Relación entre coordenadas principales y componentes principales: análisis de textos literarios

Eliseo Martínez H.

1. Coordenadas y componentes principales

Cuando tenemos la matriz de datos $\mathbf{X}$, y por ende la matriz $\widetilde{\mathbf{X}}$ podemos realizar, si la situación lo amerita, un análisis de escalado multidimensional, construyendo la matriz $\mathbf{D}$ de las distancias cuadráticas entre los puntos (filas) de la matriz $\widetilde{\mathbf{X}}$. Lo que vamos a demostrar en esta sección es que las coordenadas principales obtenidas de la matriz $\mathbf{D}$ son equivalentes a los componentes principales de las variables de la matriz $\widetilde{\mathbf{X}}$.

Ya sabemos que los componentes principales corresponden a los vectores propios de la matriz $\mathbf{S} = \frac{1}{n-1}\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$. Por otro lado, las coordenadas principales son los vectores propios estandarizados por los valores propios obtenidos de la matriz $\mathbf{Q} = \widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$. Vamos a probar que tanto la matriz $\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$ como la matriz $\widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$ tienen el mismo rango y los mismos valores propios no nulos.

Supongamos que $\mathbf{a}_i$ es un autovector de la matriz $\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$ asociado al autovalor λ_i, esto significa que

$$\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}\,\mathbf{a}_i = \lambda_i\mathbf{a}_i$$

Si multiplicamos esta igualdad por $\widetilde{\mathbf{X}}$ en ambos miembros, obtenemos

$$\widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}\,\mathbf{a}_i = \lambda_i\,\widetilde{\mathbf{X}}\,\mathbf{a}_i \tag{1}$$

es decir el vector $\widetilde{\mathbf{X}}\,\mathbf{a}_i$ es un autovector de la matriz $\widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$ asociado al autovalor λ_i. Es decir hemos concluido que todo autovalor de $\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$ es un autovalor de la matriz $\widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$.

Supongamos ahora que[1] $n > p$ y la matriz $\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$ es de rango completo p, entonces ella tendrá p autovalores no nulos que además serán autovalores no nulos de la matriz $\widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$. Además de la relación (1) los vectores propios de $\widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$ son las proyecciones de la matriz $\widetilde{\mathbf{X}}$ sobre la dirección de los autovectores de la matriz $\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$.

Ahora bien, la matriz que entrega los valores de los p componentes principales en los n individuos o unidades muestrales observadas está dada por

$$\mathbf{Z} = \widetilde{\mathbf{X}}\,\mathbf{A} \tag{2}$$

donde $\mathbf{A}$ es una matriz de $p \times p$ cuyas columnas son los autovectores de la matriz $\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$ [2], y de tal modo que $\mathbf{Z}$ es la matriz de $n \times p$ y sus columnas son los componentes principales.

---

[1] No es estrictamente necesario que esto ocurra, en cualquier caso se quiere decir que si $\mathbf{X}$ es de rango completo entonces su rango es $min\,\{n, p\}$

[2] En rigor los autovalores de la matriz $\mathbf{S} = \frac{1}{n-1}\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$

Por otro lado, la matriz $n \times p$ de coordenadas principales viene dada por

$$\mathbf{Y} = \mathbf{V}\,\mathbf{L}$$

donde la matriz $\mathbf{V}$ es de $n \times p$ y sus columnas son los autovectores de la matriz $\widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$ asociados a sus autovalores no nulos, y $\mathbf{L}$ es la matriz diagonal de $p \times p$ con diagonal $\left\{\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_p}\right\}$. De otra forma

$$\mathbf{Y} = (\mathbf{v}_1, \ldots, \mathbf{v}_p) \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_p} \end{pmatrix}$$

donde $\mathbf{v}_i$ es el autovector de $\widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$ asociado al autovalor no nulo λ_i. Pero como hemos visto que los autovectores de $\widetilde{\mathbf{X}}\,\widetilde{\mathbf{X}}^t$ son de la forma $\widetilde{\mathbf{X}}\,\mathbf{a}_i$, deducimos que

$$(\mathbf{v}_1, \ldots, \mathbf{v}_p) = \left(\widetilde{\mathbf{X}}\,\mathbf{a}_1, \ldots, \widetilde{\mathbf{X}}\,\mathbf{a}_p\right) = \widetilde{\mathbf{X}}\,\mathbf{A}$$

y entonces

$$\mathbf{Y} = \widetilde{\mathbf{X}}\,\mathbf{A}\,\mathbf{L} \tag{3}$$

Comparando las igualdades (2) y (3) vemos que, aparte del factor de escala dado por la matriz diagonal $\mathbf{L}$, ambos procedimientos conducen esencialmente al mismo resultado.

2. El ejemplo de los libros

Vamos a considerar cinco libros de escritores iberoamericanos. A saber: *Rayuela* de Cortazar, *Eva Luna* de Isabel Allende, *El túnel* de Sabato, *El coronel no tiene quien le escriba* de García Márquez, y *Palomita Blanca* de Enrique Lafourcade. La matriz de datos[3] que vamos a considerar son las frecuencias relativas de cada una de las 27 letras del abecedario español, para cada uno de los libros. De manera que la matriz de datos $\mathbf{X}$ es de dimensión 5×27. En la subsección 6.2 del artículo **Escalado multidimensional** [4] se entregan los autovalores no nulos de la matriz $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^t$, y estos son

$$\begin{aligned} \lambda_1 &= 0.001511 \\ \lambda_2 &= 0.000448 \\ \lambda_3 &= 0.000259 \\ \lambda_4 &= 0.000040 \end{aligned}$$

Y el diagrama de dispersión en las dos primera coordenadas luce como lo indica la Figura 1.

Ahora si para la matriz de datos $\mathbf{X}$ realizamos un análisis de componentes principales, en el presente trabajo se ocupó el STATGRAPHICS, empezamos con el cálculo de los au-

[3] Los datos los puede obtener en:
http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/libros.xls

[4] Ubicado en Internet:
http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/escalado.pdf

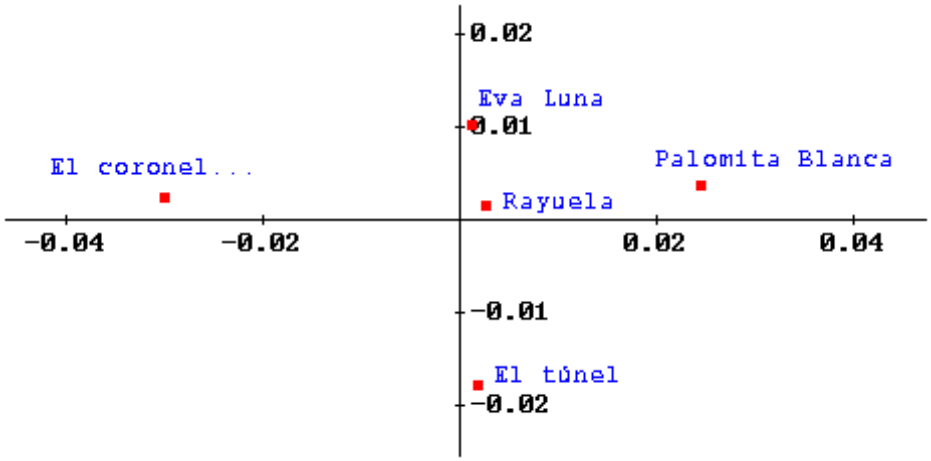

Figure 1:

tovalores de la matriz $\mathbf{S} = \frac{1}{n-1}\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$, y éstos son

$$\begin{aligned} \mu_1 &= 0.003778 \\ \mu_2 &= 0.0001121 \\ \mu_3 &= 0.00006493 \\ \mu_4 &= 0.00001006 \end{aligned}$$

Podemos notar que si multiplicamos por 4 (puesto que $n = 4$) los μ_i obtenemos los autovalores no nulos de la matriz $\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$, que son los mismos λ_i de la matriz $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^t$. Continuando con el análisis en componentes principales, selecionamos los dos componentes principales asociados a los dos mayores autovalores μ_1 y μ_2, y realizamos los cálculos para obtener las componentes, esto es

$$\mathbf{Z} = \mathbf{X}\ \mathbf{A}$$

entendiendo que $\mathbf{A}$ es la matriz que tiene los dos autovalores unitarios asociados a μ_1 y μ_2. Las componentes de $\mathbf{Z}$ se entregan en el **apéndice A**, y allí mostramos la ponderación en primera y segunda componente para cada letra.

Por otro lado, a la luz de la Figura 1, vemos que los puntos *El coronel...* y *Palomita Blanca* difieren notablemente, y es en este caso en que los componentes principales pueden dar luz sobre tal diferencia. En efecto, en el **apéndice B** agregamos la columna frecuencia promedio y las columnas de frecuencias de *El coronel ...* y *Palomita Blanca*, y allí podemos ver el significado estructural de la primera componente y así explicar la diferencia entre ambos libros. En efecto, la primera componente es una suerte de diferencia de promedio entre dos grupos de letras. A saber las letras $\{A, B, Y, Q, A\}$ son las que tienen mayor peso en la primera componente, y las de mayor peso negativo es el grupo $\{S, D, R, E, L\}$. Podemos observar que para el primer grupo las correspondientes frecuencias de *El coronel...* estan

bajo el promedio, y *Palomita Blanca* está sobre el promedio; y ocurre lo contrario para el segundo grupo de letras, allí *El coronel ...* está sobre el promedio mientras que *Palomita Blanca* está sobre el promedio. Esto es objetivamente, bajo el contexto de este análisis en componentes principales, lo que explica la ”gran distancia” entre ambos libros.

Appendix A. Primera y segunda componente

Letras	primera	segunda
A	0.200197	0.342079
B	0.349111	0.0204538
C	-0.154705	0.0214874
D	-0.208172	0.0676016
E	-0.251575	-0.531374
F	-0.0114226	-0.013698
G	-0.0307471	-0.0193097
H	0.0301127	-0.0591016
I	-0.0112432	-0.229856
J	0.000435949	0.0929849
K	0.00198435	0.00589788
L	-0.39865	0.561008
M	0.171263	-0.2351
N	-0.141599	-0.0780535
Ñ	-0.010192	0.0165176
O	0.478039	0.155697
P	-0.0915446	0.00478619
Q	0.223158	-0.155791
R	-0.241414	-0.0610521
S	-0.198181	0.211132
T	-0.0734074	-0.158802
U	0.0965764	-0.0907963
V	-0.0054949	0.0387
W	0.00175275	0.0015974
X	-0.0118145	-0.0373636
Y	0.311815	0.141139
Z	-0.0242834	-0.0107755

Appendix B. Orden descendente según primera componente

Letras	primera	segunda	frec. prom.	El coronel ...	Palomita ...
O	0.478039	0.155697	0.094457694	0.075556377	0.10100544
B	0.349111	0.0204538	0.016790821	0.001517991	0.019618475
Y	0.311815	0.141139	0.011378133	0.006243565	0.024694705
Q	0.223158	-0.155791	0.011048524	0.006507564	0.019185842
A	0.200197	0.342079	0.134250184	0.126877689	0.138010002
M	0.171263	-0.2351	0.031341744	0.028472241	0.038198631
U	0.0965764	-0.0907963	0.041925327	0.041672167	0.047549277
H	0.0301127	-0.0591016	0.010263597	0.009609546	0.011265769
K	0.00198435	0.00589788	0.000386148	3.95998E-05	8.65266E-05
W	0.00175275	0.0015974	0.000173225	1.31999E-05	8.65266E-05
J	0.000435949	0.0929849	0.006106792	0.008329153	0.009033382
V	-0.0054949	0.0387	0.011511661	0.011431136	0.011121558
Ñ	-0.010192	0.0165176	0.001564441	0.00174239	0.001165226
I	-0.0112432	-0.229856	0.062457934	0.061617255	0.060418674
F	-0.0114226	-0.013698	0.005914921	0.005121571	0.00418789
X	-0.0118145	-0.0373636	0.001030329	0.001121994	0.000357643
Z	-0.0242834	-0.0107755	0.004082823	0.004514375	0.003086117
G	-0.0307471	-0.0193097	0.010722524	0.011497136	0.0097429
T	-0.0734074	-0.158802	0.041174082	0.041764566	0.0371257
P	-0.0915446	0.00478619	0.026031908	0.028102643	0.022877646
N	-0.141599	-0.0780535	0.066726614	0.071213601	0.063349043
C	-0.154705	0.0214874	0.039557313	0.045024948	0.036710372
S	-0.198181	0.211132	0.070786581	0.075094379	0.063983571
D	-0.208172	0.0676016	0.046796686	0.05175691	0.04001569
R	-0.241414	-0.0610521	0.065862292	0.071332401	0.057470999
E	-0.251575	-0.531374	0.130730343	0.140539613	0.126582717
L	-0.39865	0.561008	0.056927357	0.07328599	0.053069677