Escalado multidimensional

Eliseo Martínez

(Basado en el libro Análisis de Datos Multivariantes de Daniel Peña, edit. MacGraw Hill, 2002)

1. La matriz de distancia

Supongamos, como antes, que la matriz de datos $\mathbf{X}$ es de n filas por p columnas. Sabemos que la forma de centrar esta matriz, esto es de que sus columnas tengan media cero es haciendo

$$\widetilde{\mathbf{X}} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right)\mathbf{X} = \mathbf{P}\cdot\mathbf{X}$$

donde $\mathbf{1}$ el vector de dimensión n con entradas de "unos", $\mathbf{I}$ es la matriz identidad de orden $n \times n$. Con esta matriz $\widetilde{\mathbf{X}}$ podemos formar dos matrices simétricas definidas positivas. A saber, la matriz de varianzas y covarianzas $\mathbf{S}$ definida por $\mathbf{X}^t\,\mathbf{X}/(n-1)$, que es de $p \times p$, y la matriz de productos cruzados, $\mathbf{Q}$, definida por

$$\mathbf{Q} = \widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^{\,t}$$

Esta matriz es de orden $n\times n$, y se interpreta como una matriz de similitud (covarianza) entre los n "individuos". Como veremos ahora, las entradas de la matriz $\mathbf{Q}$ nos permiten realizar los cálculos de las distancia euclideas entre las respuestas de los individuos (distancia entre los vectores filas). En efecto la $(i,j)-$ ésima entrada de la matriz $\mathbf{Q}$ está dada por

$$q_{ij} = \sum_{k=1}^{p} x_{ik}\,x_{jk} = \mathbf{x}_i^t\,\mathbf{x}_j$$

donde $\mathbf{x}_i$ y $\mathbf{x}_j$ son los vectores filas $i-$ésimo y $j-$ésimo respectivamente de la matriz $\mathbf{X}$. Observemos que q_{ij} es la resultante del producto punto entre $\mathbf{x}_i$ y $\mathbf{x}_j$, esto es

$$qij = |\mathbf{x}_i|\,|\mathbf{x}_j|\cos\theta_{ij}$$

Ahora si las coordenadas de $\mathbf{x}_i$ y $\mathbf{x}_j$ son muy similares entonces $\cos\theta_{ij} \approx 1$, y en consecuencia q_{ij} será muy grande. Por el contrario, si las corrdenadas de $\mathbf{x}_i$ y $\mathbf{x}_j$ difieren en mucho entonces $\cos\theta_{ij} = 0$, y en consecuencia q_{ij} será pequeño. Teniendo esto en mente podemos interpretar a la matriz $\mathbf{Q}$ como la matriz de similitud entre los elementos (individuos).

La idea de similitud también está asociada a la idea de distancia, en efecto, podemos pensar que dos elementos o individuos serán más similares si su distancia entre ellos es pequeña. Pues bien, la matriz $\mathbf{Q}$ puede generar rápidamente la distancia entre individuos. Sabemos que la distancia ueclídea al cuadrado entre dos elementos está dada por

$$d_{ij}^2 = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2 = \sum_{k=1}^{p} x_{ik}^2 + \sum_{k=1}^{p} x_{jk}^2 - 2\sum_{i=1}^{p} x_{ik}\,x_{jk}$$

de tal modo que esta distancia cuadrática puede calcularse mediante las entradas de la matriz $\mathbf{Q}$, a saber

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij} \tag{1}$$

En resumen, con la matriz $\widetilde{\mathbf{X}}$ calculamos la matriz de similitud $\mathbf{Q}$, y luego la matriz de distancias al cuadrado $\mathbf{D}$ con ayuda de la expresión anterior.

La generación de un algoritmo sencillo para el cálculo de la matriz D es como sigue. Sea $diag(\mathbf{Q})$ el vector que contiene la diagonal de la matriz $\mathbf{Q}$, y $\mathbf{1}$ el vector $n-$dimensional que contiene "unos", entonces

$$\mathbf{D} = diag(\mathbf{Q})\,\mathbf{1}^t + \mathbf{1}\,diag(\mathbf{Q})^t - 2\mathbf{Q}$$

2. El problema inverso

Entenderemos, en este contexto, el problema inverso como la reconstrucción de la matriz $\widetilde{\mathbf{X}}$ a partir de una matriz de distancias al cuadrado $\mathbf{D}$. Es claro que para la obtención de la matriz $\widetilde{\mathbf{X}}$ debemos primero obtener la matriz $\mathbf{Q}$.

Supongamos que tenemos una matriz $\mathbf{D}$ de distancias al cuadrado. Notemos que no hay pérdida de generalidad al suponer que las variables tienen media cero, toda vez que las distancias no varían si expresamos las variables desviadas respecto de la media. En efecto,

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = \sum_{k=1}^{p} ((x_{ik} - \overline{x}_k) - (x_{jk} - \overline{x}_s))^2$$

Ahora bien, suponiendo que la matriz que $\widetilde{\mathbf{X}}$ que debemos encontrar está centrada en el vector de medias, tenemos que

$$\widetilde{\mathbf{X}}^t\mathbf{1} = \mathbf{0}$$

y puesto que $\mathbf{Q} = \widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$, se tiene que también $\mathbf{Q1} = \mathbf{0}$. Lo que significa que la suma de los elementos de una fila de la matriz $\mathbf{Q}$ debe ser cero, esto es $\sum_{i=1}^{n} q_{ij} = 0$, y como es simétrica también ocurre que la suma de los elementos de una columna debe ser cero. Por otro lado, sabemos, que la relación que debe existir entre los valores d_{ij}^2 y las entradas de q_{ij} de la matriz $\mathbf{Q}$ está dada por la ecuación (1). Si en esta ecuación (1) sumamos a través de las filas, nos queda

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} q_{ii} + nq_{jj} = t + nq_{jj} \tag{2}$$

donde definimos $t = traza(\mathbf{Q}) = \sum_{i=1}^{n} q_{ii}$.

Ahora sumaremos (1) a través de las columnas, esto es

$$\sum_{j=1}^{n} d_{ij}^2 = nq_{ii} + t \tag{3}$$

Notemos que en las expresiones (2) y (3) podemos despejar q_{ii} y q_{jj}. Sumando a través de

i en (3), obtenemos

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2\,n\,t \tag{4}$$

Ahora en la ecuación (1) reemplazamos el lado derecho los valores q_{ii} y q_{jj} encontrados en (2) y (3), tenemos que

$$\begin{aligned} d_{ij}^2 &= \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{t}{n} + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{t}{n} - 2q_{ij} \\ &= \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{2\,t}{n} - 2q_{ij} \end{aligned}$$

y reemplazando $2\,t$ al despejar en (4), tenemos

$$d_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 - 2q_{ij}$$

Y hemos conseguido una expresión para las entradas q_{ij} en función de las entradas d_{ij}. Esto es

$$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2\right)$$

Si hacemos

$$\begin{aligned} d_{\cdot j}^2 &= \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 \\ d_{i\cdot}^2 &= \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 \\ d_{\cdot\cdot}^2 &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 \end{aligned}$$

obtenemos una igualdad más compacta,

$$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{\cdot j}^2 - d_{i\cdot}^2 + d_{\cdot\cdot}^2\right) \tag{5}$$

Y son estos valores q_{ij} que determinan a la matriz $\mathbf{Q}$.

Ahora nos falta determinar la matriz $\mathbf{X}$. Supongamos que $\mathbf{Q}$ es una matriz definida positiva de rango p, entonces esta matriz se puede diagonalizar. Esto es,

$$\mathbf{Q} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t \tag{6}$$

siendo $\mathbf{\Lambda}$ la matriz diagonal de orden $p \times p$ constituida por los autovalores no nulos de $\mathbf{Q}$, y $\mathbf{V}$ es una matriz de orden $n \times p$ y contiene en sus columnas a los vectores propios correspondiente a los valores propios no nulos de $\mathbf{Q}$. Puesto que $\mathbf{\Lambda} = \mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}$, tenemos que

$$\mathbf{Q} = (\mathbf{V}\mathbf{\Lambda}^{\mathbf{1/2}})(\mathbf{\Lambda}^{\mathbf{1/2}}\mathbf{V}^t) = (\mathbf{V}\mathbf{\Lambda}^{\mathbf{1/2}})(\mathbf{V}\mathbf{\Lambda}^{\mathbf{1/2}})^t$$

Si hacemos $\mathbf{Y} = (\mathbf{V}\mathbf{\Lambda}^{\mathbf{1/2}})$, tenemos que

$$\mathbf{Q} = \mathbf{Y}\,\mathbf{Y}^t$$

Esta matriz $\mathbf{Y}$ de $n \times p$ tiene sus columnas no correlacionadas (son ortogonales) que reproducen, por su propia construcción, la métrica original. ¿Es esta la matriz $\widetilde{\mathbf{X}}$ buscada?

Verifiquemos con un ejemplo. De una matriz $\mathbf{X}$ vamos a obtener la natriz de las distancias cuadráticas $\mathbf{D}$, y luego a partir de esta matriz realizaremos el procedimiento anterior y veremos si efectivamente llegamos a la matriz $\mathbf{X}$.

Ejemplo. Consideremos la matriz de datos

$$\mathbf{X} = \begin{pmatrix} 0.301 & 0.301 \\ 0.176 & -0.301 \\ -0.155 & -0.301 \\ -0.301 & 0.176 \\ -0.301 & -0.155 \\ -0.155 & -0.155 \end{pmatrix}$$

cuya matriz centrada $\widetilde{\mathbf{X}}$ es

$$\widetilde{\mathbf{X}} = \begin{pmatrix} 0.3735 & 0.3735 \\ 0.2485 & -0.2285 \\ -0.0825 & -0.2285 \\ -0.2285 & 0.2485 \\ -0.2285 & -0.0825 \\ -0.0825 & -0.0825 \end{pmatrix}$$

Calculando la matriz distancia cuadrática, $\mathbf{D}$, de estos datos, obtenemos

$$\mathbf{D} = \begin{pmatrix} 0 & 0.378029 & 0.57034 & 0.378029 & 0.57034 & 0.415872 \\ 0.378029 & 0 & 0.109561 & 0.455058 & 0.248845 & 0.130877 \\ 0.57034 & 0.109561 & 0 & 0.248845 & 0.042632 & 0.021316 \\ 0.378029 & 0.455058 & 0.248845 & 0 & 0.109561 & 0.130877 \\ 0.57034 & 0.248845 & 0.042632 & 0.109561 & 0 & 0.021316 \\ 0.415872 & 0.130877 & 0.021316 & 0.130877 & 0.021316 & 0 \end{pmatrix}$$

Todos estos tediosos cálculos los puede realizar con un sencillo programa llamado **distancia.mth** y que está a disposición en el Internet[1]. A partir de esta matriz de distancias cuadráticas, construimos la matriz $\mathbf{Q}$ cuyas entradas estarán dadas por la relación (5).

$$\mathbf{Q} = \begin{pmatrix} 0.27900 & 0.00747 & -0.11615 & 0.00747 & -0.11615 & -0.06162 \\ 0.00747 & 0.11396 & 0.03171 & -0.11356 & -0.03793 & -0.00165 \\ -0.11615 & 0.03171 & 0.05901 & -0.03793 & 0.03770 & 0.02565 \\ 0.00747 & -0.11356 & -0.03793 & 0.11396 & 0.03171 & -0.00165 \\ -0.11615 & -0.03793 & 0.03770 & 0.03171 & 0.05901 & 0.02565 \\ -0.06162 & -0.00165 & 0.02565 & -0.00165 & 0.02565 & 0.01361 \end{pmatrix}$$

Se puede verificar fácilmente que esta matriz es exactamente la misma que se puede obtener mediante $\mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^t$. Ahora bien, a la matriz $\mathbf{Q}$ le calcularemos los autovalores

[1] http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/distancia.html

no nulos y sus correspondientes autovectores asociados a fin de determinar las matrices $\mathbf{V}$ y $\mathbf{\Lambda}$ de la relación (6).

Los autovalores no nulos de la matriz $\mathbf{Q}$ son $\lambda_1 = 0.389738$ y $\lambda_2 = 0.248845$ y sus respectivos autovectores asociados

$$\mathbf{a}_1 = \begin{pmatrix} -0.84609 & -0.02265 & 0.35225 & -0.02265 & 0.35225 & 0.18688 \end{pmatrix}^t$$

$$\mathbf{a}_2 = \begin{pmatrix} 0 & 0.676143 & 0.20695 & 0.676143 & -0.20695 & 0 \end{pmatrix}^t$$

Y así estamos en condiciones de formar $\mathbf{V}$ y $\mathbf{\Lambda}^{1/2}$

$$\mathbf{V} = \begin{pmatrix} -0.84609 & 0 \\ -0.02265 & 0.676143 \\ 0.35225 & 0.20695 \\ -0.02265 & 0.676143 \\ 0.35225 & -0.20695 \\ 0.18688 & 0 \end{pmatrix}$$

$$\mathbf{\Lambda}^{1/2} = \begin{pmatrix} \sqrt{0.389738} & 0 \\ 0 & \sqrt{0.248845} \end{pmatrix}$$

De manera que

$$\mathbf{Y} = \mathbf{V}\mathbf{\Lambda}^{1/2} = \begin{pmatrix} -0.5282087655 & 0 \\ -0.01414213561 & 0.3372899345 \\ 0.2199102089 & 0.1032375900 \\ -0.01414213561 & -0.3372899345 \\ 0.2199102089 & -0.1032375900 \\ 0.1166726188 & 0 \end{pmatrix}$$

Se puede verificar que lass columnas de $\mathbf{Y}$ efectivamente son ortogonales (si no es cero el producto será casi cero por problemas de redondeo), pero el caso es que la matriz $\mathbf{Y}$ no es igual a la matriz de datos $\mathbf{X}$. Sin embargo, toda vez que $\mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^t$ podemos concluir que esta matriz $\mathbf{Y}$ es la matriz de los componentes principales de $\mathbf{X}$.

Calcularemos los componentes principales de la matriz $\mathbf{X}$. En efecto, sea $\mathbf{R}$ la matriz de correlación

$$\mathbf{R} = \begin{pmatrix} 1 & 0.2206338095 \\ 0.2206338095 & 1 \end{pmatrix}$$

Este cálculo y los que vienen a continuación se pueden hacer con el programa **multiv.mth** realizado con el DERIVE y ubicado en la red Internet[2]. Los autovalores de esta matriz son $\mu_1 = 1.220633809$ y $\mu_2 = 0.7793661904$ y los respectivos autovectores asociados (-0.7071067785, -0.7071067838)^t y (-0.7071067812, 0.7071067811)^t. Calculando los componentes principales de la matriz centrada $\widetilde{\mathbf{X}}$, tenemos

[2] http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/multiv.html

$$
\begin{aligned}
z_1 &= \begin{pmatrix} 0.3735 & 0.3735 \\ 0.2485 & -0.2285 \\ -0.0825 & -0.2285 \\ -0.2285 & 0.2485 \\ -0.2285 & -0.0825 \\ -0.0825 & -0.0825 \end{pmatrix} \cdot \begin{pmatrix} -0.7071067785 \\ -0.7071067838 \end{pmatrix} \\
&= \begin{pmatrix} -.52821 \\ -1.4142 \times 10^{-2} \\ .21991 \\ -1.4142 \times 10^{-2} \\ .21991 \\ .11667 \end{pmatrix}
\end{aligned}
$$

:

$$
\begin{aligned}
z_2 &= \begin{pmatrix} 0.3735 & 0.3735 \\ 0.2485 & -0.2285 \\ -0.0825 & -0.2285 \\ -0.2285 & 0.2485 \\ -0.2285 & -0.0825 \\ -0.0825 & -0.0825 \end{pmatrix} \cdot \begin{pmatrix} -0.7071067812 \\ 0.7071067812 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ -.33729 \\ -.10324 \\ .33729 \\ .10324 \\ 0 \end{pmatrix}
\end{aligned}
$$

Y podemos observar que, esencialmente, el problema inverso, esto es la búsqueda de la matriz de datos a partir de la matriz de distancias cuadráticas, nos lleva a la matriz de componentes principales normalizadas de la matriz de datos centrada (que siempre permanecerá desconocida). ■

Finalmente, de la relación $\mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^t$ si consideramos cualquier matriz ortogonal $\mathbf{A}$ se tiene que

$$\mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^t = \mathbf{Q} = \widetilde{\mathbf{X}}\mathbf{A}\mathbf{A}^t\widetilde{\mathbf{X}}^t$$

de tal manera que la matriz $\mathbf{Q}$ solo tiene información del espacio generado por la matriz $\mathbf{X}$ (por las variables), y puesto que la matriz $\mathbf{A}$ solo hace un efecto de rotación y estas preservan la distancia, en consecuencia cualquier rotación de las variables originales puede ser la solución. De otra forma la matriz $\mathbf{Y}$ es una rotación de la matriz $\mathbf{X}$.

3. El objetivo del escalado multidimensional[3]

Las técnicas de escalado multidimensional son una generalización de la idea de componentes principales cuando en lugar de disponer de una matriz de observaciones por variables, como en componentes principales, se dispone de una matriz $\mathbf{D}$*, cuadrada* $n \times n$ *de distancias o disimilaridades entre los* n *elementos de un conjunto. Por ejemplo, esta matriz puede representar las similitudes o distancias entre* n *productos fabricados por una empresa, las distancias percibidas entre* n *candidatos políticos, las diferencias entre* n *preguntas de un cuestionario o las distancias o similitudes entre* n *sectores industriales. Estas distancias pueden haberse obtenido a partir de ciertas variables, o pueden ser el resultado de una estimación directa, por ejemplo preguntando a un grupo de jueces por sus opiniones sobre las similaridades entre los elementos considerados.*

El objetivo que se pretende es representar esta matriz mediante un conjunto de variables ortogonales, $y_1, .., y_p$*, que llamaremos* ***coordenadas principales*** *donde* $p < n$*, de manera que las distancias euclídeas entre las coordenadas de los elementos respecto a estas variables sean iguales (o los más próximas posibles) a las distancias o disimilaridades de la matriz original. Es decir, a partir de la matriz* $\mathbf{D}$ *se pretende obtener una matriz* $\mathbf{X}$*, de dimensiones* $n \times p$*, que pueda interpretarse como la matriz de* p *variables en los* n *individuos, y donde la distancia euclídea entre los elementos reproduzca, aproximadamente, la matriz de distancias* $\mathbf{D}$ *inicial. Cuando* $p > 2$*, las variables pueden ordenarse en importancia y suelen hacerse representaciones gráficas en dos y tres dimensiones para entender la estructura existente.*

Este planteamiento presenta dos interrogantes: ¿es siempre posible encontrar estas variables? ¿Cómo construirlas? En general, no es posible encontrar p *variables que reproduzcan exactamente las distancias iniciales, sin embargo es frecuente encontrar variables que reproduzcan aproximadamente las distancias iniciales. Por otro lado, si la matriz de distancias se ha generado calculando las distancias euclídeas entre las observaciones definidas por ciertas variables, recuperaremos las componentes principales de estas variables.*

El escalado multidimensional comparte con componentes principales el objetivo de describir e interpretar los datos. Si existen muchos elementos, la matriz de similaridad será muy grande y la representación por unas pocas variables de los elementos nos permitirá entender su estructura: qué elementos tienen propiedades similares, si aparecen grupos entre los elementos, si hay elementos atípicos, etcétera. Además, si podemos interpretar las variables aumentará nuestro conocimiento del problema, al entender cómo se han generado los datos. Por ejemplo, supongamos que se realiza una encuesta para determinar qué similitudes encuentran los consumidores entre n *productos o servicios, y que la información se resume en una matriz cuadrada de similitudes entre los productos. Supongamos que descubrimos que estas similitudes pueden generarse por dos variables. Entonces, es razonable suponer que los consumidores han estimado la similitud entre dos productos utilizando estas dos variables.*

El escalado multidimensional representa un enfoque complementario a componentes

[3] Cita textual del libro de Daniel Peña

principales en el sentido siguiente. Componentes principales considera la matriz $p \times p$ de correlaciones (o covarianzas) entre variables, e investiga su estructura. El escalado multidimensional considera la matriz $n \times n$ de distancias entre individuos e investiga su estructura. Ambos enfoques están claramente relacionados, y existen técnicas gráficas que aprovechan esta dualidad para representar conjuntamente las variables y los individuos en un mismo gráfico.

4. Coordenadas principales

A la luz de las secciones anteriores, podemos resumir lo siguiente. Se tiene una matriz $\mathbf{D}$, de orden $n \times n$, que mide las distancias o disimilaridades de n individuos. Con esta matriz construimos la matriz $\mathbf{Q}$ de similaridad mediante la relación dada en (5). Luego calculamos las matrices $\mathbf{V}$ y $\mathbf{\Lambda}$, donde $\mathbf{\Lambda}$ es la matriz diagonal de orden $p \times p$ constituida por los autovalores no nulos de $\mathbf{Q}$, y $\mathbf{V}$ es una matriz de orden $n \times p$ y contiene en sus columnas a los vectores propios correspondiente a los valores propios no nulos de $\mathbf{Q}$. Entonces la matriz

$$\mathbf{Y} = (\mathbf{V}\mathbf{\Lambda}^{\mathbf{1/2}}) \tag{7}$$

tiene como columnas a las llamadas **coordenadas principales**.

Observe que para poder calcular las coordenadas principales, es absolutamente necesario que, según (7), los autovalores de la matriz $\mathbf{Q}$ sean no negativos, y esto se consigue necesariamente si $\mathbf{Q}$ es semidefinida positiva. Esto nos conduce a verificar cuándo la matriz de distancias es compatible con la métrica euclídea. En efecto, diremos que una matriz de distancias $\mathbf{D}$ es compatible con la métrica euclídea si la matriz de similitud generada por ella,

$$\mathbf{Q} = -\frac{1}{2}\mathbf{P}\,\mathbf{D}\,\mathbf{P} \tag{8}$$

es semidefinida positiva, donde $\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^t$. Lo interesante de esta proposición de que si $\mathbf{Q}$ es efectivamente una matriz definida positiva, podemos encontrar entonces una métrica euclìdea que reproduzca a la matriz $\mathbf{D}$.

Vamos a demostrar esto último. Supongamos que tenemos una matriz semidefinida positiva $\mathbf{Q}$, entonces vamos a encontrar variables $y_1, ..., y_p$ que reproduzcan las distancias observadas. Si $\mathbf{Q}$ es semidefinida positiva de rango p entonces admita la siguiente representación

$$\mathbf{Q} = \sum_{i=1}^{p} \lambda_i\, \mathbf{v}_i\, \mathbf{v}_i^t$$

donde λ_i son los autovalores y $\mathbf{v}_i$ los autovectores correspondientes. Si definimos

$$\mathbf{y}_i = \sqrt{\lambda_i}\mathbf{v}_i$$

entonces podemos escribir

$$\mathbf{Q} = \sum_{i=1}^{p} \mathbf{y}_i\, \mathbf{y}_i^t$$

Estas $\mathbf{y}_i$ representan la solución buscada de ser un conjunto de p variables incorrelacionadas entre sí y tales que el cuadrado de la distancia euclídea inducida es igual a la distancia

cuadrática original. En efecto, si $\mathbf{Q}$ es obtenida por (8) no resulta complicado verificar que se cumple la relación (1).[4]

5. Construcción de las coordenadas principales

Sea $\mathbf{D}$ la matriz de distancias al cuadrado[5]. A partir de esta matriz realice lo siguiente:

- Construya la matriz de similaridad $\mathbf{Q}$ mediante

  $$\mathbf{Q} = -\frac{1}{2}\mathbf{P}\,\mathbf{D}\,\mathbf{P}$$

  o mediante el cálculo directo de sus entradas

  $$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{\cdot j}^2 - d_{i\cdot}^2 + d_{\cdot\cdot}^2\right)$$

- Calcular los valores propios de $\mathbf{Q}$.
- Seleccionar los r valores propios mayores, donde r se escoge de manera que los restantes $n-r$ sean muy próximos a cero. Note que siempre el 0 será un autovalor de $\mathbf{Q}$, puesto que $\mathbf{P1} = \mathbf{0}$, de manera que $\mathbf{Q1} = \mathbf{0}$ y en consecuencia 0 es un autovalor de $\mathbf{Q}$ con autovector asociado $\mathbf{1}$.
- Obtener las coordenadas principales mediante $\mathbf{v}_i\sqrt{\lambda_i}$. O de manera equivalente calcular

  $$Q \approx \left(\mathbf{V}_r\mathbf{\Lambda}_r^{1/2}\right)\left(\mathbf{\Lambda}_r^{1/2}\mathbf{V}_r^{\,t}\right)$$

  donde $\mathbf{V}_r$ es la matriz que tiene en sus columnas los r autovectores asociados a los r autovalores corespondientes defnidos en $\mathbf{\Lambda}_r$. De esta forma las coordenadas principales se obtienen mediante

  $$\mathbf{Y}_r = \mathbf{V}_r\mathbf{\Lambda}_r^{1/2} \tag{9}$$

6. Aplicaciones: geografía y literatura

Vamos a estudiar dos aplicaciones [6]. En la primera aplicación se entregará la distancia entre ocho ciudades europeas, y de esta matriz de distancia calcularemos las coordenadas principales para determinar la estructura subyacente, que como ya lo imaginamos será, en definitiva, el mapa o ubicación relativa de estas siete ciudades. El segundo ejemplo, nos permitirá determinar alguna estrucrura que determine la diferencia de estilos entre cinco libros. En este último ejemplo no tenemos antecedentes de un estudio similar.

---

[4] Más detalles precisos de esta demostración, puede consultar el libro de Daniel Peña, páginas 177-178.

[5] Cuidado, debe verificarse si D es una matriz de distancias o es una matriz de distancias al cuadrado.

[6] Esta sección se aparta completamente del libro que hemos tenido como base, si bien es cierto que allí se presenta un ejemplo "geográfico" similar al de nuestras "ciudades europeas", pero no hay un tratamiento "literario" como el que aquí presentamos.

6.1 Las ciudades de europa

En la tabla siguiente se entregan las distancias, en kilómetros, entre las ocho ciudades europeas indicadas,

	Mad.	París	Brus.	Amst.	Berlín	Roma	Lisboa	Lon.
Mad.	0	1260	1556	1735	2360	2066	644	1725
París	1260	0	296	475	1100	1437	1792	465
Brus.	1556	296	0	198	789	1545	2088	374
Amst.	1735	475	198	0	685	1766	2267	344
Berlín	2360	1100	789	685	0	1529	2892	996
Roma	2066	1437	1545	1766	1529	0	2730	1902
Lisboa	644	1792	2088	2267	2892	2730	0	2257
Lon.	1725	465	374	344	996	1902	2257	0

Elevando al cuadrado cada entrada de esta matriz de manera de formar la matriz de distancias cuadráticas $\mathbf{D}$, y calculando la matriz de similaridad mediante [7] $\mathbf{Q} = -\frac{1}{2}\mathbf{P}\,\mathbf{D}\,\mathbf{P}$. Puesto que esta matriz tendrá sus entradas muy grandes, dividimos cada entrada por el factor 10^7, y obtenemos la matriz de similaridad (que por abuso de notación denotaremos con la misma letra $\mathbf{Q}$)

$$\mathbf{Q} = \begin{pmatrix} 0.141 & -0.0108 & -0.0433 & -0.0599 & -0.141 & -0.0408 & 0.206 & -0.0517 \\ -0.0108 & -0.0042 & 0.0005 & 0.0065 & 0.0038 & -0.0034 & -0.0058 & 0.0134 \\ -0.0433 & 0.0005 & 0.0139 & 0.0249 & 0.0422 & -0.0104 & -0.0541 & 0.0263 \\ -0.0599 & 0.0065 & 0.0249 & 0.0397 & 0.0628 & -0.0341 & -0.0802 & 0.0403 \\ -0.141 & 0.0038 & 0.0422 & 0.0628 & 0.132 & 0.0514 & -0.194 & 0.0431 \\ -0.0408 & -0.0034 & -0.0104 & -0.0341 & 0.0514 & 0.203 & -0.113 & -0.0526 \\ 0.206 & -0.0058 & -0.0541 & -0.0802 & -0.194 & -0.113 & 0.313 & -0.0715 \\ -0.0517 & 0.0134 & 0.0263 & 0.0403 & 0.0431 & -0.0526 & -0.0715 & 0.0526 \end{pmatrix}$$

No resulta complicado encontrar los autovalores de esta matriz, y según nuestro programa los dos mayores autovalores son

$$\lambda_1 = 0.6613132012\,;\ \lambda_2 = 0.2352207888$$

Y los correspondientes autovectores asociados a estos autovalores respectivamente son

$$\mathbf{v}_1 = \begin{pmatrix} -0.4579 \\ 0.0197 \\ 0.1264 \\ 0.1710 \\ 0.4419 \\ 0.2280 \\ 0.6871 \\ 0.1491 \end{pmatrix} \quad y \quad \mathbf{v}_2 = \begin{pmatrix} -0.1279 \\ 0.0475 \\ 0.1482 \\ 0.2943 \\ 0.0763 \\ -0.8516 \\ 0.0387 \\ 0.3745 \end{pmatrix}$$

[7] En nuestro caso la calculamos mediante $q_{ij} = -\frac{1}{2}\left(d^2_{ij} - d^2_{\cdot j} - d^2_{i\cdot} + d^2_{\cdot\cdot}\right)$. Se puede consultar el programa **mapaeuropa.dfw** ubicado en la red Internet:
http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/mapaeuropa.mth

De tal forma que formamos las matrices $\mathbf{V}_2$ y $\mathbf{\Lambda}_2$ que indica la relación (9), esto es

$$\mathbf{V}_2 = \begin{pmatrix} -0.4579 & -0.1279 \\ 0.0197 & 0.04753 \\ 0.1264 & 0.1482 \\ 0.1800 & 0.2943 \\ 0.4419 & 0.0763 \\ 0.2280 & -0.8516 \\ 0.6871 & 0.0387 \\ 0.1491 & 0.3745 \end{pmatrix}$$

$$\mathbf{\Lambda}_2^{1/2} = \begin{pmatrix} \sqrt{0.6613} & 0 \\ 0 & \sqrt{0.2352} \end{pmatrix}$$

y de esta manera calculamos

$$\mathbf{Y}_2 = \mathbf{V}_2\mathbf{\Lambda}_2^{1/2} = \begin{pmatrix} -0.4579 & -0.1279 \\ 0.0197 & 0.0475 \\ 0.1264 & 0.1482 \\ 0.1800 & 0.2943 \\ 0.4419 & 0.0763 \\ 0.2280 & -0.8516 \\ 0.6871 & 0.0387 \\ 0.1491 & 0.3745 \end{pmatrix} \begin{pmatrix} \sqrt{0.6613} & 0 \\ 0 & \sqrt{0.2352} \end{pmatrix}$$

$$= \begin{pmatrix} -.37237 & -6.2028 \times 10^{-2} \\ .01602 & 2.3036 \times 10^{-2} \\ .10279 & 7.1873 \times 10^{-2} \\ .14638 & .14273 \\ .35935 & 3.7004 \times 10^{-2} \\ .18541 & -.413 \\ .55875 & 1.8769 \times 10^{-2} \\ .12125 & .18162 \end{pmatrix}$$

De manera que, las ciudades tendrán las siguientes coordenadas:

Madrid	-0.37237	-0.0620
París	0.01602	0.0230
Brusela	0.10279	0.0718
Amsterdan	0.14638	0.14273
Berlín	0.35935	0.0370
Roma	0.18541	-0.413
Lisboa	0.55875	0.0187
Londres	0.12125	0.18162

Ahora si graficamos estas coordenadas en los ejes formado por la primera y segunda coordenada, como lo indica la figura 1, y le echamos una rápida visita a un atlas geográfico, nos damos cuanta que aproximadamente tenemos la configuración espacial de la distribución de las ciudades en al mapa de Europa.

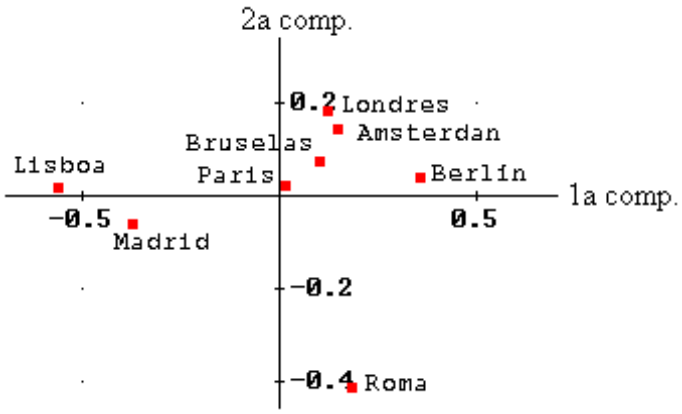

Figura 1

La bondad de este ejemplo es que vislumbra los factores latentes que estaban escondidos ante nuestros ojos pero que sin embargo subyacen en la matriz de distancia. Por lo demás este ejemplo nos da confianza para la utilización del escalado multidimensional, esto es que muestra la estructura subyacente en una matriz de distancia o en una matriz de similaridad. ■

6.2 Los libros

Vamos a considerar cinco libros de escritores sudamericanos, a saber: *Rayuela* (J. Cortázar), *Eva Luna* (Isabel Allende), *El túnel* (Ernesto Sábato), *El coronel no tiene quien le escriba* (Gabriel García Marquez) y *Palomita Blanca* (Enrique Lafourcade). Sobre cada uno de estos libros vamos a calcular la frecuencia relativa de cada una de las 27 letras del alfabeto español (a, b, c, ..., x, y, z). Y sobre esta matriz de datos vamos a calcular las distancias euclídeas entre los 5 libros, en $\mathbb{R}^{27}$, luego calcularemos la matriz de similaridad y finalmente las dos primeras coordenadas principales, para detectar una eventual estructura subyacente de "estilo" entre estos libros

Los datos de las frecuencias absolutas y relativas para cada letra en cada libro lo puede obtener del sitio Internet (fuente propia)[8], tanto en formato Excel (.xls) o en formato texto (.txt). La matriz de las distancias cuadráticas $\mathbf{D}$ así como la matriz de similaridad $\mathbf{Q}$ se pueden calcular con el programa **distancia.mth** realizado en el software DERIVE y también disponible en la red Internet[9], o por lo demás los cálculos los puede realizar en STATGRAPHICS, MINITAB o SPSS. En el programa distancia.mth obtenemos los siguientes autovalores asociados a la matriz de similaridad $\mathbf{Q}$

$$\lambda_1 \quad = \quad 0.001511200970$$

[8] http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/libros.xls
http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/libritos.txt

[9] http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/distancia.mth

$$
\begin{aligned}
\lambda_2 &= 0.0004483010966 \\
\lambda_3 &= 0.0002597095561 \\
\lambda_4 &= 4.024370405 \cdot 10^{-5} \\
\lambda_5 &= -7.257711498 \cdot 10^{-16}
\end{aligned}
$$

Como podemos observar, los autovalores están ordenados de mayor a menor y, en la práctica, los dos últimos son nulos. Vamos a encontrar los autovectores asociados a los dos primeros autovalores. Estos son

$$
\mathbf{v}_1 = \begin{pmatrix} 0.06736836919 \\ 0.03015966605 \\ 0.04481713837 \\ 0.7720335746 \\ 0.6296884000 \end{pmatrix} ; \ \mathbf{v}_2 = \begin{pmatrix} 0.06971179556 \\ 0.4845320351 \\ -0.8462580176 \\ 0.1178983993 \\ 0.1741157876 \end{pmatrix}
$$

De manera que si definimos $\mathbf{V}_2$ y $\mathbf{\Lambda}^{1/2}$, obtenemos el producto

$$
\mathbf{V}_2\mathbf{\Lambda}^{1/2} = \begin{pmatrix} 0.002618889329 & 0.001476016354 \\ 0.001172431937 & 0.01025905590 \\ 0.001742228984 & -0.01791792427 \\ -0.03001216320 & 0.002496277195 \\ 0.02447861291 & 0.003686574818 \end{pmatrix}
$$

que nos entregan las coordenadas principales. Graficamos estos puntos en el plano constituido por la primera y segunda coordenada, eje horizontal y vertical, respectivamente (Figura 2)

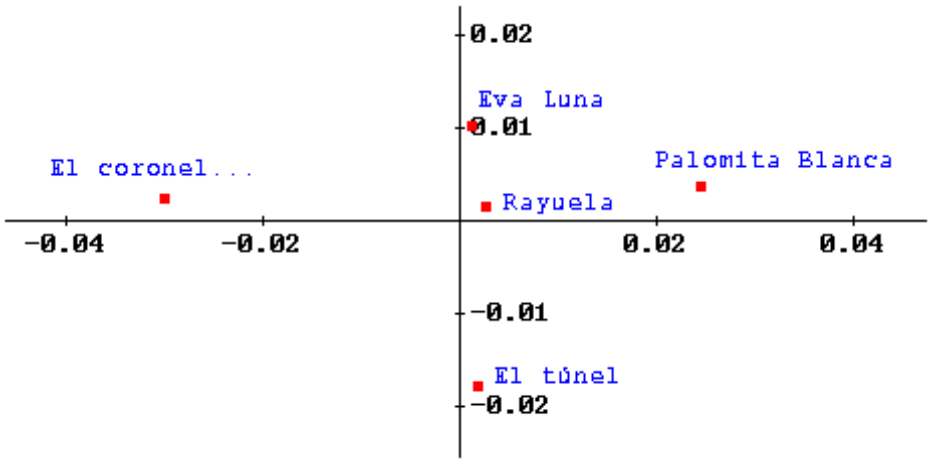

Figura 2

Podemos observar que, en relación a la primera componente *Eva Luna*, *Rayuela* y *El túnel* están al "mismo nivel". *El coronel no tiene quien le escriba* y *Palomita Blanca*, se

ubican en una suerte de antípodas...[10] ■

[10] Juzge el atento lector qué libro es mejor.