DIPLOMADO EN CIENCIA DE DATOS

Módulo: Minería de Datos

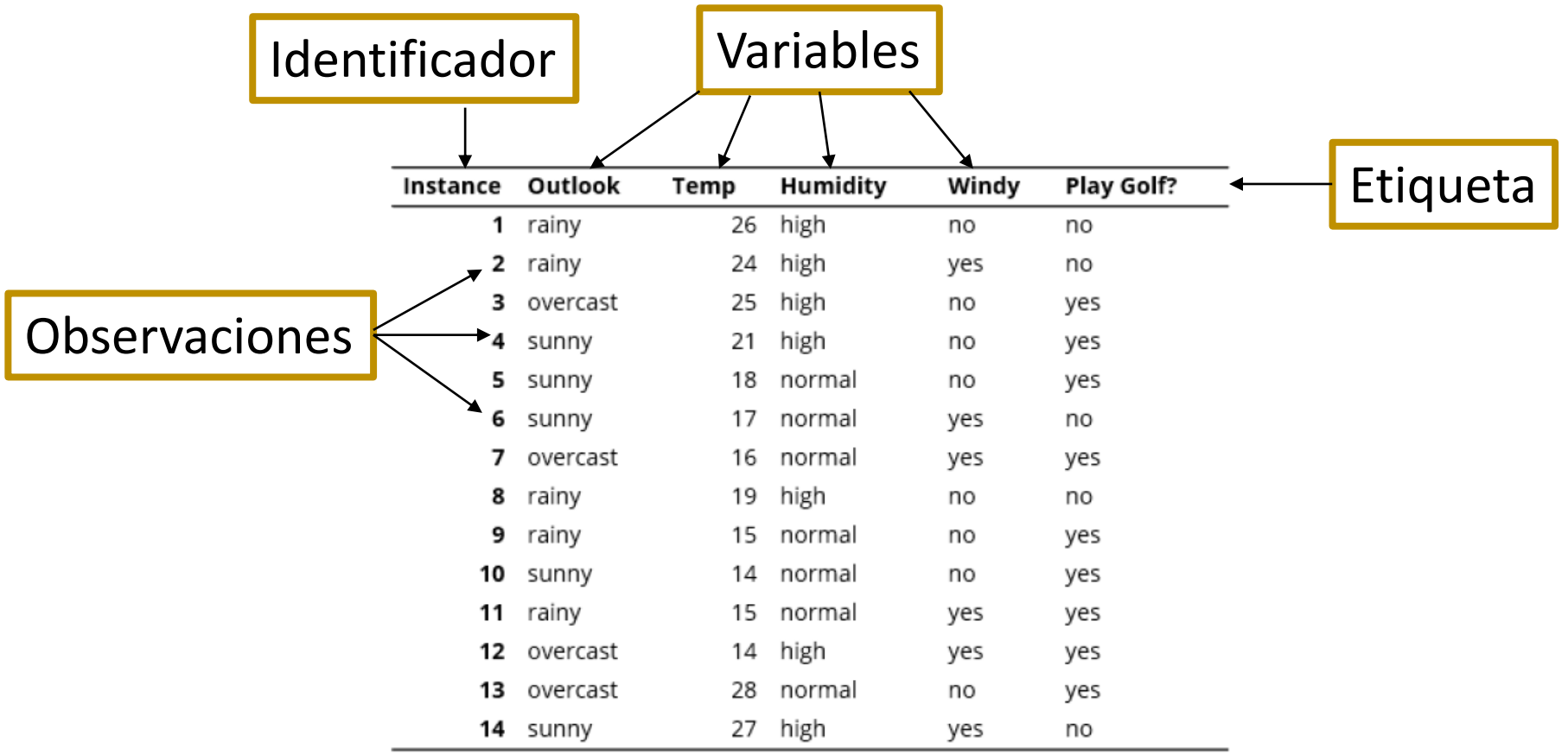
Matrices, vectores, distancias y reducción de dimensionalidad

Universidad Nacional de Colombia

Contenido

- Matrices
- Vectores
- Distancias
- Reducción de dimensionalidad

Estructura en los datos

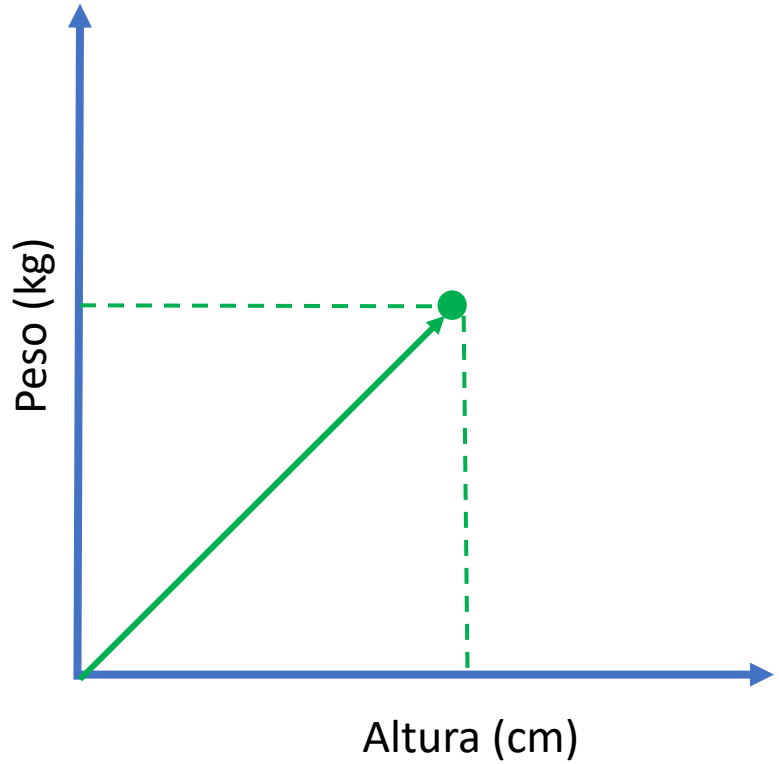


Instance	Outlook	Temp	Humidity	Windy	Play Golf?
1	rainy	26	high	no	no
2	rainy	24	high	yes	no
3	overcast	25	high	no	yes
4	sunny	21	high	no	yes
5	sunny	18	normal	no	yes
6	sunny	17	normal	yes	no
7	overcast	16	normal	yes	yes
8	rainy	19	high	no	no
9	rainy	15	normal	no	yes
10	sunny	14	normal	no	yes
11	rainy	15	normal	yes	yes
12	overcast	14	high	yes	yes
13	overcast	28	normal	no	yes
14	sunny	27	high	yes	no

Los datos como puntos n-dimensionales

Si tengo un conjunto de observaciones que tienen valores en las mismas p variables

Género	Altura	Peso	BMI
Male	174	96	4
Male	189	87	2
Female	185	110	4
Female	195	104	3
Male	149	61	3
Male	189	104	3
Male	147	92	5
Male	154	111	5
Male	174	90	3
Female	169	103	4



Cada observación se puede representar como un punto en el espacio P - dimensional
Y cada dimensión representa una variable diferente

Estructura en los datos

Género	Altura	Peso	BMI
Male	174	96	4
Male	189	87	2
Female	185	110	4
Female	195	104	3
Male	149	61	3
Male	189	104	3
Male	147	92	5
Male	154	111	5
Male	174	90	3
Female	169	103	4

Si tengo un conjunto de observaciones para las cuáles tengo valores en las mismas variables. Los datos pueden ser representados en una matriz nxp

Donde se tienen n filas, una por cada observación

Y p columnas, una por cada variable

Matrices

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Arreglo rectangular de números

La dimensión de una matriz está dada por el *número de* $filas\ (i)\ x\ el\ n$*úmero de columnas* (j)

A_{ij} es el elemento de la matriz en la *i-ésima* fila y la *j-ésima* columna

Matrices

Variables

Observaciones

$$\begin{bmatrix} X_1 & X_2 & \dots & X_j & \dots & X_p \\ X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2p} \\ & & & & & \\ X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{ip} \\ & & & & & \\ X_{n1} & X_{n2} & \dots & X_{nj} & \dots & X_{np} \end{bmatrix}$$

La matriz X tiene n observaciones y p columnas

Vector

$$b = \begin{bmatrix} 1 \\ 4 \\ 7 \end{bmatrix}$$

Matriz de tamaño *nx1*

b_i es el *i-ésimo* elemento del vector

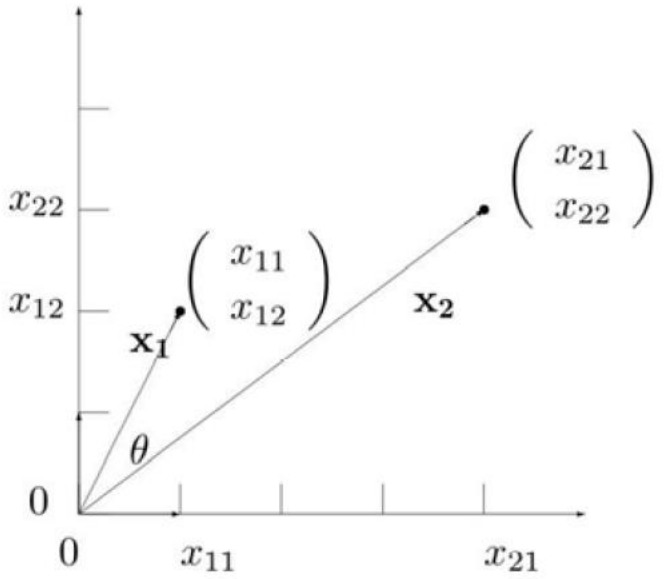
Vectores

$$X = \begin{bmatrix} X_1 & X_2 & \dots & X_j & \dots & X_p \\ X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2p} \\ & & & & & \\ X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{ip} \\ & & & & & \\ X_{n1} & X_{n2} & \dots & X_{nj} & \dots & X_{np} \end{bmatrix}$$

Estas son las dos primeras observaciones y representan dos puntos o vectores, que empiezan en el origen en un espacio p-dimensional

$$x_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1p} \end{pmatrix} \qquad x_2 = \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2p} \end{pmatrix}$$

Si $p = 2$



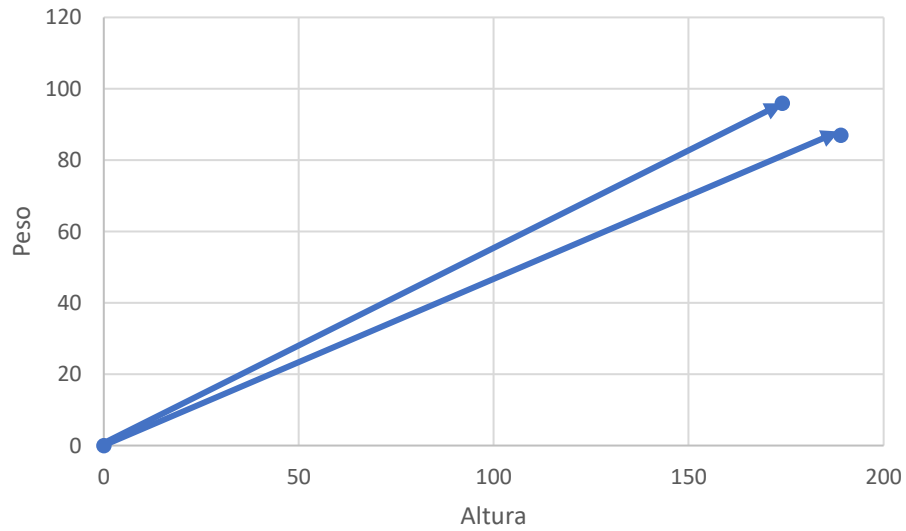
Vectores

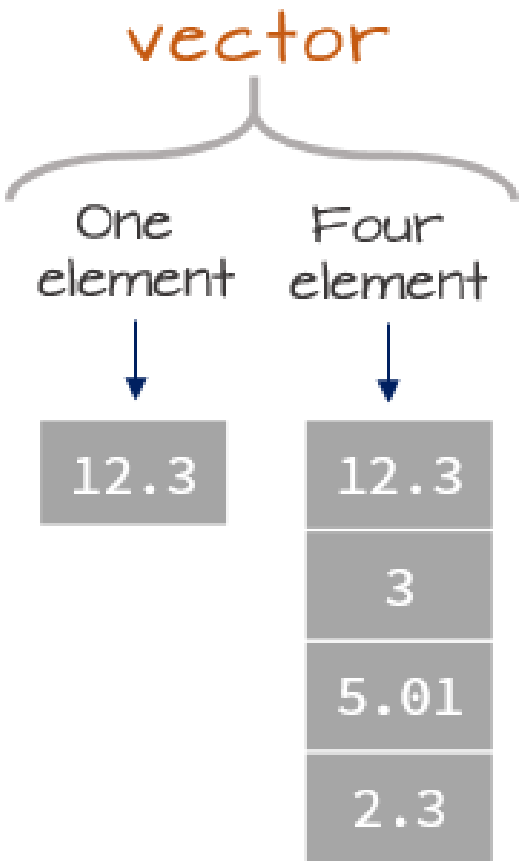
$$X = \begin{bmatrix} 174 & 96 & 4 \\ 189 & 87 & 2 \\ 185 & 110 & 4 \\ 195 & 104 & 3 \\ 149 & 61 & 3 \end{bmatrix}$$

Estas son las dos primeras observaciones en nuestro ejemplo y representan dos puntos o vectores, que empiezan en el origen en un espacio p-dimensional

$$x_1 = \begin{bmatrix} 174 \\ 96 \\ 4 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 189 \\ 87 \\ 2 \end{bmatrix}$$

Si $p = 2$





dataframe

x	y
12.3	ace
3	tea
5.01	oil
2.3	tree

matrix

12.3	0.1
3.0	5.2
5.01	3.0
2.3	0.1

list

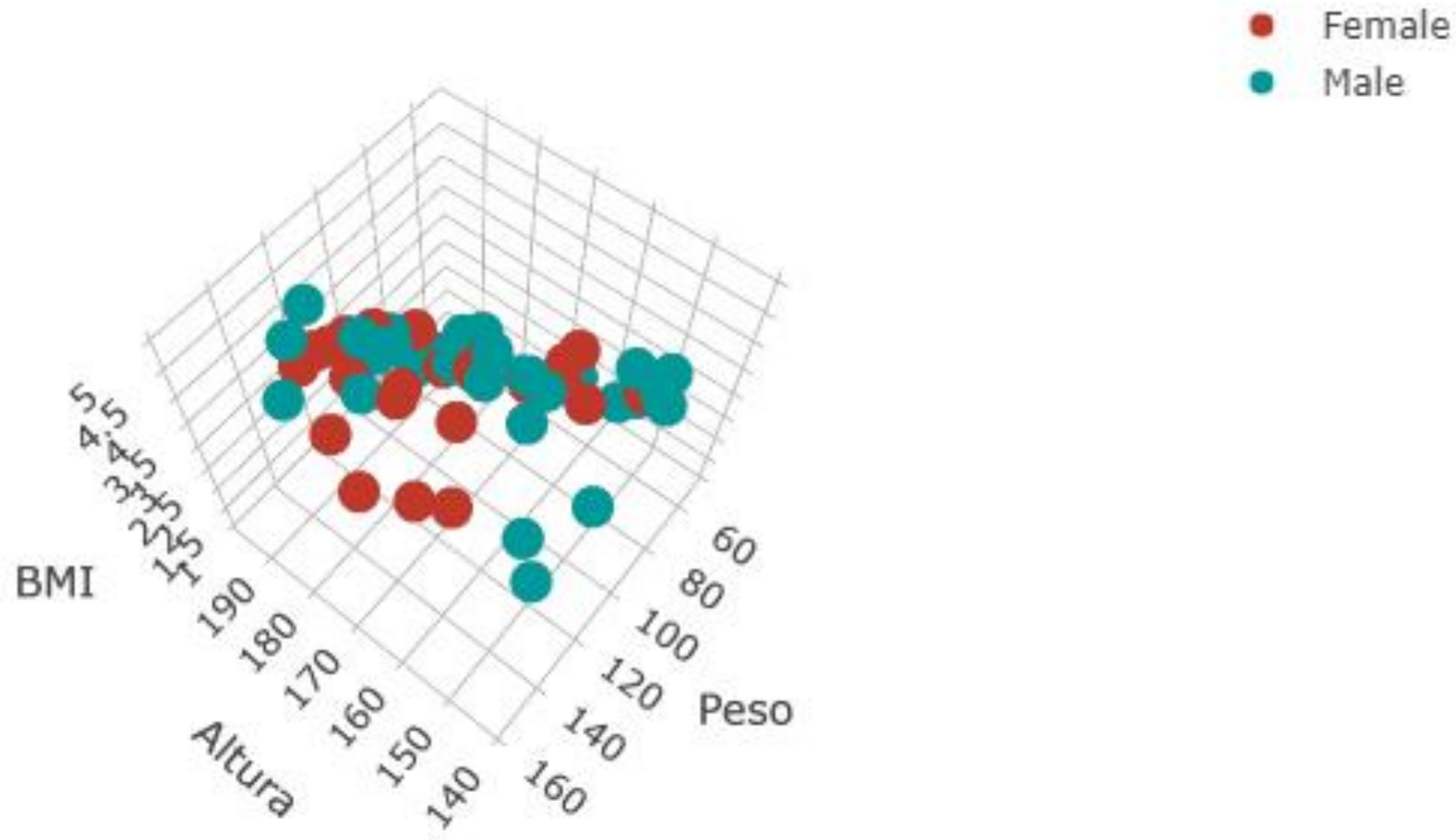
x	y
12.3	ace
3	tea
5.01	oil
2.3	tree

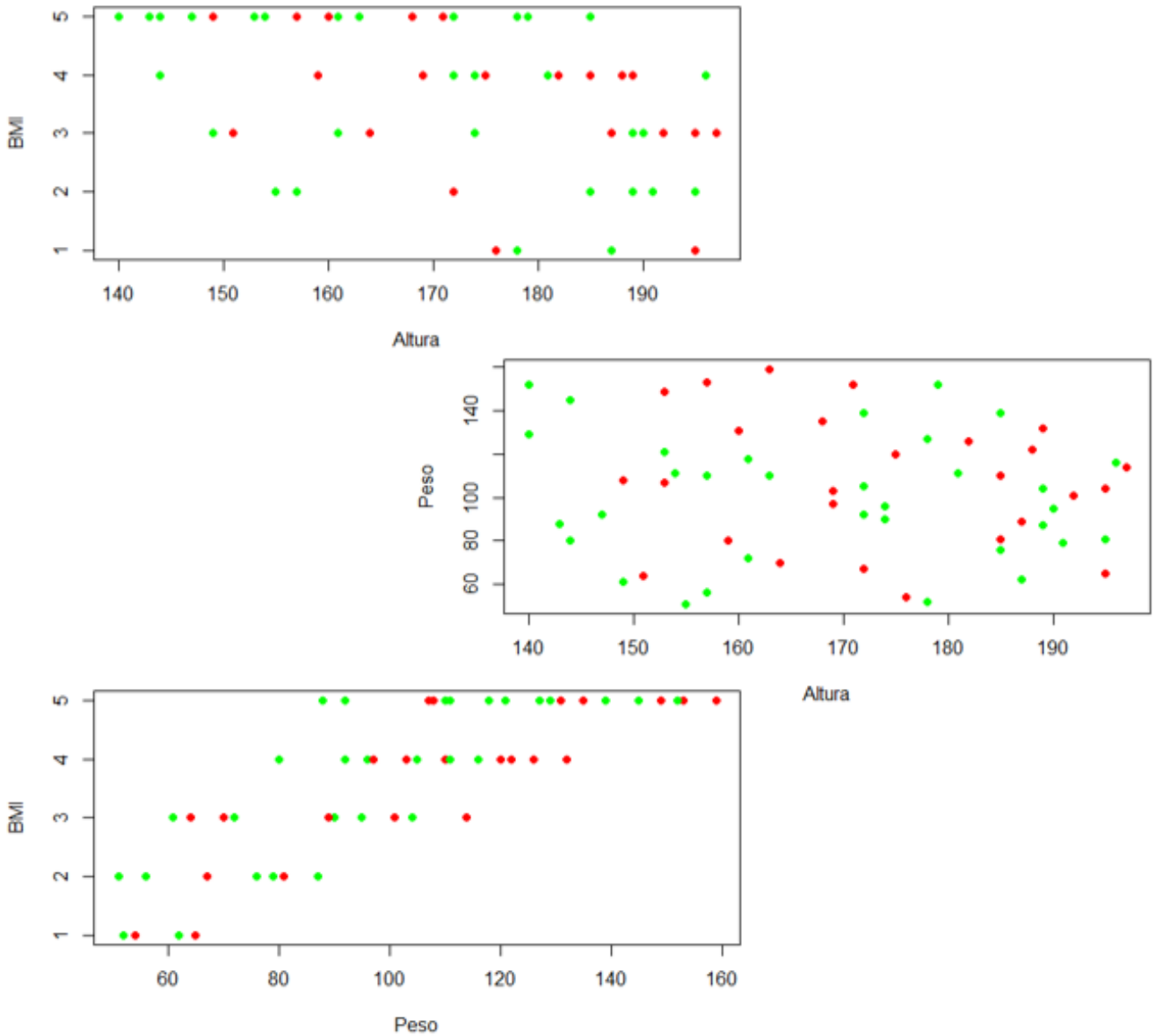
3

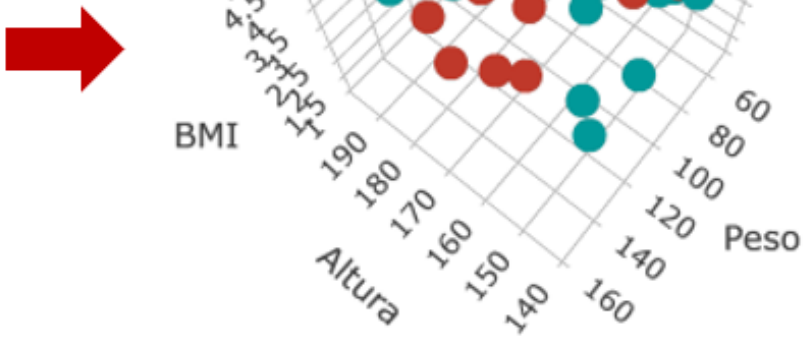
Y ~ x -1

some text

Género	Altura	Peso	BMI
Male	174	96	4
Male	189	87	2
Female	185	110	4
Female	195	104	3
Male	149	61	3
Male	189	104	3
Male	147	92	5
Male	154	111	5
Male	174	90	3
Female	169	103	4


Female
Male
BMI
5
4.5
4
3.5
3
2.5
2
1.5
1
Altura
190
180
170
160
150
140
Peso
60
80
100
120
140
160

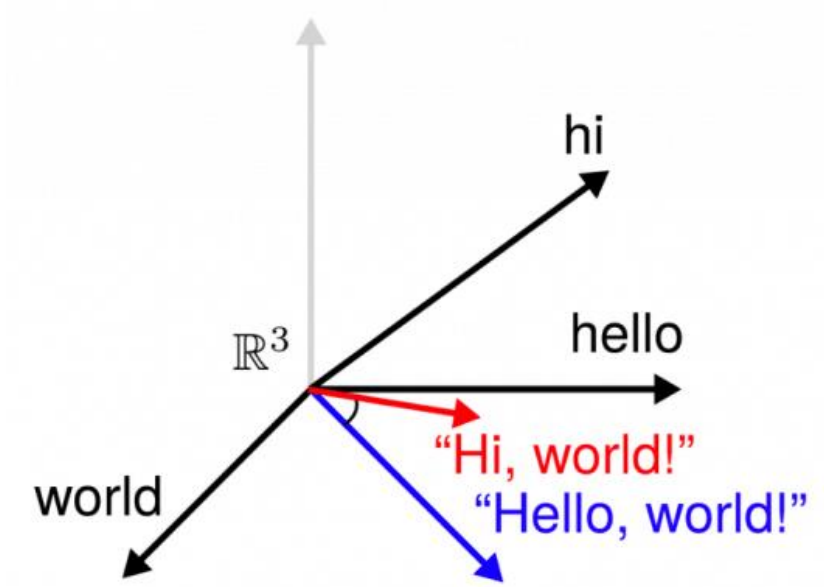

BMI
Altura
Peso
Altura
BMI
Peso


BMI
Altura
Peso

Distancias y medidas de similitud

Las distancias son medidas que describen que tan cerca dos objetos está, o qué tan similares son.

Si dos puntos se encuentran cerca, inferimos que tienen características que los hacen similares



En estadística son utilizadas como input para diferentes análisis y algoritmos.

Distancias y medidas de similitud

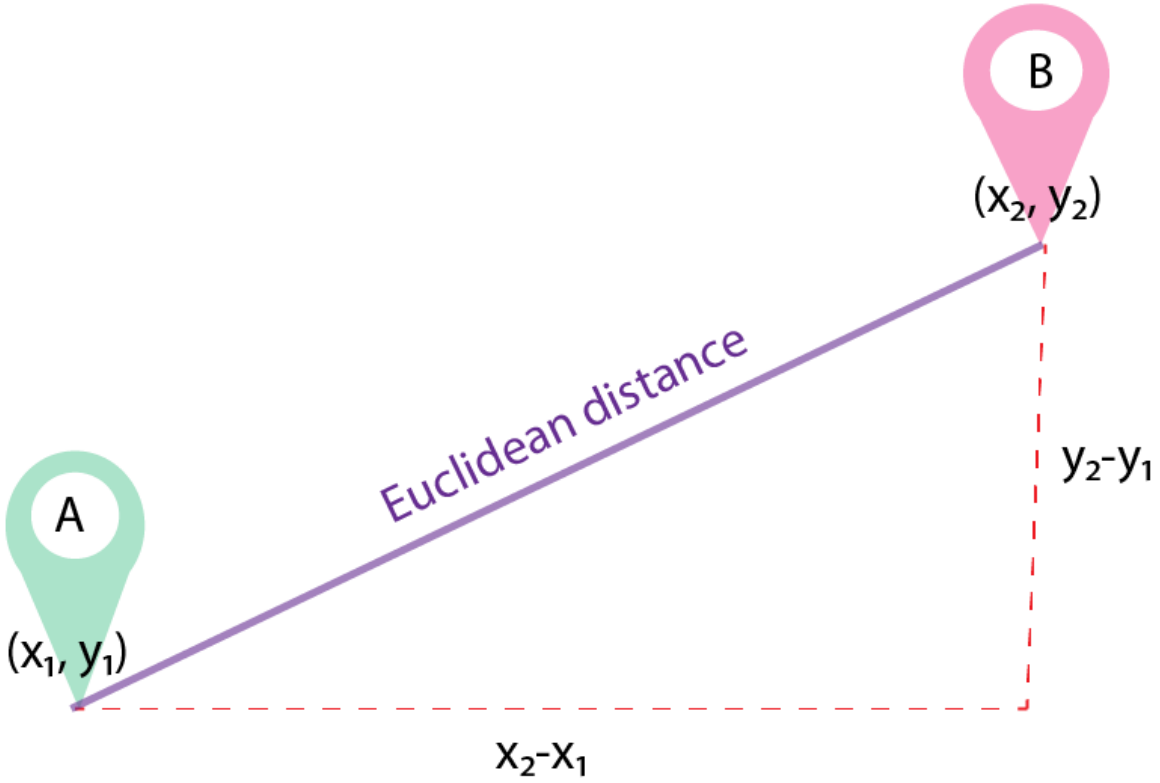
Una función de distancia o métrica debe

i.	$d(x, y) \geq 0$	No negativa
ii.	$d(x, y) = 0$	Si y solo si $x = y$
iii.	$d(x, y) = d(y, x)$	Simetría
iv.	$d(x, z) \leq d(x, y) + d(y, z)$	Desigualdad triangular

Tipos de Distancias

- Distancia euclidiana

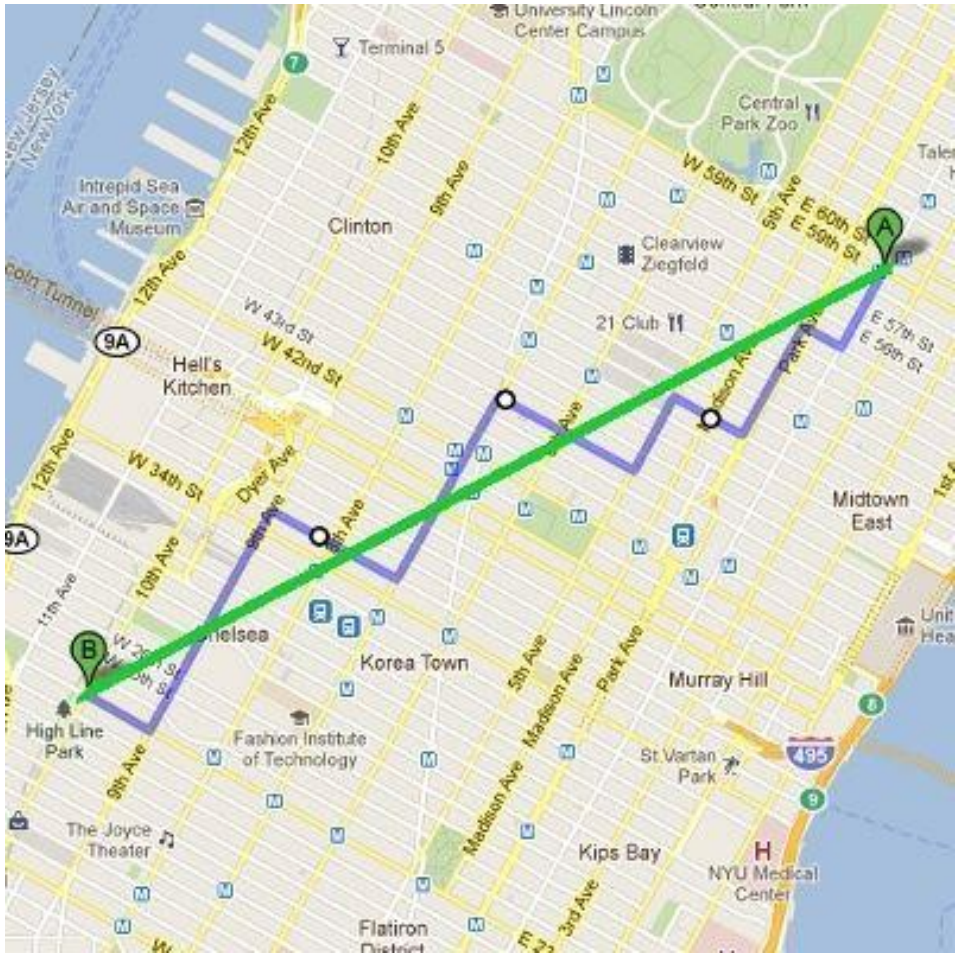
$$d\ (P1, P2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Tipos de Distancias

- Distancia Manhattan

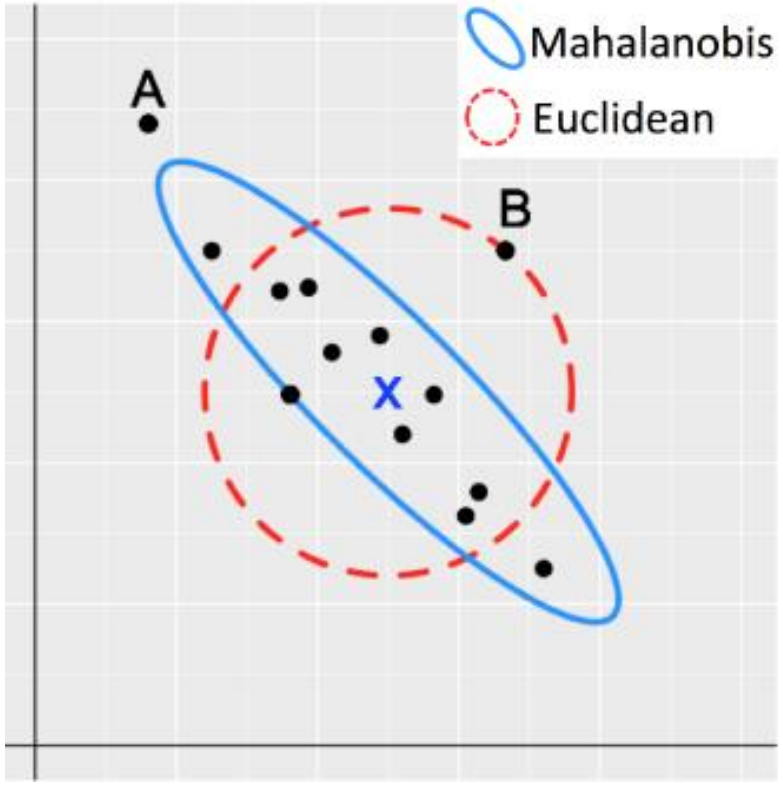
$$d\ (P1, P2) = |x_2 - x_1| + |y_2 - y_1|$$



Tipos de Distancias

- Distancia Malananobis

$$d\ (\vec{x}_1, \vec{x}_2) = \sqrt{\left(\frac{x_{11} - x_{12}}{\sigma_1}\right)^2 + \left(\frac{x_{21} - x_{22}}{\sigma_2}\right)^2}$$



Medidas de asociación para datos binarios

V / W	1	0
1	a	b
0	c	d

Donde $a + b + c + d = p$

Siendo
a: Número de atributos que *V* y *W* tienen en común
b: Número de atributos de *V* que no tiene *W*
c: Número de atributos de *W* que no tiene *V*
d: Número de atributos que no tienen ni V ni W

Una distancia como la euclidiana no resulta útil en este caso

Medidas de asociación para datos binarios

V / W	1	0
1	a	b
0	c	d

$$s_1(V,W) = \frac{a}{a+b+c}$$ (Index Jaccard)

$$s_2(V,W) = \frac{2a}{2a+b+c}$$ (Index Dice-Sorensen)

$$s_3(V,W) = \frac{a}{p}$$ (Index Russel-Rao)

Y se pueden transformar estas medidas de similaridad en medidas de disimilitud

$$d(v,w) = 1 - s(V,W) \text{ ó } d(v,w) = \frac{1}{s(V,W)}$$

Si se tienen tanto variables continuas como categóricas, se podría usar algo como la distancia Gower

https://medium.com/analytics-vidhya/gowers-distance-899f9c4bd553#:~:text=Gower's%20Distance%20can%20be%20used,and%201%20(maximally%20dissimilar).

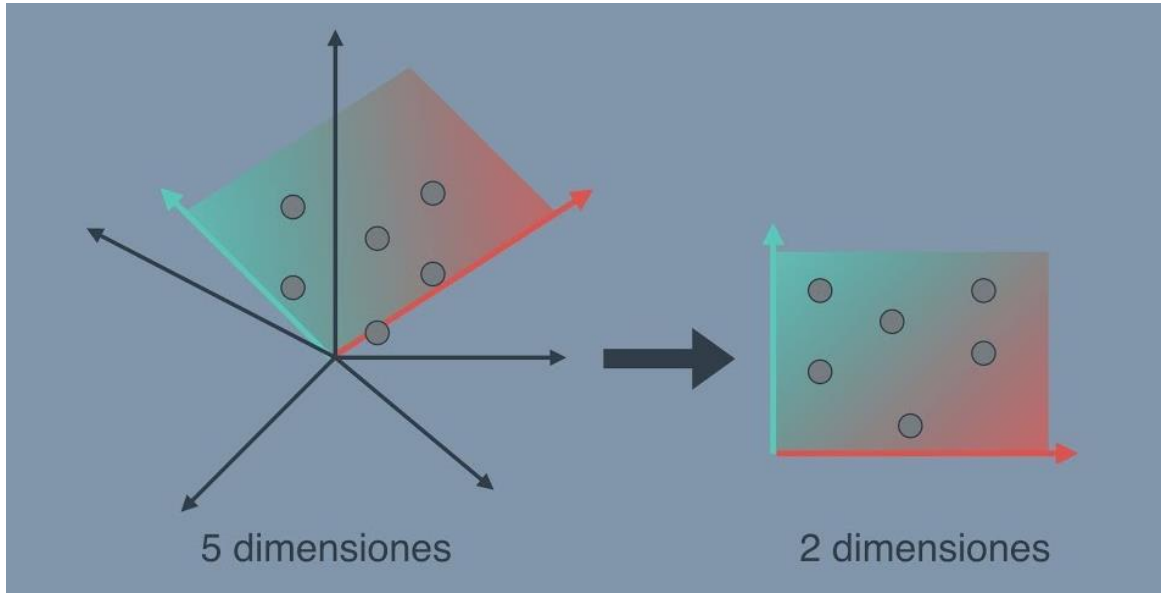
Reducción de dimensionalidad

- Maldición de dimensionalidad

- Reducir la cantidad de tiempo y el gasto computacional para analizar la data

- Visualización más sencilla de entender

- Reducir el ruido en la data



- Descartar del análisis variables que pueden ser irrelevantes

- Resumir la información en indicadores

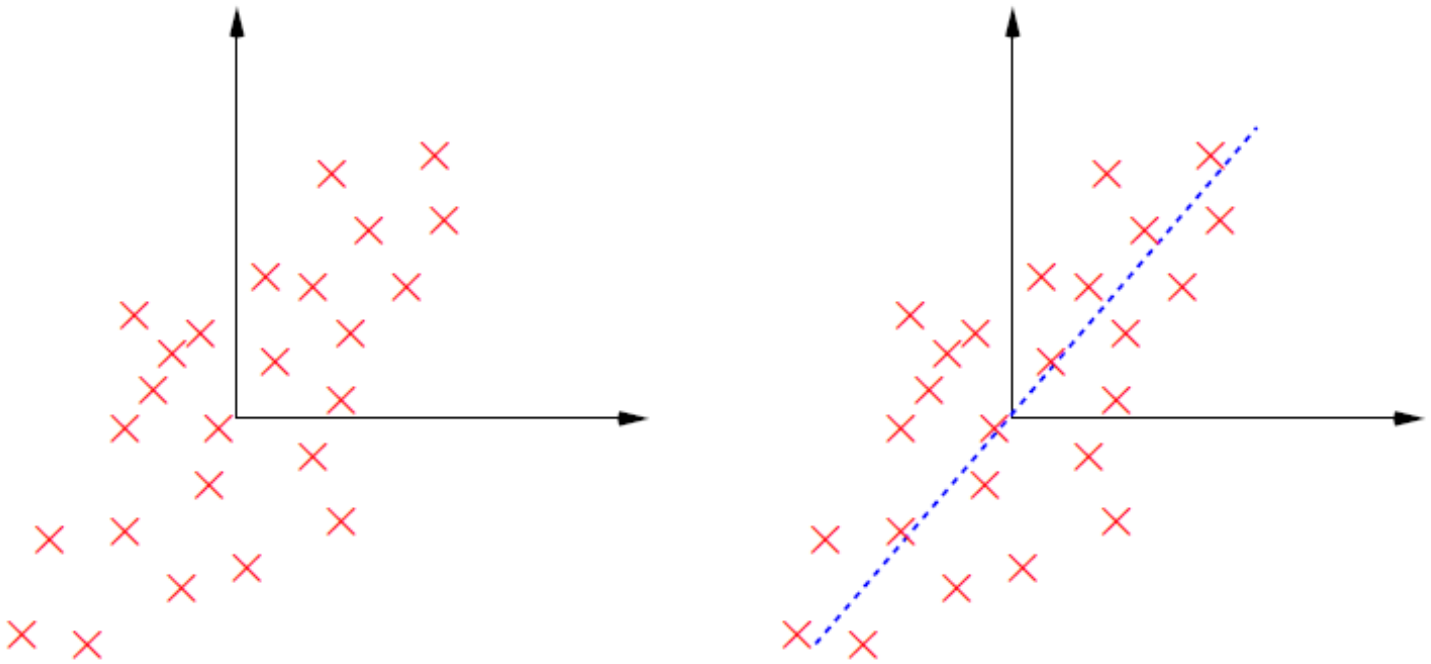
Reducción de dimensionalidad

Eliminar variables: Eliminar variables que no se consideren importantes apriori (redundantes o irrelevantes). **PERO** en ocasiones se está perdiendo información sin sustento técnico.

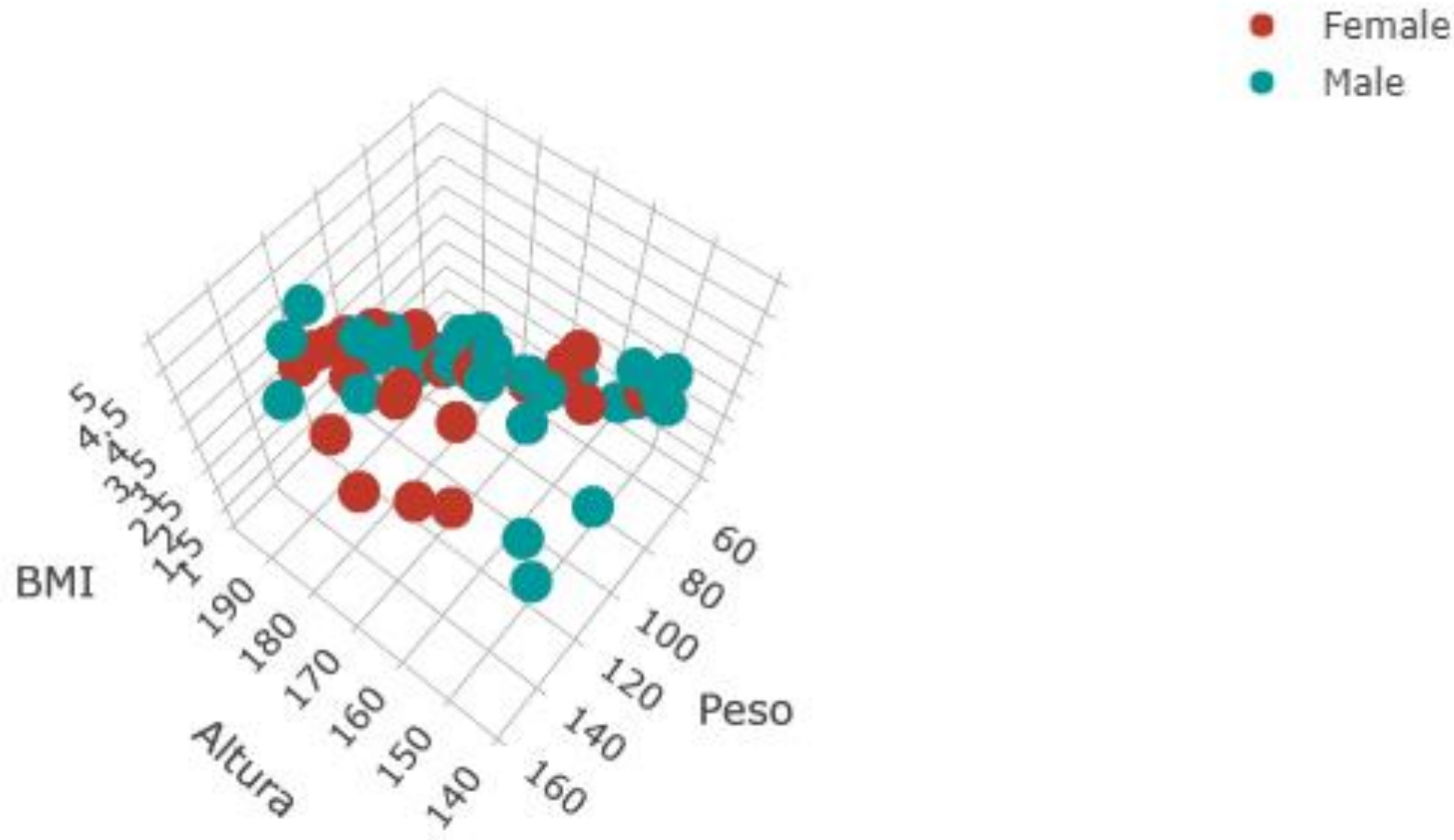
Seleccionar variables: Utilizar test estadísticos para seleccionar las variables más relevantes. Ejemplo: selección de variables en un modelo de predicción (t-test, comparación AIC, regularización). Se evita hacer sobre ajuste de modelos pero se puede igualmente perder información.

Creación de variables: Cuando dos o más variables se pueden resumir en una sola. Ejemplo: Horas de sueño = 24 – Horas de trabajo – Horas de esparcimiento. Útil cuando se sabe que se tienen variables correlacionadas, esto puede evitar problemas de multicolinearidad en los modelos.

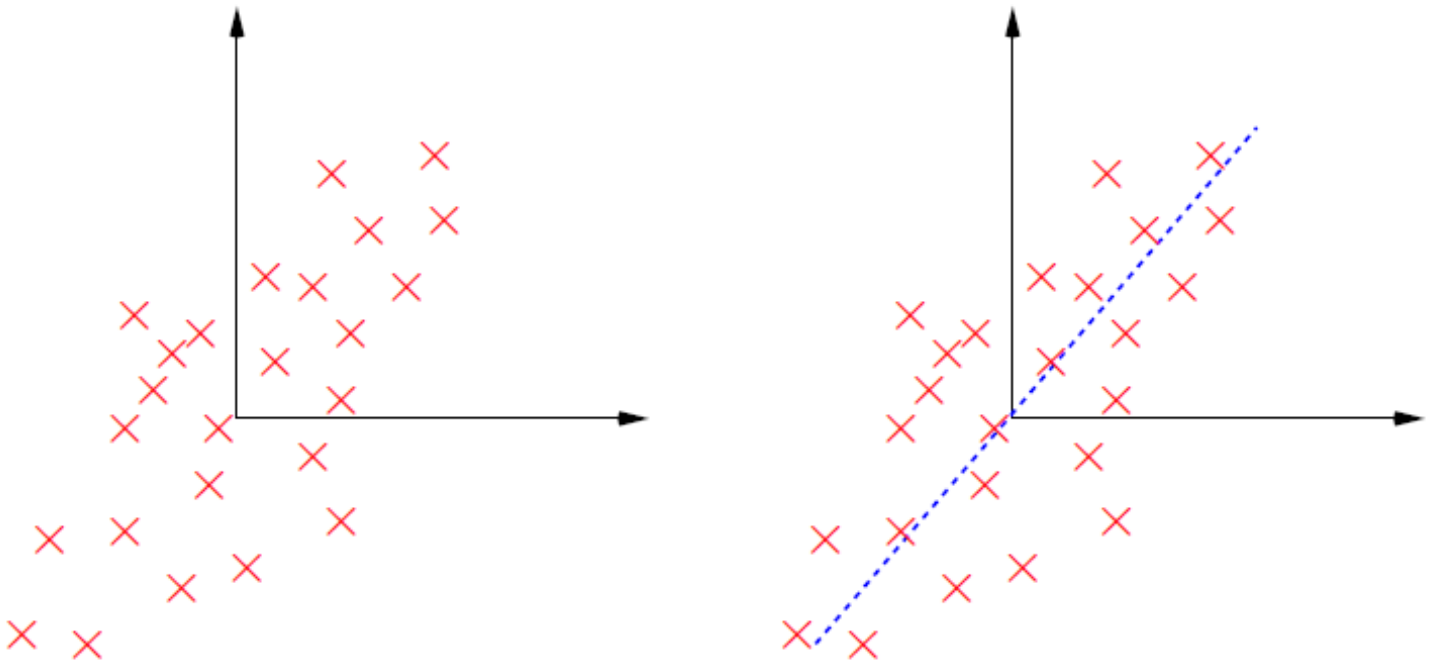
Análisis de componentes principales



Se busca reducir la dimensionalidad de una espacio de observaciones, llevando los vectores originales que pertenecen a un espacio $p - dimensional$ a unos nuevos vectores proyectados en un espacio $m - dimensional$. Donde $m < p$

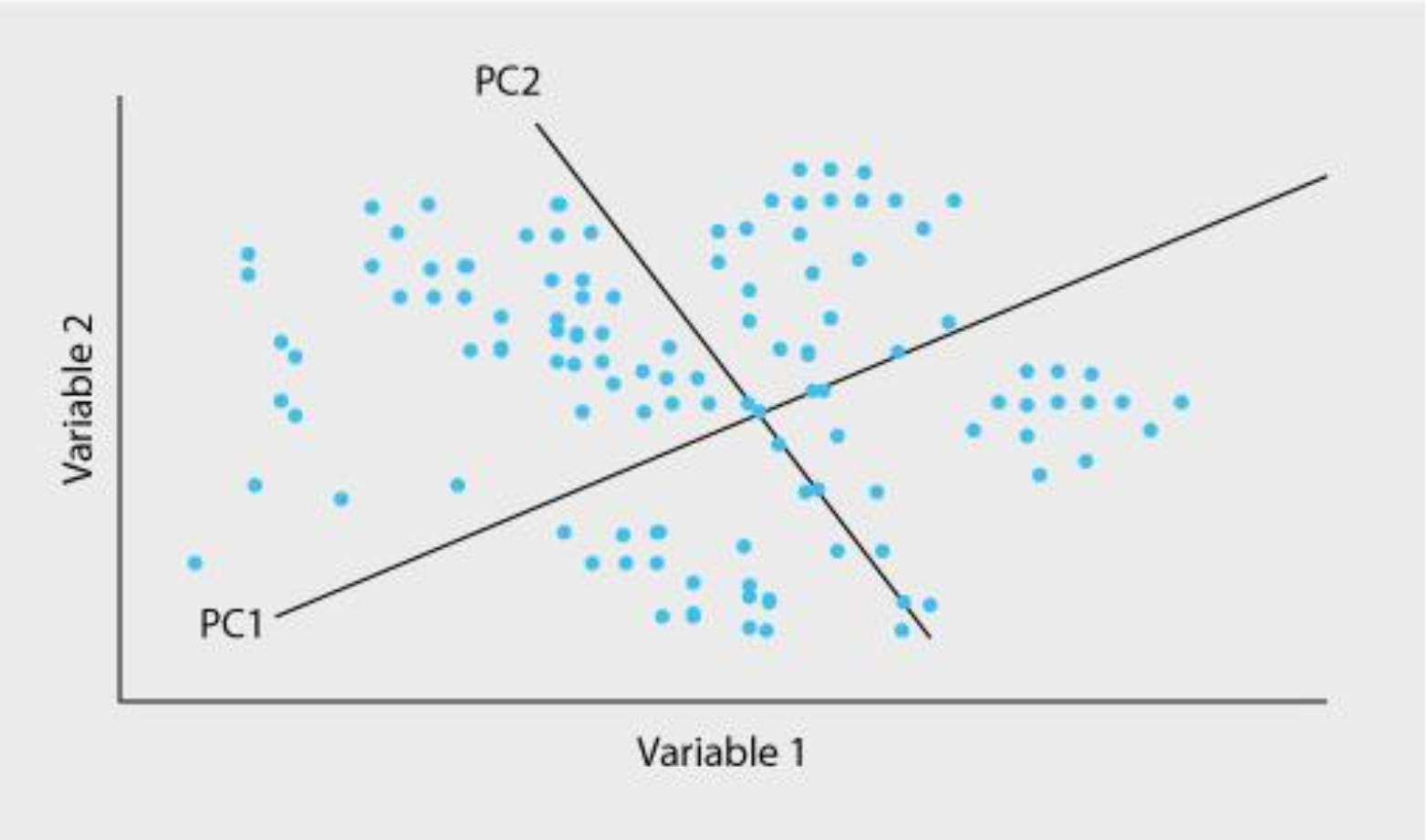

Female
Male
BMI
5
4.5
4
3.5
3
2.5
2
1.5
Altura
190
180
170
160
150
140
Peso
60
80
100
120
140
160

Análisis de componentes principales



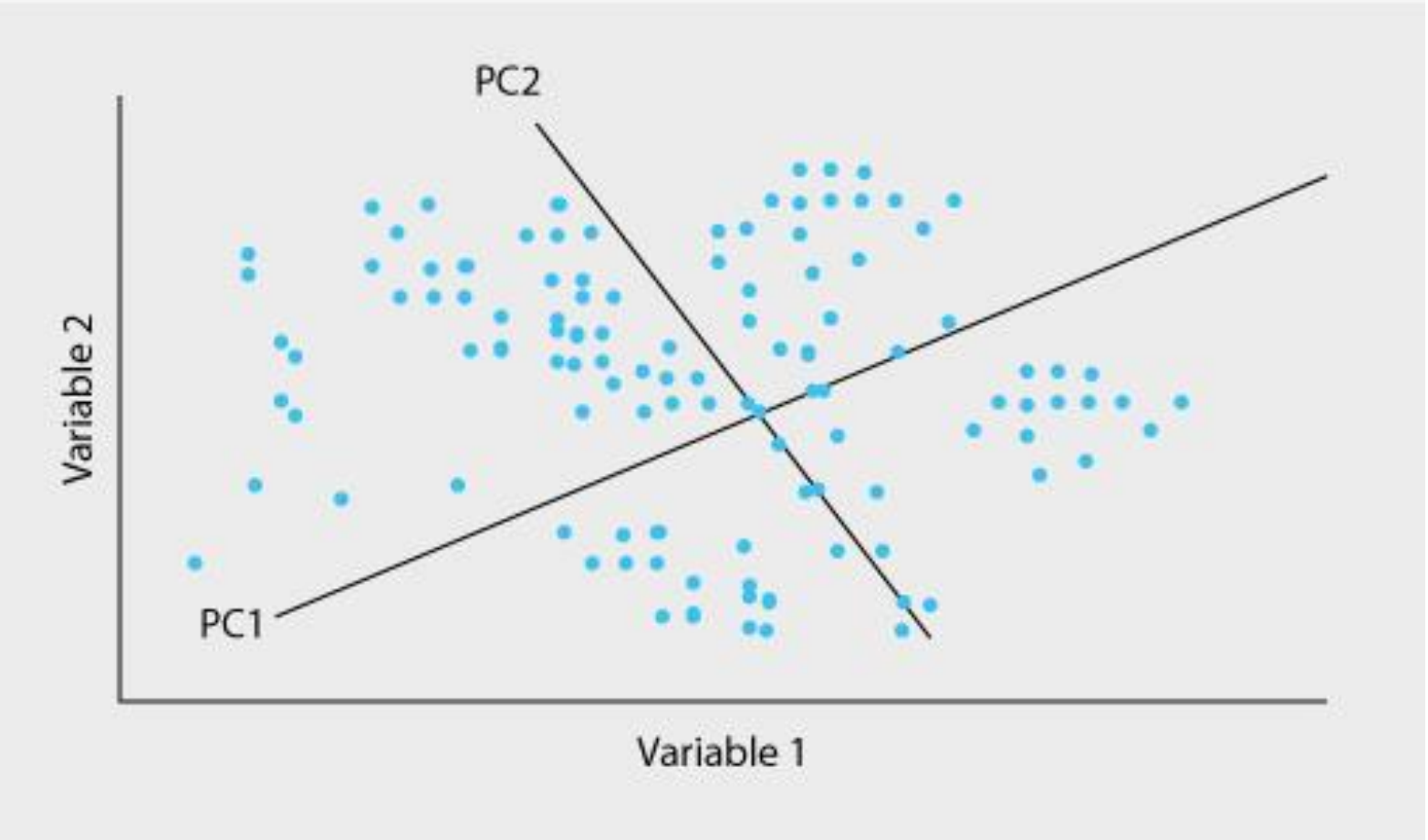
La proyección en las nuevas dimensiones se hace en la dirección donde se observa la mayor variabilidad

Análisis de componentes principales



Los vectores originales son representados como combinaciones lineales de las nuevas componentes, resumiendo la mayor cantidad de información en un menor número de dimensiones

Análisis de componentes principales



Tanto las variables como las observaciones originales tendrán unas nuevas ‘coordenadas’ en las nuevas dimensiones o componentes

Análisis de componentes principales

Sean $\boldsymbol{X_1, X_2, \dots, X_p}$ variables y $\boldsymbol{\Sigma = cov(x)}$ su correspondiente matriz de covarianza

Las componentes principales $\mathrm{CP_j}$ con $\boldsymbol{j = 1, \dots, p}$. Correspondientes a las variables $\boldsymbol{X_1, X_2, \dots, X_p}$ estarán dadas por:

$$\boldsymbol{CP_j = e_j' X = e_{j1} x_1 + e_{j2} x_2 + \cdots + e_{jp} x_p}$$

Donde los $\boldsymbol{e_j}$ son los correspondientes **vectores propios** de la matriz Σ

Es decir, cada CP estará dado por una combinación lineal de las variables originales

Análisis de componentes principales

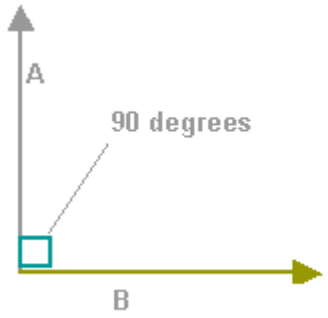
Se busca que las componentes contengan el máximo de la información original.

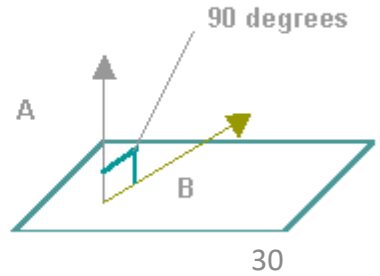
La cantidad de información que recolecta cada componente se puede definir por su varianza.

$$var(CP_j) = \lambda_j$$

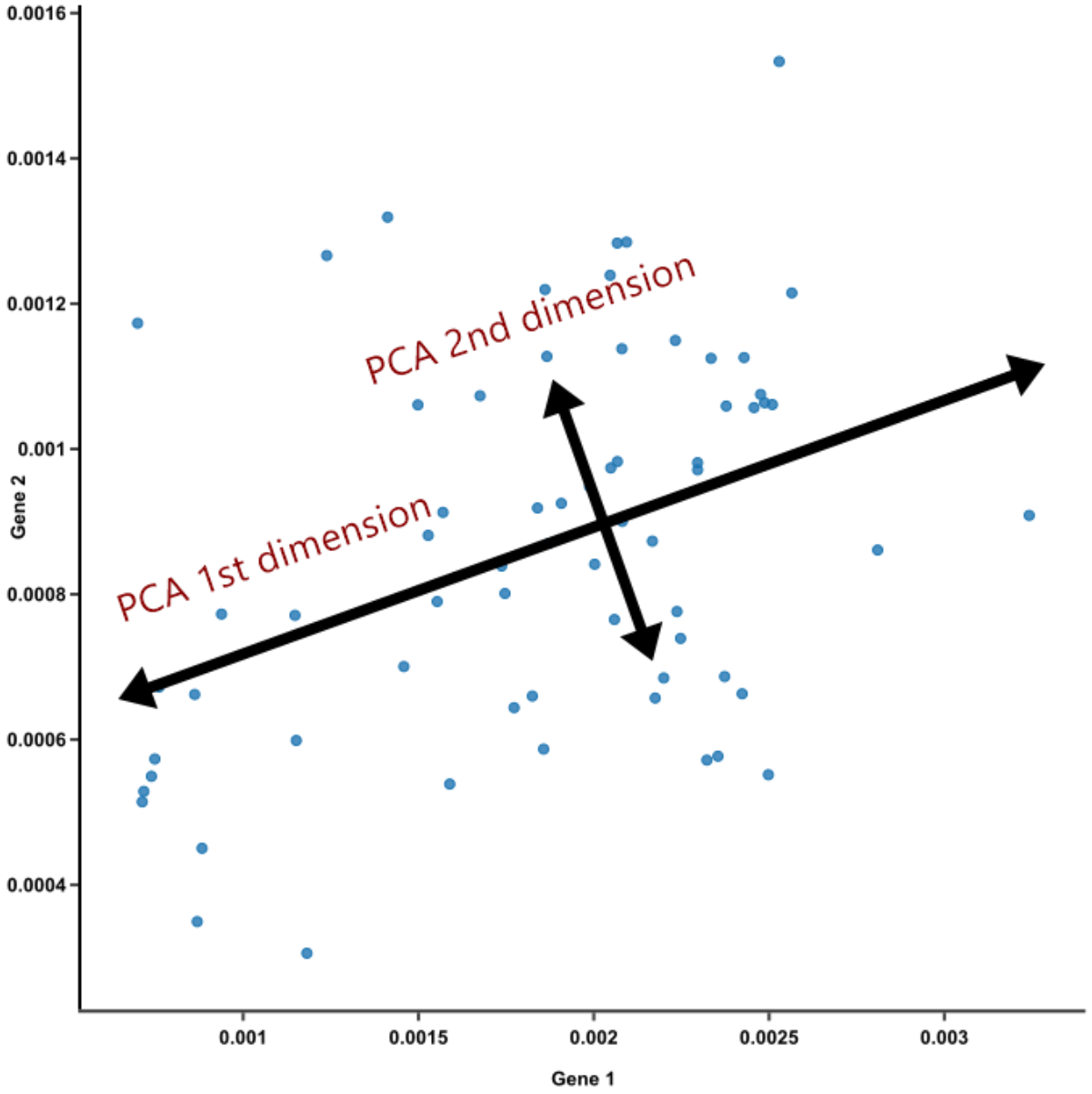
La varianza de cada uno de los componentes principales está dada por los correspondientes **valores propios λ_j** de la matriz Σ.

Además se cumple que $cov(CP_k, CP_j) = 0 \; para \; k \neq j$





Análisis de componentes principales



Elección del número de componentes principales

Siempre se tiene que la primer componente corresponde a la dirección donde hay mayor variación. El segundo componente corresponde a la dirección que abarca la mayor parte de la variación después de la primera, y así sucesivamente.

$$Información\ total = var\ (CP_1) + var\ (CP_2) + \cdots + var\ (CP_p)$$

$$= \boldsymbol{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

$$= \boldsymbol{var(X_1) + var(X_2) + \cdots + var(X_p)}$$

La primera componente será el que acumule la mayor varianza, la segunda componente tendrá la segunda mayor varianza y así sucesivamente…

$$\boldsymbol{\lambda_1 > \lambda_2 > \cdots > \lambda_p}$$

Elección del número de componentes principales

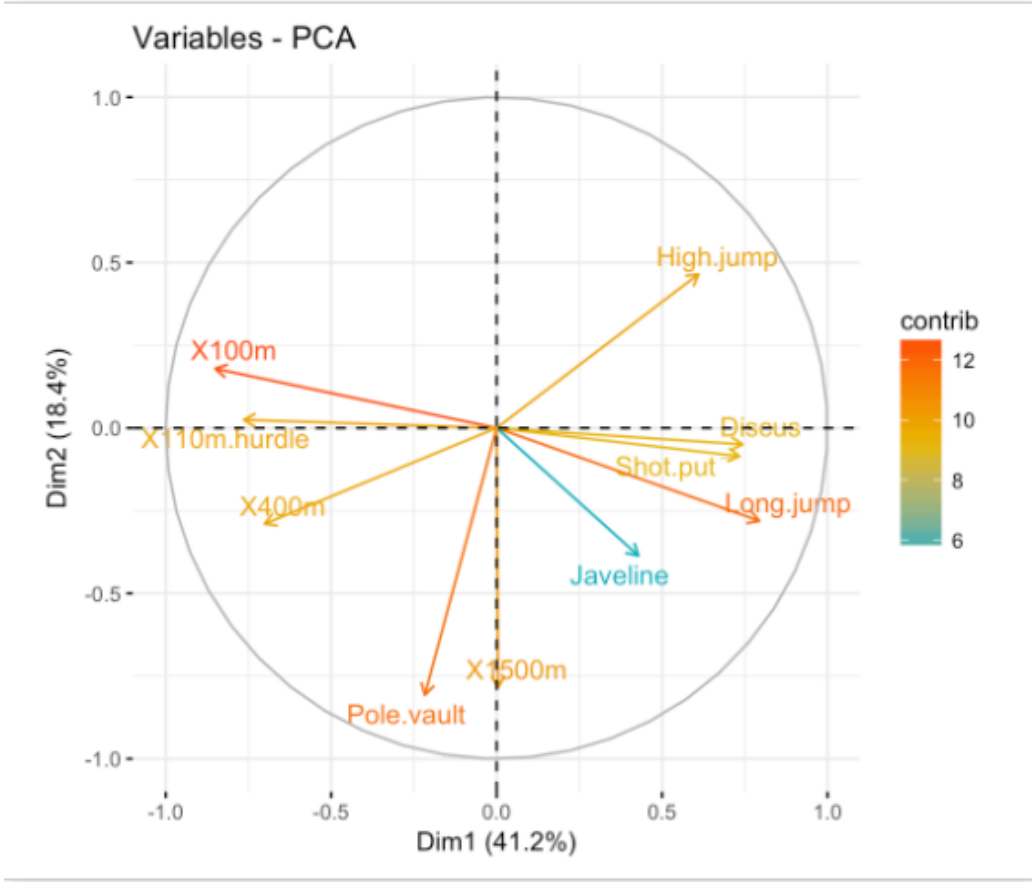
Entonces puedo determinar la proporción de información o variabilidad que recoge cada componente

$$\%var_j = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

La suma de las $\%var_j$ será 100.

Elección del número de componentes principales

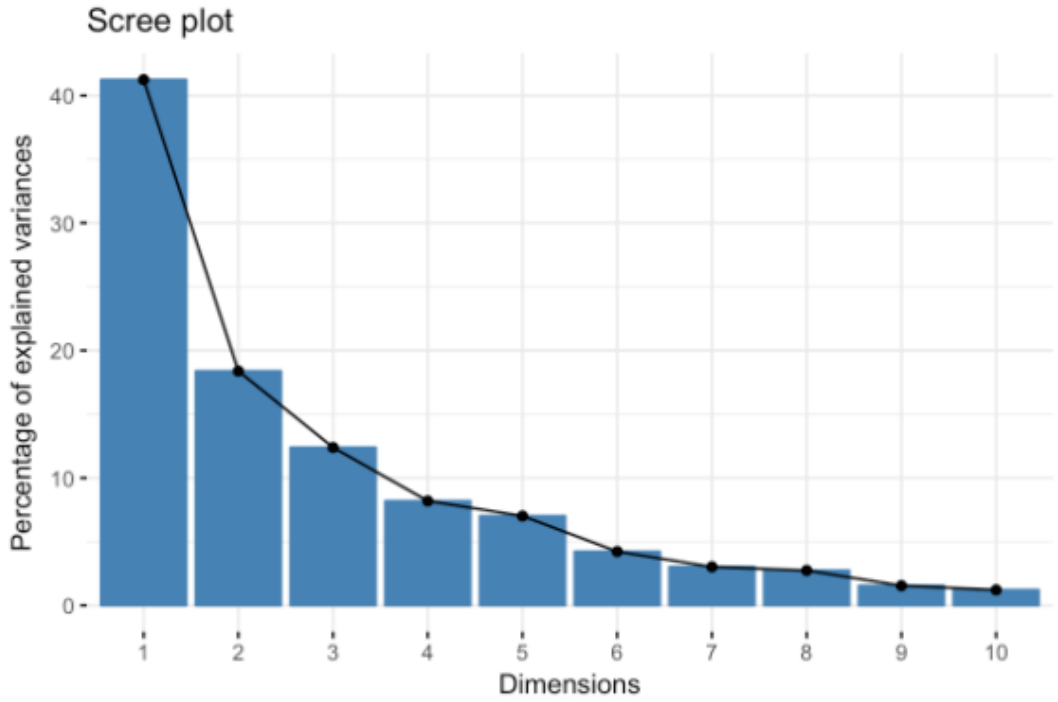
Elegimos la representación que mayor cantidad de información pueda resumir



http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

Elección del número de componentes principales

Elegimos la representación que mayor cantidad de información pueda resumir

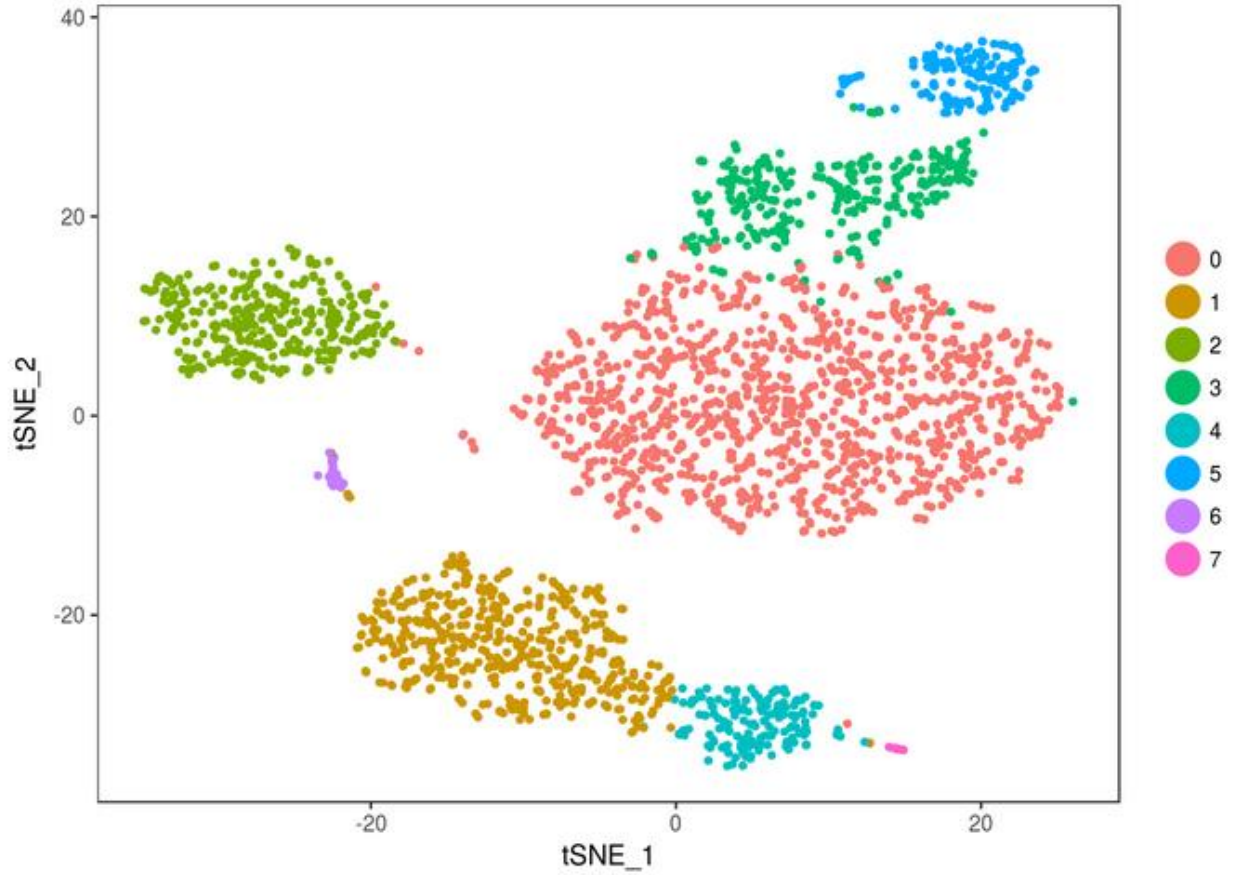


http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

Depende también del motivo por el cuál hacemos reducción de dimensionalidad

t- SNE (t-Distributed Stochastic Neighbor Embedding)

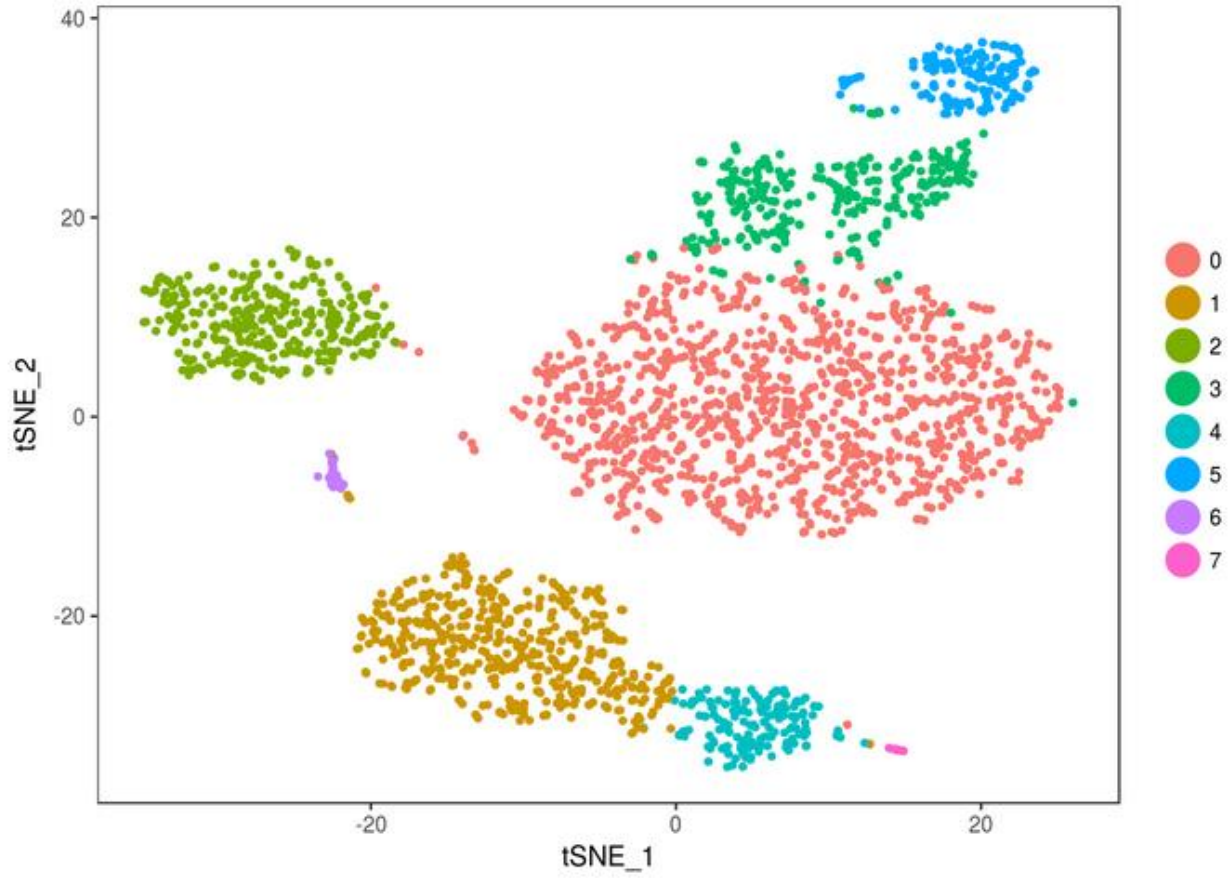
Es una técnica no lineal de reducción de dimensionalidad.



Busca minimizar la divergencia entre la distribución que mide la similitud de los puntos en el espacio original (puntos observados) y la distribución que mide la similitud de los puntos proyectados en un espacio de menor dimensiones

t- SNE (t-Distributed Stochastic Neighbor Embedding)

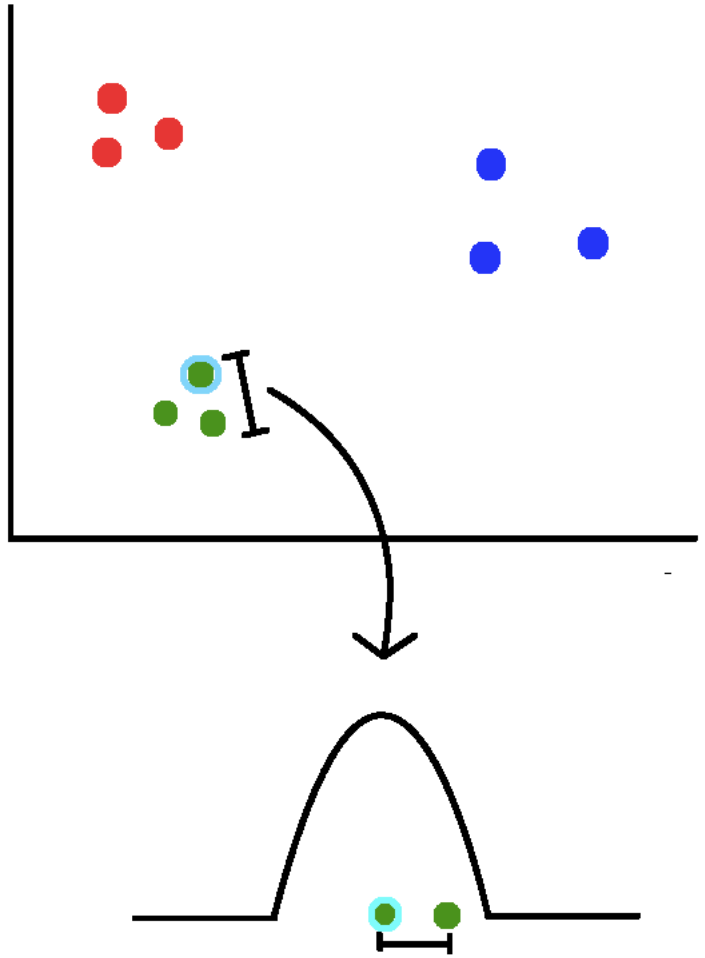
Minimiza las diferencias entre la distribución de puntos originales y la distribución de puntos proyectados en las nuevas dimensiones



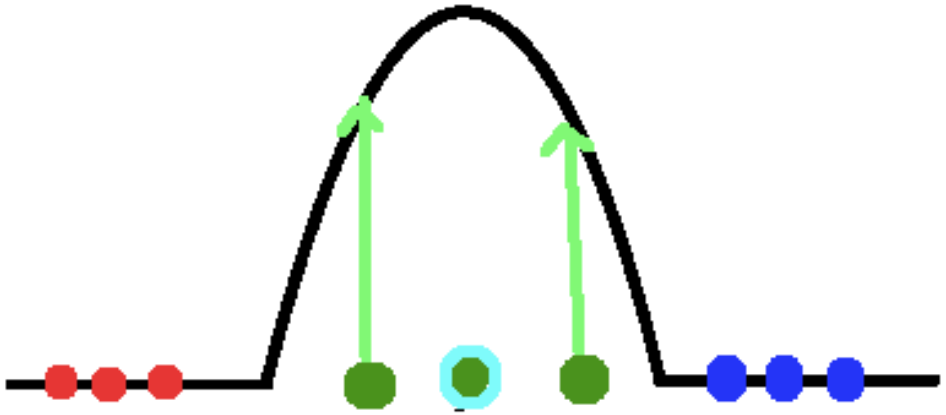
Se utiliza para visualizar no como un pre-procesamiento del cuál obtienes una salida (como en ACP)

t- SNE (t-Distributed Stochastic Neighbor Embedding)

1. En datos originales: Probabilidad de similitud de puntos, probabilidad condicional de ser vecinos



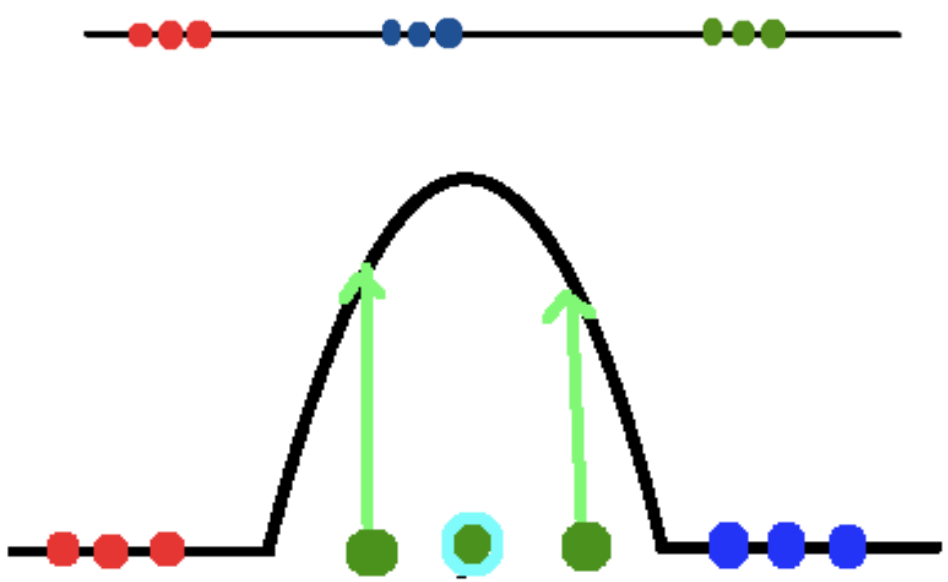
Distribución centrada en el punto con respecto al cual calculo similitudes con los restantes



https://towardsdatascience.com/t-sne-python-example-1ded9953f26

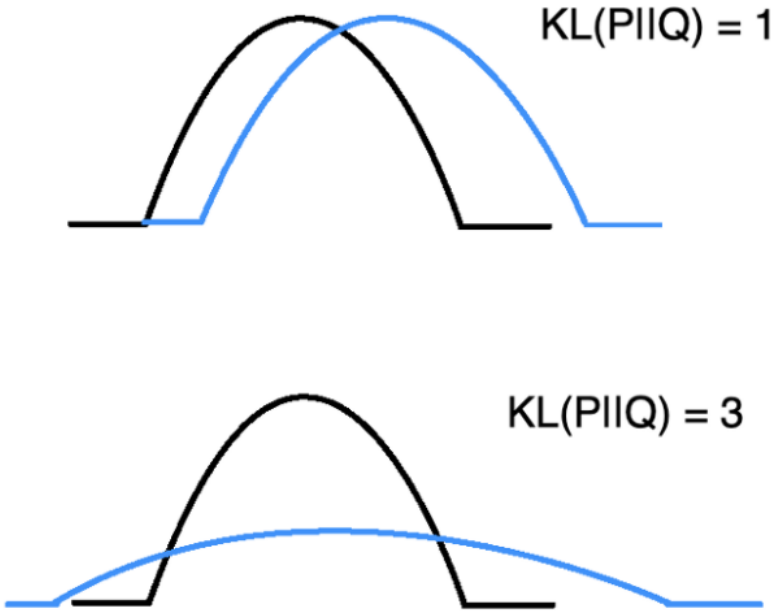
t- SNE (t-Distributed Stochastic Neighbor Embedding)

2. En el espacio de menor dimensión también se calcula la distribución de similitudes



https://towardsdatascience.com/t-sne-python-example-1ded9953f26

3. Se minimiza la diferencia entre la distribución de similitudes puntos originales y la distribución similitud de puntos proyectados en menor dimensiones



Se minimiza la medida de divergencia Kullback-Leibler (KL, mide la diferencia entre dos distribuciones de probabilidad)

ACP Vs t-SNE

- t-SNE es computacionalmente más costoso que un ACP
- ACP es un algoritmo lineal, por lo cual no siempre puede representar relaciones polinomiales entre variables. t-SNE puede ser mejor capturando patrones no lineales
- Los resultado con ACP son únicos, no necesariamente es el caso con t-SNE (hay que revisar múltiples corridas)

ACP Vs t-SNE

- t-SNE tiene una naturaleza de ‘caja negra’ donde el investigador no tiene la información para interpretar cómo se llego a los resultados. En ACP se puede interpretar e identificar numéricamente cuál es la combinación de cada componente.

- Los resultados de t-SNE funcionan en gran medida para poder visualizar la estructura de los datos, pero no para generar de este un gran análisis cuantitativo.

- En ACP es posible relacionar los componentes con dimensiones que expliquen la estructura de la data. Indicadores, variables ‘latentes’.

Conclusiones

- Identificación y manejo de matrices, vectores y data frames
- Distancias y su importancia en algoritmos de análisis de datos
- Importancia de la reducción de dimensionalidad
- Técnicas de reducción de dimensionalidad