Subject: Probability and Statistics

Class XX: Dimension Reduction: Principal Component Analysis (PCA)

UdeA

Bioengineering

Francisco José Campuzano Cardona

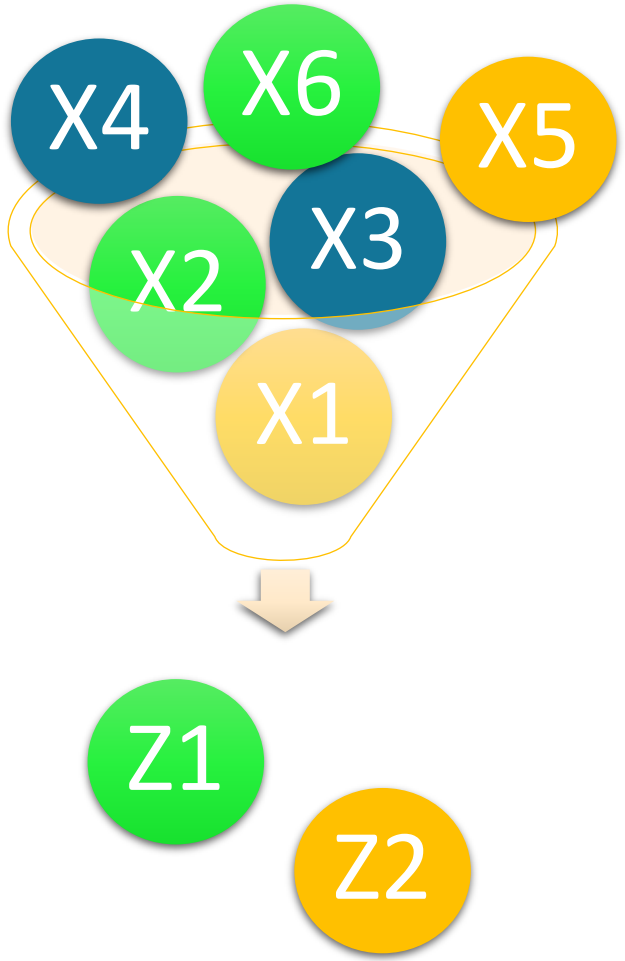
Bioengineerer, MSc in Engineering

UdeA

La última categoría de métodos de aprendizaje automático no supervisado corresponde a los métodos de reducción de dimensión, donde tenemos el más usado de estos, llamado Análisis de Compotentes Principales (PCA, siglas en inglés)

Nuevamente es un algoritmo que no tiene en cuenta ninguna etiqueta.

UdeA

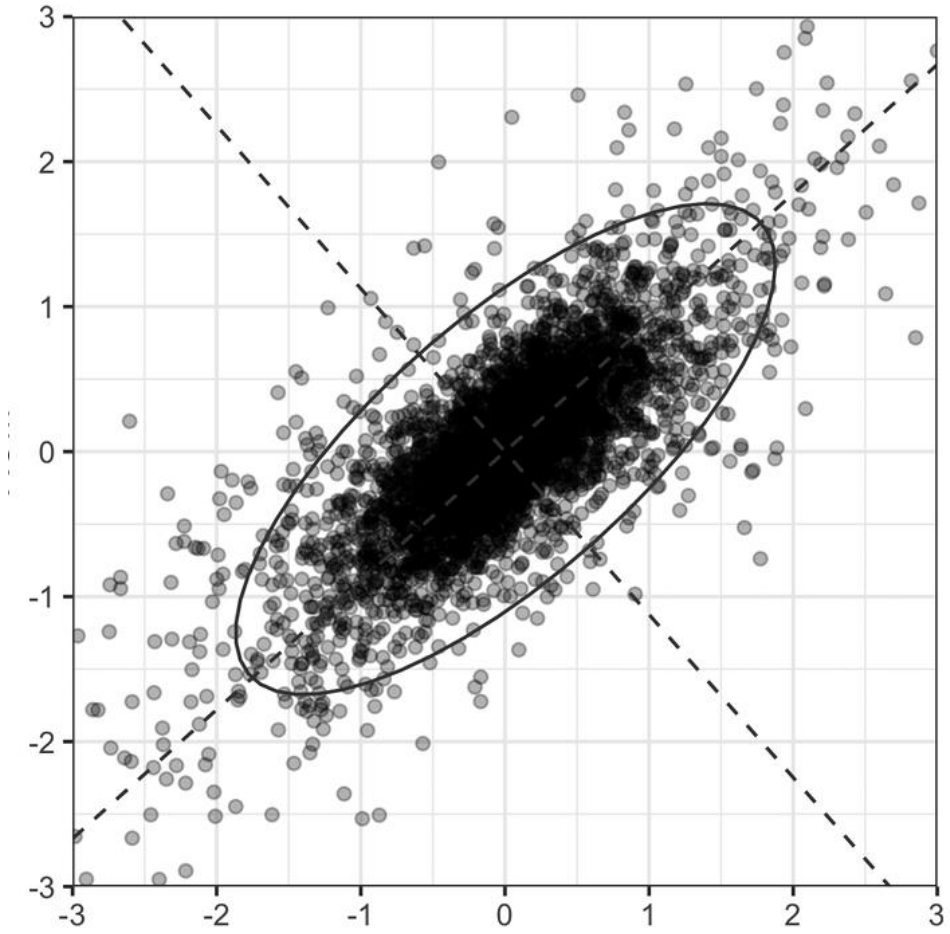


En general, la reducción de dimensión hace referencia a describir la variabilidad de los datos, con la menor cantidad de variables posibles.

UdeA

Supongamos que tenemos un set bidimensional. Lo primero es encontrar los ejes principales de los datos. Es decir, los ejes que mejor discriminan la variabilidad. Y se proyectan los datos sobre estos ejes a través de una combinación lineal

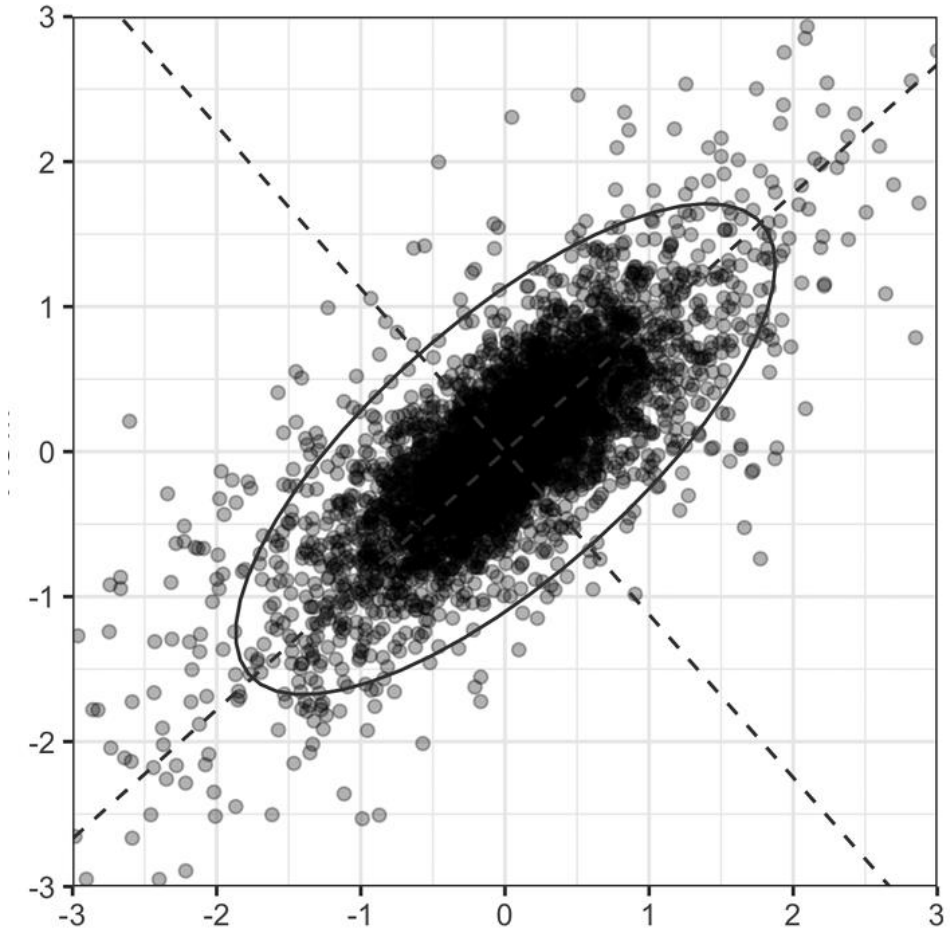
$$Z_i = w_{i,1}X_1 + w_{i,2}X_2$$



UdeA

Aquí vemos que uno de los ejes principales describe más la variabilidad de los datos que el otro. Entonces podríamos con los datos proyectados sobre ese eje, describir la mayor parte de la variabilidad de las dos variables originales.

$$Z_i = w_{i,1}X_1 + w_{i,2}X_2$$

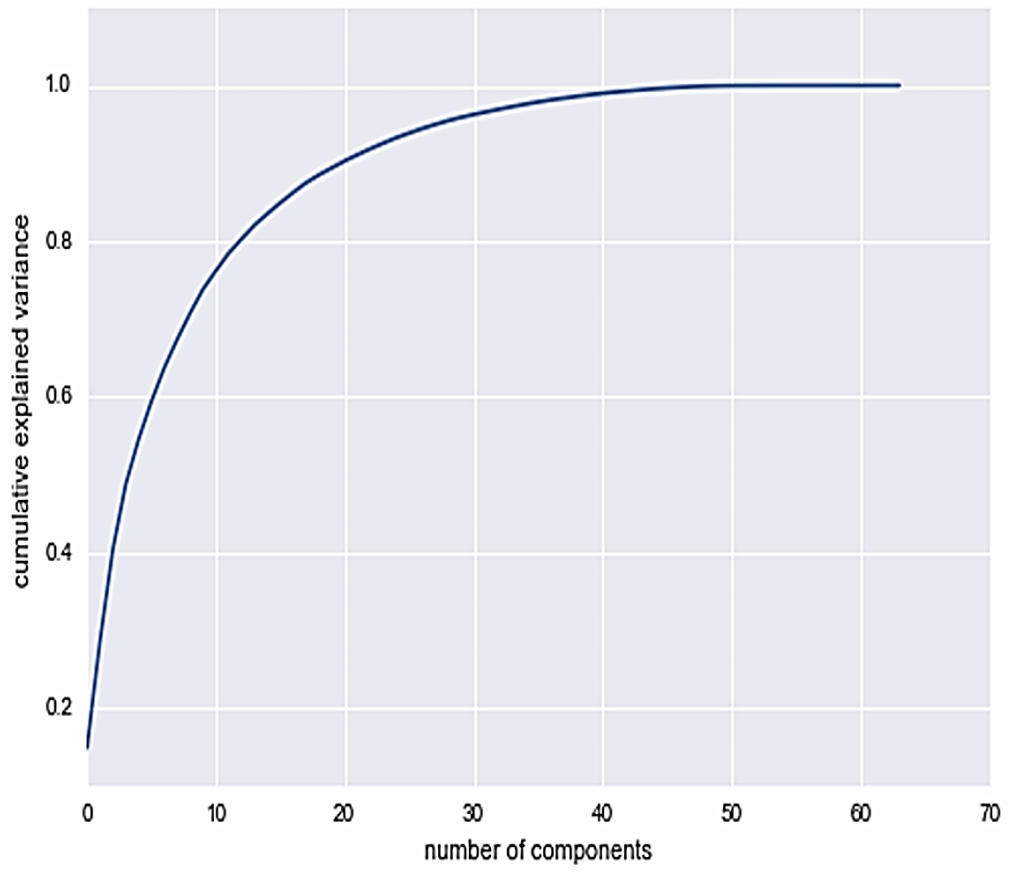


UdeA

¿cómo elegimos el numero de componentes principales?

Esta decisión puede ser guiada por una curva de varianza explicada acumulada contra el numero de componentes principales.

El 100% de la variabilidad se obtendrá con todas variables originales, pero se debe decidir cuánta variabilidad se esta dispuesto a dejar de explicar.



Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering