Capítulo 3

Escalado Multidimensional

3.1. ESCALADO MÉTRICO

Dada una matriz de distancias euclídeas al cuadrado $D = (d_{ij})$ entre n objetos, el escalado métrico consiste en encontrar las coordenadas de dichos objetos en r variables $Z_1, \ldots, Z_r$.

3.1.1. INTRODUCCIÓN

Sea la matriz X con los valores de p variables $X_1, X_2, \ldots, X_p$ para n objetos

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Llamamos Y a la matriz obtenida al centrar las columnas de X.

Matriz de productos escalares: La matriz de productos escalares Q es la matriz formada por los productos escalares entre los vectores fila de Y (correspondientes a los objetos). Llamemos $\mathbf{y}_i' = (y_{i1}, \ldots, y_{ip})$ a la fila i-ésima de Y, es decir,

$$Y = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}.$$

Entonces Q es la matriz

$$Q = \begin{pmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & \ddots & \vdots \\ q_{n1} & \cdots & q_{nn} \end{pmatrix},$$

donde el elemento en la fila i y en la columna j es

$$q_{ij} = \mathbf{y}_i'\mathbf{y}_j = \sum_{k=1}^{p} y_{ik}y_{jk}, \;\; i, j = 1, \ldots, n\,.$$

En forma matricial, Q es igual a

$$Q = YY'\,.$$

Ejemplo 3.1 Considera la matriz de datos

$$X = \begin{pmatrix} 4 & 1 & 4 & 3 \\ 5 & 5 & 4 & 4 \\ 2 & 1 & 3 & 1 \\ 1 & 1 & 1 & 4 \end{pmatrix}.$$

Calcula la matriz Y con columnas centradas y la matriz de productos escalares Q:

$$Y = \left(\right) .$$

$$Q = \left(\right) .$$

Matriz de distancias euclídeas al cuadrado: La matriz de distancias euclídeas al cuadrado es la matriz formada por las distancias (euclídeas al cuadrado) entre los n objetos:

$$D = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{pmatrix},$$

donde el elemento en la fila i y columna j es

$$d_{ij} = d_e^2(\mathbf{y}_i, \mathbf{y}_j) = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j) = \sum_{k=1}^{p} (y_{ik} - y_{jk})^2 .$$

Ejemplo 3.2 Para la matriz Y dada en el Ejemplo 3.1, calcular la matriz de distancias euclídeas al cuadrado entre las filas.

$$D = \left(\right) .$$

Relación entre Q y D: La matriz de distancias D se puede obtener a partir de la matriz de productos escalares Q, de la forma:

$$\begin{aligned} d_{ij} &= \sum_{k=1}^{p}(y_{ik}-y_{jk})^2 \\ &= \sum_{k=1}^{p} y_{ik}^2 + \sum_{k=1}^{p} y_{jk}^2 - 2\sum_{k=1}^{p} y_{ik}y_{jk} \\ &= q_{ii} + q_{jj} - 2q_{ij}, \quad i,j = 1,\ldots,n\,. \end{aligned} \tag{3.1}$$

A partir de una matriz de datos Y obtenemos la matriz de productos escalares Q, y a partir de esta matriz Q se obtiene fácilmente la matriz de distancias D.

El Escalado Métrico consiste en el proceso opuesto: Se parte de una matriz de distancias D (o más en general, de disimilaridades Δ), y se busca la matriz de coordenadas Z cuya matriz de distancias euclídeas sea exactamente D (ó Δ).

$$D \to Q \to Z \text{ tal que } D_Z = D$$
$$\Delta \to Q \to Z \text{ tal que } D_Z = \Delta$$

Vamos a ver en primer lugar el caso en el que la matriz de partida es una matriz de distancias euclídeas al cuadrado.

3.1.2. COORDENADAS PRINCIPALES A PARTIR DE DISTANCIAS

Coordenadas principales: Sea $D = (d_{ij})$ una matriz de distancias euclídeas al cuadrado entre n objetos. Las *coordenadas principales* son las coordenadas de dichos objetos en $r < n$ dimensiones, $Z_1, \ldots, Z_r$ (columnas de una matriz $Z_{n\times r}$), de manera

que la matriz de distancias euclídeas al cuadrado entre las coordenadas de dichos objetos coincida con D.

Partimos de una matriz de distancias $D = (d_{ij})$ obtenida supuestamente de una matriz de coordenadas Y. Vamos a ver que a partir de D es muy fácil obtener la matriz de productos escalares $Q = YY'$, y a partir de Q se obtiene otra matriz de coordenadas Z, cuya matriz de distancias coincide con D.

Vamos a suponer que las columnas de Z están centradas.

Proposición 3.1 La matriz de productos escalares $Q = (q_{ij})$ se puede obtener a partir de la matriz de distancias $D = (d_{ij})$, de la forma

$$q_{ij} = -\frac{1}{2}\left(d_{ij} - d_{i\cdot} - d_{\cdot j} + d_{\cdot\cdot}\right), \quad i, j = 1, \ldots, n\,, \tag{3.2}$$

donde

$$d_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} d_{ij}, \quad d_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} d_{ij}, \quad \text{y} \quad d_{\cdot\cdot} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}\,.$$

Ejemplo 3.3 Partiendo de la matriz D obtenida en el Ejemplo 3.2, obtener la matriz Q de productos escalares.

Debemos aplicar la fórmula (3.2). Para ello, en primer lugar calculamos las medias de filas y de columnas de D. Lo escribimos en forma de tabla:

d_{ij}					$d_{i\cdot}$
	0	18	9	19	...
	18	0	35	41	...
	9	35	0	14	...
	19	41	14	0	...
$d_{\cdot j}$	...	...	...	...	$d_{\cdot\cdot} = ...$

Aplicando ahora la fórmula (3.2), obtenemos

$$Q = \begin{pmatrix} & & & \\ & & & \\ & & & \\ & & & \end{pmatrix} .$$

Esta matriz coincide con la matriz de productos escalares de las filas de Y, calculada en el Ejemplo 3.1.

Ahora vamos ver cómo obtener la matriz de coordenadas principales Z a partir de Q.

Nota 3.1 Si Q se calcula a partir de una matriz de distancias D con la fórmula (3.2), entonces Q es semidefinida positiva. Es decir, todos sus valores propios son no negativos.

Proposición 3.2 Realizamos la descomposición espectral de Q,

$$Q = V\Lambda V' ,$$

donde Λ es la matriz diagonal con los valores propios de Q, y V es una matriz ortogonal $n \times r$ con los r vectores propios asociados a los valores propios de Q en sus columnas. La matriz de coordenadas principales es entonces

$$Z = V\Lambda^{1/2} .$$

Ejemplo 3.4 Obtener la matriz de coordenadas principales Z que se obtendría a partir de la matriz de distancias D calculada en el Ejemplo 3.2, sabiendo que los valores propios de la matriz Q son:

$$\lambda_1 = 23{,}33,\ \lambda_2 = 8{,}05,\ \lambda_3 = 2{,}62,\ \lambda_4 = 0\,,$$

y sus vectores propios asociados son respectivamente

$$\begin{pmatrix} -0{,}08 \\ -0{,}78 \\ 0{,}35 \\ 0{,}51 \end{pmatrix} \begin{pmatrix} -0{,}41 \\ 0{,}23 \\ -0{,}54 \\ 0{,}69 \end{pmatrix} \begin{pmatrix} 0{,}76 \\ -0{,}27 \\ -0{,}59 \\ 0{,}10 \end{pmatrix} \begin{pmatrix} 0{,}5 \\ 0{,}5 \\ 0{,}5 \\ 0{,}5 \end{pmatrix}.$$

Calcular la matriz de distancias euclídeas al cuadrado de la nueva matriz de coordenadas. Compararlas con la matriz de distancias original D.

Se definen la matriz diagonal de valores propios no nulos y la matriz de vectores propios asociados

$$\Lambda = \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}, \quad V = \begin{pmatrix} & & \\ & & \\ & & \\ & & \end{pmatrix}.$$

Así, la matriz de coordenadas es

$$Z = V\Lambda^{1/2} = \begin{pmatrix} & & \\ & & \\ & & \\ & & \end{pmatrix}.$$

La nueva matriz de distancias euclídeas al cuadrado es

$$D = \left(\right) .$$

Propiedades de las coordenadas principales:

(a) Las columnas de Z tienen media cero.

(b) Las varianzas de las columnas de Z son proporcionales a los valores propios,

$$s^2_{Z_k} = \lambda_k / n, \quad k = 1, \ldots, p.$$

(c) Las correlaciones entre los pares de columnas de Z son cero.

(d) Sea $Y_{n \times p}$ una matriz con distancias euclídeas entre sus filas igual a D. Entonces las coordenadas principales Z obtenidas a partir de D son las componentes principales de Y.

3.1.3. COORDENADAS PRINCIPALES A PARTIR DE DISIMILARIDADES

Ahora partimos de una matriz de disimilaridades $\Delta = (\delta_{ij})$ simétrica y con ceros en la diagonal, es decir, $\delta_{ij} = \delta_{ji}$, y $\delta_{ii} = 0$, pero no exigimos que verifique la desigualdad triangular.

Nota 3.2 Las disimilaridades pueden atender a preferencias. En ese caso, $\delta_{A,B} = 0$ cuando no hay preferencia por A sobre B ni al contrario y $\delta_{A,B} > 0$ cuando hay preferencia por A sobre B o por B sobre A.

La matriz de disimilaridades puede haber sido obtenida de una de las siguientes maneras:

(a) Estimación directa: Un juez o conjunto de jueces estima directamente las disimilaridades entre los elementos. Una escala muy utilizada es la escala 0-100. Con n elementos para comparar, se requieren $n(n-1)/2$ evaluaciones.

(b) Rangos por pares: Se presentan al juez los $n(n-1)/2$ pares posibles, y se le pide que los ordene de menor a mayor disimilaridad, asignando rangos de 1 a $n(n-1)/2$.

Si las disimilaridades se han obtenido como en (b), es usual transformarlas de la siguiente forma: se calcula el rango medio para cada objeto promediando los rangos de los pares donde aparece. La disimilaridad final entre dos objetos va a ser la diferencia en valor absoluto entre sus rangos.

Ejemplo 3.5 Supongamos que un experto ha comparado cuatro objetos A, B, C y D ordenando los pares de la forma:

Par	(C,D)	(B,C)	(B,D)	(A,D)	(A,B)	(A,C)
Rango	1	2	3	4	5	6

Los rangos medios son

Objeto	A	B	C	D
Rango	5	3.33	3	2.66

Así, la matriz de disimilaridades es

	A	B	C	D
A	0			
B	1.66	0		
C	2	0.33	0	
D	2.33	0.66	0.33	0

Deseamos encontrar la matriz de coordenadas $Z = (z_{ij})$ tal que su matriz de distancias euclídeas al cuadrado $D = (d_{ij})$ coincida con Δ. Esto se puede realizar siguiendo el mismo proceso que en la Sección 3.1.2, pero partiendo de Δ.

(a) Calculamos Q a partir de la fórmula

$$q_{ij} = -\frac{1}{2}\left(\delta_{ij} - \delta_{i\cdot} - \delta_{\cdot j} + \delta_{\cdot\cdot}\right), \quad i, j = 1, \ldots, n\,, \tag{3.3}$$

donde

$$\delta_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n}\delta_{ij}, \quad \delta_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\delta_{ij}, \quad \text{y} \quad \delta_{\cdot\cdot} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}\,.$$

(b) Si los valores propios de Q son no negativos, entonces se puede realizar su descomposición espectral igual que antes, $Q = V\Lambda V'$, lo que proporciona la matriz de coordenadas $Z = V\Lambda^{1/2}$.

Proposición 3.3 Supongamos que a partir de la matriz de disimilaridades $\Delta = (\delta_{ij})$, obtenemos Q como en (a). Supongamos que Q tiene valores propios no negativos, y obtenemos la matriz de coordenadas Z aplicando (b). Entonces estas coordenadas son las coordenadas principales, es decir, su matriz de distancias euclídeas es la matriz Δ de partida.

3.1.4. SELECCIÓN DEL NÚMERO DE DIMENSIONES

En lugar de quedarnos con p coordenadas principales (p columnas de Z), seleccionaremos las $r < p$ que tengan los mayores valores propios, de manera que la representación sea en la menor dimensión posible. Así, en lugar de $Z = V\Lambda^{1/2}$, tomaremos $Z_r = V_r\Lambda_r^{1/2}$, donde Λ_r es la matriz diagonal con los r mayores valores propios, y V_r es la matriz con los r vectores propios correspondientes en las columnas.

Si existe algún valor propio negativo, se tomarán los r mayores valores propios positivos para construir Λ_r igual que antes, y sus r vectores propios asociados que forman las columnas de V_r, con lo cual las coordenadas principales serán $Z_r = V_r\Lambda_r^{1/2}$.

Una forma de asesorar con cuántas dimensiones quedarnos es con la ayuda de la medida

$$100 \cdot \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{n} |\lambda_i|} .$$

Ejemplo 3.6 Sea la matriz de disimilaridades

$$\Delta = \begin{pmatrix} 0 & 3 & 0 \\ 3 & 0 & 6 \\ 0 & 6 & 0 \end{pmatrix} .$$

Obtener la matriz de coordenadas Z, cuya matriz de distancias euclídeas se aproxime a Δ.

3.2. ESCALADO NO MÉTRICO

Partiendo de la matriz de disimilaridades $\Delta = (\delta_{ij})$, el Escalado no métrico consiste en encontrar unas coordenadas, cuyas distancias euclídeas al cuadrado mantengan el órden de las disimilaridades. Es decir, el Escalado no Métrico solo tiene en cuenta la información referente al órden entre las disimilaridades δ_{ij}, y no su magnitud.

3.2.1. COORDENADAS INICIALES

Se calculan unas coordenadas iniciales $Z^{(0)}$, por ejemplo aplicando el Escalado métrico a la matriz de disimilaridades $\Delta = (\delta_{ij})$. Esto consiste en calcular la matriz de productos escalares Q, realizar la descomposición espectral $Q = V\Lambda V'$, y tomar $Z^{(0)} = V_r\Lambda_r^{1/2}$, donde Λ_r contiene los r mayores valores propios, y V_r sus vectores propios asociados en columnas. Se calcula la matriz de distancias $D^{(0)} = (d_{ij}^{(0)})$ a partir de las coordenadas iniciales $Z^{(0)}$.

3.2.2. CÁLCULO DE DISPARIDADES

Las disparidades $\widehat{d}_{ij}$ son una transformación de las distancias actuales d_{ij} que mantienen la misma ordenación que las disimilaridades δ_{ij}, es decir,

$$\widehat{d}_{ij} = f(d_{ij})\,,$$

donde f es una función monótona que verifica

$$\text{Si } \delta_{ij} \leq \delta_{kl}, \text{entonces } \widehat{d}_{ij} \leq \widehat{d}_{kl}\,.$$

Ejemplo 3.7 Considérese la matriz de disimilaridades

$$\Delta = (\delta_{ij}) = \begin{pmatrix} 0 & 2{,}1 & 3 & 2{,}4 \\ & 0 & 1{,}7 & 3{,}9 \\ & & 0 & 3{,}2 \\ & & & 0 \end{pmatrix}$$

Supongamos que hemos obtenido una matriz de coordenadas inicial $Z^{(0)}$, cuya matriz de distancias es

$$D^{(0)} = (d_{ij}^{(0)}) = \begin{pmatrix} 0 & 1{,}6 & 4{,}5 & 5{,}7 \\ & 0 & 3{,}3 & 4{,}3 \\ & & 0 & 1{,}3 \\ & & & 0 \end{pmatrix}$$

Obtener las disparidades:

Escribimos las disimilaridades en órden creciente

$$\delta_{23} = 1{,}7 \;\; \delta_{12} = 2{,}1 \;\; \delta_{14} = 2{,}4 \;\; \delta_{13} = 3 \;\; \delta_{34} = 3{,}2 \;\; \delta_{24} = 3{,}9$$

Ahora escribimos las distancias correspondientes

$$d_{23} = ... \quad d_{12} = ... \quad d_{14} = ... \quad d_{13} = ... \quad d_{34} = ... \quad d_{24} = ...$$

Si manduviesen el mismo órden que las disimilaridades, estarían ordenadas de menor a mayor, y en ese caso, las disparidades serían iguales a las distancias. Entonces $Z^{(0)}$ sería una solución válida.

Una transformación monótona de estas distancias que preserva el órden de las disimilaridades se calcula de la siguiente forma: cuando existe una secuencia de distancias que están ordenadas al contrario de lo deseado, se reemplazan todas estas distancias por la media

de las distancias de dicha secuencia. Así, las disparidades son:

$$\widehat{d}_{23} = \widehat{d}_{12} = \frac{1}{2}(d_{23} + d_{12}) = ... \quad ,$$
$$\widehat{d}_{14} = \widehat{d}_{13} = \widehat{d}_{34} = \frac{1}{3}(d_{14} + d_{13} + d_{34}) = ... \quad ,$$
$$\widehat{d}_{24} = d_{24} = ... \quad .$$

3.2.3. MEDIDAS DE BONDAD DE AJUSTE DE LA SOLUCIÓN OBTENIDA

- STRESS

$$S = \left[\frac{\sum_{i<j}(d_{ij} - \widehat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}\right]^{1/2} .$$

  ✓ $S \in (0, 0.01] \Rightarrow$ Solución muy buena;

  ✓ $S \in (0.01, 0.05] \Rightarrow$ Solución buena;

  ✓ $S \in (0.05, 0.1] \Rightarrow$ Solución aceptable.

- S-STRESS

$$SS = \left[\frac{\sum_{i<j}(d_{ij}^2 - \widehat{d}_{ij}^2)^2}{\sum_{i<j} d_{ij}^4}\right]^{1/2} .$$

  Esta medida está entre 0 y 1, siendo valores cercanos a cero indicadores de un buen ajuste, y valores cercanos a 1 indicadores de un mal ajuste.

- RSQ

  Es el coeficiente de correlación al cuadrado entre las distancias y las disparidades. El ajuste es aceptable para RSQ $\geq$ 0,6.

Ejemplo 3.8 Calcular el STRESS para las matrices de distancia y de disparidades obtenidas en el Ejemplo 3.7.

$$\sum_{i<j}(d_{ij} - \widehat{d}_{ij})^2 = \qquad , \qquad \sum_{i<j} d_{ij}^2 = \qquad .$$

Por tanto,

$$S = \left(\text{——}\right)^{1/2} = \qquad .$$

Si para las coordenadas actuales, estas medidas no son satisfactorias, entonces pasamos a la búsqueda de una nueva solución. Esta solución se busca minimizando una de las medidas de ajuste respecto a las coordenadas, generalmente se utiliza el STRESS o el SSTRESS.

3.2.4. MINIMIZACIÓN DEL STRESS

Sea $\mathbf{z} = (\mathbf{z}_1', \ldots, \mathbf{z}_n')'$ el vector formado por las n filas de Z (nuestras incógnitas). El problema es encontrar $\mathbf{z}$ que minimice

$$S = \left[\frac{\sum_{i<j}(d_{ij}(\mathbf{z}) - \widehat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2(\mathbf{z})}\right]^{1/2} ,$$

donde

$$d_{ij}^2(\mathbf{z}) = d_e^2(\mathbf{z}_i, \mathbf{z}_j) = \sum_{k=1}^{p}(z_{ik} - z_{jk})^2 .$$

Este problema es no lineal, con lo cual es necesario recurrir a métodos de resolución numéricos.