

Análisis Multivariante

Tema 5. Distancias estadísticas

uc3m | Universidad Carlos III de Madrid

Aurea Grané
Dpto. Estadística
aurea.grane@uc3m.es

Distancias estadísticas

1. Introducción
2. Distancias estadísticas
3. Definición de distancia
4. Datos cuantitativos
5. Datos cualitativos y similaridades
6. Datos de tipo mixto

1. Introducción

El análisis multivariante es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar un número $p > 1$ de variables estadísticas sobre una muestra de n individuos.

Las variables observables son **homogéneas** y **correlacionadas**, sin que ninguna predomine sobre las demás.

Generalmente la información multivariante es una **matriz de datos**. Aunque, a menudo, también puede ser una **matriz de distancias (o similaridades)**, que miden el grado de discrepancia (o similitud) entre los individuos.

Notación habitual

Supondremos que hemos observado p variables en un conjunto de n elementos o individuos. Cada una de estas p variables es una variable **univariante** y el conjunto de las p variables forma una **variable multivariante**.

La matriz de datos \mathbf{X} es la representación de estas p variables medidas en los n individuos:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

La matriz \mathbf{X} puede representarse de dos formas distintas: por filas y por columnas.

Representación por filas:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix}$$

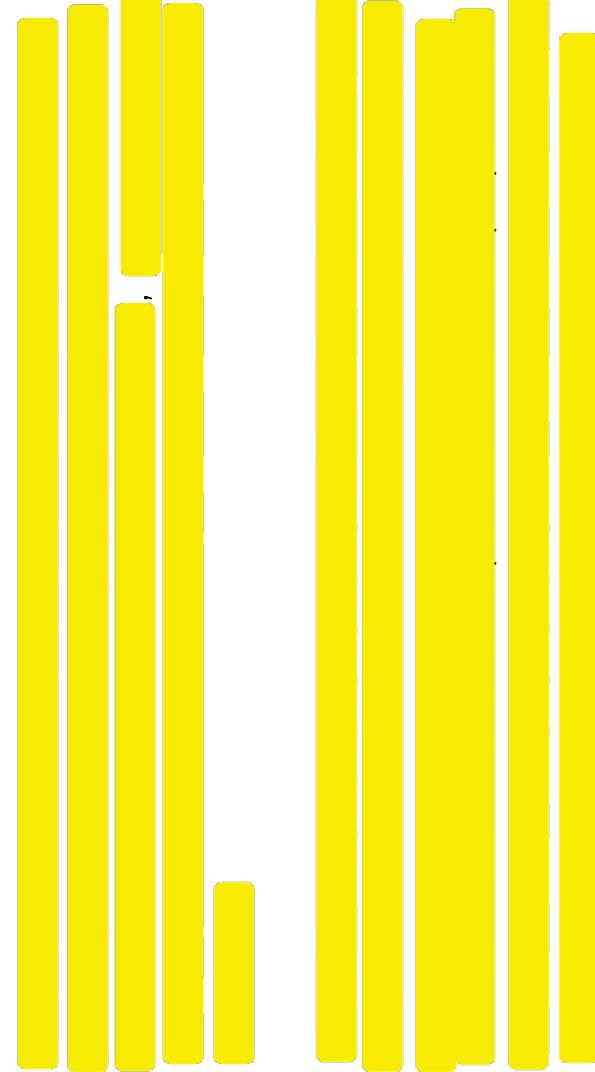
donde $\mathbf{x}'_i \in \mathbb{R}^p$ ($i = 1, \dots, n$) representa los valores observados para el individuo i -ésimo en las p variables.

Representación por columnas:

$$\mathbf{X} = (X_1, X_2, \dots, X_p),$$

donde $X_j \in \mathbb{R}^n$ ($j = 1, \dots, p$) representa la variable univariante j -ésima medida sobre todos los individuos de la muestra.

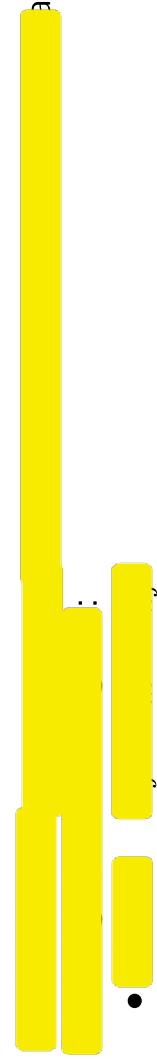
2. Distancias estadísticas



Definición de distancia

Otra medida de proximidad es la **disimilaridad**, que se define como:

- $\forall i, j, d_{ij} \geq 0$
- $\forall i, j, d_{ii} = 0$
- $\forall i, j, d_{ij} = d_{ji}$



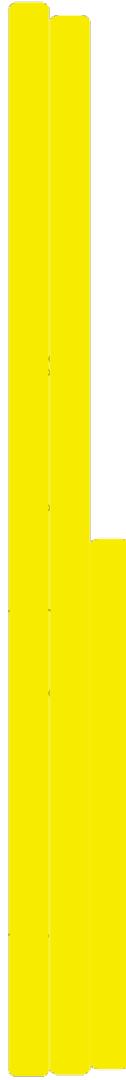
Finalmente, se define:

- $\forall i, j, d_{ij} = 1 - \rho_{ij}$

En general,

$$\begin{aligned} d_{ij} &= \sqrt{(0 - \delta_{21})^2 + (\delta_{12} - 0)^2 + \dots + (\delta_{1n} - \delta_{2n})^2}, \quad \text{con } \delta_{ij} = \delta_{ji}. \\ d_{ij} &= \sqrt{\delta_{21}^2 + \delta_{12}^2 + \dots + \delta_{2n}^2}, \quad \text{con } \delta_{ij} = \delta_{ji}, \\ d_{ij} &= \sqrt{\delta_{21}^2 + \delta_{n1}^2 + \dots + \delta_{n2}^2}, \quad \text{con } \delta_{ij} = \delta_{ji}, \end{aligned}$$

3.



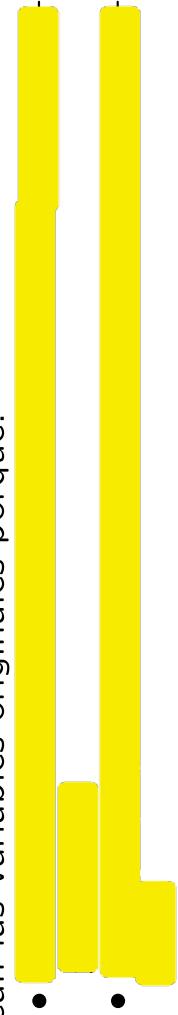
La **l_p** y la familia ℓ^q

Una de las distancias más conocidas y utilizadas es la **ℓ_p**

$$\|x - y\|_p = \left(\sum_{k=1}^n |x_k - y_k|^p \right)^{1/p}$$

← si $p > 1$, esta elevada al cuadrado

Sin embargo, su uso **no es recomendable** cuando las X_j sean las variables originales porque:



Consideremos el **ℓ^α** donde $\alpha \in \mathbb{R}$, $\alpha \neq 1$. Ahora las puntuaciones de los individuos i y j son $y_i = \alpha x_i$ e $y_j = \alpha x_j$, y la distancia euclídea es

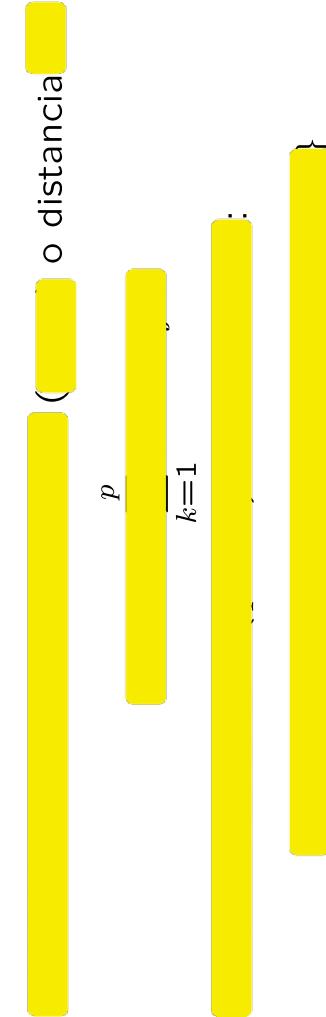
$$\|y_i - y_j\| = \|(y_i - y_j)'(y_i - y_j)\|^{1/2} = \|(\alpha x_i - \alpha x_j)'(\alpha x_i - \alpha x_j)\|^{1/2} = \|\alpha(x_i - x_j)\|$$

Observación: Los paquetes estadísticos suelen usar, por defecto, la distancia euclídea en aquellos métodos que necesitan del uso de una distancia. Quizás antes de utilizarlos, convendría reflexionar si, en el caso que nos ocupe, el uso de la distancia euclídea está justificado. Sin ir más lejos, esto ocurre en una de las herramientas más usadas en *classification rating* como es el algoritmo de *k*-medias ...

La L_p o de Minkowski es:

$$d(x, y) = \left(\sum_{k=1}^p |x_k - y_k|^q \right)^{1/q}, \quad q > 0.$$

además, veremos este concepto más adelante). Minkowski son:



Distancias invariantes frente a cambios de escala

$$d(x, y) = \sqrt[p]{\sum_{k=1}^p |x_k - y_k|^p} \quad (\text{euclídea}):$$

$$d(x, y) = \sqrt[p]{\sum_{k=1}^p \frac{|x_k - y_k|^p}{s_k^p}} \quad \text{donde}$$

Esta expresión equivale a reescalar cada variable en unidades de desviación típica. Pero

uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)

donde **es la**

porque

• *→ no singulares es que la transformación lineal no sea multiplicar por cero.*

- y
- Por ejemplo,
- , sino que
-
-
-
-

uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)

Cálculo de distancias con Matlab

La función pdist(X) devuelve la distancia euclídea entre las filas de la matriz X. Cuidado, esta función nos devuelve la distancia, no el cuadrado de la distancia.

La opción pdist(X,'distance') permite además obtener otros tipos de distancias según el argumento que se indique como 'distance'. Entre las opciones que admite están 'seuclidean', ('distancia de Pearson'), 'mahalanobis', 'cityblock', ('distancia ℓ^1 '), 'minkowski' ('distancia ℓ^q ', donde hay que indicar el valor de q).

En realidad, la función pdist devuelve un vector fila con la parte triangular superior de la matriz de distancias. Si se quiere ver propiamente la matriz de distancias, hay que utilizar la función squareform.

Ejercicio 1 Escribir una función matlab que, dada una matriz de datos X , $n \times p$, devuelva la matriz de cuadrados de distancias de Mahalanobis (sin usar la función pdist).

```
% La funcion D=maha(X) calcula una matriz de cuadrados de
% distancias. El elemento (i,j) de la matriz D contiene el
% cuadrado de la distancia de Mahalanobis entre la fila "i"
% y la fila "j" de la matriz X.
%
% Entradas: una matriz X de dimension nxp.
% Salidas: una matriz D de dimension nxn.
%
function D=maha(X)
[n,p]=size(X);
% calculo del vector de medias y de la matriz de covarianzas
% de X:
S=cov(X,1);
% calculo de las distancias de Mahalanobis (al cuadrado):
D=zeros(n);
invS=inv(S);
for i=1:n
    for j=i+1:n
        D(i,j)=(X(i,:)-X(j,:))*invS*(X(i,:)-X(j,:))';
    end
    D=D+D';
end
```

Ejemplo 1 Para ver cómo difieren las distancias ℓ^2 o euclídea, ℓ^1 , ℓ^4 , Mahalanobis y Pearson, generamos una muestra de tamaño $n = 10$ a partir de una ley normal multivariante con vector de medias μ y matriz de covarianzas Σ :

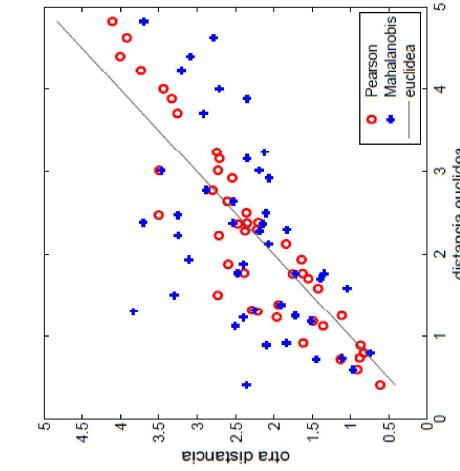
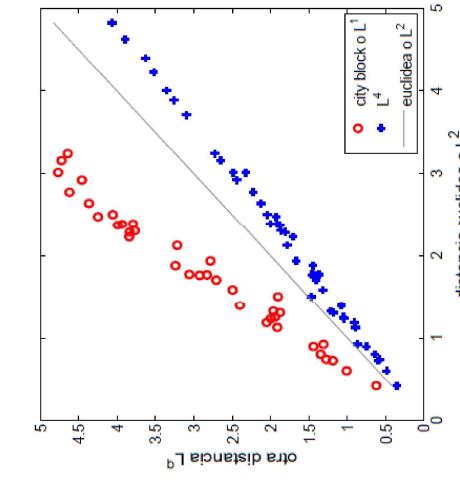
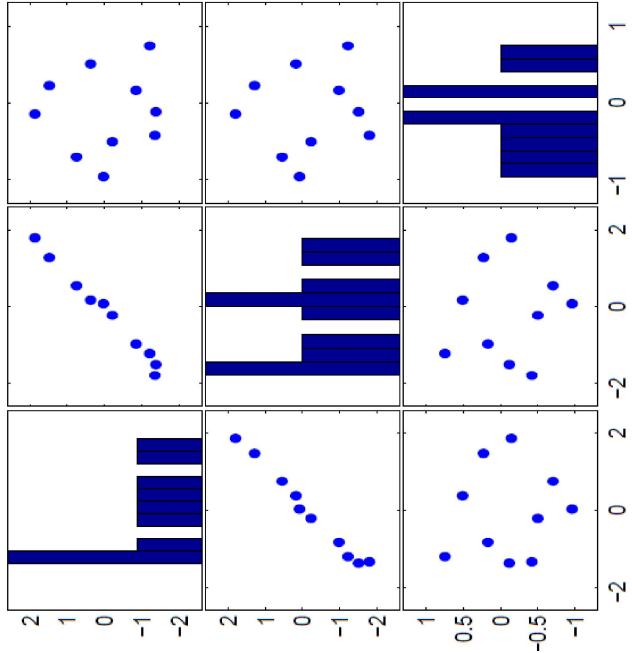
$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.98 & 0 \\ 0.98 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Observad que las variables X_1 y X_2 están altamente correladas, mientras que X_3 está incorrelada con X_1 y X_2 .

Para generar los datos utilizamos la función `[m,S,R,X]=nmult2(mu,Sigma,n)` que, además de los datos devuelve una estimación del vector de medias y las matrices de covarianzas y de correlaciones.

```
[m,S,R,X]=nmult2(mu,Sigma,10)
X =
   -1.3745    -1.5027   -0.1170
   -0.8393    -0.9752   0.1685
   -0.2086    -0.2258   -0.5012
    0.7559     0.5464   -0.7051
    0.3757     0.1764   0.5082
   -1.3454    -1.7920   -0.4209
    1.4819     1.2854   0.2291
    0.0327     0.0833   -0.9595
    1.8705     1.7965   -0.1460
   -1.2090    -1.2182   0.7445
Yeuclid=pdist(X);
Ycity= pdist(X,'cityblock');
Ymink= pdist(X,'minkowski',4);
Ymaha= pdist(X,'mahal');
Ypearson= pdist(X,'seuclidean');
```

Plotmatrix(X)



Generamos una nueva muestra de datos a partir de nuevos valores para μ y Σ y mezclamos ambas muestras en una misma matriz de datos X

```

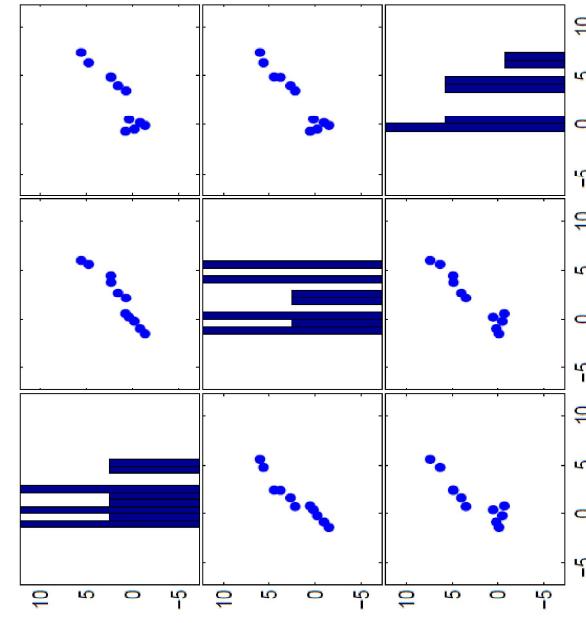
mu1' =
  0          0          0
  1.00000   0.98000   0
  0.98000   1.00000   0
  0          0          1.00000

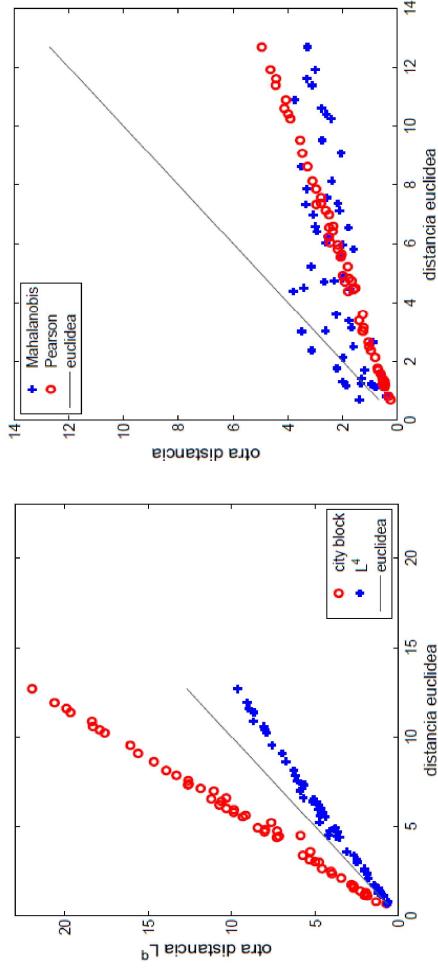
mu2' =
  2          3          4
  1.00000   0.50000   0.7000
  0.50000   1.00000   0.8000
  0.70000   0.80000   1.00000

X =
 -1.3745   -1.5027   -0.1170
 -0.8393   -0.9752   0.1685
 -0.2086   -0.2258   -0.5012
 0.7559    0.5464   -0.7051
 0.3757    0.1764   0.5082
 2.3188    3.7784   4.8340
 0.6923    2.1687   3.4850
 1.5664    2.6757   3.9881
 2.3426    4.4614   4.8653
 5.5784    6.0095   7.3810
 4.7694    5.6120   6.2893

```

plotmatrix(X)



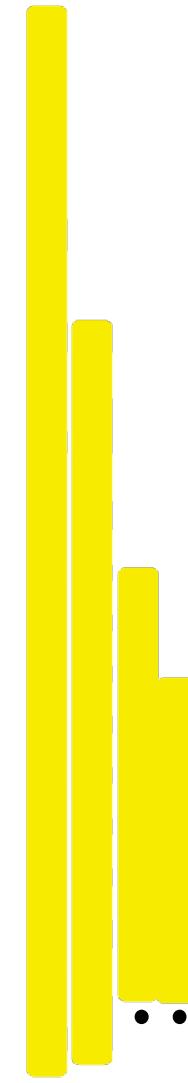
**uc3m**

Grado en Estadística y Empresa

Aurea Grané (Estadística)

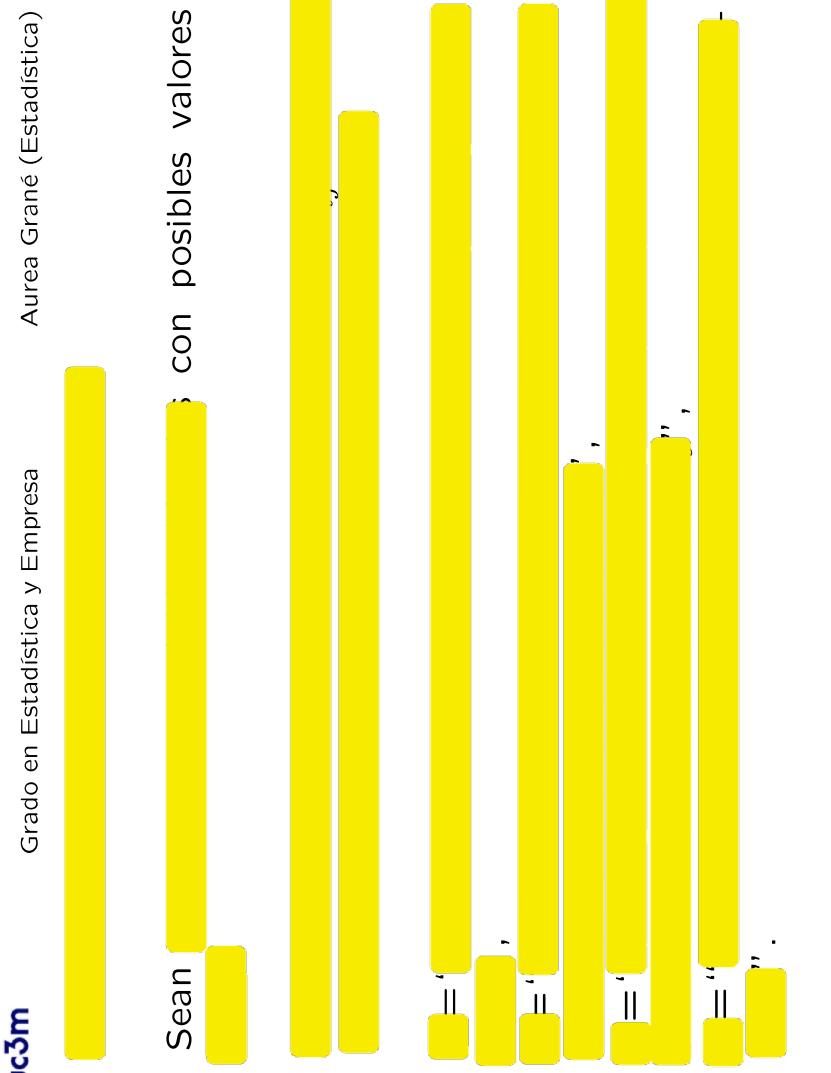
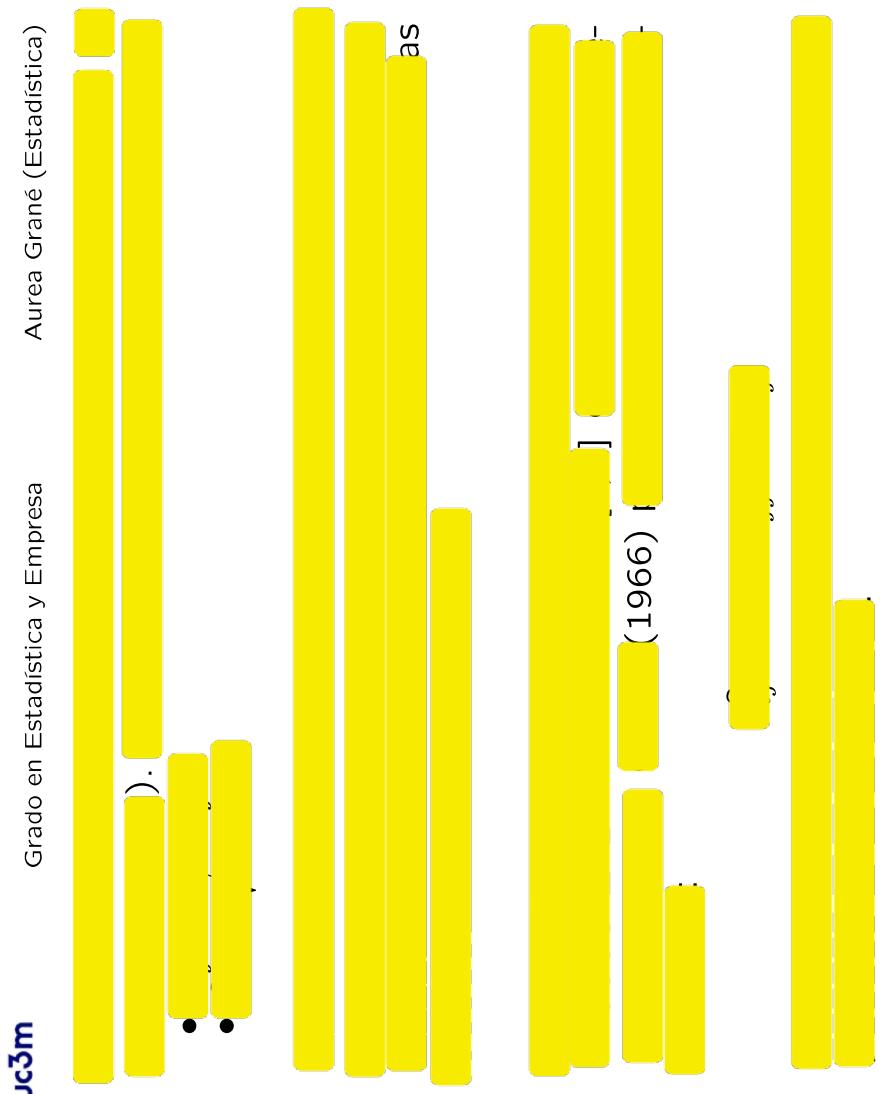
5.

En muchas aplicaciones resultará más cómodo trabajar con **distancias** en lugar de *distancias*.



Y en este caso, tendremos una

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}, \text{ con}$$



Algunos  son:

: 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
2	3	2	0	1
3	2	3	2	2

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

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

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

(tema 6)

...  Propiedad Euclídea

Coeficiente	Valores	Prop. euclídea
$\frac{a}{b+c}$	$(0, \infty)$	
$\frac{a}{a+b+c+d}$	$(0, 1)$	sí
$\frac{a}{a+b+c+d}$	$(0, 1)$	sí
$\frac{a}{a+b+c+d}$	$(0, 1)$	sí
$\frac{a+2(b+c)}{a+d}$	$(0, 1)$	sí
$\frac{a+2(b+c)+d}{a+d}$	$(0, 1)$	sí
$\frac{a+0.5(b+c)}{a+d}$	$(0, 1)$	sí
$\frac{a+0.5(b+c)+d}{a+d}$	$(0, 1)$	no
...		

Jaccard \rightarrow Sokal \rightarrow Sneath-Sokal \rightarrow

Coeficiente Valores Prop. euclídea

\dots	$\frac{a+d-(b+c)}{a+b+c+d}$	$(-1, 1)$	sí
Kulczyński \rightarrow	$\frac{1}{2} \left(\frac{a}{a+b} + \frac{a+c}{a+c} \right)$	$(0, 1)$	no
	$\frac{1}{2} \left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d} \right)$	$(0, 1)$	no
	$\frac{\sqrt{(a+b)(a+c)}}{\sqrt{(a+b)(a+c)}}$	$(0, 1)$	sí
	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	$(0, 1)$	sí
	$\frac{ad-bc}{\sqrt{ad+bc}}$	$(-1, 1)$	sí
	$\frac{ad-bc}{ad+bc}$	$(-1, 1)$	no

Ejercicio 2 Dada una matriz de datos X , $n \times p$,

- Escribir una función matlab que devuelva la matriz de similaridades según el coeficiente de Sokal y Michener (S_4).
- Escribir una función matlab que devuelva la matriz de similaridades según el coeficiente de Jaccard (S_3).

% SOKAL

% Dada una matriz de datos binarios X (n,p), la función
% $S=sokal(X)$ devuelve la matriz de similaridades, segun el
% coeficiente de similaridad de Sokal y Michener, entre los
% n individuos.

```
function S=sokal(X)
[n,p]=size(X);
J=ones(n,p);
a=X*X';
d=(J-X)*(J-X)';
S=(a+d)/p;
```

Para obtener la matriz de cuadrados de distancias, haremos $D^2=2*(ones(n)-S)$

```
% JACCARD
%
% Dada una matriz de datos binarios X (n,p), la función
% S=jaccard(X) devuelve la matriz de similaridades, segun
% el coeficiente de similaridad de Jaccard, entre los n
% individuos.
%
function S=jaccard(X)
[n,p]=size(X);
J=ones(n,p);
a=X*X';
d=(J-X)*(J-X)';
%
[i0,j0]=find(d==p);
for i=1:length(i0)
    d(i0(i),j0(i))=p-1;
end
S=a./(p*ones(n)-d);
```

Para obtener la matriz de cuadrados de distancias, haremos $D^2 = 2 * (\text{ones}(n) - S)$

uc3m

Ejemplo 2 Analizar cómo difieren las distancias obtenidas mediante las similaridades de Sokal-Michener y Jaccard respecto de la distancia euclídea para la siguiente matriz de datos binarios:

$$X = \begin{matrix} & & 1 & 1 & 1 \\ & & 1 & 0 & 0 \\ & & 0 & 1 & 0 \\ & & 0 & 0 & 1 \\ & & 1 & 1 & 0 \\ & & 0 & 1 & 1 \\ & & 1 & 0 & 1 \\ & & 0 & 0 & 0 \end{matrix}$$

Para obtener las distancias, usar la transformación $\delta_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$ y, por tanto, $\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$.

Observad que todos los individuos son distintos. Para obtener la distancia euclídea, podemos hacer `Deuclid=squareform(pdist(X))`.

uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)

```
Deuclid= 0 1.4142 1.4142 1.4142 1.0000 1.0000 1.0000 1.7321  
1.4142 0 1.4142 1.4142 1.0000 1.7321 1.0000 1.0000  
1.4142 1.4142 0 1.4142 1.0000 1.7321 1.0000 1.0000  
1.4142 1.4142 1.4142 0 1.0000 1.4142 1.4142 1.4142  
1.0000 1.0000 1.0000 1.7321 0 1.4142 1.4142 1.4142  
1.0000 1.7321 1.0000 1.0000 1.4142 0 1.4142 1.4142  
1.0000 1.0000 1.7321 1.0000 1.4142 1.4142 0 1.4142  
1.7321 1.0000 1.0000 1.4142 1.4142 1.4142 0 0
```

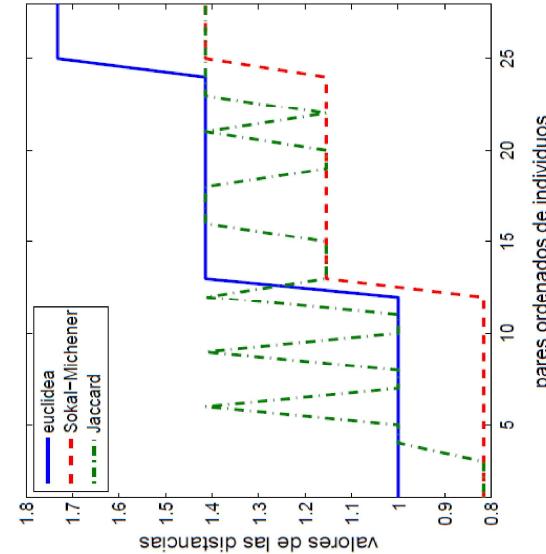
```
Dsokal= 0 1.1547 1.1547 1.1547 0.8165 0.8165 0.8165 1.4142  
1.1547 0 1.1547 1.1547 0.8165 1.4142 0.8165 0.8165  
1.1547 1.1547 0 1.1547 0.8165 0.8165 1.4142 0.8165  
1.1547 1.1547 1.1547 0 1.4142 0.8165 0.8165 0.8165  
0.8165 0.8165 0.8165 1.4142 0 1.1547 1.1547 1.1547  
0.8165 1.4142 0.8165 0.8165 1.1547 0 1.1547 1.1547  
0.8165 0.8165 1.4142 0.8165 1.1547 1.1547 0 1.1547  
1.4142 0.8165 0.8165 0.8165 1.1547 1.1547 1.1547 0
```

```
Djaccard= 0 1.1547 1.1547 1.1547 0.8165 0.8165 0.8165 1.4142  
1.1547 0 1.4142 1.4142 1.0000 1.4142 1.0000 1.4142  
1.1547 1.4142 0 1.4142 1.0000 1.4142 1.0000 1.4142  
1.1547 1.4142 1.4142 0 1.0000 1.1547 1.1547 1.1547  
0.8165 1.0000 1.0000 1.4142 0 1.1547 1.1547 1.1547  
0.8165 1.4142 1.0000 1.4142 1.1547 1.1547 0 1.1547  
1.4142 1.4142 1.4142 1.4142 1.4142 1.4142 0 0
```

uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)



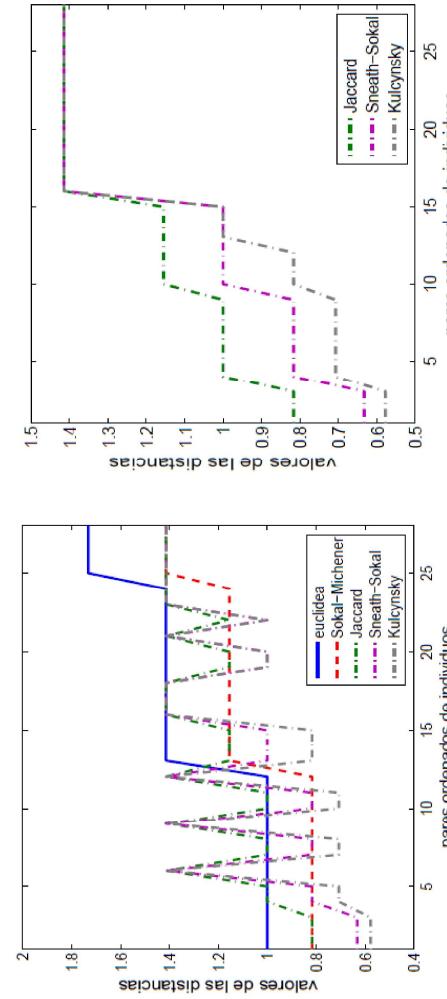
Ejemplo 3 Obtener las distancias asociadas a los coeficientes de similaridad de Sneath-Sokal (S_7)

$$s_{ij} = \frac{a}{a + \frac{1}{2}(b + c)}$$

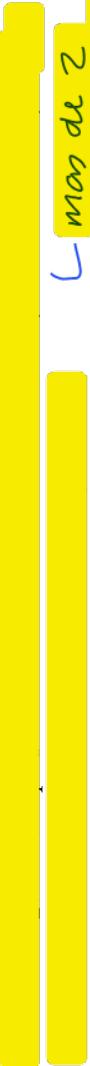
y Kulcynsky (S_{10})

$$s_{ij} = \frac{1}{2} \left(\frac{a}{a + b} + \frac{a}{a + c} \right)$$

y compararlas con las del Ejemplo 2.



Similaridades con variables categóricas

L mas de 2 categorias

Supongamos que al estudiar la similitud entre los individuos i, j, α es el número de coincidencias para las p variables cualitativas, por lo que $p - \alpha$ serán las no coincidencias.

La similaridad más habitual es considerar el **coeficiente de coincidencias**:

$$s_{ij} = \frac{\alpha}{p}.$$

Observad que cuando este coeficiente se aplica sobre variables binarias coincide con el coeficiente de Sokal-Michener (S_4), puesto que $\alpha = a + d$.

$$\begin{matrix} X_1 & X_2 \\ 1 & 4 \\ 2 & 2 \\ 3 & 1 \end{matrix} \quad \begin{matrix} \alpha_{12} = 1 \\ \alpha_{13} = 0 \\ \alpha_{23} = 0 \end{matrix}$$

$$S_{12} = \frac{1}{3}, \quad S_{23} = \frac{0}{3} = 0$$

Coeficientes de similaridad para variables categóricas

$p - \alpha = n - \text{no coincidencias}$

Cuando se aplica sobre binarias ...

Coeficiente	Valores	
SC_1	$(0, 1)$	$= S_4$
SC_2	$(0, \infty)$	
SC_3	$(-1, 1)$	$= S_9$
SC_4	$(0, 1)$	$= S_6$

Ejercicio 3 Dada una matriz de datos X , $n \times p$,

- Escribir una función matlab que devuelva la matriz de similaridades según el coeficiente de coincidencias (SC_1).
- Escribir una función matlab que devuelva el coeficiente SC_4 .

```
% COINCIDENCIAS
%
% Dada una matriz de datos categoricos X (n,p), la funcion
% S=coincidencias(X) calcula la matriz de similaridades, segun
% el coeficiente de similaridad de coincidencias entre los n individuos.
%
function S=coincidencias(X)
[n,p]=size(X);
S=p*eye(n);
for i=1:n
    for j=i+1:n
        S(i,j)=sum(X(i,:)==X(j,:));
        S(j,i)=S(i,j);
    end
end
S=S/p;
```

Para obtener la matriz de cuadrados de distancias, haremos $D2=2*(ones(n)-S)$

```
% COINCIDENCIAS4
%
% Dada una matriz de datos categoricos X (n,p), la funcion
% S=coincidencias4(X) calcula la matriz de similaridades, segun
% el coeficiente de similaridad s(i,j)=alpha/(alpha+2*(p-alpha)).
%
function S=coincidencias4(X)
[n,p]=size(X);
for i=1:n
    alpha(i,i)=p;
    for j=i+1:n
        alpha(i,j)=sum(X(i,:)==X(j,:));
        alpha(j,i)=alpha(i,j);
    end
end
S=alpha./(alpha+2*(p*ones(n)-alpha));
```

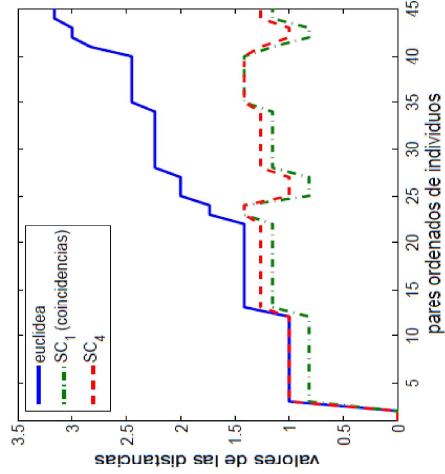
Para obtener la matriz de cuadrados de distancias, haremos $D^2 = 2 * (\text{ones}(n) - S)$

Ejemplo 4 Dada la matriz de datos categóricos:

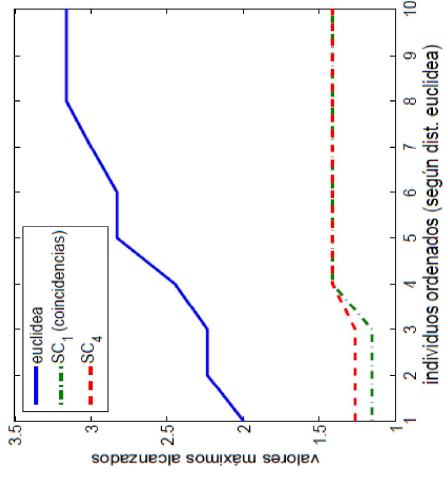
$$X = \begin{matrix} & & 1 & 1 & 3 \\ & & 1 & 1 & 2 \\ & & 3 & 1 & 2 \\ & & 1 & 0 & 3 \\ & & 2 & 1 & 3 \\ & & 0 & 1 & 2 \\ & & 1 & 1 & 2 \\ & & 0 & 1 & 3 \\ & & 3 & 1 & 2 \\ & & 1 & 1 & 3 \\ & & 1 & 2 & 1 \end{matrix}$$

obtener las distancias asociadas a los coeficientes de similaridad SC_1 y SC_4 y compararlas con la distancia euclídea.

Para obtener las distancias, usar la transformación $\delta_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$ y por tanto, $\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$.



Analiso para interpretación ↗



uc3m

6. Datos de tipo mixto

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 el coeficiente de similaridad de Gower (1971) como

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3}, \quad 0 \leq s_{ij} \leq 1, \quad (1)$$

donde p_1 es el número de variables cuantitativas, a y d son, respect., el número de coincidencias (1, 1) y (0, 0) para las p_2 variables binarias, α es el número de coincidencias en las p_3 variables cualitativas (no binarias) y G_h es el rango (o recorrido) de la h -ésima variable cuantitativa. $\max(x_h) - \min(x_h)$

A partir de s_{ij} se obtiene la distancia de Gower como $d_{ij}^2 = 1 - s_{ij}$.

Ejercicio 4 Escribir una función matlab que, dada una matriz de datos X , $n \times p$, devuelva la matriz de distancias según el coeficiente de similaridad de Gower.

uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)

```
% La funcion S=gower2(X,p1,p2,p3) calcula una matriz de similitudes,  
% segun el coeficiente de similaridad de Gower.  
  
% Entradas:  
% X matriz de datos mixtos, cuyas columnas deben estar ordenadas  
% de la forma: continuas, binarias, categoricas,  
% p1 numero de variables continuas,  
% p2 numero de variables binarias,  
% p3 numero de variables categoricas (no binarias),  
  
function S=gower2(X,p1,p2,p3)  
[n,p]=size(X);  
% matriz de variables cuantitativas  
X1=X(:,1:p1);  
% matriz de variables binarias  
X2=X(:,p1+1:p1+p2);  
% matriz de variables categoricas  
X3=X(:,p1+p2+1:p);  
  
% calculos para las variables continuas  
rango=max(X1)-min(X1);  
for i=1:n  
    c(i,i)=p1;
```

```

for j=1:i-1
    c(i,j)=p1-sum(abs(X1(i,:)-X1(j,:))./rango);
    c(j,i)=c(i,j);
end

% calculo de las matrices a y d para las variables binarias
J=ones(size(X2));
a=X2*X2';
d=(J-X2)*(J-X2)';
% calculos de la matriz alpha para las variables categoricas
for i=1:n
    for j=i:n
        alpha(i,j)=sum(X3(i,:)==X3(j,:));
        alpha(j,i)=alpha(i,j);
    end
    % calculo del coeficiente de similaridad de Gower
    S=(c+a+alpha)./(p*ones(n)-d);
end

```

uc3m

Grado en Estadística y Empresa

Aurea Grané (Estadística)

Ejemplo 5 El fichero jugadores2007.txt contiene siete variables observadas sobre 50 jugadores de la liga española de fútbol 2006/07:

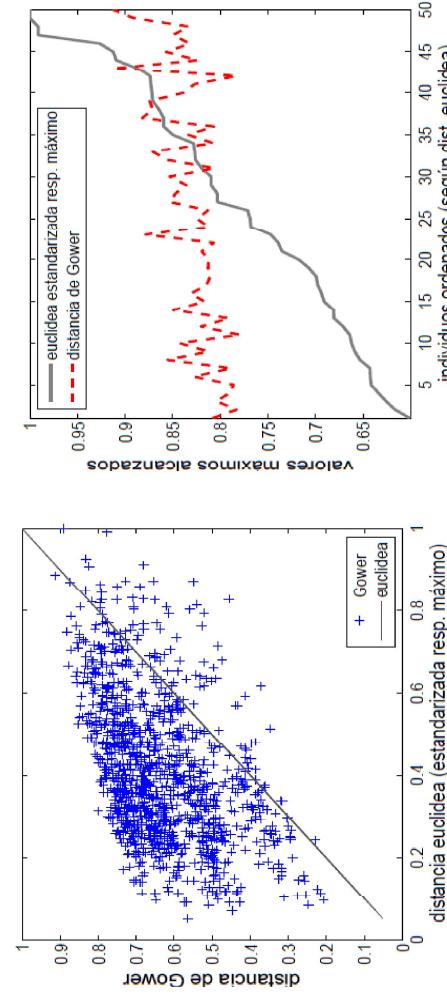
Jugador	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1. Ronaldinho	15	26	1.78	71	1	2	2
2. Etoo	21	25	1.8	75	0	3	2
3. Xavi	6	26	1.7	68	0	5	4
4. Messi	7	19	1.69	67	0	1	3
5. Puyol	1	28	1.78	78	0	5	3
6. Raúl	7	29	1.8	73.5	1	5	3
7. Ronaldo	18	30	1.83	82	0	2	1
8. Beckham	4	31	1.8	67	0	9	3
50. Dóbłas	0	25	1.84	78	0	5	3

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

Calcular las distancias de Gower entre estos 50 jugadores y compararlas con la distancia euclídea.

Llamamos X a la matriz de datos que contiene la información del fichero jugadores2007.txt.

```
% distancia de Gower
p1=4; p2=1; p3=2;
S=gower2(X,p1,p2,p3);
n=size(X,1);
Dgower=sqrt(ones(n)-S);
Ygower=squareform(Dgower); % toma valores en (0,1)
% distancia euclídea
Y=pdist(X);
m=max(Y);
Y=Y/m; % para poder comparar con Ygower
```



Observemos detenidamente el coeficiente de Gower

El coeficiente de similaridad Gower es la suma de diferentes coeficientes apropiados para cada tipo de variables. Por ejemplo:

- Si sólo disponemos de variables cuantitativas, la distancia que se obtiene es

$$\delta_{ij}^2 = \frac{1}{p} \sum_{k=1}^p \frac{|x_{ik} - x_{jk}|}{G_k}$$

*← distancia L₁
entremariza por el paro*

- Si solamente se dispone de variables binarias, el coeficiente de similaridad de Gower coincide con el de Jaccard.
- Si solamente se tienen variables cualitativas (no binarias), el coeficiente de similaridad de Gower coincide con el coeficiente de coincidencias.

Con esta idea se pueden construir fácilmente otros coeficientes de similaridad para datos de tipo mixto. Algunas recomendaciones para ello son:

- Si se quiere que el coeficiente resultante tenga la propiedad euclídea, todos los coeficientes que se combinen deben cumplir dicha propiedad por sí solos.
- Para variables cualitativas, deben utilizarse coeficientes que dividan cada comparación por un factor de normalización antes de sumar.
- Para variables binarias y cualitativas, serán preferibles aquellos coeficientes que tomen valores en [0, 1] para evitar rescalar las similaridades antes de sumar.

Ejercicio 5 Escoger una medida de distancias para variables cuantitativas (distinta de la distancia euclídea) y dos coeficientes de similaridad para variables binarias y cualitativas y construir un nuevo coeficiente de similaridad para datos mixtos, siguiendo la idea del coeficiente de similaridad de Gower. Escribir una función matlab que, dada una matriz de datos \mathbf{X} , $n \times p$, devuelva la matriz de distancias según este nuevo coeficiente.

Referencias

- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 325–338.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27, 857–874.