ANÁLISIS DE TABLAS DE DISIMILITUDES REPRESENTACIÓN GEOMÉTRICA DE LA POBLACIÓN

Gérard Defives
A.D.E.M.
Grupo de Análisis de Datos y
Estadísticas Multidimensional
Escuela Básica de Ingeniería
Universidad de Los Andes

RESUMEN. Dada una población, cuyos individuos no se pueden describir mediante medidas cuantitativas, queda posible definir similitudes o disimilitudes entre los objetos a comparar, y luego se pueden obtener árboles de clasificación jerárquica, árbol máximo (o mínimo), grafo G3, filtrante, etc. En este trabajo se propone una representación geométrica de esta población. Ésta se basa en la construcción de un producto escalar entre los objetos del estudio, lo cual, mediante la diagonalización de la matriz simétrica, semi-definida positiva, de dicho producto, desemboca sobre una representación de la población parecida a las representaciones que se obtienen en Análisis de Componentes Principales.

0 INTRODUCCIÓN

En muchas situaciones, se desea comparar objetos mediante el cálculo de índices de similitud o de disimilitud, lo cual conduce generalmente a aplicar métodos de clasificación jerárquica (Benzécri, 1973 y Roux, 1985) o derivados de la teoría de grafos (G3 y filtrantes), (Degenne y Vergès, 1973). Varios métodos han sido propuestos para obtener una traducción geométrica de

una tabla de similitudes o disimilitudes. Cuando los objetos a comparar son modalidades de variables cualitativas, el análisis de Correspondencias puede aplicarse (Benzécri, 1973 y Di Giacomo, 1981). Nótese que en el ejemplo tratado por Di Giacomo, el índice de similitud no es otra cosa que la comparación de los perfiles en un tabla de contingencia. En este caso resulta fácil relacionar Análisis de Correspondencias y Clasificación. No es el caso si sólo se dispone de las similitudes o disimilitudes.

Un método consiste en considerar la matriz de similitudes **S (nxn)** como la propia matriz del producto escalar entre los objetos. Este método lo expone Escoufier (Escoufier, 1979). El paso de disimilitudes a similitudes no presenta dificultades. También se puede operar en base a la fórmula de W.S. Torgerson (Rao, 1975). Estos métodos se justifican principalmente cuando los objetos a comparar presentan el carácter de variables. En efecto, los términos de la diagonal de **S** son todos la similitud máxima $\mathbf{s_m}$. Dividir por $\mathbf{s_m}$, conduce a una matriz análoga a la de las variables en Análisis de Componentes Principales (ACP).

En lo que sigue, se propone un método mejor adaptado a la representación de los individuos. En un artículo posterior, se describirá, de nuevo, este método propuesto y se ilustrará con un ejemplo.

1 ANÁLISIS DE UNA TABLA DE DISTANCIAS

Sean $\mathbf{\{A_i\}_{i=1...n}}$ **, n** puntos ponderados por pesos respectivos $\mathbf{p_i}$, tales que $\sum_i \mathbf{p_i = 1}$, para los cuales sólo se conocen las distancias entre puntos $\mathbf{d(A_i,A_j) = d(i,j)}$, distancia euclídea. Suponiendo la nube centrada en 0 (cero), se quiere reconstruir la nube de puntos en un espacio **q**-dimensional.

Al igual que en el ACP, se procede a partir de la matriz del producto escalar correspondiente a la distancia **d**. Al producto escalar está asociada la norma $\|X\|^2 = (X|X)$, y la distancia se define por d(A,B) = $\|OA - OB\|$.

Como

$$\|X - Y\|^2 = \|X\|^2 + \|Y\|^2 - 2(X|Y)$$

$$(X|Y) = \frac{1}{2}\left[\|X\|^2 + \|Y\|^2 - \|X - Y\|^2\right]$$

resulta que los productos escalares entre individuos se expresan en función de las distancias por:

$$(A_i|A_J) = \frac{1}{2}\left[d^2(o,i) + d^2(o,j) - d^2(i,j)\right] = W_{ij}$$

Pero en la situación presente, se conocen los **d(i,j)** distancias entre los individuos $\mathbf{A_i}$ y $\mathbf{A_j}$ pero no las distancias **d(o,i)** y **d(o,j),** distancias entre estos individuos y el centro de gravedad **0** (cero).

Para obtener estas distancias **d(o,i),** se puede proceder de la siguiente manera:

Sea $\mathbf{I_i}$ la inercia de la nube de puntos con respecto al punto $\mathbf{A_i}$.

$$I_i = \sum_{j=1}^{n} p_j\, d^2(i,j) = d^2(o,i) + I \qquad \text{(teorema de Huygens)} \qquad \mathbf{(1)}$$

donde **I** representa la inercia de la nube de puntos con respecto a su centro de gravedad **0** (cero).

$$\sum_{i=1}^{n} p_i\, I_i = \sum_{i=1}^{n} p_i\, d^2(o,i) + \sum_{i=1}^{n} p_i\, I$$

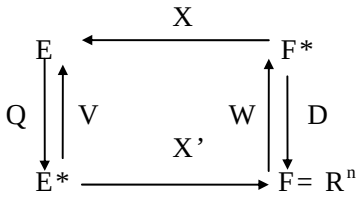
pero $\mathbf{I = \sum_{i=1}^{n} p_i\ d^2\,(o\,,\,i)}$ y $\mathbf{\sum_{i=1}^{n} p_i = 1}$, luego $\mathbf{\sum_{i=1}^{n} p_i\ I_i = 2\,I}$

$$\mathbf{I = \frac{1}{2}\sum_{i=1}^{n} p_i\ I_i} \qquad \mathbf{(2)}$$

se introduce en (1):

$$d^2\,(o\,,\,i) = I_i - I = I_i - \tfrac{1}{2}\sum_{j=1}^{n} p_j\ I_j$$

Una vez calculados los **d(o,i),** se puede calcular la tabla de los productos escalares $\mathbf{W = [w_{ij}]}$ y hacer jugar a esta matriz **W** el papel de la matriz $\mathbf{W=X'QX}$ del esquema de dualidad:



Si **d** es una distancia euclídea, la matriz **W** obtenida es simétrica, semidefinida positiva, y verifica, si **D** denota la matriz **(nxn)** diagonal de los pesos, y $\underline{1}$ el vector **(nx1)** cuyos términos son todos **1** (unos) **:** $\mathbf{WD\underline{1} = 0.}$

En efecto, la simetría resulta de la simetría de **d: d(i,j) = d(j,i),** y **W** sería definida positiva si no fuera por el centrado de la nuve de puntos, lo que trae como consecuencia $\mathbf{WD\underline{1} = 0.}$

$$\mathbf{2\sum_{i=1}^{n} w_{ij}p_i = \sum_{i=1}^{n}[p_i d^2(o,i) + p_i d^2(o,j) - p_i d^2(i,j)] = I + d^2(o,j) - I_j = 0}$$

En el caso presente, no se dispone de la matriz $\mathbf{X}_r$ que describe los individuos en un espacio vectorial $\mathbf{E = R^r}$, ni **X'** que describe las variables en **F**. Sin embargo, se puede

obtener un mapa que represente la población $\{A_i\}$ y que distorsione, lo menos posible, las distancias entre los individuos A_i.

Se trata entonces de hallar la matriz **Y (qxn)**, cuyas columnas representarán a los individuos en un espacio vectorial de dimensión **q**, referido a una base ortogonal, y tal que los productos escalares calculados entre estas representaciones **(Y'Y)** sean lo más cercano posible a la matriz **W**. Tomando en cuenta los pesos atribuidos a los individuos, esto consiste en maximizar:

$$RV(WD, Y'YD) = \frac{Tr(WDY'YD)}{\sqrt{Tr(WD)^2 Tr(Y'YD)^2}}$$

y la solución es la matriz **Y**, cuyas filas son los vectores propios de **WD** asociados a los **q** mayores valores propios (Escoufier, 1979). Se normalizarán estos vectores propios de tal forma que $YDY'=\Lambda$, la matriz diagonal **(qxq)** de los **q** mayores valores propios de **WD** para que la inercia de la proyección de la nube corresponda a la inercia de la nube. (**Y** cuando **q** = **p** = dimensión del más pequeño espacio vectorial que contiene la nube, la inercia inicial está totalmente reconstituida. Esta dimensión **p** es el número de valores propios de **WD** diferentes de cero (Defives, 1983).

2 ANÁLISIS DE UNA TABLA DE DISIMILITUDES

En la práctica, el proceso arrancará de una disimilitud, que no es una distancia. Si esta disimilitud no verifica la desigualdad del triángulo, puede resultar que algunos $d^2(o,i)$ sean negativos, o que la matriz **W** no sea positiva.

Por ejemplo, para tres puntos a,b,c, cuyas disimilitudes están dadas por la tabla siguiente:

d	a	b	c
a	0	1	2
b		0	4
c			0

provistos de pesos idénticos (1/3), se tienen las inercias:

$\mathbf{I_a}$ = 5/3 ; $\mathbf{I_b}$ = 17/3 ; $\mathbf{I_c}$ = 20/3, y la inercia con respecto al origen **0**, colocado en el centro de gravedad:

$$\mathbf{I = \frac{1}{6}(I_a + I_b + I_c) = 7/3}$$

Luego: $\mathbf{d^2(o,a)}$ = -2/3 ; $\mathbf{d^2(o,b)}$ = 10/3 ; $\mathbf{d^2(o,c)}$ = 13/3.

El paliativo a esta situación consiste en cambiar la disimilitud **d** por otro índice de disimilitud δ, mediante la agregación a todos los **d(i,j)**, fuera de la diagonal, de una misma longitud **l**, de tal forma que cualquiera sean **i, j, k**

$$\mathbf{d(i,k) + l \leq d(i,j) + l + d(j,k) + l}$$

o sea: $\mathbf{d(i,k) \leq d(i,j) + d(j,k) + l}$

El uso de la disimilitud $\mathbf{\delta(i,j) = d(i,j) + l}$ si $\mathbf{i \neq j}$, $\mathbf{\delta(i,i) = 0}$, conserva la misma ordenanza, o sea: las clasificaciones resultantes de **d** y δ son idénticas. Sólo los nudos están trasladados de una longitud **l**. Pero δ es una distancia, luego la matriz $\hat{\mathbf{W}} = \left[\hat{\mathbf{W}}_{\mathbf{ij}}\right]$, de los productos escalares asociados a δ es simétrica, semi-definida positiva, y $\hat{\mathbf{W}}\,\mathbf{D\underline{1}} = \mathbf{0}$.

Se puede observar que de no verificarse la desigualdad triangular, la matriz W no resulta definida positiva. En el caso del ejemplo, usando

$\mathbf{W_{ij} = \frac{1}{2}[d^2(o,i) + d^2(o,j) - d^2(i,j)]}$, se obtiene:

$$W = \frac{1}{6}\begin{bmatrix} -4 & 5 & -1 \\ 5 & 20 & -25 \\ -1 & -25 & 26 \end{bmatrix}$$

Si se agrega **l = 1** a todas las disimilitudes **d(i,j),** se obtiene la distancia **δ**:

δ	a	b	c
a	0	2	3
b		0	5
c			0

Ahora, $\mathbf{I_a}$ = 13/3 ; $\mathbf{I_b}$ = 29/3 ; $\mathbf{I_c}$ = 43/3 ; **I** =38/9

$\mathbf{\delta^2}$ **(o,a)** = 1/9 ; $\mathbf{\delta^2}$ (o,b) = 49/9 ; $\mathbf{\delta^2}$ (o,c) = 64/9

y el producto escalar está dado por la matriz:

$$\hat{\mathbf{W}} = \frac{1}{6}\begin{bmatrix} 1 & 7 & -8 \\ 7 & 49 & -56 \\ -8 & -56 & 64 \end{bmatrix}$$

$\hat{\mathbf{W}}$ es semi-definida positiva; $\hat{\mathbf{W}}$**D1 = 0.**

Se verifica también que tr $\hat{\mathbf{W}}$**D** = **I** = 38/9.

Esto es general, ya que se ha construido $\hat{\mathbf{W}}$ sobre δ que es una verdadera distancia.

Del punto de vista de las disimilitudes, agregar $\mathbf{L}$ a **d** para obtener δ, es perfectamente legítimo. Esto equivale a no tomar en consideración un elemento común a todos los individuos a comparar. La dificultad radica en hallar el más pequeño $\mathbf{l}$ tal que $\delta = \mathbf{d} + \mathbf{l}$ sea una distancia.

Para todo **i = 1,...,n**, y todo **k = i + 1,...,n**, y todo **j=l,...,n**, se calcula $\mathbf{l(i,j,k)} = \mathbf{d(i,k) - d(i,j) - d(j,k)}$ y se toma el mayor de los $\mathbf{l(i,j,k)}$ como $\mathbf{l}$.

Otra estrategia consiste en despreocuparse de hallar una distancia, y buscar directamente la matriz del producto escalar. Para esto, se diagonaliza la matriz **W**, calculada con la disimilitud **d**. Sea **k** el más pequeño valor propio (negativo) de **W**, la matriz **M = W - kI** es una matriz simétrica semidefinida positiva, porque si **X** es vector propio de **W** para algún valor propio **h** : $\mathbf{MX} = \mathbf{WX} - \mathbf{kX} = \mathbf{(h - k)X}$, luego **X** es vector propio de **M** para el valor propio **(h - k)**. Siendo **k** el más pequeño de los valores propios de **w**, $\mathbf{h-k \geq 0}$.

Si $\mathbf{m_{ij}}$ denota el término de la i-ésima fila y j-ésima columna de **M**, la distancia entre individuos, inducida por la matriz **M** es:

$$\delta^2(\mathbf{i},\mathbf{j}) = \mathbf{m_{ii}} + \mathbf{m_{jj}} - \mathbf{k} - 2\mathbf{w_{ij}}$$

$$\delta^2(\mathbf{i},\mathbf{j}) = \mathbf{w_{ii}} - \mathbf{k} + \mathbf{w_{jj}} - \mathbf{k} - 2\mathbf{w_{ij}} = \mathbf{d}^2(\mathbf{i},\mathbf{j}) - 2\mathbf{k}$$

$\boldsymbol{\delta}$**(i,j)** es una distancia que produce la misma ordenanza que la disimilitud **d** porque si:

$$\mathbf{d(i,j)} \ < \ \mathbf{d} \ \ \mathbf{(i',j')}$$

$$d^2(i,j) < d^2(i',j')$$

$$d^2(i,j) - 2k < d^2(i',j') - 2k$$

es decir: $$\delta^2(i,j) < \delta^2(i',j')$$

$$\delta^2(i,j) < \delta^2(i',j') \text{ ya que } \delta > 0$$

Pero **M** no verifica la relación $\mathbf{MD\underline{1}} = \mathbf{0}$ ya que $\mathbf{WD\underline{1} - kD\,\underline{1}} = \mathbf{-kD\,\underline{1}.}$

Esto se puede corregir fácilmente, porque si **M** es semi-definida positiva, la matriz $\hat{\mathbf{M}}=\mathbf{(I - \underline{11}'D)M(I-D\,\underline{11}')}$ lo es también. Además $\hat{\mathbf{M}}\mathbf{D\underline{1}}= \mathbf{0}$ y

$$\hat{m}_{ii} + \hat{m}_{jj} - 2\hat{m}_{ij} = m_{ii} + m_{jj} - 2m_{ij}$$

luego, al usar $\hat{\mathbf{M}}$ en lugar de **M**, resultan las mismas distancias δ.

Demostración:

$\mathbf{M = (I - \underline{11}'\,D)\,M\,(I - D\,\underline{11}') = M - \underline{11}'\,DM - MD\,\underline{11}' + \underline{11}'\,DM\,D\underline{11}'}$

$$\hat{m}_{ij} = m_{ij} - \sum_{k=1}^{n} p_k m_{kj} - \sum_{k=1}^{n} p_k m_{ik} + \sum_{k=1}^{n}\sum_{h=1}^{n} p_k p_h m_{kh}$$

de ahí sale: $\hat{m}_{ii} + \hat{m}_{jj} - 2\,\hat{m}_{ij} = m_{ii} + m_{jj} - 2\,m_{ij}$

Sustituyendo y observando que

$$\sum_{k=1}^{n} p_k m_{ki} = \sum_{k=1}^{n} p_k m_{ik}$$

por la simetría de M:

$$\sum_{j=1}^{n} p_j \hat{m}_{ij} = \sum_{j=1}^{n} p_j m_{ij} - \sum_{j=1}^{n}\sum_{k=1}^{n} p_j p_k m_{kj} - \sum_{j=1}^{n} p_j \sum_{k=1}^{n} p_k m_{ik} + \sum_{k=1}^{n}\sum_{h=1}^{n} p_k p_h m_{kh} = 0$$

lo que comprueba que $\hat{\mathbf{M}}\mathbf{D}\underline{\mathbf{1}} = \mathbf{0}$.

Por otra parte, **M** siendo simétrica, $\hat{\mathbf{M}}$ también porque $\mathbf{(I-\underline{1}\,\underline{1}'D)'=I - D\,\underline{1}\,\underline{1}'}$ **y** $\mathbf{M = P\,\Delta\,P^{-1} = P\,\Delta\,P'}$ conduce a $\hat{\mathbf{M}} = \mathbf{(I - \underline{l}\,\underline{1}'D)\,P\Delta P'\,(I - D\,\underline{1}\,\underline{1}')}$, o sea

$\mathbf{M = Q\,\Delta\,Q'}$ con $\mathbf{Q = (I - \underline{1}\,\underline{1}'D)P}$ **y** $\mathbf{Q^{-1} = Q'.}$

M y $\hat{\mathbf{M}}$ tienen los mismos valores propios, luego si **M** es semi-definida positiva, $\hat{\mathbf{M}}$ también lo es.

El primer método modifica las disimilitudes **d** agregándoles **l** para obtener la distancia δ, y el segundo método equivale a modificar $\mathbf{d^2}$ restándole **2 k**, (**k** siendo el más pequeño valor propio de **W**), para obtener δ^2 .

Ambos métodos desembocan sobre una matriz $\hat{\mathbf{W}}$ o $\hat{\mathbf{M}}$, simétricas, semi-definidas positivas, que verifican $\hat{\mathbf{W}}\mathbf{D}\underline{\mathbf{1}} = \mathbf{0}$, $\hat{\mathbf{M}}\mathbf{D}\underline{\mathbf{1}} = \mathbf{0}$, y pueden jugar el papel de la matriz de productos escalares entre individuos, cuya diagonalización conduce a la matriz **Y,** que representa los individuos en un espacio vectorial referido a una base ortogonal.

¿A cuál método dar la preferencia?

En el ejemplo, la disimilitud **d** conducía a una matriz **W**, no positiva, cuyos valores propios son 8,34 ; 0 ; -1.34. Luego la matriz **M** es:

$$\mathbf{M} = \frac{\mathbf{1}}{\mathbf{6}} \begin{bmatrix} \mathbf{4{,}05} & \mathbf{5} & \mathbf{-1} \\ \mathbf{5} & \mathbf{28{,}05} & \mathbf{-25} \\ \mathbf{-1} & \mathbf{-25} & \mathbf{34{,}05} \end{bmatrix}$$

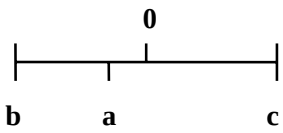
Resultan las distancias:

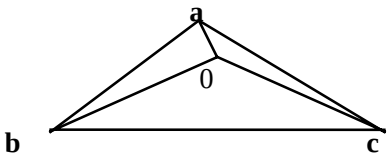
δ^2	**a**	**b**	**c**
a	0	3,68	6,68
b		0	18,68
c			0

δ	**a**	**b**	**c**
a	0	1,91	2,59
b		0	4,32
c			0

En este caso, las distancias δ obtenidas por el segundo método son más cercanas a las disimilitudes originales **d** que por el primer método. No se puede afirmar que será general. Restar el doble del menor valor propio de **W** de los cuadrados de las disimilitudes **d,** afecta menos a las disimilitudes altas que a las bajas, y esto es difícil de justificar. Sin embargo, como no afecta la ordenanza, no tiene mucha importancia visto lo que la selección de un índice de disimilitud contiene ya de arbitrario.

Las configuraciones geométricas resultantes de ambos métodos se dan a continuación:





3 CONCLUSIÓN

A partir de una tabla de disimilitudes **d(i,j)** se pueden calcular las **d(o,j)** que faltan para construir la matriz **W** cuyos términos son:

$$w_{ij} = \frac{1}{2}[\, d^2(o,i) + d^2(o,j) - d^2(i,j)]$$

Si **d** resulta ser una distancia euclídea, **W** es un producto escalar. De lo contrario, la construcción de la distancia δ por la agregación a **d** de la menor longitud **L** que permita verificar la desigualdad triangular, o restar de la diagonal de **W** el menor de sus valores propios negativos, permite definir sobre la población un producto escalar cuya matriz es $\hat{\mathbf{W}}$ en el primer caso, $\hat{\mathbf{M}}$ en el segundo.

Los vectores propios **Y** de $\hat{\mathbf{W}}\mathbf{D}$ o de $\hat{\mathbf{M}}\mathbf{D}$ según la táctica escogida, normados de tal forma que $\mathbf{YY'} = \Lambda$ (la diagonal de los **q** mayores valores propios) dan las **q** coordenadas de cada individuo en una representación **q**-dimensional de la población. Los criterios de apreciación de la calidad de la representación serán los mismos que en el ACP, en particular podrá considerarse la fracción de la inercia total reconstituida.

BIBLIOGRAFÍA

Benzécri J.P. (1973): **L'Analyse des données**. París, Dunod.

Defives G. (1983): "Inmersión de un conjunto de puntos en un espacio vectorial". **Sistemas** Mérida: ULA. No. 2. 6-8.

Degenne A. y Vergès P. (1973): "Introduction á l'analyse des similitudes". **Revue Fracaise de Sociologie**. XIV. 471-512.

Di Giacomo J.P. (1981): "Aspects méthodologiques de l'analyse des representations sociales". **Cahiers de Psychologie Congnitive.** 1-397-422.

Escoufier Y. (1979): **Cours d'Analuse des données.** Cap. 2. U.S.T.L. Montpellier.

Rao C.R. (1965): "The use and interpretation of P.C.A. in applied research". **Sankhya The Indian Journal of Statiscs**, A 26. 329-358.

Roux M. (1985): **Algorithmes de classification**. París: Masson.