Miscelánea Matemática **41** (2005) 1–7 SMM

Identificación de la dimensión en una matriz de distancias euclidianas

José M. González Barrios y Raúl Rueda

IIMAS, UNAM

Circuito Exterior

Ciudad Universitaria

04510 México, D.F.

México

pinky@sigma.iimas.unam.mx

A Juanjo

Resumen

Dada una matriz D de distancias euclidianas entre m puntos, donde la dimensión n de los puntos es desconocida, presentamos un algoritmo inductivo simple para determinar el valor de n. Con este algoritmo, siempre es posible encontrar m puntos cuya matriz de distancias siempre sea D. La única restricción que imponemos es que $m \geq n + 1$.

1. Introducción

Un problema clásico en estadística multivariada es el de escalamiento multidimensional y que puede ser resumido como sigue: Supongamos que tenemos un conjunto de vectores $\mathbf{x}_1, \ldots, \mathbf{x}_m$ de dimensión n y sus respectivas distancias $\delta_{rs} = \|\mathbf{x}_r - \mathbf{x}_s\|$ y tratamos de encontrar un conjunto de vectores k-dimensionales $\mathbf{y}_1, \ldots, \mathbf{y}_m$ tal que sus distancias $d_{rs} = \|\mathbf{y}_s - \mathbf{y}_r\|$ satisfagan $\delta_{rs} \approx d_{rs}$. En general los valores de las $\mathbf{x}_i$'s no se dan y δ_{rs} es sólo una medida de la proximidad entre los objetos r y s. La solución a este problema es lo que se conoce como escalamiento multidimensional. Cuando D es una matriz euclidiana, el problema ha

sido resuelto, véase por ejemplo Gower (1966), Mardia (1978) y Mardia *et al.* (1979). En Seber (1984) se tratan otros ejemplos.

De la teoría de escalamiento multidimensional, sabemos que existe una solución al problema de identificar la dimensión propuesto por Schoenberg y por Young y Householder, ver por ejemplo Mardia *et al.* (1979). Su método demuesta que si $D = (d_{ij})_{i,j\in\{1,2,\ldots,m\}}$ es una matriz euclidiana y si definimos $A = (-(1/2)d_{ij}^2)_{i,j\in\{1,2,\ldots,n\}}$ y

$$B := (I - \frac{1}{m}\mathbf{1}\mathbf{1}')A(I - \frac{1}{m}\mathbf{1}\mathbf{1}')$$

donde $I_{m\times m}$ es la matriz identidad y $\mathbf{1}' = (1, 1, \ldots, 1)$ es un vector de unos, entonces D es una matriz euclidiana si y solo si B es semi-definida positiva, y en este caso, si el rango de B es n sabemos que D es una matriz de distancias entre m puntos de dimensión n. Por supuesto, si m es muy grande, tendremos problemas computacionales.

Recientemente Brito, Quiroz y Yukich (2002) consideraron el problema de identificar la dimensión en la que una muestra de puntos vive, desde el punto de vista probabilista, en donde sólo las distancias entre los puntos son conocidas.

En este trabajo, proponemos una solución exacta y muy sencilla para identificar la dimensión correcta de una matriz de distancias euclidianas. Específicamente, en el caso en el que tengamos una muestra grande proveniente de una distribución contínua en Re^n, daremos un método inductivo para detectar la dimensión en la que los puntos viven.

Vamos a suponer que solamente conocemos la matriz D de m vectores $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ de dimensión n, donde no sólo las $\boldsymbol{x}_i$ son desconocidas, sino también desconocemos la dimensión n. En este caso, sólo tenemos una matriz D cuyas entradas están dadas por

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\| = \left(\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2\right)^{\frac{1}{2}}, \tag{1}$$

donde $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,n})$ para $i = 1, 2, \ldots, m$. En la siguiente sección veremos que podemos encontrar la dimensión n y un conjunto de m puntos para los que D es su matriz de distancias, suponiendo únicamente que tenemos al menos $n+1$ puntos. El procedimiento para encontrar la dimensión es inductivo.

2. Resultados principales y ejemplo

Analizaremos primero el caso más sencillo, dimensión uno. Supongamos que D es generada por $m \geq 3$ puntos en los reales. Tomemos $1 \leq i < j < k \leq m$, arreglando las entradas de la matriz D podemos suponer que $i = 1, j = 2$ and $k = 3$ y que las entradas respectivas son d_{12}, d_{13}, d_{23}; podemos también suponer, rearreglando si es necesario, que $d_{12} \geq \text{máx}\{d_{13}, d_{23}\}$. Dado que para cualesquier números reales $-\infty < a < b < c < \infty$, $|c - a| = |c - b| + |b - a|$, tenemos que

$$d_{12} = d_{13} + d_{23}. \tag{2}$$

Por lo tanto D es una matriz euclidiana de dimensión uno, si y sólo si para cualesquier tres entradas de D seleccionadas como arriba podemos encontrar una permutación de índices tal que (2) se satisface. Supongamos que hemos identificado que la dimensión de D es uno, para encontrar m puntos en Re tales que D sea su matriz de distancias, podemos suponer que $d_{12} \geq \text{máx}\{d_{13}, d_{23}\}$; entonces, si hacemos $x_1 = 0$, $x_2 = d_{12}$ y $x_3 = d_{13}$ tenemos que $\|x_i - x_j\| = d_{ij}$ para $1 \leq i, j \leq 3$, y entonces $x_i = \pm d_{1i}$ para $4 \leq i \leq m$, dependiendo de si $d_{2i} \leq d_{1i}$ o $d_{2i} > d_{1i}$.

Notemos que los puntos pueden estar en un espacio de dimensión mayor a uno, pero si encontramos que la dimensión de D es uno, significa que los puntos están localizados en un subespacio de dimensión uno.

El método descrito nos da la clave de cómo proceder en dimensiones mayores. Supongamos que tenemos $m \geq 4$ puntos $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ de dimensión dos y que D es su matriz euclidiana. consideremos las entradas d_{ij} para $1 \leq i, j \leq 4$, ya que la distancia euclidiana es invariante ante rotaciones y traslaciones, podemos tomar $\mathbf{x}_1 = (0, 0)$ y $\mathbf{x}_2 = (d_{12}, 0)$, y obviamente $\|\mathbf{x}_1 - \mathbf{x}_2\| = d_{12}$. Todo lo que tenemos que hacer ahora es encontrar puntos $\mathbf{x}_3 = (x_3, y_3)$ y $\mathbf{x}_4 = (x_4, y_4)$ en los que sus coordenadas puedan escribirse en términos de las entradas de la matriz. Para esto, notemos que las siguiente ecuaciones deben de satisfacerse para estos puntos

$$d_{13}^2 = x_3^2 + y_3^2 \quad \text{y} \quad d_{14}^2 = x_4^2 + y_4^2 \tag{3}$$

$$d_{23}^2 = d_{12}^2 + d_{13}^2 - 2d_{12}x_3 \quad \text{y} \quad d_{24}^2 = d_{12}^2 + d_{14}^2 - 2d_{12}x_4 \tag{4}$$

$$d_{34}^2 = d_{13}^2 + d_{14}^2 - 2x_3x_4 - 2y_3y_4. \tag{5}$$

resolviendo para x_3 y x_4 en la ecuación (4) obtenemos

$$x_3 = \frac{d_{12}^2 + d_{13}^2 - d_{23}^2}{2d_{12}} \quad \text{y} \quad x_4 = \frac{d_{12}^2 + d_{14}^2 - d_{24}^2}{2d_{12}}. \tag{6}$$

Ahora, usando x_3 y x_4 obtenemos y_3 y y_4 de la ecuación (3)

$$y_3 = \pm\sqrt{d_{13}^2 - x_3^2} \qquad y_4 = \pm\sqrt{d_{14}^2 - x_4^2}. \tag{7}$$

De hecho, la ecuación (5) es la clave para demostrar si los puntos están en un espacio euclideano de dimensión dos. Podemos decir que estos cuatro puntos pertenecen a Re^2 si

$$\text{mín}\{d_{34}^2 - \left(d_{13}^2 + d_{14}^2 - 2x_3x_4 \pm 2y_3y_4\right)\} = 0, \tag{8}$$

donde x_3, x_4 están dados por la ecuación (6) y y_3, y_4 se seleccionan a partir de (7) usando el signo positivo. Notemos también que si el mínimo es alcanzado con signo negativo, tendremos dos soluciones para las coordenadas y, esto es, $y_3 < 0$ y $y_4 > 0$, o $y_3 > 0$ y $y_4 < 0$, y si el mínimo es alcanzado con signo positivo, también tendremos dos soluciones para las coordenadas y, que son $y_3 < 0$ y $y_4 < 0$, o $y_3 > 0$ y $y_4 > 0$; en cada caso, obtenemos dos soluciones posibles. Si tenemos más de cuatro puntos, hagamos $\mathbf{x}_5 = (x_5, y_5)$ y a partir de $d_{15}^2 = x_5^2 + y_5^2$ y $d_{25}^2 = d_{15}^2 + d_{12}^2 - 2d_{12}x_5$, tendremos que $x_5 = (d_{15}^2 + d_{12}^2 - d_{25}^2)/2d_{12}$ y $y_5 = \pm\sqrt{d_{15}^2 - x_5^2}$. Entonces, habiendo asignado los signos de y_3 y y_4 podemos obtener el signo de y_5 usando d_{35} y d_{45}. Para los puntos restantes procedemos de la misma manera para obtener las coordenadas. Por lo tanto, siempre podemos obtener dos conjuntos diferentes de puntos tales que D es su matriz de distancias.

Para ver si D es una matriz de dimensión dos procedemos como sigue:

- Si encontramos tres entradas de D digamos d_{12}, d_{13} y d_{23}, donde $d_{1,2} \geq \text{máx}\{d_{13}, d_{23}\}$ tales que (2) no se cumple, sabemos que la matriz D no tiene dimensión dos.
- Si para cualquier d_{ij} donde $1 \leq i, j \leq 4$ para todos los posibles rearreglos de las entradas de D la ecuación (8) se cumple, entonces D es una matriz euclidiana de dimensión dos.

Notemos también que las ecuaciones en (7) proporcionan desigualdades no triviales que deben cumplir los puntos de Re^2:

$$d_{13}^2 \geq \left(\frac{d_{12}^2 + d_{13}^2 - d_{23}^2}{2d_{12}}\right)^2 \quad \text{y} \quad d_{14}^2 \geq \left(\frac{d_{12}^2 + d_{14}^2 - d_{24}^2}{2d_{12}}\right)^2. \tag{9}$$

Finalmente, si las coordenadas de los m puntos son generadas aleatoriamente de una distribución continua, es suficiente probar que la ecuación (8) se cumple para $d_{ij}, 1 \leq i, j \leq 4$.

El proceso usado en dimensión dos puede ser generalizado de manera muy sencilla a dimensiones mayores. Por ejemplo, en el caso de dimensión tres, si tenemos $m \geq 5$ puntos con matriz de distancias D, trasladando y rotando podemos suponer que $\mathbf{x}_1 = (0,0,0), \mathbf{x}_2 = (d_{12},0,0), \mathbf{x}_3 = (x_3, y_3, 0), \mathbf{x}_4 = (x_4, y_4, z_4)$ y $\mathbf{x}_5 = (x_5, y_5, z_5)$ y tendremos ecuaciones equivalentes a (3),(4) and (5) dadas por

$$d_{13}^2 = x_3^2 + y_3^2 \quad \text{y} \quad d_{1j}^2 = x_j^2 + y_j^2 + z_j^2 \text{ para } j = 4,5 \tag{10}$$

$$d_{2j}^2 = d_{12}^2 + d_{1j}^2 - 2d_{12}x_j \quad \text{para} \quad j = 3,4,5 \tag{11}$$

$$d_{3j}^2 = d_{13}^2 + d_{1j}^2 - 2x_3x_j - 2y_3y_j \quad \text{para} \quad j = 4,5, \tag{12}$$

y una ecuación extra dada por

$$d_{45}^2 = d_{14}^2 + d_{15}^2 - 2x_4x_5 - 2y_4y_5 - 2z_4z_5. \tag{13}$$

Por (11) tenemos que

$$x_j = \frac{d_{12}^2 + d_{1j}^2 - d_{2j}^2}{2d_{12}} \quad \text{para} \quad j = 3,4,5 \tag{14}$$

entonces, de (10) $y_3 = \pm\sqrt{d_{13}^2 - x_3^2}$ y usando (12) obtenemos

$$y_j = \frac{d_{13}^2 + d_{1j}^2 - d_{3j}^2 - 2x_3x_j}{2y_3} \quad \text{para} \quad j = 4,5. \tag{15}$$

De nuevo, por (10) tenemos que

$$z_j = \pm\sqrt{d_{1j}^2 - x_j^2 - y_j^2} \quad \text{para} \quad j = 4,5.$$

Por lo tanto, los puntos están en localizados en un espacio euclideano de dimensión tres si

$$\text{mín}\{d_{45}^2 - d_{14}^2 - d_{15}^2 - 2x_4x_5 \pm 2y_4y_5 \pm 2z_4z_5\} = 0, \tag{15}$$

donde x_4, x_5 están dadas por la ecuación (14), y_4, y_5 y z_4, z_5 se obtienen de (15) y (16) con signo positivo. En este caso, tendremos cuatro soluciones dependiendo de los signos de y_3 y de y_4.

por lo tanto, para ver si D es una matriz euclidiana de dimensión tres, procedemos como sigue:

- Si encontramos tres entradas de D, digamos d_{12}, d_{13} y d_{23}, donde $d_{1,2} \geq \text{máx}\{d_{13}, d_{23}\}$ tales que no cumplen con (2), sabremos que D no es una matriz de distancias con dimensión uno.

- Si encontramos seis entradas de D, digamos $d_{12}, d_{13}, d_{14}, d_{23}, d_{24}$ y d_{34} tales que (8) no se cumple, entonces D no es de dimensión dos.

- Si para cualquier d_{ij} donde $1 \leq i, j \leq 5$ se cumple (17) para todos los posibles arreglos de entradas de D, entonces D es una matriz de distancias euclidiana de dimensión tres.

Como en el caso de dimensión dos. la ecuación (16) genera desigualdades no triviales que deben de cumplir puntos de dimensión tres

$$d_{1j}^2 \geq \left(\frac{d_{1j}^2 + d_{ij}^2 - d_{2j}^2}{2d_{12}}\right)^2 + \left(\frac{d_{13}^2 + d_{1j}^2 - d_{3j}^2 - 2x_3x_j}{2\sqrt{d_{13}^2 - x_3^2}}\right)^2 \text{ para } j = 4, 5, \tag{18}$$

donde x_3 y x_j están dadas en (14).

Si tenemos más de cinco puntos, podemos resolver para los restantes puntos, como en el caso de dimensión dos, y obtendremos cuatro diferentes soluciones.

El procedimiento anterior puede ser extendido a cualquier dimensión. Aunque el proceso para encontrar la verdadera dimensión de la matriz de distancias es inductivo, puede ser implementado fácilmente, en especial, en el caso en que los vectores sean generados aleatoriamente de una distribución continua.

Para terminar, presentamos un ejemplo de seis puntos de dimensión tres generados de una uniforme $(0, 1)$, en el que seguiremos el proceso para detectar la dimensión y generación de puntos cuya matriz de distancia sea precisamente D. Los puntos generados fueron

$$\boldsymbol{x}_1 = (0{,}36492, 0{,}21188, 0{,}77431), \quad \boldsymbol{x}_2 = (0{,}13833, 0{,}65251, 0{,}70783),$$

$$\boldsymbol{x}_3 = (0{,}88871, 0{,}18157, 0{,}62118), \quad \boldsymbol{x}_4 = (0{,}76361, 0{,}52763, 0{,}05633),$$

$$\boldsymbol{x}_5 = (0{,}53709, 0{,}11589, 0{,}11998), \quad \boldsymbol{x}_6 = (0{,}21213, 0{,}40094, 0{,}84482)$$

cuya matriz de distancias resulta en

$$D = \begin{pmatrix} 0 & 0{,}50198 & 0{,}54225 & 0{,}87783 & 0{,}60614 & 0{,}25584 \\ 0{,}50198 & 0 & 0{,}89015 & 0{,}91161 & 0{,}83969 & 0{,}29581 \\ 0{,}54225 & 0{,}89015 & 0 & 0{,}67414 & 0{,}55274 & 0{,}74559 \\ 0{,}87783 & 0{,}91161 & 0{,}67414 & 0 & 0{,}49135 & 0{,}97052 \\ 0{,}60614 & 0{,}83969 & 0{,}55274 & 0{,}49135 & 0 & 0{,}77648 \\ 0{,}25584 & 0{,}29581 & 0{,}74559 & 0{,}97052 & 0{,}77648 & 0 \end{pmatrix}$$

Como $d_{23} \geq \max\{d_{12}, d_{13}\}$ pero $d_{23} - (d_{12} + d_{13}) = -0{,}15407$, por (2) sabemos que la matriz no corresponde a dimensión uno.

Tomando x_3, x_4 de (6) y y_3, y_4 de (7) obtenemos $d_{34}^2 - (d_{13}^2 + d_{14}^2 - 2x_3x_4 - 2y_3y_4) = 0{,}34454$ y $d_{34}^2 - (d_{13}^2 + d_{14}^2 - 2x_3x_4 + 2y_3y_4) = -1{,}3127$. Entonces, usando (8) sabemos que los puntos no están en Re^2.

Ahora, definiendo x_4, x_5 como en (14), y_4, y_5 como en (15) y z_4, z_5 como en (16), tenemos que $d_{45}^2 - (d_{14}^2 + d_{15}^2 - 2x_4x_5 - 2y_4y_5 - 2z_4z_5) = 0$, que nos dice que los puntos tienen dimensión tres. Finalmente, a partir de las ecuaciones (10) a (16) y estimando de la misma ecuación al último punto, resulta que

$$\boldsymbol{y}_1 = (0,0,0), \qquad \boldsymbol{y}_2 = (0{,}50198, 0, 0),$$

$$\boldsymbol{y}_3 = (-0{,}24539, 0{,}48354, 0), \qquad \boldsymbol{y}_4 = (0{,}19077, 0{,}72773, 0{,}45232),$$

$$\boldsymbol{y}_5 = (-0{,}08537, 0{,}32471, 0{,}50466) \text{ y } \boldsymbol{y}_6 = (0{,}22903, -0{,}08688, -0{,}07385)$$

satisfacen con tener como matriz de distancias a D.

Referencias

[1] M.R. Brito, J.A. Quiroz, and J.E. Yukich, *Graph-theoretic procedures for dimension identification*, J. of Mult. Anal. **81**, (2002) 67–84.

[2] J.C. Gowe, *Some latent properties of latent root and vector methods used in multivariate analysis*, Biometrika **23**, (1966), 623–628.

[3] K.V. Mardia, *Some properties of classical multi-dimensional scaling*, Commun. Stat. Theor. Methods A, **7** (1978), 1233–1241.

[4] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis*, Academic Press, London 1979.

[5] G.A.F. Seber, *Multivariate Observations*, Wiley Ser. in Prob. and Math. Stat., New York 1984.