Componentes principales

Módulo 4

—

Análisis de datos multivariantes

Análisis de datos multivariantes

Uno de los objetivos principales del análisis de datos multivariantes es la reducción de la dimensionalidad que consiste en describir los valores de p variables por un pequeño subconjunto r de ellas. De esta forma, **se reduce la dimensión del problema con una pequeña perdida de información**

Componentes principales

El análisis de componentes principales (PCA) tiene como objetivo analizar si es posible representar adecuadamente la información de n observaciones de $\boldsymbol{p}$ variables con un número menor de variables conocidas como **Componentes Principales**, que son combinaciones lineales de las variables originales y que son independientes entre sí

$$
\begin{aligned}
Y_1 &= a_1' X = a_{11} X_1 + a_{12} X_2 + \cdots + a_{1p} X_p \\
Y_2 &= a_2' X = a_{21} X_1 + a_{22} X_2 + \cdots + a_{2p} X_p \\
&\vdots \qquad\qquad\qquad \vdots \\
Y_p &= a_p' X = a_{p1} X_1 + a_{p2} X_2 + \cdots + a_{pp} X_p
\end{aligned}
$$

Componentes principales

Cuando queremos analizar una gran cantidad de variables no es posible graficarlas, lo que dificulta tener una idea de las posibles tendencias que se presentan en los datos. Por esta razón, es útil aplicar el análisis de componentes principales ya que al reducir la dimensionalidad de los datos, será posible visualizarlos e identificar si hay grupos de individuos que son muy similares o muy diferentes entre si

Pasos para calcular los componentes principales

1. **<u>Estandarizar las variables originales</u>**

$$z_i = \frac{x_i - \bar{X}}{S}$$

Pasos para calcular los componentes principales

2.<u>Calcular la matriz de varianzas y covarianzas de la matriz de datos estandarizados</u>

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \cdots & s_{1p} \\ \vdots & & \vdots \\ s_{p1} & \cdots & s_p^2 \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{bmatrix}$$

Pasos para calcular los componentes principales

4.<u>Calcular los valores propios y vectores propios de la matriz de varianzas y covarianzas</u>

$\lambda_i \rightarrow$ Valores propios

$e_i \rightarrow$ Vectores propios

$\rightarrow (\lambda_i, e_i)$

La matriz de varianzas y covarianzas tiene tantos valores propios como filas (o columnas). Cada valor propio tiene un vector propio asociado.

Pasos para calcular los componentes principales

5.Ordenar las parejas de valores y vectores propios de mayor a menor según el valor del valor propio

$$(\lambda_1, e_1), \quad (\lambda_2, e_2), \quad \ldots, \quad (\lambda_p, e_p)$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

Pasos para calcular los componentes principales

6. Calcular los componentes principales

La i-ésima componente principal está dada por la siguiente combinación lineal:

$$Y_i = e_i^{'} X = e_{i1} X_1 + e_{i2} X_2 + \cdots + e_{ip} X_p$$