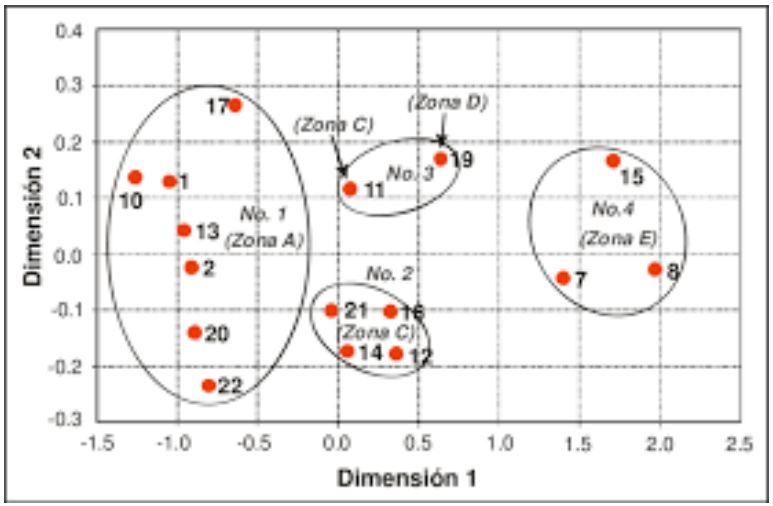
Escalamiento Multidimensional

Técnicas de Visualización y Reducción de Datos

Jerónimo Barragán y Manuel Horn

Tercer Bimestre de 2024



Problema

Disponemos de una matriz cuadrada $\mathbf{D}$ de $n \times n$, de distancias o disimilaridades entre n elementos de un conjunto.
El objetivo es representar esta matriz $\mathbf{D}$ con una $\mathbf{X}$ de $n \times p$, que pueda interpretarse como la matriz de n observaciones con coordenadas en p variables decorrelacionadas, cuyas distancias euclídeas aproximen las de la matriz original $\mathbf{D}$.
Entonces, buscamos encontrar un conjunto de variables $y_1, ..., y_p$ ortogonales, con $p < n$ que cumplan lo mencionado.

Métodos

Existen dos métodos de escalamiento multidimensional.

- **métricos** : la matriz inicial es de distancias.
- **no métricos**: la matriz inicial es de similaridades.

Escalamiento métrico: caso euclídeo

Dada la matriz $\mathbf{X}$ de observaciones por filas en $n \times p$ (de rango completo p), obtenemos variables con media cero mediante

$$\widetilde{\mathbf{X}} = \underbrace{(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')}_{:=\mathbf{P}}\mathbf{X}$$

A partir de esta matriz podemos derivar las matrices

$$\mathbf{S} = \frac{1}{n}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} \text{ y } \mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}' \text{ en } n \times n$$

de covarianzas y similitudes, respectivamente. Esta última es de interés ya que contiene los productos escalares entre pares de observaciones, que vienen a representar similitud coseno:

$$q_{ij} = \sum_{s=1}^{p} x_{is}x_{js} = \mathbf{x}_i'\mathbf{x}_j = |\mathbf{x}_i||\mathbf{x}_j| \cos\theta_{ij}$$

Inicialmente el problema que se nos presenta es construir la matriz $\mathbf{X}$ a partir de la matriz de distancias euclídeas al cuadrado $\mathbf{D}$.

Escalamiento métrico: caso euclídeo

Se puede mostrar primero que $\mathbf{D} = diag(\mathbf{Q})\mathbf{1}' + \mathbf{1}diag(\mathbf{Q})' - 2\mathbf{Q}$, y luego que

$$\mathbf{Q} = -\frac{1}{2}\mathbf{PDP}$$

Observando que $\mathbf{Q}$ tiene rango p igual al de $\mathbf{X}$, se puede escribir

$$\mathbf{Q} = \mathbf{V}\Lambda\mathbf{V}'$$

- $\mathbf{V}$ es $n \times p$ y contiene en sus columnas a los autovectores correspondientes a los autovalores no nulos de $\mathbf{Q}$, todos ortogonales entre sí.
- Λ es diagonal $p \times p$ y contiene los autovalores correspondientes.

Si además $\mathbf{Q}$ es semidefinida positiva, se tiene

$$\mathbf{Q} = (\mathbf{V}\Lambda^{1/2})(\Lambda^{1/2}\mathbf{V}')$$

De donde tomando

$$\mathbf{Y} = \mathbf{V}\Lambda^{1/2}$$

tenemos una matriz $n \times p$ con p variables no correlacionadas. Definiendo las p **coordenadas principales** $\mathbf{z}_i' = (\mathbf{y}_{i1}, ..., \mathbf{y}_{ip})$, las distancias al cuadrado entre dos puntos cualesquiera $\delta_{ij} = (\mathbf{z}_i - \mathbf{z}_j)'(\mathbf{z}_i - \mathbf{z}_j)$ satisfacen $\delta_{ij} = d_{ij}$

Escalamiento métrico: caso euclídeo

- ¿Obtendremos en $\mathbf{z}_i$ las variables originales $\mathbf{x}_i$? No, estamos obteniendo sus **componentes principales**. Esto es inevitable, porque las distancias entre puntos no varían si modificamos las medias de las variables, o rotamos los puntos, es decir que distintas matrices de datos pueden llevar a las mismas coordenadas principales.
- ¿La solución al problema es única? Tampoco, por ejemplo si transformamos los puntos con una matriz ortogonal $\mathbf{T}$:

$$\delta_{ij} = (\mathbf{T}\mathbf{z}_i - \mathbf{T}\mathbf{z}_j)'(\mathbf{T}\mathbf{z}_i - \mathbf{T}\mathbf{z}_j) = (\mathbf{z}_i - \mathbf{z}_j)'\mathbf{T}'\mathbf{T}(\mathbf{z}_i - \mathbf{z}_j) = (\mathbf{z}_i - \mathbf{z}_j)'(\mathbf{z}_i - \mathbf{z}_j) = d_{ij}$$

- ¿$\mathbf{Q}$ es siempre semidefinida positiva? A priori podría no serlo...

Escalamiento métrico como reducción de la dimensión

Hallamos $\mathbf{Y}$ de $n \times p$ tal que $\mathbf{Q} = \mathbf{Y}\mathbf{Y}'$, donde $\mathbf{Y}$ tiene en cada fila las p coordenadas principales de cada observación.
Si $\mathbf{Q} = \mathbf{V}\Lambda\mathbf{V}'$ es descomposición espectral, entonces la matriz de rango $r < p$ que mejor aproxima a $\mathbf{Q}$ en norma de Frobenius resulta ser

$$\mathbf{Q}_r = \mathbf{V}_r\Lambda_r\mathbf{V}_r'$$

que resultan de quedarse con las primeras r columnas de cada matriz (suponiendo que los autovalores están ordenados decrecientemente). Luego, tomando

$$\mathbf{Y}_r = \mathbf{V}_r\Lambda_r^{1/2}$$

obtenemos un conjunto de n puntos de dimensión r cuyas distancias euclídeas mejor aproximan a las distancias originales, es decir, las r coordenadas principales de cada observación original.

Escalamiento métrico: caso general

Matrices compatibles con métricas euclídeas

Dada una matriz de distancias $\mathbf{D}$ en $n \times n$, la matriz $\mathbf{Q} = -\frac{1}{2}\mathbf{PDP}$ de rango p es semidefinida positiva si y sólo si existen n vectores de dimensión p cuyas distancias euclídeas reproducen $\mathbf{D}$. En ese caso decimos que $\mathbf{D}$ es compatible con una métrica euclídea.

Esta proposición es útil incluso cuando $\mathbf{D}$ no sea compatible con una métrica euclídea. En ese caso, si los autovalores de Λ están ordenados decrecientemente, nos quedamos con las primeras r columnas de $\mathbf{V}$ asociadas a autovectores de autovalor positivo, obteniendo una matriz semidefinida positiva

$$\mathbf{Q}_r = \mathbf{V}_r \Lambda_r \mathbf{V}_r'$$

Si los autovalores negativos son chicos en valor absoluto, las distancias entre $\mathbf{z}_i$ y $\mathbf{z}_j$ deberían ser una buena aproximación a las distancias originales d_{ij}.

Escalamiento métrico: ejemplo

	Beijing	Cape Town	Hong Kong	Honolulu	London	Melbourne
Cape Town	12947					
Hong Kong	1972	11867				
Honolulu	8171	18562	8945			
London	8160	9635	9646	11653		
Melbourne	9093	10338	7392	8862	16902	
Mexico	12478	13703	14155	6098	8947	13557
Montreal	10490	12744	12462	7915	5240	16730
Moscow	5809	10101	7158	11342	2506	14418
New Delhi	3788	9284	3770	11930	6724	10192
New York	11012	12551	12984	7996	5586	16671
Paris	8236	9307	9650	11988	341	16793
Rio de Janeiro	17325	6075	17710	13343	9254	13227
Rome	8144	8417	9300	12936	1434	15987
San Francisco	9524	16487	11121	3857	8640	12644
Singapore	4465	9671	2575	10824	10860	6050
Stockholm	6725	10334	8243	11059	1436	15593
Tokyo	2104	14737	2893	6208	9585	8159

Fuente: "Modern Multivariate Statistical Techniques". Alan Julian Izenman. Springer. 2008.

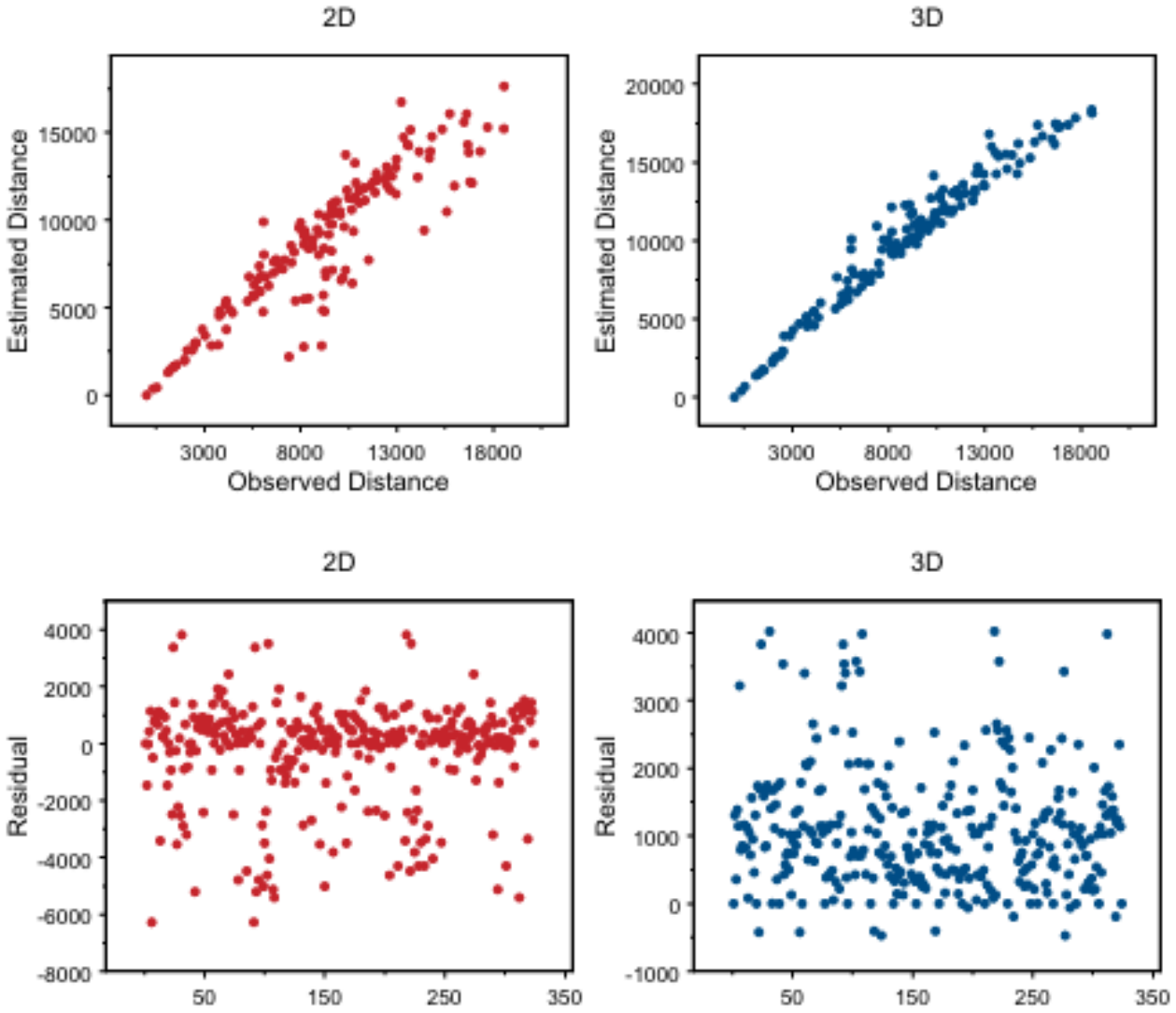
Escalamiento métrico: ejemplo

	Eigenvalues	Eigenvectors		
1	471582511	0.245	–0.072	0.183
2	316824787	0.003	0.502	-0.347
3	253943687	0.323	–0.017	0.103
4	–98466163	0.044	–0.487	-0.080
5	–74912121	–0.145	0.144	0.205
6	–47505097	0.366	–0.128	–0.569
7	31736348	–0.281	–0.275	–0.174
8	–7508328	–0.272	–0.115	0.094
9	4338497	–0.010	0.134	0.202
10	1747583	0.209	0.195	0.110
11	–1498641	–0.292	–0.117	0.061
12	145113	–0.141	0.163	0.196
13	–102966	–0.364	0.172	–0.473
14	60477	–0.104	0.220	0.163
15	–6334	–0.140	–0.356	–0.009
16	–1362	0.375	0.139	–0.054
17	100	–0.074	0.112	0.215
18	0	0.260	–0.214	0.173

Autovalores y autovectores de $\mathbf{Q}$.

Fuente: "Modern Multivariate Statistical Techniques". Alan Julian Izenman. Springer. 2008.

Escalamiento métrico: ejemplo



Distancias observadas vs. distancias entre las coordenadas principales.

Fuente: "Modern Multivariate Statistical Techniques". Alan Julian Izenman. Springer. 2008.

Escalamiento métrico: ejemplo

"We should not really be surprised at the results in this example. The differences occur because of the fact that the estimated airline distances are taken to be Euclidean. Airline distances are measured over a curved surface rather than a flat one. We should, therefore, expect to see a certain amount of distortion when we use a Euclidean metric to estimate distances between cities distributed across the surface of a globe"

Fuente: "Modern Multivariate Statistical Techniques". Alan Julian Izenman. Springer. 2008

Escalado no métrico

La obtención de la matriz de diferencias entre objetos en este caso es distinta. Generalmente se obtienen por alguno de los siguientes procedimientos:

- Estimación directa
- Estimación de rangos
- Rangos por pares

Se establece que las disimilitudes δ_{ij} entre elementos está relacionada con las distancias euclídeas d_{ij} mediante una función desconocida

$$\delta_{ij} = f(d_{ij})$$

donde se impone que f es monótona:

$$\delta_{ij} > \delta_{ih} \iff d_{ij} > d_{ih}$$

Escalado no métrico

Buscamos encontrar coordenadas que sean capaces de reproducir estas distancias bajo esta condición para f. Necesitamos:

- Un criterio de bondad de ajuste que sea invariante ante transformaciones monótonas de los datos.
- Un algoritmo para obtener las coordenadas, optimizando el criterio establecido.

Criterios STRESS y S-STRESS

Dadas las disimilitudes ordenadas crecientemente, $\delta_{i_1 j_1} < \delta_{i_2 j_2} < ... < \delta_{i_m j_m}$, con $m = n(n-1)/2$, buscamos representar los n puntos y_i en una dimensión t tal que $d_{i_1 j_1} < d_{i_2 j_2} < ... < d_{i_m j_m}$ (mismo orden que la disimilitudes), con $d_{ij} = ||y_i - y_j||$.

Esto no siempre va a ser posible, y por eso buscamos aproximar d_{ij} por $\hat{d}_{ij}$ (disparidades), que están relacionadas a d_{ij} de manera monótona y cumplen $\hat{d}_{i_1 j_1} \leq \hat{d}_{i_2 j_2} \leq ... \leq \hat{d}_{i_m j_m}$

Las $\hat{d}_{ij}$ pueden no ser distancias, y puede no existir una configuración de puntos y_i tal que d_{ij} son sus distancias.

La función objetivo a minimizar para obtener esto será el STRESS:

$$STRESS = S(y_1, ..., y_n) = \left\{ \frac{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2} \right\}^{1/2}$$

o alternativamente el S-STRESS:

$$S - STRESS = SS(y_1, ..., y_n) = \sum_{i<j} (d_{ij}^2 - \hat{d}_{ij}^2)^2$$

Algoritmo general para escalamiento no métrico

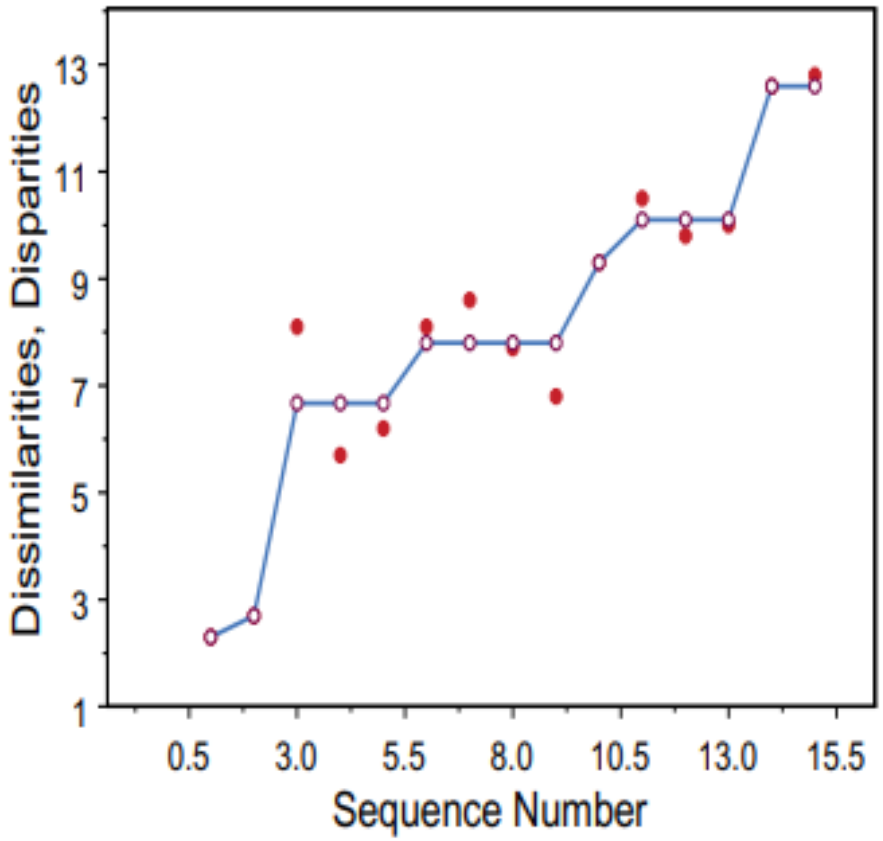
1. Ordenar las $m = n(n-1)/2$ disimilaridades δ_{ij} decrecientemente
2. Fijar t, el número de dimensiones, y elegir una configuración inicial de puntos $\mathbf{y}_i \in \mathcal{R}^t, i = 1, ..., n$
3. Computar distancias d_{ij} entre pares
4. Usar regresión isotónica (preserva monotonía) para producir $\hat{d}_{ij}$, computar el valor inicial de STRESS
5. Usar gradient descent con respecto a las coordenadas $\mathbf{y}_{ij}$ y función objetivo STRESS; esto produce un nuevo conjunto de d_{ij}
6. Nuevamente usar regresión isotónica para actualizar $\hat{d}_{ij}$
7. Repetir 5 y 6 hasta algún criterio de parada (puede ser que no mejore STRESS)
8. Repetir pasos anteriores para valores diferentes de t. Con la regla del codo decidimos el t final.

TABLE 13.8. *Finding the disparities by isotonic regression for an artificial example with $n = 6$ and $m = 15$. The columns I, II, III, IV, V, and VI display a sequence of trial solutions for the disparities. The cells in red indicate the active block at each trial solution. The value of S is 6.85%.*

rank	d_{ij}	I	II	III	IV	V	VI	$\hat{d}_{ij}$
1	2.3	2.3	2.3	2.30	2.30	2.30	2.30	2.30
2	2.7	2.7	2.7	2.70	2.70	2.70	2.70	2.70
3	8.1	**8.1**	**6.9**	6.67	6.67	6.67	6.67	6.67
4	5.7	**5.7**	**6.9**	6.67	6.67	6.67	6.67	6.67
5	6.2	6.2	**6.2**	6.67	6.67	6.67	6.67	6.67
6	8.1	8.1	8.1	**8.10**	**8.13**	7.80	7.80	7.80
7	8.6	8.6	8.6	**8.60**	**8.13**	7.80	7.80	7.80
8	7.7	7.7	7.7	**7.70**	**8.13**	7.80	7.80	7.80
9	6.8	6.8	6.8	6.80	**6.80**	7.80	7.80	7.80
10	9.3	9.3	9.3	9.30	9.30	9.30	9.30	9.30
11	10.5	10.5	10.5	10.50	10.50	**10.50**	**10.15**	10.10
12	9.8	9.8	9.8	9.80	9.80	**9.80**	**10.15**	10.10
13	10.0	10.0	10.0	10.00	10.00	10.00	**10.00**	10.10
14	12.6	12.6	12.6	12.60	12.60	12.60	12.60	12.60
15	12.8	12.8	12.8	12.80	12.80	12.80	12.80	12.60

Fuente: "Modern Multivariate Statistical Techniques". Alan Julian Izenman. Springer. 2008.

Disparidades usando regresión isotónica



Puntos rojos: distancias euclídeas halladas con el algoritmo (disimilaridades, no monótono)
Curva azul: aproximaciones de distancias halladas con el algoritmo (disparidades, monótono)
Fuente: "Modern Multivariate Statistical Techniques". Alan Julian Izenman. Springer. 2008.

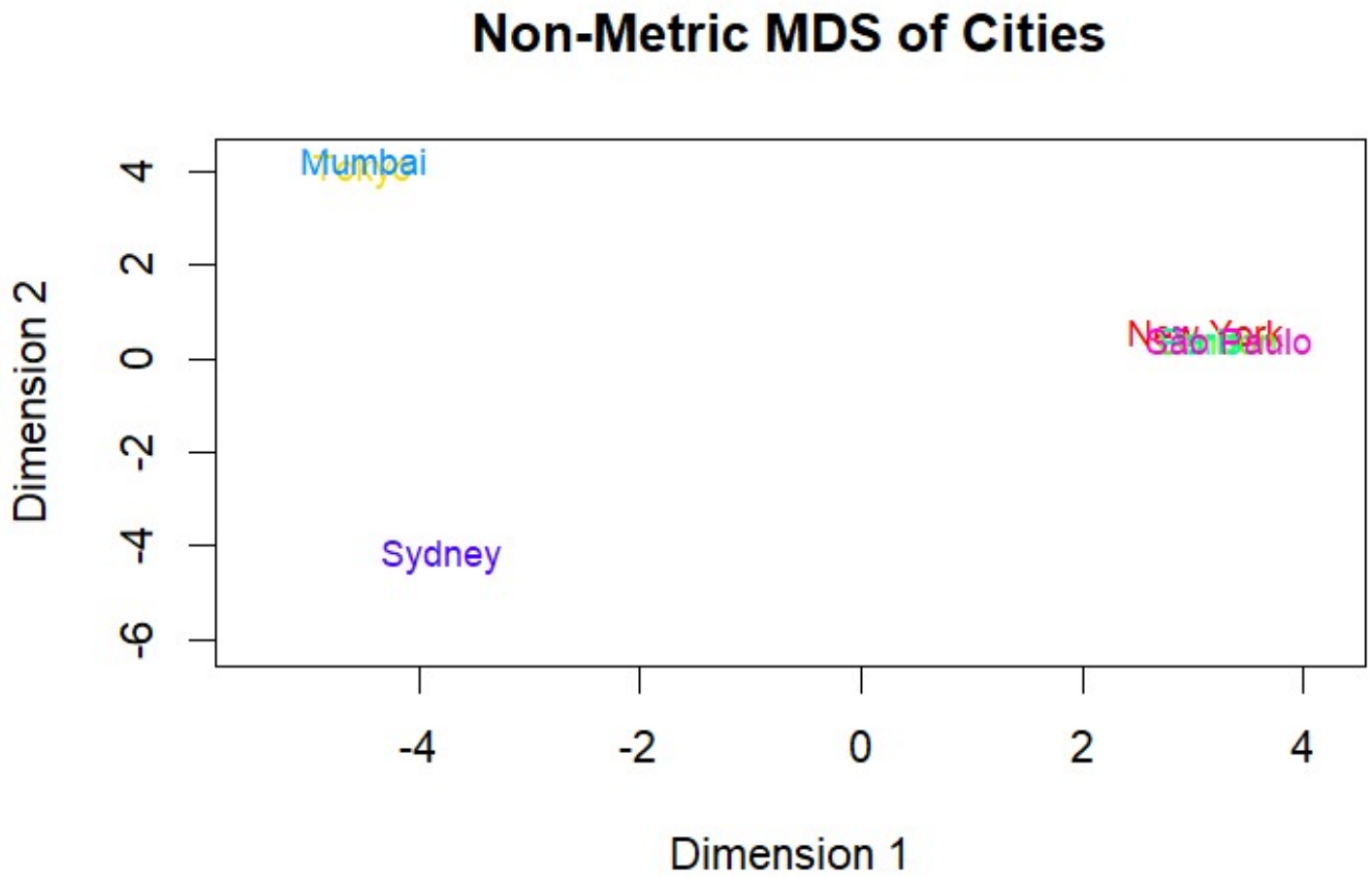
Escalamiento no métrico: ejemplo

Asumimos disimilaridades hipotéticas pero realistas basadas en factores como: diferencias culturales, poder económico, distancia geográfica, idioma, diferencias culturales y diferencia horaria.

	New York	Tokyo	London	Paris	Mumbai	Sydney	São Paulo
New York	0	9.5	3.2	4.0	8.0	9.0	6.5
Tokyo	9.5	0	7.5	7.8	5.2	7.2	9.0
London	3.2	7.5	0	2.5	7.8	8.5	7.0
Paris	4.0	7.8	2.5	0	7.6	8.8	6.8
Mumbai	8.0	5.2	7.8	7.6	0	8.5	8.7
Sydney	9.0	7.2	8.5	8.8	8.5	0	8.9
São Paulo	6.5	9.0	7.0	6.8	8.7	8.9	0

Matriz de disimilaridades entre ciudades.

Escalamiento no métrico: ejemplo



Escalado no métrico: coordenadas principales

Preguntas?