Seminario de Analítica de datos e Inteligencia Artificial

Santiago Murillo Rendón - Reinel Tabares Soto

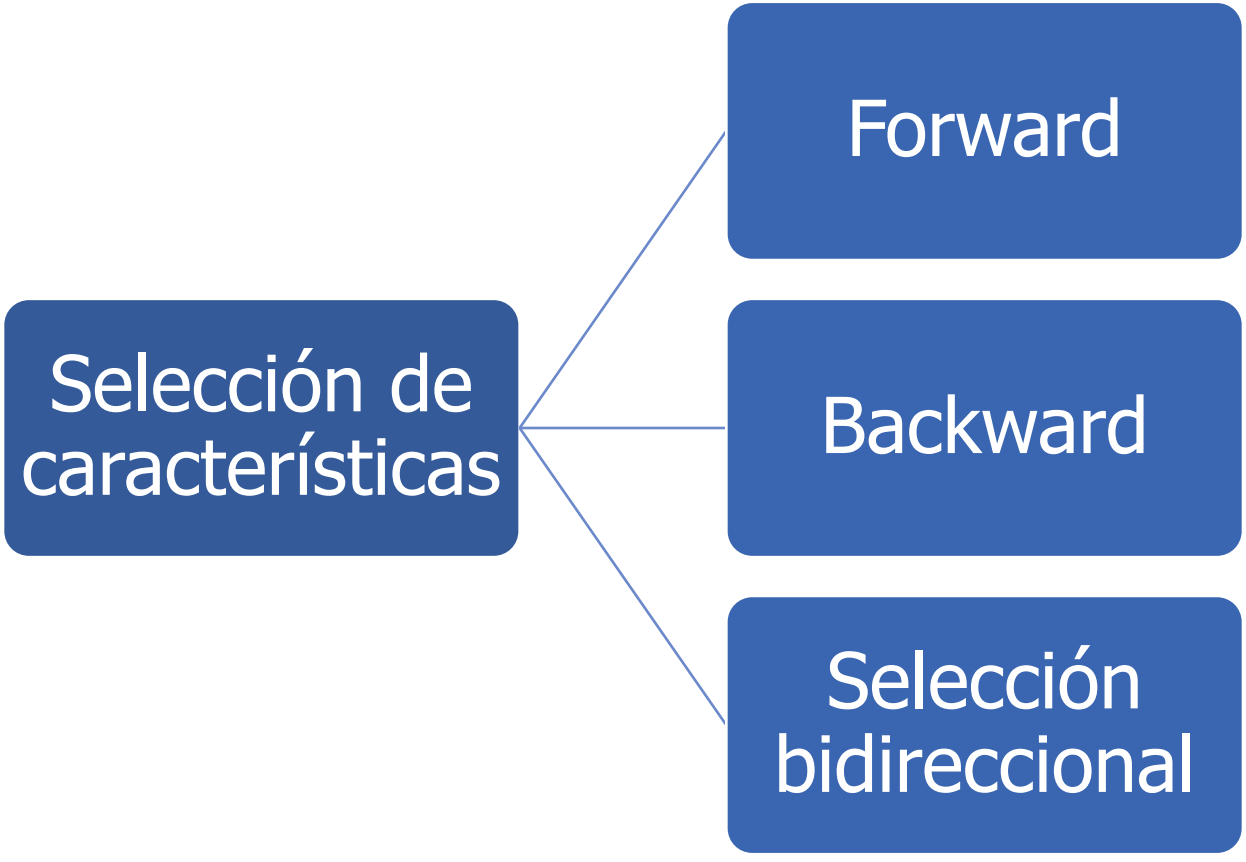
Agenda

1. Técnicas de extracción y selección de características
2. Backward selection
3. Fordward selection
4. PCA

Técnicas de selección y extracción de características

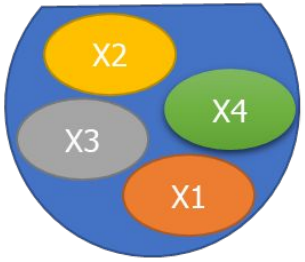
- No todas las características aportan en un proceso de clasificación o regresión.
- Deben encontrarse aquellas características más relevantes para el proceso
- La extracción permite representar en un espacio diferente las características originales. Dicho espacio permite mejorar la representación de los datos. Algunas técnicas son el análisis de componentes principales o la reducción de dimensión multifactorial.
- La selección permite, del conjunto de características iniciales conservar únicamente aquellas que aporten al proceso. Existen múltiples conjuntos de técnicas siendo las tradicionales el análisis hacia adelante o el análisis hacia atrás.

Selección de características


Selección de características
Forward
Backward
Selección bidireccional

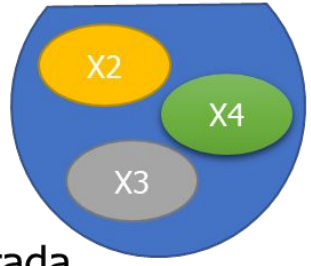
Selección forward

Modelo vacío



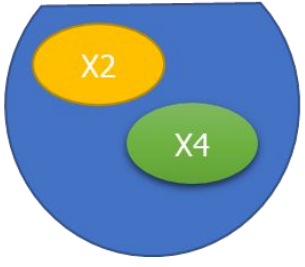
Agregar la variable más significativa

Modelo con una variable



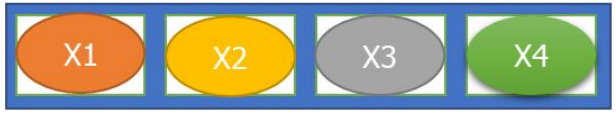
Continuar agregando la variable más significativa hasta alcanzar la regla de parada

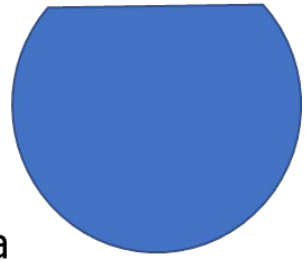
Modelo con dos variables



Selección backward

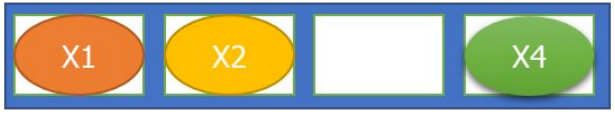
Modelo completo

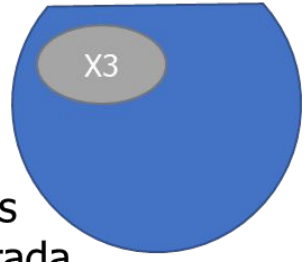




Remover la variable menos significativa

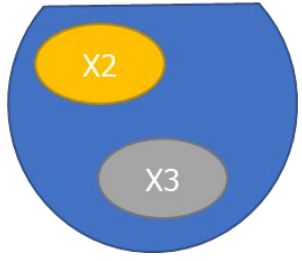
Modelo con una variable





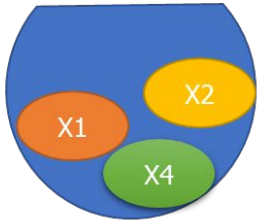
Continuar removiendo la variable menos significativa hasta alcanzar la regla de parada

Modelo con dos variables



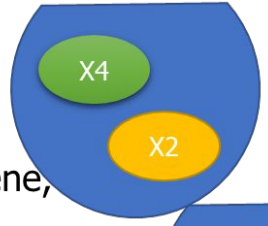
Selección bidireccional

Agregar la variable más significativa



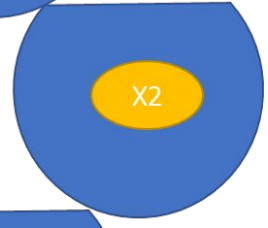
Agregar la variable más significativa

Modelo con una variable

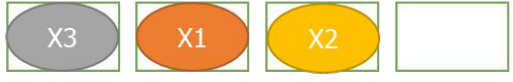


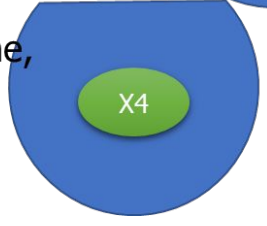
Si la variable mejora el modelo se mantiene, sino se retira. Se agrega la siguiente

Modelo con dos variables

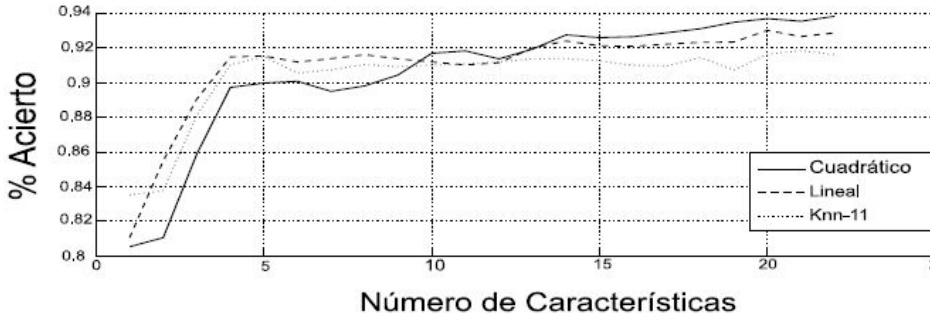


Si la variable mejora el modelo se mantiene, sino se retira. Se agrega la siguiente





Ejemplos



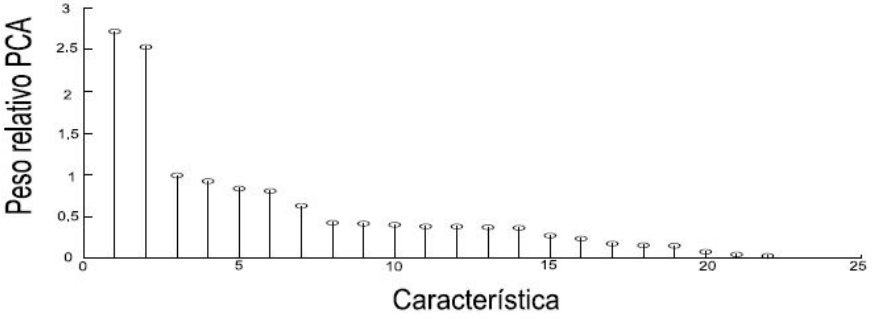


Figura 2. *Clasificación por componentes ordenados según relevancia PCA*

Característica	Jitter	Shimmer	HNR	CHNR	NNE	GNE	11MFCC
Índice Medias	1	2	3	4	5	6	7 a 17
Índice Des.Estándar	18	19	20	21	22	23	24 a 34

Tabla 1. *Indexación de las características*

Extracción de características

Análisis de componentes principales PCA

- Es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.
- Supóngase que existe una muestra con n individuos cada uno con p variables (X1, X2, …, Xp), es decir, el espacio muestral tiene p dimensiones.
- PCA permite encontrar un número de factores subyacentes ($z<p$) que explican aproximadamente lo mismo que las p variables originales.
- Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores.
- Cada una de estas z nuevas variables recibe el nombre de **componente principal.**

Los eigenvectores y los egenvalue

Los **eigenvectores** son un caso particular de multiplicación entre una matriz y un vector. Obsérvese la siguiente multiplicación:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} x \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4\, x \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Cuando se multiplica una matriz por alguno de sus **eigenvectores** se obtiene un múltiplo del vector original, es decir, el resultado es ese mismo vector multiplicado por un número.

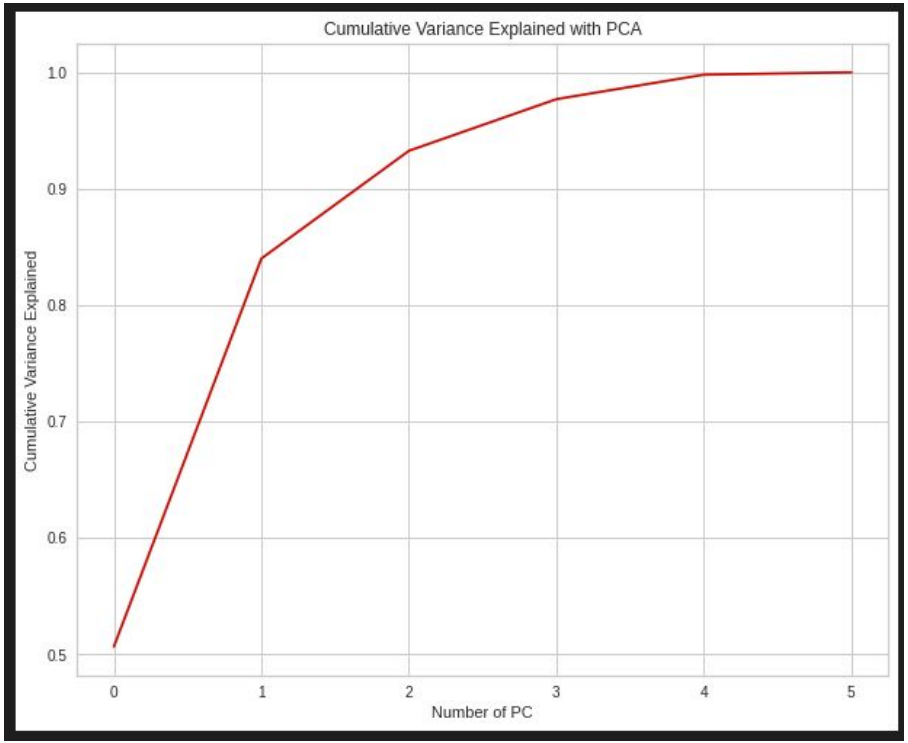
Al valor por el que se multiplica el **eigenvector** resultante se le conoce como **eigenvalor**.

A todo **eigenvector** le corresponde un **eigenvalor** y viceversa.

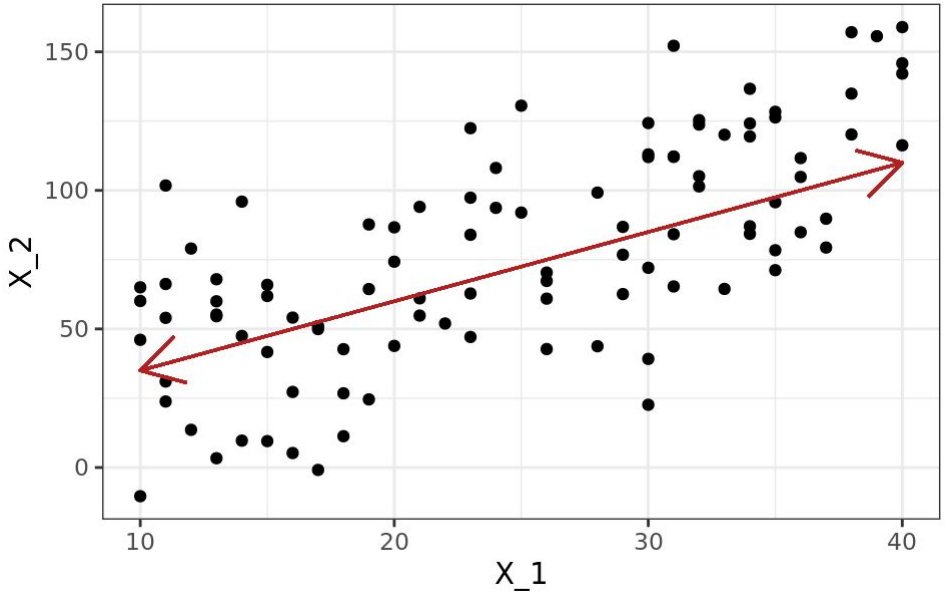
- En el método PCA, cada una de las componentes se corresponde con un **eigenvector**, y el orden de componente se establece por orden decreciente de **eigenvalor**. Así pues, la primera componente es el **eigenvector** con el **eigenvalor** asociado más alto.

- Para obtener los **eigenvalores** se realiza un proceso de determinar la varianza del conjunto original de características

Componentes y varianza acumulada

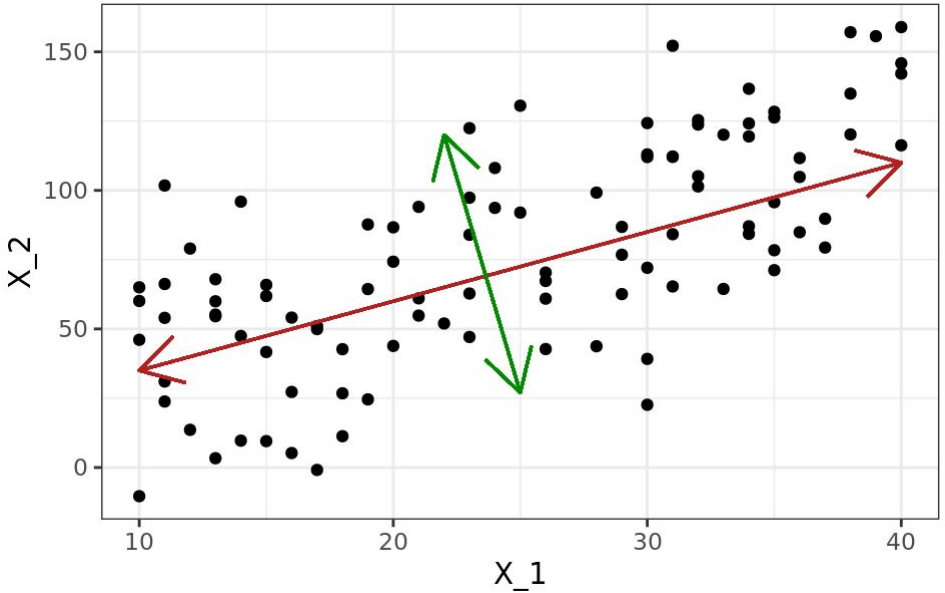


Interpretación geométrica de las componentes principales



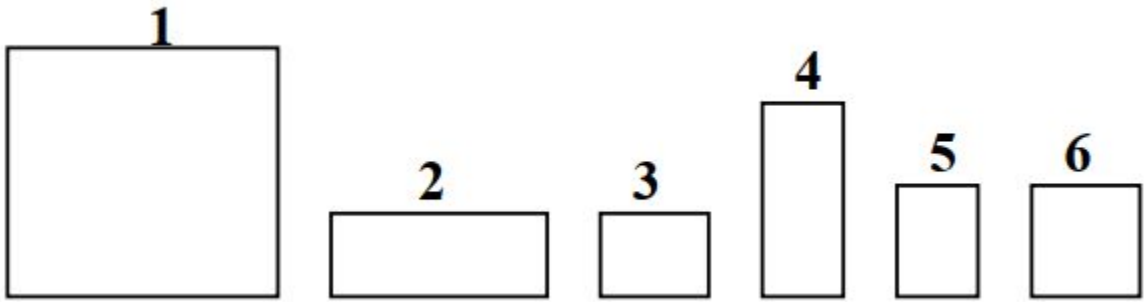
Tomado de Análisis de Componentes Principales. Joaquín Amat Rodrigo

Interpretación geométrica de las componentes principales



Tomado de Análisis de Componentes Principales. Joaquín Amat Rodrigo

¿Qué son los componentes principales?



$$X = \begin{bmatrix} 2 & 2 \\ 1.5 & 0.5 \\ 0.7 & 0.5 \\ 0.5 & 1.5 \\ 0.5 & 0.7 \\ 0.7 & 0.7 \end{bmatrix}$$

Tomado de Análisis de Datos multivariantes. Daniel Peña

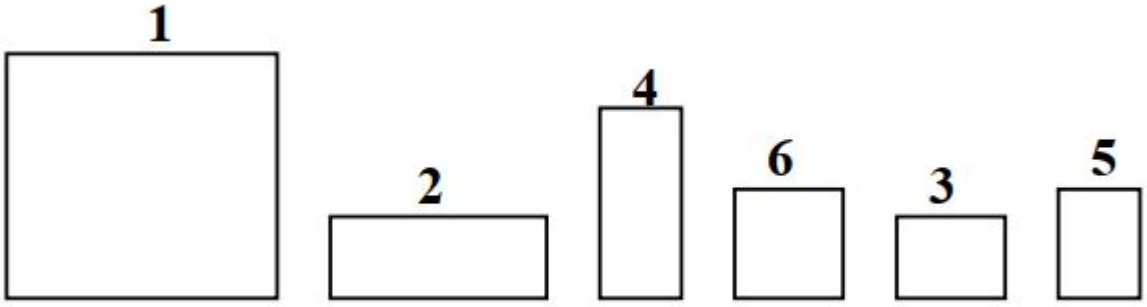
¿Qué son los componentes principales?

$$Z_1 = Xa_1 = 0.707\log(X_1) + 0.707\log(X_2) = 0.707\log(X_1X_2) = \begin{bmatrix} 0.426 \\ -0.088 \\ -0.322 \\ -0.088 \\ -0.322 \\ -0.219 \end{bmatrix}$$

$$Z_2 = Xa_2 = 0.707\log(X_1) - 0.707\log(X_2) = 0.707\log(\frac{X_1}{X_2}) = \begin{bmatrix} 0 \\ 0.337 \\ 0.103 \\ -0.337 \\ -0.103 \\ 0 \end{bmatrix}$$

Tomado de Análisis de Datos multivariantes. Daniel Peña

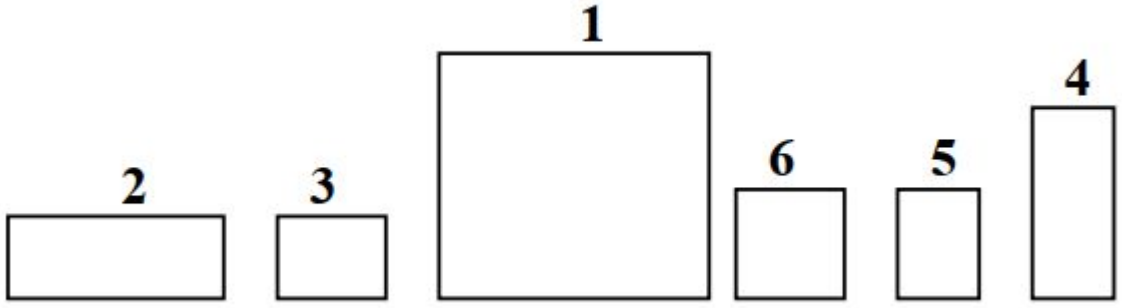
Ordenando los elementos según las componentes



Coincide con la inducida por el volumen de los rectángulos, es una transformación creciente del producto de la base por la altura, y el primer componente describe el tamaño.

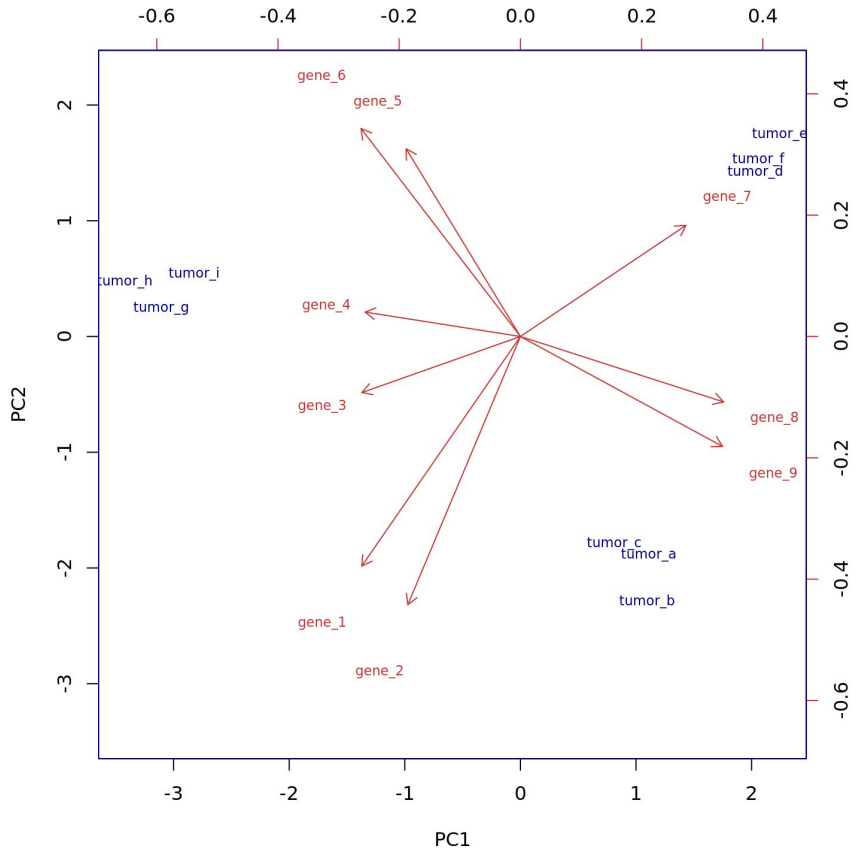
Tomado de Análisis de Datos multivariantes. Daniel Peña

Ordenando los elementos según las componentes



El segundo componente relaciona la base con la altura y ordena las observaciones en función de su forma.

Tomado de Análisis de Datos multivariantes. Daniel Peña


-0.6
-0.4
-0.2
0.0
0.2
0.4
gene_6
gene_5
tumor_e
tumor_f
tumor_d
gene_7
tumor_h
tumor_i
tumor_g
gene_4
gene_3
gene_8
gene_9
tumor_c
tumor_a
tumor_b
gene_1
gene_2
2
1
0
-1
-2
-3
PC2
0.4
0.2
0.0
-0.2
-0.4
-0.6
-3
-2
-1
0
1
2
PC1