Estadística Multivariada
Cuestionario Final

Valeria Zamora Israel Garcia Mateo Valencia Marian Medina André Chavéz

26 de Mayo de 2024

Ejercicio 1.

Explica en qué consiste el analisis por correspondencias. Realiza una tabla comparativa con respecto al análisis por componentes principales.

Respuesta:

El análisis de correspondencias es método estadístico que se usa para explorar la relación entre dos o más variables categóricas. Sirve para analizar tablas de contingencia, donde se registran las frecuencias de ocurrencia de varias categorías en dos o más variables. Su objetivo principal es poder visualizar y comprender la estructura de como se asocian las categorías de las variables entre sí, lo que puede ayudar a identificar patrones, tendencias o relaciones latentes en los datos.

Análisis de correspondencias	**Análisis PCA**
Variables categóricas	Variables cuantitativas
Relaciones entre variables	Reduce dimensión de datos
Tabla de contingencia	matriz de covarianza
Coeficientes de asociación	correlaciones lineales
Resultado: mapa de coordenadas	componentes principales
Sin procesamiento	normalización de datos

(1)

Ejercicio 2.

Realiza un análisis de correspondencias múltiple con los datos de iris. ¿Cuáles son tus conclusiones?

Al realizar el análisis por correspondencias múltiples se obtuvo lo siguiente:
Conclusiones:

- La especie Setosa esta separada debido a sus características distintivas en términos de longitud y anchura de sépalo y pétalo. Lo cual nos indica que es bastante diferente a las demás especies.
- Hay una entre las categorías de longitud y anchura de sépalo y pétalo, pues hay una clara tendencia sobre que las flores con pétalos largos tienden a tener sépalos largos.
- Las especies Versicolor y Virginica comparten características más similares entre ellas en comparación con Setosa.

Ejercicio 3.

Realiza un análisis del análisis factorial y su significado. Puedes tomar como referencia: `https://online.stat.psu.edu/stat505/lesson/12`

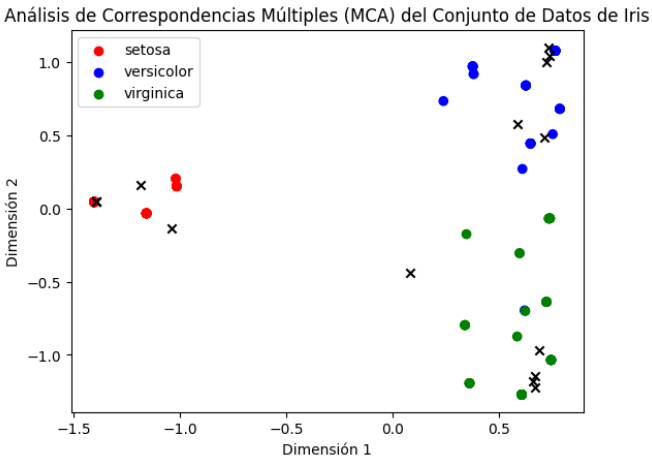


Figura 1: MCA Datos IRIS

El análisis factorial es un método estadístico que se utiliza para analizar las relaciones subyacentes o latentes entre variables observadas.
Su objetivo principal es identificar patrones en estas variables para así poder obtener factores y reducir la dimensión de los datos al resumir la información de todo el conjunto en estos factores.

Los factores capturan la información común entre las variables observadas y ayudan a entender más facilmente la estructura de los datos. Cada variable observada es una combinación lineal de estos.
Hay dos tipos:

- **Análisis factorial exploratorio:** Se usa cuando no hay hipótesis claras sobre el número de factores o la relación entre las variables,
- **Análisis factorial confirmatorio:** Se usa para ver que tan adecuado es un número de factores que ya se sospeche.

Proceso:

1. Selección de variables.
2. Se determina el número de factores: con el método de componentes principales, entre otros.
3. Extracción de factores: con el método de máxima verosimilitud.
4. Rotación de factores para facilitar su interpretación.
5. Interpretación.

Este análisis se usa en áreas como la piscología, la sociología, la economía y muchos otros.

Ejercicio 4. 13.1

Mostrar que los supuestos llevan a (13.2), $\text{var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \psi_i$.
Queremos demostrar que $\text{var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \psi_i$. Primero, recordamos el modelo de factores:

$$y_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{im} f_m + \epsilon_i$$

La varianza de y_i se expresa como:

$$\text{var}(y_i) = \text{var}(\lambda_{i1}f_1 + \lambda_{i2}f_2 + \cdots + \lambda_{im}f_m + \epsilon_i)$$

Dado que los factores f_j y el término de error ϵ_i son independientes, podemos separar las varianzas:

$$\text{var}(y_i) = \text{var}(\lambda_{i1}f_1) + \text{var}(\lambda_{i2}f_2) + \cdots + \text{var}(\lambda_{im}f_m) + \text{var}(\epsilon_i)$$

Para cada término $\lambda_{ij}f_j$, la varianza es $\lambda_{ij}^2\text{var}(f_j)$. Dado que $\text{var}(f_j) = 1$:

$$\text{var}(\lambda_{ij}f_j) = \lambda_{ij}^2 \cdot 1 = \lambda_{ij}^2$$

La varianza del término de error ϵ_i es ψ_i, es decir:

$$\text{var}(\epsilon_i) = \psi_i$$

Sumando todas estas varianzas obtenemos:

$$\text{var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \psi_i$$

Ejercicio 4. 13.2

Verify directly that cov(y, f) = Λ as in (13.13).

$$\begin{aligned} cov(y,f) &= cov(\Lambda f + \varepsilon, f) && (2)\\ &= cov(\Lambda f, f) && (3)\\ &= E[\Lambda f - E(\Lambda f)][f - E(f)]' && (4)\\ &= E[-(f)][f - E(f)]' && (5)\\ &= [f - E(f)][f - E(f)]' && (6)\\ &= \Lambda cov(f) = \Lambda && (7) \end{aligned}$$

Ejercicio 4. 13.3

Show that $f* = T'f$ in 13,18 satisfies the assumptions 13,6 and 13,7, $Ef* = 0$, $Cov(f*) = I$.

1. Calculamos la media de f*

$$E(f*) = E(T'f) \tag{8}$$

Pero T' es una matriz constante y E(f)=0, entonces

$$E(f*) = T'E(f) = T' * 0 = 0 \tag{9}$$

2. Calculamos la Cov de f*

$$Cov(f*) = Cov(T'f) \tag{10}$$

Pero T' es una matriz constante y Cov(f)=I

$$Cov(f*) = T'Cov(f)T = T'IT = T'T = I \tag{11}$$

porque T' es ortogonal

Y se satisface 13.6 y 13.7

Ejercicio 4. 13.5

Show that the sum $\sum_{i=1}^{p}\sum_{j=1}^{m}(\hat{\lambda}_{ij})^2$ is equal to the sum of the first m eigenvalues and also equal to the sum of all p communalities, as in $\sum_{i=1}^{p}\sum_{j=1}^{m}(\hat{\lambda}_{ij})^2 = \sum_{i=1}^{p}(h_i)^2 = \sum_{j=1}^{m}\theta_j$.

Repuesta: Sean:

- h_i^2: comunalidad de la variable i
- $\hat{\lambda}_{ij}^2$: elemento ij de la matriz de cargas factoriales
- θ_i: eigenvalor i

$$\sum_{i=1}^{p}\left[\sum_{j=1}^{m}(\hat{\lambda}_{ij})^2\right] = \sum_{i=1}^{p}\hat{h}_i^2$$

Ahora, cambiando el orden de la suma, se tiene que:

$$\sum_{j=1}^{m}\left[\sum_{i=1}^{p}(\hat{\lambda}_{ij})^2\right] = \sum_{j=1}^{m}\theta_j$$

Por lo tanto, podemos concluir que la triple igualdad se cumple.

Ejercicio 5.

Considera el siguiente conjunto de puntos (0, 0), (0, 1), (-1, 2), (2, 0), (3, 0), (4,-1).

1. Calcula la matriz de disimilaridades.
2. Realiza un K-means. Usa (0, 0) y (4,-1) como centroides iniciales (K = 2). ¿A qué cluster pertenece el punto (1, 1)?

Repuesta:

1. Matriz de disimilaridades:

$$\begin{bmatrix} 0{,}0 & 1{,}0 & 2{,}2360 & 2{,}0 & 3{,}0 & 4{,}1231 \\ 1{,}0 & 0{,}0 & 1{,}4142 & 2{,}2360 & 3{,}1622 & 4{,}4721 \\ 2{,}2360 & 1{,}4142 & 0{,}0 & 3{,}6055 & 4{,}4721 & 5{,}8309 \\ 2{,}0 & 2{,}2360 & 3{,}6055 & 0{,}0 & 1{,}0 & 2{,}2360 \\ 3{,}0 & 3{,}1622 & 4{,}4721 & 1{,}0 & 0{,}0 & 1{,}4142 \\ 4{,}1231 & 4{,}4721 & 5{,}8309 & 2{,}2360 & 1{,}4142 & 0{,}0 \end{bmatrix}$$

2. El punto (1, 1) pertenece al clúster 0.

Ejercicio 6.

Para un conjunto de puntos $(x_i)_{i=1}^{n}$ en R^m, demuestra que la media muestral $\hat{\mu}$ es la solución al problema de optimización:

$$\hat{\mu} = \arg\min_{\mu \; in R^m} \sum_{i=1}^{n} d_2(x_n,\mu)^2 \tag{12}$$

Consideremos la función de coste $J(\mu)$ que queremos minimizar:

$$J(\mu) = \sum_{i=1}^{n} d_2\left(x_i,\mu\right)^2 = \sum_{i=1}^{n}\left\|x_i-\mu\right\|^2$$

Donde $\|x_i - \mu\|$ es la distancia euclidiana entre el punto x_i y μ. Para minimizar $J(\mu)$, calculamos el gradiente de J con respecto a μ y lo igualamos a cero:

$$\nabla_\mu J(\mu) = \nabla_\mu \sum_{i=1}^{n} \|x_i - \mu\|^2 = \nabla_\mu \sum_{i=1}^{n} (x_i - \mu)^T (x_i - \mu).$$

Calculando el gradiente obtenemos:

$$\nabla_\mu J(\mu) = \nabla_\mu \sum_{i=1}^{n} \left(\|x_i\|^2 - 2x_i^T \mu + \|\mu\|^2\right) = \sum_{i=1}^{n} (-2x_i + 2\mu).$$

Igualando a cero:

$$\sum_{i=1}^{n} (-2x_i + 2\mu) = 0 \Rightarrow \sum_{i=1}^{n} \mu - \sum_{i=1}^{n} x_i = 0 \Rightarrow n\mu = \sum_{i=1}^{n} x_i.$$

Resolviendo para μ :

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Por lo tanto, la media muestral $\hat{\mu}$ es la solución al problema de optimización ya que minimiza la suma de las distancias al cuadrado de los puntos al promedio.

Ejercicio 7.

Supongamos que se tienen n observaciones, cada una con p características. Determina cuáles de los siguientes enunciados son verdaderos con respecto al análisis de clusters:

1. Podemos agrupar las n observaciones sobre la base de las p caracteríssticas para identificar subgrupos entre las observaciones.
2. Podemos agrupar las p características sobre la base de las n- observaciones para descubrir subgrupos entre las características.
3. El análisis por clusters es parte del aprendizaje supervisado y es parte del análisis exploratorio de datos.

Justifica tus respuestas.

Repuesta:

1. **Verdadero.** El análisis de cluesters se usa precisamente para agrupar observaciones de acuerdo a sus características para poder diferenciar grupos, encontrar patroes y estructuras latentes en los datos.
2. **Verdadero.** Para el caso del análisis de clústers en las características, donde las caracteristicas son las que se agrupan según su comportamiento en las observaciones y ayuda a identificar si hay grupos de características que se comportan de forma similar o si hay algunas que estén correlacionadas.
3. **Falso.** El análisis de clústers no es parte del aprendizaje supervisado, ya que este usa datos etiquetados que entrenar el modelo de predicción, como lo hace la regresión. De hecho, es una técnica de aprendizaje no supervisado pues no usa etiquetas que agrupen datos. Pero lo que es **Verdadero** es que es parte del análisis exploratorio de datos, pues el EDA incluye técnicas que resumen las principales características de un conjunto de datos.

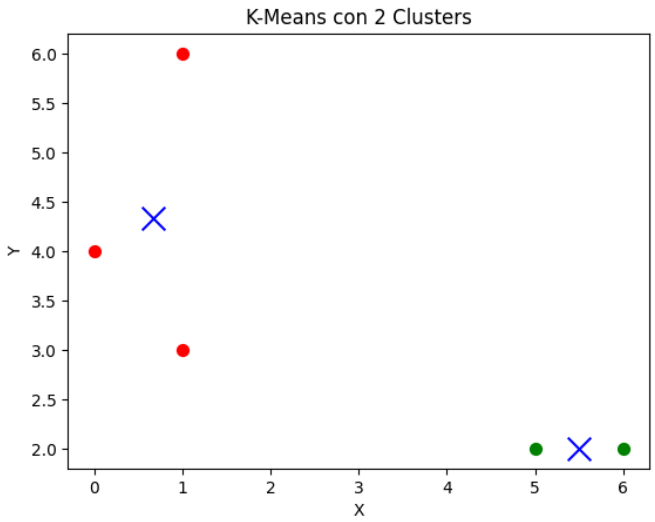
Ejercicio 8.

(*) Realiza un algoritmo de K-means (K = 2) dadas las asignaciones siguientes:

Observación	X_1	X_2	Clúster inicial
1	1	3	2
2	0	4	1
3	6	2	2
4	5	2	2
5	1	6	1

Determina las asignaciones finales de los agrupamientos.

Aplicando el agoritmo de k-means con K=2 y se obtuvieron las asignaciones siguientes:



Centroides en (0.66,4.33) y (5.5,2)

X1	X2	Clúster final
1	3	1
0	4	1
6	2	2
5	2	2
1	6	1

Ejercicio 9.

Realiza este ejercicio a mano y en Python. Considera los siguientes puntos:

(2, 10), (2, 5), (8, 4), (5, 8), (7, 5), (6, 4), (1, 2), (4, 9)

1. Usando el algoritmo de k-means, agrupa los puntos en 3 clústers.

2. Realiza un análisis de clúster usando single-link, complete-link y average-link para agrupar los puntos dados.

(1) Inicializamos a los centroides. Seleccionamos aleatoriamente los siguientes centroides iniciales:

$$C_1 : (2, 10)$$
$$C_2 : (8, 4)$$
$$C_3 : (1, 2)$$

Iteración 1

Asignación de Puntos a Clústers

Calculamos las distancias de cada punto a los centroides usando la fórmula:

$$\text{Distancia} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Punto	**Distancia a C_1**	**Distancia a C_2**	**Distancia a C_3**	**Clúster Asignado**
(2, 10)	0	8.49	8.06	1
(2, 5)	5	6.08	3.16	3
(8, 4)	8.49	0	7.28	2
(5, 8)	3.61	5	7.21	1
(7, 5)	7.07	1.41	6.71	2
(6, 4)	7.21	2	5.39	2
(1, 2)	8.06	7.28	0	3
(4, 9)	2.24	6.4	7.62	1

Cuadro 1: Distancias de los puntos a los centroides y clúster asignado

Actualización de Centroides

Calculamos los nuevos centroides utilizando la media de los puntos asignados a cada clúster:

$$C_1 : \left(\frac{2+5+4}{3}, \frac{10+8+9}{3}\right) = (3{,}67, 9{,}0)$$
$$C_2 : \left(\frac{8+7+6}{3}, \frac{4+5+4}{3}\right) = (7{,}0, 4{,}33)$$
$$C_3 : \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1{,}5, 3{,}5)$$

Iteración 2

Repetimos el proceso de asignación y actualización con los nuevos centroides. Como los clústers no cambian, detenemos el proceso, ya que hemos alcanzado la convergencia.

Resultados

- Clúster 1: (2, 10), (5, 8), (4, 9)
- Clúster 2: (8, 4), (7, 5), (6, 4)
- Clúster 3: (2, 5), (1, 2)

(2)

Single link

Clúster	Punto
1	$(2, 10)$
2	$(2, 5)$
3	$(8, 4)$
4	$(5, 8)$
5	$(7, 5)$
6	$(6, 4)$
7	$(1, 2)$
8	$(4, 9)$

Paso 1: Calcular distancias mínimas y fusionar clústers más cercanos Se utiliza la distancia euclidiana para calcular las distancias entre clústers.

Clústers	Distancia	Distancia mínima
$(1, 2)$	5	
$(1, 3)$	$\sqrt{72}$	
$(1, 4)$	$\sqrt{13}$	
$(1, 5)$	$\sqrt{50}$	
$(1, 6)$	$\sqrt{52}$	
$(1, 7)$	$\sqrt{65}$	
$(1, 8)$	$\sqrt{5}$	mínima
$(2, 3)$	$\sqrt{37}$	
$(2, 4)$	$\sqrt{18}$	
$(2, 5)$	5	
$(2, 6)$	$\sqrt{17}$	
$(2, 7)$	$\sqrt{10}$	
$(2, 8)$	$\sqrt{20}$	
$(3, 4)$	$\sqrt{25}$	
$(3, 5)$	$\sqrt{2}$	mínima
$(3, 6)$	2	mínima
$(3, 7)$	$\sqrt{53}$	
$(3, 8)$	$\sqrt{41}$	
$(4, 5)$	$\sqrt{13}$	
$(4, 6)$	$\sqrt{17}$	
$(4, 7)$	$\sqrt{52}$	
$(4, 8)$	$\sqrt{2}$	mínima
$(5, 6)$	$\sqrt{2}$	mínima
$(5, 7)$	$\sqrt{45}$	
$(5, 8)$	$\sqrt{25}$	
$(6, 7)$	$\sqrt{29}$	
$(6, 8)$	$\sqrt{29}$	
$(7, 8)$	$\sqrt{58}$	

Fusionamos los clústers (3, 5) y (3, 6) porque tienen la distancia mínima $\sqrt{2}$:

Clúster	Puntos
1	$(2, 10)$
2	$(2, 5)$
3	$(8, 4), (7, 5), (6, 4)$
4	$(5, 8)$
7	$(1, 2)$
8	$(4, 9)$

Paso 2: Calcular distancias mínimas y fusionar clústers más cercanos. Repetimos el cálculo de las distancias mínimas y fusionamos los clústers más cercanos: - Distancia mínima: Distancia entre (4, 8) = $\sqrt{2}$ Fusionamos los clústers (4, 8):

Clúster	Puntos
1	$(2, 10)$
2	$(2, 5)$
3	$(8, 4), (7, 5), (6, 4)$
4	$(5, 8), (4, 9)$
7	$(1, 2)$

Repetimos los pasos de la misma manera
Paso 3:
- Distancia mínima: Distancia entre (2, 7) = $\sqrt{10}$ Fusionamos los clústers (2, 7):

Clúster	Puntos
1	$(2, 10)$
2	$(2, 5), (1, 2)$
3	$(8, 4), (7, 5), (6, 4)$
4	$(5, 8), (4, 9)$

Paso 4:
- Distancia mínima: Distancia entre (1, 4) = $\sqrt{5}$ Fusionamos los clústers (1, 4):

Clúster	Puntos
1	$(2, 10), (5, 8), (4, 9)$
2	$(2, 5), (1, 2)$
3	$(8, 4), (7, 5), (6, 4)$

Paso 5:
- Distancia mínima: Distancia entre (2, 3) = $\sqrt{2}$ Fusionamos los clústers (2, 3):

Clúster	Puntos
1	$(2, 10), (5, 8), (4, 9)$
2	$(2, 5), (1, 2), (8, 4), (7, 5), (6, 4)$

Paso 6:
- Distancia mínima: Distancia entre (1, 2) = $\sqrt{5}$ Fusionamos los clústers (1, 2):

Clúster	Puntos
1	$(2, 10), (5, 8), (4, 9), (2, 5), (1, 2), (8, 4), (7, 5), (6, 4)$

Average link

Clúster	Punto
1	$(2, 10)$
2	$(2, 5)$
3	$(8, 4)$
4	$(5, 8)$
5	$(7, 5)$
6	$(6, 4)$
7	$(1, 2)$
8	$(4, 9)$

Paso 1: Calcular distancias promedio y fusionar clústers más cercanos Se utiliza la distancia euclidiana promedio para calcular las distancias entre clústers.

Clústers	Distancia	Distancia mínima
$(1, 2)$	5	
$(1, 3)$	$\sqrt{72}$	
$(1, 4)$	$\sqrt{13}$	
$(1, 5)$	$\sqrt{50}$	
$(1, 6)$	$\sqrt{52}$	
$(1, 7)$	$\sqrt{65}$	
$(1, 8)$	$\sqrt{5}$	
$(2, 3)$	$\sqrt{37}$	
$(2, 4)$	$\sqrt{18}$	
$(2, 5)$	5	
$(2, 6)$	$\sqrt{17}$	
$(2, 7)$	$\sqrt{10}$	
$(2, 8)$	$\sqrt{20}$	
$(3, 4)$	$\sqrt{25}$	
$(3, 5)$	$\sqrt{2}$	mínima
$(3, 6)$	2	mínima
$(3, 7)$	$\sqrt{53}$	
$(3, 8)$	$\sqrt{41}$	
$(4, 5)$	$\sqrt{13}$	
$(4, 6)$	$\sqrt{17}$	
$(4, 7)$	$\sqrt{52}$	
$(4, 8)$	$\sqrt{2}$	mínima
$(5, 6)$	$\sqrt{2}$	mínima
$(5, 7)$	$\sqrt{45}$	
$(5, 8)$	$\sqrt{25}$	
$(6, 7)$	$\sqrt{29}$	
$(6, 8)$	$\sqrt{29}$	
$(7, 8)$	$\sqrt{58}$	

Fusionamos los clústers (3, 5) y (3, 6) porque tienen la distancia mínima $\sqrt{2}$:

Clúster	Puntos
1	$(2, 10)$
2	$(2, 5)$
3	$(8, 4), (7, 5), (6, 4)$
4	$(5, 8)$
7	$(1, 2)$
8	$(4, 9)$

Paso 2: Repetimos el cálculo de las distancias promedio y fusionamos los clústers más cercanos:

Clústers	Distancia	Distancia mínima
$(1, 2)$	5	
$(1, 3)$	$\sqrt{72}$	
$(1, 4)$	$\sqrt{13}$	
$(1, 7)$	$\sqrt{65}$	
$(1, 8)$	$\sqrt{5}$	mínima
$(2, 3)$	$\sqrt{37}$	
$(2, 4)$	$\sqrt{18}$	
$(2, 7)$	$\sqrt{10}$	
$(2, 8)$	$\sqrt{20}$	
$(3, 4)$	$\sqrt{25}$	
$(3, 7)$	$\sqrt{53}$	
$(3, 8)$	$\sqrt{41}$	
$(4, 7)$	$\sqrt{52}$	
$(4, 8)$	$\sqrt{2}$	mínima
$(7, 8)$	$\sqrt{58}$	

Fusionamos los clústers (4, 8):

Clúster	Puntos
1	$(2, 10)$
2	$(2, 5)$
3	$(8, 4), (7, 5), (6, 4)$
4	$(5, 8), (4, 9)$
7	$(1, 2)$

Paso 3:

Clústers	Distancia	Distancia mínima
$(1, 2)$	5	
$(1, 3)$	$\sqrt{72}$	
$(1, 4)$	$\sqrt{13}$	
$(1, 7)$	$\sqrt{65}$	
$(2, 3)$	$\sqrt{37}$	
$(2, 4)$	$\sqrt{18}$	
$(2, 7)$	$\sqrt{10}$	mínima
$(3, 4)$	$\sqrt{25}$	
$(3, 7)$	$\sqrt{53}$	
$(4, 7)$	$\sqrt{52}$	

Fusionamos los clústers (2, 7):

Clúster	Puntos
1	$(2, 10)$
2	$(2, 5), (1, 2)$
3	$(8, 4), (7, 5), (6, 4)$
4	$(5, 8), (4, 9)$

Paso 4:

Clústers	Distancia	Distancia mínima
$(1, 2)$	5	
$(1, 3)$	$\sqrt{72}$	
$(1, 4)$	$\sqrt{13}$	mínima
$(2, 3)$	$\sqrt{37}$	
$(2, 4)$	$\sqrt{18}$	
$(3, 4)$	$\sqrt{25}$	

Fusionamos los clústers (1, 4):

Clúster	Puntos
1	$(2, 10), (5, 8), (4, 9)$
2	$(2, 5), (1, 2)$
3	$(8, 4), (7, 5), (6, 4)$

Paso 5:

Clústers	Distancia	Distancia mínima
$(1, 2)$	$\sqrt{5}$	mínima
$(1, 3)$	$\sqrt{72}$	
$(2, 3)$	$\sqrt{37}$	

Fusionamos los clústers (1, 2):

Clúster	Puntos
1	$(2, 10), (5, 8), (4, 9), (2, 5), (1, 2)$
3	$(8, 4), (7, 5), (6, 4)$

Paso 6:

Clústers	Distancia	Distancia mínima
$(1, 3)$	$\sqrt{5}$	mínima

Fusionamos los clústers (1, 3):

Clúster	Puntos
1	$(2, 10), (5, 8), (4, 9), (2, 5), (1, 2), (8, 4), (7, 5), (6, 4)$

Complete link

Clúster	Punto
1	$(2, 10)$
2	$(2, 5)$
3	$(8, 4)$
4	$(5, 8)$
5	$(7, 5)$
6	$(6, 4)$
7	$(1, 2)$
8	$(4, 9)$

Paso 1: Calcular distancias máximas y fusionar clústers más cercanos Se utiliza la distancia euclidiana máxima para calcular las distancias entre clústers.

Clústers	Distancia
$(1, 2)$	5
$(1, 3)$	$\sqrt{72} \approx 8{,}49$
$(1, 4)$	$\sqrt{13} \approx 3{,}61$
$(1, 5)$	$\sqrt{50} \approx 7{,}07$
$(1, 6)$	$\sqrt{52} \approx 7{,}21$
$(1, 7)$	$\sqrt{65} \approx 8{,}06$
$(1, 8)$	$\sqrt{5} \approx 2{,}24$
$(2, 3)$	$\sqrt{37} \approx 6{,}08$
$(2, 4)$	$\sqrt{18} \approx 4{,}24$
$(2, 5)$	5
$(2, 6)$	$\sqrt{17} \approx 4{,}12$
$(2, 7)$	$\sqrt{10} \approx 3{,}16$
$(2, 8)$	$\sqrt{20} \approx 4{,}47$
$(3, 4)$	$\sqrt{25} = 5$
$(3, 5)$	$\sqrt{2} \approx 1{,}41$
$(3, 6)$	2
$(3, 7)$	$\sqrt{53} \approx 7{,}28$
$(3, 8)$	$\sqrt{41} \approx 6{,}40$
$(4, 5)$	$\sqrt{13} \approx 3{,}61$
$(4, 6)$	$\sqrt{17} \approx 4{,}12$
$(4, 7)$	$\sqrt{52} \approx 7{,}21$
$(4, 8)$	$\sqrt{2} \approx 1{,}41$
$(5, 6)$	$\sqrt{2} \approx 1{,}41$
$(5, 7)$	$\sqrt{45} \approx 6{,}71$
$(5, 8)$	$\sqrt{25} = 5$
$(6, 7)$	$\sqrt{29} \approx 5{,}39$
$(6, 8)$	$\sqrt{29} \approx 5{,}39$
$(7, 8)$	$\sqrt{58} \approx 7{,}62$

Fusionamos los clústers (3, 5) y (5, 6) porque tienen la menor distancia máxima $\sqrt{2}$:

Clúster	Puntos
1	$(2, 10)$
2	$(2, 5)$
3	$(8, 4), (7, 5), (6, 4)$
4	$(5, 8)$
6	$(1, 2)$
7	$(4, 9)$

Paso 2: Calcular distancias máximas y fusionar clústers más cercanos Repetimos

Clústers	Distancia
$(1, 2)$	5
$(1, 3)$	$\sqrt{72} \approx 8{,}49$
$(1, 4)$	$\sqrt{13} \approx 3{,}61$
$(1, 6)$	$\sqrt{65} \approx 8{,}06$
$(1, 7)$	$\sqrt{5} \approx 2{,}24$
$(2, 3)$	$\sqrt{37} \approx 6{,}08$
$(2, 4)$	$\sqrt{18} \approx 4{,}24$
$(2, 6)$	$\sqrt{10} \approx 3{,}16$
$(2, 7)$	$\sqrt{20} \approx 4{,}47$
$(3, 4)$	$\sqrt{25} = 5$
$(3, 6)$	$\sqrt{53} \approx 7{,}28$
$(3, 7)$	$\sqrt{41} \approx 6{,}40$
$(4, 6)$	$\sqrt{52} \approx 7{,}21$
$(4, 7)$	$\sqrt{2} \approx 1{,}41$
$(6, 7)$	$\sqrt{58} \approx 7{,}62$

Fusionamos los clústers (4, 7) porque tienen la menor distancia máxima $\sqrt{2}$:

Clúster	Puntos
1	$(2, 10)$
2	$(2, 5)$
3	$(8, 4), (7, 5), (6, 4)$
4	$(5, 8), (4, 9)$
6	$(1, 2)$

Paso 3:

Clústers	Distancia
$(1, 2)$	5
$(1, 3)$	$\sqrt{72} \approx 8{,}49$
$(1, 4)$	$\sqrt{13} \approx 3{,}61$
$(1, 6)$	$\sqrt{65} \approx 8{,}06$
$(2, 3)$	$\sqrt{37} \approx 6{,}08$
$(2, 4)$	$\sqrt{18} \approx 4{,}24$
$(2, 6)$	$\sqrt{10} \approx 3{,}16$
$(3, 4)$	$\sqrt{25} = 5$
$(3, 6)$	$\sqrt{53} \approx 7{,}28$
$(4, 6)$	$\sqrt{52} \approx 7{,}21$

Fusionamos los clústers (1, 4) porque tienen la menor distancia máxima $\sqrt{13}$:

Clúster	Puntos
1	$(2,10),(5,8),(4,9)$
2	$(2,5)$
3	$(8,4),(7,5),(6,4)$
6	$(1,2)$

Paso 4:

Clústers	Distancia
$(1,2)$	5
$(1,3)$	$\sqrt{72} \approx 8{,}49$
$(1,6)$	$\sqrt{65} \approx 8{,}06$
$(2,3)$	$\sqrt{37} \approx 6{,}08$
$(2,6)$	$\sqrt{10} \approx 3{,}16$
$(3,6)$	$\sqrt{53} \approx 7{,}28$

Fusionamos los clústers (2, 6) porque tienen la menor distancia máxima $\sqrt{10}$:

Clúster	Puntos
1	$(2,10),(5,8),(4,9)$
2	$(2,5),(1,2)$
3	$(8,4),(7,5),(6,4)$

Paso 5:

Clústers	Distancia
$(1,2)$	5
$(1,3)$	$\sqrt{72} \approx 8{,}49$
$(2,3)$	$\sqrt{37} \approx 6{,}08$

Fusionamos los clústers (1, 2) porque tienen la menor distancia máxima 5:

Clúster	Puntos
1	$(2,10),(5,8),(4,9),(2,5),(1,2)$
3	$(8,4),(7,5),(6,4)$

Paso 6:

Clústers	Distancia
$(1,3)$	$\sqrt{72} \approx 8{,}49$

Fusionamos los clústers (1, 3):

Clúster	Puntos
1	$(2,10),(5,8),(4,9),(2,5),(1,2),(8,4),(7,5),(6,4)$

Ejercicio 10.

Construye una base de datos de 20 tickers (que incluyan 10 empresas con tickers terminación.MX) en un umbral de tiempo de 6 años. Considera las siguientes estadísticas:

- Movimientos finales. Se obtienen a través de los precios diarios tomando el precio más alto menos el precio más bajo en ese día.

- Rendimientos. Se obtiene como el precio a la fecha corriente entre el precio del día anterior menos 1.

Realiza un análisis de cluster con respecto a tales estadísticas. ¿Cuáles son tus conclusiones? También, realiza una simulación Monte-Carlo para obtener la frontera eficiente de Markowitz. ¿Cuál es el índice de Sharpe?, ¿Qué pesos son los óptimos en tu portafolio?

Nota:En el Colab adjunto se encuentran las soluciones correspondientes a este ejercicio.