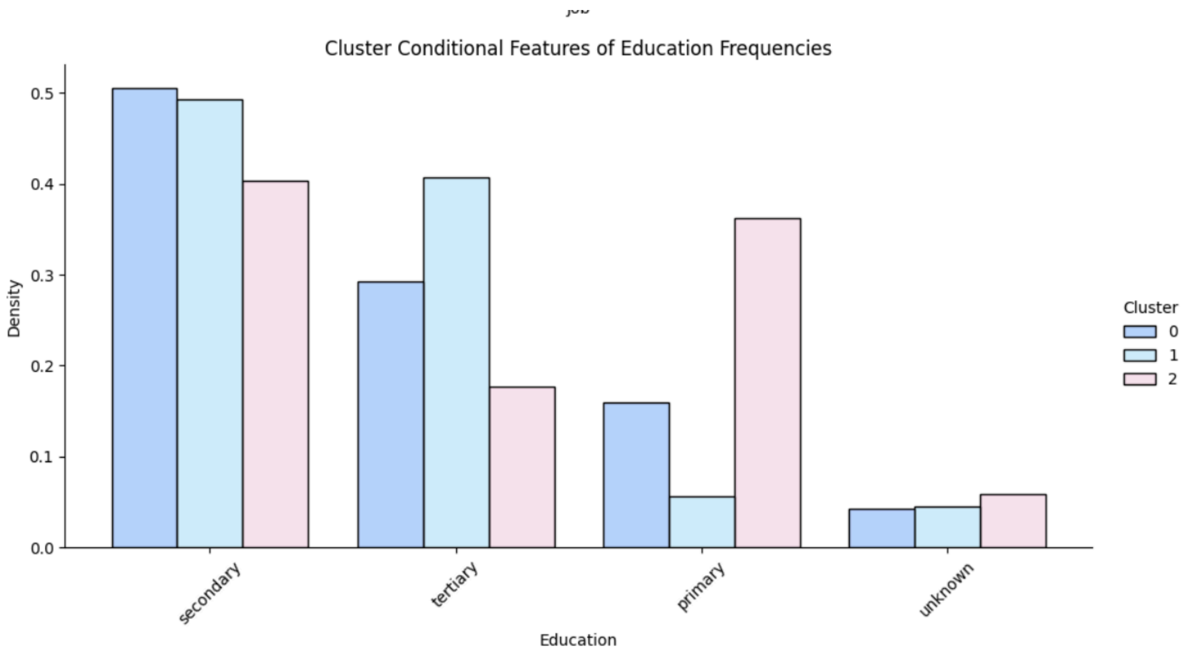


c. [2v] Plot the cluster conditional features of the frequencies of “job” and ”education” according to the clusters obtained in the previous question (2b.). Use `sns.displot` (see Data Exploration notebook), with `multiple="dodge"`, `stat='density'`, `shrink=0.8` and `common_norm=False`. Describe the main differences between the clusters in no more than half a page.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans

data = pd.read_csv('accounts.csv')

data = data.iloc[:, :8].drop_duplicates().dropna()

data_categorical = data[['job', 'education']].copy()

data_encoded = pd.get_dummies(data, drop_first=True)

scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_encoded)

pca = PCA(n_components=2)
data_pca = pca.fit_transform(data_scaled)

kmeans = KMeans(n_clusters=3, random_state=42)
clusters  =  kmeans.fit_predict(data_scaled)    #  Use  scaled  data  for
clustering

data_categorical['Cluster'] = clusters
```

```
palette = ["#A1C4FD", "#C2E9FB", "#F6D8E8"]

sns.displot(data=data_categorical, x='job', hue='Cluster', kind="hist",
multiple="dodge",
                        stat='density',  shrink=0.8,  common_norm=False,
palette=palette, aspect=2.0)
plt.xticks(rotation=45)
plt.title("Cluster Conditional Features of Job Frequencies")
plt.xlabel("Job")
plt.ylabel("Density")
plt.show()

sns.displot(data=data_categorical,    x='education',    hue='Cluster',
kind="hist", multiple="dodge",
                        stat='density',  shrink=0.8,  common_norm=False,
palette=palette, aspect=2.0)
plt.xticks(rotation=45)
plt.title("Cluster Conditional Features of Education Frequencies")
plt.xlabel("Education")
plt.ylabel("Density")
plt.show()
```





Beginning with the "Job" distribution across clusters, each bar indicates the relative density of various job categories within each cluster. For instance, if Cluster 0 exhibits a high concentration of roles such as "Engineer," it suggests that this group is composed of individuals with specialized skills or careers in technology. On the other hand, if Cluster 1 shows a significant number of lower-skilled positions, like "Attendant" or "Laborer," it may imply that this cluster includes customers with less formal education or those in entry-level jobs. Additionally, if Cluster 2 features a wider variety of professions or a noticeable presence of roles in creative or service sectors, this could reflect a more diverse occupational background among its members.

Now, turning to the "Education" distribution by clusters, a similar interpretation can be applied. The bars in this graph represent the density of different educational attainment levels for each cluster. For example, if Cluster 0 predominantly consists of individuals with higher education degrees, it may indicate a group of customers with greater socioeconomic status or better access to educational resources. In contrast, if a cluster contains a large proportion of individuals with only a high school education, it might represent a demographic facing more limited educational opportunities. If Cluster 2 displays a more balanced range of educational backgrounds, including vocational training, it could suggest that this segment values practical skills and training.

Overall, the variations in the "job" and "education" distributions across the clusters provide valuable insights into the demographic and professional profiles of the customers. Understanding these characteristics is vital for effectively segmenting the market and tailoring marketing strategies. The findings from these distributions can guide decisions on the types of products and services to offer each group, leading to more personalized and impactful marketing efforts.