Aplicación de la descomposición en valores singulares al análisis de datos

David Moreno Maldonado

Tutora: Amparo Baíllo Moreno

Trabajo de Fin de Grado
Doble Grado en Ingeniería Informática y Matemáticas
Universidad Autónoma de Madrid

2 Junio, 2020

- Nuestra capacidad de recolectar y almacenar grandes cantidades de datos observados ha aumentado enormemente.
- Reducir la dimensión de los datos puede facilitar el análisis de la información muestral.
- La descomposición en valores singulares (SVD) es una técnica que permite resumir y comprimir la información contenida en la matriz de datos.
- Las técnicas estudiadas que hacen uso de la SVD son:
  - Componentes principales.
  - Correlaciones canónicas.
  - Aproximación y compleción de matrices.
- Se ha estudiado la base teórica detrás de cada una de las técnicas y, después, se han aplicado en datos reales y actuales.

Definición

La **descomposición en valores singulares** *(SVD) de una matriz* $\mathbf{A}$ *de dimensiones* $m \times n$ *es la factorización*

$$\mathbf{A} = \mathbf{UDV}'$$

donde:

- $\mathbf{U}$ *es una matriz* $m \times n$ *ortogonal* $(\mathbf{U}'\mathbf{U} = \mathbf{I}_n)$ *cuyas columnas* $\mathbf{u}_j \in \mathbb{R}^m$ *son los autovectores ortonormales de* $\mathbf{AA}'$*;*
- $\mathbf{V}$ *es una matriz* $n \times n$ *ortogonal* $(\mathbf{V}'\mathbf{V} = \mathbf{I}_n)$ *cuyas columnas* $\mathbf{v}_j \in \mathbb{R}^n$ *son los autovectores ortonormales de* $\mathbf{A}'\mathbf{A}$*;*
- $\mathbf{D}$ *es* $n \times n$ *y diagonal, cuyos componentes de la diagonal* $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ *son los valores singulares de la matriz.*

Los valores singulares de $\mathbf{A}$ que son no nulos coinciden con la raíz cuadrada de los autovalores no nulos de $\mathbf{A}'\mathbf{A}$ o $\mathbf{AA}'$.

Componentes principales

Definición

Las **componentes principales** *de una matriz de datos* $\mathbf{X}$ *con* n *observaciones* p*-variantes son las combinaciones lineales incorreladas* $\mathbf{Y}_1 = \mathbf{X}\mathbf{t}_1, \ldots, \mathbf{Y}_p = \mathbf{X}\mathbf{t}_p$ *tales que la varianza muestral de cada* $\mathbf{Y}_i$ *es máxima condicionado a* $\mathbf{t}_i'\mathbf{t}_i = 1$ *y a que* $cov(\mathbf{Y}_i, \mathbf{Y}_j) = 0$ *para todo* $j < i$

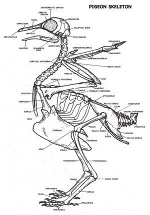
Utilizamos la descomposición espectral, un caso particular de la SVD para matrices simétricas, para determinar de las componentes principales.

Teorema

Dada la matriz de covarianzas muestral $\mathbf{S}$ *de la matriz de datos* $\mathbf{X}$*, si obtenemos su descomposición espectral* $\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{U}'$*, las componentes principales de* $\mathbf{X}$ *son las combinaciones lineales* $\mathbf{Y} = \mathbf{X}\mathbf{U}\mathbf{D}$.

Descripción de los datos

- Se disponen un total de 413 observaciones de diferentes tipos de aves.
- Las medidas disponibles se corresponden a la longitud y el diámetro de estos huesos:
    - Húmero
    - Cúbito
    - Fémur
    - Tibiotarso
    - Tarsometatarso



- Se realizará un análisis por separado de las longitudes y de los diámetros.

Resultados longitudes de los huesos

Criterio del porcentaje: Dados los porcentajes acumulados que cada componente principal $\mathbf{Y}_i$ explica de la variabilidad total, $P_i = 100\frac{\sigma_1+\cdots+\sigma_m}{\sigma_1+\cdots+\sigma_p}$; podemos especificar un porcentaje r y tomar las m primeras componentes principales tal que $P_m > r$

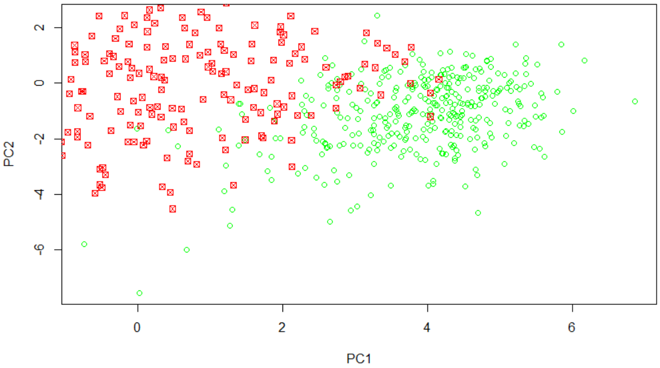
Variable	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**
Húmero	0.602	-0.200	0.472	-0.266	-0.551
Cúbito	0.649	-0.434	-0.369	0.304	0.403
Fémur	0.187	0.257	-0.674	-0.649	-0.154
Tibiotarso	0.375	0.668	0.351	-0.117	0.525
Tarsometatarso	0.202	0.509	-0.250	0.635	-0.484
Porcentaje de variación	89.81%	7.98%	1.19%	0.68%	0.31%
Porcentaje acumulado	89.81%	97.79%	98.99%	99.68%	100%

Descripción de los datos

Imágenes digitalizadas de biopsias de senos. Las medidas tomadas se corresponden con el núcleo celular y son las siguientes:

- Radio: Media de las distancias del centro al perímetro.
- Textura: Desviación estándar de la escala de grises de la imagen.
- Perímetro.
- Área.
- Suavidad: Variación local de la longitud del radio.
- Compacidad: $\frac{\text{perimetro}^2}{\text{area}-1}$.
- Concavidad: Agudeza de las porciones cóncavas del contorno.
- Puntos de concavidad: Número de porciones cóncavas del contorno.
- Simetría.
- Dimensión fractal.

Aplicación de la transformación por componentes principales de células malignas.



Correlaciones canónicas (1/2)

Sean $\mathbf{X}$ e $\mathbf{Y}$ dos matrices $n \times p$ y $n \times q$, respectivamente, de observaciones de dos vectores en una muestra de n individuos. Existen $m = \min(p, q)$ pares de vectores canónicos $\mathbf{a}_1, \mathbf{b}_1, \ldots, \mathbf{a}_m, \mathbf{b}_m$ que proporcionan:

$$
\begin{array}{ccc}
U_1 = \mathbf{X}\mathbf{a}_1, & V_1 = \mathbf{Y}\mathbf{b}_1, & r_1 = \mathrm{cor}(U_1, V_1) \\
\vdots & \vdots & \vdots \\
U_m = \mathbf{X}\mathbf{a}_m, & V_m = \mathbf{Y}\mathbf{b}_m, & r_m = \mathrm{cor}(U_m, V_m)
\end{array}
$$

de tal manera que cada r_i es máxima en cada caso y $\mathrm{cor}(U_i, U_j) = 0$ y $\mathrm{cor}(V_i, V_j) = 0$ para todo $j < i$.
Cada par de U_i y V_i definen las i-ésimas **variables canónicas** y cada r_i es la i-ésima **correlación canónica**.

Correlaciones canónicas (2/2)

Denotamos $\mathbf{S}_{11}$ y $\mathbf{S}_{22}$ las matrices de covarianzas muestrales de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente. Y las matrices de covarianzas cruzadas $\mathbf{S}_{12} = \mathbf{S}'_{21}$. Consideramos la matriz $p \times q$:

$$\mathbf{Q} = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2},$$

Sea $\mathbf{Q} = \mathbf{UDV}'$ la descomposición en valores singulares de esta matriz.

Teorema

Los vectores canónicos y correlaciones canónicas son

$$\mathbf{a}_i = \mathbf{S}_{11}^{-1/2}\mathbf{u}_i, \quad \mathbf{b}_i = \mathbf{S}_{22}^{-1/2}\mathbf{u}_i, \quad r_i = \sigma_i,$$

donde $\mathbf{u}_i$ *son las columnas de* $\mathbf{U}$ *y* $\mathbf{D} = diag(\sigma_1, \sigma_2, \ldots)$ *con* $\sigma_i \geq \sigma_{i+1}$ *para todo* i.

Índices sobre el nivel educativo

- ALF-JOVEN = Ratio de alfabetismo en mujeres jóvenes (% de mujeres entre 15 y 24 años).
- ALF-ADULTA = Ratio de alfabetismo en mujeres adultas (% de mujeres mayores de 15 años).
- EST-PROF-PRIM = Ratio estudiante-profesor en educación primaria.
- EST-PROF-SEC = Ratio estudiante-profesor en educación secundaria.
- INSC-PRIM = Porcentaje de inscripción femenino en educación primaria.
- INST-SEC = Porcentaje de inscripción femenino en educación secundaria.
- PROFESORAS = Porcentaje de profesoras en la educación secundaria.

Índices sobre los derechos y libertades de las mujeres

Indican el porcentaje de mujeres que creen que está justificado que su marido la golpee si:

- Ella discute con él. (DISCUTIR)
- A ella se le quema la comida.(COMIDA)
- Ella desatiende a los hijos. (HIJOS)
- Ella sale de casa sin consultárselo. (SALIR)
- Ella se niega a mantener relaciones sexuales con él. (RELACIONES)

Resultados de la aplicación de correlaciones canónicas

Variable	**CC1**	**CC2**
ALF-JOVEN	-0.023	-0.030
ALF-ADULTA	0.009	0.019
EST-PROF-PRIM	0.006	-0.010
EST-PROF-SEC	0.004	0.007
INSC-PRIM	-0.002	0.011
INSC-SEC	0.008	0.000
PROFESORAS	-0.001	0.006
DISCUTIR	-0.027	-0.028
COMIDA	0.032	0.010
HIJOS	-0.031	0.001
SALIR	0.040	0.024
RELACIONES	-0.002	-0.019
Correlación canónica	0.928	0.856

Problema aproximación de matrices

Sea una matriz $\mathbf{A}$ de dimensiones $m \times n$, el problema de optimización para aproximarla es

$$\hat{\mathbf{A}} = \underset{\mathbf{M}\in\mathbb{R}^{m\times n}}{\arg\min} \|\mathbf{A} - \mathbf{M}\|^2 \text{ sujeto a } \Phi(\mathbf{M}) \leq c,$$

donde la función Φ sirve para establecer una restricción que hace que la matriz $\hat{\mathbf{A}}$ tenga muchos ceros (*sparse*).

Teorema de la aproximación

Teorema

Dada una matriz $\mathbf{A}$ de dimensiones $m \times n$, la matriz de rango menor o igual que k (expresada en la forma $\sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i'$) que mejor aproxima a $\mathbf{A}$, en el sentido de que minimiza el error de aproximación

$$\left|\left|\mathbf{A} - \sum_{i=1}^{k} \mathbf{x}_i \mathbf{y}_i'\right|\right|$$

donde $||\mathbf{A}||$ es la norma de Frobenius de $\mathbf{A}$, viene dada por los k primeros términos de la SVD de la matriz $\mathbf{A}$, $\mathbf{A}_k$.

Resolvemos el problema de la aproximación tomando:

$$\mathbf{M} = \mathbf{A}_k,\ \ \Phi(\mathbf{M}) = \mathrm{rank}(\mathbf{M}) \text{ y } c = k \leq \mathrm{rank}(\mathbf{A})$$

Problema de compleción de matrices

Sea $\mathbf{A}$ la matriz con datos en un subconjunto $\Omega \subset \{1, \ldots, m\} \times \{1, \ldots, n\}$ y $\mathbf{M}$ una aproximación, el problema de optimización para la compleción de matrices es

$$\min_{\operatorname{rank}(\mathbf{M}) \leq r} \sum_{(i,j) \in \Omega} (a_{ij} - m_{ij})^2.$$

- El problema es no convexo y las soluciones globales no son posibles de manera general.
- Existen heurísticas que permiten encontrar mínimos locales de manera efectiva utilizando una estimación inicial.

Algoritmo iterativo del paquete `SoftImpute` de R para compleción de matrices

Sea $\mathbf{A} = (a_{ij})$ la matriz con datos observados en un subconjunto $\Omega \subset \{1, \ldots, m\} \times \{1, \ldots, n\}$, $\mathbf{M} = (m_{ij})$ una estimación inicial de los valores faltantes de $\mathbf{A}$ y $P_\Omega(\mathbf{A})$ la matriz con los valores de Ω mantenidos como en $\mathbf{A}$ y el resto igualados a 0. El algoritmo seguido es:

1. $\mathbf{Z} = P_\Omega(\mathbf{A}) + P_{\Omega^\perp}(\mathbf{M})$
2. Hallamos la SVD $\mathbf{Z} = \mathbf{UDV}'$
3. Obtenemos la matriz $\mathbf{D}_k$, igual a la matriz $\mathbf{D}$, pero igualando a cero todos los valores singulares que no sean los k primeros.
4. Construimos $\mathbf{Z}_k = \mathbf{UD}_k\mathbf{V}'$
5. Actualizamos la matriz $\mathbf{M} = \mathbf{Z}_k$

Datos de películas utilizados

Pulp fiction	Los juegos del hambre	Interestellar	El corredor del laberinto	El lobo de Wall Street	El viaje de Chihiro	Seven
4	3	5	2	4	5	5
4	2	5	2	5	4	4
5	3	5	3	4	4	4
3	3	5	2	4	5	4
4	3	1	1	3	4	N/A
4	3	N/A	N/A	3	4	4
5	2	5	2	4	5	5
4	3	5	3	4	3	3
5	3	4	1	3	4	4
4	2	5	4	N/A	5	N/A
4	4	4	4	4	4	4
5	4	5	N/A	5	5	5
5	1	4	N/A	4	N/A	4

Resultados de la aplicación de `SoftImpute`

Rango máximo	Iteraciones hasta convergencia	Diferencia total	Diferencia media	Diferencia mínima	Diferencia máxima
1	6	8.505	1.063	0.288	1.891
2	12	17.157	2.145	0.254	6.112
3	26	19.073	2.384	0.052	7.626
4	24	29.593	3.699	0.248	6.662
5	50	27.593	3.449	1.234	6.491
6	15	30.165	3.771	0.661	6.542

Muchas gracias.