Aprendizaje No Supervisado

Maestría en Ciencia de Datos

Lucas Fernández Piana

Primavera 2022

Universidad de San Andrés

Multidimensional Scaling

INTRO

Recordemos que los procedimientos de reducción de dimensión tienen como objetivo encontrar una representación de los datos en un espacio de dimensión menor.

INTRO

Supongamos que tenemos una matriz de distancias o disimilaridades $d = d_{ij}$ de tamaño $n \times n$ que se construye a partir de un conjunto de datos D.

INTRO

Supongamos que tenemos una matriz de distancias o disimilaridades $d = d_{ij}$ de tamaño $n \times n$ que se construye a partir de un conjunto de datos D.

El objetivo de Multidimensional Scaling (MDS) es contruir una representación en el plano o en el espacio que preserve la distancia o disimilaridad utilizando la distancia natural del plano (norma eucídea).

INTRO

El estudio de Ekman (1954) nos muestra un típico problema de Multidimensional Scaling.

Es un trabajo en la percepción del color en la visión humana:

INTRO

El estudio de Ekman (1954) nos muestra un típico problema de Multidimensional Scaling.

Es un trabajo en la percepción del color en la visión humana:

Se consideraron 14 colores que difieren solamente en el matiz con longitudes de onda que varian desde 434 μm a 674μm.

INTRO

El estudio de Ekman (1954) nos muestra un típico problema de Multidimensional Scaling.

Es un trabajo en la percepción del color en la visión humana:

Se consideraron 14 colores que difieren solamente en el matiz con longitudes de onda que varian desde 434 μm a 674μm.

Le pidieron a 31 personas que califiquen en una escala del 0 (no hay similaridad) al 4 (son idénticos) para cada par de las $\binom{14}{2}$ combinaciones posibles.

INTRO

Los autores obtienen un similaridad tomando el promedio de las calificaciones para cada par de colores.

INTRO

Los autores obtienen un similaridad tomando el promedio de las calificaciones para cada par de colores.

Para construir la disimilaridad primero escalaron y luego le restaron a 1 la entrada correspondiente en la similaridad escalada.

INTRO

Los autores obtienen un similaridad tomando el promedio de las calificaciones para cada par de colores.

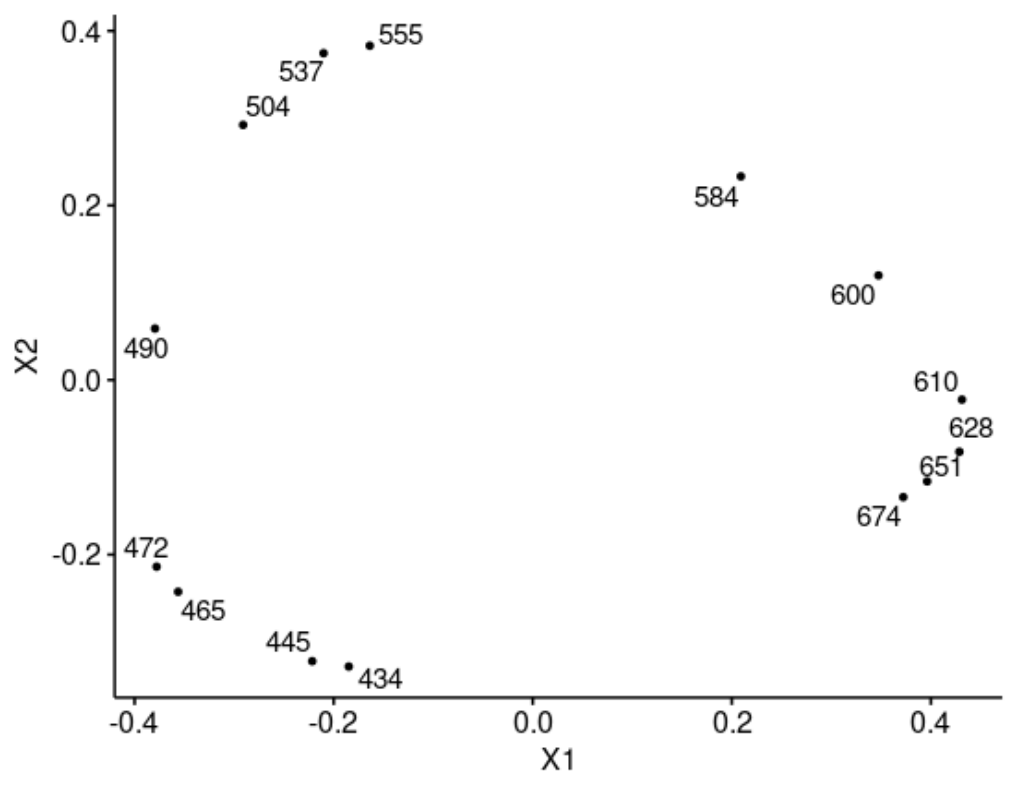
Para construir la disimilaridad primero escalaron y luego le restaron a 1 la entrada correspondiente en la similaridad escalada.

El resultado es una matriz simétrica de 14×14 que representa disimilaridad entre los pares de colores según la percepción de los participantes del estudio.

	434	445	465	472	490	504	537	555	584	600	610	628	651	674
434	0.00	0.86	0.42	0.42	0.18	0.06	0.07	0.04	0.02	0.07	0.09	0.12	0.13	0.16
445	0.86	0.00	0.50	0.44	0.22	0.09	0.07	0.07	0.02	0.04	0.07	0.11	0.13	0.14
465	0.42	0.50	0.00	0.81	0.47	0.17	0.10	0.08	0.02	0.01	0.02	0.01	0.05	0.03
472	0.42	0.44	0.81	0.00	0.54	0.25	0.10	0.09	0.02	0.01	0.00	0.01	0.02	0.04
490	0.18	0.22	0.47	0.54	0.00	0.61	0.31	0.26	0.07	0.02	0.02	0.01	0.02	0.00
504	0.06	0.09	0.17	0.25	0.61	0.00	0.62	0.45	0.14	0.08	0.02	0.02	0.02	0.01
537	0.07	0.07	0.10	0.10	0.31	0.62	0.00	0.73	0.22	0.14	0.05	0.02	0.02	0.00
555	0.04	0.07	0.08	0.09	0.26	0.45	0.73	0.00	0.33	0.19	0.04	0.03	0.02	0.02
584	0.02	0.02	0.02	0.02	0.07	0.14	0.22	0.33	0.00	0.58	0.37	0.27	0.20	0.23
600	0.07	0.04	0.01	0.01	0.02	0.08	0.14	0.19	0.58	0.00	0.74	0.50	0.41	0.28
610	0.09	0.07	0.02	0.00	0.02	0.02	0.05	0.04	0.37	0.74	0.00	0.76	0.62	0.55
628	0.12	0.11	0.01	0.01	0.01	0.02	0.02	0.03	0.27	0.50	0.76	0.00	0.85	0.68
651	0.13	0.13	0.05	0.02	0.02	0.02	0.02	0.02	0.20	0.41	0.62	0.85	0.00	0.76
674	0.16	0.14	0.03	0.04	0.00	0.01	0.00	0.02	0.23	0.28	0.55	0.68	0.76	0.00

INTRO

MDS nos permitirá encontrar una configuración para poder representar estas disimilaridades entre los colores en un plano.



INTRO


Espectro visible por el ojo humano (Luz)
Ultravioleta
Infrarrojo
400 nm
450 nm
500 nm
550 nm
600 nm
650 nm
700 nm
750 nm

INTRO

En general cuando nos referimos a MDS hablamos de una familia de algoritmos diseñados para lograr una configuración óptima en $\mathbb{R}^p$ para $p = 2$ o 3.

En general cuando nos referimos a MDS hablamos de una familia de algoritmos diseñados para lograr una configuración óptima en $\mathbb{R}^p$ para $p = 2$ o 3.

Tenemos tres clases algoritmos en esta familia:

- MDS clásico.
- MDS métrico.
- MDS no métrico.

MDS trata de encontrar $x_1, \ldots, x_n$ en $\mathbb{R}^p$ tal que

$$d_{ij} \approx ||x_i - x_j||_2 \text{ lo más posible.}$$

En algunos casos para p suficientemente grande, existe una configuración $x_1, \ldots, x_n$ tal que se cumple la igualdad, $d_{ij} = ||x_i - x_j||_2$. En este caso, decimos que d es euclídea.

En caso contrario, decimos que d es no-euclídea. Es decir, para cada p, existen i, j tal que $d_{ij} \neq ||x_i - x_j||_2$.

MDS CLÁSICO

Supongamos por ahora que la d es euclídea. El objetivo de MDS Clásico (cMDS) es encontrar una matrix $X = [x_1, \ldots, x_n]$ tal que $||x_i - x_j||_2 = d_{ij}$. Observemos que la solución no es única.

MDS CLÁSICO

Supongamos por ahora que la d es euclídea. El objetivo de MDS Clásico (cMDS) es encontrar una matrix $X = [x_1, \ldots, x_n]$ tal que $||x_i - x_j||_2 = d_{ij}$. Observemos que la solución no es única.

Si $X \in \mathbb{R}^{q \times n}$ es solución, entonces $X^* = X + c$ con $c \in \mathbb{R}^q$ también lo es, pues

$$d_{ij} = ||x_i - x_j||_2 = ||(x_i + c) - (x_j + c)||_2 = ||x_i^* - x_j^*||_2.$$

Nos quedaremos con la solución centrada, es decir

$$\sum_{i=1}^{n} x_{il} = 0 \quad 1 \leq l \leq q. \tag{1}$$

MDS CLÁSICO

Resumiendo cMDS busca $x_1, \ldots, x_n$ en $\mathbb{R}^q$ centrados para $q \geq n - 1$ tal que las distancias coincidan con la matriz original d.

MDS CLÁSICO

Resumiendo cMDS busca $x_1, \ldots, x_n$ en $\mathbb{R}^q$ centrados para $q \geq n-1$ tal que las distancias coincidan con la matriz original d.

El enfoque no es encontrar directamente X, sino la matriz de Gram $B = X'X$

Observar que la matriz B contiene los productos internos,

$$B = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix} = \begin{bmatrix} x_1'x_1 & x_1'x_2 & \ldots & x_1'x_n \\ x_2'x_1 & x_2'x_2 & \ldots & x_2'x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_n'x_1 & x_n'x_2 & \ldots & x_n'x_n \end{bmatrix}$$

MDS CLÁSICO

Como suponemos que las disimilaridades o distancias son euclídeas,

$$d_{ij}^2 = ||x_i - x_j||_2^2 = x_i' x_i + x_j' x_j - 2x_i' x_j = b_{ii} + b_{jj} - 2b_{ij}$$

MDS CLÁSICO

Como suponemos que las disimilaridades o distancias son euclídeas,

$$d_{ij}^2 = ||x_i - x_j||_2^2 = x_i' x_i + x_j' x_j - 2x_i' x_j = b_{ii} + b_{jj} - 2b_{ij}$$

Observar que de (1)

$$\sum_{i=1}^{n} b_{ij} = \sum_{i=1}^{n} \sum_{l=1}^{q} x_{il} x_{jl} = \sum_{l=1}^{q} x_{jl} \sum_{i=1}^{n} x_{il} = 0.$$

MDS CLÁSICO

Como suponemos que las disimilaridades o distancias son euclídeas,

$$d_{ij}^2 = ||x_i - x_j||_2^2 = x_i' x_i + x_j' x_j - 2x_i' x_j = b_{ii} + b_{jj} - 2b_{ij}$$

Observar que de (1)

$$\sum_{i=1}^{n} b_{ij} = \sum_{i=1}^{n} \sum_{l=1}^{q} x_{il} x_{jl} = \sum_{l=1}^{q} x_{jl} \sum_{i=1}^{n} x_{il} = 0.$$

Notemos $T = traza(B) = \sum_{i=1}^{n} b_{ii}$, entonces

MDS CLÁSICO

$$d_{\cdot j}^2 = \sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} \left(b_{ii} + b_{jj} - 2b_{ij} \right) = T + nb_{jj}. \tag{2}$$

$$d_{i\cdot}^2 = \sum_{j=1}^{n} d_{ij}^2 = \sum_{j=1}^{n} \left(b_{ii} + b_{jj} - 2b_{ij} \right) = nb_{ii} + T. \tag{3}$$

MDS CLÁSICO

$$d_{\cdot j}^2 = \sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} \left(b_{ii} + b_{jj} - 2b_{ij} \right) = T + nb_{jj}. \tag{2}$$

$$d_{i\cdot}^2 = \sum_{j=1}^{n} d_{ij}^2 = \sum_{j=1}^{n} \left(b_{ii} + b_{jj} - 2b_{ij} \right) = nb_{ii} + T. \tag{3}$$

Juntando (2) y (3),

$$d^2_{\cdot\cdot} = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2nT \tag{4}$$

Finalmente,

$$
\begin{aligned}
d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij} &= (d_{i}^2. - T)\frac{1}{n} + (d_{.j}^2 - T)\frac{1}{n} - 2b_{ij} \\
&= \frac{d_{i}^2.}{n} + \frac{d_{.j}^2}{n} - 2\frac{T}{n} - 2b_{ij} = \frac{d_{i}^2.}{n} + \frac{d_{.j}^2}{n} - \frac{d..}{n^2} - 2b_{ij} \\
b_{ij} &= -\frac{1}{2}\left[d_{ij}^2 - \frac{d_{i}^2.}{n} - \frac{d_{.j}^2}{n} + \frac{d..}{n^2}\right].
\end{aligned}
$$

Lo cual también demuestra que tenemos solución única para B.

MDS CLÁSICO

Hemos podido construir una matriz simétrica B a partir de las disimilaridades o distancias dadas. X lo encontramos a partir de la descomposición espectral de B.

Es decir, existen $\{\gamma_1, \ldots, \gamma_n\}$ matriz de autovectores de B con autovalores asociados $\lambda_1, \ldots, \lambda_n$ tal que $B = \Gamma\Lambda\Gamma'$ donde

- $\Gamma = [\gamma_1\gamma_2 \ldots \gamma_n]$ es la matriz con autovectores como columnas.
- $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ es la matriz diagonal de autovalores.

MDS CLÁSICO

Finalmente la solución es tomar $X = \Lambda^{1/2}\Gamma'$, pues

$$B = \Gamma\Lambda\Gamma' = \Gamma\Lambda^{1/2}\Lambda^{1/2}\Gamma' = (\Gamma\Lambda^{1/2})(\Gamma\Lambda^{1/2})' = X'X.$$

Si queremos reducir la dimensión a $p \leq q$, las primeras p columnas de X son las que mejor preservan las d_{ij} (autovalores más grandes). Tomamos $X_{(p)} = \Lambda_p \Gamma'_p$ que se corresponde con

- Λ_p la submatriz superior de $p \times p$.
- Γ_p las primeras p columnas de Γ.

MDS MÉTRICO

Estos algoritmos permiten relajar la hipótesis $d_{ij} \approx ||x_i - x_j||_2$ por $f(d_{ij}) \approx ||x_i - x_j||_2 = \widehat{d}_{ij}$ donde f es una función monótona.

MDS MÉTRICO

Estos algoritmos permiten relajar la hipótesis $d_{ij} \approx ||x_i - x_j||_2$ por $f(d_{ij}) \approx ||x_i - x_j||_2 = \widehat{d}_{ij}$ donde f es una función monótona.

Los métodos de MDS métrico se basan en algoritmos de optimización que tratan de minimizar una función que se llama stress.

$$stress(\widehat{d}_{ij}) = \left(\sum_{i<j} \frac{\left(\widehat{d}_{ij} - f(d_{ij})\right)^2}{\sum_{j=1} d_{ij}} \right)^{1/2}$$

Usualmente la f se construye de forma paramétrica, por ejemplo $f(d_{ij}) = \alpha + \beta d_{ij}$.

MDS MÉTRICO - SAMMON

El algoritmo de Sammon es un ejemplo de MDS métrico. Toma la f como la identidad y normaliza las distancias.

MDS MÉTRICO - SAMMON

El algoritmo de Sammon es un ejemplo de MDS métrico. Toma la f como la identidad y normaliza las distancias.

Se basa en la función de stress:

$$Sammon - stress(\hat{d}_{ij}) = \frac{1}{\sum_{l<k} d_{lk}} \sum_{i<j} \frac{(\hat{d}_{ij} - d_{ij})^2}{d_{ij}}.$$

MDS MÉTRICO - SAMMON

El algoritmo de Sammon es un ejemplo de MDS métrico. Toma la f como la identidad y normaliza las distancias.

Se basa en la función de stress:

$$Sammon - stress(\hat{d}_{ij}) = \frac{1}{\sum_{l<k} d_{lk}} \sum_{i<j} \frac{(\hat{d}_{ij} - d_{ij})^2}{d_{ij}}.$$

Sammon preserva las d_{ij} mas chicas dandoles el mismo peso que a las más grandes.

La solución se obtiene por procedimientos numéricos.

MDS NO MÉTRICO

Hasta el momento hemos asumido que las disimilaridades son conocidas.

MDS NO MÉTRICO

Hasta el momento hemos asumido que las disimilaridades son conocidas.

No siempre es el caso: los objetos son rankeados. Por ejemplo, podemos tener un problema donde las disimilaridades entre los objetos no son conocidas, pero sí su orden.

MDS NO MÉTRICO

Hasta el momento hemos asumido que las disimilaridades son conocidas.

No siempre es el caso: los objetos son rankeados. Por ejemplo, podemos tener un problema donde las disimilaridades entre los objetos no son conocidas, pero sí su orden.

En este caso vamos a querer asignar valores a las proximidades que mantengan la estructura de orden dado.

MDS NO MÉTRICO

Consideremos el siguiente ejemplo con rangos en la matriz de disimilaridades en marcas de auto (según Lucas):

MDS NO MÉTRICO

Consideremos el siguiente ejemplo con rangos en la matriz de disimilaridades en marcas de auto (según Lucas):

i / j	1	2	3	4	
		Mercedes	Jaguar	Ferrari	VW
1	Mercedes				
2	Jaguar	3			
3	Ferrari	2	1		
4	VW	5	4	6	

MDS NO MÉTRICO

Supongamos que tomamos unas coordenas que nos parecen.

Marca	x_1	x_2
Mercedes	3	2
Jaguar	2	7
Ferrari	1	3
VW	10	4

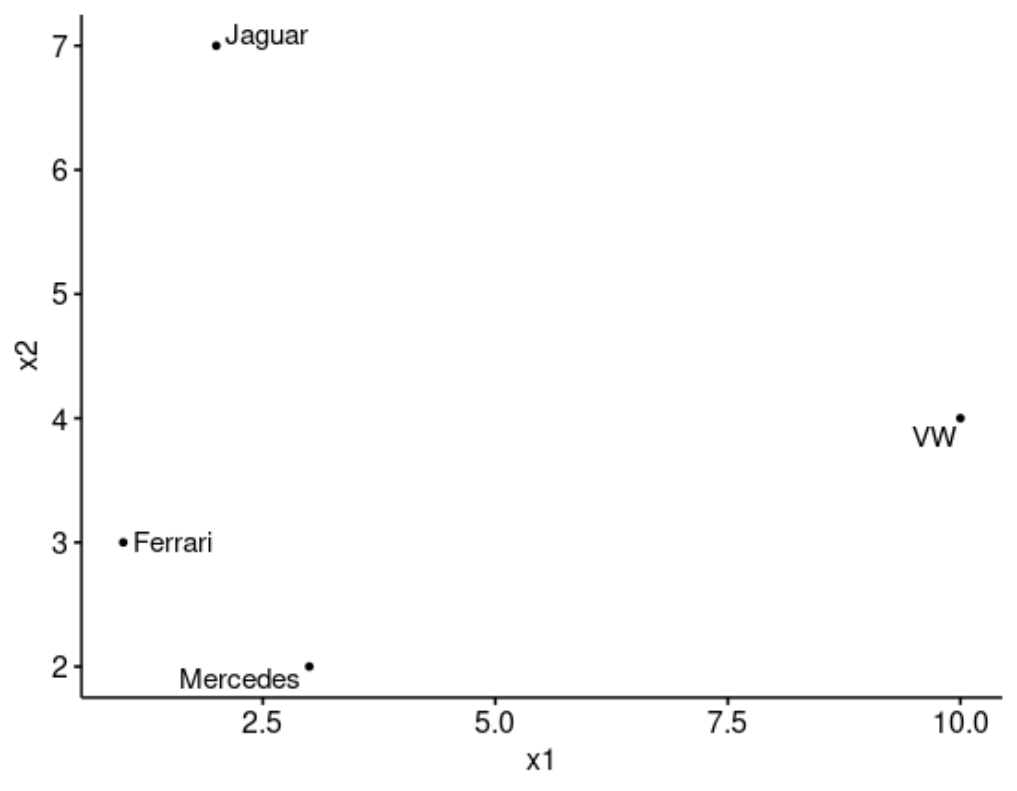
MDS NO MÉTRICO

Supongamos que tomamos unas coordenas que nos parecen.

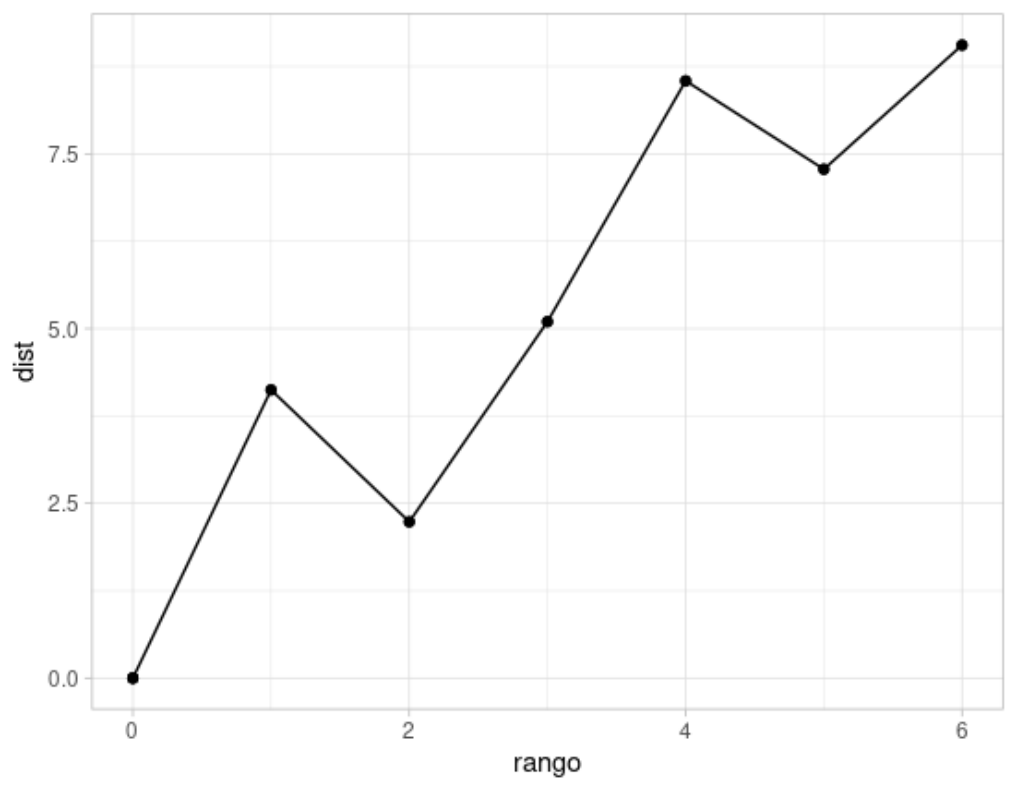
Marca	x_1	x_2
Mercedes	3	2
Jaguar	2	7
Ferrari	1	3
VW	10	4

¿Qué tienen que preservar?

MDS NO MÉTRICO



MDS NO MÉTRICO



MDS NO MÉTRICO

En general la f está definida de forma implícita.

En este caso $f(d_{ij}) = d^*_{ij}$ llevan el nombre de disparidad, soló preservan el orden:

$$d_{ij} < d_{kl} \Leftrightarrow f(d_{ij}) \leq f(d_{kl}) \Leftrightarrow d^*_{ij} \leq d^*_{kl}. \tag{5}$$

MDS NO MÉTRICO

En general la f está definida de forma implícita.

En este caso $f(d_{ij}) = d^*_{ij}$ llevan el nombre de disparidad, soló preservan el orden:

$$d_{ij} < d_{kl} \Leftrightarrow f(d_{ij}) \leq f(d_{kl}) \Leftrightarrow d^*_{ij} \leq d^*_{kl}. \tag{5}$$

Tenemos doble tarea construir las disparidades y las coordenadas.

MDS NO MÉTRICO - KRUSKAL

Llamemos $\hat{d}_{ij} = ||x_i - x_j||^2$ el algoritmo de Kruskal se basa en minimizar la función de stress:

$$Kruskall - stress(\hat{d}_{ij}, d^*_{ij}) = \left(\sum_{i<j} \frac{(\hat{d}_{ij} - d^*_{ij})^2}{\sum_{i<j} \hat{d}^2_{ij}} \right)^{1/2}.$$

MDS NO MÉTRICO - KRUSKAL

Llamemos $\hat{d}_{ij} = ||x_i - x_j||^2$ el algoritmo de Kruskal se basa en minimizar la función de stress:

$$Kruskall - stress(\hat{d}_{ij}, d^*_{ij}) = \left(\sum_{i<j} \frac{(\hat{d}_{ij} - d^*_{ij})^2}{\sum_{i<j} \hat{d}^2_{ij}} \right)^{1/2}.$$

Notar que las disimilaridades originales sólo se usan para asegurar que se cumpla (5).

MDS NO MÉTRICO - KRUSKAL

Llamemos $\hat{d}_{ij} = ||x_i - x_j||^2$ el algoritmo de Kruskal se basa en minimizar la función de stress:

$$Kruskall - stress(\hat{d}_{ij}, d^*_{ij}) = \left(\sum_{i<j} \frac{(\hat{d}_{ij} - d^*_{ij})^2}{\sum_{i<j} \hat{d}^2_{ij}} \right)^{1/2}.$$

Notar que las disimilaridades originales sólo se usan para asegurar que se cumpla (5).

La función f se puede interpretar como la curva de regresión entre $\hat{d}_{ij}$ y d^*_{ij}.