



RECONOCIMIENTO DE PATRONES EN IMÁGENES TICS 585

FACULTAD DE INGENIERÍA Y CIENCIAS
UNIVERSIDAD ADOLFO IBÁÑEZ

SEGUNDO SEMESTRE 2021

PROFESOR: MIGUEL CARRASCO

COMPRESIÓN DE DATOS

■ Compresión

- Análisis de Componentes Principales (PCA)
- Vector Quantisation

■ Selección

- Discriminante de Fisher
- Búsqueda exhaustiva y secuencial
- “Plus L-take away R”
- Branch & Bound

Ya obtuve mis
características,
¿cómo selecciono
las mejores?





¿Qué es el Análisis de Componentes Principales (PCA)?

Objetivo:

Representar los datos en una menor dimensión

¿Qué realiza?

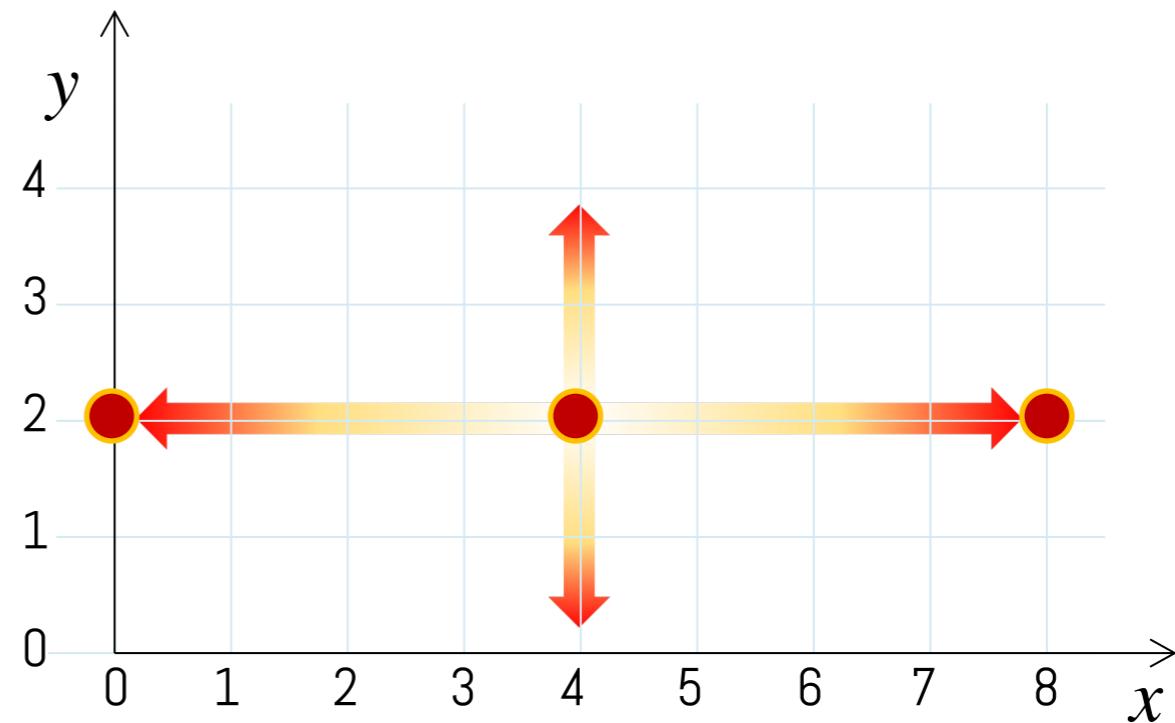
En un procedimiento matemático que **permite identificar patrones en los datos y reducir el número de dimensiones con baja pérdida de información**. Esto lo realiza a través de transformación lineal a otro espacio empleando los vectores de máxima varianza.

¿Cómo funciona?

Transforma las variables correlacionadas a un limitado número de variables no correlacionadas a través de los vectores y valores propios.

variables

variables	
x	y
0	2
4	2
8	2

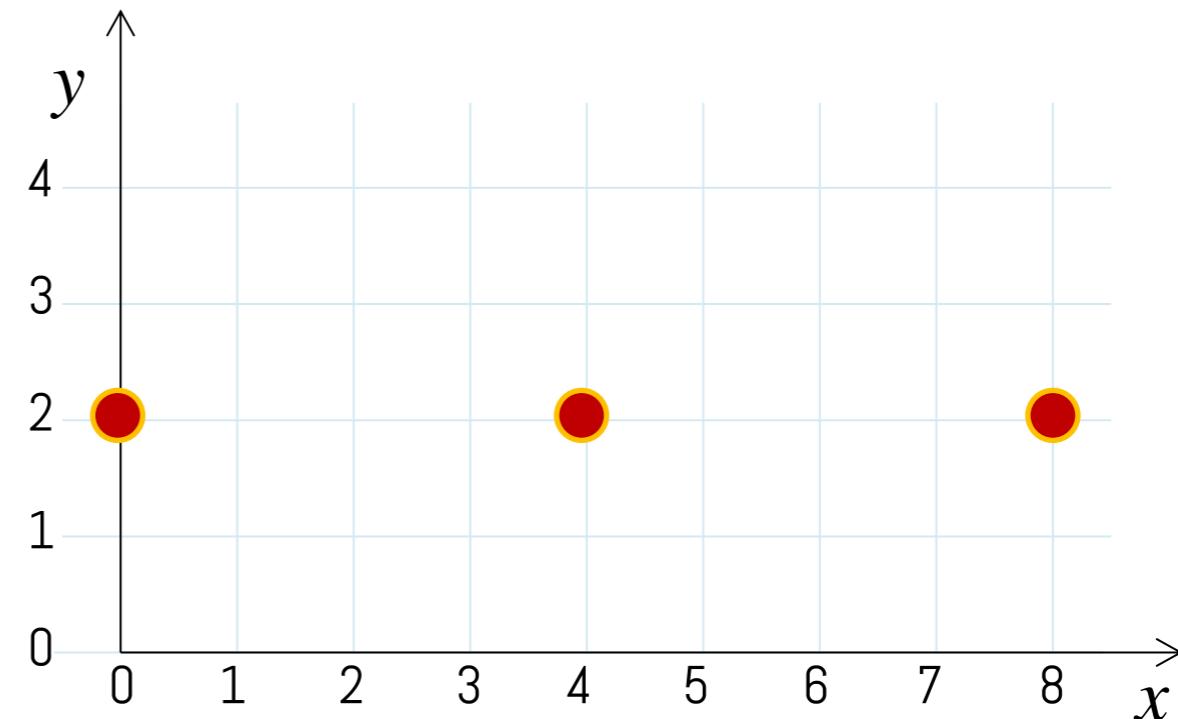


¿En qué **eje** se presenta variación de los datos?

Observamos una dispersión en el eje X. Para ello estimaremos la varianza muestral en cada variable

variables

		variables	
		<i>x</i>	<i>y</i>
datos		0	2
$\mu =$		4	2
$S(X) =$		16	0



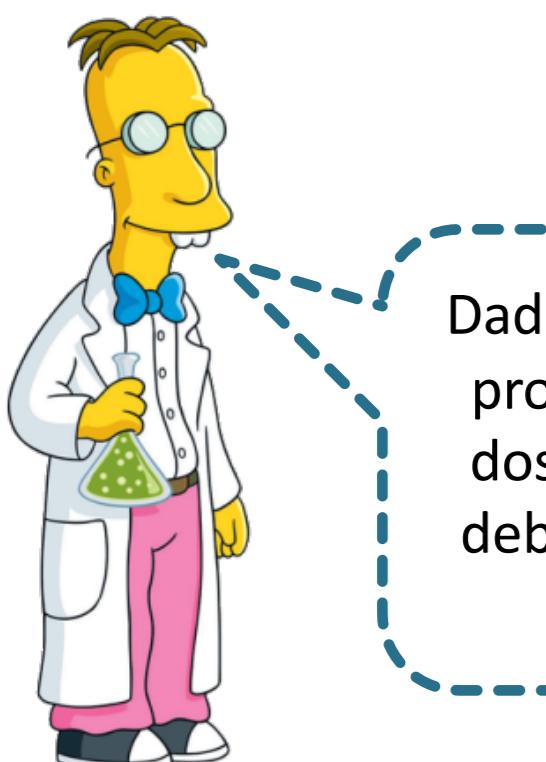
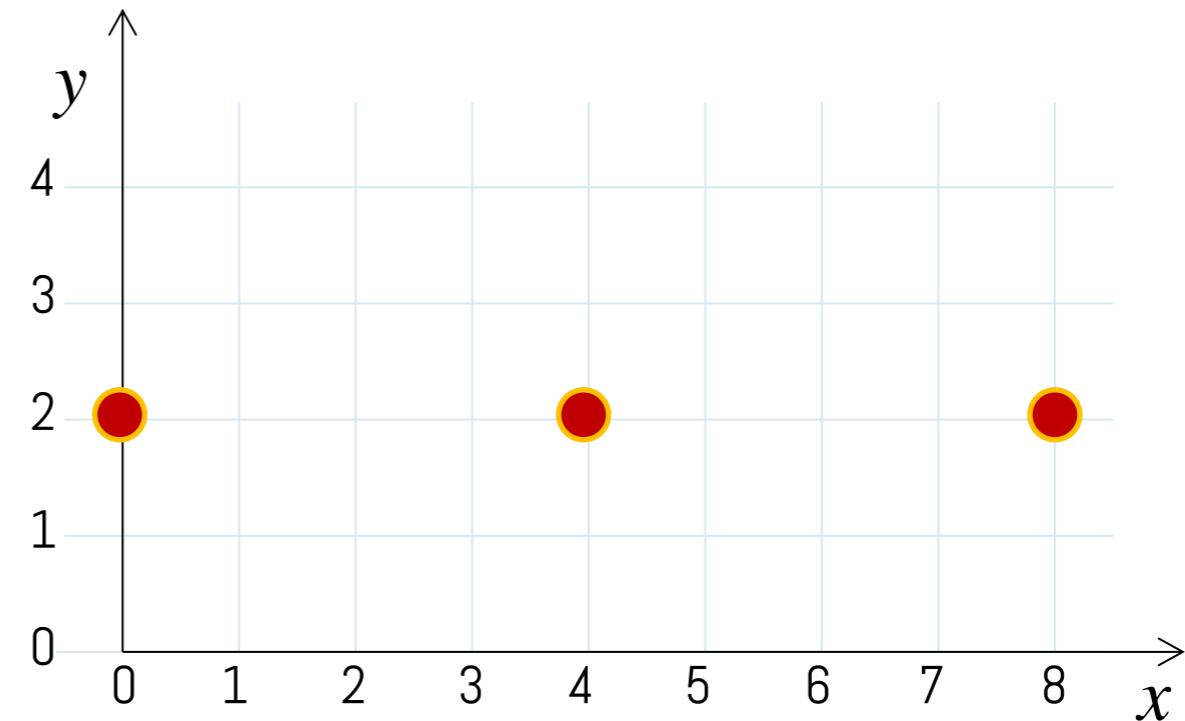
Ya sabemos en qué variable hay mayor variación, pero ¿cómo podemos estimar la variación entre variables ?

$$S(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{2} ((0-4)^2 + (4-4)^2 + (8-4)^2) = 16$$

$$S(y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \mu)^2 = \frac{1}{2} ((2-2)^2 + (2-2)^2 + (2-2)^2) = 0$$

variables

		variables	
		x	y
datos		0	2
		4	2
		8	2
$\bar{x} = 4$		$\bar{y} = 2$	
$S(X) =$		16	0



Dado que en nuestro problema tenemos dos variables $\{x, y\}$, debemos calcular la covarianza

$$\begin{aligned} \text{cov}(x, y) &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ \text{cov}(x, x) &= \frac{1}{2} ((0-4)^2 + (4-4)^2 + (8-4)^2) = 16 \end{aligned}$$

$$C = \begin{bmatrix} \text{cov}(x, x) & \text{cov}(x, y) \\ \text{cov}(y, x) & \text{cov}(y, y) \end{bmatrix} = \begin{bmatrix} 16 & 0 \\ 0 & 0 \end{bmatrix}$$

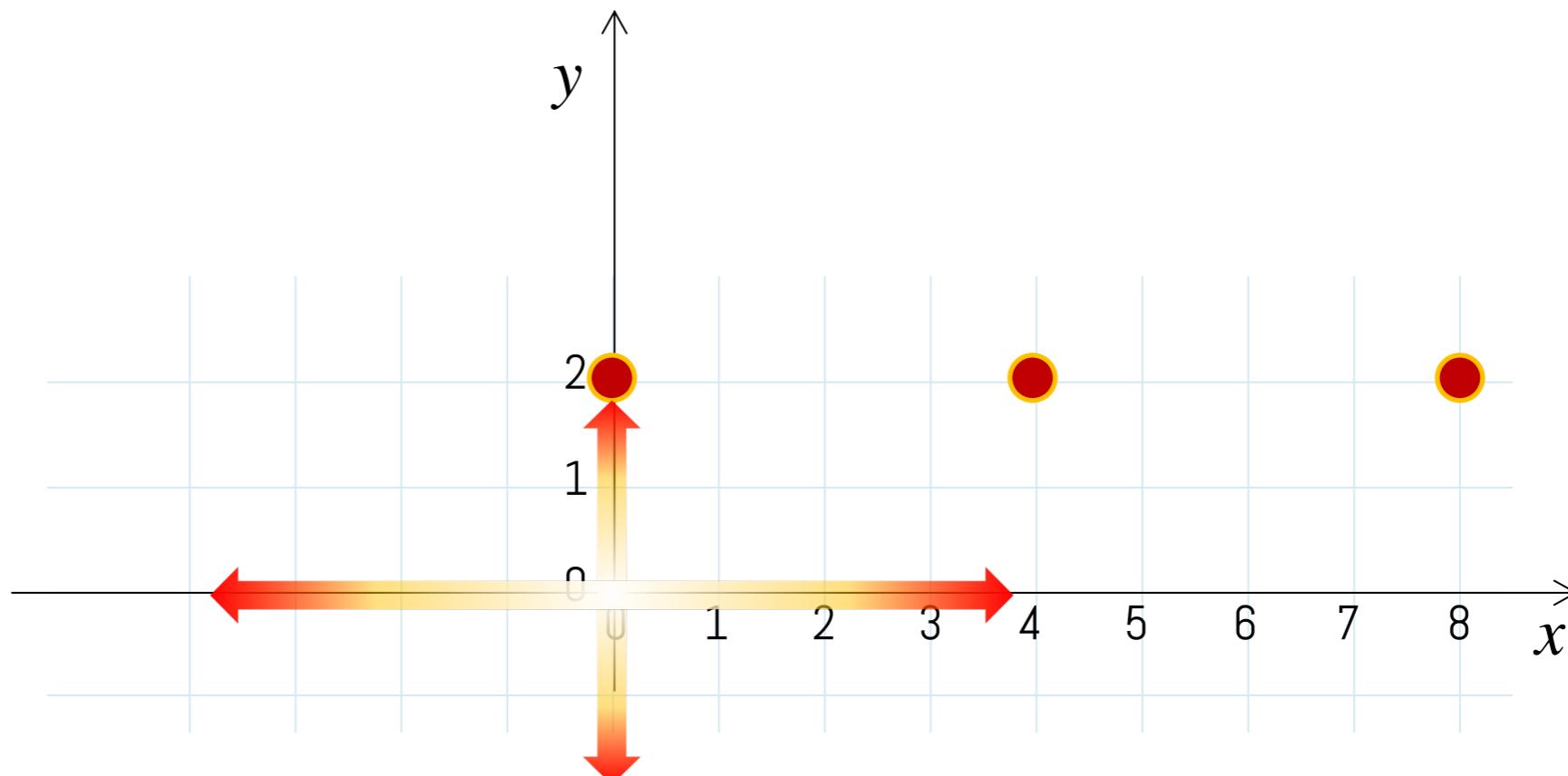
En esta matriz solo vemos una componente en el eje X

Análisis de componentes principales (PCA)

variables

x	y
0	2
4	2
8	2
$\bar{x} = 4$	$\bar{y} = 2$

datos



variables

x	y
0	2
4	2
8	2

datos

Restamos las medias

$x - \mu_x$	$y - \mu_y$
0 - 4	2 - 2
4 - 4	2 - 2
8 - 4	2 - 2

resultado

x'	y'
-4	0
0	0
4	0

¿Qué ocurre con la matriz de covarianza al centrar los datos?

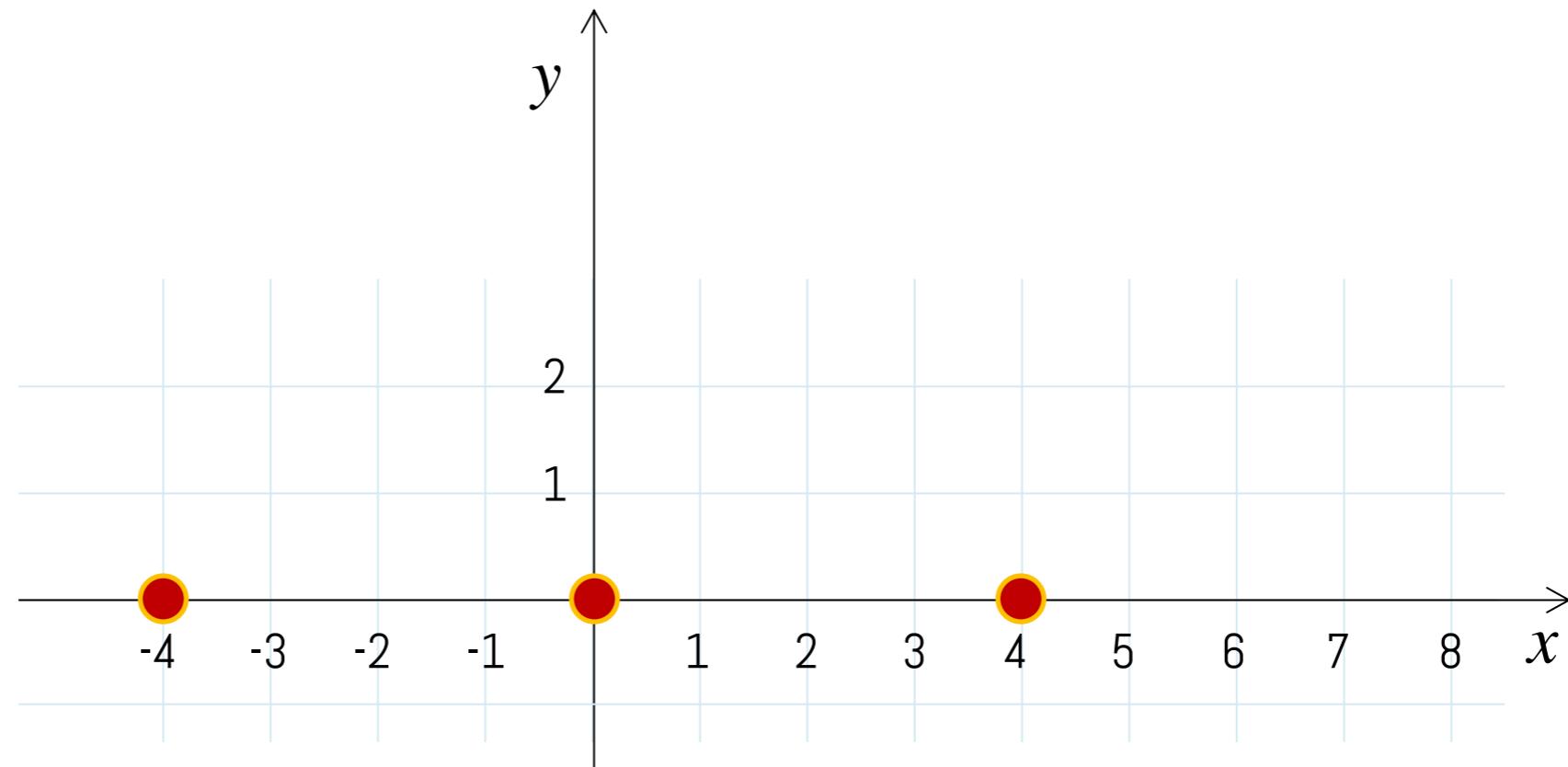


Análisis de componentes principales (PCA)

variables

x	y
-4	0
0	0
4	0

$\bar{x} = 0$ $\bar{y} = 0$



$$\text{cov}(x, x) = \frac{1}{2} ((-4 - 0)^2 + (0 - 0)^2 + (4 - 0)^2) = 16$$

$$\text{cov}(x, y) = \frac{1}{2} ((-4 - 0)(0 - 0) + (0 - 0)(0 - 0) + (4 - 0)(0 - 0)) = 0$$

$$C = \begin{bmatrix} \text{cov}(x, x) & \text{cov}(x, y) \\ \text{cov}(y, x) & \text{cov}(y, y) \end{bmatrix} = \begin{bmatrix} 16 & 0 \\ 0 & 0 \end{bmatrix}$$



Centrar los datos no genera un cambio en la dispersión de los datos

Como podemos observar, no hay cambios en la covarianza



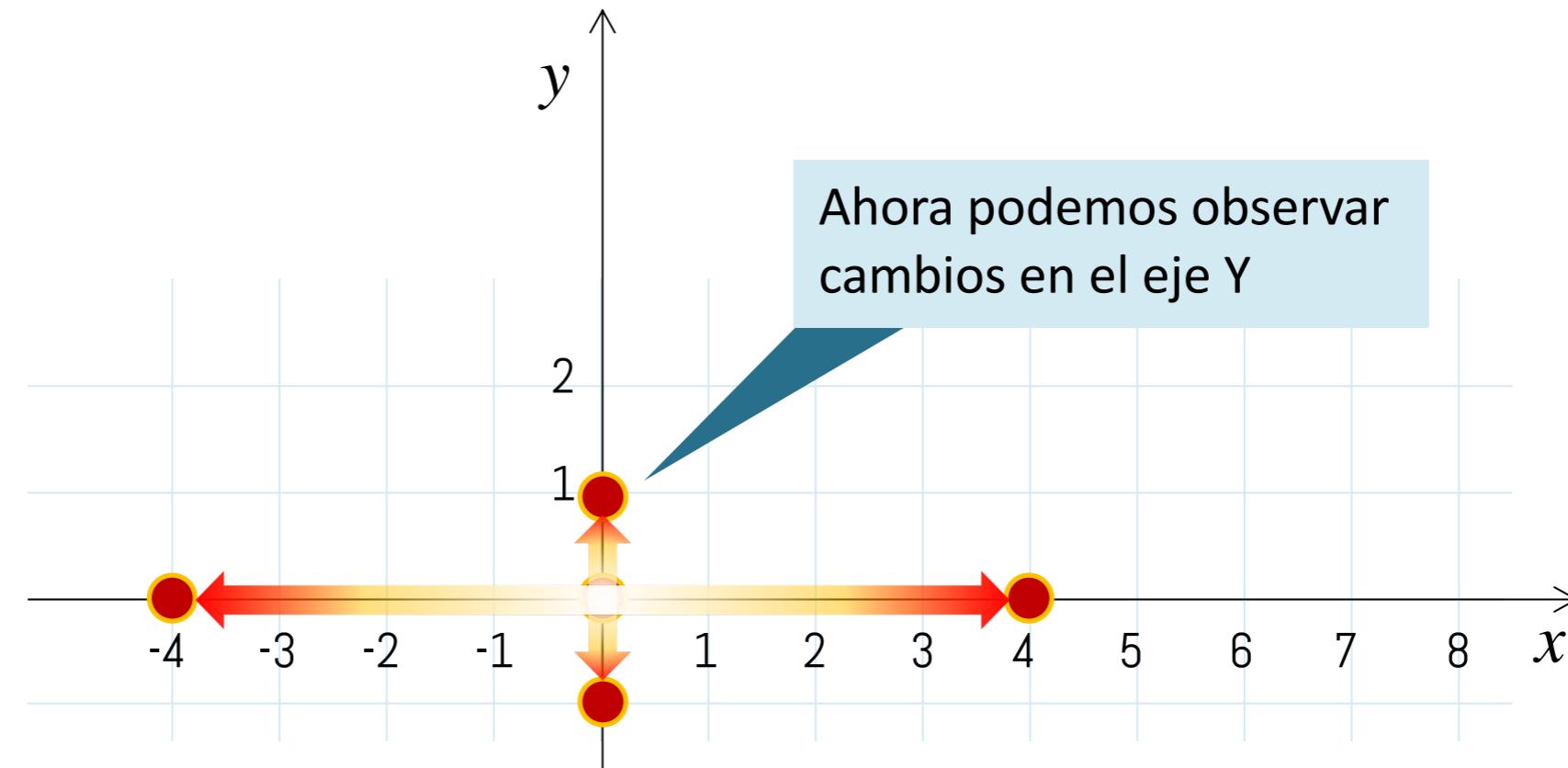
Análisis de componentes principales (PCA)

variables

x	y
-4	0
0	0
4	0
0	1
0	-1
$\bar{x} = 0$ $\bar{y} = 0$	



Qué ocurre si
agregamos dos puntos
en la coordenada Y
¿Hay un cambio en la
covarianza?

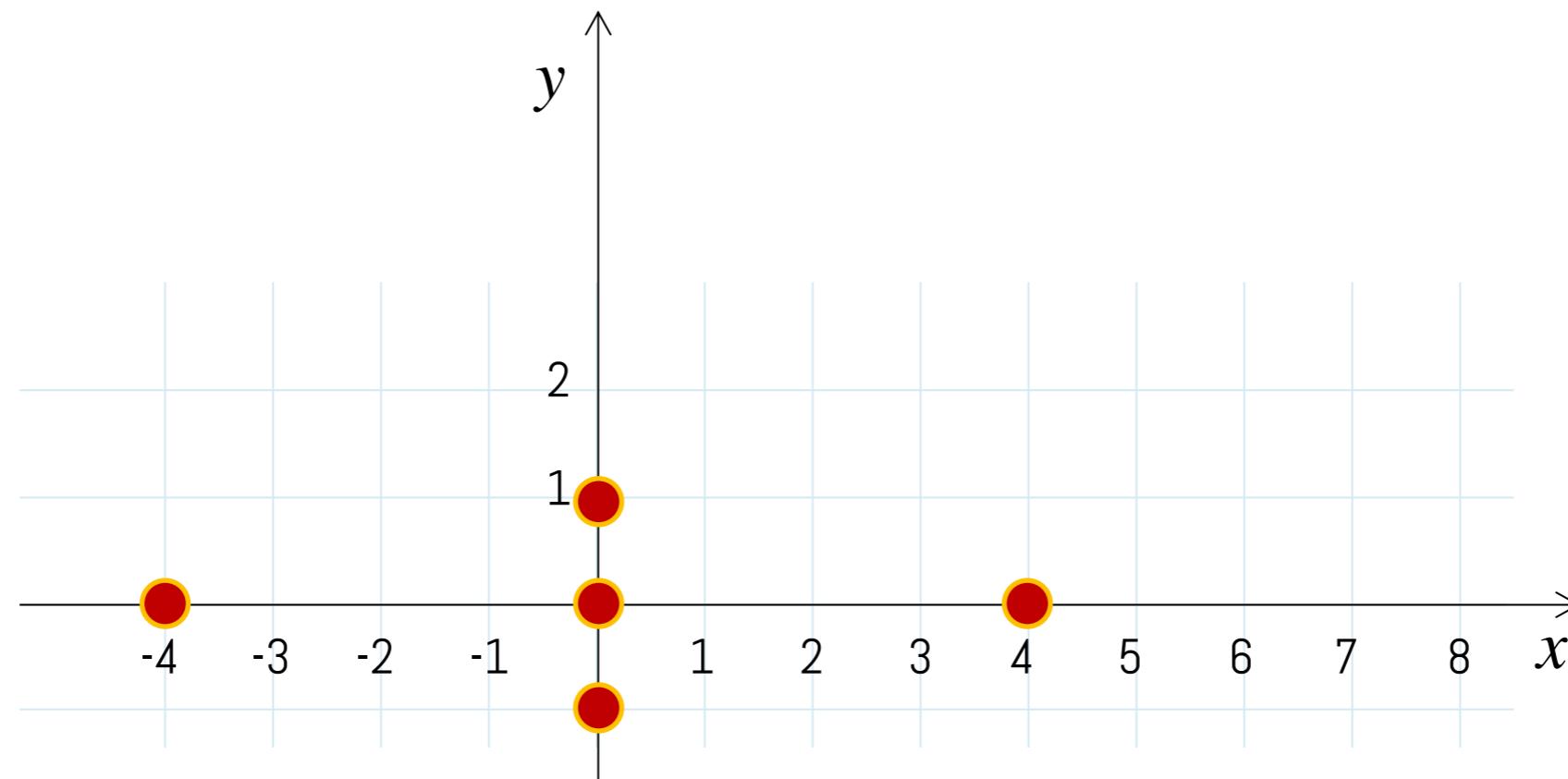


$$cov(x, x) = \frac{1}{4}((-4-0)^2 + (0-0)^2 + (4-0)^2 + (0-0)^2 + (0-0)^2) = 8$$

$$cov(y, y) = \frac{1}{4}((-0-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (-1-0)^2) = 0.5$$

$$C = \begin{bmatrix} cov(x, x) & cov(x, y) \\ cov(y, x) & cov(y, y) \end{bmatrix} = \begin{bmatrix} 8 & 0 \\ 0 & 0.5 \end{bmatrix}$$

Ahora podemos observar cambios en el eje Y

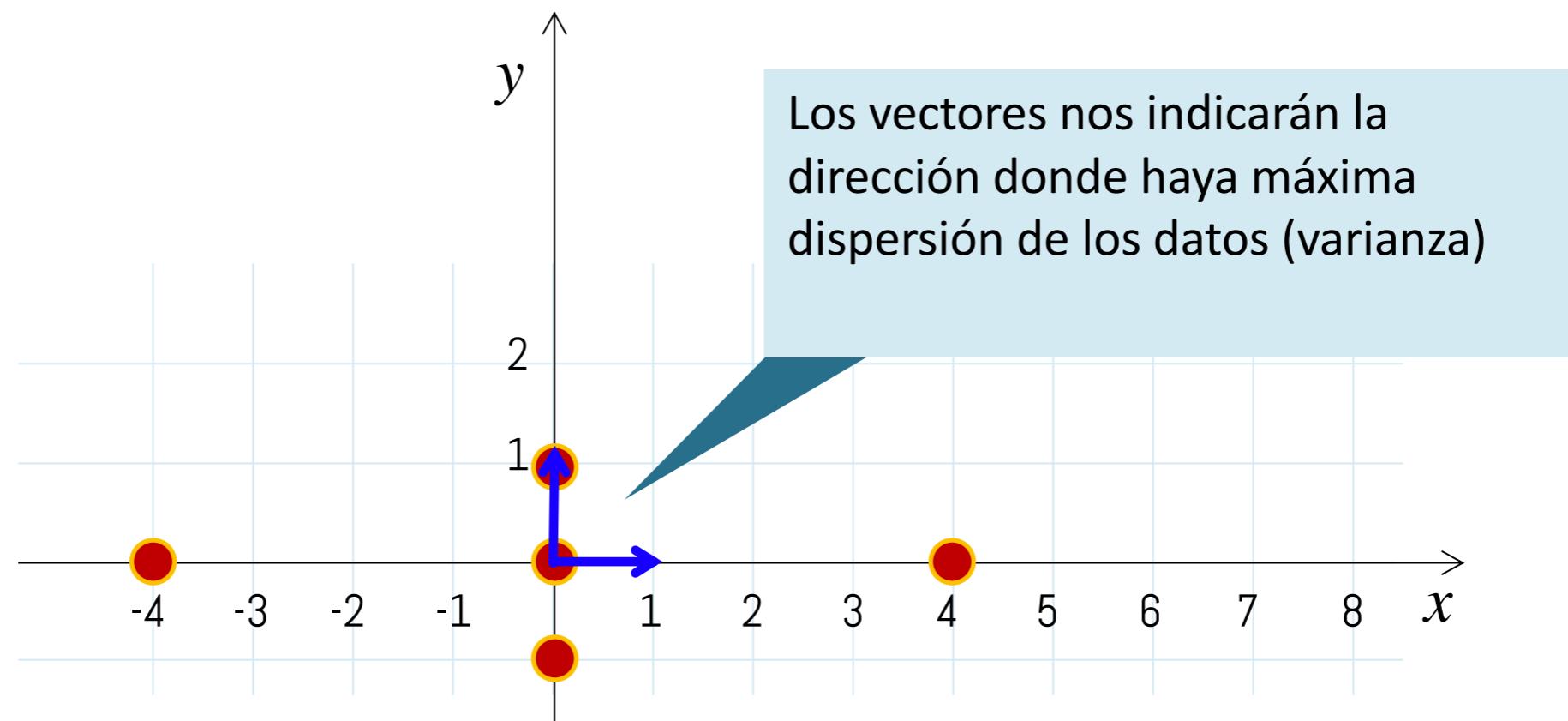


Dado que conocemos la dispersión entre los ejes, ahora necesitamos calcular un vector dirección

$$C = \begin{bmatrix} cov(x, x) & cov(x, y) \\ cov(y, x) & cov(y, y) \end{bmatrix} = \begin{bmatrix} 8 & 0 \\ 0 & 0.5 \end{bmatrix}$$

cambios en el eje x
cambios en el eje y

¿Cómo podemos calcular un vector que nos indique la dirección y magnitud de la variación por cada eje?



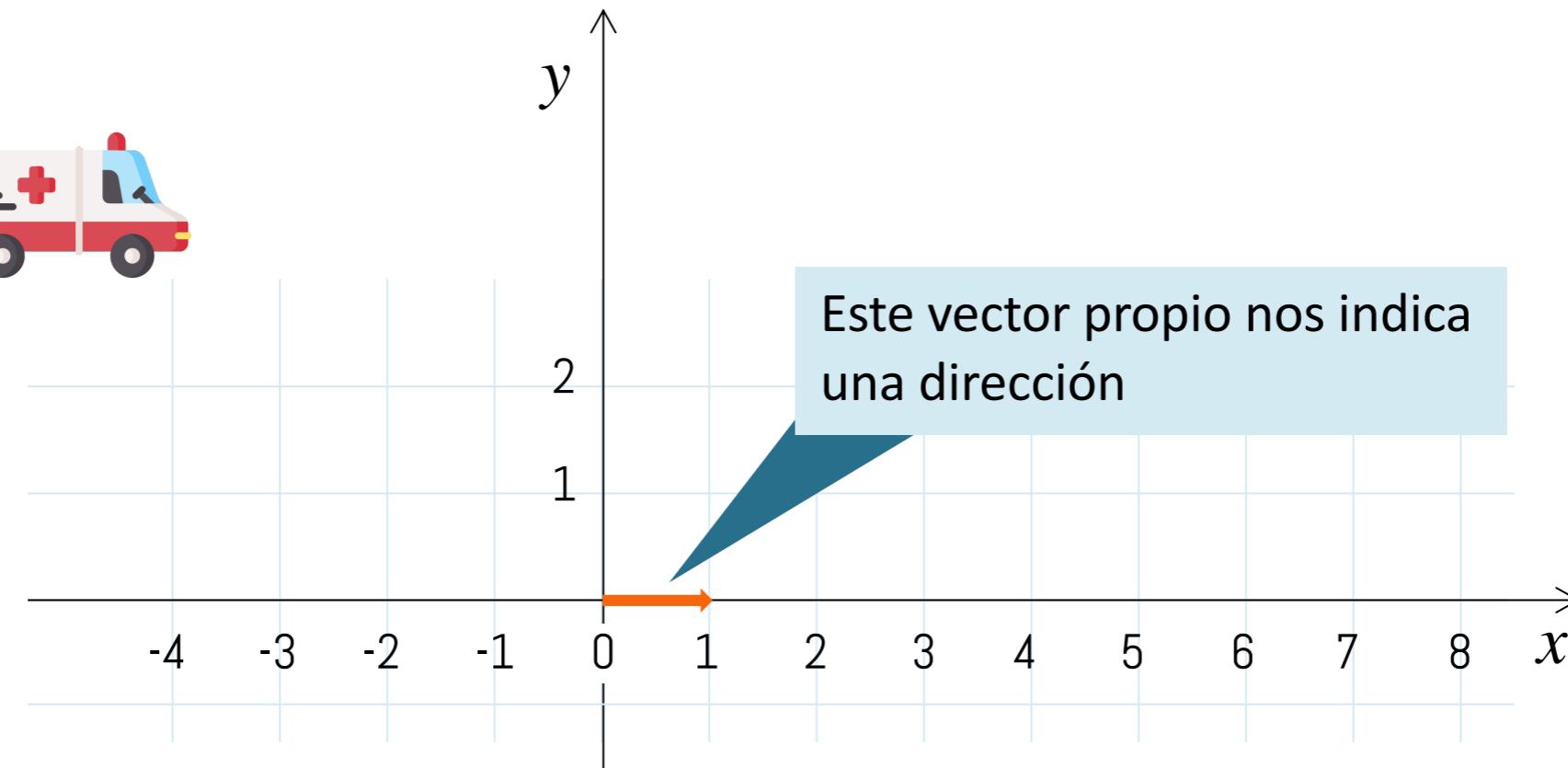
Necesitamos encontrar un vector que indique las magnitudes de los cambios por cada eje.



¿Cómo lo calculamos?
¿cómo?
¿Te acuerdas de Álgebra?



Vectores propios al rescate



$$\begin{bmatrix} 8 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \end{bmatrix}$$

Vectores y valores propios

$$\begin{bmatrix} 8 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 8 \\ 0 \end{bmatrix} = 8 \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$



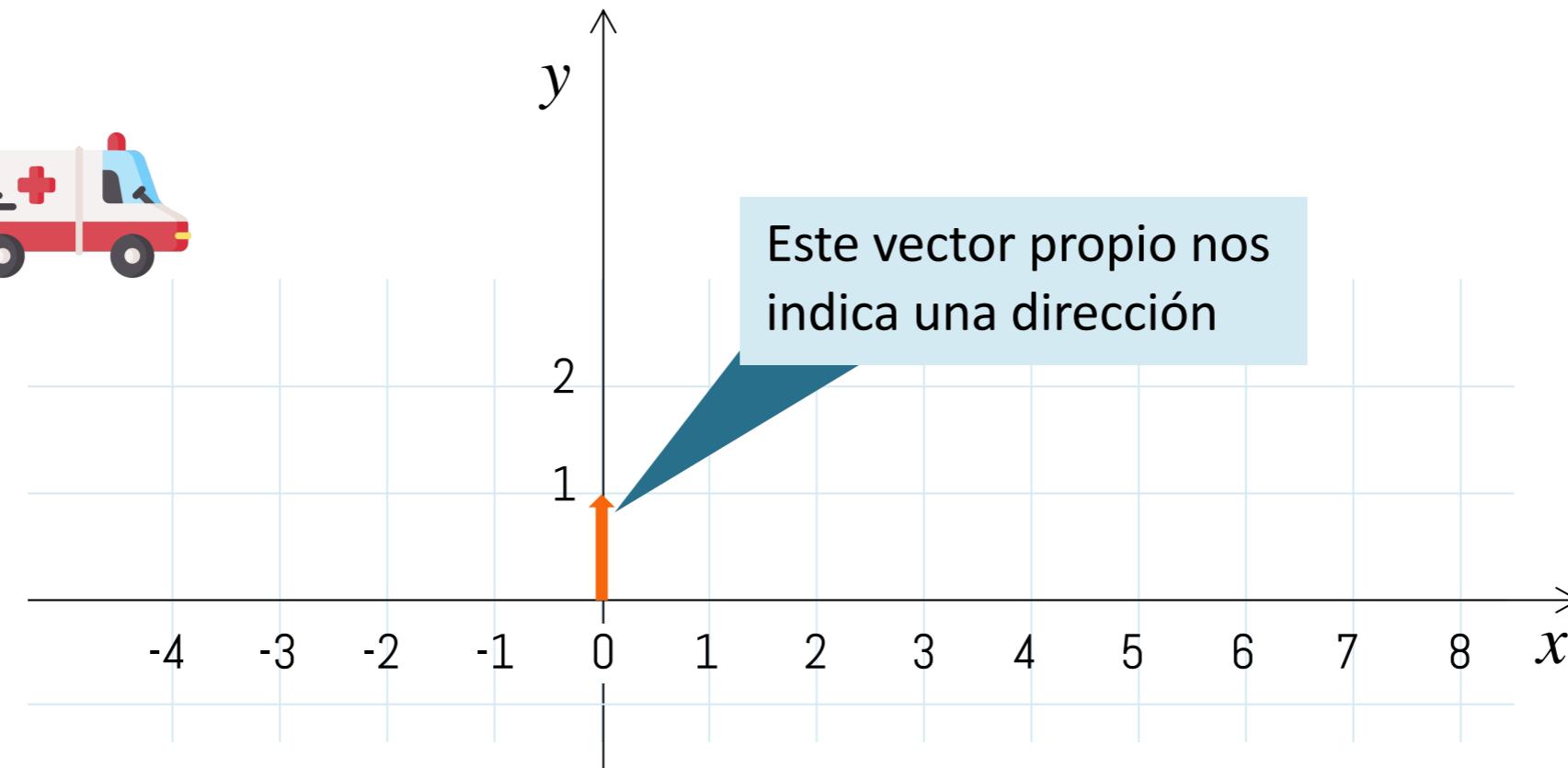
¿Qué vector $\begin{bmatrix} a \\ b \end{bmatrix}$ existe para que se cumpla dicha igualdad?

Probemos con el vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

El vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ es un **vector propio** y 8 es su **valor propio**

indica la magnitud

Vectores propios al rescate



$$\begin{bmatrix} 8 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \end{bmatrix}$$

Vectores y valores propios

$$\begin{bmatrix} 8 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} = 0.5 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

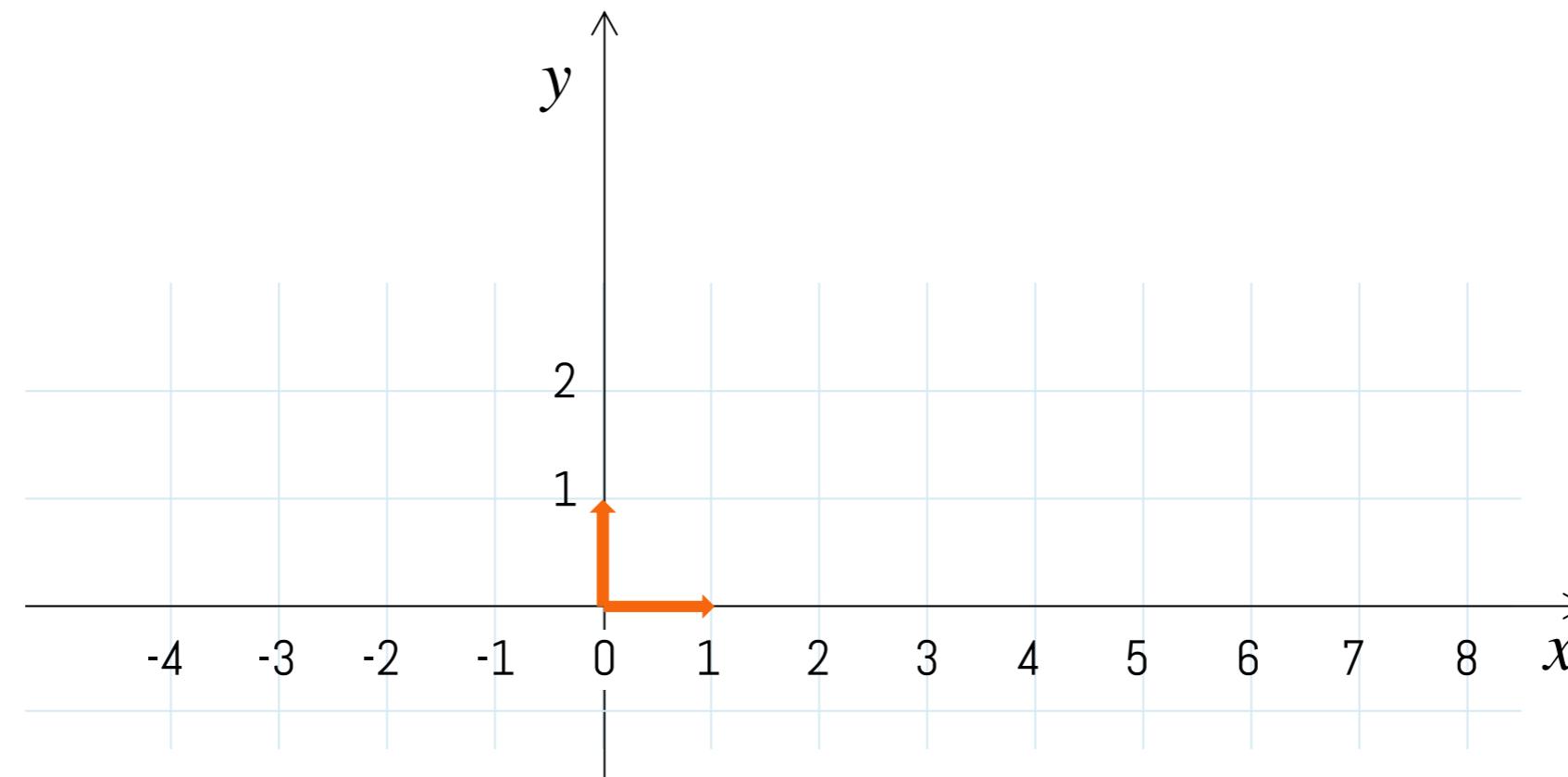


¿Qué vector $\begin{bmatrix} a \\ b \end{bmatrix}$ existe para que se cumpla dicha igualdad?

Probemos con el vector $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

El vector $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ es un **vector propio** y 0.5 es su **valor propio**

indica la magnitud

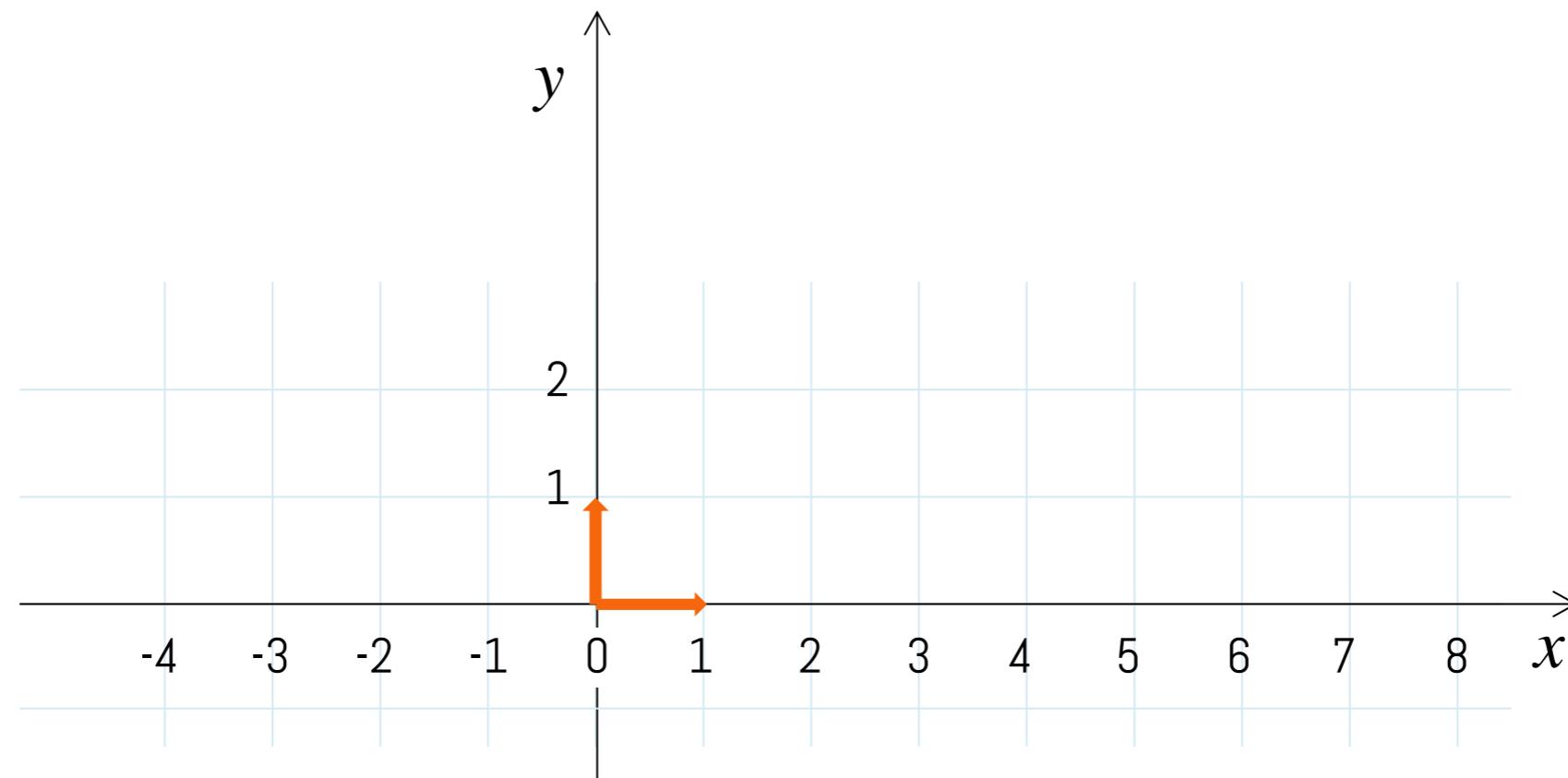


¿Qué información nos entregan los valores propios?

Valores propios

$[8 \quad 0.5]$

Nos entrega la cantidad de variación explicada por cada dirección. Pero debemos normalizarla para que sume 1



¿Qué información nos entregan los valores propios?

Valores propios

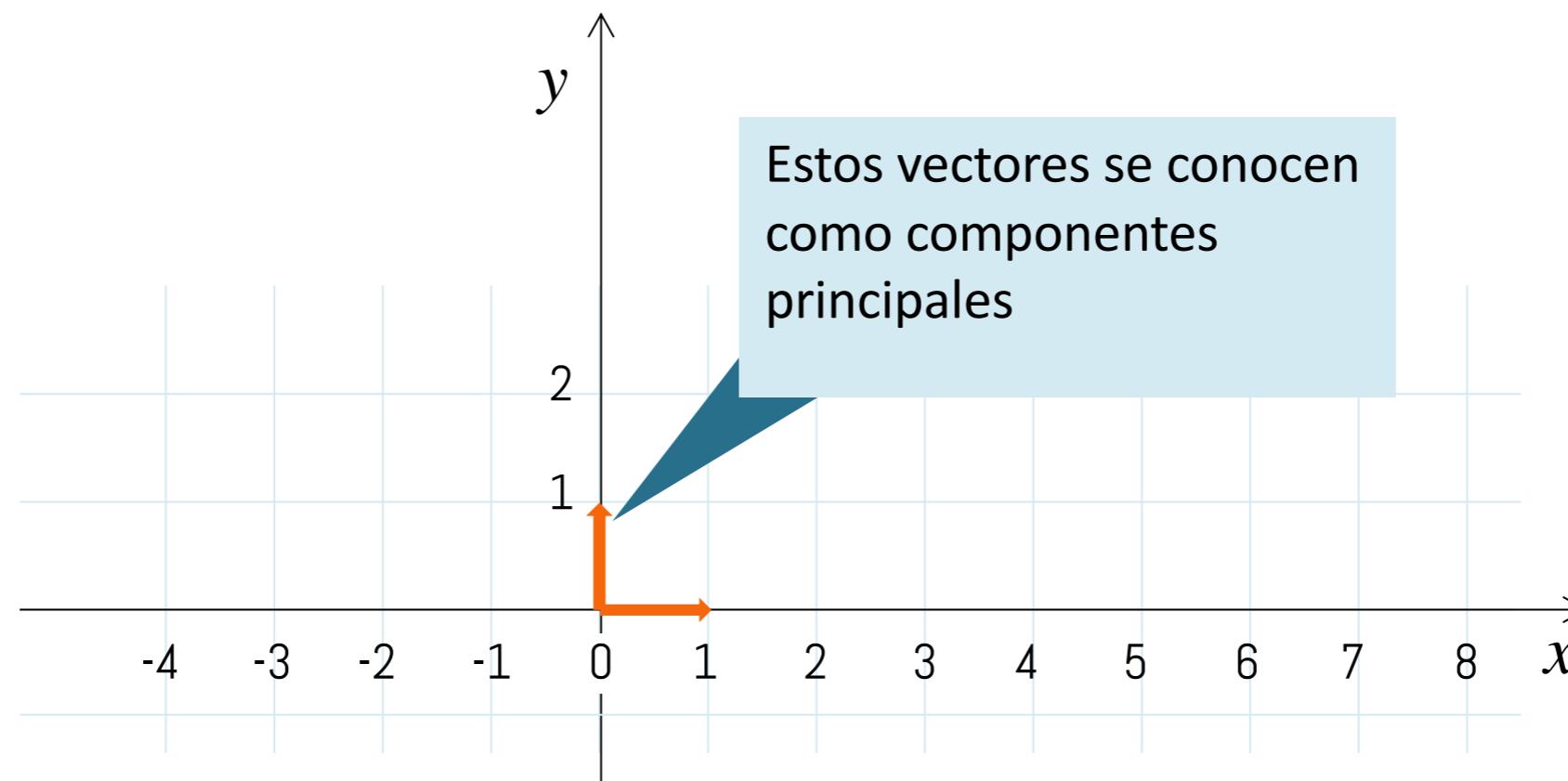
$[8 \quad 0.5]$



El primer vector propio explica un 94% de la varianza



$$\left[\frac{8}{8+0.5}, \frac{0.5}{8+0.5} \right]$$



¿Qué información nos entregan los vectores propios?

Vectores propios

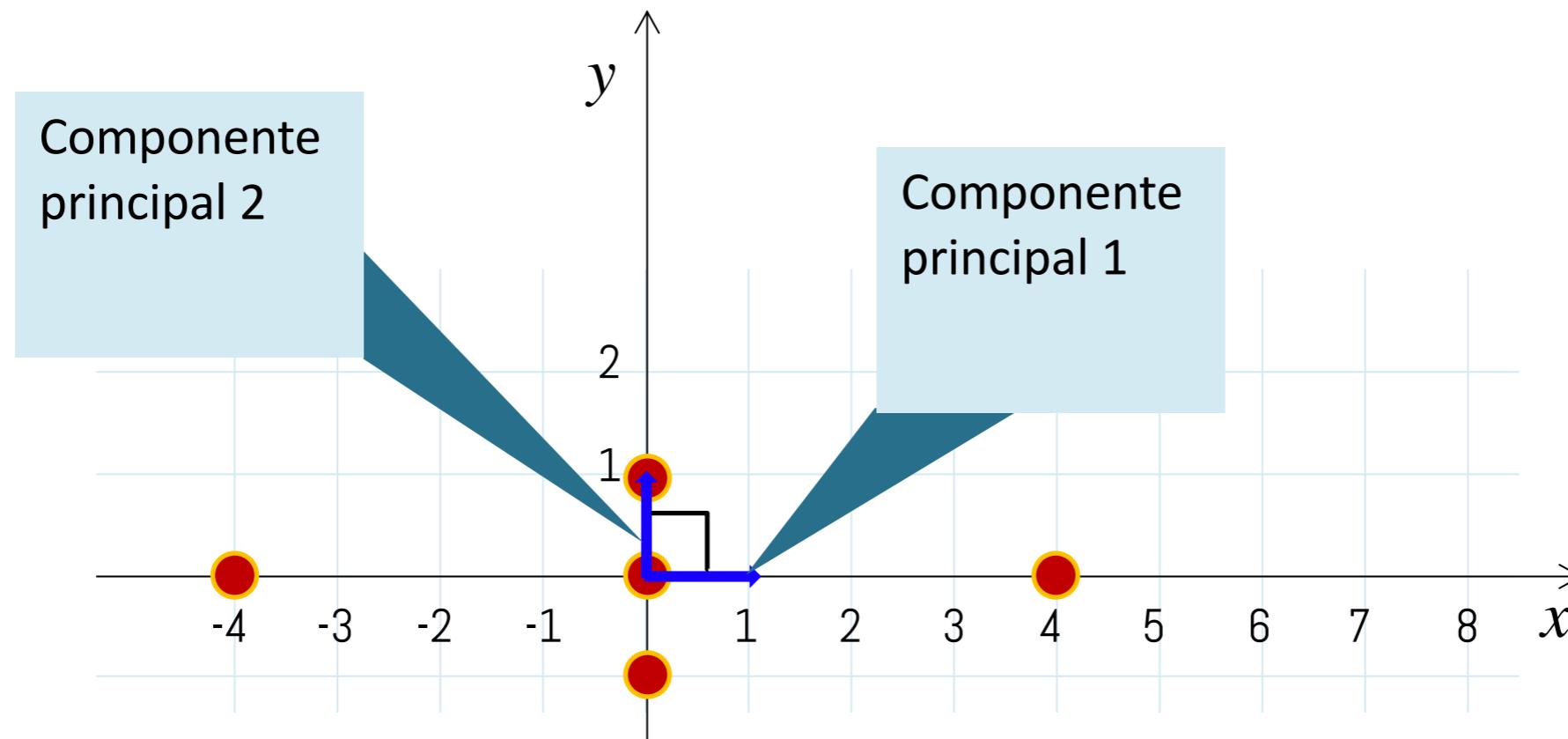
$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Componente Principal 1

Componente Principal 2

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Cada vector propio nos entrega la dirección donde existe mayor varianza.



¿Qué información nos entregan los vectores propios?

Vectores propios

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Componente Principal 1

Componente Principal 2

Cada vector propio nos entrega la dirección donde existe mayor varianza.

Los vectores son ortogonales (es decir, el producto escalar entre los vectores es cero)

$$\begin{bmatrix} 1 & 0 \end{bmatrix} * \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0$$



¿Cómo puedo generar una representación en una menor dimensión?

$$\begin{array}{c} \text{variables} \\ \text{datos} \end{array} \left\{ \begin{array}{|c|c|} \hline x & y \\ \hline -4 & 0 \\ \hline 0 & 0 \\ \hline 4 & 0 \\ \hline 0 & 1 \\ \hline 0 & -1 \\ \hline \end{array} \right. \times \begin{array}{c} pc_1 \\ 2 \times 1 \end{array} = \begin{array}{c} KLT \\ 4 \times 1 \end{array}$$

Esta matriz se conoce como transformación Karhunen-Loeve



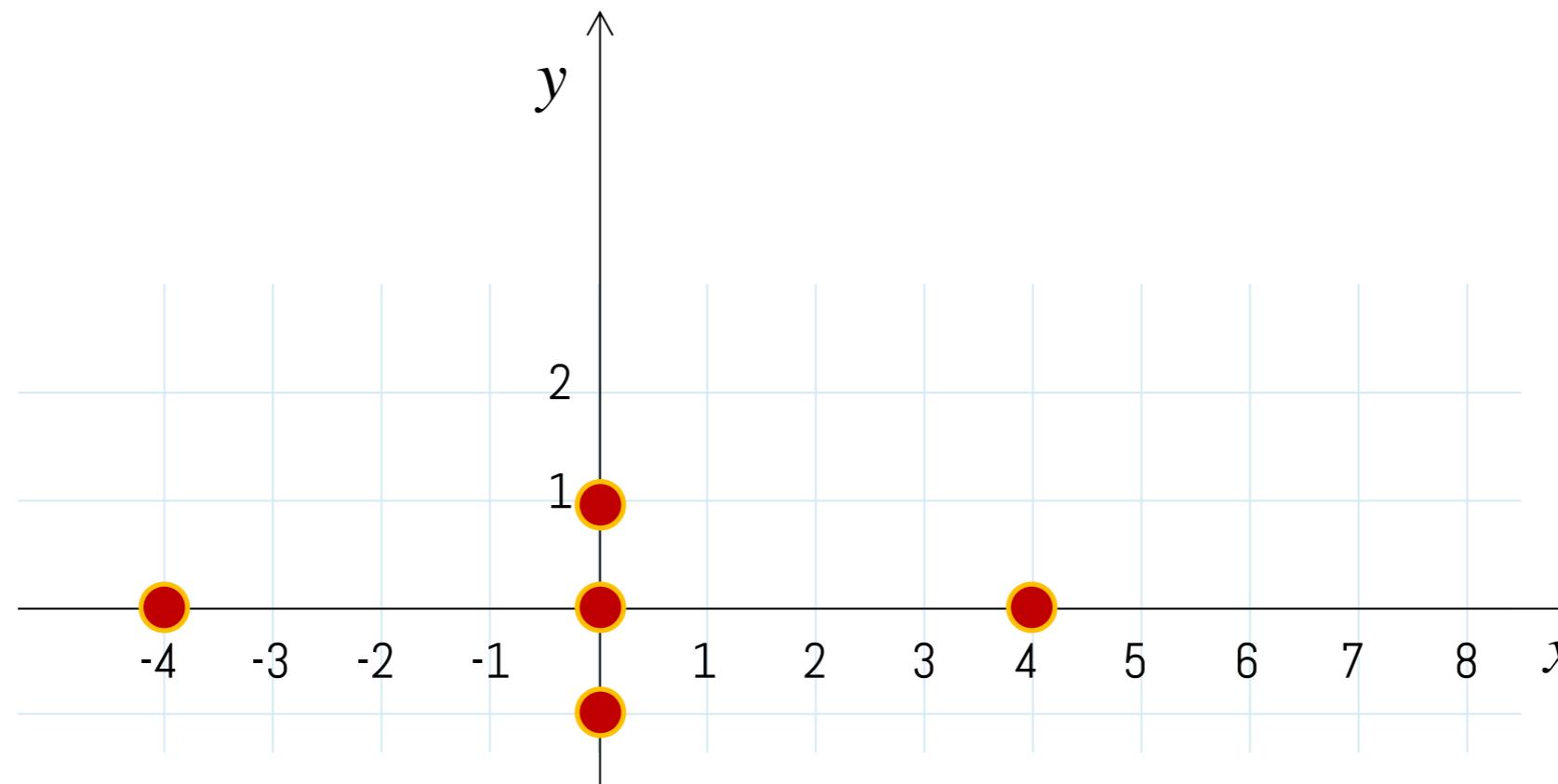
¿Cómo puedo generar una representación en una menor dimensión?



Análisis de componentes principales (PCA)

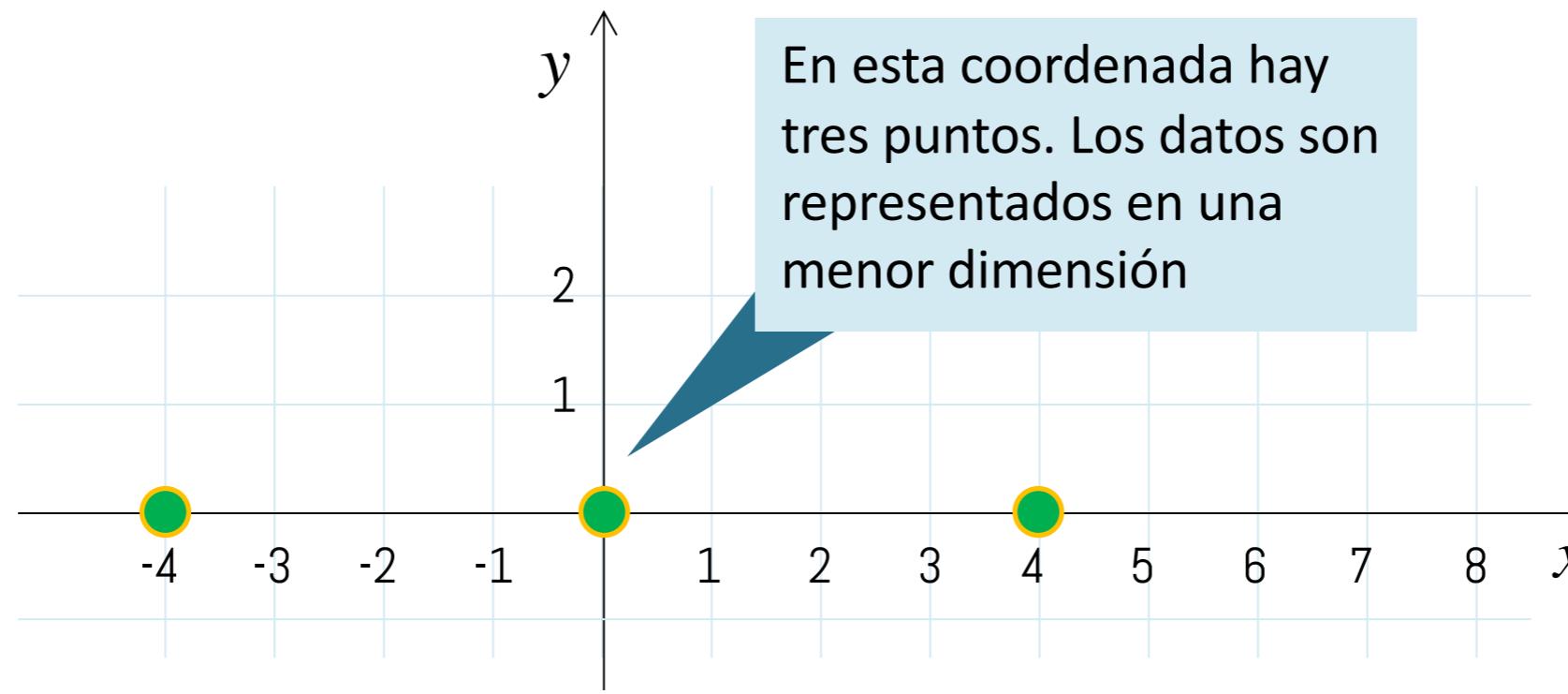
valores
originales

x	y
-4	0
0	0
4	0
0	1
0	-1



datos con
pérdida

x'	y'
-4	0
0	0
4	0
0	0
0	0





Luego del análisis, ¿quedan más claro los conceptos?

Objetivo:

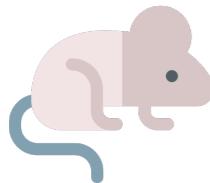
Representar los datos en una menor dimensión

¿Qué realiza?

En un procedimiento matemático que **permite identificar patrones en los datos y reducir el número de dimensiones con baja pérdida de información**. Esto lo realiza a través de transformación lineal a otro espacio empleando los vectores de máxima varianza.

¿Cómo funciona?

Transforma las variables correlacionadas a un limitado número de variables no correlacionadas a través de los vectores y valores propios.



Utilice los siguientes datos para estudiar otro ejemplo

variables

x	y
2.5	24
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

datos





PRIMERO: Reste la media de los datos.

1

$$B = X - \mu$$

- Recuerde calcular la media por cada característica, o columna.

$$\begin{array}{c}
 \left. \begin{array}{|c|c|} \hline x & y \\ \hline 2.5 & 24 \\ 0.5 & 0.7 \\ 2.2 & 2.9 \\ 1.9 & 2.2 \\ 3.1 & 3.0 \\ 2.3 & 2.7 \\ 2 & 1.6 \\ 1 & 1.1 \\ 1.5 & 1.6 \\ 1.1 & 0.9 \\ \hline \end{array} \right\} X - \left. \begin{array}{|c|c|} \hline \mu & \\ \hline 1.81 & 1.91 \\ \hline \end{array} \right\} = \left. \begin{array}{|c|c|} \hline B = X - \mu & \\ \hline 0.69 & 0.49 \\ -1.31 & -1.21 \\ .039 & 0.99 \\ 0.09 & 0.29 \\ 1.29 & 1.09 \\ 0.49 & 0.79 \\ 0.19 & -0.31 \\ -0.81 & -0.81 \\ -0.31 & -0.31 \\ -0.71 & -1.01 \\ \hline \end{array} \right\}
 \end{array}$$



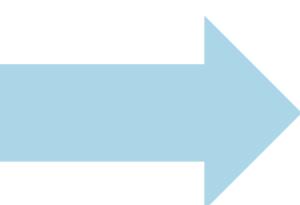
SEGUNDO: Calcule la covarianza. La covarianza es una medida de dispersión entre dos variables.

2

$$C = \frac{1}{n-1} B^T B$$

n es la cantidad de del problema

$B = X - \mu$	
0.69	0.49
-1.31	-1.21
.039	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.71	-1.01



C	
0.6166	0.6154
0.6154	0.7166



TERCERO: Calcule la valores y vectores propios (Eigenvalues, Eigenvectors) de la matriz resultante C

3

$$V^{-1} \cdot C \cdot V = D$$

- En Python:

```
import numpy as np
from numpy import linalg as LA

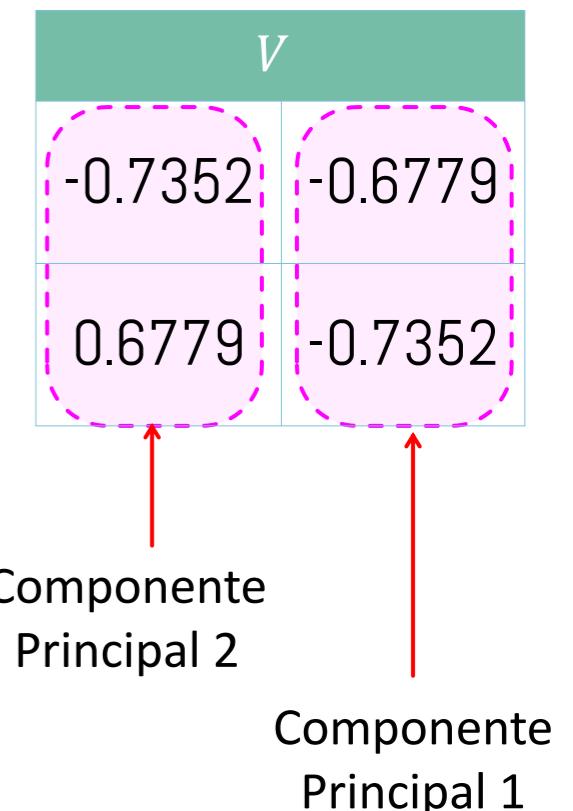
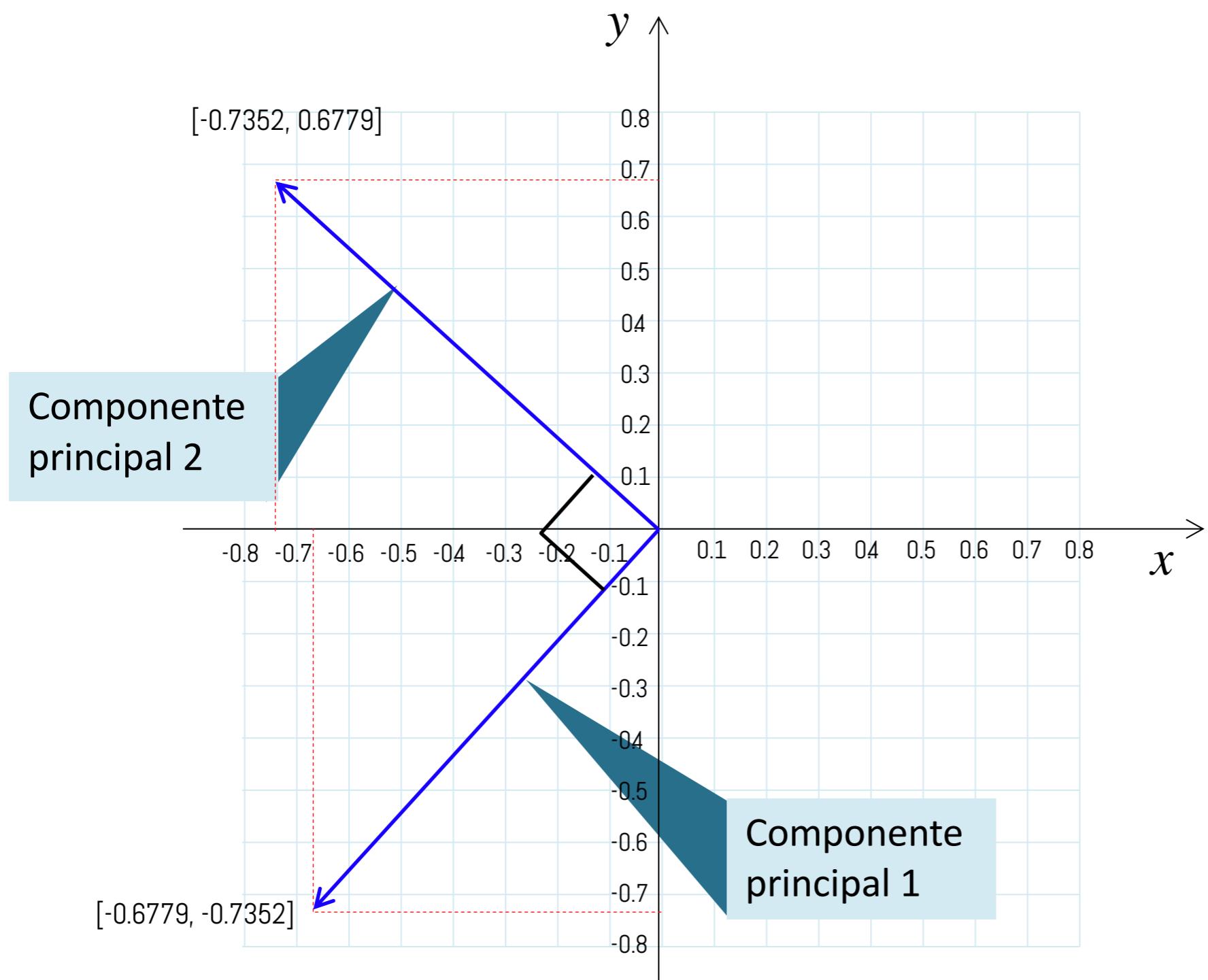
co= np.cov(x.T)
w, v = LA.eig(co)
```

donde (V) corresponde a la matriz de vectores propios y la matriz D corresponde a los valores propios.

<i>C</i>		<i>V</i>		<i>D</i>	
0.6166	0.6154	-0.7352	-0.6779	0.0491	0
0.6154	0.7166	0.6779	-0.7352	0	1.2840
→					



Como puede observar, los componentes principales son ortogonales





CUARTO: Ordene los valores propios en orden decreciente

4

Ordenar los valores propios en orden decreciente

D	
0.0491	0
0	1.2840

$$\begin{aligned} D &= [1.284, 0.0491] \\ \text{Ord} &= [2, 1] \end{aligned}$$

- Luego calcule la información acumulada de cada valor propio

El largo de este vector depende del número de columnas

$$I_V(l) = \frac{\sum_{i=1}^L D(i)}{\sum_{i=1}^n D(i)} \quad \forall i \in \{1, \dots, n\}$$

$$Iv = [0.963, 1]$$



QUINTO: Seleccione un mínimo valor de columnas de vectores propios de tal forma que la energía sea máxima.

5

Seleccione las columnas con mayor varianza

- Para nuestro ejemplo, seleccionar una cantidad menor o igual a 98% de energía implica que sólo es necesario ocupar la columna con mayor valor propio.

$$\begin{aligned} D &= [1.284, 0.0491] \\ Iv &= [96.3\%, 100\%]; \\ Ord &= [2, 1] \end{aligned}$$

El 96.3% de la varianza se logra con la columna 2. Para lograr un 100% ocupamos las dos columnas.

V	
COL 1	COL 2
-0.7352	-0.6779
0.6779	-0.7352

W
-0.6779
-0.7352

Este paso es clave





SEXTO: La transformación Karhunen-Loeve (PCA) se obtiene a partir de una combinación lineal correspondiente a:

6

$$KLT = B \cdot W$$

<i>B</i>	
0.69	0.49
-1.31	-1.21
.039	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.71	-1.01

×

<i>W</i>	
-0.6779	
	-0.7352

=

KLT	
-0.8280	
	1.7776
-0.9922	
	-0.2742
-1.6758	
	-0.9129
0.0991	
	1.1446
0.4380	
	1.2238

Ésta es la
compresión
PCA



Si quiero recuperar la información original (con pérdida) simplemente utilizamos la siguiente ecuación

7

$$P = KLT \cdot W^T + \mu$$

 P

Corresponde a la
media de los
datos originales

x	y
2.5	24
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

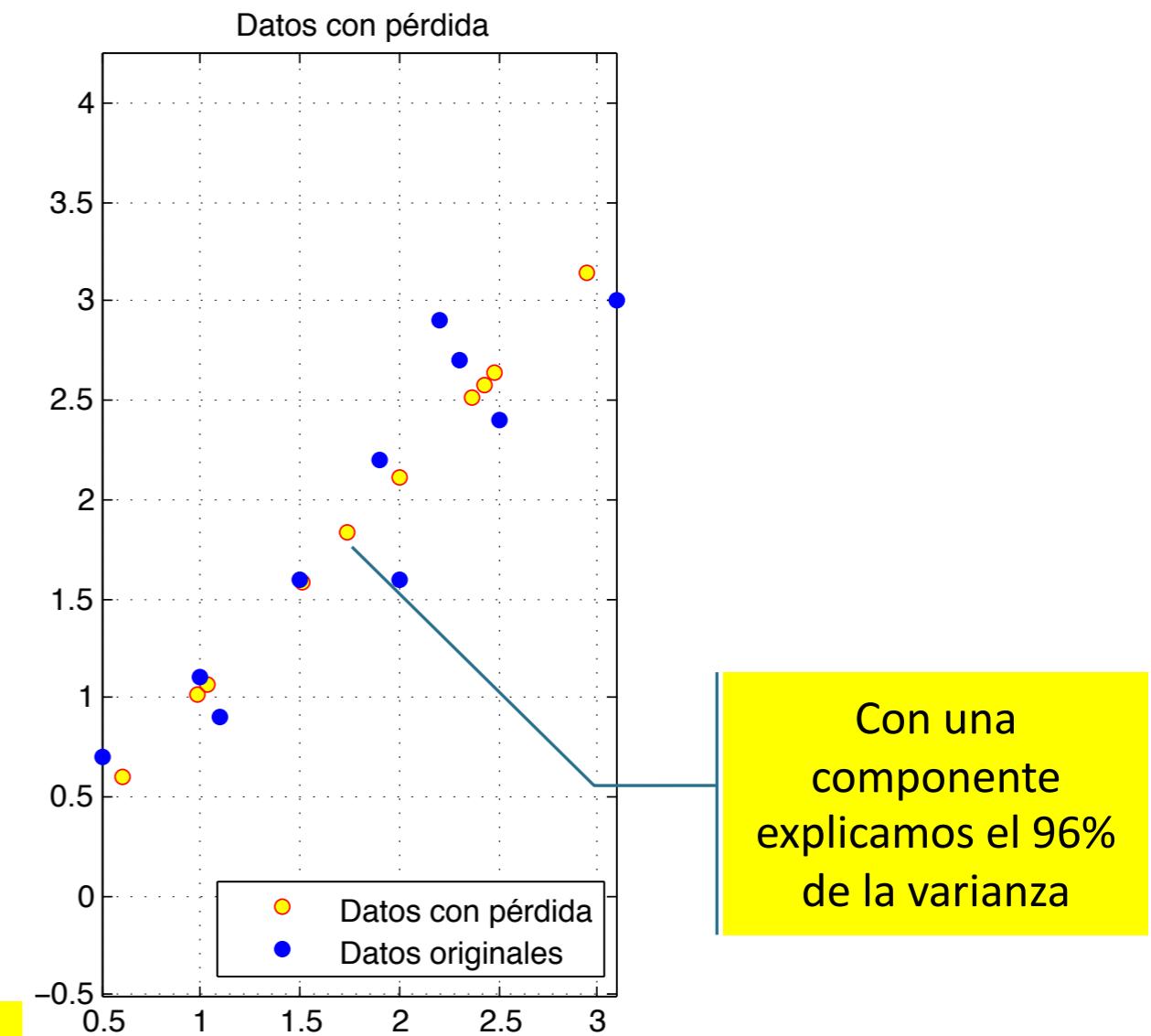
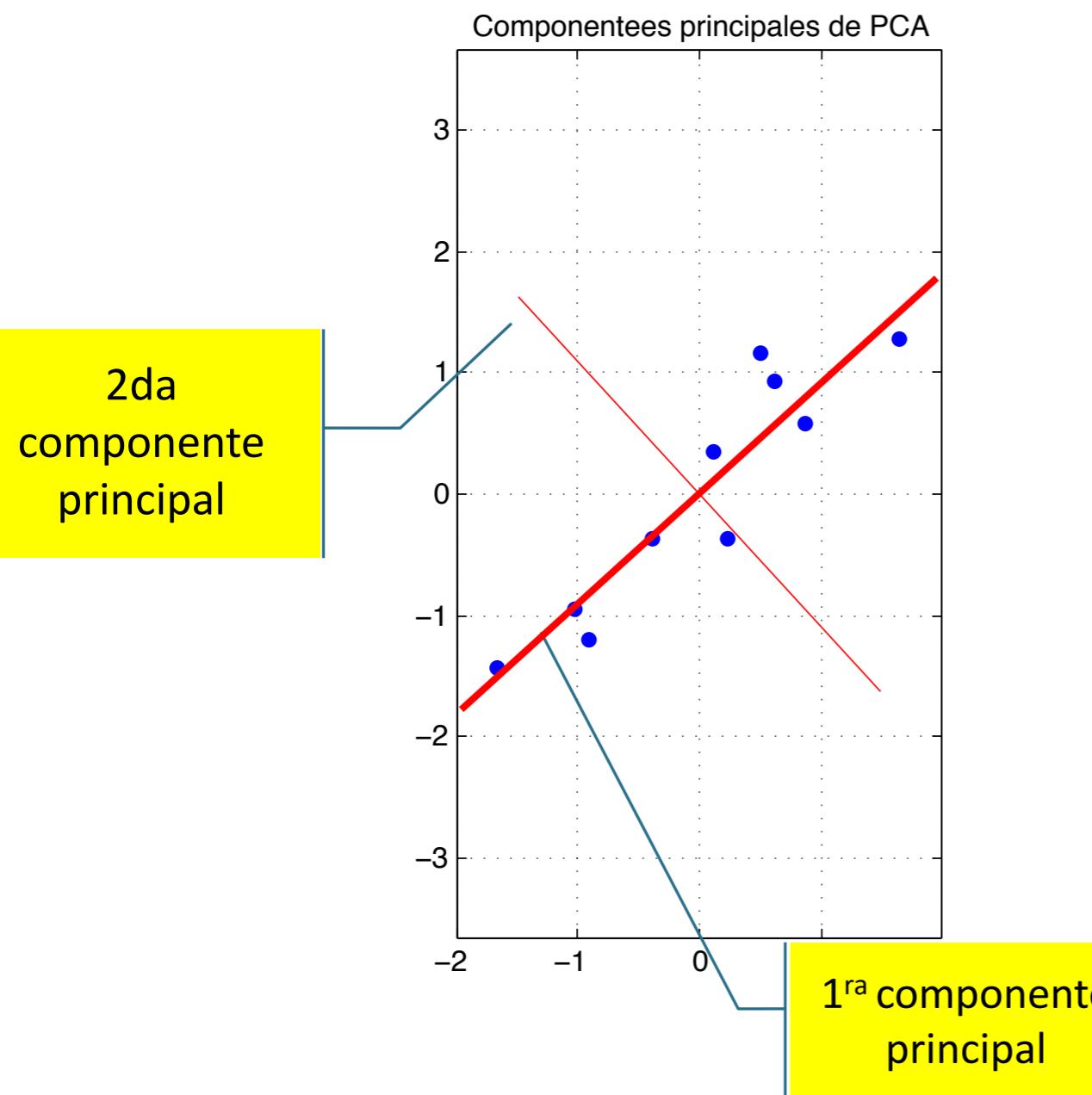
datos originales

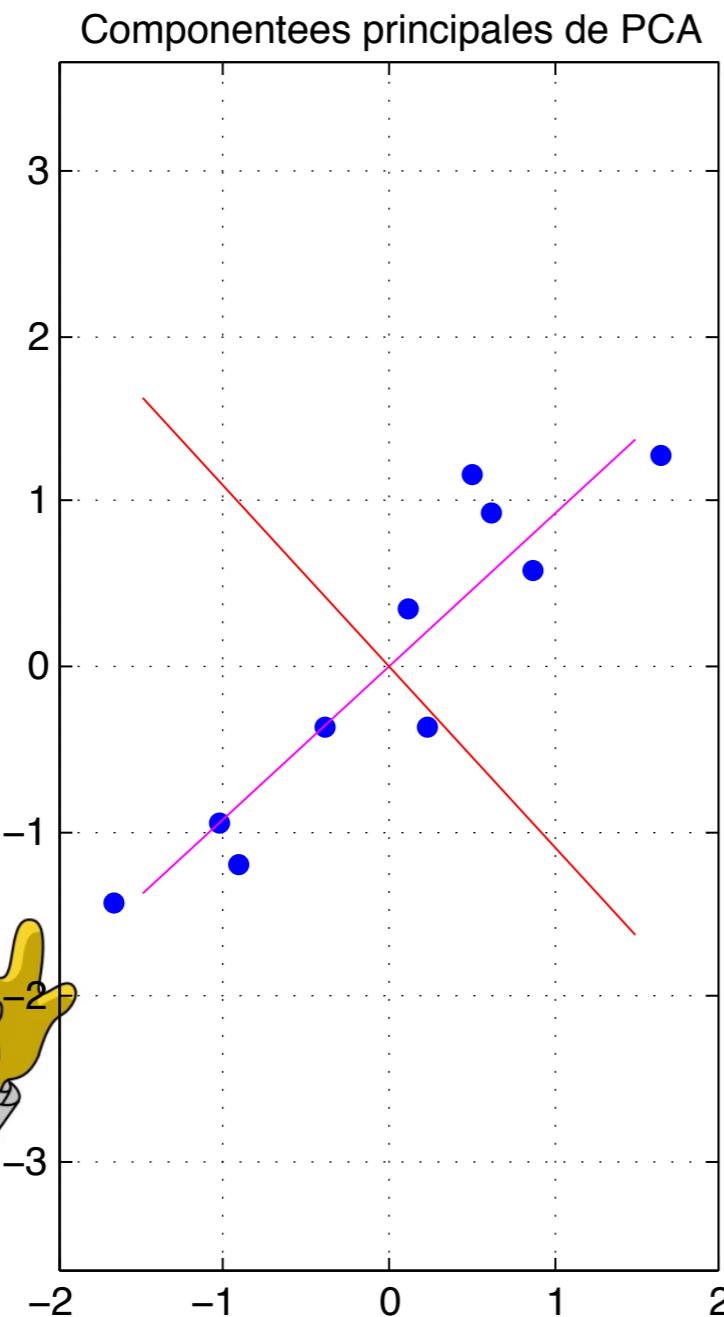
datos con pérdida

x'	y'
24	2.5
0.6	0.6
2.5	2.6
2.0	2.1
2.9	3.1
24	2.6
1.7	1.8
1.0	1.0
1.5	1.6
1.0	1.0



Veamos una representación bidimensional de los datos y de las componentes principales.





- La componente principal conserva la varianza de mayor tamaño. La segunda es ortogonal a la primera y así sucesivamente
- PCA reduce la dimensión de los datos reteniendo aquellas características que más contribuyan a maximizar la varianza.

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

KLT
-0.8280
1.7776
-0.9922
-0.2742
-1.6758
-0.9129
0.0991
1.1446
0.4380
1.2238

■ Análisis de Componentes Principales (PCA)

- El mismo algoritmo lo podemos aplicar a imágenes. Sólo que ahora en vez de comprimir datos, reducimos la calidad de la imagen.
- Observe como la compresión de información genera un degradamiento en la calidad de la imagen.

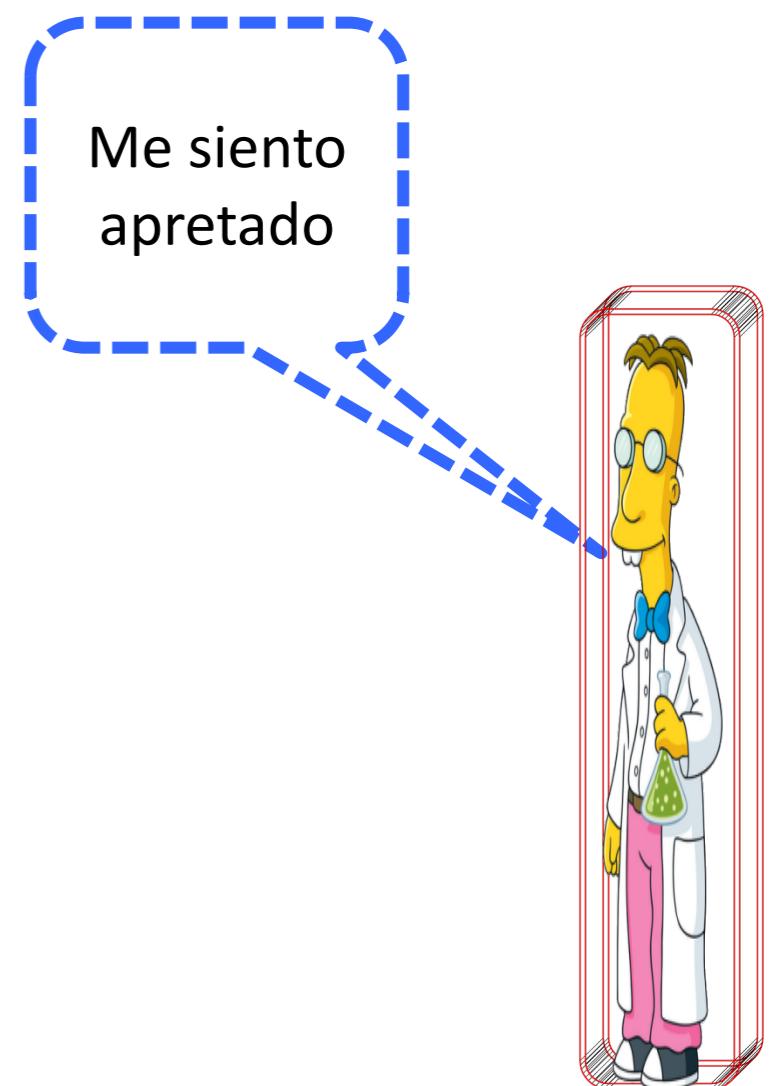


■ Compresión

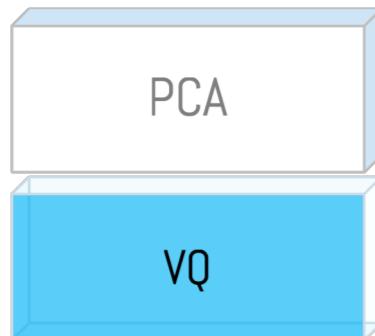
- Análisis de Componentes Principales (PCA)
- Vector Quantisation

■ Selección

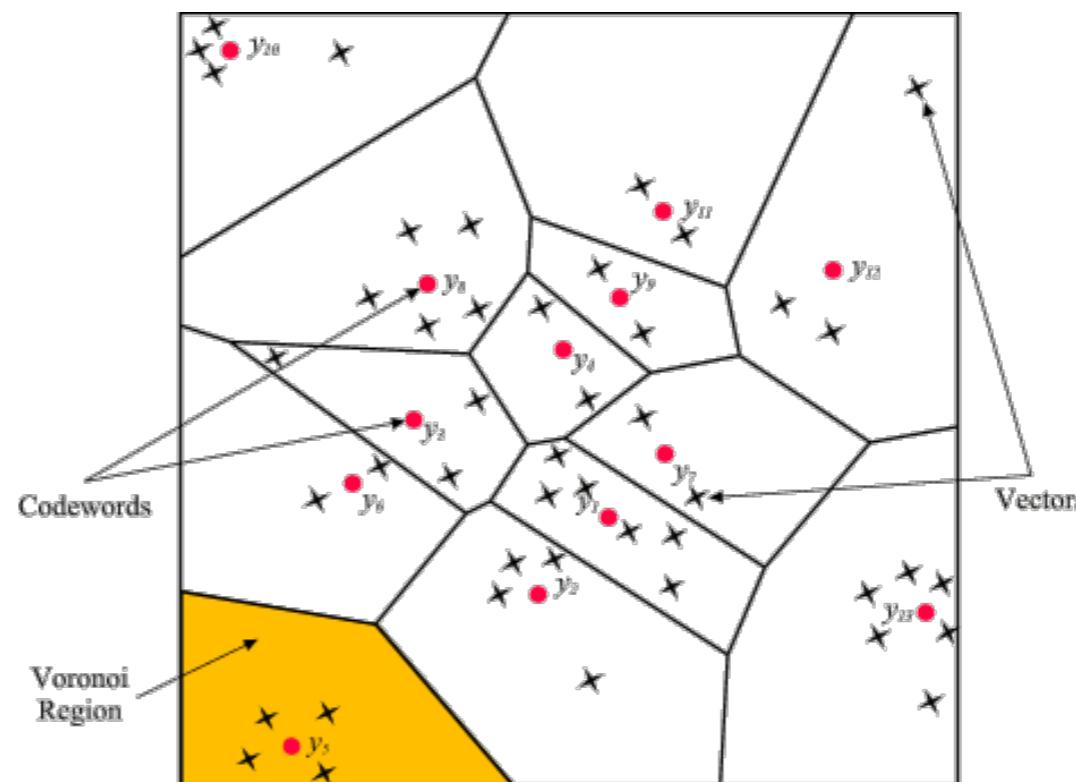
- Discriminante de Fisher
- Búsqueda exhaustiva y secuencial
- “Plus L-take away R”
- Branch & Bound



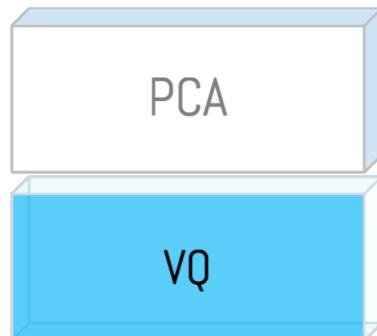
■ Vector Quantization (VQ)



- **Objetivo:**
Representar los datos en una menor dimensión. Fue propuesto por Linde, Buzo, y Gray (LBG) en 1980. El algoritmo VQ también es conocido LBG-VQ.
- **¿Qué realiza?**
En un procedimiento matemático iterativo que modela el espacio buscando puntos representativos (conocido como *codewords*)



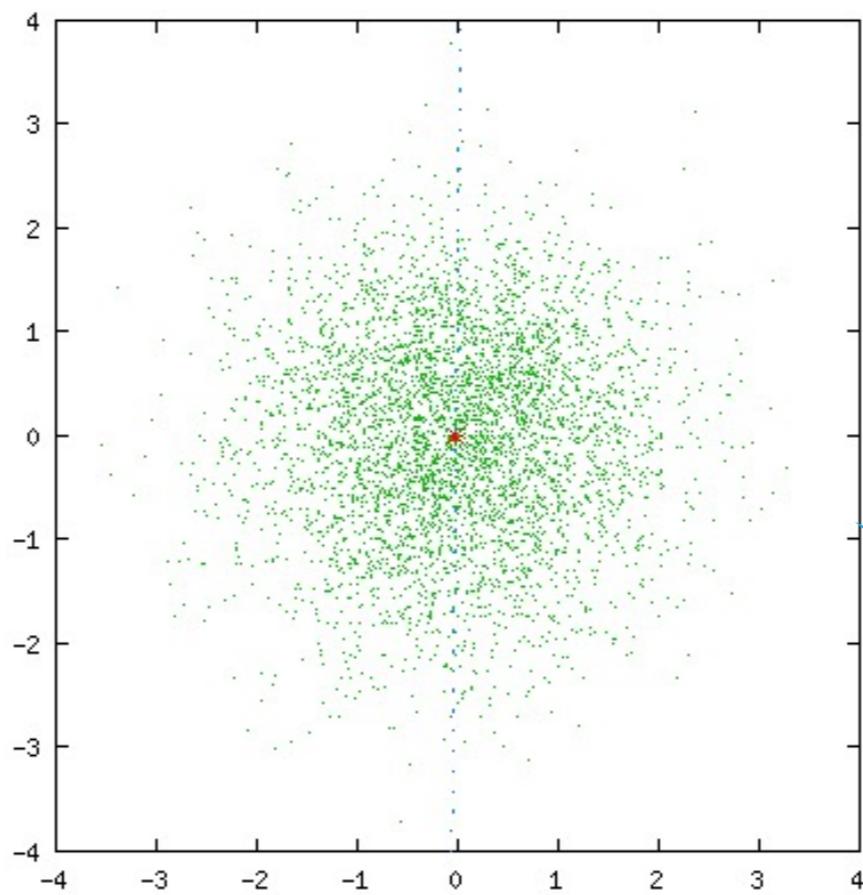
■ Vector Quantization (VQ)



■ *Cómo funciona*

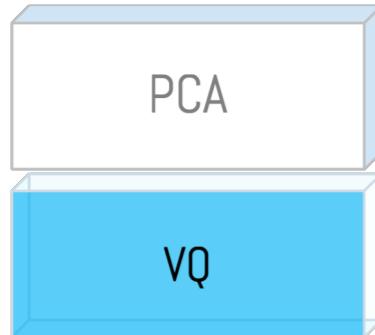
El algoritmo va dividiendo el espacio de puntos en grupos, manteniendo aproximadamente el mismo número de puntos cercanos a cada grupo.

Al finalizar, cada grupo es representado por un centroide.



Vea la
animación

■ Vector Quantization (VQ)

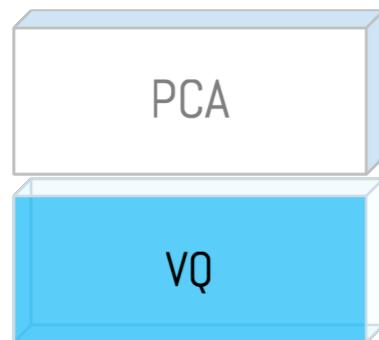


- *PRIMERO:* Calcule la media de los datos.
- Considere para el ejemplo que los datos están expuestos en columnas.

X	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

m	1.81
	1.91

■ Vector Quantization (VQ)



- **SEGUNDO:** Genere dos centros a partir de la media del paso anterior de la siguiente forma.

m	1.81
	1.91

$$\mathbf{c}_1 = (1 + \varepsilon) \cdot \mathbf{m}$$
$$\mathbf{c}_2 = (1 - \varepsilon) \cdot \mathbf{m}$$

$$\varepsilon = 0.01$$

c ₁	(1+0.01)·1.81
	(1-0.01)·1.91

An arrow points from the C₁ table to the C₁ result table.

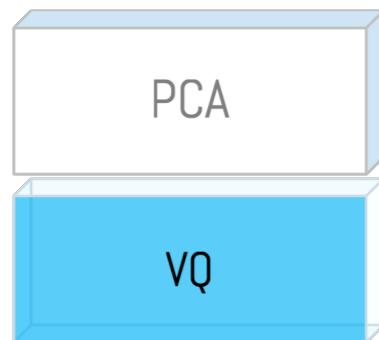
c ₁	1.8281
	1.9291

c ₂	(1-0.01)·1.81
	(1-0.01)·1.91

An arrow points from the C₂ table to the C₂ result table.

c ₂	1.8098
	1.9098

■ Vector Quantization (VQ)



- **TERCERO:** Por cada centroide, calculamos la distancia euclídea respecto a los datos originales.

C ₁	1.8281	C ₂	1.8098
	1.9291		1.9098

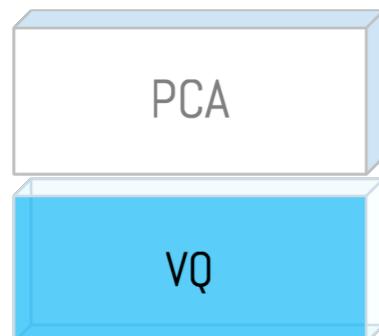
X	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

- **Ejemplo**

$$\sqrt{(1.8281 - 2.5)^2 + (1.9291 - 2.4)^2} = 0.82$$

d ₁	0.82	1.80	1.03	0.28	1.66	0.90	0.37	1.17	0.46	1.26
d ₂	0.84	1.78	1.06	0.30	1.68	0.92	0.36	1.14	0.43	1.23

■ Vector Quantization (VQ)



- **CUARTO:** Determinamos los valores mínimos de cada columna y determinamos a qué fila pertenecen cada mínimo.

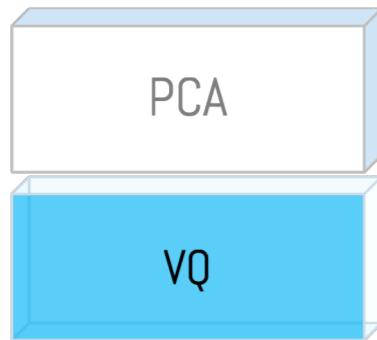
d_1	0.82	1.80	1.03	0.28	1.66	0.90	0.37	1.17	0.46	1.26
d_2	0.84	1.78	1.06	0.30	1.68	0.92	0.36	1.14	0.43	1.23

tmp	0.82	1.78	1.03	0.28	1.66	0.90	0.36	1.14	0.43	1.23
-----	------	------	------	------	------	------	------	------	------	------

Corresponde a la fila donde se encuentra el mínimo

c	1	2	1	1	1	1	2	2	2	2
---	---	---	---	---	---	---	---	---	---	---

■ Vector Quantization (VQ)



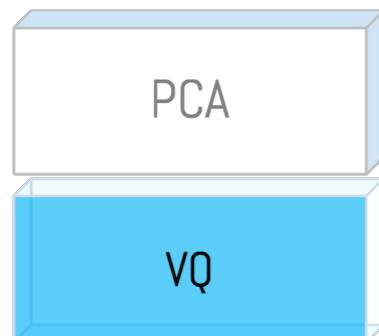
- **QUINTO:** Según la clase de cada mínimo, calculamos los nuevos centroides promediando los valores según cada grupo.
- Los pasos 3ro al 5to los realizamos número fijo de veces (5 o más), para mejorar la estabilidad del centroide.

c	1	2	1	1	1	1	2	2	2	2
---	---	---	---	---	---	---	---	---	---	---

X	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

C ₁	2.400	C ₂	1.22
	2.64		1.18

■ Vector Quantization (VQ)



- **SEXTO:** Duplicamos el número de centroides mientras alcancemos el número de centroides que buscamos. Para ello empleamos la siguiente ecuación (La variable n se duplica cada vez). El proceso comienza nuevamente desde el tercer paso

M es la cantidad de filas o características del problema

$$\mathbf{c}_i = (1 + \varepsilon) \cdot \mathbf{c}_i$$

$$\mathbf{c}_{i+n} = (1 - \varepsilon) \cdot \mathbf{c}_i$$

$$\varepsilon = 0.01$$

$$n = 2$$

$i = 1, 2$	
c_1	2.400
c_2	1.22

c_1	2.64
c_2	1.18

$i=1$

c_1	2.4240
	2.664

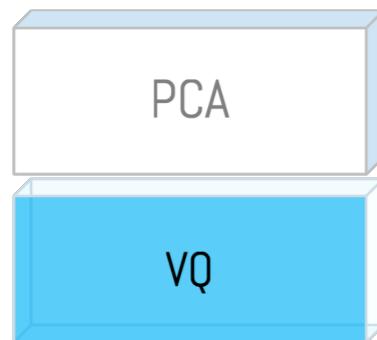
c_3	2.3998
	2.6397

$i=2$

c_2	1.2322
	1.1928

c_4	1.2199
	1.1799

■ Vector Quantization (VQ)



- Veamos que sucede con cuatro centroides. Determinamos la distancia euclidiana (3er paso) y luego determinamos la fila que pertenece cada mínimo

C_1	2.4240	C_2	1.2322	C_3	2.3998	C_4	1.2199
	2.664		1.1928		2.6397		1.1799

Tercer
paso

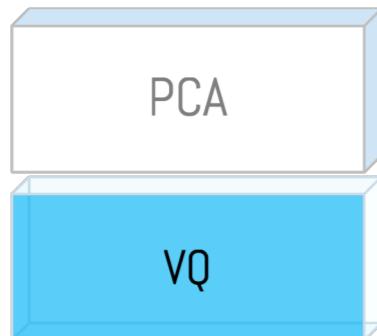
X	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

Cuarto
paso

d_1	0.27	2.75	0.32	0.70	0.75	0.12	1.14	2.11	1.41	2.20
d_2	1.75	0.88	1.96	1.20	2.59	1.84	0.86	0.24	0.48	0.32
d_3	0.25	2.71	0.32	0.66	0.78	0.11	1.11	2.08	1.37	2.17
d_4	1.76	0.86	1.97	1.22	2.61	1.86	0.88	0.23	0.50	0.30

c	3	4	1	3	1	3	2	4	2	4
-----	---	---	---	---	---	---	---	---	---	---

■ Vector Quantization (VQ)



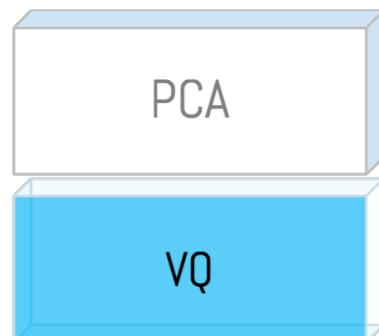
- Veamos que sucede con cuatro centroides. Determinamos la distancia euclidiana (Tercer paso) y luego determinamos la fila que pertenece cada mínimo

Quinto paso

A diagram showing the fifth step of VQ. A red curly brace on the left groups three tables: one for centroids (C), one for data points (X), and one for their distances. The text "Quinto paso" is placed to the left of the first table.

C	3	4	1	3	1	3	2	4	2	4
X	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
X	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9
c_1	2.65		c_2	1.75		c_3	2.33		c_4	0.86
	2.95			1.60			2.43			0.90

■ Vector Quantization (VQ)



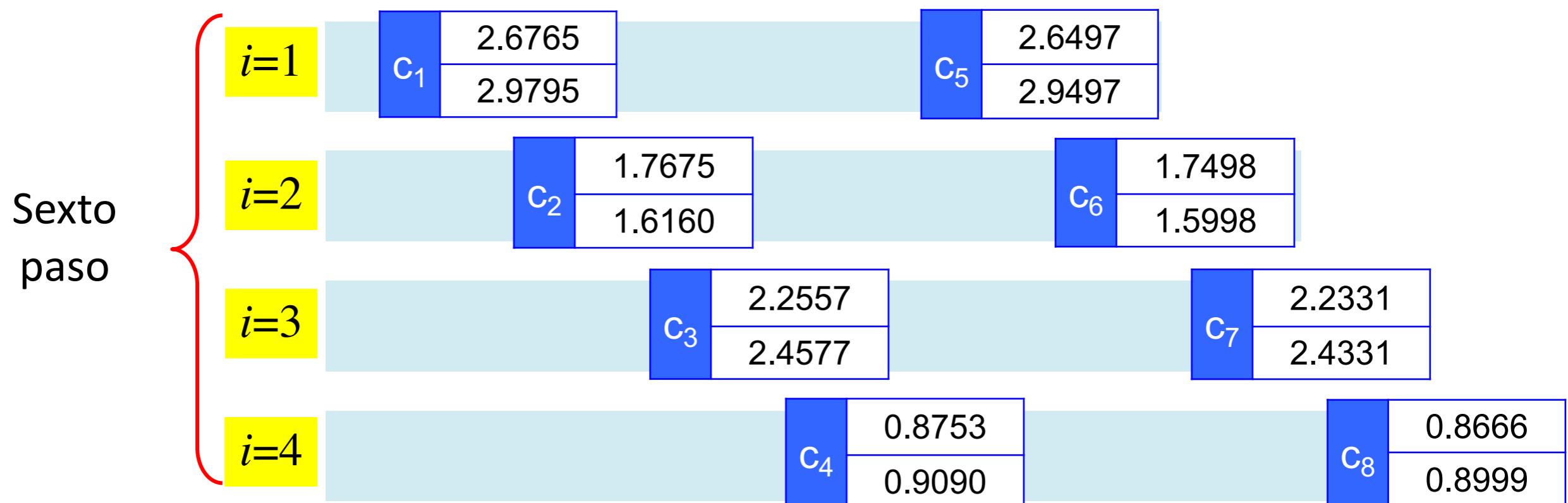
$$\begin{aligned}\varepsilon &= 0.01 \\ n &= 4\end{aligned}$$

$$\mathbf{c}_i = (1 + \varepsilon) \cdot \mathbf{c}_i$$

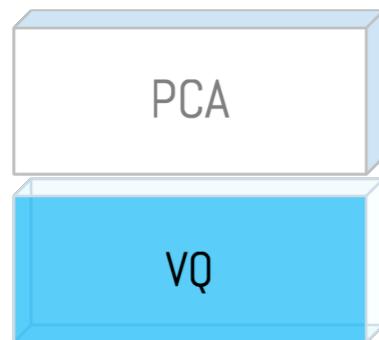
$$\mathbf{c}_{i+n} = (1 - \varepsilon) \cdot \mathbf{c}_i$$

- Veamos que sucede con cuatro centroides. Determinamos la distancia euclídea (Tercer paso) y luego determinamos la fila que pertenece cada mínimo

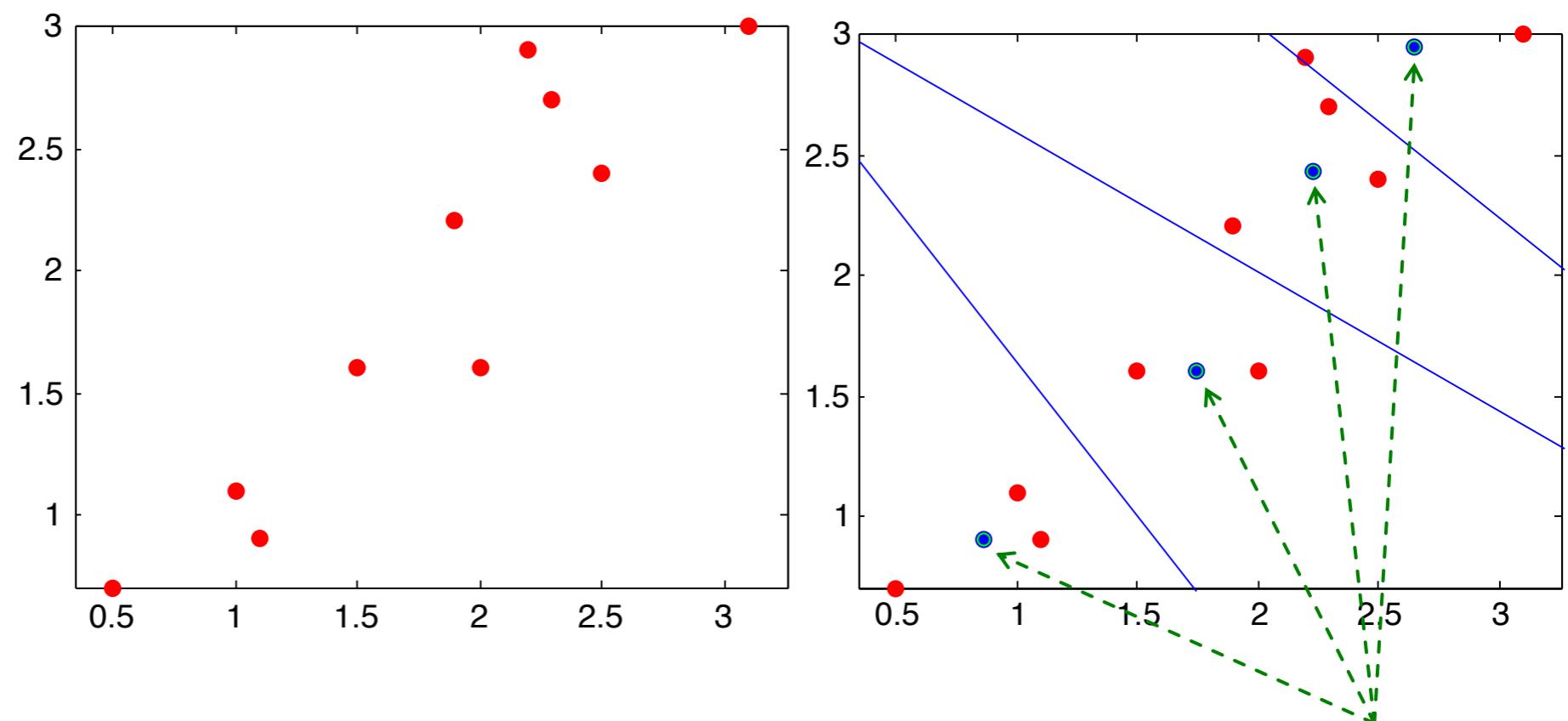
	$i = 1, \dots, 4$	$n = 4$
\mathbf{c}_1	2.65	
\mathbf{c}_2	1.75	
\mathbf{c}_3	2.33	
\mathbf{c}_4	0.86	
\mathbf{c}_1	2.95	
\mathbf{c}_2	1.60	
\mathbf{c}_3	2.43	
\mathbf{c}_4	0.90	



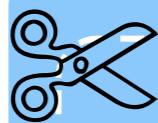
■ Vector Quantization (VQ)



- Si analizamos los datos, y empleando sólo 4 *codebooks*, generamos el siguiente gráfico
- Las líneas fueron generadas a partir de un diagrama de Voronoi



Centroide
generado con el
algoritmo VQ.



Anexo

Vectores y valores propios



■ Vectores y valores propios.

- La idea matemática consiste en contrar un vector v y un escalar λ que cumpla la siguiente condición

$$A \cdot v = \lambda \cdot v$$

- donde v son **vectores propios** (no nulos) de la matriz cuadrada A y λ son los **valores propios** (correspondiente a cada vector propio).
- El vector propio corresponde a una transformación lineal de los datos originales y el valor propio es el factor de escala por el cual ha sido multiplicado. Podemos visualizar al vector propio como una dirección desde el origen.
- Para determinar si un problema tiene un vector y valor propio, determinamos la siguiente ecuación:

$$(A - \lambda I) \cdot v = 0$$

■ Vectores y valores propios.

- Observamos que el sistema $(\mathbf{A} - \lambda \mathbf{I}) \cdot \mathbf{v} = 0$ corresponde a $\mathbf{B} \cdot \mathbf{v} = 0$
- Una solución del sistema anterior es cuando $v=0$ (es nulo), pero dicha solución no nos interesa. Por el contrario, buscamos determinar una solución que implique que $\det(\mathbf{B}) = 0$. Esto significa que buscamos que $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$
- El polinomio $\det(\mathbf{A} - \lambda \mathbf{I})$ se denomina **polinomio característico** de la matriz A . Dicho polinomio es de grado n , ya que la matriz A es de dimensión $n \times n$
- **Definición:** si λ es un valor propio de A y si existe un vector no nulo tal que $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$. Entonces v es un vector propio de A correspondiente al valor propio λ

- **Importante.** No todas las matrices poseen vector y valor propio

- Vectores y valores propios.

- Veamos un ejemplo. ¿Es $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ un vector propio de la matriz $\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$?

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 11 \\ 5 \end{bmatrix} \quad [\text{R}]. \text{ No es un vector propio}$$

- Un segundo ejemplo. ¿Es $\begin{bmatrix} 3 \\ 2 \end{bmatrix}$ un vector propio de la matriz $\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$?

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \cdot \begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad [\text{R}]. \text{ Sí, es un vector propio}$$

¿Cómo calculamos el vector propio? Resolviendo el sistema $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$

$$\det\left(\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0$$

$$\det\left(\begin{bmatrix} 2-\lambda & 3 \\ 2 & 1-\lambda \end{bmatrix}\right) = 0 \quad \xrightarrow{\hspace{1cm}} \quad (2-\lambda)(1-\lambda) - 6 = 0$$
$$2 - 3\lambda + \lambda^2 - 6 = 0 \quad \xrightarrow{\hspace{1cm}} \quad \lambda^2 - 3\lambda - 4 = (\lambda+1)(\lambda-4) = 0$$

Por lo tanto, los dos **valores propios** son: $\lambda_1 = -1, \lambda_2 = 4$

Vectores y valores propios.

- **Primero.** Busquemos el vector propio para el valor propio $\lambda_1 = -1$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_1 = \lambda_1 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_1$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_1 = - \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_1 \quad \Rightarrow \quad \begin{array}{l} 2v_1 + 3v_2 = -v_1 \\ 2v_1 + v_2 = -v_2 \end{array} \quad \Rightarrow \quad \begin{array}{l} 3v_1 + 3v_2 = 0 \\ 2v_1 + 2v_2 = 0 \end{array} \quad \Rightarrow \quad v_2 = -v_1$$

- El sistema tiene infinitas soluciones de la forma $\begin{bmatrix} v_1 & -v_1 \end{bmatrix}^T$. Por ejemplo si tomamos $v_1 = 1$, entonces $v_2 = -1$. Así nuestro **vector propio** queda de la siguiente forma:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}_1$$

$$\Rightarrow = -1 \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix}_1 \quad \Rightarrow \quad \text{valor propio}_1 \quad \lambda_1 = -1$$

$$\text{vector propio}_1 \quad \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

- Vectores y valores propios.

- Segundo. Busquemos el vector propio para el valor propio $\lambda_2 = 4$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_2 = \lambda_2 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_2$$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_2 = 4 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_2 \quad \rightarrow \quad \begin{array}{l} 2v_1 + 3v_2 = 4v_1 \\ 2v_1 + v_2 = 4v_2 \end{array} \quad \rightarrow \quad \begin{array}{l} -2v_1 + 3v_2 = 0 \\ 2v_1 + -3v_2 = 0 \end{array} \quad \rightarrow \quad v_2 = \frac{2v_1}{3}$$

- El sistema tiene infinitas soluciones de la forma $\begin{bmatrix} v_1 & \frac{2v_1}{3} \end{bmatrix}^T$. Por ejemplo si tomamos $v_1 = 3$, entonces $v_2 = 2$. Así nuestro **vector propio** queda de la siguiente forma:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix}_2 = \begin{bmatrix} 12 \\ 8 \end{bmatrix}_2$$

$$\rightarrow = 4 \cdot \begin{bmatrix} 3 \\ 2 \end{bmatrix}_1 \quad \rightarrow \quad \text{valor propio}_2 \quad \lambda_2 = 4$$

$$\text{vector propio}_2 \quad \mathbf{v}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

■ Vectores y valores propios.

- Una importante propiedad, denominada **transformación de similaridad**, se expresa con la siguiente igualdad:

$$\mathbf{V}^{-1} \cdot \mathbf{A} \cdot \mathbf{V} = \mathbf{B}$$

- Donde la matriz V corresponde a los **vectores propios** y la matriz B corresponde a la **diagonal de los valores propios**. Verifiquemos dicho resultado con nuestro ejemplo.

$$\begin{bmatrix} 1 & 3 \\ -1 & 2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 3 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 0.4 & -0.6 \\ 0.2 & 0.2 \end{bmatrix} \cdot \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 3 \\ -1 & 2 \end{bmatrix}$$
$$= \begin{bmatrix} 0.4 & -0.6 \\ 0.2 & 0.2 \end{bmatrix} \cdot \begin{bmatrix} -1 & 12 \\ 1 & 8 \end{bmatrix}$$
$$= \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix} \quad \rightarrow \quad \mathbf{B} = \begin{bmatrix} -1 & 0 \\ 0 & 4 \end{bmatrix}$$

- Este resultado concuerda con nuestros resultados previos, ya que el primer valor propio es $\lambda_1 = -1$ y el segundo valor propio es $\lambda_2 = 4$ los cuales corresponden a los valores almacenados en la diagonal de la matriz B



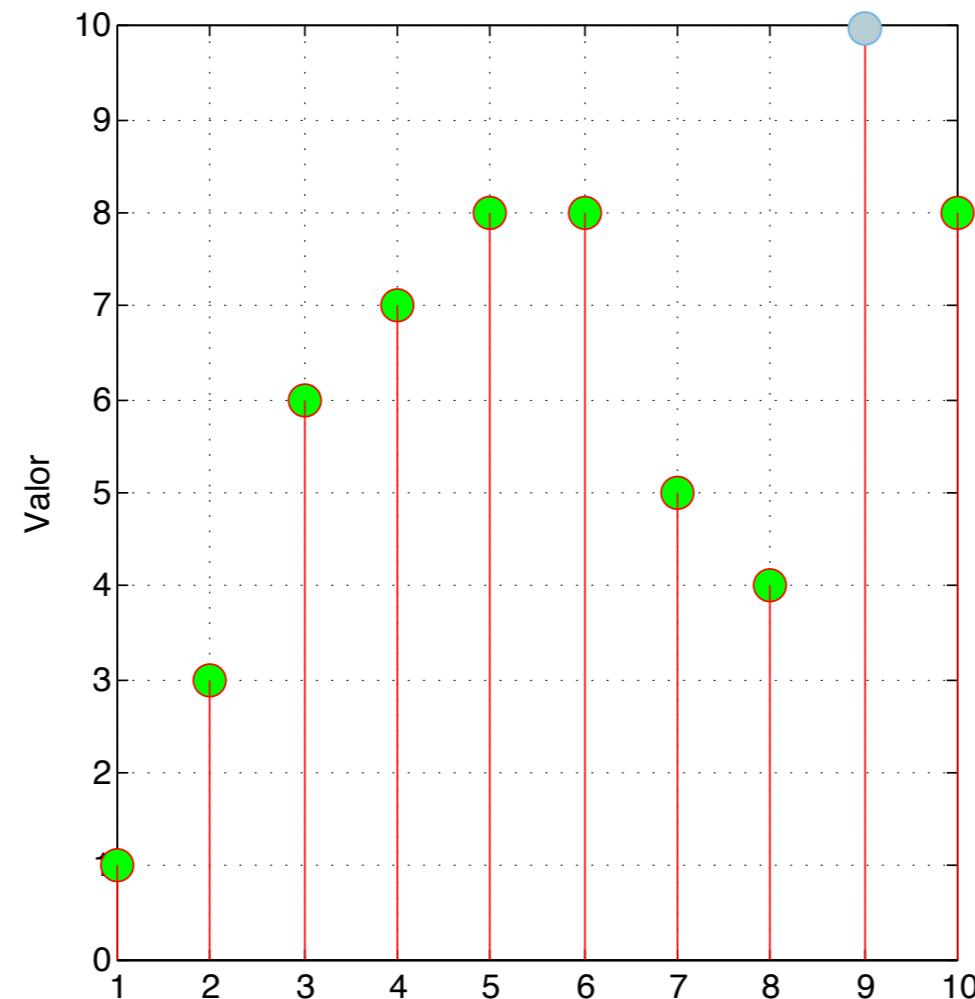
Anexo

Desviación estándar,
Varianza y Covarianza



■ Media:

- Es una medida estadística que mide el centro de gravedad de una distribución o conjunto de datos.
- Supongamos los siguientes datos: $X = [\begin{array}{cccccccccc} 1 & 3 & 6 & 7 & 8 & 8 & 5 & 4 & 10 & 8 \end{array}]$



Ejemplo

En el gráfico, los valores están en el eje de la X

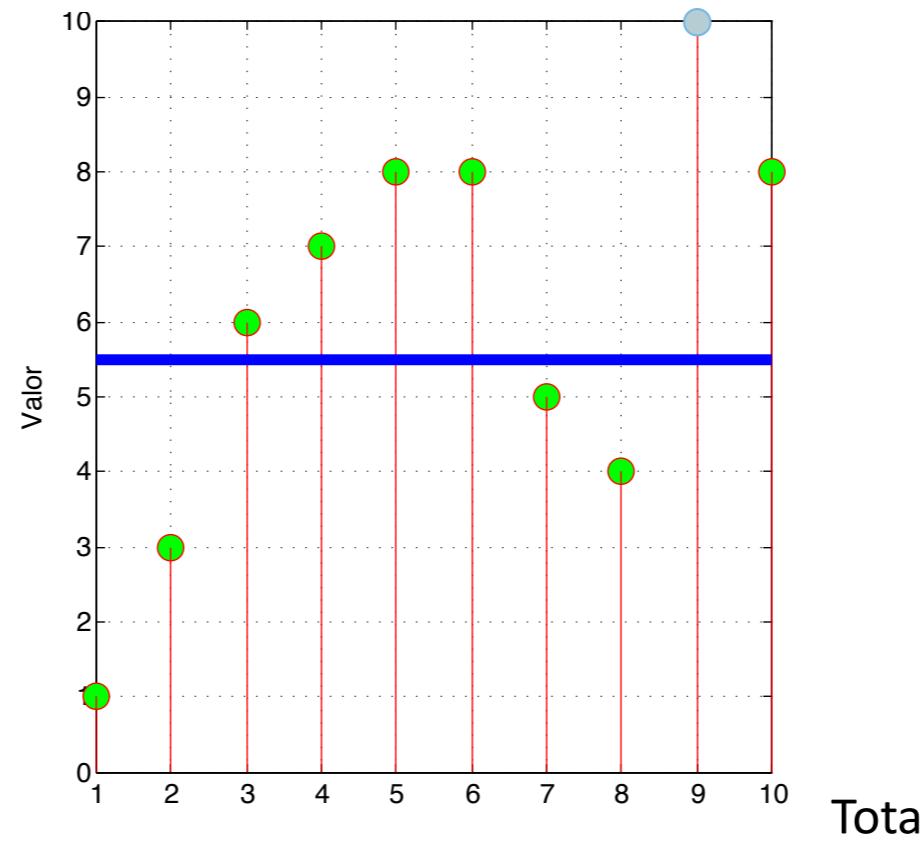


■ Media:

- Es una medida estadística que mide el centro de gravedad de una distribución o conjunto de datos.
- Supongamos los siguientes datos: $X = [\quad 1 \quad 3 \quad 6 \quad 7 \quad 8 \quad 8 \quad 5 \quad 4 \quad 10 \quad 8 \quad]$

Media

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$



X
1
3
6
7
8
8
5
4
10
8

60

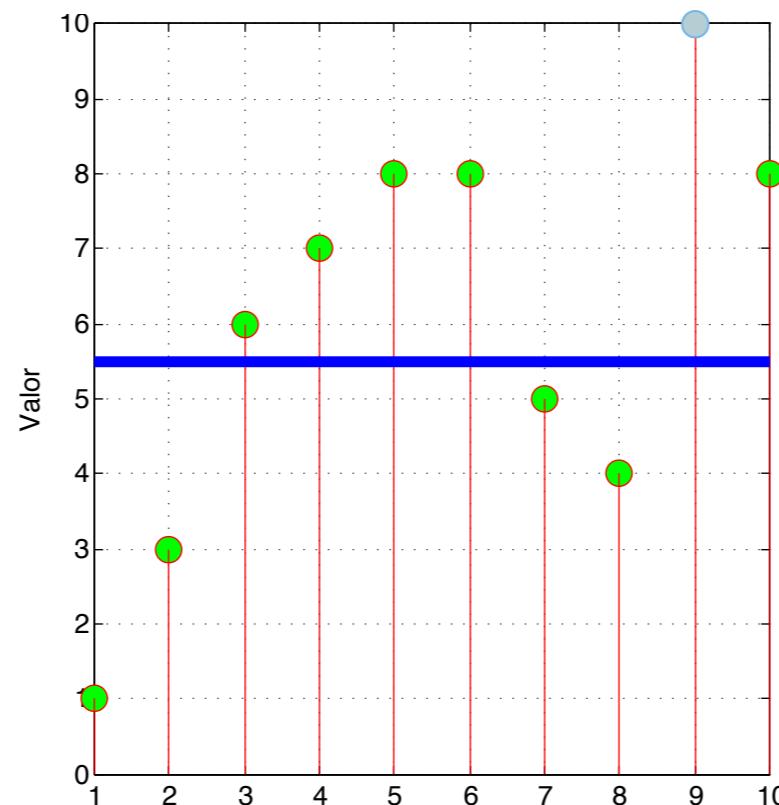
$$\bar{x} = \frac{60}{10}$$

■ Varianza

- Es la media aritmética de las desviaciones respecto a la media elevada al cuadrado. Si calculamos respecto a todos los datos se denomina **varianza poblacional** (con n datos). Si empleamos una muestra de los datos, se denomina **varianza muestral** el cual es un valor imparcial de los datos. En este último caso dividimos por $n-1$.

Varianza

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$$



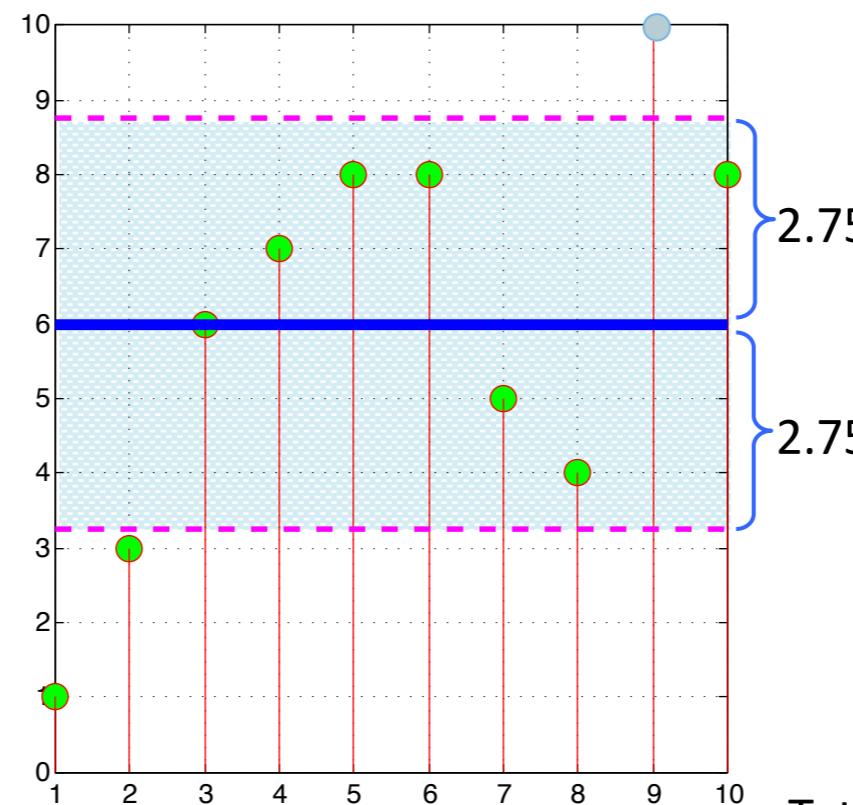
X	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	-5	25
3	-3	9
6	0	0
7	1	1
8	2	4
8	2	4
5	-1	1
4	-2	4
10	4	16
8	2	4
Total	60	68
División ($n-1$)		$s^2 = 7.55$

■ Desviación Estándar:

- Es una medida estadística que mide la dispersión de una variable unidimensional respecto a la media, o bien el promedio de la distancia a la media. En nuestro ejemplo, siempre emplearemos la desviación estándar respecto a la **varianza poblacional**.

Desviación estándar

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2}$$



X	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	-5	25
3	-3	9
6	0	0
7	1	1
8	2	4
8	2	4
5	-1	1
4	-2	4
10	4	16
8	2	4
Total		68
División (n-1)		$s^2 = 7.55$
Raíz cuadrada		$s = 2.75$

■ Matriz de Covarianza

- Es una medida estadística que mide la dispersión entre **dos variables estadísticas**.
- Por ejemplo, si queremos medir la covarianza entre **dos** variables (x,y), debemos medir la covarianza entre los pares ($x-x$), ($x-y$) e ($y-y$).

$$C = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{cov}(y,y) \end{bmatrix}_{2 \times 2}$$

Covarianza (X-Y)

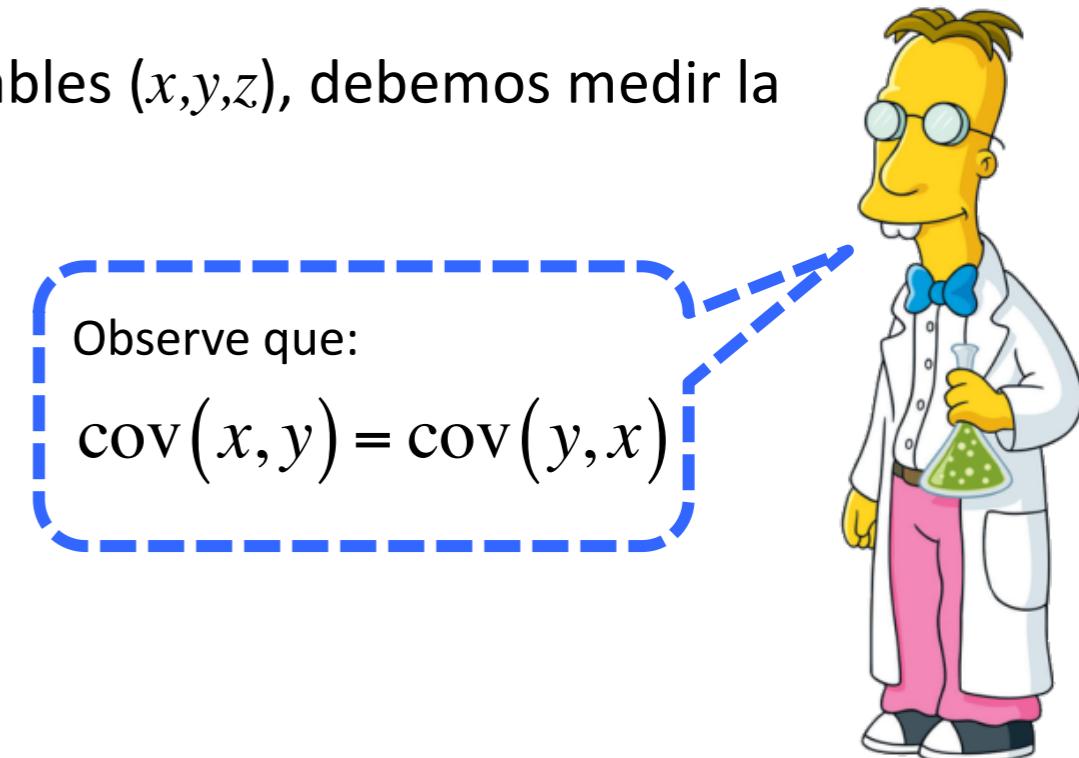
$$\text{cov}(x,y) = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Si queremos medir la covarianza entre **tres** variables (x,y,z), debemos medir la covarianza entre los pares ($x-y$), ($x-z$) e ($y-z$).

$$C = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{bmatrix}_{3 \times 3}$$

Observe que:

$$\text{cov}(x,y) = \text{cov}(y,x)$$



■ Matriz de Covarianza

- **Pregunta:** ¿Cuántas covarianzas calculamos si queremos medir la matriz de covarianza entre n variables?

➤ **Respuesta:**

Corresponde al número de covarianzas en su misma dimensión más el coeficiente binomial

$$C = \begin{bmatrix} \text{cov}(\text{dim}_1, \text{dim}_1) & \text{cov}(\text{dim}_1, \text{dim}_2) & \cdots & \text{cov}(\text{dim}_1, \text{dim}_n) \\ \text{cov}(\text{dim}_2, \text{dim}_1) & \text{cov}(\text{dim}_2, \text{dim}_2) & \cdots & \text{cov}(\text{dim}_2, \text{dim}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\text{dim}_n, \text{dim}_1) & \text{cov}(\text{dim}_n, \text{dim}_2) & \cdots & \text{cov}(\text{dim}_n, \text{dim}_n) \end{bmatrix}_{n \times n}$$

$$\text{combinaciones} = n + \frac{n!}{2 \cdot (n - 2)!}$$

- **Pregunta:** ¿Qué indica el valor de la covarianza?

si $\text{cov}(x, y) > 0$

Significa dependencia positiva

✧ valores grandes en x corresponden valores grandes en y

si $\text{cov}(x, y) = 0$

Significa que no existe una relación lineal entre las variables x e y

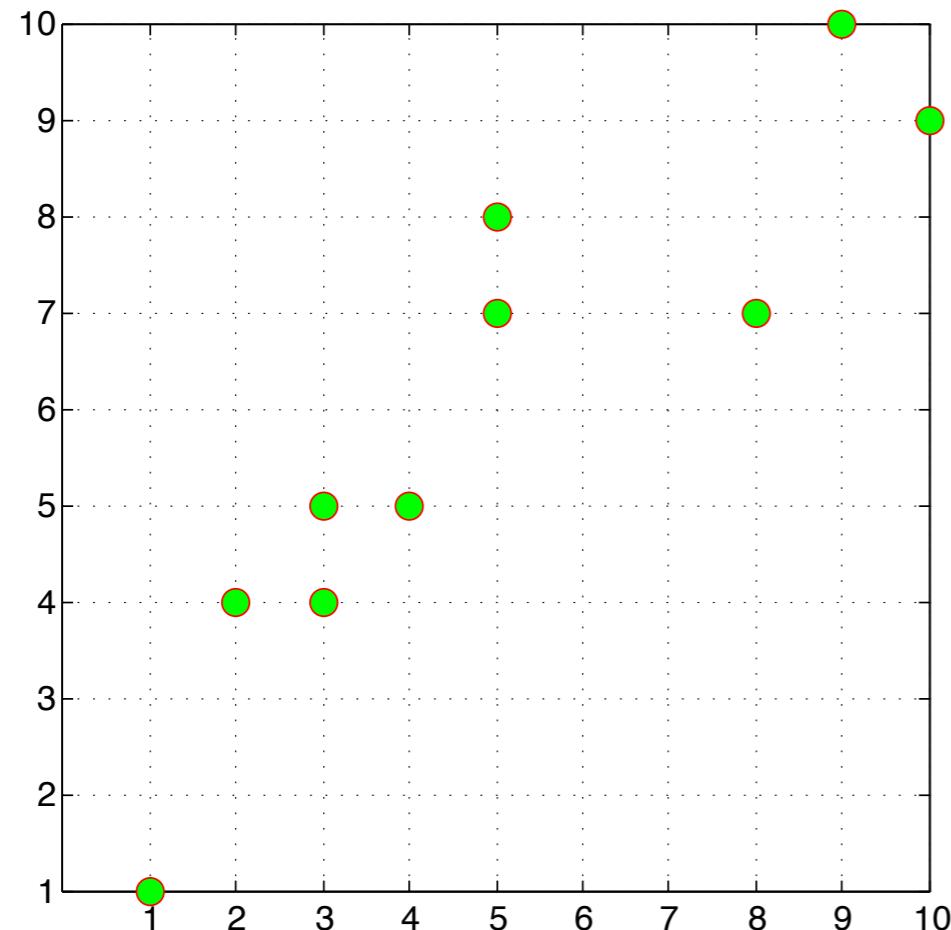
si $\text{cov}(x, y) < 0$

Significa dependencia negativa o inversa

✧ valores grandes en x corresponden valores pequeños en y

Covarianza

- Es una medida estadística que mide la dispersión entre **dos variables estadísticas**. En el ejemplo, suponga que tenemos **dos** características, y deseamos analizar si existe una relación entre ambas variables.



X	Y
1	1
2	4
3	5
5	7
8	7
5	8
4	5
3	4
9	10
10	9

Σ 50 60

Primero:
calculamos las medias

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$X \longrightarrow \bar{x} = \frac{50}{10} = 5$$

$$Y \longrightarrow \bar{y} = \frac{60}{10} = 6$$

Covarianza

- Es una medida estadística que mide la dispersión entre **dos variables estadísticas**. En el ejemplo, suponga que tenemos **dos** características, y deseamos analizar si existe una relación entre ambas variables.

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$
1	1	-4	-5
2	4	-3	-2
3	5	-2	-1
5	7	0	1
8	7	3	1
5	8	0	2
4	5	-1	-1
3	4	-2	-2
9	10	4	4
10	9	5	3
Σ	50	60	
	$\bar{x} = 5$	$\bar{y} = 6$	

Segundo:
Restamos cada media
según cada columna
(esto centra los datos)

—————> $(x_i - \bar{x})$

—————> $(y_i - \bar{y})$

Covarianza

- Es una medida estadística que mide la dispersión entre **dos variables estadísticas**. En el ejemplo, suponga que tenemos **dos** características, y deseamos analizar si existe una relación entre ambas variables.

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	1	-4	-5	16	20	25
2	4	-3	-2	9	6	4
3	5	-2	-1	4	2	1
5	7	0	1	0	0	1
8	7	3	1	9	3	1
5	8	0	2	0	0	4
4	5	-1	-1	1	1	1
3	4	-2	-2	4	4	4
9	10	4	4	16	16	16
10	9	5	3	25	15	9
Σ	50	60				
	$\bar{x} = 5$	$\bar{y} = 6$				

Tercero:
Multiplicamos cada
valor fila por fila

$$(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Covarianza

- Es una medida estadística que mide la dispersión entre **dos variables estadísticas**. En el ejemplo, suponga que tenemos **dos** características, y deseamos analizar si existe una relación entre ambas variables.

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	1	-4	-5	16	20	25
2	4	-3	-2	9	6	4
3	5	-2	-1	4	2	1
5	7	0	1	0	0	1
8	7	3	1	9	3	1
5	8	0	2	0	0	4
4	5	-1	-1	1	1	1
3	4	-2	-2	4	4	4
9	10	4	4	16	16	16
10	9	5	3	25	15	9
Σ	50	60		84	67	66
Cuarto: Sumamos cada columna				$\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$		

Covarianza

- Es una medida estadística que mide la dispersión entre **dos variables estadísticas**. En el ejemplo, suponga que tenemos **dos** características, y deseamos analizar si existe una relación entre ambas variables.

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})^2$	Quinto: Dividimos por $(n-1)$
1	1	-4	-5	16	20	25	$\frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
2	4	-3	-2	9	6	4	
3	5	-2	-1	4	2	1	
5	7	0	1	0	0	1	
8	7	3	1	9	3	1	
5	8	0	2	0	0	4	
4	5	-1	-1	1	1	1	
3	4	-2	-2	4	4	4	
9	10	4	4	16	16	16	
10	9	5	3	25	15	9	
Σ	50	60		84	67	66	
				$\frac{1}{n-1}$	9.33	7.44	7.33

■ Matriz de Covarianza

- Es una medida estadística que mide la dispersión entre **dos variables estadísticas**. En el ejemplo, suponga que tenemos **dos** características, y deseamos analizar si existe una relación entre ambas variables.

Covarianza

$$\text{cov}(x, y) = S_{xy} = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{cov}(x, x) = 9.33$$

$$\text{cov}(x, y) = 7.44$$

$$\text{cov}(y, y) = 7.33$$

Matriz de Covarianza

$$C = \begin{bmatrix} \text{cov}(x, x) & \text{cov}(x, y) \\ \text{cov}(y, x) & \text{cov}(y, y) \end{bmatrix}_{2 \times 2}$$

$$C = \begin{bmatrix} 9.33 & 7.44 \\ 7.44 & 7.33 \end{bmatrix}_{2 \times 2}$$