

# Distancias estadísticas y Escalado Multidimensional (Análisis de Coordenadas Principales)

Aurea Grané  
Departamento de Estadística  
Universidad Carlos III de Madrid

## Distancias estadísticas

El concepto de distancia entre objetos o individuos permite interpretar geoméricamente muchas técnicas clásicas del análisis multivariante, equivalentes a representar estos objetos como puntos de un espacio métrico adecuado.

Esta interpretación es posible no solamente cuando se dispone de variables cuantitativas, sino también, y sobretodo, cuando las variables observadas son de tipo más general, o incluso cuando no se dispone de variables propiamente dichas, siempre que tenga sentido obtener una medida de proximidad entre los objetos o individuos.

## I. Distancias para variables cuantitativas

Sean  $\mathbf{x}'_i = (x_{i1}, \dots, x_{ip})$   $\mathbf{x}'_j = (x_{j1}, \dots, x_{jp})$  las observaciones de dos objetos o individuos  $i, j$ , resultado de medir  $p$  variables  $X_1, \dots, X_p$  sobre ellos.

La distancia euclídea

$$\delta_E^2(i, j) = \sum_{k=1}^p (x_{ik} - x_{jk})^2 = (\mathbf{x}_i - \mathbf{x}_j)' (\mathbf{x}_i - \mathbf{x}_j),$$

**no es recomendable** cuando las  $X_j$  son las variables originales porque:

- no es invariante frente a cambios de escala de las variables,
- presupone que las variables son incorrelacionadas y de varianza unidad.

La distancia de Minkowski

$$\delta_{m_q}(i, j) = \left( \sum_{k=1}^p |x_{ik} - x_{jk}|^q \right)^{1/q}, \quad q > 0.$$

Presenta los mismos inconvenientes que  $\delta_E$  ( $\delta_E = \delta_{m_2}$ ) y, además, es difícilmente euclidianizable (veremos este concepto más adelante).

Casos particulares de la distancia de Minkowski son:

Distancia ciudad o de Manhattan ( $q = 1$ )

$$\delta_{m_1}(i, j) = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

Distancia dominante ( $q \rightarrow \infty$ )

$$\delta_{m_\infty}(i, j) = \max\{|x_{i1} - x_{j1}|, \dots, |x_{ip} - x_{jp}|\}$$

### Distancias invariantes frente a cambios de escala:

Distancia de Canberra (modificación de la distancia ciudad):

$$\delta_C(i, j) = \sum_{k=1}^p \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}$$

Distancia de Karl Pearson (modificación de la distancia euclídea):

$$\delta_K^2(i, j) = \sum_{k=1}^p \frac{(x_{ik} - x_{jk})^2}{s_k^2} = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}_0^{-1} (\mathbf{x}_i - \mathbf{x}_j),$$

donde  $\mathbf{S}_0 = \text{diag}(s_1^2, \dots, s_p^2)$  es la matriz diagonal que contiene las varianzas de  $X_1, \dots, X_p$ .

Esta expresión equivale a reescalar cada variable en unidades de desviación típica. El peso que se atribuye a la diferencia entre individuos es mayor cuanto menor es la dispersión en esa variable. Pero sigue suponiendo que las variables están incorrelacionadas.

Distancia de Mahalanobis:

$$\delta_M^2(i, j) = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j),$$

donde  $\mathbf{S}$  es la matriz de covarianzas de la matriz de datos  $\mathbf{X}$ .

Es adecuada como medida de discrepancia entre datos, porque

- es invariante frente a transformaciones lineales no singulares de las variables,
- $\delta_E = \delta_M$  cuando  $\mathbf{S} = \mathbf{I}$ , y  $\delta_K = \delta_M$  cuando  $\mathbf{S} = \text{diag}(s_1^2, \dots, s_p^2)$ ,
- tiene en cuenta las correlaciones entre las variables. Por ejemplo, no aumenta por el simple hecho de aumentar el número de variables observadas, sino que solamente aumentará cuando las nuevas variables no sean redundantes con respecto de la información aportada por las anteriores.

## II. Distancias para variables binarias

Sean  $X_1, \dots, X_p$   $p$  variables binarias con posibles valores  $\{0, 1\}$ .

Existen muchísimos coeficientes de similaridad  $s_{ij}$  entre dos individuos  $i, j$ , calculados a partir de las frecuencias:

$a$  = “número de variables con respuesta 1 en ambos individuos”,

$b$  = “número de variables con respuesta 0 en el individuo  $i$  y con respuesta 1 en el individuo  $j$ ”,

$c$  = “número de variables con respuesta 1 en el individuo  $i$  y con respuesta 0 en el individuo  $j$ ”,

$d$  = “número de variables con respuesta 0 en ambos individuos”.

Observad que  $a + b + c + d = p$ .

Algunos coeficientes de similaridad son:

$$\text{Sokal y Michener: } s_{ij} = \frac{a + d}{p}, \quad \text{Jaccard: } s_{ij} = \frac{a}{a + b + c}.$$

Aplicando uno de estos coeficientes a un conjunto de  $n$  objetos se obtiene una matriz de similaridades  $\mathcal{S} = (s_{ij})_{n \times n}$ .

**Ejemplo 1:** Se han medido 6 variables sobre 3 individuos:

ind.	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
1	1	1	0	0	1	1
2	1	1	1	0	0	1
3	1	0	0	1	0	1

ind.	a (1,1)			b (0,1)			c (1,0)			d (0,0)		
	1	2	3	1	2	3	1	2	3	1	2	3
1	4	3	2	0	1	1	0	1	2	2	1	1
2	3	4	2	1	0	1	1	0	2	1	2	1
3	2	2	3	2	2	0	1	1	0	1	1	3

ind.	a (1,1)			b (0,1)			c (1,0)			d (0,0)		
	1	2	3	1	2	3	1	2	3	1	2	3
1	4	3	2	0	1	1	0	1	2	2	1	1
2	3	4	2	1	0	1	1	0	2	1	2	1
3	2	2	3	2	2	0	1	1	0	1	1	3

La matrices de similitudes de Sokal y Michener y de Jaccard son:

$$S_{Sokal} = \begin{pmatrix} 1 & 0.6667 & 0.5 \\ 0.6667 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}, \quad S_{Jaccard} = \begin{pmatrix} 1 & 0.6 & 0.4 \\ 0.6 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix}$$

### III. Distancias para variables categóricas

Se mide una variable categórica nominal con  $k$  estados excluyentes sobre una muestra de  $n = n_1 + \dots + n_g$  individuos provenientes de  $g$  poblaciones diferentes. Se desea obtener una medida de disimilaridad entre estas poblaciones.

En estas condiciones, el vector de frecuencias de cada población  $\mathbf{n}_\alpha = (n_{\alpha 1}, \dots, n_{\alpha k})$ , para  $\alpha = 1, \dots, g$ , tiene una distribución conjunta multinomial con parámetros  $(n_\alpha, \mathbf{p}_\alpha)$ , donde  $n_\alpha = n_{\alpha 1} + \dots + n_{\alpha k}$  y  $\mathbf{p}_\alpha = (p_{\alpha 1}, \dots, p_{\alpha k})$  es el vector de probabilidades de los  $k$  estados en la población  $\alpha$  (con  $p_{\alpha 1} + \dots + p_{\alpha k} = 1$ ).

### Ejemplo 2. Problema 5.3

La siguiente tabla contiene las proporciones génicas observadas entre 10 poblaciones.

	Población	grupo A	grupo AB	grupo B	grupo O
1.	francesa	0.21	0.06	0.06	0.67
2.	checa	0.25	0.04	0.14	0.57
3.	germánica	0.22	0.06	0.08	0.64
4.	vasca	0.19	0.04	0.02	0.75
5.	china	0.18	0.00	0.15	0.67
6.	ainu	0.23	0.00	0.28	0.49
7.	esquimal	0.30	0.00	0.06	0.64
8.	negra USA	0.10	0.06	0.13	0.71
9.	española	0.27	0.04	0.06	0.63
10.	egipcia	0.21	0.05	0.20	0.54

Observad que las filas suman 1.

Dos medidas de disimilaridad para este tipo de variables son:

la distancia de Bhattacharyya, conocida en genética como distancia de Cavalli-Sforza:

$$d_{ij}^2 = \arccos \left( \sum_{l=1}^k \sqrt{p_{il} p_{jl}} \right)$$

y la distancia de Balakrishnan-Sanghvi:

$$d_{ij}^2 = 2 \sum_{l=1}^k \frac{(p_{il} - p_{jl})^2}{p_{il} + p_{jl}}$$

Para los datos del ejemplo 2, la matriz de cuadrados de distancias de Bhattacharyya es:

0	0.1567	0.0435	0.1246	0.2863	0.3966	0.2622	0.1850	0.0800	0.2204
0.1567	0	0.1156	0.2665	0.2240	0.2605	0.2476	0.2093	0.1364	0.0897
0.0435	0.1156	0	0.1660	0.2715	0.3636	0.2608	0.1769	0.0778	0.1787
0.1246	0.2665	0.1660	0.0000	0.3221	0.4732	0.2607	0.2555	0.1517	0.3359
0.2863	0.2240	0.2715	0.3221	0	0.1933	0.1896	0.2710	0.2653	0.2491
0.3966	0.2605	0.3636	0.4732	0.1933	0	0.3101	0.3701	0.3642	0.2422
0.2622	0.2476	0.2608	0.2607	0.1896	0.3101	0	0.3608	0.2024	0.3226
0.1850	0.2093	0.1769	0.2555	0.2710	0.3701	0.3608	0.0000	0.2438	0.1997
0.0800	0.1364	0.0778	0.1517	0.2653	0.3642	0.2024	0.2438	0	0.2211
0.2204	0.0897	0.1787	0.3359	0.2491	0.2422	0.3226	0.1997	0.2211	0

Los individuos más cercanos (según la distancia de Bhattacharyya medida sobre sus proporciones génicas) son las poblaciones *francesa* y *germánica* con  $\delta_{1,3}^2 = 0.0435$ , mientras que los más alejados son las poblaciones *francesa* y *ainu* con  $\delta_{1,6}^2 = 0.3966$ .

#### IV. Distancias para variables mixtas

Se dispone de un conjunto de datos mixto, es decir, un conjunto de individuos sobre los que se han observado tanto variables cuantitativas como cualitativas (o categóricas).

Se define la distancia de Gower como  $d_{ij}^2 = 1 - s_{ij}$ , donde

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3} \tag{1}$$

es el coeficiente de similitud de Gower,  $p_1$  es el número de variables cuantitativas continuas,  $p_2$  es el número de variables binarias,  $p_3$  es el número de variables cualitativas (no binarias),  $a$  es el número de coincidencias (1, 1) en las variables binarias,  $d$  es el número de coincidencias (0, 0) en las variables binarias,  $\alpha$  es el número de coincidencias en las variables cualitativas (no binarias) y  $G_h$  es el rango (o recorrido) de la  $h$ -ésima variable cuantitativa.

**Ejemplo 3. Problema 5.5:** Siete variables observadas sobre 50 jugadores de la liga española de fútbol 2006/07.

Jugador	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$
1. Ronaldinho	15	26	1.78	71	1	2	2
2. Etoo	21	25	1.8	75	0	3	2
3. Xavi	6	26	1.7	68	0	5	4
4. Messi	7	19	1.69	67	0	1	3
5. Puyol	1	28	1.78	78	0	5	3
6. Raúl	7	29	1.8	73.5	1	5	3
7. Ronaldo	18	30	1.83	82	0	2	1
8. Beckham	4	31	1.8	67	0	9	3
...				...			
50. Doblás	0	25	1.84	78	0	5	3

$X_1$  = número de goles marcados,  $X_2$  = edad (años),  $X_3$  = altura (m),  $X_4$  = peso (kg),  $X_5$  = pierna buena del jugador (1 = derecha, 0 = izquierda),  $X_6$  = nacionalidad (1 = Argentina, 2 = Brasil, 3 = Camerun, 4 = Italia, 5 = España, 6 = Francia, 7 = Uruguay, 8 = Portugal, 9 = Inglaterra),  $X_7$  = tipo de estudios (1 = sin estudios, 2 = básicos, 3 = medios, 4 = superiores).

#### Propiedades generales de las Distancias

$\delta : \mathcal{E} \times \mathcal{E} \rightarrow \mathbb{R}_+$  es una *disimilitud* o *casi-métrica* si

- $\forall i, j, \delta_{ij} = \delta_{ji}$ ,
- $\forall i, \delta_{ii} = 0$ .

Una *semi-métrica* es una disimilitud tal que

- $\forall i, j, k, \delta_{ij} \leq \delta_{ik} + \delta_{kj}$ .

Una *métrica* es una semi-métrica que cumple

- $\forall i, j, \delta_{ij} = 0 \Leftrightarrow i = j$ .

La palabra *distancia* puede hacer referencia tanto a una métrica como a una semi-métrica.

### Propiedades generales de las Similaridades

$s : \mathcal{E} \times \mathcal{E} \rightarrow \mathbb{R}$  es una *similaridad* si

- $\forall i, j, \quad 0 \leq s_{ij} \leq s_{ii} = 1,$
- $\forall i, j, \quad s_{ij} = s_{ji}.$

La transformación  $\delta_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$  permite obtener una distancia  $\delta$  de forma natural a partir de una similaridad  $s$ . En notación matricial,

$$\mathbf{D}^{(2)} = 2(\mathbf{1}\mathbf{1}' - \mathcal{S}),$$

donde  $\mathcal{S}$  es la matriz de similaridades y  $\mathbf{D}^{(2)}$  denota la matriz de cuadrados de distancias.

### Escalado Multidimensional

**Objetivo:** Obtener una representación euclídea, exacta o aproximada, de los elementos de un conjunto  $\mathcal{E}$  de  $n$  objetos o individuos, a partir de una matriz de distancias  $\mathbf{D}$  sobre  $\mathcal{E}$ .

**Atención:** No disponemos de una matriz de datos, sino de una matriz de distancias entre individuos. Y buscamos representar estos individuos en un plano.

**Representación euclídea exacta** en dimensión  $p \geq 0$  de  $(\mathcal{E}, \mathbf{D})$  es un conjunto de  $n$  puntos  $\mathbf{x}_1, \dots, \mathbf{x}_n$  del espacio euclídeo  $\mathbb{R}^p$ , que verifica que las distancias euclídeas entre los  $\mathbf{x}_i$  son iguales a los elementos correspondientes de la matriz  $\mathbf{D}$ , es decir:

$$\delta_{i,j}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j), \quad 1 \leq i, j \leq n.$$

donde  $\delta_{i,j}$  son los elementos de la matriz  $\mathbf{D}$ .

### ¿Cuándo una distancia es euclídea?

$\mathbf{D} = (\delta_{ij})_{1 \leq i, j \leq n}$  una matriz de distancias,  
 $\mathbf{1}$  el vector columna de unos de dimensión  $n$ ,  
 $\mathbf{I}$  la matriz identidad de dimensión  $n$ ,  
 $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$  la matriz de centrado.

**Teorema 1.** La matriz de distancias  $\mathbf{D}$  tiene una representación euclídea de dimensión  $p \leq n - 1$  si, y sólo si, la matriz

$$\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{D}^{(2)}\mathbf{H}$$

es semidefinida positiva con  $p = \text{rang}(\mathbf{B})$ , donde  $\mathbf{D}^{(2)}$  denota la matriz de cuadrados de distancias.

### Obtención de las coordenadas principales

Si  $\mathbf{B}$  es semidefinida positiva, entonces utilizando su descomposición espectral, podemos escribir:

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}' = \mathbf{X}\mathbf{X}',$$

siendo  $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}$  las *coordenadas principales*, donde  $\mathbf{\Lambda}$  es la matriz diagonal que contiene los autovalores de  $\mathbf{B}$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > \lambda_{p+1} = \dots = \lambda_n = 0,$$

y  $\mathbf{U}$  es una matriz  $n \times p$  ortogonal cuyas columnas son los autovectores de  $\mathbf{B}$ .

Las  $n$  filas de la matriz  $\mathbf{X}$  son las *coordenadas* de los individuos cuya matriz de distancias era  $\mathbf{D}$ . Las dos primeras columnas de  $\mathbf{X}$  dan lugar a una representación de los  $n$  individuos sobre un plano.

### Euclideanización de una distancia

Según el Teorema 1, si  $\mathbf{B}$  no es semidefinida positiva no existe una configuración euclídea para la matriz de distancias  $\mathbf{D}$ , es decir, no existe ninguna matriz  $\mathbf{X}$  cuyas filas sean las coordenadas de los individuos.

**Teorema 2.** Si  $\mathbf{B}$  tiene valores propios negativos, la transformación

$$\tilde{\delta}_{ij}^2 = \begin{cases} \delta_{ij}^2 + c, & i \neq j, \\ 0, & i = j, \end{cases} \quad (2)$$

donde  $c \geq 2|\lambda|$ ,  $\lambda$  es el valor propio negativo de módulo máximo, da lugar a una nueva matriz de distancias  $\tilde{\mathbf{D}}$  que admite una representación euclídea.

### Algoritmo de obtención

El Teorema 1 proporciona un algoritmo para la obtención de las coordenadas principales a partir de una matriz  $\mathbf{D}$  euclídea sobre los  $n$  individuos de un conjunto  $\mathcal{E}$ :

- Calcular la matriz de cuadrados de distancias  $\mathbf{D}^{(2)}$ .
- Construir la matriz  $\mathbf{B} = -\frac{1}{2} \mathbf{H} \mathbf{D}^{(2)} \mathbf{H}$ .
- Diagonalizar  $\mathbf{B} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}'$ .
- Las filas de  $\mathbf{X} = \mathbf{U} \mathbf{\Lambda}^{1/2}$  son las coordenadas principales (euclídeas) de los elementos del conjunto  $\mathcal{E}$ .

### Propiedades

- Las filas de  $\mathbf{X}$  verifican que  $\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$ .
- Las columnas de  $\mathbf{X}$ ,  $X_1, \dots, X_p$ , tienen media cero. Puesto que  $\mathbf{X} = \mathbf{U} \mathbf{\Lambda}^{1/2}$ , donde  $\mathbf{U}$  son los autovectores de  $\mathbf{B}$ , entonces

$$\bar{\mathbf{x}} = \frac{1}{n} \mathbf{1}' \mathbf{X} = \frac{1}{n} \mathbf{1}' \mathbf{U} \mathbf{\Lambda}^{1/2} = \mathbf{0},$$

al ser  $\mathbf{1}$  un autovector de  $\mathbf{B}$  de autovalor 0.

- Las variables  $X_j$ ,  $j = 1, \dots, p$ , son incorrelacionadas y sus varianzas son proporcionales a los autovalores de  $\mathbf{B}$ .

$$\text{Var}(\mathbf{X}) = \frac{1}{n} \mathbf{X}' \mathbf{X} = \frac{1}{n} \mathbf{\Lambda}^{1/2} \mathbf{U}' \mathbf{U} \mathbf{\Lambda}^{1/2} = \frac{1}{n} \mathbf{\Lambda}$$

- Sea  $\mathbf{X}$  una configuración exacta centrada de  $\mathbf{D}$ ,  $n \times p$ , donde  $p = \text{rang}(\mathbf{B})$ . Las columnas de  $\mathbf{X}$  pueden interpretarse como componentes principales.

La matriz de covarianzas de  $\mathbf{X}$  es  $\mathbf{S} = \mathbf{X}' \mathbf{X} / n$ . Puesto que  $\mathbf{X}' \mathbf{X}$  y  $\mathbf{B} = \mathbf{X} \mathbf{X}'$  tienen los mismos autovalores no nulos,

$$\mathbf{X}' \mathbf{X} = \mathbf{T} \mathbf{\Lambda} \mathbf{T}',$$

donde  $\mathbf{T}$  es la matriz ortogonal de autovectores de  $\mathbf{S}$ . La matriz de componentes principales de  $\mathbf{X}$  es

$$\mathbf{Y} = \mathbf{X} \mathbf{T}.$$

$\mathbf{Y}$  coincide con la solución de coordenadas principales de  $\mathbf{D}$ , excepto un posible factor  $\pm 1$  que afecta a las columnas, puesto que

$$\mathbf{Y} \mathbf{Y}' = \mathbf{X} \mathbf{T} \mathbf{T}' \mathbf{X}' = \mathbf{X} \mathbf{X}' = \mathbf{B}.$$

5. Si  $\mathbf{D}$  tiene una representación euclídea exacta  $\mathbf{X}$  de dimensión  $p = \text{rang}(\mathbf{B}) \leq n - 1$ , entonces para cada  $r < p$  puede obtenerse una representación euclídea aproximada  $\mathbf{X}(r)$  de dimensión  $r$ , tomando las  $r$  primeras columnas de  $\mathbf{X}$

$$\mathbf{X}(r) = (\mathbf{X}_1, \dots, \mathbf{X}_r).$$

La variabilidad total de  $\mathbf{X}(r)$  es  $(\lambda_1 + \dots + \lambda_r)/n$ , y el porcentaje de variabilidad explicado por  $\mathbf{X}(r)$  respecto de  $\mathbf{X}$  es

$$P_r = \frac{\lambda_1 + \dots + \lambda_r}{\lambda_1 + \dots + \lambda_p} \times 100.$$

### Ejemplo 3: Problemas 5.4 y 5.11

Sobre el conjunto de individuos  $\mathcal{E} = \{\text{león, girafa, vaca, oveja, gato, hombre}\}$  se han medido las siguientes variables binarias:

$X_1 = \text{¿tiene cola?}$ ,  $X_2 = \text{¿es salvaje?}$ ,  $X_3 = \text{¿tiene el cuello largo?}$ ,  
 $X_4 = \text{¿es animal de granja?}$ ,  $X_5 = \text{¿es carnívoro?}$ ,  $X_6 = \text{¿camina sobre cuatro patas?}$

La matriz de datos es

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{array}{l} \text{león} \\ \text{girafa} \\ \text{vaca} \\ \text{oveja} \\ \text{gato} \\ \text{hombre} \end{array}$$

Coefficiente de similaridad de Sokal y Michener

$$S = (\mathbf{a} + \mathbf{d})/p,$$

donde  $\mathbf{a} = \mathbf{X}\mathbf{X}'$ ,  $\mathbf{d} = (\mathbf{1}_n\mathbf{1}'_p - \mathbf{X})(\mathbf{1}_n\mathbf{1}'_p - \mathbf{X})'$ ,  $p = 6$  es el número de variables observadas y  $n = 6$  es el número de individuos.

La matriz de similaridades es

$$S = \begin{pmatrix} 1.00 & 0.67 & 0.50 & 0.50 & 0.83 & 0.50 \\ 0.67 & 1.00 & 0.50 & 0.50 & 0.50 & 0.17 \\ 0.50 & 0.50 & 1.00 & 1.00 & 0.67 & 0.33 \\ 0.50 & 0.50 & 1.00 & 1.00 & 0.67 & 0.33 \\ 0.83 & 0.50 & 0.67 & 0.67 & 1.00 & 0.67 \\ 0.50 & 0.17 & 0.33 & 0.33 & 0.67 & 1.00 \end{pmatrix}$$

Utilizando la transformación

$$\delta_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij},$$

que, en notación matricial es

$$\mathbf{D}^{(2)} = 2(\mathbf{1}_n\mathbf{1}'_n - S),$$

se obtiene la matriz de distancias (al cuadrado)

$$\mathbf{D}^{(2)} = \begin{pmatrix} 0 & 0.67 & 1.00 & 1.00 & 0.33 & 1.00 \\ 0.67 & 0 & 1.00 & 1.00 & 1.00 & 1.67 \\ 1.00 & 1.00 & 0 & 0 & 0.67 & 1.33 \\ 1.00 & 1.00 & 0 & 0 & 0.67 & 1.33 \\ 0.33 & 1.00 & 0.67 & 0.67 & 0 & 0.67 \\ 1.00 & 1.67 & 1.33 & 1.33 & 0.67 & 0 \end{pmatrix}$$

Los autovalores de  $\mathbf{B} = -\mathbf{HD}^{(2)}\mathbf{H}/2$  son:

1 0.7958 0.3333 0.0931 0.0000 0.0000

Existe una configuración euclídea de  $\mathbf{D}$  de dimensión 4. Las coordenadas principales son la filas de la matriz

$$\begin{pmatrix} 0.22361 & -0.35823 & 0.86603 & 1.9993 \\ -0.22361 & -0.61643 & -0.86603 & -0.77460 \\ -0.44721 & 0.30822 & 0 & 0.38730 \\ -0.44721 & 0.30822 & 0 & 0.38730 \\ 0.22361 & 0.050016 & 0.86603 & -0.23866 \\ 0.67082 & 0.30822 & -0.86603 & 0.38730 \end{pmatrix}$$

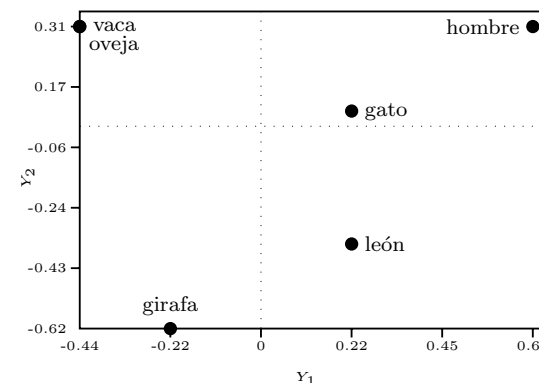
### Ejemplo 4: Problema 5.9

Con los datos del Ejemplo 2:

	Población	grupo A	grupo AB	grupo B	grupo O
1.	francesa	0.21	0.06	0.06	0.67
2.	checa	0.25	0.04	0.14	0.57
3.	germánica	0.22	0.06	0.08	0.64
4.	vasca	0.19	0.04	0.02	0.75
5.	china	0.18	0.00	0.15	0.67
6.	ainu	0.23	0.00	0.28	0.49
7.	esquimal	0.30	0.00	0.06	0.64
8.	negra USA	0.10	0.06	0.13	0.71
9.	española	0.27	0.04	0.06	0.63
10.	egipcia	0.21	0.05	0.20	0.54

obtener una representación en coordenadas principales utilizando la matriz de distancias de Bhattacharyya. Determina cuál es el porcentaje de variabilidad explicado por las dos primeras coordenadas principales.

### Representación de los individuos en dimensión 2



Porcentaje de variabilidad explicada:

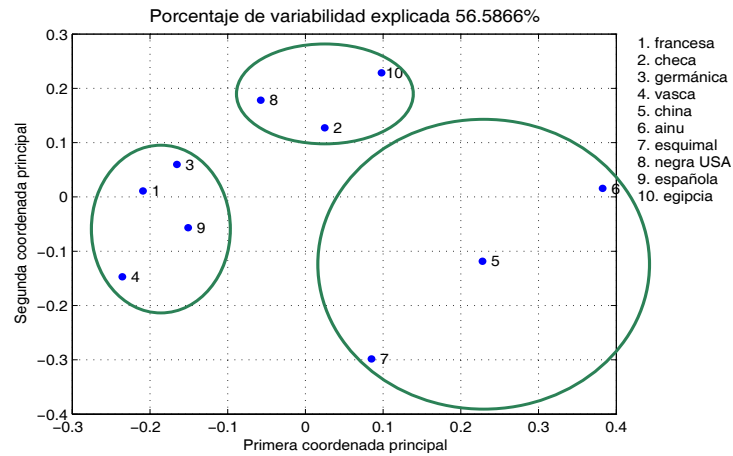
$$P_2 = \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^p \lambda_i} \times 100 = \frac{1.7958}{2.2222} \times 100 = 80.811\%$$

La matriz de cuadrados de distancias de Bhattacharyya es:

0	0.1567	0.0435	0.1246	0.2863	0.3966	0.2622	0.1850	0.0800	0.2204
0.1567	0	0.1156	0.2665	0.2240	0.2605	0.2476	0.2093	0.1364	0.0897
0.0435	0.1156	0	0.1660	0.2715	0.3636	0.2608	0.1769	0.0778	0.1787
0.1246	0.2665	0.1660	0.0000	0.3221	0.4732	0.2607	0.2555	0.1517	0.3359
0.2863	0.2240	0.2715	0.3221	0	0.1933	0.1896	0.2710	0.2653	0.2491
0.3966	0.2605	0.3636	0.4732	0.1933	0	0.3101	0.3701	0.3642	0.2422
0.2622	0.2476	0.2608	0.2607	0.1896	0.3101	0	0.3608	0.2024	0.3226
0.1850	0.2093	0.1769	0.2555	0.2710	0.3701	0.3608	0.0000	0.2438	0.1997
0.0800	0.1364	0.0778	0.1517	0.2653	0.3642	0.2024	0.2438	0	0.2211
0.2204	0.0897	0.1787	0.3359	0.2491	0.2422	0.3226	0.1997	0.2211	0

La función Matlab `[X,vaps,percent,acum] = coorp(D2)` realiza la representación en coordenadas principales de un conjunto de elementos cuya matriz de cuadrados distancias es D2.





### Ejemplo 5. Distancias por carretera entre algunas capitales españolas

	Barcelona	Madrid	San Sebastián	Sevilla	Valencia
Barcelona	0	639	606	1181	364
Madrid	639	0	474	542	355
San Sebastián	606	474	0	908	597
Sevilla	1181	542	908	0	679
Valencia	364	350	597	679	0

Construir un mapa según las distancias por carretera utilizando el análisis de coordenadas principales. ¿Coincide con el mapa geográfico?

Llamamos  $D$  a la matriz que contiene las distancias por carretera y construimos la matriz de cuadrados de distancias  $D_2 = D.^2$

0	408321	367236	1394761	132496
408321	0	224676	293764	126025
367236	224676	0	824464	356409
1394761	293764	824464	0	461041
132496	122500	356409	461041	0

La matriz de centrado es  $H = \text{eye}(5) - \text{ones}(5)/5$ . Construimos la matriz  $B$  y comprobamos si es semidefinida positiva:

```
B=-1/2*H*D2*H;
sort(eig(B))=
-66467.1272    0    24552.7156    215462.7239    743937.7877
```

Directamente, no pueden obtenerse las coordenadas principales de  $D$ .

Euclidianización de  $D$ :

```
lambda=min(eig(B));
D2_E=D2+2*abs(lambda)*ones(5)-2*abs(lambda)*eye(5);
```

y recalculamos la matriz  $B$ :

```
B=-1/2*H*D2_E*H;
sort(eig(B))=
810404.9149    281929.8511    91019.8428    0    0
```

Por tanto, existe una configuración euclídea de  $D_E$  en dimensión 3.

Las coordenadas principales se obtienen:

$$[U,D]=\text{eig}(B);$$

$$X=U*D.^{(1/2)};$$

Las dos primeras columnas de  $X$  son las coordenadas en el plano de las 5 ciudades:

551.85	-159.71	Barcelona
-114.51	041.24	Madrid
176.19	429.19	San Sebastian
-676.56	-49.91	Sevilla
63.03	-260.81	Valencia

Porcentaje de variabilidad explicada 92.31%

