Análisis de datos. Técnicas aplicadas a datos de proximidad

Tema 2: MDS clásico.

Solución clásica en MDS

Consideremos una matriz simétrica de datos de disimilaridad $\Delta=(\delta_{ij})$ entre n objetos. El modelo métrico del MDS pretende encontrar un conjunto de puntos en un espacio de dimensión reducida, de forma que las distancias entre los puntos $\{d_{ij}\}$, se aproximen cuantitativamente tanto como sea posible a las disimilaridades de partida, es decir, $d_{ij} \approx f(\delta_{ij})$, con f una función monótona, paramétrica y continua. Desde un punto de vista matemático, supongamos que los objetos pertenecen a un conjunto O y sea $\delta_{ij} = \delta(o_i, o_j)$ un elemento de la matriz de disimilaridades Δ entre cada par de objetos $o_i, o_j \in O$. Consideremos Φ una aplicación arbitraria, donde $\Phi(o_i) = x_i$, es un vector de dimensión K que contiene las coordenadas del objeto o_i en el espacio de representación que supondremos de dimensión K. Así, el objetivo será encontrar una aplicación Φ, representada a través de una matriz de configuración X, para la que $d_{ij} \approx f(\delta_{ij}), \forall i, j$.

Una solución al problema puede obtenerse mediante el procedimiento clásico de Torgerson (1958), que está basado en el siguiente resultado debido a Schoenberg (1935) y a Young & Houselholder (1938), descrito en el tema anterior y cuya demostración puede verse en Mardia et al. [1980], pg(397).

Teorema:

Sea $D_{(n\times n)}$ una matriz de distancias entre n puntos en un espacio de configuración de dimensión K y sea $B_{(n\times n)}$ la matriz dada por $B = HAH$, siendo $H_{(n\times n)}$ dada por $H = I - n^{-1}\mathbf{1}\mathbf{1}^t$ y $A_{(n\times n)}$ la matriz cuyos elementos vienen dados a través de $a_{rs} = -d_{rs}^2/2$. Entonces, D es una matriz de distancias Euclídeas, si y solo si, B es semidefinida positiva. Además se tiene:

1. Si D es la matriz de distancias Euclídeas para una configuración dada por $Z_{(n\times K)} = (z_1,.,z_n)^t$, entonces $B = (HZ)(HZ)^t$, es decir $b_{rs} = (z_r - \overline{z})^t(z_s - \overline{z}), \quad \forall r,s = 1,\ldots,n,$ de donde $B \geq 0.$ B será la matriz centrada de productos escalares de Z.

2. Inversamente, si B es semidefinida positiva de rango K, entonces puede construirse una configuración asociada a B de la siguiente

forma: Sean $\lambda_1 > \cdots > \lambda_K$ los K valores propios positivos de B correspondientes a los vectores propios $X_{(n\times K)} = (x_{(1)},.,x_{(K)})$, normalizados según la condición $x'_{(i)}x_{(i)} = \lambda_i,\ \forall i = 1,.,K$. Los puntos de $\mathbb{R}^K$ de coordenadas $x_r = (x_{r1},.,x_{rp})^t$ (donde x_r representa la r-ésima fila de la matriz X), tienen matriz de distancias D. Además esa configuración está centrada en $\bar{x} = 0$ y B es la matriz de productos escalares de esa configuración.

Ejercicio 2.1;

Apoyándote en el libro de Mardia et al. (1980), describe razonadamente la demostración del teorema.

Tal y como se comentó en el tema anterior, ese resultado indica que existe solución única para distancias Euclídeas en un espacio de dimensión $K = rang(B)$, salvo isometrías, donde K como máximo será $(n-1)$, puesto que **1** es un vector propio de B asociado al valor propio 0. El problema que resuelve MDS surge cuando se pretende una representación en un espacio de dimensión menor que K y/o la información de la que se dispone entre cada par de elementos a representar no viene dada en términos de distancia Euclídea sino en términos de una matriz de disimilaridades.

Asistido por Ledyard Tucker, Torgerson propuso uno de los primeros algoritmos de MDS, basado en la relación,

$$\delta_{ij} = d_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2},\ \forall i, j.$$

Suponiendo que los datos están centrados para cada dimensión, es decir, $\sum_i x_{ik} = \sum_j x_{jk} = 0,\ \forall i, j$, se construye una matriz doblemente centrada *B* con elementos

$$b_{ij} = -\frac{1}{2}(\delta_{ij}^2 - \delta_{i.}^2 - \delta_{.j}^2 + \delta_{..}^2),$$

Donde $\delta_{i.}^2 = \frac{1}{n}\sum_j \delta_{ij}^2$, $\delta_{.j}^2 = \frac{1}{n}\sum_i \delta_{ij}^2$ y $\delta_{..}^2 = \frac{1}{n^2}\sum_i\sum_j \delta_{ij}^2$. Si *B* es *semidefinida positiva* de rango *K*, tendrá *K* valores propios no negativos y (*n*-*K*) cero, por lo que podemos considerar la descomposición espectral de *B* en función de esos *K* valores propios ordenados de mayor a menor $(\lambda_1 \geq \cdots \geq \lambda_K)$,

$$B = V\Lambda V^t = XX^t,$$

siendo

$$X = V\Lambda^{1/2},$$

donde $\Lambda = diag(\lambda_1,\cdots,\lambda_K)$ es la matriz diagonal de valores propios de B y $V = [v_1,\cdots,v_K]$ la matriz de los correspondientes vectores propios normalizados de forma que $v_i v_i^t = 1$ y donde $\Lambda^{1/2} = (\lambda_1^{1/2},\cdots,\lambda_K^{1/2})$. La matriz B será la matriz de productos escalares asociada a la configuración encontrada X. Se trata de ver que la matriz de distancias D que se obtiene a partir de X coincide con Δ. Pero

$$(x_r - x_s)(x_r - x_s)^t = x_r^t x_r - 2x_r^t x_s + x_s^t x_s = b_{rr} - 2b_{rs} + b_{ss} = (*),$$

y $B = HAH$, con $H = I - n^{-1}\mathbf{1}\mathbf{1}^t$ donde A la matriz dada por, $a_{rs} = \frac{-1}{2}\delta_{rs}^2$. Así pues, sustituyendo las b_{rs} por su expresión se obtiene,

$$(*) = -\frac{1}{2}\delta_{rr}^2 + \frac{1}{2}\delta_{r.}^2 + \frac{1}{2}\delta_{.r}^2 - \frac{1}{2}\delta^2 + \delta_{rs}^2 - \delta_{r.}^2 - \delta_{.s}^2 + \delta_{ss}^2$$

$$-\frac{1}{2}\delta_{ss}^2 + \frac{1}{2}\delta_{s.}^2 + \frac{1}{2}\delta_{.s}^2 - \frac{1}{2}\delta^2 = -\frac{1}{2}\delta_{rr}^2 + \delta_{rs}^2 - \frac{1}{2}\delta_{ss}^2.$$

Hay que tener en cuenta que para obtener esta última igualdad se supone que $\delta_{.r} = \delta_{r.}$, $\forall r = 1,\ldots,n$ lo cual equivale a suponer que la matriz de disimilaridades Δ es *simétrica*. Puesto que se supone un origen definido en 0 para las disimilaridades, entonces la diagonal de Δ será nula, es decir $\delta_{rr} = 0$, $\forall r$, de donde se obtiene finalmente

$$(x_r - x_s)(x_r - x_s)^t = \delta_{rs}^2,$$

de forma que la matriz de distancias obtenidas D es precisamente Δ.

Ejercicio 2.2:
La siguiente tabla recoge las distancias entre las siguientes 10 ciudades Europeas: Londres, Estocolmo, Lisboa, Madrid, Paris, Amsterdam, Berlin, Praga, Roma y Dublin. Escribe el código necesario usando R o MatLab, para obtener los vectores propios normalizados ($v_i' v_i = \lambda_i$) para la matriz B de productos escalares centrados asociada.

Tabla 2.1. Distancia entre ciudades Europeas.

c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
0	569	667	530	141	140	357	396	570	190
569	0	1212	1043	617	446	325	423	787	648
667	1212	0	201	596	768	923	882	714	714
530	1043	201	0	431	608	740	690	516	622

141	617	596	431	0	177	340	337	436	320
140	446	768	608	177	0	218	272	519	302
357	325	923	740	340	218	0	114	472	514
396	423	882	690	337	272	114	0	364	573
569	787	714	516	436	519	472	364	0	755
190	648	714	622	320	302	514	573	755	0

Representación en dimensión baja

Del Teorema se desprende que al menos un valor propio es cero, ya que $B\mathbf{1} = V\Lambda V\mathbf{1} = 0$. Luego sabemos que al menos habrá solución en un espacio de dimensión $(n-1)$. La configuración obtenida podría rotarse hasta hacerla coincidir sobre sus ejes principales, o sea, las primeras coordenadas de los puntos serían las proyecciones sobre el primer eje principal, las segundas las del segundo, etc... Así, solo elegimos los primeros K ejes principales para representar la configuración, con $K = rang(B)$. No obstante esa rotación no es necesaria ya que la configuración, por construcción, viene dada en el sentido de los ejes principales, puesto que $XX^t = \Lambda$, y Λ es una matriz diagonal.

El problema estadístico surge si deseamos fijar la representación en K dimensiones con $K < rang(B)$. El problema tendrá solución ya que si los K primeros valores propios son suficientemente grandes, la contribución del resto de valores propios y por tanto del resto de dimensiones subyacentes será insignificante, por lo que el problema de representación en un espacio de dimensión baja, normalmente 2 o 3, quedaría resuelto en el sentido de que la solución en un espacio de dimensión menor, elegido en virtud de los primeros mayores valores propios, es la mejor configuración posible para representar los puntos en espacios de dimensión menor que el rango de B. Formalmente, el problema puede enunciarse de la siguiente forma:

Dada una matriz de disimilaridades simétrica, con diagonal nula, $\Delta_{(n\times n)}$, se trata de encontrar una configuración $\hat{X}_{(n\times K)}$ en un espacio de dimensión $K < p$, siendo $p = rang(B)$ y $B_{(n\times n)} = HAH$ la matriz de productos escalares definida según el teorema con configuración asociada $X_{(n\times p)}$, de forma que la matriz de distancias asociada a $\hat{X}$, que denotaremos por $\hat{D}_{(n\times n)}$, de elementos $\hat{d}_{rs}^2 = (\hat{x}_r - \hat{x}_s)^t(\hat{x}_r - \hat{x}_s)$, se aproxime lo máximo posible a Δ. La matriz de productos escalares asociada a $\hat{X}$ la denotaremos por $\hat{B}_{(n\times n)}$. Una medida de discrepancia entre B y $\hat{B}$ fue definida por Mardia de la forma:

$$\Psi = \sum_{r,s=1}^{n} (b_{rs} - \hat{b}_{rs})^2 = tr((B - \hat{B})^2).$$

La solución clásica de MDS es óptima para esa medida en virtud del siguiente resultado, cuya demostración puede verse en Mardia et al. [1980], pg(408).

Teorema:

Si Δ es una matriz de distancias no necesariamente Euclídeas, (por ejemplo una matriz Δ de disimilaridades simétrica, con diagonal nula) entonces para una dimensión fija K, la medida Ψ definida por la expresión anterior, alcanzará su mínimo en el conjunto de todas las configuraciones $\hat{X}$ en dimensión K, para la configuración $\hat{X}$, obtenida en la solución clásica de MDS.

Desde un punto de vista aplicado, el método clásico puede describirse de la siguiente forma:

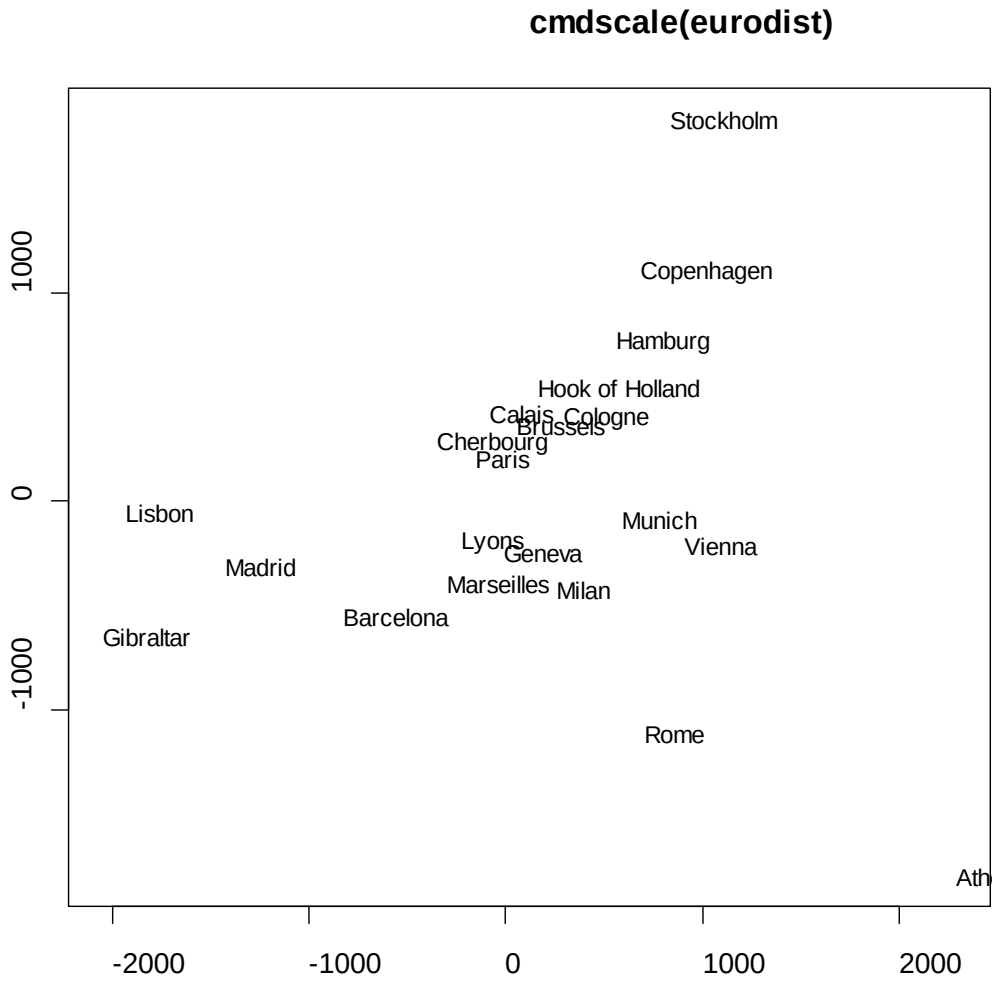
1. Se obtienen las disimilaridades δ_{ij}.
2. Se calcula la matriz $A = -\delta_{ij}^2 / 2$
3. Se calcula la matriz $B = HAH$
4. Se obtienen (n-1) los valores propios $\lambda_1, \ldots, \lambda_{n-1}$ asociados a los correspondientes vectores propios, $v_1, \ldots, v_{n-1}$, normalizados de forma que $v_i^t v_i = \lambda_i$, $\forall i = 1, \ldots, n-1$. Si B es no es semidefinida positiva entonces o bien se ignoran los valores propios negativos o se añade una constante aditiva c a las disimilaridades tal y como se describirá a continuación, volviendo al paso 2.
5. Se elige la dimensión apropiada $k < p$.
6. Las coordenadas de los n puntos en el espacio Euclídeo de dimensión K, vendrán dadas por los K vectores propios asociados a los K mayores valores propios. Así, $x_{ik} = v_{ik}$, $i = 1, \ldots, n$, $k = 1, \ldots, K$.

Ejemplo 2.1: El conjunto de datos “eurodist” incorporado en el paquete “stats” de R contiene las distancias por carretera medidas en Kilómetros entre 21 ciudades europeas los datos de distancias entre capitales europeas. Vamos a efectuar un MDS clásico en dimensión 2 usando R. Para ello bastará con ejecutar el siguiente fichero de sintaxis.

```
#Cargamos los datos.
library(stats)
eurodist
datos=eurodist
#Se comprueba que B no es semidefinida positiva.
Xfull=cmdscale(datos, k = 20, eig = TRUE, add = FALSE, x.ret = FALSE)
#La solución en dos dimensiones explica aproximadamente el 87%.
X=cmdscale(datos, k = 2, eig = TRUE, add = FALSE, x.ret = FALSE)
X
#Se representa la configuración después de hacer una reflexión en y.
require(graphics)
x = X$points[,1]
```

```
y = X$points[,2]
plot(x, -y, type="n", xlab="", ylab="", main="cmdscale(eurodist)")
text(x, -y, rownames(X$points), cex=0.8)
```

La representación obtenida en dos dimensiones se aprecia en la siguiente figura.



Tal y como puede apreciarse, la representación obtenida mediante distancias por carretera (salvo algunas excepciones) se parece al mapa geográfico real.

Ejercicio 2.3:
Usando R o MatLab, construye un fichero de sintaxis mediante el cual se efectúe MDS clásico sobre los datos de las ciudades de la Tabla 1, determinando la dimensión adecuada para la representación.
Compara los resultados con los obtenidos en el Ejemplo anterior.

El problema de la constante aditiva

Si B no es semidefinida positiva, puede añadirse una constante a todas las disimilaridades excepto a los valores diagonales (δ_{ii}), que hará que B sea semidefinida positiva. Este es el problema conocido como el de la constante aditiva (Calliez(1983)).

Si las disimilaridades están medidas en una escala de razón existe una relación natural entre éstas y las distancias Euclídeas, pero si están medidas en una escala de tipo intervalo, en las que no hay un origen definido, entonces dicha relación no existe. El problema de la constante aditiva consiste en estimar la constante c de forma que $\delta_{rs} + c(1-\delta_{KR}^{rs})$ puedan considerarse datos de tipo razón, siendo δ_{KR}^{rs} el delta de Kroneker.

Una solución sencilla al problema puede encontrarse fácilmente si se añade una constante a las disimilaridades al cuadrado en lugar de a las disimilaridades en sí mismas. Así

$$\delta_{rs}^{2(c)} = \delta_{rs}^2 + c(1-\delta_{KR}^{rs}) .$$

El menor valor de *c* que hace que B sea semidefinida positiva será $-2\lambda_n$, donde λ_n es el menor valor propio de B.

Ejercicio 2.2.
Sobre los datos de **eurodist**, construir un fichero de sintaxis en R o MatLab que permita realizar MDS clásico después de transformar las distancias mediante el procedimiento de la constante aditiva. Construye la representación en dos dimensiones y compárala con la anterior. ¿Cuántas dimensiones resultan necesarias para explicar los datos?

Análisis de coordenadas principales

Hasta ahora se ha tratado la matriz de distancias D entre n objetos como el punto de partida en nuestro análisis. No obstante, en muchas situaciones se parte de una matriz $X_{n\times p}$ y resulta necesario efectuar una elección de la función de distancias. Existen muchas posibilidades para ello, siendo la más simple utilizar la distancia Euclídea, en cuyo caso existe una estrecha conexión con el Análisis en Componentes Principales (ACP).

Sea $X_{n\times p}$ una matriz de datos y sean $\lambda_1 \le \cdots \le \lambda_p$ los valores propios de la matriz $nS = X^{'}HX$, donde S será la matriz de covarianza muestral. Supondremos sin pérdida de generalidad que los valores propios son todos no nulos y distintos. Así, estos valores propios también lo serán de la matriz $B = HXX^{'}H$. Nótese que las filas de *HX*

son las filas centradas de la matriz X, por lo que B no es más que la matriz centrada de productos escalares,

$$b_{rs} = (x_r - \overline{x})'(x_s - \overline{x}).$$

Si denotamos por v_i al i-ésimo vector propio de B, normalizado por $v_i'v_i = \lambda_i$, entonces para un valor K fijo, con $1 \leq k \leq p$, las filas de la matriz por columnas $V_k = [v_1, \ldots, v_k]$ son denominadas las **coordenadas principales** de X en K dimensiones.

Así, por el Teorema clásico, si D es la matriz de distancias Euclídeas entre las filas de X, entonces la solución clásica del MDS en dimensión K viene dada por las coordenadas principales de X en K dimensiones.

Bibliografía

- Calliez, F. (1983). The analytical solution to the additive constant problem. *Psychometrika*, 48, 305-308.
- Mardia, K.V., Kent, J.T. & Bibby, J.M. (1980).- Multivariate Analysis. *Academic Press*.
- Schönemman, P. H. (1972).- An algebraic solution for a class of subjective metrics models. *Psychometrika*, 37, 441-451.
- Torgerson, W. S. (1958). Theory and Methods of Scaling. *Wiley, New York.*
- Young, G. & Householder, A. S., (1938).- Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3, 19-22.