# Análisis de datos. Técnicas aplicadas a datos de proximidad

Nazaret Pacheco Vázquez

Actividad 1

**Teorema 1.** *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 $H_{nxn}$ dada por $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^t$ y $A_{n\times n}$ la matriz cuyos elementos vienen dados por $a_{rs} = -\frac{1}{2}d_{rs}^2$. 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 $Z_{nxK} = (z_1, \ldots, z_n)^t \in \mathbb{R}^{n\times K}$, entonces $B = (HZ)(HZ)^t$, es decir, $b_{rs} = (z_r - \bar{z})(z_s - \bar{z})^t, \quad \forall r, s = 1, \ldots, n$, de donde $B \geq 0$. En este caso, $B$ 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 $X = (x^1, \ldots, x^K)$, normalizados según la condición*

$$(x^i)^t x^i = \lambda_i \ , \quad \forall i = 1, \ldots, K.$$

*Entonces, los puntos $K \in \mathbb{R}^K$ de coordenadas $x_r = (x_{r1}, \ldots, x_{rK})^t$, 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.*

*Demostración.* Vamos a denotar comor $\|\cdot\|$ la norma euclídea en $\mathbb{R}^K$ y por $\cdot$ el producto escalares usual.

**Propiedades y notación que vamos a tener en cuenta:**

- $H$ es simétrica y $H^2 = H$ (proyección ortogonal sobre el subespacio ortogonal a $\mathbf{1}$).
- $H\mathbf{1} = 0$.
- Para una matriz $M$, escribiremos $\mathrm{diag}(M)$ para el vector columna formado por las entradas diagonales de $M$.

Podemos dividir la demostración en dos implicaciones.

**(1)(⇒) Si $D$ es euclídea, es decir, proviene de puntos en $\mathbb{R}^K$ entonces $B \geq 0$.**

Supongamos que existen puntos $z_1, \ldots, z_n \in \mathbb{R}^K$ tales que $d_{rs} = \|z_r - z_s\|$ para todo $r, s$. Es decir, $d_{rs} := -2a_{rs} = (z_r - z_s)(z_r - z_s)$. Formemos la matriz

$$Z = \begin{pmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_n^T \end{pmatrix} \in \mathbb{R}^{n \times K}.$$

Definimos la matriz Gramiana $G = ZZ^T$, con $g_{rs} = z_r \cdot z_s$. Por la identidad del producto escalar, tenemos que

$$d_{rs}^2 = \|z_r - z_s\|^2 = g_{rr} + g_{ss} - 2g_{rs}.$$

y de aquí, obtenemos

$$a_{rs} = -\tfrac{1}{2} d_{rs}^2 = g_{rs} - \tfrac{1}{2} g_{rr} - \tfrac{1}{2} g_{ss}.$$

lo que en forma matricial quedaría como

$$A = G - \tfrac{1}{2}\big( \operatorname{diag}(G)\, \mathbf{1}^T + \mathbf{1}\, \operatorname{diag}(G)^T \big).$$

Ahora, multiplicando por $H$ en ambos lados y usando la propiedad que comentamos al principio de que $H\mathbf{1} = 0$, se anulan los términos lineales en $\mathbf{1}$:

$$B = HAH = HGH.$$

Pero tenemos que $G = ZZ^T$, así que, sustituyendo, nos quedaría que

$$B = HZZ^TH = (HZ)(HZ)^T.$$

La expresión $B = MM^T$ con $M = HZ$ muestra inmediatamente que $B$ es simétrica semidefinida positiva, es decir $B \succeq 0$. Además $\operatorname{rang}(B) \leq \operatorname{rang}(Z) \leq K$. Finalmente, las filas de $HZ$ son $\tilde{z}_r = z_r - \bar{z}$ con $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$, por lo que

$$B_{rs} = \tilde{z}_r \cdot \tilde{z}_s = (z_r - \bar{z}) \cdot (z_s - \bar{z}).$$

y esto es lo que queríamos demostrar para la primera implicación directa y la primera propiedad anunciada.

**(2)(⇐) Si $B \succeq 0$ (semidefinida positiva) entonces $D$ es euclídea.**

Supongamos ahora que $B = HAH$ es semidefinida positiva y que $\operatorname{rang}(B) = K$. Como $B$ es simétrica y semidefinida positiva, existe descomposición espectral

$$B = \Gamma \Lambda \Gamma^T,$$

donde $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, y $\Gamma$ ortonormal.
Sean $\lambda_1, \ldots, \lambda_K$ los autovalores estrictamente positivos y $v_1, \ldots, v_K$ los vectores propios

asociados.
Definimos $\Gamma_K$ como $\Gamma_K = [v_1 \ \cdots \ v_K] \in \mathbb{R}^{n\times K}$, $\quad \Lambda_K = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ y si denotamos

$$X = \Gamma_K \Lambda_K^{1/2} \in \mathbb{R}^{n\times K}$$

se cumple que $B = XX^T$.
Denotemos por $x_r \in \mathbb{R}^K$ la $r$-ésima fila de $X$. Entonces $B_{rs} = x_r \cdot x_s$. Consideremos la distancia euclídea entre $x_r$ y $x_s$:

$$\|x_r - x_s\|^2 = x_r \cdot x_r + x_s \cdot x_s - 2x_r \cdot x_s = B_{rr} + B_{ss} - 2B_{rs}.$$

Para comprobar que $\|x_r - x_s\| = d_{rs}$ calculamos la forma entrada a entrada de $B$. Si definimos las medias de $A$

$$\bar{a}_{r\cdot} = \tfrac{1}{n}\sum_{t=1}^{n} a_{rt}, \qquad \bar{a}_{\cdot s} = \tfrac{1}{n}\sum_{t=1}^{n} a_{ts}, \qquad \bar{a}_{\cdot\cdot} = \tfrac{1}{n^2}\sum_{u,t=1}^{n} a_{ut},$$

la expansión de $B = HAH$ da la conocida fórmula de elementos

$$B_{rs} = a_{rs} - \bar{a}_{r\cdot} - \bar{a}_{\cdot s} + \bar{a}_{\cdot\cdot}.$$

Entonces tenemos que

$$\begin{aligned} B_{rr} + B_{ss} - 2B_{rs} &= \big(a_{rr} - 2\bar{a}_{r\cdot} + \bar{a}_{\cdot\cdot}\big) + \big(a_{ss} - 2\bar{a}_{\cdot s} + \bar{a}_{\cdot\cdot}\big) - 2\big(a_{rs} - \bar{a}_{r\cdot} - \bar{a}_{\cdot s} + \bar{a}_{\cdot\cdot}\big) \\ &= a_{rr} + a_{ss} - 2a_{rs}. \end{aligned}$$

Ahora tenemos que los términos con medias se cancelan. Como $a_{rr} = -\frac{1}{2}d_{rr}^2 = 0$ (y análogamente $a_{ss} = 0$), se tiene que

$$B_{rr} + B_{ss} - 2B_{rs} = -2a_{rs} = d_{rs}^2.$$

Por tanto $\|x_r - x_s\|^2 = d_{rs}^2$ y, tomando raíces no negativas, $\|x_r - x_s\| = d_{rs}$. Esto muestra que las filas de $X$ son coordenadas en $\mathbb{R}^K$ que da como resultado exactamente la matriz de distancias $D$.
Además, la configuración está centrada, es decir, $B\mathbf{1} = HAH\mathbf{1} = HA0 = 0$, así que $B\mathbf{1} = 0$. Dado que $B = XX^T$, se tiene $XX^T\mathbf{1} = 0$ y, multiplicando $X^T$ por la izquierda, obtenemos

$$X^T X X^T \mathbf{1} = 0.$$

Como $\mathrm{rang}(X) = K$, la matriz $X^TX \in \mathbb{R}^{K\times K}$ es invertible, luego $X^T\mathbf{1} = 0$. Pero $X^T\mathbf{1}$ es exactamente la suma $\sum_{r=1}^{n} x_r$, por lo que la suma de las filas es cero, esto es, la configuración está centrada en el origen.

**Propiedades y observaciones**

- Puesto que $B\mathbf{1} = 0$, el vector $\mathbf{1}$ pertenece al núcleo de $B$ y por tanto $\mathrm{rang}(B) \leq n - 1$. La dimensión mínima necesaria para representar las distancias es $K = \mathrm{rang}(B)$.

- *Unicidad salvo isometrías.* Si $X, Y \in \mathbb{R}^{n\times K}$ son dos matrices cuyas filas dan configuraciones con la misma matriz de productos $XX^T = YY^T = B$ y además $\text{rang}(X) = \text{rang}(Y) = K$, entonces existen rotación/reflexión (matriz ortogonal) $Q \in O(K)$ tal que $Y = XQ$. De hecho, como $\text{col}(Y) \subseteq \text{col}(X)$ y recíprocamente (debido a $XX^T = YY^T$), existe una matriz invertible $R \in \mathbb{R}^{K\times K}$ con $Y = XR$. Sustituyendo en $YY^T = XX^T$ se obtiene $XRR^TX^T = XX^T$. Multiplicando por la izquierda y derecha por matrices apropiadas (o bien multiplicando por $X^T$ y usando que $X^TX$ es invertible) sigue que $RR^T = I_K$, es decir $R$ es ortogonal. Así $Y = XR$ con $R$ ortogonal, lo que expresa la unicidad de la configuración salvo isometrías (rotaciones y reflexiones).

Con esto queda demostrada la equivalencia y las propiedades anunciadas y concluimos nuestra demostración. □