Pub. Mat. UAB
Nº 22 Nov. 1989
Actes VII JMHL

UNA CONTRIBUCION AL ANALISIS DE PROXIMIDADES

Carles M. Cuadras, Carmen Ruiz-Rivas

Dpto. de Bioestadística. Universidad de Barcelona
Dpto. de Estadística. Universidad Autónoma de Madrid

1. COORDENADAS PRINCIPALES

Sea $\Delta = (\delta_{ij})$ una matriz de distancias no euclideas sobre un conjunto finito $\Omega = \{1,..,i...,j,..,n\}$. Se llama preordenación sobre Ω asociada a δ_{ij} a la relación binaria definida en $\Omega x \Omega$:

$$(i,j) \preccurlyeq (i',j') \quad \text{si y sólo si} \quad \delta_{ij} \leq \delta_{i'j'}$$

Sean $P_1,...,P_n$ puntos del espacio euclídeo R^m de coordenadas euclídeas $X = (x_{ij})$ y distancia

$$d^2_{ij} = d^2(P_i,P_j) = \sum_{h=1}^{m} (x_{ih}-x_{jh})^2 .$$

Se dice que la configuración X realiza la preordenación si se verifica

$$d_{ij} \leq d_{i'j'} \quad \text{si y sólo si} \quad \delta_{ij} \leq \delta_{i'j'} .$$

La obtención de una configuración que realice total o aproximadamente una preordenación es un problema clásico de Análisis de Proximidades. El análisis de coordenadas principales (Gower,1966) puede sintetizarse en el siguiente resultado:

<u>Teorema 1</u>: Sean I_n (matriz identidad), $E=(1,\overset{(n}{\ldots},1)'$, $H=I_n - \frac{1}{n}E.E'$, $A=(a_{ij})$, con $a_{ij}=-\frac{1}{2}\delta^2_{ij}$, y $B = HAH$. Supongamos B semid. positiva de rango k ($\leq n-1$), sean $\lambda_1,...,\lambda_k$ los valores propios positivos y

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \qquad \lambda_i = \sum_{h=1}^{n} x^2_{hi}$$

la matriz de vectores propios (por columnas) λ-normalizados. Entonces X es una configuración euclídea, con centroide en el origen, cuyas distancias euclídeas reproducen exactamente δ_{ij}.

El anál. de coord. princ.(ACP) es el mejor método para reproducir, en dimensión reducida d, una distancia euclídea. El ACP toma las d primeras coordenadas X_d de X. Se verifica $B = X'X$, mientras que $B^* = X_d'X_d$ representa una aproximación a B en el sentido de los mínimos cuadrados. Eckart y Young(1936) demuestran que este mínimo es

$$D = \min \operatorname{tr}(B - B^*)^2 = \sum_{i=d+1}^{k} \lambda_i^2 \quad .$$

D se toma como una medida de distorsión al representar en dimensión d .

2. SOLUCION DE LINGOES

Si δ_{ij} no es euclídea B tiene valores propios positivos y negativos y al menos uno nulo. Sean estos

$$\lambda_1 \geqslant \ldots \geqslant \lambda_r \geqslant 0 > \lambda_1' \geqslant \ldots \geqslant \lambda_s' \qquad (1)$$

Tomando d coord. princ. $(d \leqslant r)$, la distorsión es

$$D = \sum_{i=d+1}^{r} \lambda_i^2 + \sum_{j=1}^{s} \lambda_j'^2 \qquad (2)$$

Por transformación monótona de δ_{ij} es posible obtener una distancia euclídea que realice la preordenación. Una solución de este tipo se debe a Lingoes (1971), que puede enunciarse en la forma siguiente:

Teorema 2: Con las mismas notaciones del teorema 1, sean (1) los valores propios de B y sea la matriz $A^* = A - \lambda_s'(I_n - \frac{1}{n} E.E')$. Se verifica:

1) La matriz $B^* = HA^*H$ es semidef. posit. de rango n-2.
2) Diagonalizando B^*, en la forma $B^* = X'.X$, se obtiene una configuración euclídea X que realiza la preordenación.
3) La distancia euclídea es

$$d_{ij}^{*2} = \delta_{ij}^2 - 2\lambda_s' \qquad \text{si } i \neq j ,$$
$$= 0 \qquad \text{si } i = j.$$

La solución de Lingoes tiene una distorsión $D = (n-1)\lambda_s'^2$, que en general será notablemente superior a (2). Es una solución que realiza exactamente la preordenación, pero a costa de incrementar la dimensión.

3. SOLUCION DE MARDIA

Si se considera la distorsión, no respecto a la distancia original, sino respecto a la distancia transformada

$$d_{ij}^{*2} = \delta_{ij}^2 - 2a \qquad \text{si } i \neq j ,$$
$$= 0 \qquad \text{si } i = j,$$

la distorsión es entonces (indicando por $\lambda_1, \ldots, \lambda_{n-1}$ los val. prop. de B)

$$D(a) = \sum_{i=d+1}^{n-1} (\lambda_i - a)^2$$

Mardia(1978) propone tomar como valor de a el que hace mínimo D(a), es decir,

$$a_d = \left(\sum_{i=d+1}^{n-1} \lambda_i \right) / (n-d+1)$$

Esta solución proporciona menos distorsión que la de Lingoes, la cual toma $a = \lambda_s'$. Sin embargo, para d igual a 2 , 3 u otros valores pequeños, la solución a veces no existe (d_{ij}^* resulta negativa).

4. SOLUCION LINEAL

En el presente trabajo se propone la transformación lineal

$$\hat{d}_{ij} = a\delta_{ij} + b \qquad \text{si } i \neq j ,$$
$$= 0 \qquad \text{si } i = j,$$

a fin de obtener una realización euclídea de la preordenación. Sin embargo, la diagonalización de B(a,b) relativa a $\hat{d}_{ij}$ es mucha más compleja que la diagonalización de B(a) relativa a d_{ij}^*. Como alternativa para resolver este problema, se propone el siguiente algoritmo iterativo:

Paso 1: Se ajusta δ_{ij} a una distancia euclídea $d_{ij}^{(0)}$ por ACP y se obtiene $\hat{d}_{ij}^{(1)}$ por transformación monótona lineal de δ_{ij}

$$\hat{d}_{ij}^{(1)} = a_1\delta_{ij} + b_1$$

con la condición de que $G_1 = \sum_{i<j} (\hat{d}_{ij}^{(1)} - d_{ij}^{(0)})^2$ sea mínimo.

Paso N: Se ajusta $\hat{d}_{ij}^{(N)}$ a una distancia euclídea $d_{ij}^{(N)}$ por ACP y se obtiene $\hat{d}_{ij}^{(N+1)}$ por transformación monótona lineal

$$\hat{d}_{ij}^{(N+1)} = a_N \hat{d}_{ij}^{(N)} + b_N$$

con la condición de que $G_N = \sum_{i<j} (\hat{d}_{ij}^{(N+1)} - d_{ij}^{(N)})^2$ sea mínimo.

Se verifica

$$G_1 \geq G_2 \geq \dots \geq G_N \geq \dots$$

tomándose como solución aquella configuración euclídea $X^{(N)}$, con distancias $d_{ij}^{(N)}$ para la que G_N se estabiliza. Los coeficientes a_N y b_N que minimizan G_N son

$$\hat{a}_N = \frac{\sum_{i<j} \hat{d}_{ij}^{(N)} \cdot d_{ij}^{(N)} - p\bar{\delta}\cdot\bar{d}}{\sum_{i<j} (d_{ij}^{(N)})^2 - p\,\bar{\delta}^2} \quad , \quad \hat{b}_N = \bar{d} - a_N\bar{\delta} ,$$

siendo $\quad p = n(n-1)/2$, $\bar{\delta} = \frac{1}{p}\sum_{i<j} d_{ij}^{(N)}$, $\bar{d} = \frac{1}{p}\sum_{i<j} d_{ij}^{(N)}$.

Si los coeficientes hallados verifican

$$\hat{d}_{ij}^{(N+1)} = \hat{a}_N \hat{d}_{ij}^{(N)} + \hat{b}_N < 0$$

para algún $\hat{d}_{ij}^{(N)}$, la solución no es válida. Si $\hat{d}_{min}$ es el menor de los $\hat{d}_{ij}^{(N)}$, bastará minimizar G_N sujeto a la condición

$$a_N \hat{d}_{min} + b_N = 0.$$

La solución es entonces

$$\hat{a}_N = \frac{\sum_{i<j} d_{ij}^{(N)} (\hat{d}_{ij}^{(N)} - \hat{d}_{min})}{\sum_{i<j} (\hat{d}_{ij}^{(N)} - \hat{d}_{min})^2} \quad , \quad \hat{b}_N = -\hat{a}_N . \hat{d}_{min} .$$

5. CONCLUSIONES

Esta solución lineal ha sido ensayada con ejemplos numéricos y aplicaciones prácticas(distancias genéticas no euclídeas), con referencia a cuatro medidas diferentes de distorsión, llegándose a las siguientes conclusiones:

1) En general son necesarias pocas iteraciones para que G_N se estabilice.
2) La distorsión obtenida en dimensión reducida resulta sensiblemente inferior a las soluciones de Lingoes, Mardia y por coordenadas principales.
3) En dimensión n-2 se obtiene una realización euclídea de la preordenación asociada a δ_{ij}.

6. BIBLIOGRAFIA

Eckart, C., Young, G. (1936) The aproximation of one matrix by another of lower rank. Psychometrika,1,211-218.

Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53,325-338.

Lingoes, J.C. (1971) Some boundary conditions for a monotone analysis of symmetric matrices. Psychometrika, 36, 195-203.

Mardia, K.V. (1978) Some properties o f classical multi-dimensional scaling. Commun.Statist.-Theor.Meth.,A7(13), 1233-1241.