Análisis de Componentes Principales

Dr. Cs. Miguel Pari Soto y Dr. Cs. Claver Pari Soto

(PCA - Principal Component Analysis)

El **Análisis de Componentes Principales (PCA)** y la **Descomposición en Valores Singulares (SVD)** son técnicas fundamentales en álgebra lineal y machine learning que permiten:

- **Reducir la dimensionalidad** de datos complejos
- **Comprimir información** manteniendo la esencia de los datos
- **Identificar patrones** ocultos en conjuntos de datos
- **Optimizar el almacenamiento** y procesamiento de información

Introducción - Matemática

Producto punto de vectores

Sean los vectores $\vec{x}$ y $\vec{y}$ con m componentes

$$\vec{x} = [x_1, x_2, \ldots, x_m]$$

$$\vec{y} = [y_1, y_2, \ldots, y_m],$$

entonces el producto punto se define como

$$\vec{x} \cdot \vec{y} = x_1 y_1 + x_2 y_2 + \ldots + x_m y_m$$

Matrices

La suma de matrices se hace sumando los elementos correspondientes de las dos matrices del mismo tamaño

$$A = \begin{bmatrix} -1 & 2 & 4 \\ 2 & -3 & -1 \end{bmatrix}_{2\times 3} \quad B = \begin{bmatrix} 3 & 4 \\ 1 & 7 \\ -1 & 1 \end{bmatrix}_{3\times 2}$$

$$AB = \begin{bmatrix} [-1 \;\; 2 \;\; 4] \cdot [3 \;\; 1 \;\; -1] & [-1 \;\; 2 \;\; 4] \cdot [4 \;\; 7 \;\; 1] \\ [2 \;\; -3 \;\; -1] \cdot [3 \;\; 1 \;\; -1] & [2 \;\; -3 \;\; -1] \cdot [4 \;\; 7 \;\; 1] \end{bmatrix}$$

$$AB = \begin{bmatrix} -5 & 14 \\ 4 & -14 \end{bmatrix}_{2\times 2}$$

$$AB = A \times B$$

Multiplicación de vectores (o matrices) por escalar

$$2\begin{bmatrix} 2 & -4 \end{bmatrix} = \begin{bmatrix} 4 & -8 \end{bmatrix}$$

$$3\begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 3 \\ -6 \end{bmatrix}$$

$$-2\begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} -2 & 4 \\ -6 & -8 \end{bmatrix}$$

Aplicaciones

- En procesamiento de señales, una señal de tiempo discreto con n muestras puede representarse como un vector en el espacio vectorial $\mathbb{R}^n$
- En control, estados del sistema pueden representarse en espacios vectoriales de dimensión igual al número de variables de estado.

Espacio vectorial

- Es un conjunto de vectores donde cualquier combinación lineal de vectores de ese conjunto también pertenece al espacio.
- Una combinación lineal se hace multiplicando escalares a los vectores y sumando los vectores resultantes, de la forma $a\vec{x} + b\vec{y}$, donde a y b son escalares

Base de espacio vectorial

- Es un conjunto mínimo de vectores linealmente independientes que generan todo el espacio.
- Ningún conjunto con menos vectores podrá generar el espacio.
- Ejemplo: $[1, 0]$ y $[0, 1]$ forman una base para el espacio 2-D. Cualquier $[x, y]$ en el espacio 2-D se puede escribir como $[x, y] = x[1, 0] + y[0, 1]$

Autovalores y autovectores

- Un autovector es un vector $\vec{x}$ distinto de cero que satisface $A\vec{x} = \lambda\vec{x}$
- λ es el autovalor correspondiente

Sean

$$A = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} \quad \text{y} \quad \vec{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \implies A\vec{x} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}; \qquad \vec{x} \text{ no es un autovector}$$

Por otro lado, si

$$A = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} \quad \text{y} \quad \vec{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \implies A\vec{x} = \begin{bmatrix} 4 \\ 4 \end{bmatrix} = 4\vec{x} = \lambda\vec{x}; \qquad \vec{x} \text{ es un autovector}$$

y $\lambda = 4$ es su autovalor correspondiente

Autovalores y autovectores

- Una matriz M puede ser descompuesta usando autovectores
- Los autovectores forman una base
- Los autovalores más grandes son los más *influyentes*
- Podemos reducir la dimensionalidad ignorando los autovalores *pequeños*

Como calcular los autovectores e autovalores

Estadística

Sean dos vectores

$$\vec{x} = [x_1 \; x_2 \; \dots \; x_n] \text{ y } \vec{y} = [y_1 \; y_2 \; \dots \; y_n]$$

Media:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Varianza:

$$\sigma_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

Covarianza:

$$\text{cov}(\vec{x}, \vec{y}) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n-1}$$

Estadística - Algunas propiedades

- La *media* es una medida de tendencia central
- La *varianza* mide la dispersión de los datos respecto a la media
- La *covarianza* relaciona las variaciones en $\vec{x}$ con las de $\vec{y}$, en relación con sus respectivas medias

Covarianzas

Las cosas se simplifican cuando las medias son 0

- Si $\vec{x} = [-1, 2, 1, -2]$ y $\vec{y} = [1, -1, 1, -1]$

$$\implies \text{cov}(\vec{x}, \vec{y}) = \frac{(-1)(1) + (2)(-1) + (1)(1) + (-2)(-1)}{3} = \frac{0}{3} = 0$$

- Si $\vec{x} = [-1, 2, 1, -2]$ y $\vec{y} = [-1, 1, 1, -1]$

$$\implies \text{cov}(\vec{x}, \vec{y}) = \frac{(-1)(-1) + (2)(1) + (1)(1) + (-2)(-1)}{3} = \frac{6}{3} = 2$$

- El signo de covarianza es la pendiente de la relación
- La covarianza igual a 0 implica no correlacionamiento

Matriz de Covarianza

Sea

$$M_{m\times n} = \begin{bmatrix} \text{altura}_1 & \text{altura}_2 & \dots & \text{altura}_n \\ \text{peso}_1 & \text{peso}_2 & \dots & \text{peso}_n \end{bmatrix}$$

y

$$C_{m\times m} = \{c_{ij}\} = \frac{1}{n-1} M M^T$$

Para esta definición de matriz de covarianza, la media de cada tipo de medición deve ser 0

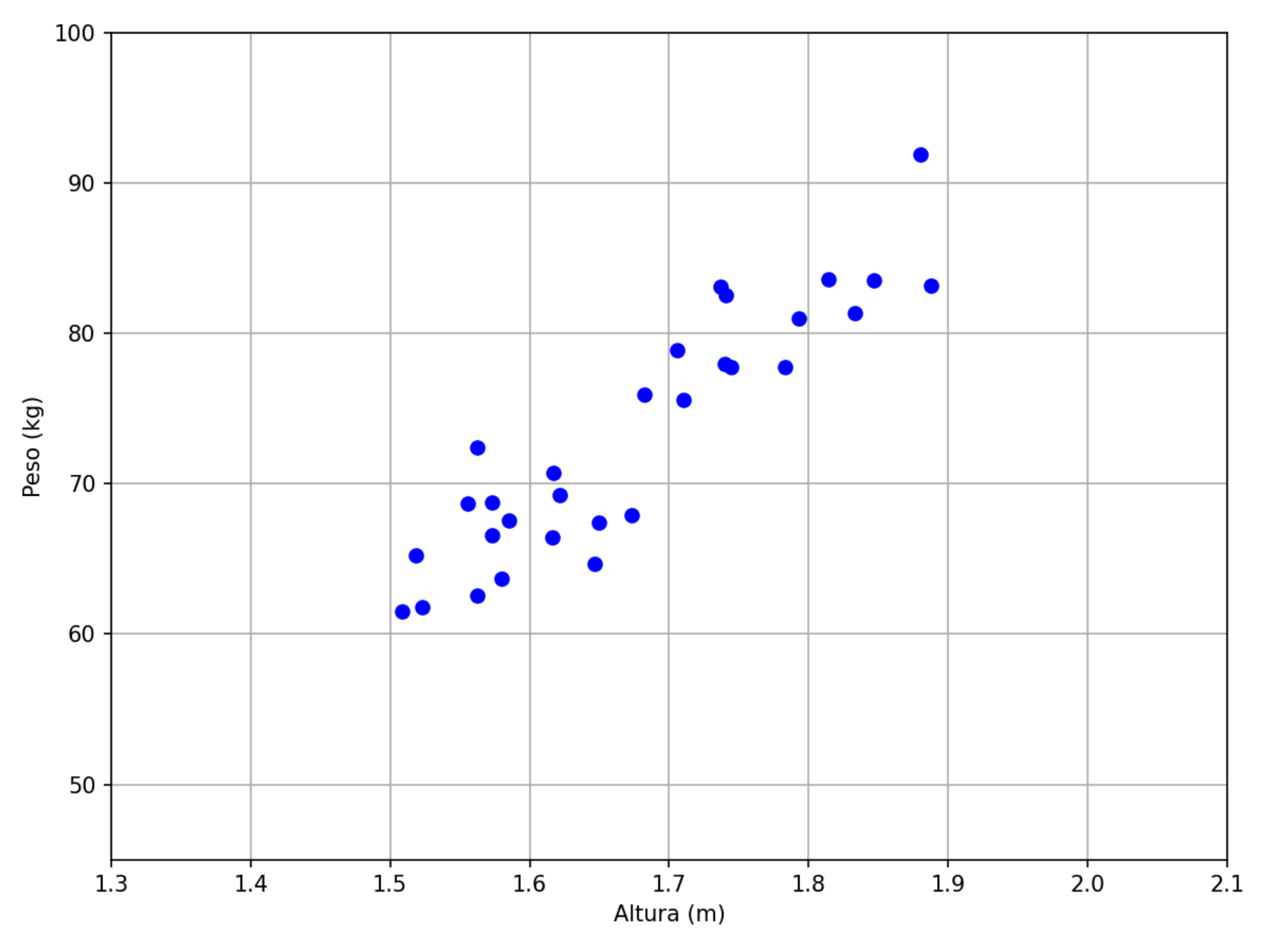
Matriz de Covarianza (continuación)

- Elementos en la diagonal de C
    - Por lo general, las grandes variaciones son las más interesantes
    - Lo ideal: algunos grandes y la mayoría pequeños
- Elementos fuera de la diagonal de C
    - Covarianza entre tipos de medición
    - Si $\text{cov} = 0 \implies \text{no correlacionada}$
    - Si $\text{cov} \neq 0 \implies \text{redundancia}$
    - Lo ideal: que los elementos fuera de la diagonal sean todos 0

Análisis de Componentes Principales

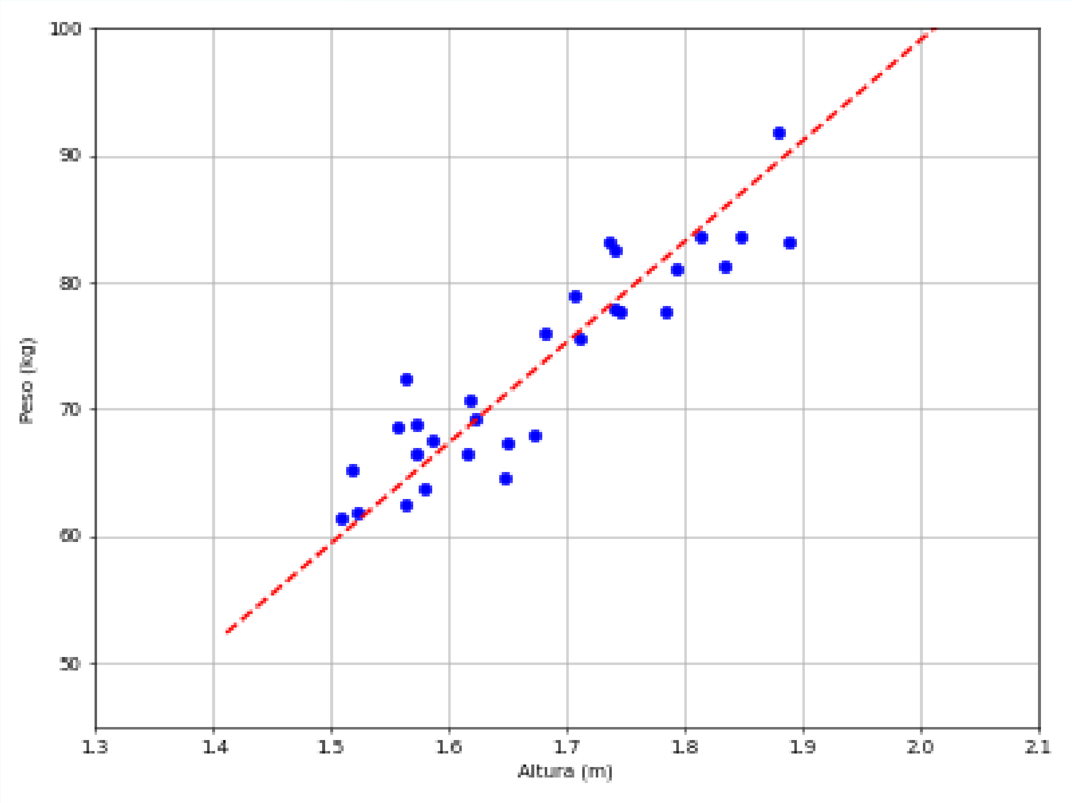
- PCA es una poderosa técnica de aprendizaje automático basada en métodos de álgebra lineal.
- Encuentra las dimensiones más significativas en un conjunto de datos. Reduce la dimensionalidad del problema con una pérdida mínima de información.
- La descomposición en valores singulares (SVD) es la forma más común de implementar el PCA.

La base vetorial original no es la más informativa



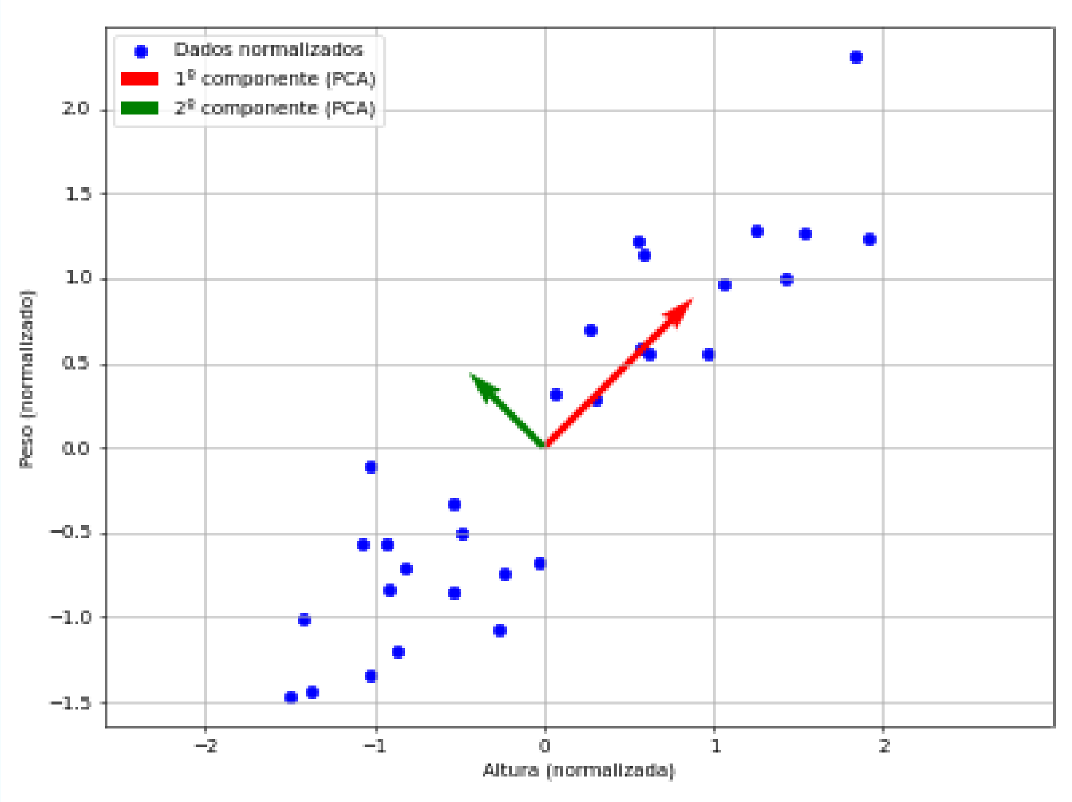
Alinear la base vectorial con los datos

- Línea roja es la dirección de la varianza máxima
- Reduce el "ruido"



Análisis de Componentes Principales (PCA)

- La estructura de los datos define la dirección de los vectores
- La longitud cuantifica la varianza en cada dirección



Idea básica del PCA

PCA alinea la base con las varianzas. Diagonalizando la matriz de covarianza C.

Para la diagonalización:

1. Elegir la dirección (dimensión) con desviación máxima
2. Encontrar la dirección con la varianza máxima que sea ortogonal a todas las direcciones seleccionadas anteriormente
3. Volver a 2 (hasta que no queden dimensiones)

Los vectores resultantes son los componentes principales

Intuición detrás del PCA

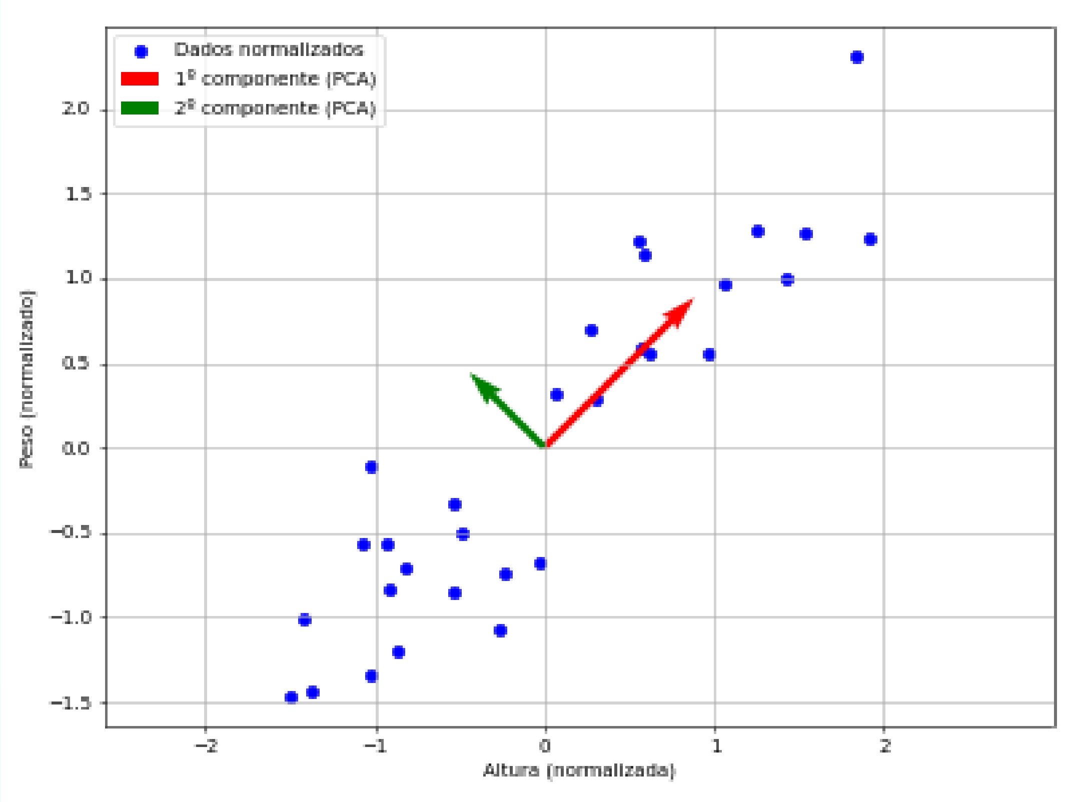
- Para conocer una ciudad, no es necesario visitar todas las calles
- Podemos reducir la dimensionalidad del problema

Hipótesis fuertes del PCA

1. Linealidad
   - El cambio de base es una operación lineal
   - Pero, algunos procesos son inherentemente no lineales
2. Grandes varianzas son más *interesantes*
   - La varianza grande es *señal*, la pequeña es *ruido*
   - Pero, puede no ser válido para algunos problemas
3. Los componentes principales son ortogonales
   - Hace que el problema se resuelva de manera eficiente
   - Pero, no ortogonal es mejor en algunos casos

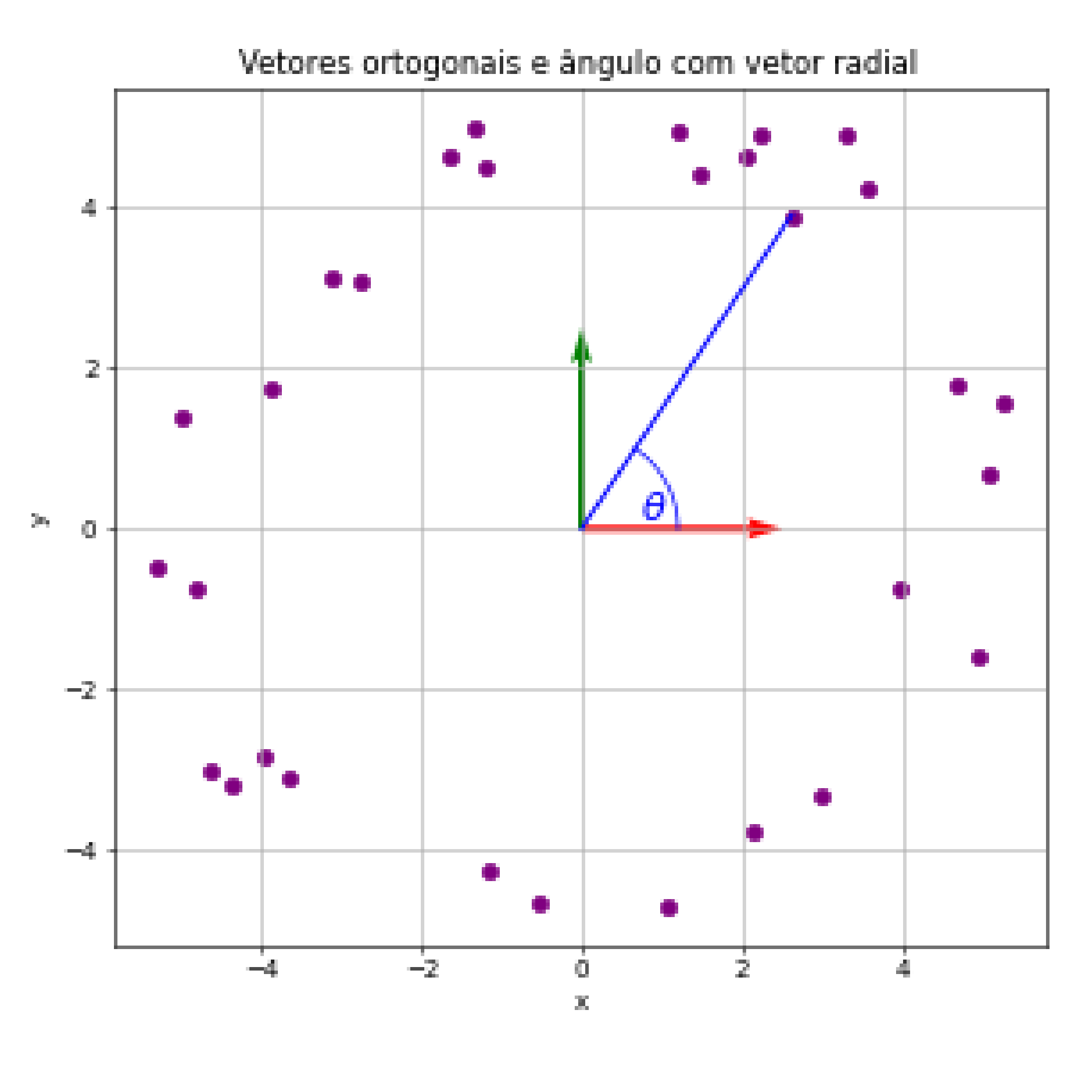
Éxito del PCA

- El vector rojo más largo es la *señal*. Más informativo que el vector corto
- Podemos ignorar el vector corto. Puede ser considerado "ruido"
- Reduce el problema de 2-D a 1-D



PCA fallará cuando

- Datos distribuidos en formato de rueda. Resultados de PCA inútiles
  - El ángulo θ tiene toda la información
  - Pero θ es no lineal en relación a la base (x, y)
- Datos con padrones complejos que no pueden ser descritos por un modelo linear
- PCA asume linealidad



Resumen de PCA

- Organizar los datos en una matriz centralizada $M_{m \times n}$
  - Donde n es el número de *experimentos*
  - m *mediciones* por experimento
- Formar la matriz de covarianza $C_{m \times m} = \frac{1}{n-1} M M^T$
- Calcular los autovalores y autovectores de esa matriz de covarianza C. Cada autovector es de dimensión m
- Formar la matriz de autovectores V donde cada columna es un autovector
- El producto $M^T V$ es la matriz transformada de los datos, con matriz de covarianza diagonal

Descomposición en Valores Singulares (SVD - Singular Values Decomposition)

- Es una forma elegante de encontrar los autovectores
- Sea la matriz de datos a ser analisados: $M_{m \times n}$
- SVD descompone la matriz como

  $$M = U\Sigma V^T$$

Esto funciona en un entorno muy general.

SVD es una técnica general para encontrar componentes principales.

SVD y PCA

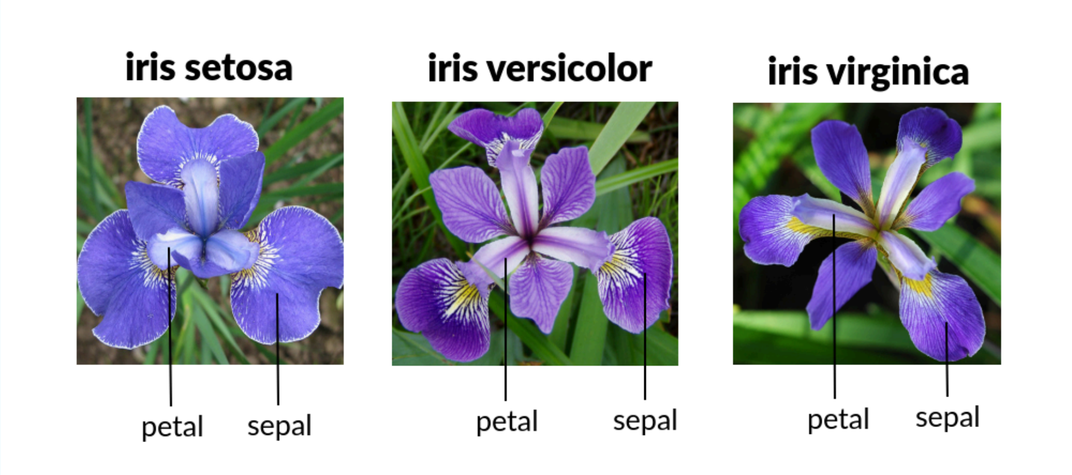
- SVD es una (mejor) manera de hacer PCA
- Una forma de calcular autovectores
- En la SVD, $M = U\Sigma V^T$, donde
  - Las columnas de U contienen los autovectores de MM^T
  - Las columnas de V contienen los autovectores de M^TM
  - Σ es diagonal y sus elementos son los valores singulares. Cada valor singular es la raiz cuadrada de uno de los autovalores.

Ejemplo

Tenemos una matriz de datos $X \in \mathbb{R}^{30\times 4}$. Cuatro medidas para treinta muestras :

$$X = \begin{bmatrix} 5.7 & 2.8 & 4.5 & 1.3 \\ 5.0 & 3.4 & 1.5 & 0.2 \\ 6.4 & 3.2 & 4.5 & 1.5 \\ 5.1 & 3.5 & 1.4 & 0.2 \\ 6.3 & 2.9 & 5.6 & 1.8 \\ 4.9 & 2.4 & 3.3 & 1.0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$$Y = \begin{bmatrix} 1 & 0 & 1 & 0 & 2 & 1 & \cdots \end{bmatrix}$$

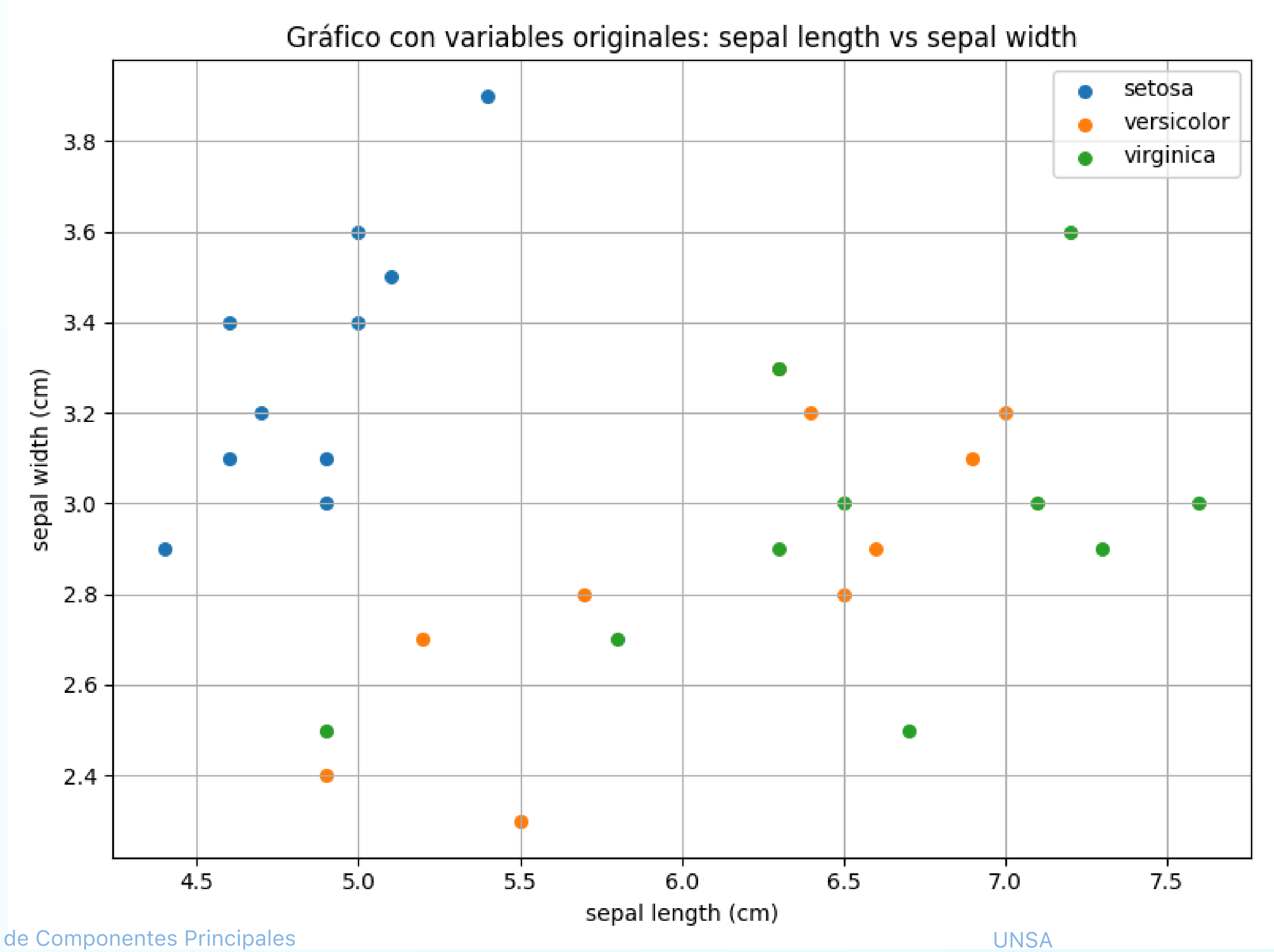


$col0$ y $col1$: longitud y ancho del sépalo

$col2$ y $col3$: longitud y ancho del pétalo

En Y: 0: setosa; 1: versicolor; 2: virginica

Dataset Iris con 30 muestras. Gráfico de dispersión segun la longitud y el ancho de los sépalos



1. Cargar las librerias y los datos

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.decomposition import PCA

X, Y = load_iris(return_X_y=True)
```

- $X \in \mathbb{R}^{30\times 4}$: matriz de datos con 30 muestras y 4 características.
- $Y \in \{0, 1, 2\}^{30}$: etiquetas de clase (no se usan para el PCA).

2. Paso interno clave: centrado de datos

Antes de aplicar la descomposición, los datos se **centran**, es decir, se le resta la media a cada columna:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_{i,:} \quad \text{(media por columna)}$$

$$X_c = X - \bar{x} \quad \text{(matriz centrada)}$$

donde

$$X_c \in \mathbb{R}^{30 \times 4}$$

La librería `scikit-learn` implementa el PCA usando la **descomposición SVD** de la matriz centrada X_c:

$$X_c = U \Sigma V^T$$

Donde:

- $U \in \mathbb{R}^{30 \times 4}$: matriz de vectores singulares izquierdos (no la guarda `scikit-learn`)
- $\Sigma \in \mathbb{R}^{4 \times 4}$: matriz diagonal con los valores singulares $\sigma_1, \sigma_2, \sigma_3, \sigma_4$
- $V \in \mathbb{R}^{4 \times 4}$: matriz cuyos **filas son los componentes principales**

Proyección a 2 dimensiones

Como pedimos solo 2 componentes principales:

Se toman las primeras dos columnas de V^T, o sea, las primeras dos **filas** de V

$$V_2 \in \mathbb{R}^{2\times 4}$$

V_2 contiene los dos vectores principales (componentes).

La nueva representación de los datos se obtiene proyectando X_c sobre estos vectores:

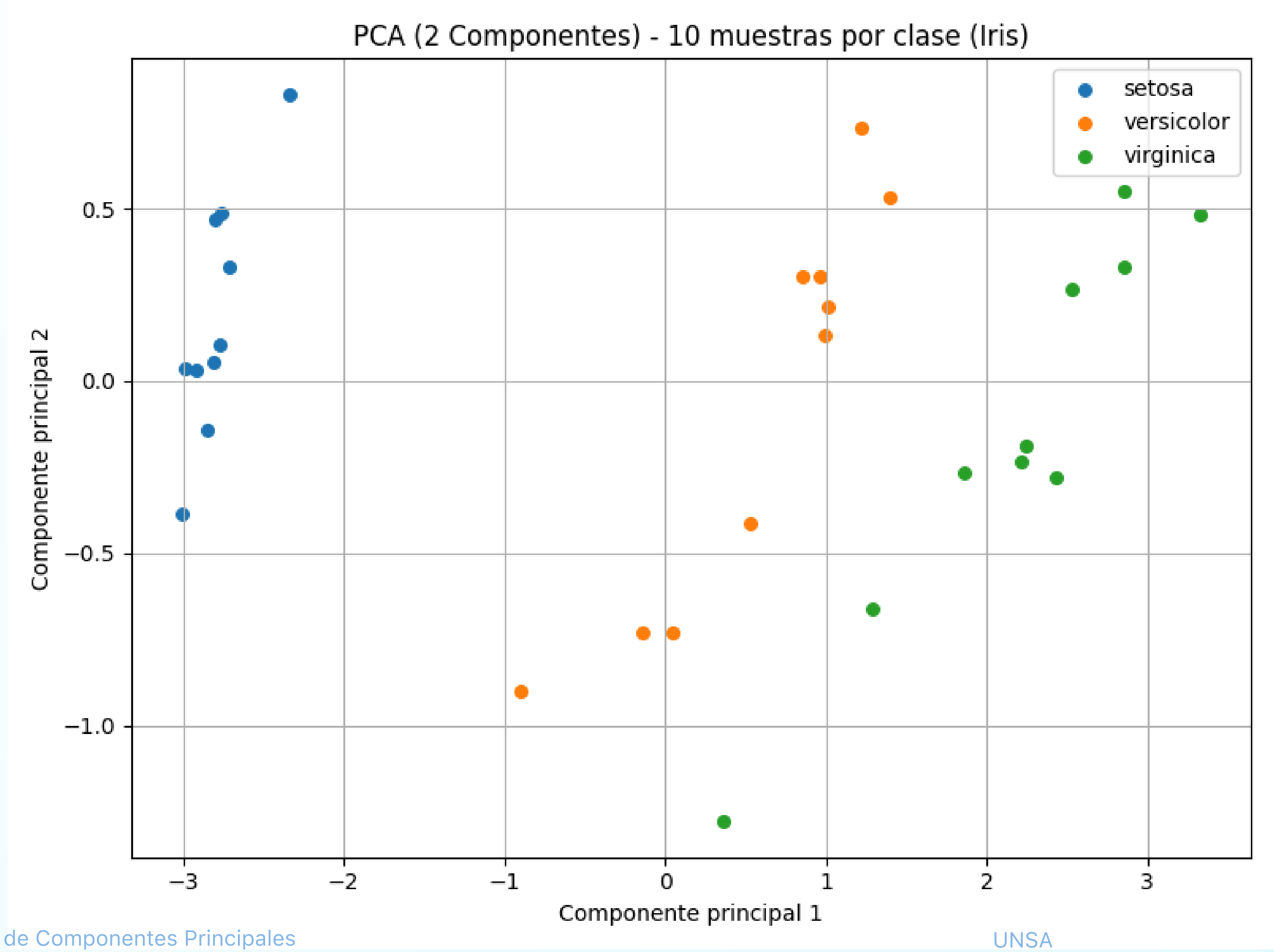
$$X_{\mathrm{PCA}} = X_c \cdot V_2^T \in \mathbb{R}^{30\times 2}$$

3. PCA con 2 componentes (asumiendo que X es el X_c)

```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)
```

Esto realiza un PCA y transforma los datos de $X \in \mathbb{R}^{30\times 4}$ a una representación de 2 dimensiones $X_{\text{PCA}} \in \mathbb{R}^{30\times 2}$

Dataset Iris con 30 muestras. Gráfico de dispersión segun las dos dimensiones encontradas por el PCA



Fin