

Modelos Estadísticos para la Regresión y la Clasificación

Práctico 7 - Componentes Principales

Micaela Long

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

9 de octubre de 2024

Práctico 7

Componentes Principales

Material para hacer práctico 7 (disponible en EVA):

- Teóricos Análisis Componentes Principales (Tema 8)

Será de utilidad tener a mano el libro :

"An Introduction to Statistical Learning with Applications in R"

o su versión en Python:

"An Introduction to Statistical Learning with Applications in Python"

Ambos pueden ser descargados aquí: <https://www.statlearning.com/>

Análisis de Componentes Principales (PCA)

Si quiero bajar de 3 a 2 dimensiones y perder la menor cantidad de información posible:

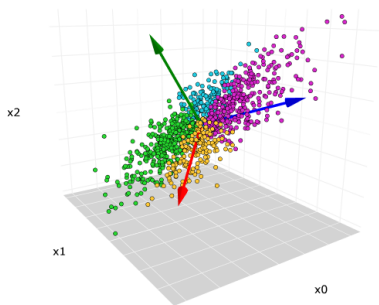


Figura: towardsdatascience.com

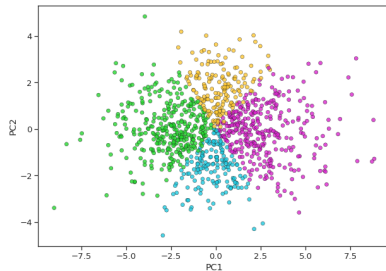


Figura: towardsdatascience.com

Análisis de Componentes Principales (PCA)

Si quiero bajar de 3 a 2 dimensiones y perder la menor cantidad de información posible:

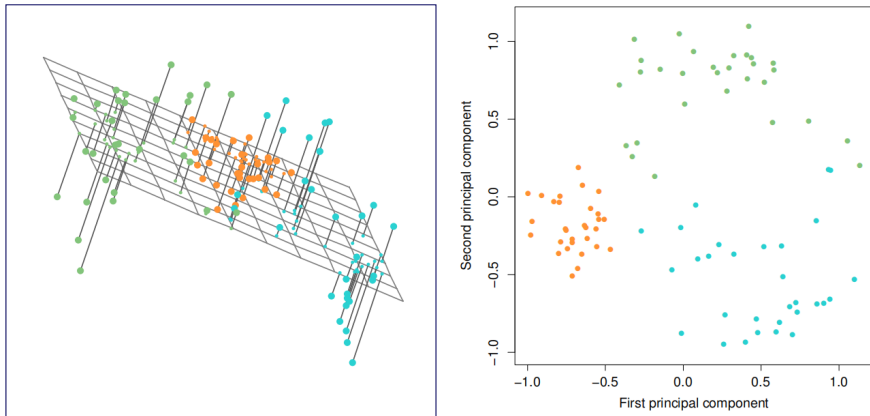


Figura: <https://www.statlearning.com/>

- El Análisis de Componentes Principales (PCA) es una técnica que en general se utiliza para reducir la dimensionalidad.
- Transforma a las variables originales en componentes principales (combinaciones lineales de las primeras)
- Las componentes principales (PC) capturan la mayor parte de la varianza en los datos, y son ortogonales
- Los componentes principales se ordenan por la cantidad de varianza que explican:
 - 1 PC1 es la dirección en el espacio de los datos que captura la mayor cantidad de varianza posible en los datos.
 - 2 PC2 es la segunda dirección más importante, que es ortogonal a PC1.
 - 3 PC3 es la tercera, ortogonal a PC1 y PC2, y así sucesivamente.

Análisis de Componentes Principales (PCA)

X_1, X_2, X_3 variables originales.

Si quiero considerar menos dimensiones, una forma inteligente es considerar

$$Z_1 = z_{11}X_1 + z_{12}X_2 + z_{13}X_3$$

A Z_1 le llamamos componente principal.

Queremos $z_1 = (z_{11}, z_{12}, z_{13})$ de forma tal que maximiza la varianza

$$z_1 = \arg \max_{\|z_1\|^2=1} \text{Var}(Z_1)$$

Se prueba que z_1 es el vector propio asociado al valor propio más grande de S la matriz de varianzas y covarianzas.

Luego

$$Z_2 = z_{21}X_1 + z_{22}X_2 + z_{23}X_3$$

donde $z_2 = (z_{21}, z_{22}, z_{23})$

$$z_2 = \arg \max_{\substack{\|z_2\|^2=1 \\ z_1 \perp z_2}} \text{Var}(Z_2)$$

que es el vector propio asociado al segundo valor propio más grande de S la matriz de varianzas y covarianzas.

Y así lo mismo para Z_3

Ejercicio 2

Consideramos la siguiente matriz de datos:

	x_1	x_2	x_3
x_1	1	0	0
x_2	1	2	0
x_3	2	2	2
x_4	0	0	2

- 1 Centrar y reducir la matriz.
- 2 Calcular la matriz de varianzas-covarianzas, sus valores propios y vectores propios.
- 3 Completar la tabla siguiente:

	Z_1	Z_2	Z_3	cal Z_1	cal Z_2	cal Z_3	ctr Z_1	ctr Z_2	ctr Z_3
x_1									
x_2									
x_3									
x_4									

donde cal Z_i es la calidad sobre el eje i (en %), y ctr Z_i es la contribución a la construcción del eje i (en %).

- 4 Hacer la representación gráfica de los individuos sobre el plano dado por Z_1 y Z_2 .

Ejercicio 2

	x_1	x_2	x_3
x_1	1	0	0
x_2	1	2	0
x_3	2	2	2
x_4	0	0	2

- 1 Filas \rightarrow observaciones
- 2 Columnas \rightarrow variables \rightarrow queremos estandarizarlas

$$\mu_{x_1} = \frac{1 + 1 + 2 + 0}{4} = 1$$

$$\sigma_{x_1} = \sqrt{\frac{(1-1)^2 + (1-1)^2 + (2-1)^2 + (0-1)^2}{3}} = 0,816$$

$$Y =$$

	x_1	x_2	x_3
x_1	$\frac{1-1}{0,816}$	—	—
x_2	$\frac{1-1}{0,816}$	—	—
x_3	$\frac{2-1}{0,816}$	—	—
x_4	$\frac{0-1}{0,816}$	—	—

Completar el resto de las columnas!

Y es la matriz centrada y reducida:

$$Y = \begin{pmatrix} \frac{1-1}{0,816} & - & - \\ \frac{1-1}{0,816} & - & - \\ \frac{2-1}{0,816} & - & - \\ \frac{0-1}{0,816} & - & - \end{pmatrix}$$

La matriz de varianzas y covarianzas es

$$S = \frac{1}{3} Y^T Y$$

Hay que hallar valores propios $\lambda_1 \geq \lambda_2 \geq \lambda_3$ y los vectores propios asociados v_1, v_2, v_3 de S .

$$Z_1 = Yv_1$$

- Z_1 es la primera componente principal.
- Maximiza la varianza y es una combinación lineal de las variables originales.
- Se calcula mediante la proyección de los datos sobre el primer vector propio de la matriz de covarianzas.

La calidad sobre el eje i (en %) se define como

$$\text{cal}Z_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} \times 100$$

y es el porcentaje de la varianza explicada por PC1.

La contribución a la construcción del eje i es

$$\text{ctr}Z_i = \frac{z_{ij}^2}{\sum_j z_{ij}^2}$$

La representación gráfica se hace proyectando los puntos en el plano definido por las componentes principales Z_1, Z_2 .