Reducción de dimensionalidad

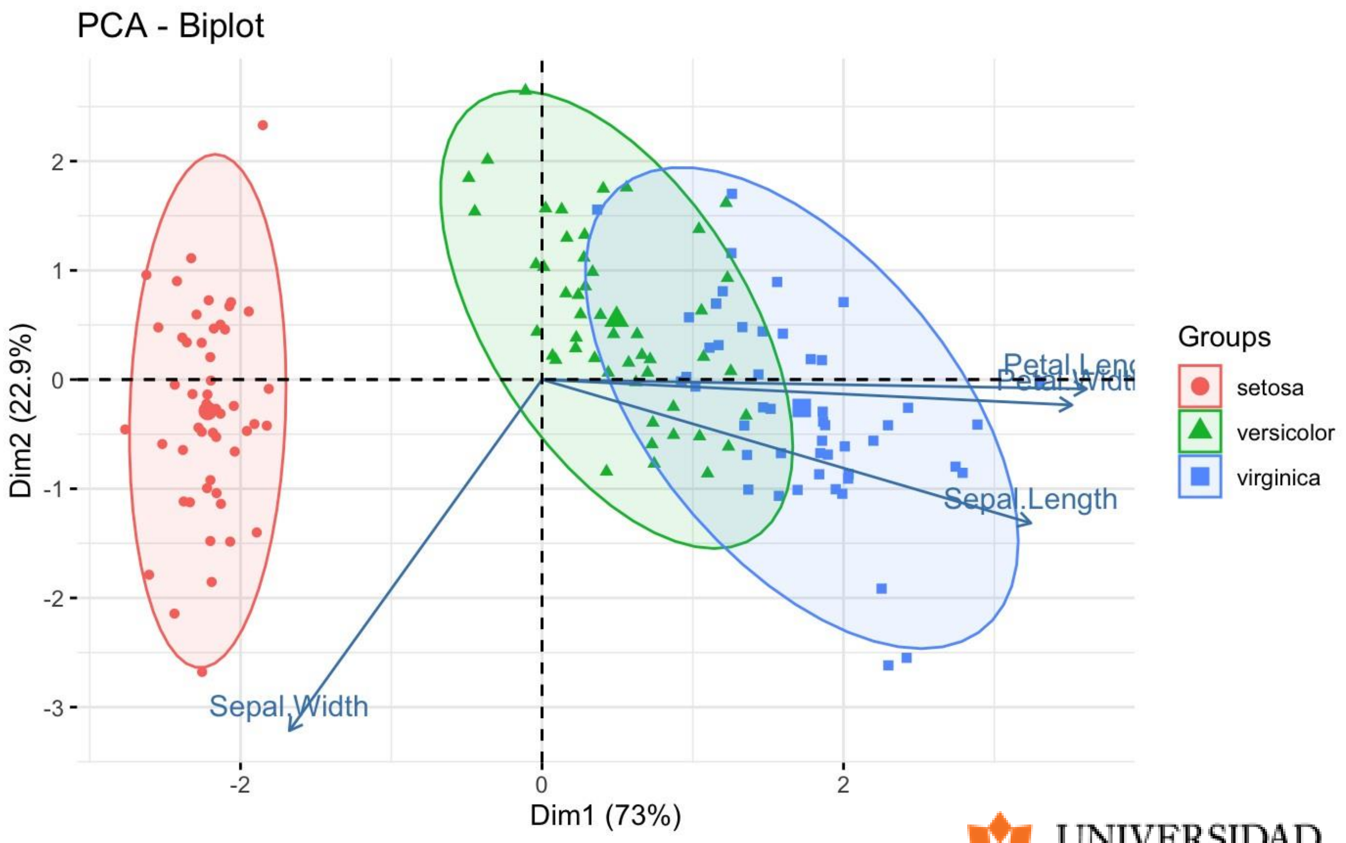
UNIVERSIDAD
DE LIMA

UNIVERSIDAD DE LIMA

ANÁLISIS DE COMPONENTES PRINCIPALES

Mg Sc. Meza Rodríguez, Aldo Richard

armeza@ulima.edu.pe



UNIVERSIDAD DE LIMA

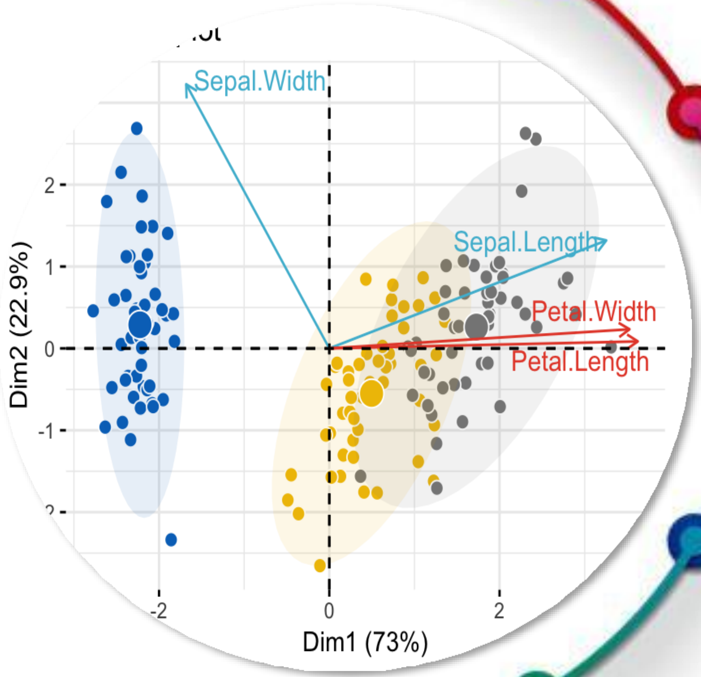
El análisis componentes principales

El análisis de componentes principales es una técnica multivariada que permite encontrar nuevas variables que en menor número simplifican la información contenida en las variables originales

Objetivos

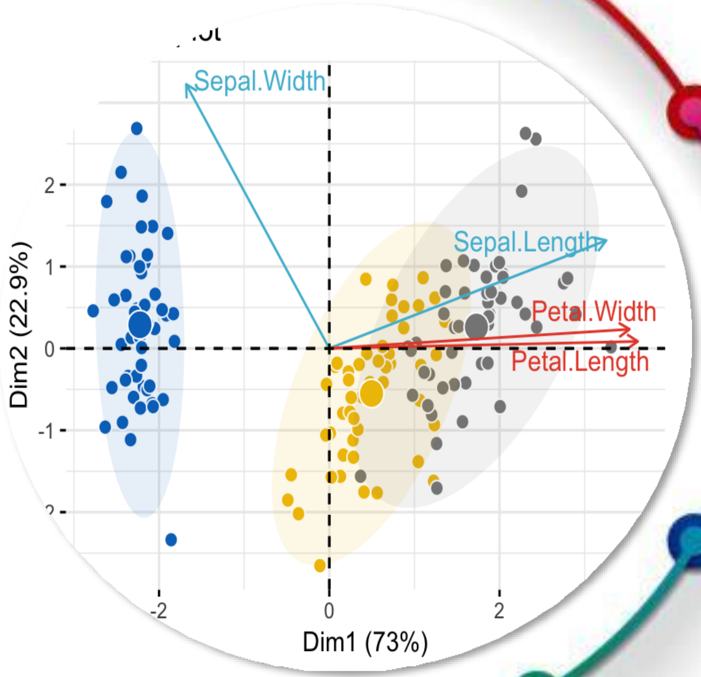
- Identificar patrones ocultos en un conjunto de datos.
- Reducir la dimensionalidad de los datos eliminando el ruido y la redundancia de los mismos.
- Identificar variables correlacionadas.
- Sirve como paso previo para otras técnicas

Características y requisitos



- El ACP se realiza con variables cuantitativas (aunque también puede realizarse reducción de dimensiones con diferentes variables)
- Si las variables originales no están correlacionadas el análisis no tiene sentido.
- Como medida de calidad de la información incorporada en cada componente se utiliza la varianza.
- El ACP determina los pesos que tienen c/u de las variables en cada componente, es decir las CP se explican en función de variables observables.
- Los componentes se expresan como una combinación lineal de las variables.

Características y requisitos



- El ACP generalmente se trabaja con variables tipificadas para evitar problemas derivados de las escalas.
- El nuevo conjunto de componentes es igual al número de variables.
- Como medida de calidad de la información incorporada en cada componente se utiliza la varianza.
- Cada componente se ordena de mayor a menor según la cantidad de varianza.
- Los componentes se expresan como una combinación lineal de las variables.

UNIVERSIDAD DE LIMA

Cálculo de los autovalores y autovectores

Para el cálculo de los autovalores λj y autovectores vj muestrales se utiliza el Teorema de Descomposición Espectral. Siendo A la matriz de variancias-covariancia muestral que estima a $\sum$, entonces expresando en términos de los estimadores se tiene:

$$\underbrace{(A - \lambda I)}_{n \times n} . v = 0_V$$

donde $\hat{\lambda_j}$ $y \, v_j$ son los estimadores del autovalor λj y autovector vj asociado respectivamente.

Análisis de Componentes Principales

$$
\begin{aligned}
Y_1 &= W_{11}X_1 + W_{12}X_2 + \ldots + W_{1p}X_p \\
Y_2 &= W_{21}X_1 + W_{22}X_2 + \ldots + W_{2p}X_p \\
&. \\
&. \\
&. \\
Y_p &= W_{p1}X_1 + W_{p2}X_2 + \ldots + W_{pp}X_p
\end{aligned}
$$

$$
\begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_P \end{bmatrix} = \begin{bmatrix} w_{11} & \cdots & w_{1P} \\ \vdots & \ddots & \vdots \\ w_{P1} & \cdots & w_{PP} \end{bmatrix} \begin{bmatrix} X_1 \\ . \\ . \\ . \\ X_P \end{bmatrix}
$$

$$
\text{siendo} \begin{cases} Var(y_1) = \lambda_1 \\ Var(y_2) = \lambda_2 \\ \vdots \\ Var(y_p) = \lambda_p \end{cases}
$$

Para que Y1 capte la mayor variabilidad de los datos sus coeficientes deben ser las coordenadas del autovector u1 de la matriz de covarianzas Σde las variables originales que corresponden al mayor autovalor λ1. De igual modo, para que el segundo componente principal sea no correlacionado con el primero y capte la variabilidad que no ha sido explicada por éste, sus coeficientes deben ser las coordenadas del autovector u2 de la matriz Σcuyo autovalor λ2 es el que le sigue en valor al primer autovalor λ1, y así sucesivamente hasta obtener p componentes principales. Se considera que cada autovector ui debe ser unitario, es decir, de longitud 1.

Algoritmo del ACP

1. Calcular la matriz de covarianciа global $\sum_X$
2. Calcular los autovalores de $\sum_X$, $\lambda_1, \lambda_2, \ldots, \lambda_p$
3. Calcular los autovectores1 1, 2,…., p asociados a $\lambda_1, \lambda_2, \ldots, \lambda_p$
4. Formar la matriz W = [1, 2,…., p]

Reducción de Dimensionalidad

Propiedades

1

$$Var(Y_i) = \lambda_i$$

•La varianza de la i-ésima componente principal es igual a su autovalor asociado.

2

$$V(Y1) \geq V(Y2) \geq \ldots V(Yp)$$

•Las varianzas están en orden decrecientes

3

$$\sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

$$\sum_{i=1}^{p} Var(Y_i) = \lambda_1 + .\lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} Var(Z_i) = p$$

•La suma de las varianzas de los componentes es igual a la de las variables originales

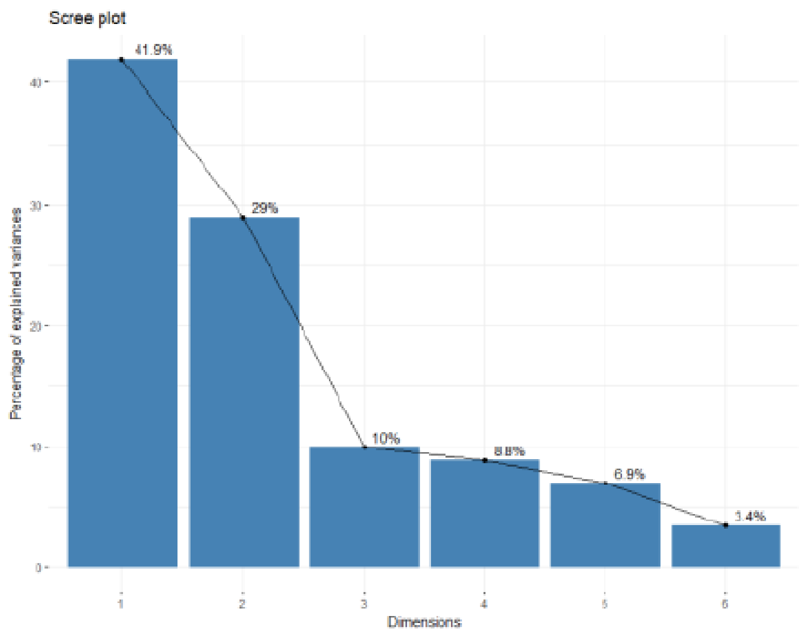
4

$$\frac{\lambda_k}{\lambda_1 + \ldots + \lambda_p}, \text{ k=1,......,p} \qquad \frac{\lambda_k}{p}, \text{ k=1,......,p}$$

La proporción explicada por la por las k primeras componentes respecto a la variabilidad total es:

Número de componentes a retener

- **Criterio del grafico de sedimentación (Scree Plot)**



- **Criterio de la media aritmética**

Se debe seleccionar aquellas componentes cuyo autovalor excede de la media de los autovalores.