Análisis de componentes principales

Matías Carrasco

9 de octubre de 2023

Índice

1. Introducción

El análisis de componentes principales (PCA en inglés) es una técnica de reducción de la dimensión que se aplica a tablas de datos donde las filas se consideran observaciones y las columnas deben ser variables continuas:

Denotamos $x_i^{(j)}$ el valor tomado por el individuo i para la variable j, donde i varía de 1 a N y j de 1 a D. De este modo la matriz de diseño es:

$$\boldsymbol{X} = \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(D)} \\ \vdots & x_i^{(j)} & \vdots \\ x_N^{(1)} & \cdots & x_N^{(D)} \end{bmatrix}$$

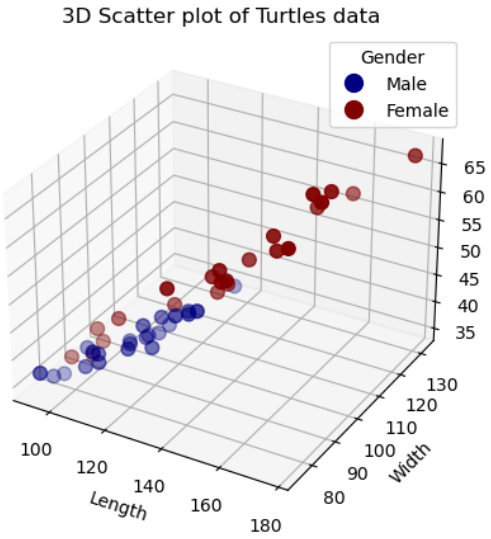

Figura 1: Izquierda: Tortuga Pintada (Fuente: Wikipedia); Derecha: Scatter plot 3D de la nube de individuos

También vamos usar la media y el desvío estándar de cada variable:

$$\bar{x}^{(j)} = \frac{1}{N}\sum_{i=1}^{N} x_i^{(j)}, \quad s_j^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i^{(j)} - \bar{x}^{(j)}\right)^2$$

A modo de ejemplo, utilicemos un conjunto de datos que contiene mediciones del caparazón de 24 tortugas pintadas machos y 24 hembras (Chrysemys picta marginata). El mismo contiene las siguientes 4 variables:

- Gender: Male/Female
- Length: carapace length
- Width: carapace width
- Height: carapace height

En este ejemplo la matriz de diseño $\boldsymbol{X}$ consta de $D = 3$ columnas o variables y $N = 48$ individuos u observaciones. También disponemos de una cuarta variable categórica que indica el sexo de la tortuga.

Cada fila de la matriz de datos es un vector de $\mathbb{R}^D$ que representa las mediciones realizadas sobre un mismo individuo. El conjunto de N puntos en el espacio $\mathbb{R}^D$ se conoce como *la nube de individuos*. El espacio $\mathbb{R}^D$ se llama el *espacio de individuos*. Ver la Fig. 1.

La distancia euclídea entre los individuos i y l está dada por

$$d(i,l) = \sqrt{\sum_{j=1}^{D} \left(x_i^{(j)} - x_l^{(j)} \right)^2}$$

Si dos individuos tienen valores similares en las D variables de la tabla, también están cerca en el espacio $\mathbb{R}^D$.

La forma de la nube sigue siendo la misma incluso cuando se la traslada. En PCA siempre es conveniente trabajar con los datos centrados, lo que corresponde a considerar $x_i^{(j)} - \bar{x}^{(j)}$ en lugar de $x_i^{(j)}$. El PCA estandarizado es cuando trabajamos con $\left(x_i^{(j)} - \bar{x}^{(j)} \right) / s_j$ que sí modifica la forma de la nube. Es la opción por defecto en la mayoría de los casos.

El objetivo del PCA es representar la nube de puntos en un espacio de dimensiones reducidas de forma óptima, es decir, distorsionando lo menos posible las distancias entre las observaciones.

2. La mejor representación en 1D

Para obtener esta representación en 1D, la nube se proyecta sobre una recta de $\mathbb{R}^D$ denotada r, elegida de tal manera que se minimice la distorsión de la nube de puntos, es decir, tal que las distancias entre los puntos proyectados sean lo más cercanas posible a las distancias entre los puntos iniciales.

Es decir, queremos minimizar

$$\min_r \left\{ \sum_{\{i,l\}} \left| d(i,l)^2 - d_r(i,l)^2 \right| \right\}$$

en donde $d_r(i,l)$ es la distancia entre los puntos proyectados.

Dado que, en la proyección ortogonal, las distancias solo pueden disminuir:

$$\begin{aligned} \min_r \left\{ \sum_{\{i,l\}} \left| d(i,l)^2 - d_r(i,l)^2 \right| \right\} &= \min_r \left\{ \sum_{\{i,l\}} d(i,l)^2 - d_r(i,l)^2 \right\} \\ &= \sum_{\{i,l\}} d(i,l)^2 - \max_r \left\{ \sum_{\{i,l\}} d_r(i,l)^2 \right\} \end{aligned}$$

Es decir, el problema es equivalente a maximizar $\sum_{\{i,l\}} d_r(i,l)^2$.

Llamemos r_i a la proyección ortogonal sobre r del individuo i. Entonces

$$\begin{aligned}\sum_{\{i,l\}} d_r(i,l)^2 &= \sum_{\{i,l\}} \|r_i - r_l\|^2 \\ &= \frac{1}{2}\sum_{i,l} \|r_i - r_l\|^2 = \frac{1}{2}\sum_{i,l} \|r_i\|^2 + \frac{1}{2}\sum_{i,l} \|r_l\|^2 - \sum_{i,l} r_i \cdot r_l \\ &= N\sum_{i} \|r_i\|^2 - \sum_{i,l} r_i \cdot r_l = N\sum_{i} \|r_i\|^2 - \left(\sum_i r_i\right)\cdot\left(\sum_l r_l\right) \\ &= N^2\left(\frac{1}{N}\sum_i \|r_i\|^2 - \left\|\frac{1}{N}\sum_i r_i\right\|^2\right) = N^2 \mathrm{Var}\left(\{r_i\}\right)\end{aligned}$$

En conclusión, la recta que distorsiona lo menos posible las distancias entre los puntos proyectados coincide con **la recta que maximiza la varianza** de los puntos proyectados.

3. La mejor representación en nD

Para obtener la representación en nD, la nube se proyecta sobre un hiperplano n-dimensional de $\mathbb{R}^D$ denotado H, elegido de tal manera que se minimice la distorsión de la nube de puntos. Al igual que en 1D, esto equivale a buscar el hiperplano que maximiza la variabilidad:

$$\min_H \left\{\sum_{\{i,l\}} \left|d(i,l)^2 - d_H(i,l)^2\right|\right\} = N^2 \max_H \left\{\mathrm{Var}\left(\{p_H(i)\}\right)\right\}$$

en donde p_H indica la proyección ortogonal sobre H.

Recordar que al proyectar ortogonalmente un vector A sobre un vector B, la longitud de la proyección está dada por $\|A\|\cos(\theta)$ en donde θ es el ángulo entre A y B, ver Fig 2. Si el vector B tiene norma 1, esta longitud es igual al producto escalar entre A y B:

$$A \cdot B = \langle A, B\rangle = A^t B$$

si pensamos a los vectores como matrices de una sola columna.

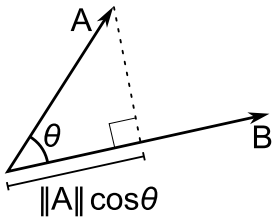

Figura 2: Producto escalar entre dos vectores

4. ¿Cómo encontrar la recta óptima?

Un individuo viene representado en el espacio de individuos $\mathbb{R}^D$ por una fila de la matriz de diseño $\boldsymbol{X}$. Sea $\boldsymbol{u}$ un vector unitario (una dirección) en $\mathbb{R}^D$. La proyección ortogonal del individuo i sobre $\boldsymbol{u}$ está dada por el producto escalar (fila i de $\boldsymbol{X}$)$\cdot\boldsymbol{u}$. Entonces, la proyección de la nube de individuos entera viene dada por el vector $\boldsymbol{X}\boldsymbol{u}$.

Cuando $\boldsymbol{X}$ está centrada, lo mismo ocurre con la proyección $\boldsymbol{X}\boldsymbol{u}$. De este modo la varianza de la proyección viene dada por $\frac{1}{N}\|\boldsymbol{X}\boldsymbol{u}\|^2$. Esto quiere decir que la recta de mayor variabilidad será aquella que maximiza

$$\max_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \|\boldsymbol{X}\boldsymbol{u}\|^2$$

Vamos a asumir de ahora en mas que X está centrada. Por lo que dijimos la varianza de la proyección está dada por

$$\|\boldsymbol{X}\boldsymbol{u}\|^2 = (\boldsymbol{X}\boldsymbol{u})^\top \boldsymbol{X}\boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{u}$$

La entrada k, l de la matriz $\boldsymbol{X}^\top\boldsymbol{X}$ es (recordar que $\boldsymbol{X}$ está centrada):

$$(\boldsymbol{X}^\top\boldsymbol{X})_{kl} = \sum_{i=1}^{N} \left(x_i^{(k)} - \bar{x}^{(k)}\right)\left(x_i^{(l)} - \bar{x}^{(l)}\right) = N\,\mathrm{cov}(k,l)$$

que es (a menos del factor N) la covarianza entre la variable k y la variable l.

Definimos la matriz de covarianzas S cuya entrada kl es

$$S_{kl} = \mathrm{cov}(k,l).$$

En el caso en que trabajemos con la matriz de diseño centrada, la matriz S puede calcularse como

$$\boldsymbol{S} = \frac{1}{N}\boldsymbol{X}^t\boldsymbol{X}$$

	Length	Width	Height		Length	Width	Height
Length	419.496	253.991	165.830	Length	1.000	0.978	0.965
Width	253.991	160.677	102.192	Width	0.978	1.000	0.961
Height	165.830	102.196	70.440	Height	0.965	0.961	1.000

Cuadro 1: Izquierda: La matriz $\boldsymbol{S}$ de covarianzas. Derecha: La matriz $\boldsymbol{R}$ de correlaciones.

Recordar que estamos buscando la dirección $\boldsymbol{u} \in \mathbb{R}^D$ que maximiza $\frac{1}{N}\|\boldsymbol{X}\boldsymbol{u}\|^2$, y por lo que vimos antes esto equivale a maximizar

$$\max_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u}$$

ya que $\frac{1}{N}\|\boldsymbol{X}\boldsymbol{u}\|^2 = \boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u}$.

Observar que en la diagonal de $\boldsymbol{S}$ aparecen las respectivas varianzas de las tres variables Length, Width, y Height. Observar también la simetría de ambas matrices. La matriz de correlaciones $\boldsymbol{R}$ es la matriz de covarianzas de la matriz de diseño estandarizada.

Por simplicidad, supongamos $D = 2$ variables. Así la matriz de covarianzas es una matriz 2×2:

$$\boldsymbol{S} = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix}$$

y queremos maximizar

$$F(\boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} = [u_1, u_2] \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = s_x^2 u_1^2 + 2 s_{xy} u_1 u_2 + s_y^2 u_2^2$$

con la restricción $N(\boldsymbol{u}) = u_1^2 + u_2^2 = 1$.

Recordar que el método de los multiplicadores de Lagrange consiste en introducir un multiplicador λ y resolver el sistema

$$\nabla F - \lambda \nabla N = 0$$

Calculamos los gradientes:

$$\nabla F = 2\left(s_x^2 u_1 + s_{xy} u_2, s_{xy} u_1 + s_y^2 u_2\right) = 2 \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 2\boldsymbol{S}\boldsymbol{u}$$

$$\nabla N = 2(u_1, u_2) = 2\boldsymbol{u}$$

de donde obtenemos la ecuación $\boldsymbol{S}\boldsymbol{u} = \lambda \boldsymbol{u}$. Esto quiere decir que la dirección óptima está dada por el vector propio de valor propio maximal.

5. Los valores y vectores propios de S

Toda matriz simétrica es diagonalizable en una base ortonormal. Además, si la matriz es definida no negativa, sus valores propios son no negativos. Esto implica que existe una base ortonormal $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_D\}$ de vectores propios de $\mathbb{R}^D$ (vendría a ser como rotar los ejes coordenados) con valores propios

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_D \geq 0$$

tales que $\boldsymbol{S}\,\boldsymbol{c}_j = \lambda_j\,\boldsymbol{c}_j$.

Si trabajamos en esta base, la varianza proyectada en la dirección $\boldsymbol{u}$ queda

$$\boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} = \lambda_1 u_1^2 + \cdots + \lambda_D u_D^2$$

Notar por un lado que

$$\boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} = \lambda_1 u_1^2 + \cdots + \lambda_D u_D^2 \geq \lambda_1 u_1^2$$

y tomando máximo en ambos lados

$$\max_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} \geq \max_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \lambda_1 u_1^2 \geq \lambda_1.$$

Por otro lado

$$\boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} = \lambda_1 u_1^2 + \cdots + \lambda_D u_D^2 \leq \lambda_1 (u_1^2 + \cdots + u_D^2) = \lambda_1,$$

y tomando máximo en ambos lados

$$\max_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} \leq \lambda_1.$$

Es decir,

$$\boxed{\max_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u} = \lambda_1 \text{ y la dirección óptima es } \boldsymbol{u} = \pm \boldsymbol{c}_1.}$$

En resumen, para encontrar la dirección de mejor representación, o lo que es equivalente de mayor variabilidad, debemos:

1. A partir de la matriz de diseño, calculamos la matriz de covarianzas $\boldsymbol{S}$. En caso de que $\boldsymbol{X}$ esté centrada, el cálculo es simplemente $\boldsymbol{S} = \frac{1}{N}\boldsymbol{X}^\top \boldsymbol{X}$.
2. Hallamos la base de vectores propios $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_D\}$ de $\boldsymbol{S}$ y los valores propios
$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_D \geq 0$$
3. La dirección óptima es $u = \pm \boldsymbol{c}_1$.
4. Los valores proyectados en esta dirección se obtienen haciendo $\boldsymbol{X}\,\boldsymbol{c}_1$.

La Fig.3 muestra el resultado en el ejemplo de las tortugas.

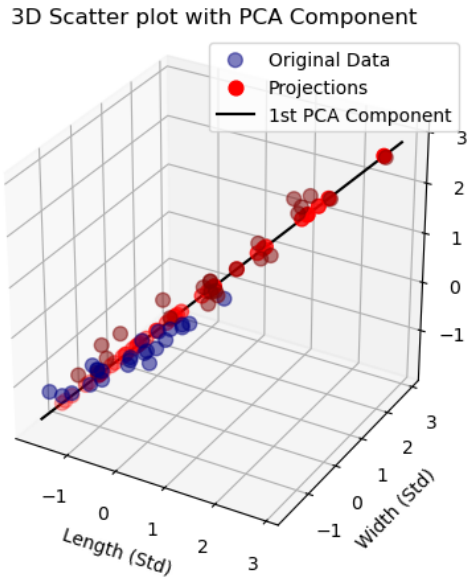

Figura 3: Primera componente principal y datos proyectados en el ejemplo de las tortugas.

6. Usando n componentes

Resulta que si en lugar de buscar la mejor recta buscamos el mejor hiperplano de dimensión n, el mismo argumento que antes muestra que se obtiene tomando el espacio generado por las primeras n componentes $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$. En la Fig. 4 se muestra cómo es el plano ($n = 2$) generado por las dos primeras componentes principales en el ejemplo de las tortugas.

7. Variabilidad y calidad de la representación

Una medida usual para la variabilidad total de la nube de puntos es la suma de las varianzas de cada variable:

$$\text{Variabilidad total} = \text{var}(1) + \cdots + \text{var}(D)$$

Este número es igual a la traza de la matriz de covarianzas $\boldsymbol{S}$ y se puede calcular como:

$$\text{tr}(\boldsymbol{S}) = \lambda_1 + \cdots + \lambda_D$$

Si usamos una representación con $n \leq D$ componentes principales, el coeficiente

$$\frac{\lambda_1 + \cdots + \lambda_n}{\text{tr}(\boldsymbol{S})} \times 100$$

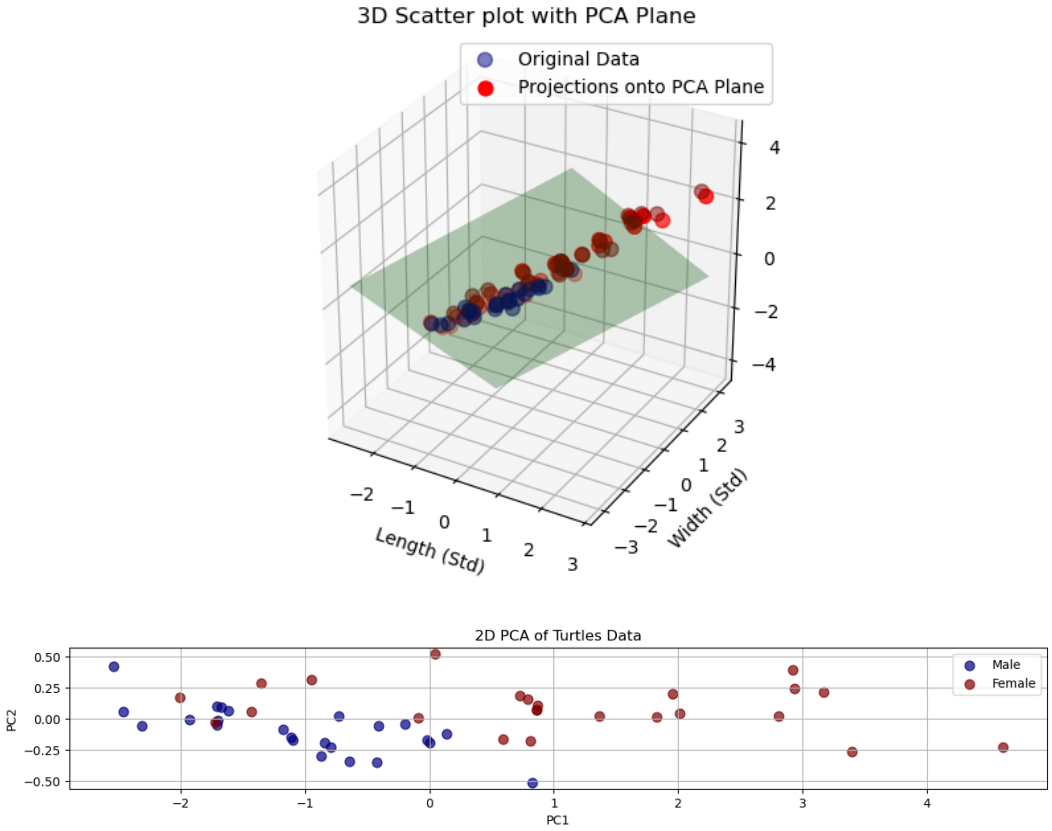

Figura 4: Arriba: plano generado por las dos primeras componentes. Abajo: scatter 2D de los puntos proyectados sobre dicho plano.

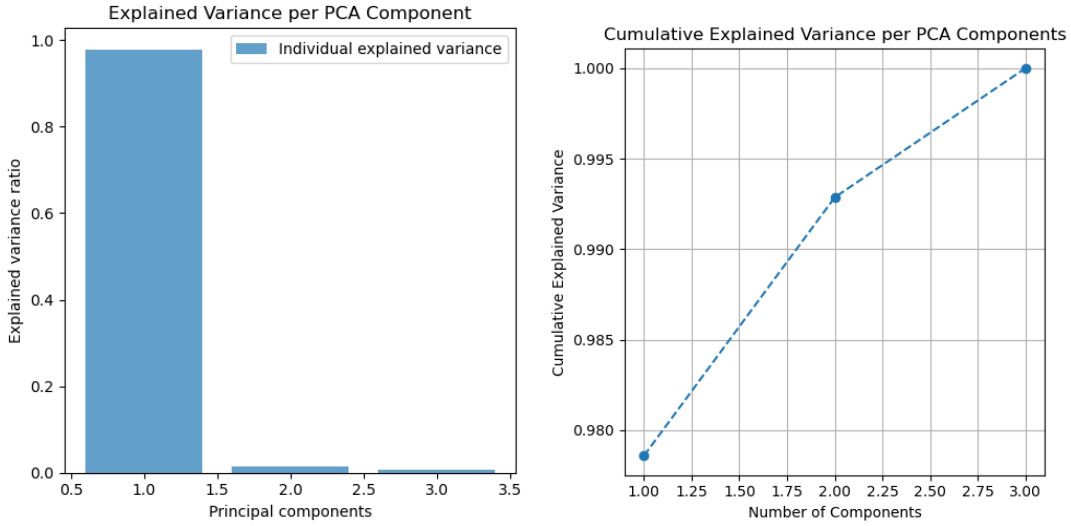

Figura 5: Varianza individual y varianza acumulada en función de las componentes.

es un indicador de la calidad de la representación. En la Fig. 5 se muestra la varianza individual de cada componente así como la varianza acumulada en función del número de componentes.