Análisis de datos. Técnicas aplicadas a datos de proximidad

Tema 3: MDS métrico y no métrico.

Introducción

Los modelos de MDS requieren que cada valor de proximidad sea representado exactamente por su correspondiente distancia. No obstante, en la práctica la presencia de errores de medida incluso en las distancias hace que la relación de "igualdad" sea relajada por la de "aproximadamente igual". Así, una medida de la bondad de ajuste es el error total cometido en la aproximación, de modo que no resulta necesaria la representación exacta, bastando una buena aproximación de la solución. Por tanto, la estimación de los parámetros del modelo se realiza empleando el usual concepto de error estadístico, para el cual los procedimientos más empleados de mínimos cuadrados y máxima verosimilitud han sido los más empleados. Así, en MDS se hace corresponder una disimilaridad δ_{ij} con una distancia $d_{ij}(X)$ obtenida en un espacio X, mediante una función monótona f, de forma que $f(\delta_{ij}) \approx d_{ij}(X)$. Esa función determina el modelo particular de MDS.

Desde un punto de vista exploratorio, se define el **STRESS** bruto (σ_r) como una medida de bondad de ajuste dada por la suma de los errores de la representación al cuadrado,

$$\sigma_r(X) = \sum_{ij} e_{ij}^2 = \sum_{ij} [f(\delta_{ij}) - d_{ij}(X)]^2 ,$$

donde $f(\delta_{ij})$ es una medida de disimilaridad y $d_{ij}(X)$ la distancia entre los correspondientes puntos de la configuración X mediante la cual es aproximada la disimilaridad.

La raíz cuadrada de ese valor, normalizado por la suma de las distancias al cuadrado es lo que se denomina STRESS-1 o Stress de Kruskal (Kruskal, 1964ª). Minimizar el STRESS-1 requiere encontrar una configuración X óptima en dimensión k. Si f es una función paramétrica, los valores de la función también deberán ser estimados. Así en MDS de tipo intervalo, los parámetros son estimados mediante regresión lineal, mientras que el MDS de tipo ordinal se utiliza regresión monótona.

El método de estimación empleado y las diferentes medidas de bondad de ajuste, son los principales factores que determinan la existencia de los diferentes métodos de MDS que se conocen en la actualidad. Uno de los procedimientos más empleado es el de mínimos cuadrados y en particular, el procedimiento SMACOF (Scaling by Majorizing a Complicated Function) desarrollado por de Leeuw and Heiser (1980).

MDS métrico. El método SMACOF

Dada una matriz de disimilaridades $\Delta=(\delta_{ij})$ entre *n* objetos *(i,j=1,2,...,n)*, SMACOF determina la configuración X de orden *n* x *K*, de tal manera que las distancias Euclídeas $d_{ij}(X)$ medidas entre los puntos de la representación obtenida sean lo más parecidas posible a las disimilaridades δ_{ij} de inicio $\forall$ *i,j = 1…n*. Formalmente, el problema que resuelve SMACOF en el caso más general es la minimización de la función de pérdidas de mínimos cuadrados STRESS, de Kruskal (1964),

$$\sigma(X)=\sum_{i<j} w_{ij}(\delta_{ij}-d_{ij}(X))^2 ,$$

siendo

- $d_{ij}\left(X\right)^2=\left(x_i-x_j\right)^t\left(x_i-x_j\right)$, con x_i y x_j, vectores columna de orden *K* x *1*, definidos a partir de las filas *i* y *j* de la matriz de X.

- $W=(w_{ij})$, una matriz conocida de pesos no negativos, que se asumirá simétrica y no negativa. Estas ponderaciones permiten la manipulación de datos faltantes (en el caso de faltar alguna medida de disimilaridad, la ponderación en tal caso vale 0), aunque también puede se pueden asignar pesos positivos a las disimilaridades.

Siguiendo a de Leeuw (1977a), la función STRESS puede descomponerse de la forma:

$$\sigma(X)=\sum_{i<j} w_{ij}(\delta_{ij}-d_{ij}(X))^2 =\sum_{i<j} w_{ij}\delta_{ij}^2+\sum_{i<j} w_{ij}d_{ij}^2(X)-2\sum_{i<j} w_{ij}\delta_{ij}d_{ij}(X)$$
$$=\eta_\delta^2+\eta^2(X)-2\rho(X)$$

Bajo ciertas condiciones se comprueba que el STRESS queda acotado superiormente de la forma:

$$\sigma(X)\le n(n-1)/2+\mathbf{tr}X^tVX-2\mathbf{tr}X^tB(Y)Y ,$$

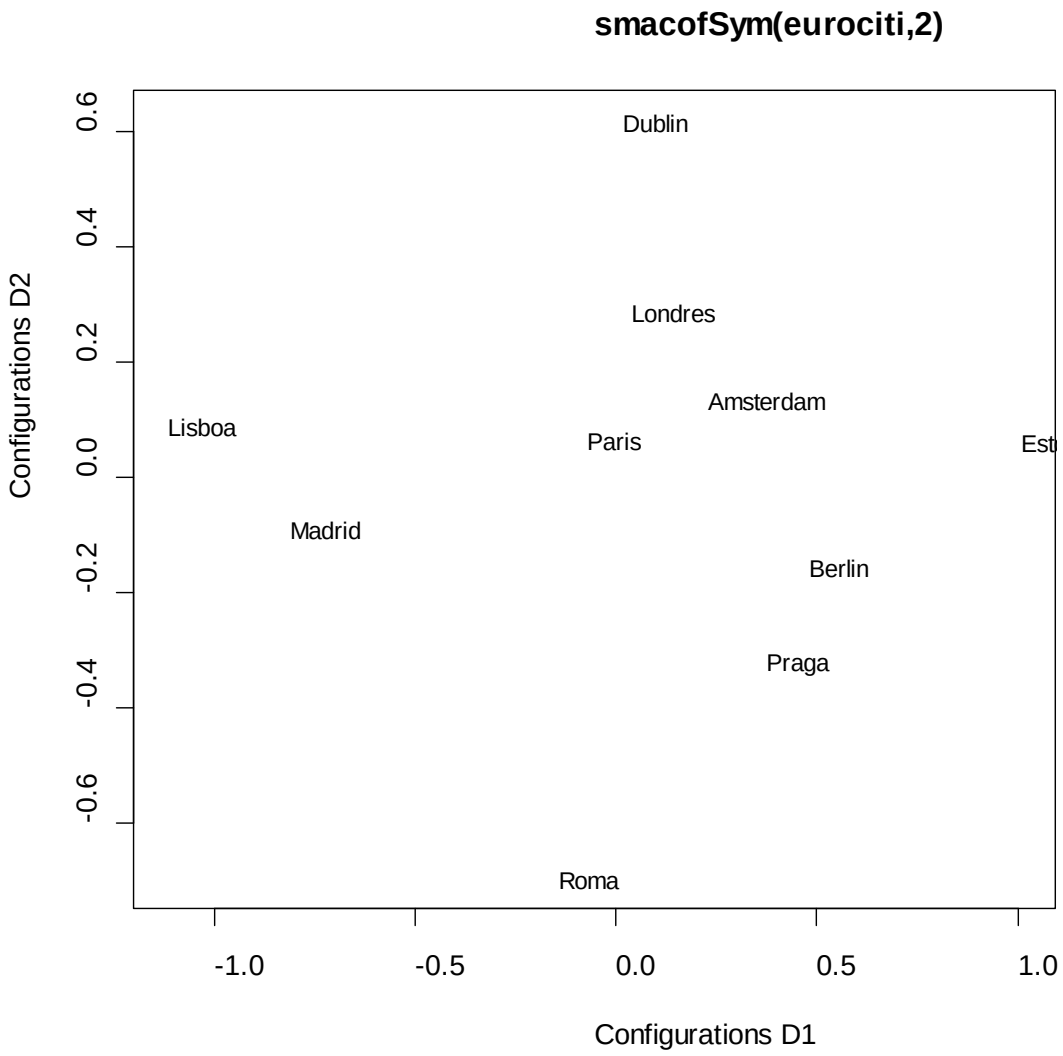
donde $V=\sum_{i<j} w_{ij}A_{ij}$, siendo $Aij=(e_i-e_j)(e_i-e_j)^t$, $B(Y)=\sum_{i<j} w_{ij}s_{ij}(Y)A_{ij}$, donde $s_{ij}=\delta_{ij}/d_{ij}$, si $d_{ij}>0$, y cero en caso contrario, e Y cualquier configuración. Así, la configuración óptima se obtiene a partir de otra configuración Y mediante la denominada transformación de Guttman (Gutman, 1968), de la forma:

$$X=V^+B(Y)Y$$

donde $V^+=(V+n^{-1}\mathbf{1}\mathbf{1}^t)^{-1}$ es la inversa de Moore-Penrose de V. Puesto que el procedimiento de mayorización es iterativo, dada una configuración inicial se obtienen nuevas configuraciones hasta alcanzar el criterio de convergencia.

Ejemplo 3.1
Consideremos los datos (**eurocitis**) de distancias entre 10 ciudades de la Tabla 1. En primer lugar se leen los datos y se calculan los valores propios mediante la solución clásica, apreciándose que los datos no son distancias Euclídeas. A continuación hacemos MDS métrico usando SMACOF.

```
datos=read.table("european_cities1.dat",header=TRUE,sep="",quote="")
eurociti=as.dist(datos) #Lo ponemos en la clase dist.
cmdscale(eurociti, k = 9, eig = TRUE, add = FALSE, x.ret = FALSE)$eig
#Hacemos MDS métrico.
resm.eurociti=smacofSym(eurociti,2,)
resm.eurociti
summary(resm.eurociti)
plot(resm.eurociti, main="smacofSym(eurociti,2)")
```



Tal y como puede apreciarse, la configuración se ajusta bastante bien a la

realidad.

Ejercicio 3.1 Efectuar MDS métrico usando SMACOF par los datos **eurodist**. Compara los resultados con los obtenidos mediante el procedimiento clásico.

MDS no métrico con SMACOF

En la función de pérdida anterior, no se observa ninguna transformación para las disimilaridades δ_{ij}. Si las disimilaridades están en una escala ordinal, podemos pensar en una transformación que preserve el orden de éstas. Si además esa transformación, sólo obedece a la restricción de monotonía, es decir, $\delta_{ij} < \delta_{rs} \Rightarrow f(\delta_{ij}) < f(\delta_{rs})$, entonces hablamos de MDS no métrico.

Los valores de disimilaridad transformados, $\hat{\delta}_{ij} = f(\delta_{ij})$ son denominados *disparidades* y deberán ser estimadas mediante regresión monótona. Así, la función STRESS adopta en este caso de la expresión,

$$\sigma(X, \hat{D}) = \sum_{i<j} w_{ij}(\hat{\delta}_{ij} - d_{ij}(X))^2,$$

que tendrá que ser minimizada con respecto a X y a $\hat{D}$. En el algoritmo de mayorización representa un nuevo ciclo iterativo después de la etapa relativa a la transformada de Guttman. Así, en la iteración t, si el orden de $d_{ij}(X^t)$ es el mismo que $\hat{\delta}^{t-1}$ entonces la actualización óptima es $\hat{\delta}^t = d_{ij}(X^t)$, si no, la actualización óptima se realiza mediante regresión monótona.

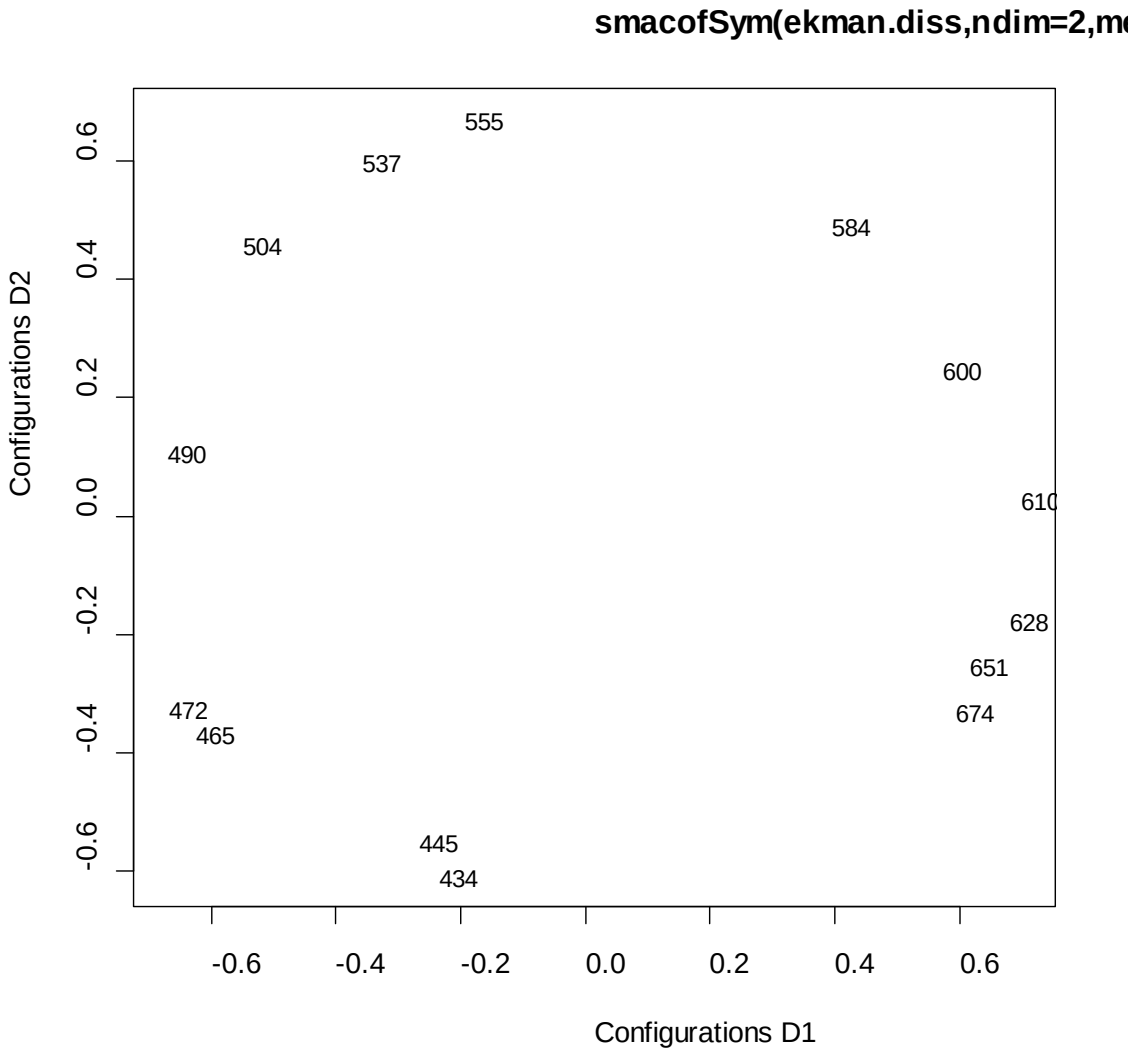
Para ello, en primer lugar es necesario considerar el caso de empates (tied) en la matriz de disimilaridades ordinales Δ, es decir, cuando $\delta_{ij} = \delta_{rs}$. En ese caso pueden distinguirse tres aproximaciones: la aproximación *primaria*, en la que no resulta necesario que $d_{ij}(X^t) = d_{rs}(X^t)$, en contra de lo que ocurre en la aproximación *secundaria*. En la aproximación *terciaria*, solo se requiere que las medias de los bloques de empates estén en el orden correcto.

La implementación en R de SMACOF resuelve el problema de la regresión monótona mediante el algoritmo PAVA (*pooled-adjacent-violators algorithm:* Ayer, Brunk, Ewing, Reid and Silverman 1955; Barlow, Bartholomew, Bremner, and Brunk 1972). Este algoritmo realiza regresión monótona mediante el uso de medias ponderadas.

Ejemplo 3.2 (Datos de color de Ekman)
Ekman (1954) analizó los datos de similaridad entre 14 colores (con longitudes de onda desde 434 a 674 nm). Las similaridades se obtuvieron mediante la clasificación por parte de 31 individuos de los pares de colores en una escala de 5 modalidades (0=no similares hasta 5=idénticos). Después de

promediarlas, las similaridades obtenidas fueron divididas por 4 para que estuviesen en el intervalo unidad. Usando la función smacofSym(), vamos a efectuar un MDS métrico en dos dimensiones para los datos de Ekman. Para ello, en primer lugar se convierten las similaridades en disimilaridades de la forma $\delta_{ij} = 1 - s_{ij}$, mediante la función sim2diss().

```
data(ekman)
ekman.diss=sim2diss(ekman,1)
resnm.ekman=smacofSym(ekman.diss,ndim=2, type="ordinal")
resnm.ekman
summary(resnm.ekman)
plot(resnm.ekman,main='smacofSym(ekman.diss,ndim=2, type="ordinal")')
```



En la configuración anterior se aprecia cómo las longitudes de onda se sitúan en forma circular. Así, puede verse como los pares 434-445 y 465-472 se

corresponden con colores azulados, 490 con turquesa, el conjunto 504-555 con verdosos, 584-610 con amarillos-naranjas y finalmente 628-674 con rojizos.

Ejercicio 3.2:
Efectuar un análisis no métrico de los datos **eurocitis** de la Tabla 1 usando SMACOF. Compara los resultados obtenidos con los de la solución métrica.

Bibliografía

- Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E (1955). An Empirical Distribution Function for Sampling with Incomplete Information." The Annals of Mathematical Statistics, 26, 641{647.

- Barlow RE, Bartholomew RJ, Bremner JM, Brunk HD (1972). Statistical Inference Under order Restrictions. John Wiley & Sons, New York.

- de Leeuw J (1977a). Applications of Convex Analysis to Multidimensional Scaling." In JR Barra, F Brodeau, G Romier, B van Cutsem (eds.), Recent Developments in Statistics, pp. 133{145. North Holland Publishing Company, Amsterdam.

- de Leeuw J, Heiser WJ (1980). Multidimensional Scaling with Restrictions on the con_guration. In P Krishnaiah (ed.), Multivariate Analysis, Volume V, pp. 501{522. North Holland Publishing Company, Amsterdam.

- Ekman G. (1954). Dimensions of Color Vision." Journal of Psychology, 38, 467-474.

- Guttman L (1968). \A General Nonmetric Technique for Fitting the Smallest Coordinate Space for a Con_guration of Points." Psychometrika, 33, 469{506.

- Kruskal, J. B. (1964).- Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika,* 29, 1-28, 115-129.