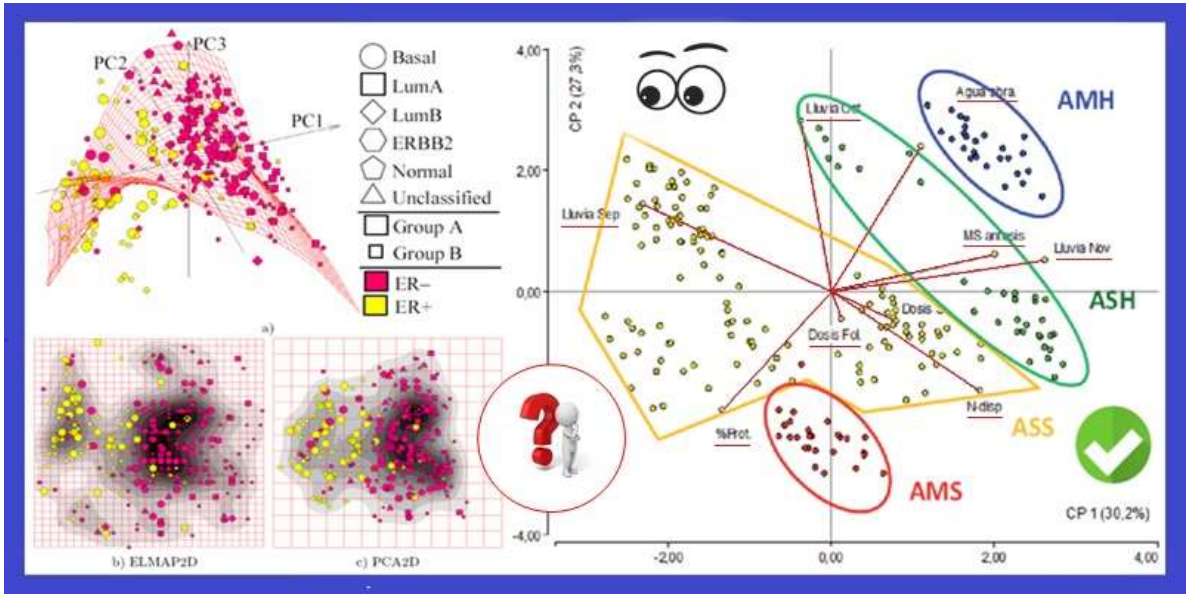
Análisis de componentes principales (ACP)

pip3 --version



Luis Raúl Figueroa Martínez

TEMAS SELECTOS DE ESTADÍSTICA I
APRENDIZAJE AUTOMÁTICO Y APRENDIZAJE GEOMÉTRICO PROFUNDO

La inteligencia, lo que consideramos acciones inteligentes,
se modifica a lo largo de la historia…
es una colección de potencialidades que se completan.
Howard Gardner

Historia del PCA

El **Análisis de Componentes Principales (PCA)** fue introducido por primera vez por el matemático **Karl Pearson** en **1901**, cuando publicó un artículo titulado *"On Lines and Planes of Closest Fit to Systems of Points in Space"*. Pearson estaba interesado en encontrar una manera de simplificar datos complejos, y su trabajo sentó las bases para lo que hoy llamamos **reducción de dimensionalidad**.

Más tarde, en **1933**, el estadístico **Harold Hotelling** extendió y generalizó el trabajo de Pearson, dándole un enfoque más formal y matemáticamente riguroso, lo que permitió que PCA fuera aplicado a una amplia variedad de problemas en la estadística multivariante y otras disciplinas.

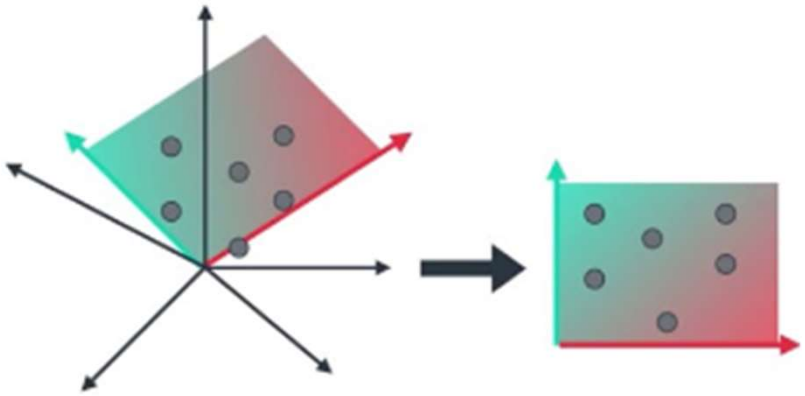
El problema de los datos de alta dimensionalidad

- **Ruido**: Muchas de las características pueden ser irrelevantes o redundantes.
- **Dificultad de visualización**: Los datos en más de tres dimensiones son imposibles de visualizar directamente.
- **Costos computacionales**: Los algoritmos pueden volverse ineficientes al trabajar con muchas características.

es útil encontrar una manera de **reducir la dimensionalidad** de los datos sin perder la información importante.

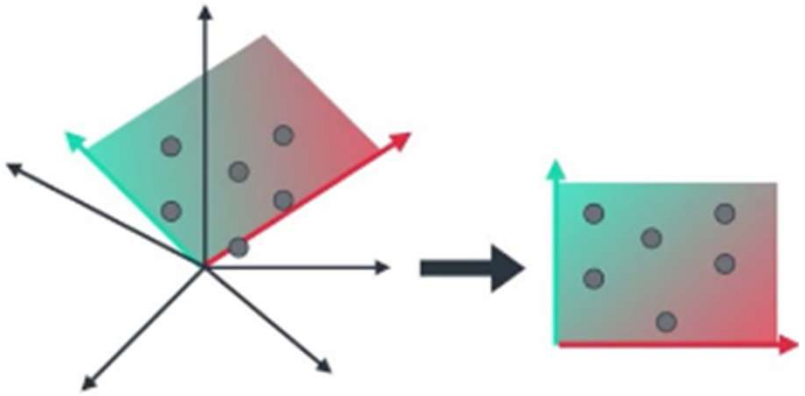
¿Cómo encontramos las nuevas dimensiones?

En vez de trabajar con las características originales, queremos proyectar los datos en un nuevo sistema de coordenadas donde las direcciones principales (componentes) capturen la máxima variabilidad. Es decir, buscamos transformar los datos originales a un nuevo espacio donde la información esencial esté representada en **menos dimensiones**.



Análisis de Componentes Principales (PCA)

Es una técnica de reducción de dimensionalidad que transforma un conjunto de variables posiblemente correlacionadas en un nuevo conjunto de variables no correlacionadas, conocidas como **componentes principales**. Estas nuevas variables están ordenadas de manera que los primeros componentes capturan la mayor cantidad de variabilidad presente en los datos originales.

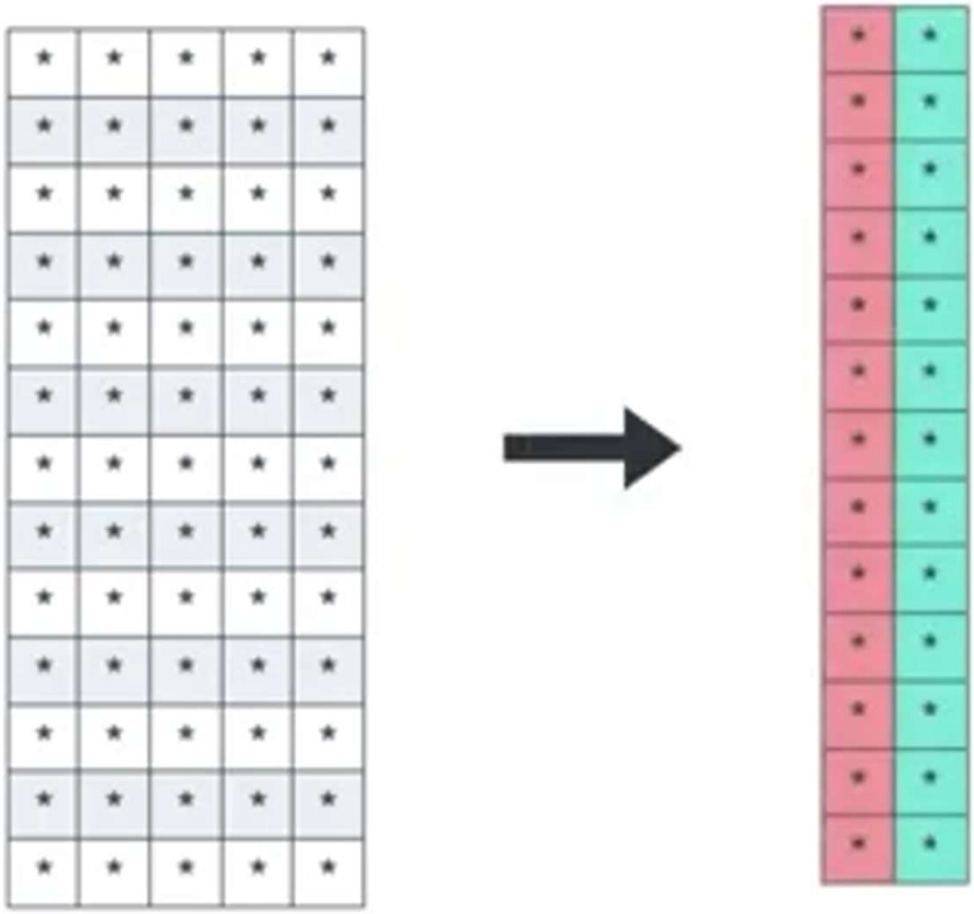


Matemáticamente, PCA busca:

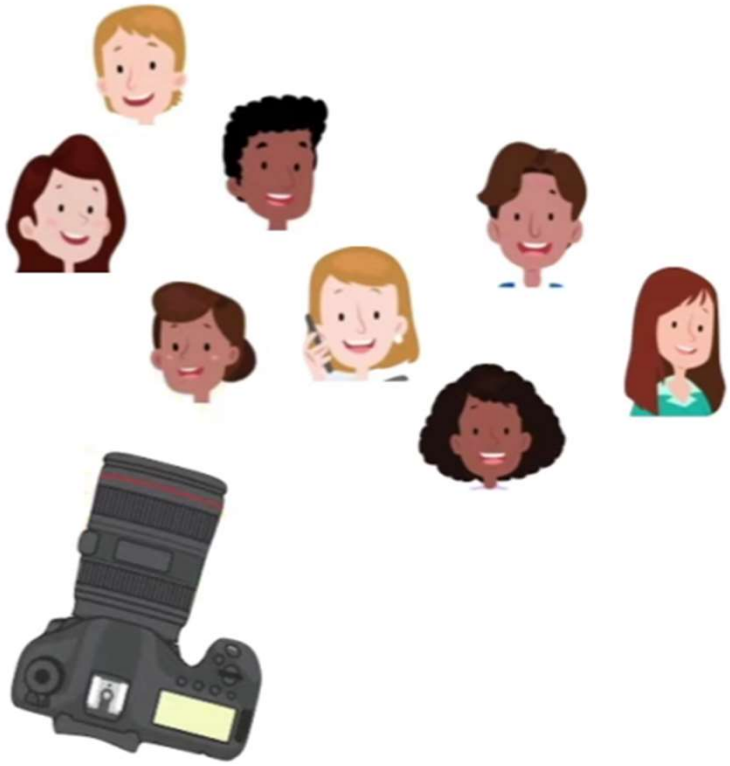
- Encontrar una proyección de los datos en un nuevo espacio de menor dimensión.
- Este nuevo espacio está definido por los **autovectores** de la matriz de covarianza de los datos.
- Los autovalores correspondientes a estos autovectores nos indican cuánta varianza de los datos originales captura cada componente principal.

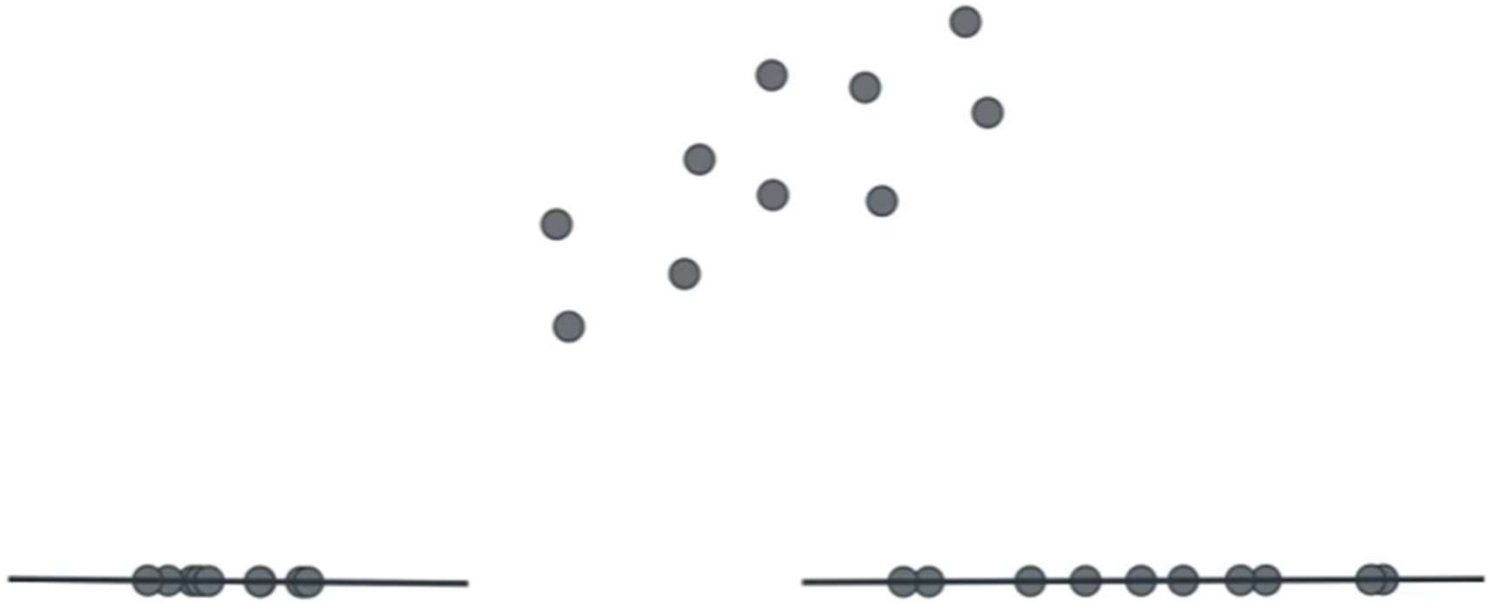
Construcción de PCA

Reducción de dimensionalidad



Tomar una foto es la forma más simple de reducir dimensionalidad





Ejemplo con datos de casas

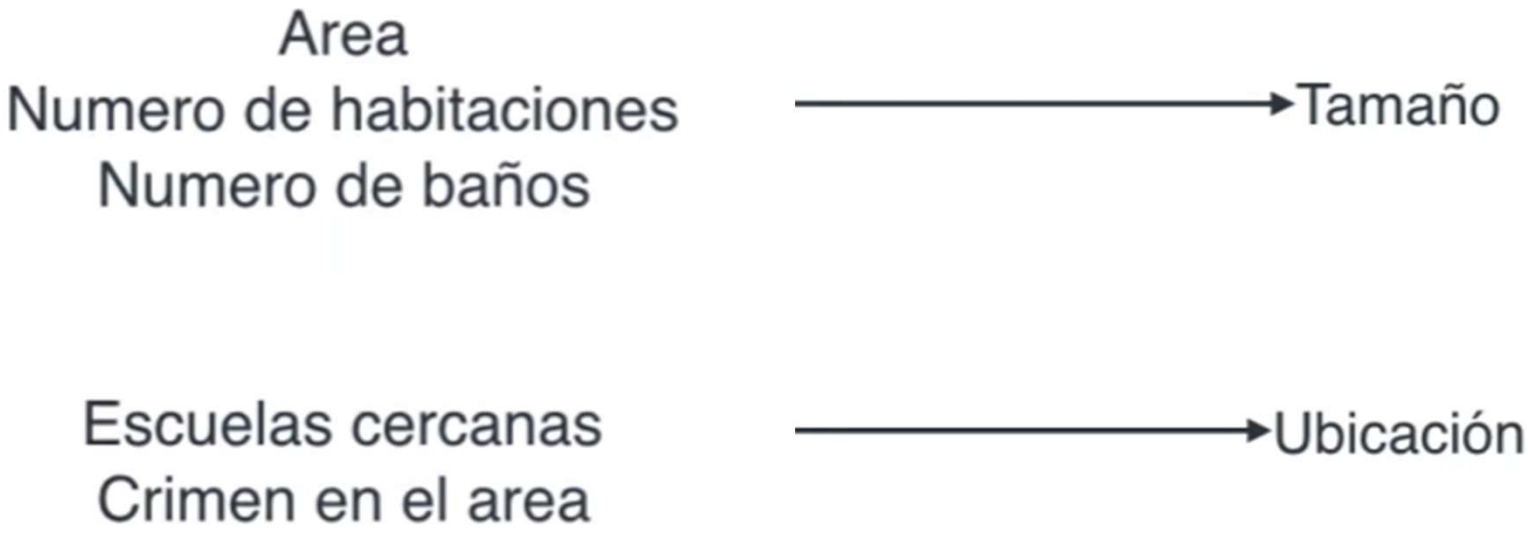
Area
Numero de habitaciones
Numero de baños

Tamaño

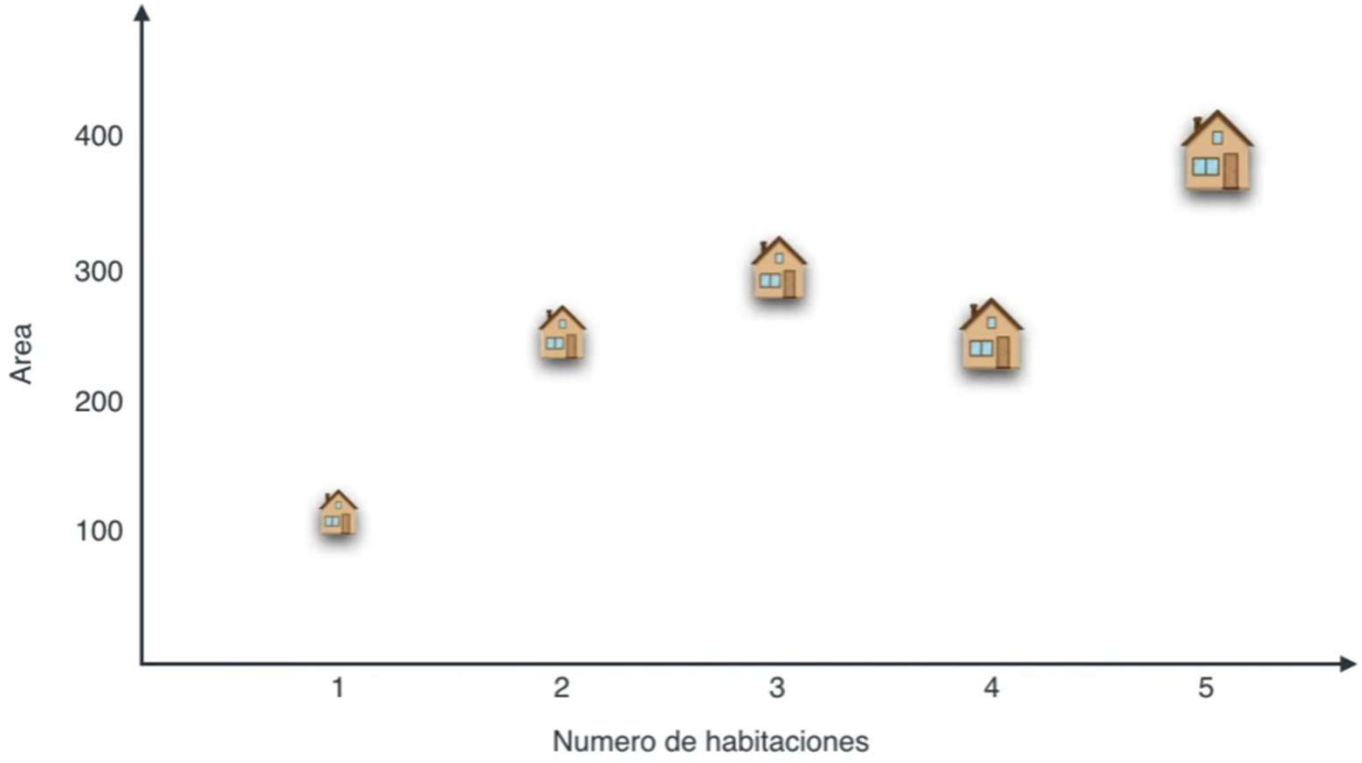
Escuelas cercanas
Crimen en el area

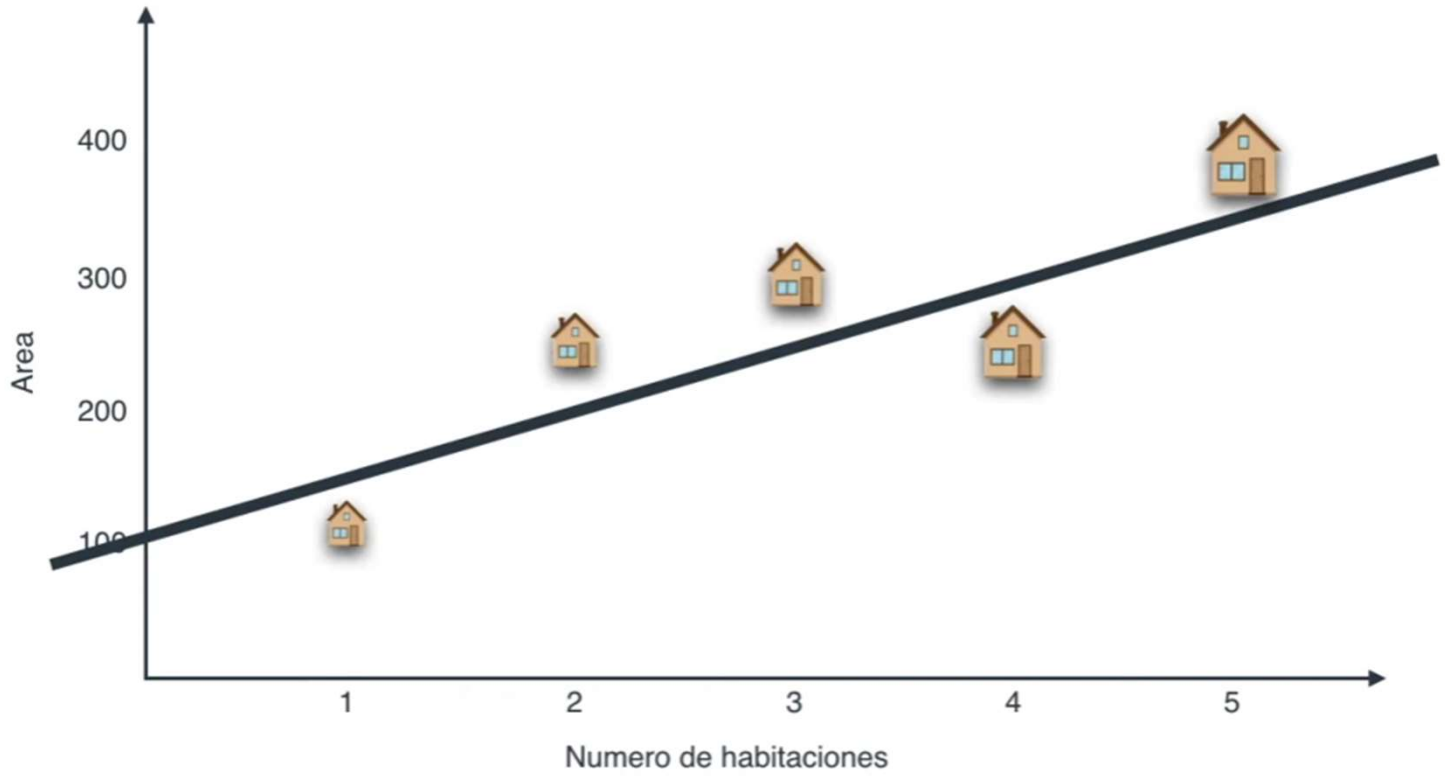
Ubicación

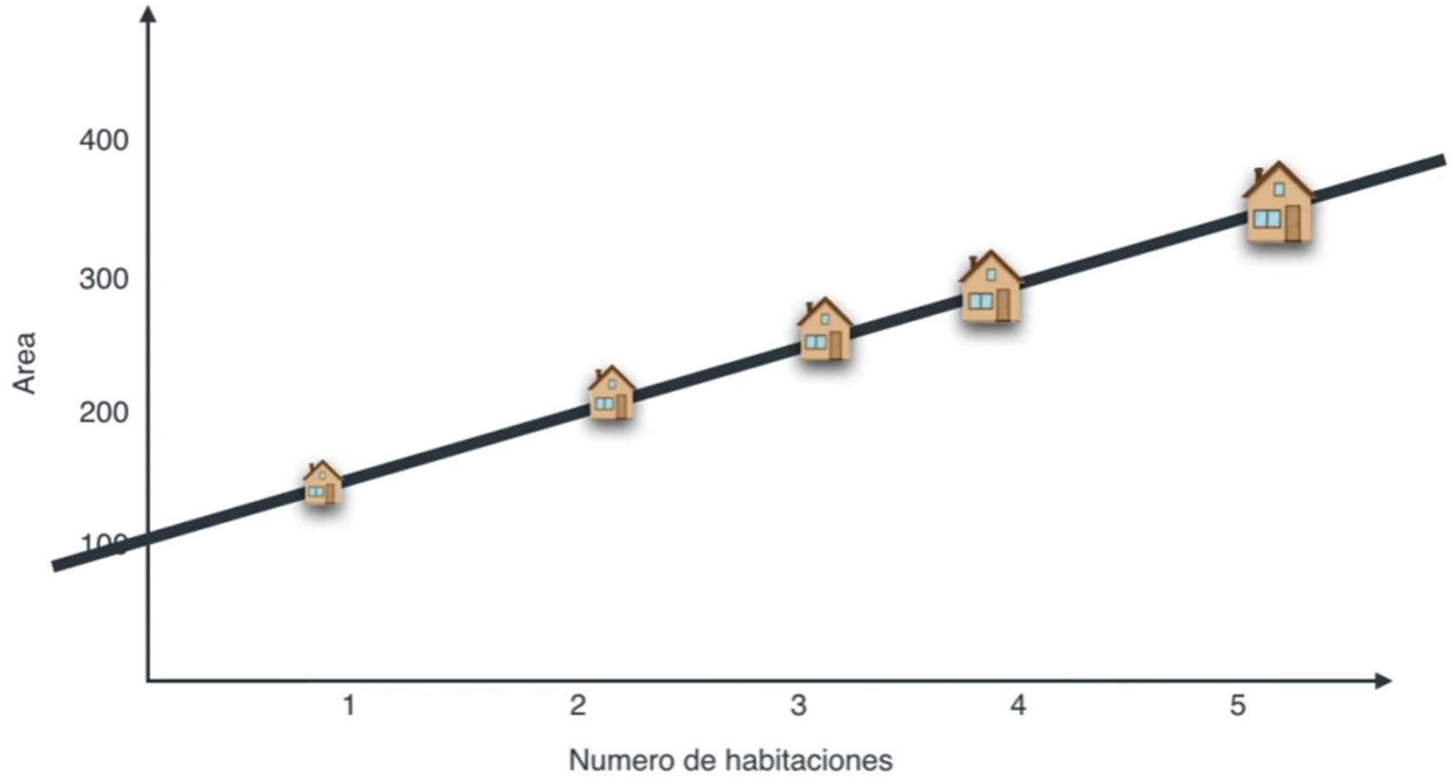
Ejemplo con datos de casas



Representacion geometrica 2-1

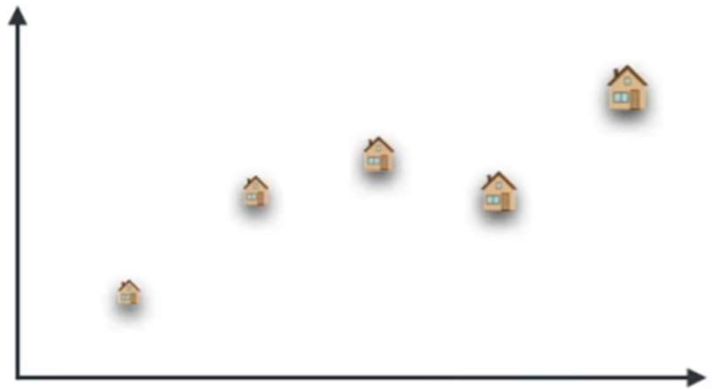



400
300
200
Area
1
2
3
4
5
Numero de habitaciones


400
300
200
100
Area
1
2
3
4
5
Numero de habitaciones

2 dimensiones

area
numero de habitaciones



1 dimension

tamaño

Ejemplo con datos de casas

5 dimensiones	2 dimensions
Area	
Numero de habitaciones	Tamaño
Numero de baños	
Escuelas cercanas	Ubicación
crimen en el area	

Conceptos de estadística (vista geométrica)

El promedio, también conocido como media aritmética, se define como:

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

donde $\bar{x}$ es el promedio, n es el número de elementos, y x_i representa cada uno de los elementos del conjunto.

La varianza de un conjunto de datos se define como:

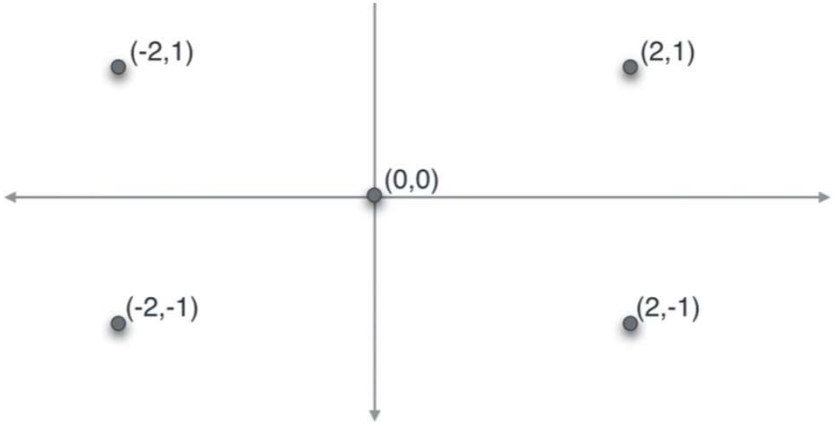
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

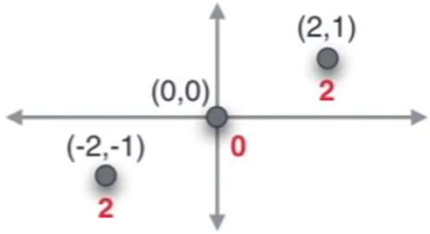
donde: - σ^2 es la varianza, - n es el número de elementos, - x_i son los elementos del conjunto, y - $\bar{x}$ es el promedio de los elementos.

La covarianza entre dos variables aleatorias X y Y se define como:

$$\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - X)(Y_i - Y)$$

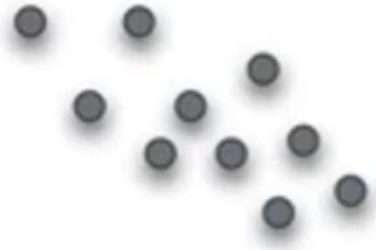
donde: - $\text{Cov}(X,Y)$ es la covarianza, - n es el número de pares de datos, - X_i y Y_i son los valores de las variables X y Y, - X es el promedio de X, y - Y es el promedio de Y.



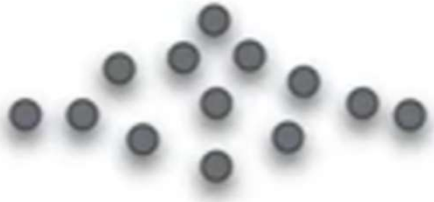


covarianza = $\frac{(-2) + 0 + (-2)}{3} = -4/3$

covarianza = $\frac{2 + 0 + 2}{3} = 4/3$



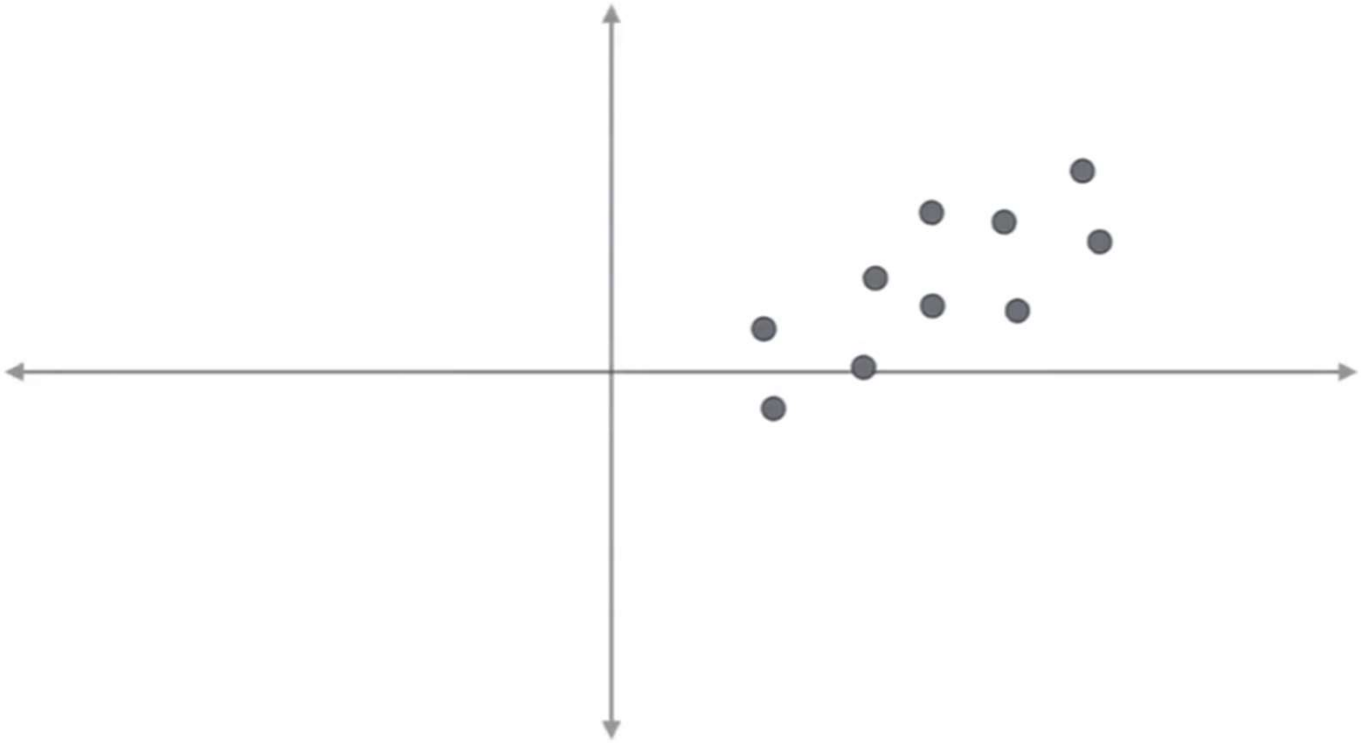
covarianza
negativa

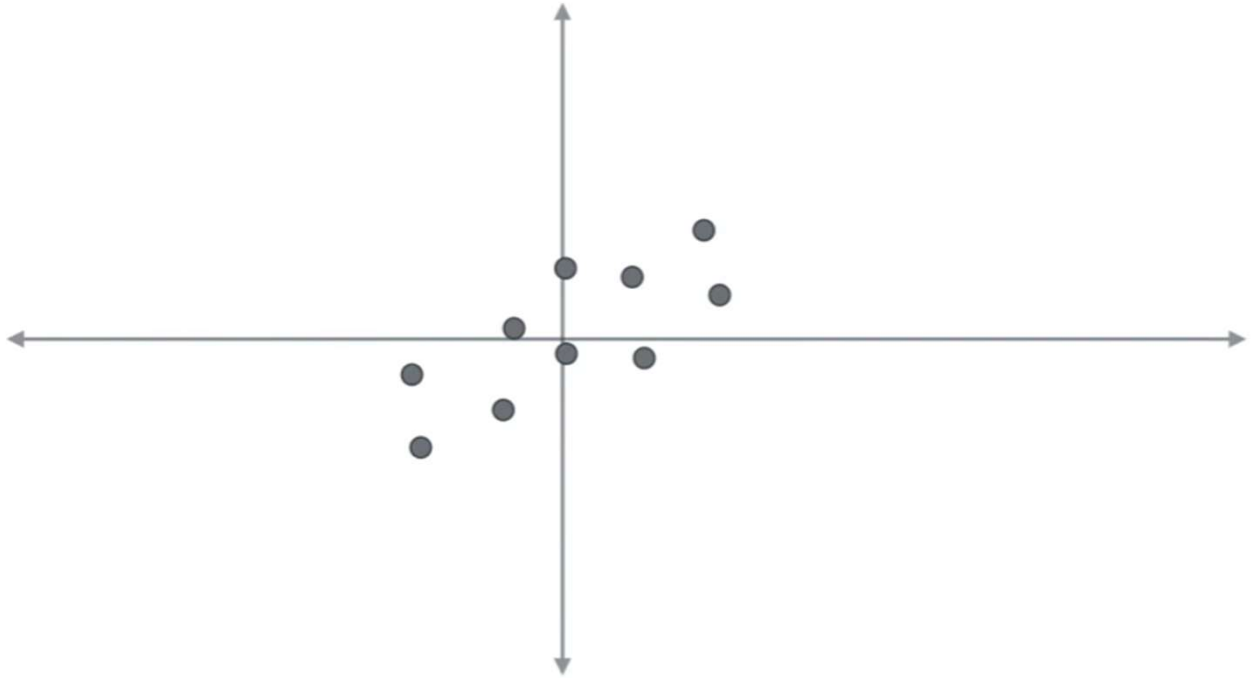


covarianza cero
(o muy pequena)

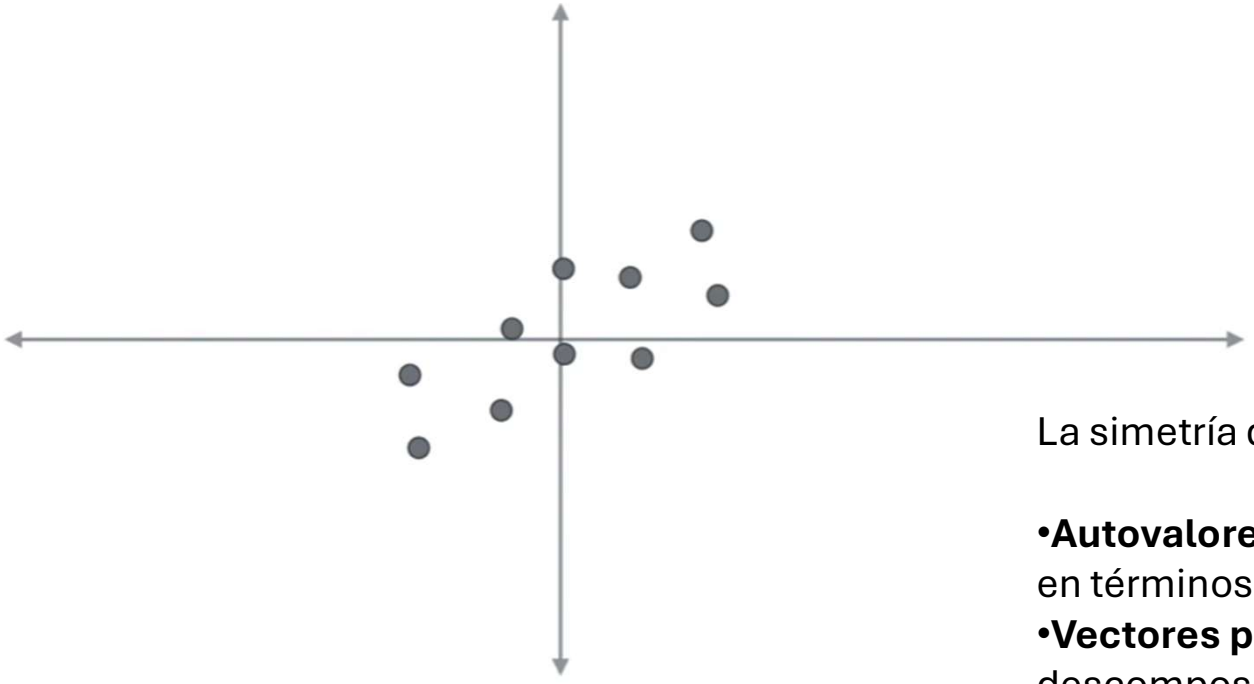
covarianza
positiva

Valores y vectores propios





Matriz de covarianza

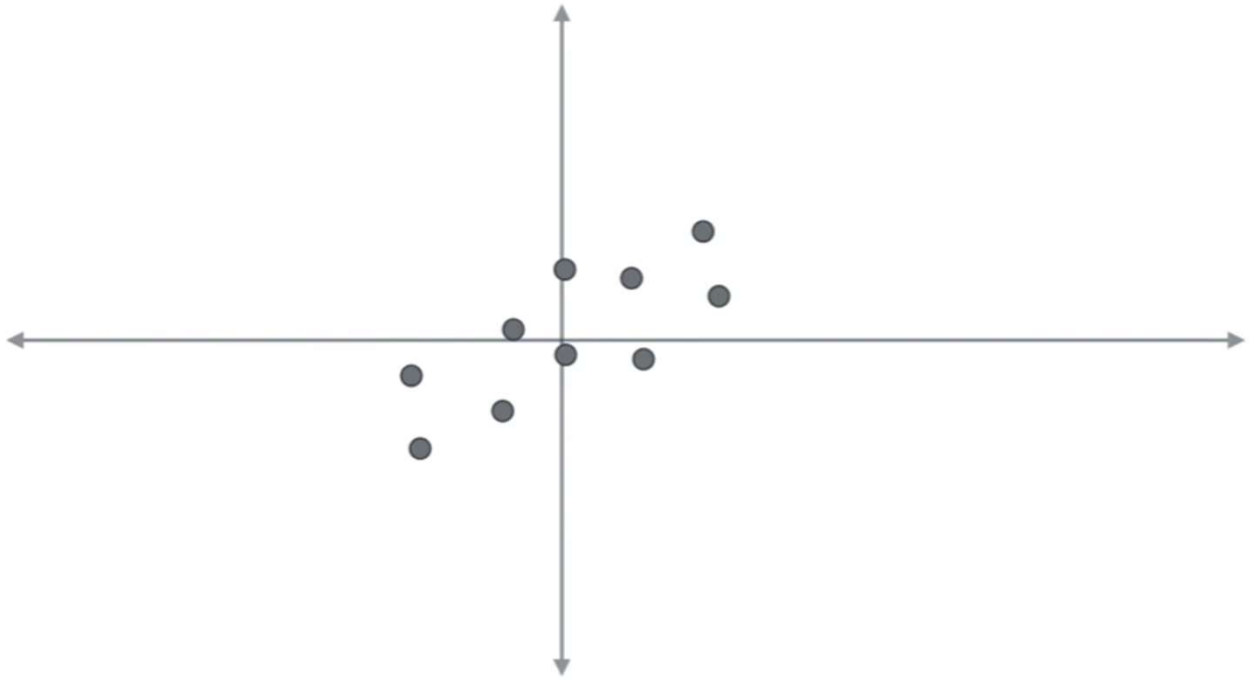


$$\Sigma = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(X,Y) & \text{Var}(Y) \end{pmatrix}$$

La simetría de la matriz de covarianza asegura que tenga:

- **Autovalores reales**, lo que permite interpretaciones claras en términos de varianza.
- **Vectores propios ortogonales**, que facilitan la descomposición en componentes principales y la visualización de la variación en los datos.

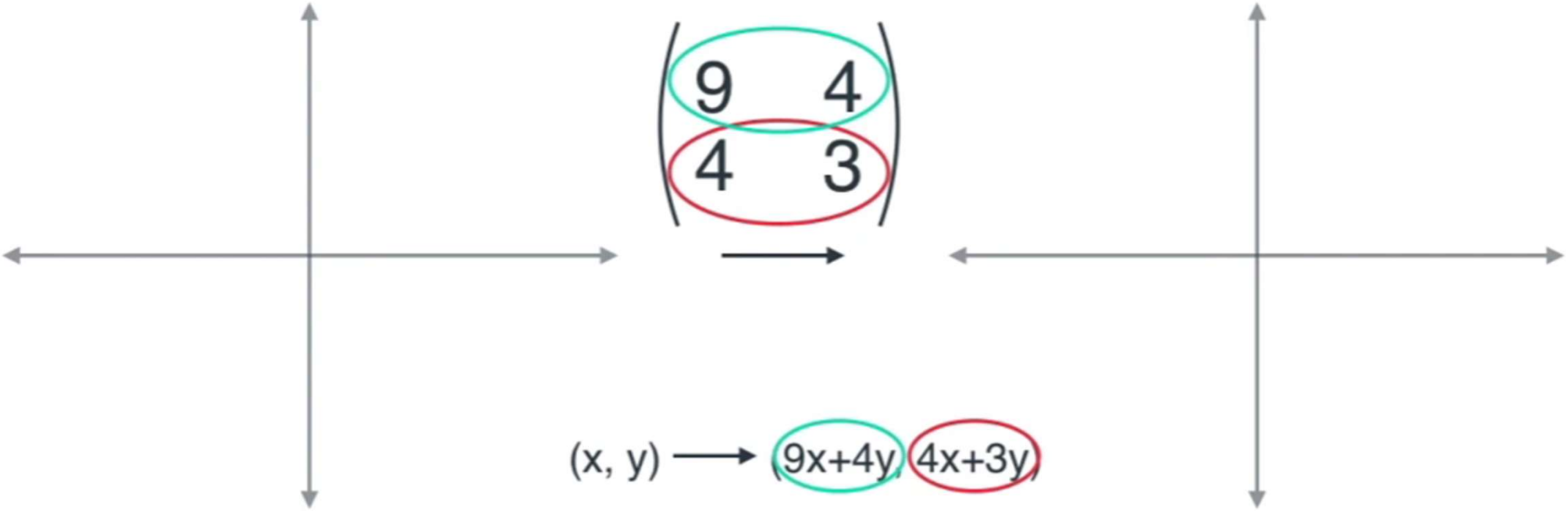
Matriz de covarianza



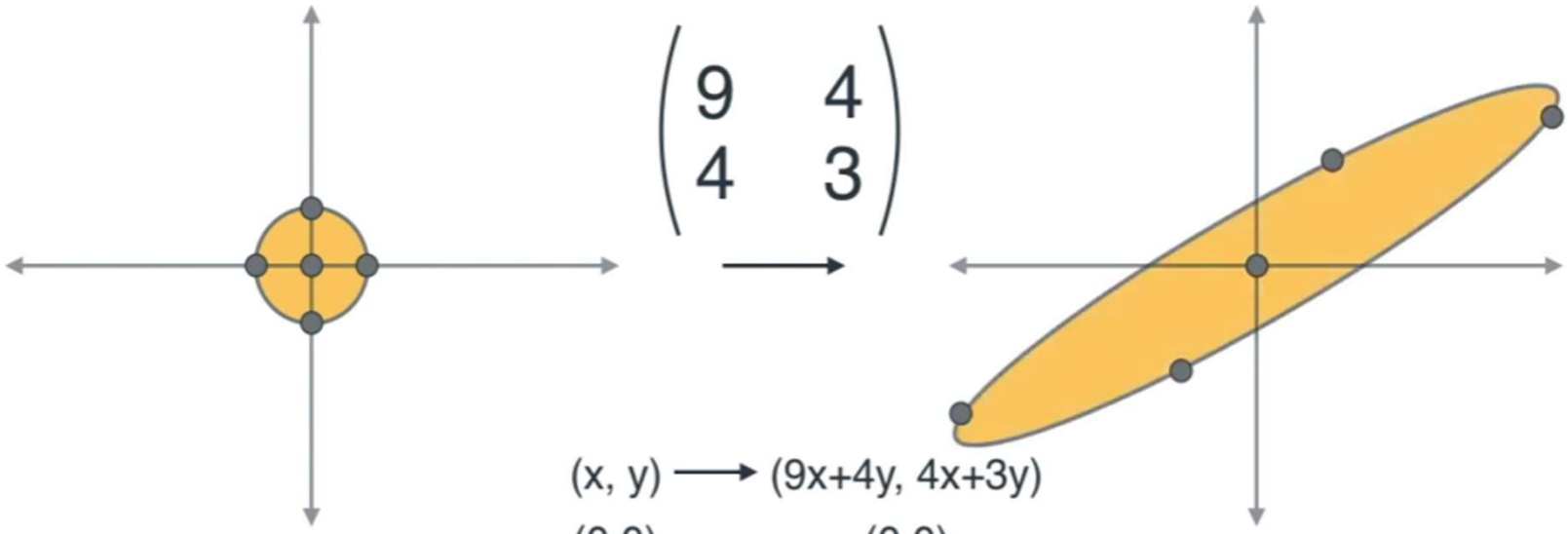
$$\Sigma = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(X,Y) & \text{Var}(Y) \end{pmatrix}$$

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

Transformación lineal

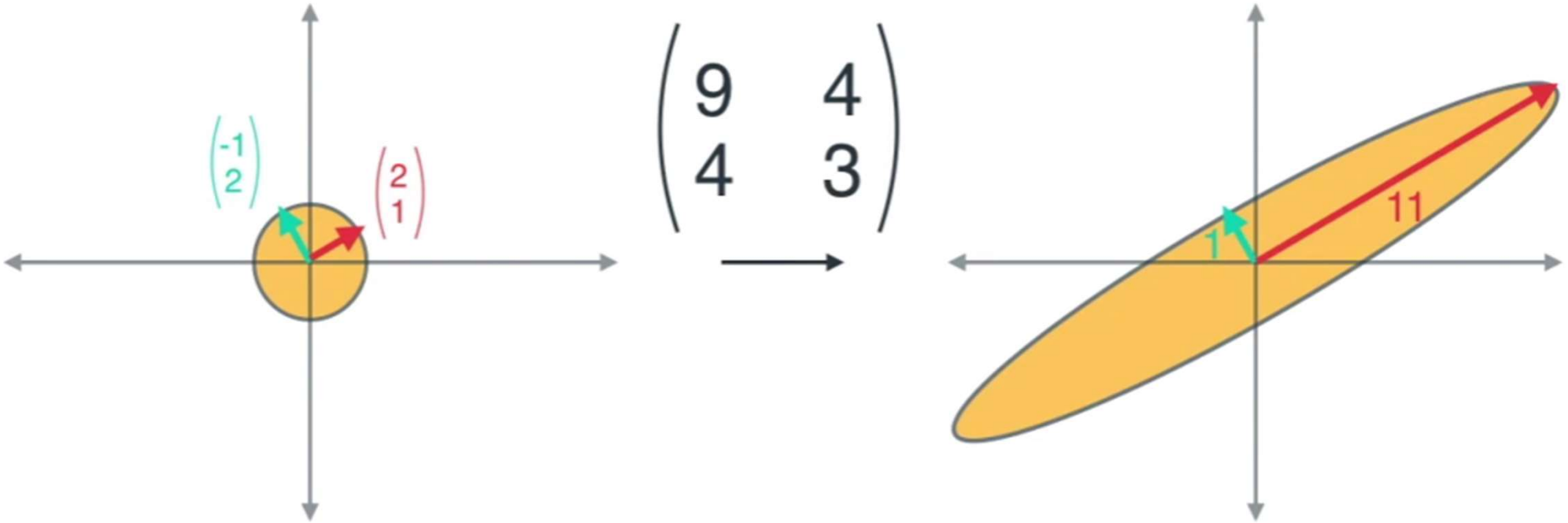


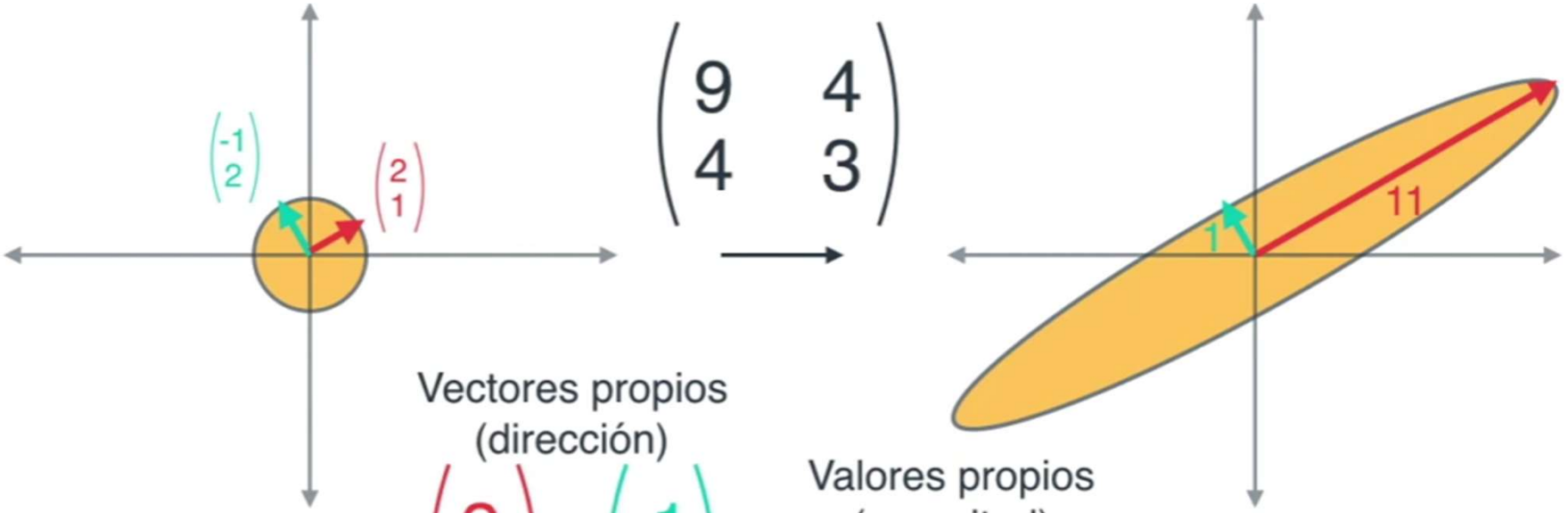
Transformación lineal



$(x, y) \longrightarrow (9x+4y,\ 4x+3y)$

(0,0)	(0,0)
(1,0)	(9,4)
(0,1)	(4,3)
(-1,0)	(-9,-4)
(0,-1)	(-4,-3)


-1
2
2
1
9 4
4 3
1
11



Vectores propios
(dirección)

$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$

Valores propios
(magnitud)

11 1

Valores Propios

Los valores propios de una matriz A se encuentran resolviendo el **polinomio característico**, que se obtiene de la siguiente ecuación:

$$\det(A - \lambda I) = 0$$

donde:

- $\det$ denota el determinante,
- λ es un escalar (valor propio),
- I es la matriz identidad del mismo tamaño que A.

Vectores Propios

Una vez que se han encontrado los valores propios λ, los vectores propios $\mathbf{v}$ se obtienen resolviendo el sistema de ecuaciones lineales:

$$(A - \lambda I)\mathbf{v} = 0$$

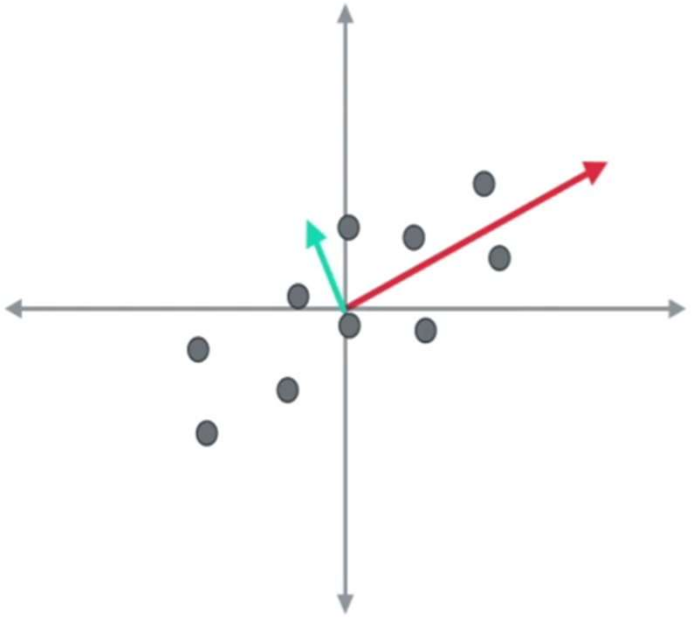
Esto implica que para cada valor propio λ, debes resolver:

1. Sustituir λ en $A - \lambda I$.
2. Encontrar los vectores $\mathbf{v}$ que satisfacen la ecuación.

Resumen del Proceso

1. **Calcular el polinomio característico**: $\det(A - \lambda I) = 0$.
2. **Resolver para λ**: Obtienes los valores propios.
3. **Para cada λ**: Sustituir en $(A - \lambda I)\mathbf{v} = 0$ para encontrar los vectores propios.

Análisis de componentes principales (ACP)



$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

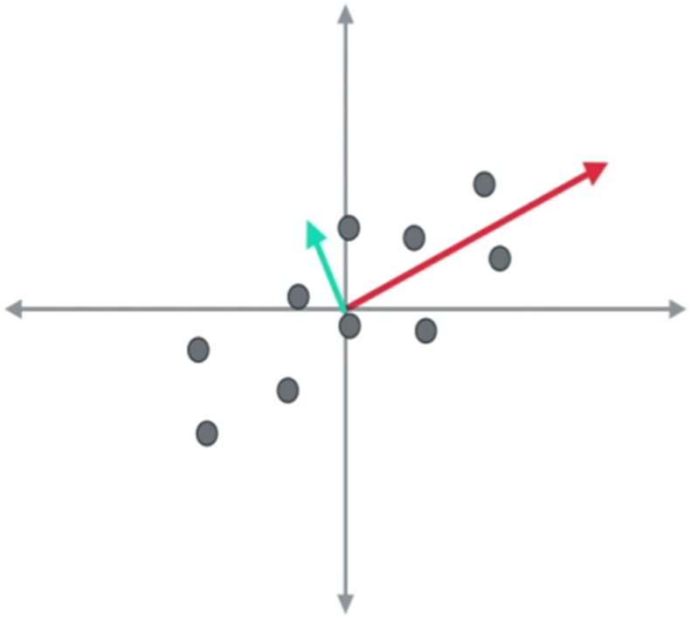
$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

Vectores propios (dirección)

11 1

Valores propios (magnitud)

Proyectar los datos



$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

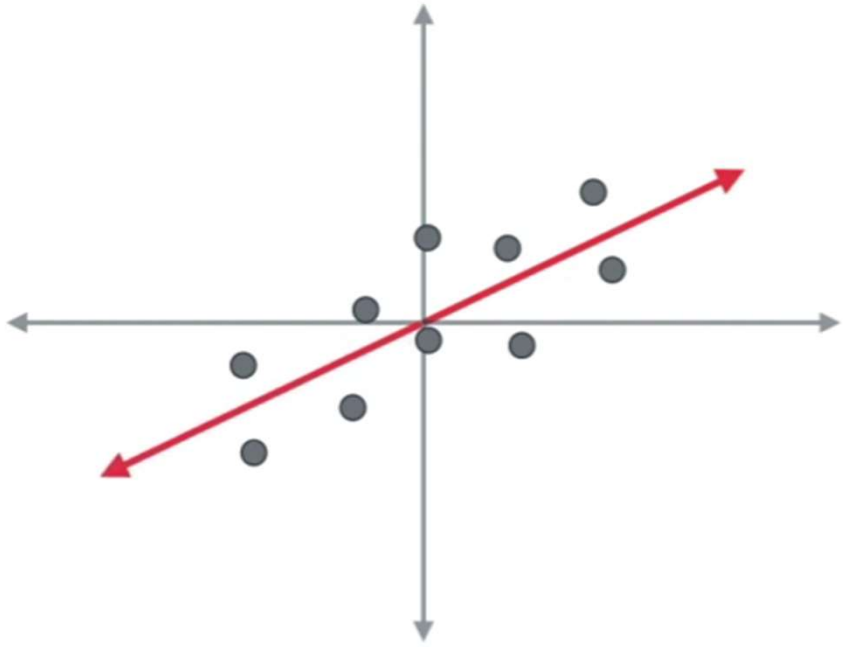
$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

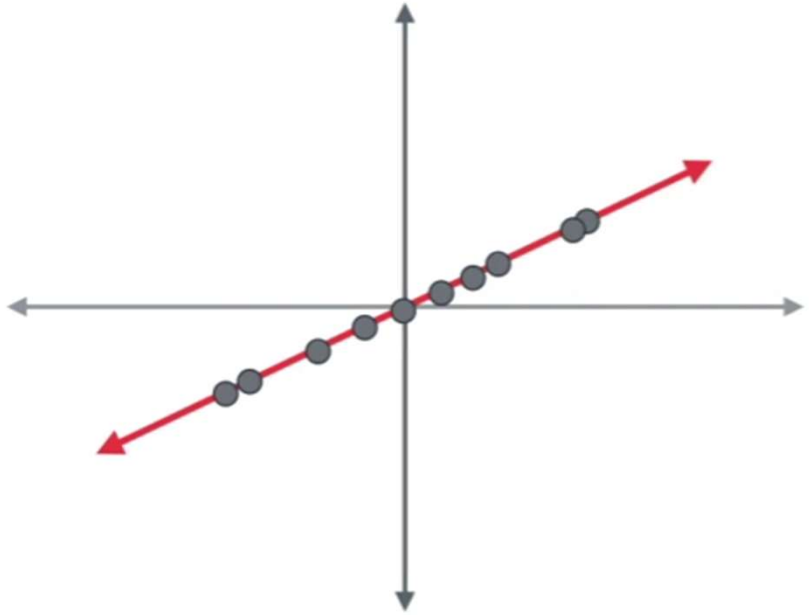
Vectores propios
(dirección)

11 1

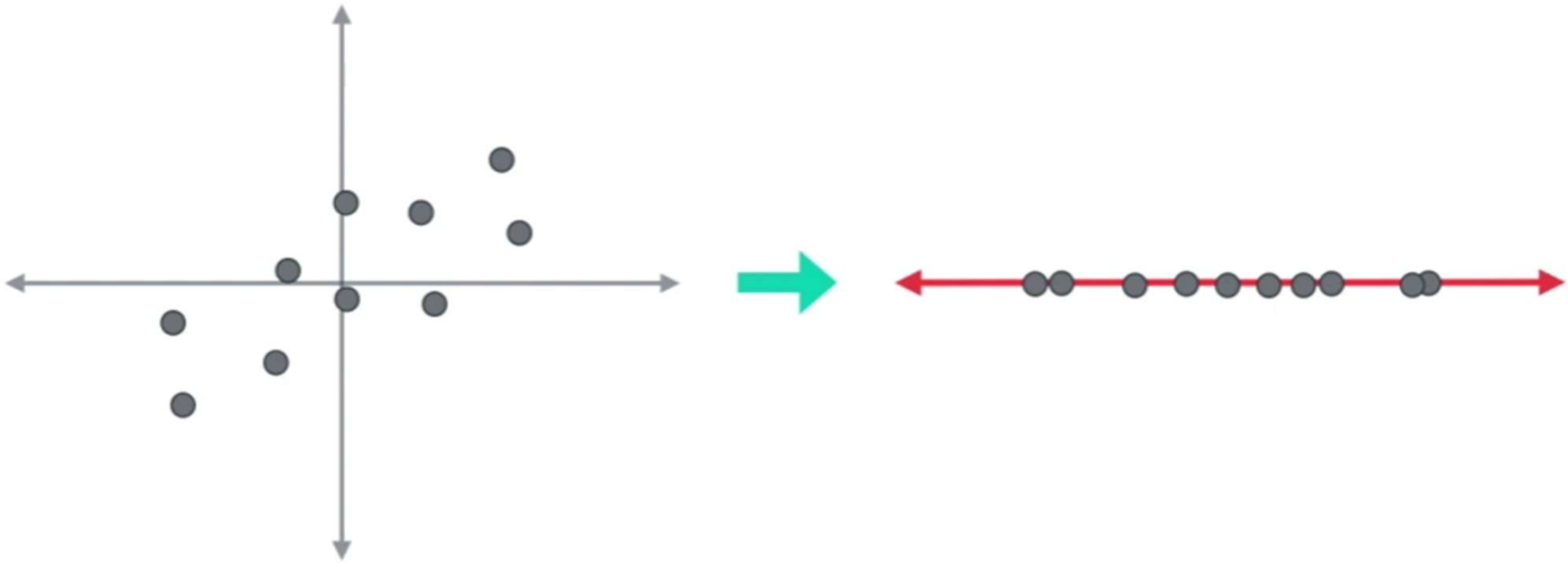
Valores propios
(magnitud)

Proyectar los datos





Proyectar los datos



¿Cómo funciona el PCA? Guía en 5 pasos

Tabla grande

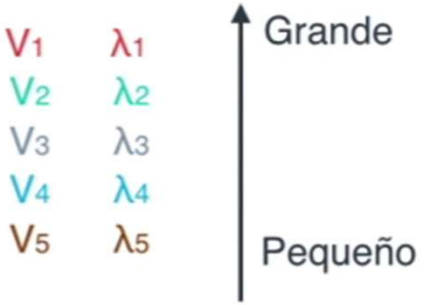
X_1	X_2	X_3	X_4	X_5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

Tabla grande

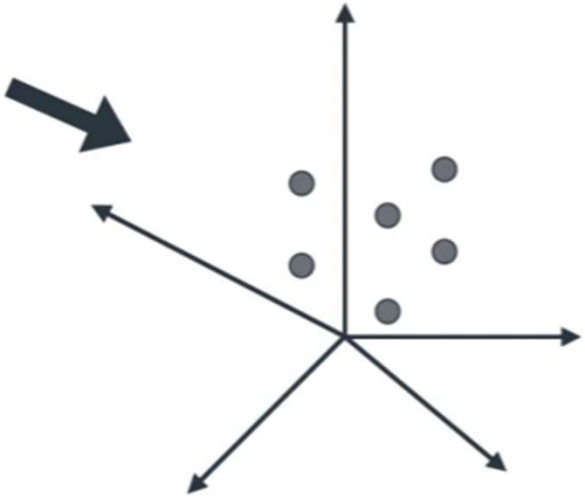
x_1	x_2	x_3	x_4	x_5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

Matriz de covarianza

Vectores y valores propios



V_1 λ_1
V_2 λ_2



5 dimensiones

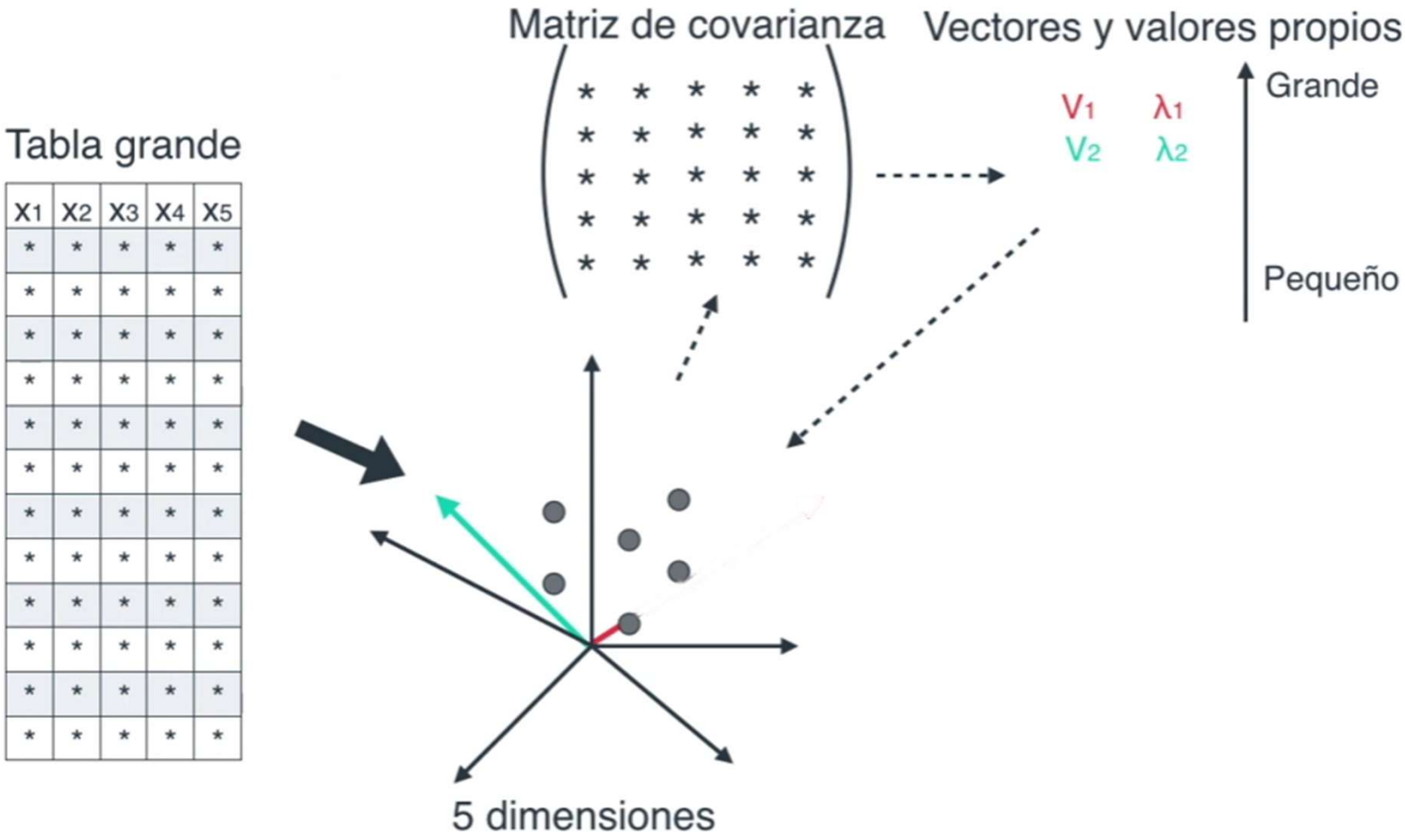


Tabla grande
X1 X2 X3 X4 X5
Matriz de covarianza
Vectores y valores propios
V1 λ1
V2 λ2
Grande
Pequeño
5 dimensiones

Tabla grande

x_1	x_2	x_3	x_4	x_5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

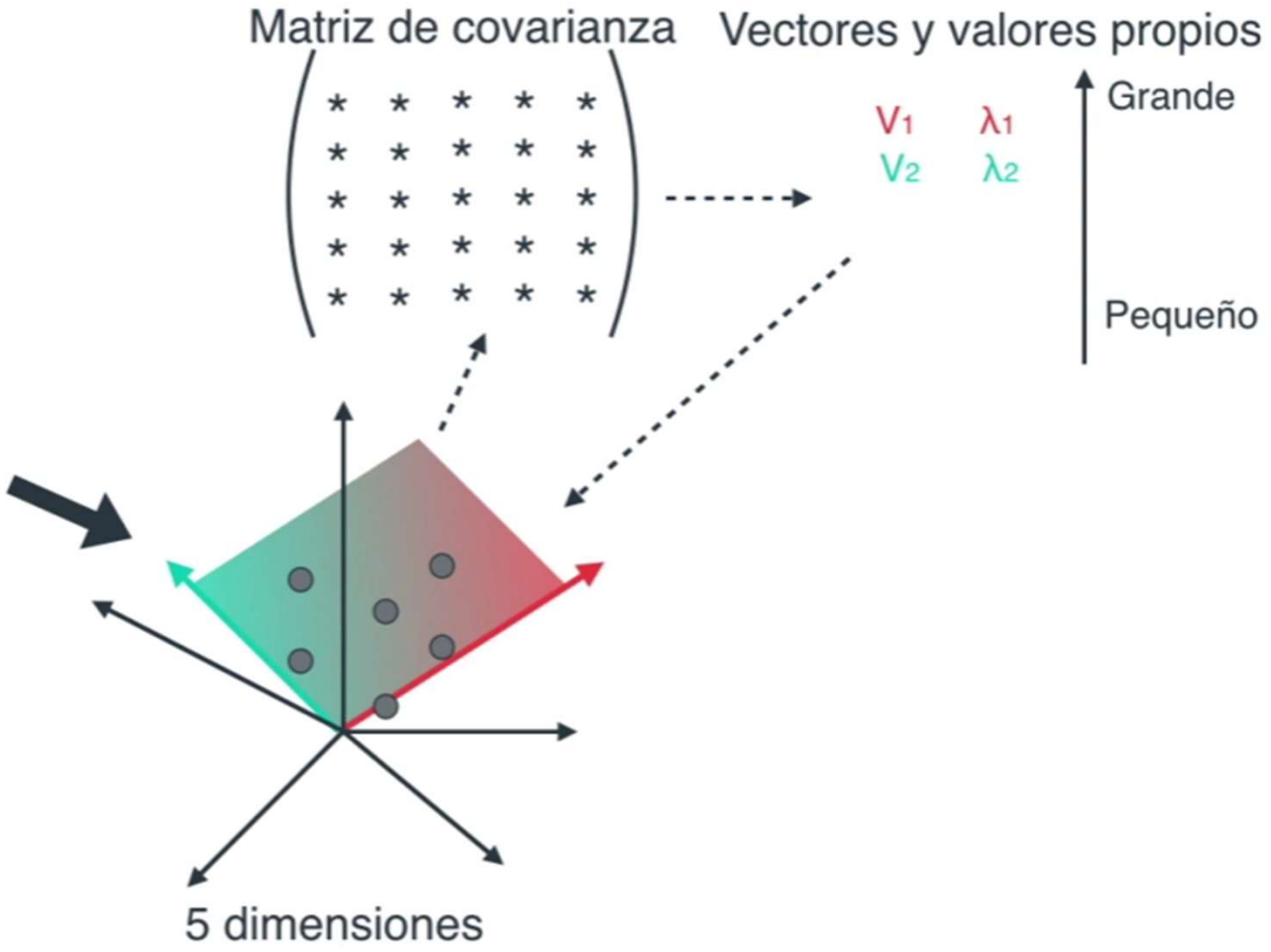
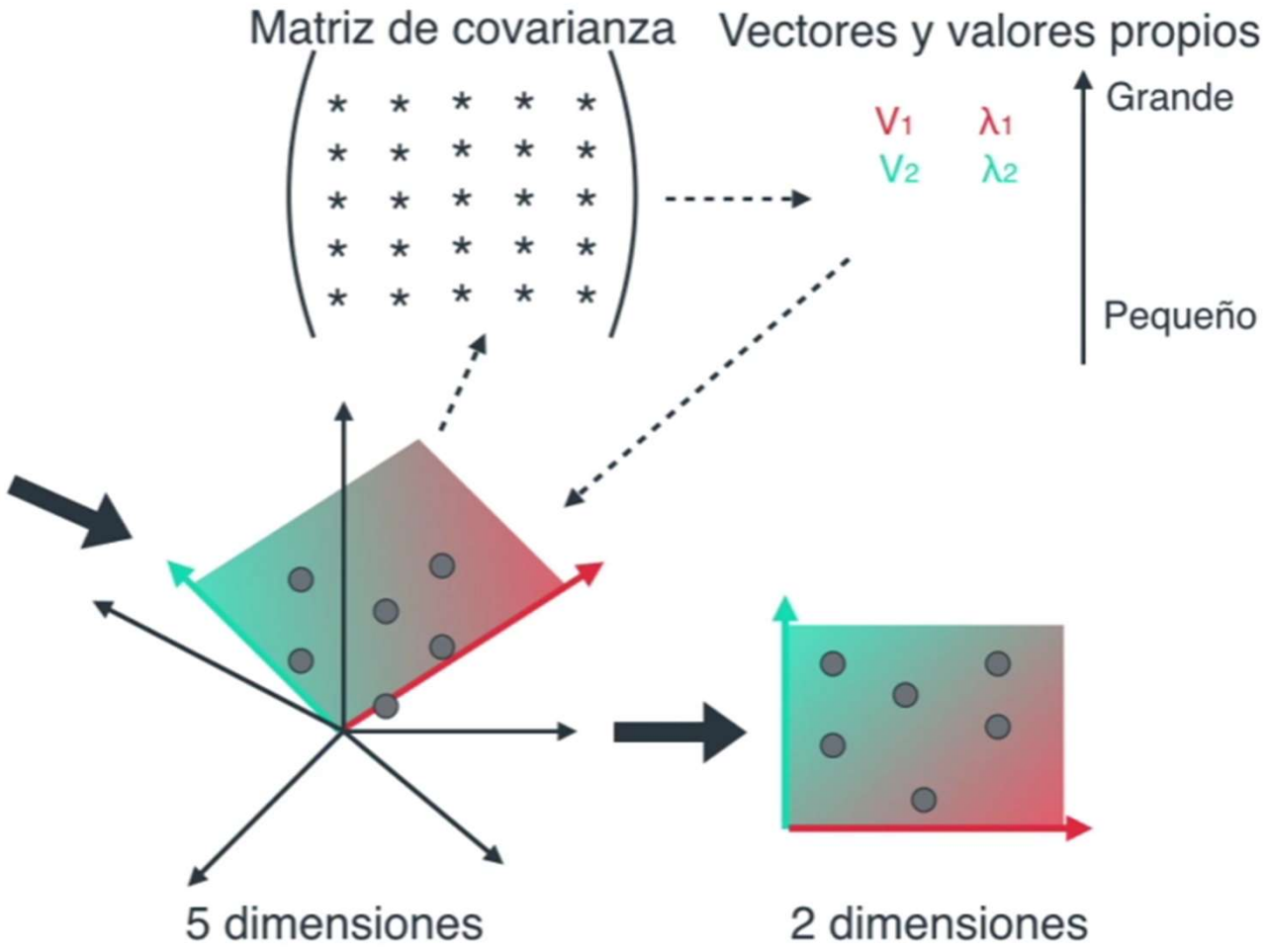


Tabla grande

X_1	X_2	X_3	X_4	X_5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*



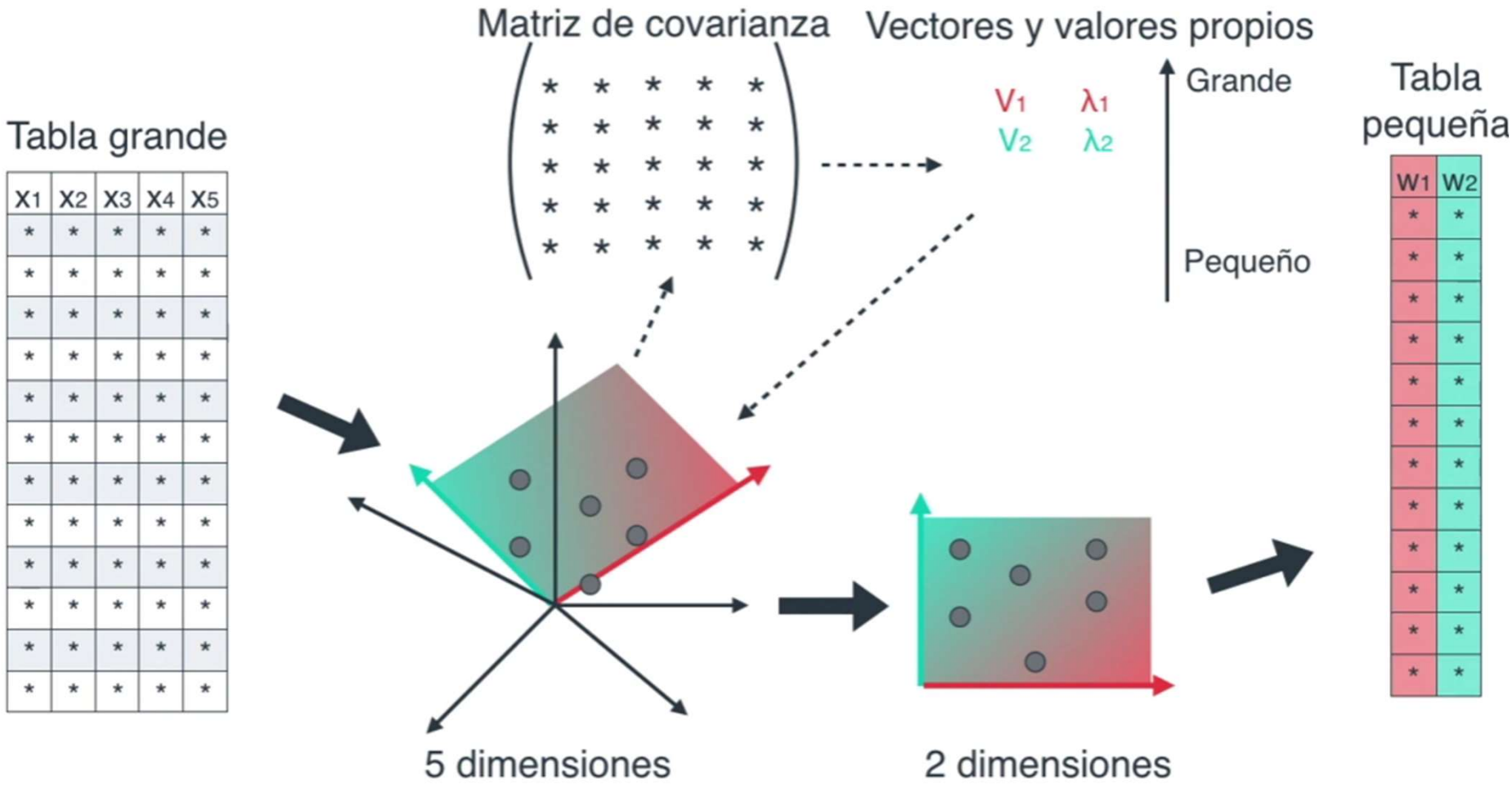
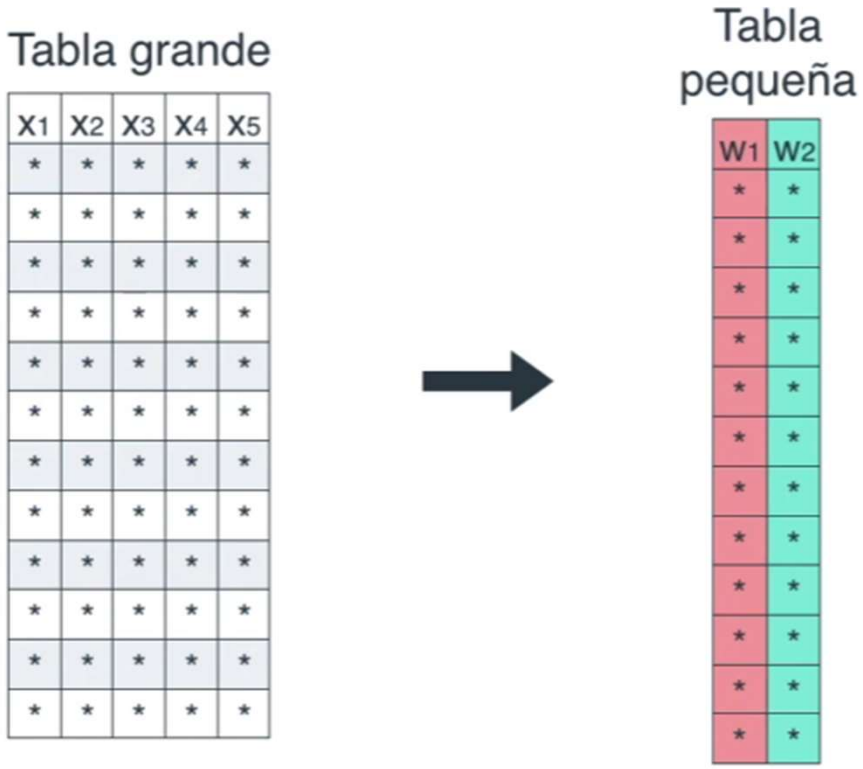


Tabla grande
X1 X2 X3 X4 X5
Matriz de covarianza
Vectores y valores propios
V1 λ1
V2 λ2
Grande
Pequeño
Tabla pequeña
W1 W2
5 dimensiones
2 dimensiones

Resultado



Ventajas y desventajas

Ventajas:
- Reducción de dimensionalidad
- eliminación de ruido
- mejora de la eficiencia computacional
- mejor visualización
- aplicaciones diversas.

Desventajas:
- Pérdida de interpretabilidad
- asume linealidad
- escalabilidad limitada en grandes datos
- sensible a la escala de las variables
- pérdida de información y problemas con datos faltantes.

Aplicaciones del análisis de componentes principales

Finanzas

Prever los precios de las acciones a partir de precios pasados es una noción utilizada en la investigación desde hace años. **El PCA puede utilizarse** para reducir la dimensionalidad y analizar los datos para ayudar a los expertos a encontrar componentes relevantes que expliquen la mayor parte de la variabilidad de los datos.

Puedes obtener más información sobre la **reducción de la dimensionalidad en R** en nuestro curso específico.

Procesamiento de imágenes

Una imagen está formada por varias características. El PCA se aplica principalmente en la compresión de imágenes para conservar los detalles esenciales de una imagen dada, reduciendo al mismo tiempo el número de dimensiones. Además, el PCA puede utilizarse para tareas más complicadas, como el reconocimiento de imágenes.

Aplicaciones del análisis de componentes principales

Sanidad

En la misma lógica de la compresión de imágenes. El PCA se utiliza en la imagen por resonancia magnética (MRI) para reducir la dimensionalidad de las imágenes con el fin de mejorar la visualización y el análisis médico. También puede integrarse en tecnologías médicas utilizadas, por ejemplo, para reconocer una determinada enfermedad a partir de imágenes.

Seguridad

Los sistemas biométricos utilizados para el reconocimiento de huellas dactilares pueden integrar tecnologías que aprovechan el análisis de componentes principales para extraer las características más relevantes, como la textura de la huella dactilar e información adicional.

Ejercicio

En un pequeño estudio de biología, se desea analizar las características de tres plantas diferentes. Se midieron dos variables para cada planta: la **altura** (en cm) y el **diámetro del tallo** (en mm). A continuación, se presentan los datos de las tres plantas:

Planta	Altura (cm)	Diámetro del tallo (mm)
1	15	6
2	17	8
3	16	7

Aplicar PCA a mano para reducir la dimensionalidad de los datos, encontrando los componentes principales. Luego, interpreta los resultados en términos de la relación entre las dos características medidas.

Visualización

- El diagrama de dispersión nos sirve para ver los valores de las observaciones respecto de los dos componentes principales y resaltamos las diferentes observaciones por su especie.
-

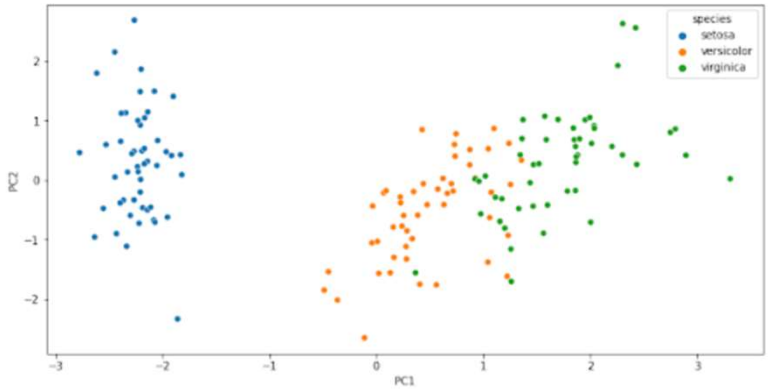


Figura 1. Diagrama de dispersión de las dos componentes principales. En azul las observaciones que corresponden a la especie setosa, naranja a versicolor y verde a virgínica.

Visualización

- El biplot es un gráfico que permite representar las variables del dataset original y las observaciones transformadas en los ejes de los dos componentes principales. Las flechas representan las variables originales y es importante hacia dónde apuntan

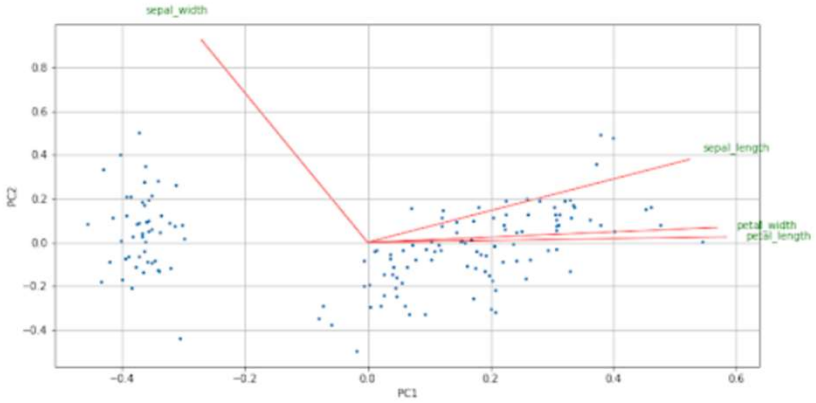


Figura 2. Biplot de los componentes principales. Los puntos son las observaciones transformadas en los ejes de los dos componentes principales y las flechas representan las variables originales.

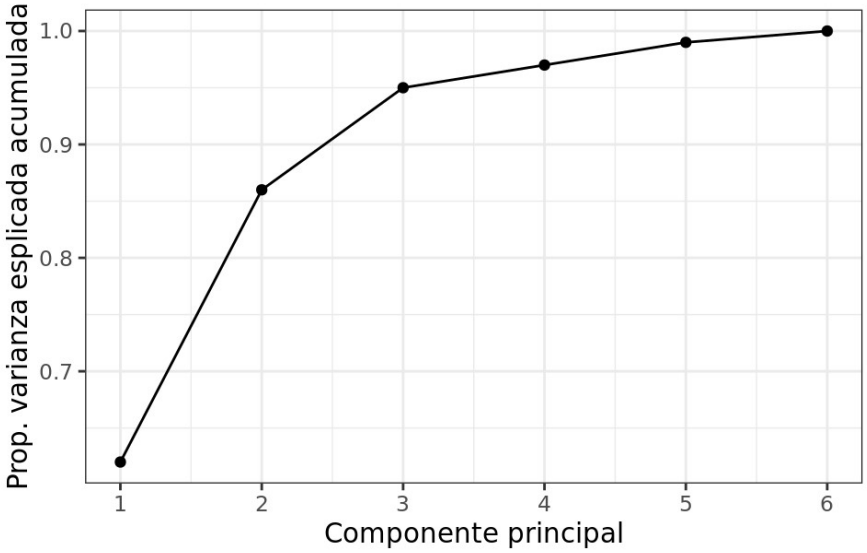
Número óptimo de componentes principales

- Por lo general, dada una matriz de datos de dimensiones *n* x *p*, el número de componentes principales que se pueden calcular es como máximo de *n*-1 o *p* (el menor de los dos valores es el limitante)

No existe una respuesta o método único que permita identificar cual es el número óptimo de componentes principales a utilizar.

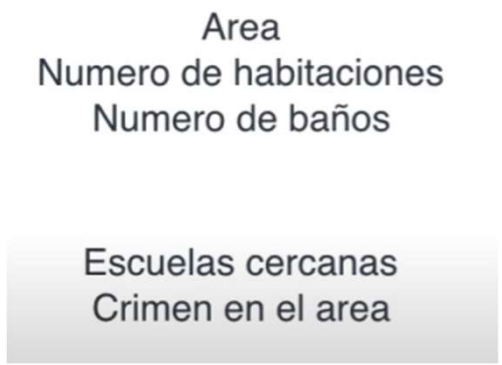
Número óptimo de componentes principales

- Una forma de proceder muy extendida consiste en evaluar la proporción de varianza explicada acumulada y seleccionar el número de componentes mínimo a partir del cual el incremento deja de ser sustancial.



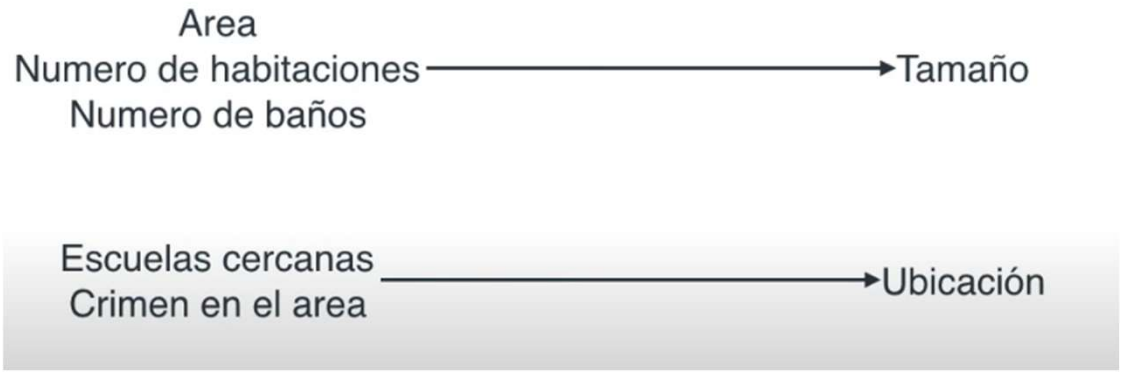
Conceptos previos

- Reducción de dimensionalidad



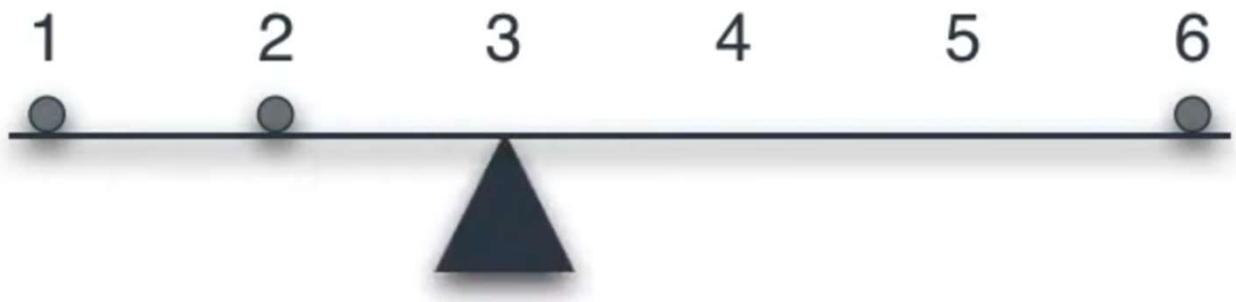
Conceptos previos

- Reducción de dimensionalidad



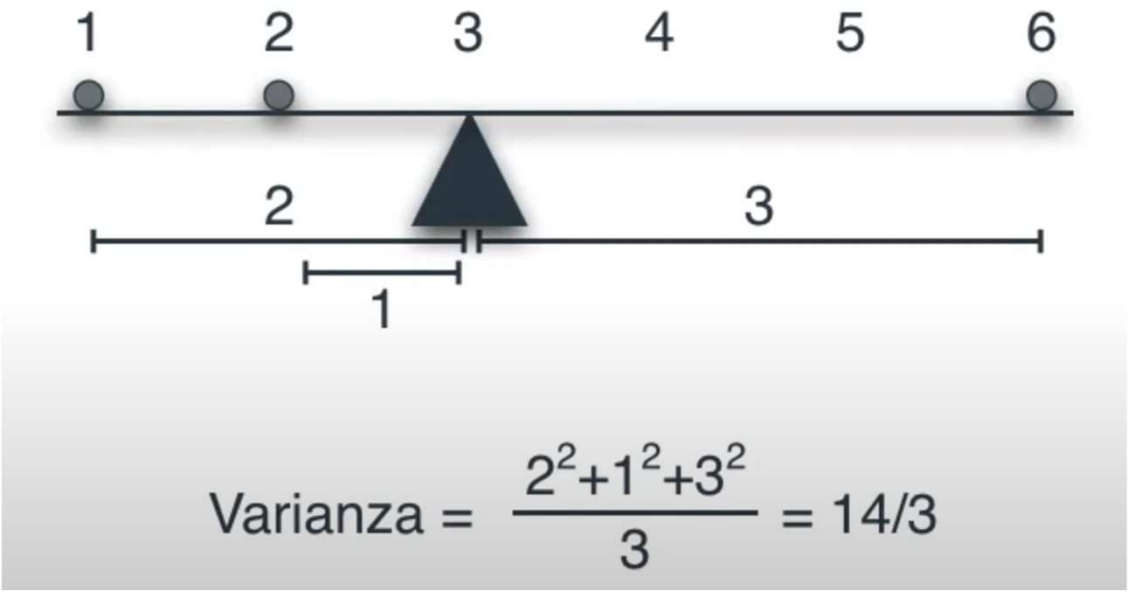
Conceptos previos

- Reducción de dimensionalidad
- Promedio



Conceptos previos

- Reducción de dimensionalidad
- Promedio
- Varianza



Conceptos previos

- Reducción de dimensionalidad