5 Principal Component Analysis

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica estadística utilizada para la reducción de dimensionalidad de conjuntos de datos, preservando la mayor cantidad de información posible. A diferencia de otros métodos de reducción, PCA busca direcciones de máxima varianza, permitiendo representar los datos en un nuevo sistema de coordenadas más compacto.

5.1. Definición

Formalmente, PCA transforma un conjunto de datos correlacionados en un nuevo conjunto de variables no correlacionadas llamadas **componentes principales**. Cada componente principal es una combinación lineal de las variables originales y representa una dirección de máxima varianza en los datos.

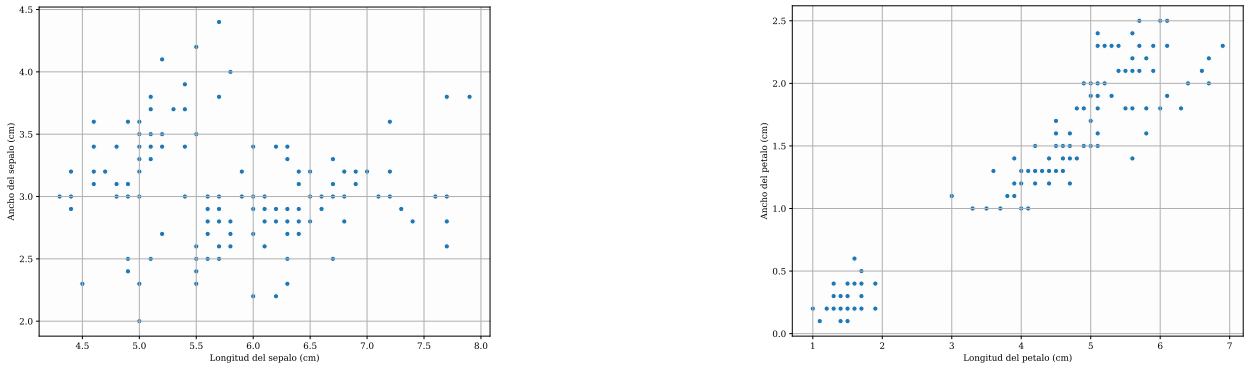
Matemáticamente, dado un conjunto de datos $X \in \mathbb{R}^{n \times p}$, donde n es el número de muestras y p el número de variables, el procedimiento para aplicar PCA es el siguiente:

1. **Estandarización de los datos:** Centrar los datos restando la media de cada variable.
2. **Cálculo de la matriz de covarianza:**
$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^\top \mathbf{X}$$
3. **Obtención de eigenvectores y eigenvalores:** Resolver el problema de autovalores
$$\mathbf{C}\mathbf{v} = \lambda \mathbf{v}$$
donde λ son los eigenvalores (varianzas explicadas) y $\mathbf{v}$ los eigenvectores (direcciones principales).
4. **Ordenamiento:** Clasificar los eigenvectores en orden descendente según sus eigenvalores.
5. **Proyección:** Transformar los datos originales sobre el subespacio generado por los primeros k componentes:
$$\mathbf{Z} = \mathbf{X}\mathbf{V}_k$$
donde $\mathbf{V}_k$ contiene los primeros k eigenvectores.

Cada componente principal puede interpretarse como una nueva dimensión que captura una parte de la variabilidad total de los datos. El primer componente explica la mayor cantidad de varianza posible, el segundo explica la mayor cantidad de varianza restante, y así sucesivamente.

Es importante recalcar que este método al emplear principalmente **operaciones matriciales** tiene un costo computacional sustancialmente menor con respecto a otros métodos de reducción de dimensionalidad, pero esto viene con una limitación importante: PCA es un método lineal, lo que significa que para relaciones no lineales la interpretación de la conexión entre los componentes puede ser compleja; para solucionar esto existen extensiones como **Kernel PCA** o técnicas contemporáneas como **Uniform Manifold Approximation (UMAP)**.

Tomemos como ejemplo el ya conocido Iris Dataset [2] un conjunto de datos sobre la morfología de tres especies de flores de lirio, donde se midieron tanto el ancho como el largo de los sépalos y pétalos de 50 flores distintas. Supongamos que tenemos estos datos sin conocer la especie individual

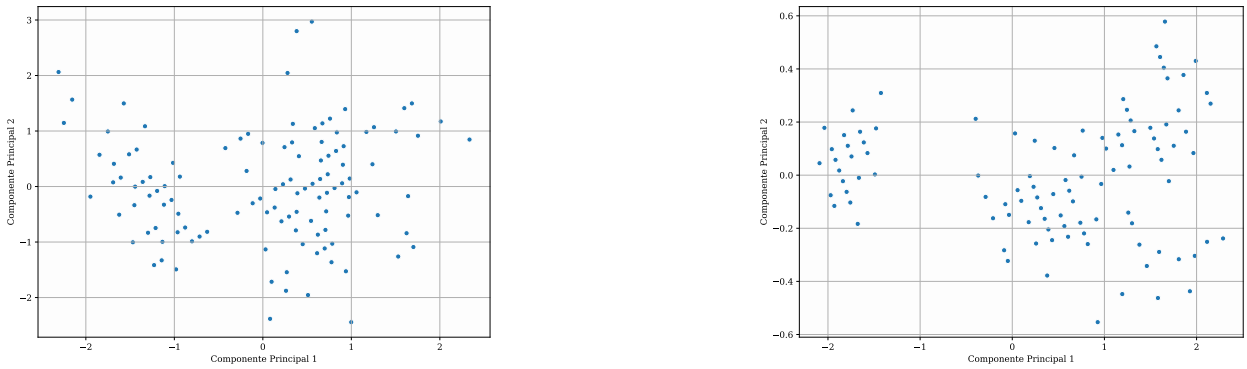


(a) Longitudes de Sépalo **(b)** Longitudes de Pétalo

Figura 5.1: Diagramas de dispersión de las categorias de datos

de flor; al realizar un diagrama de dispersión de estos datos obtenemos el resultado mostrado en la figura 5.1.

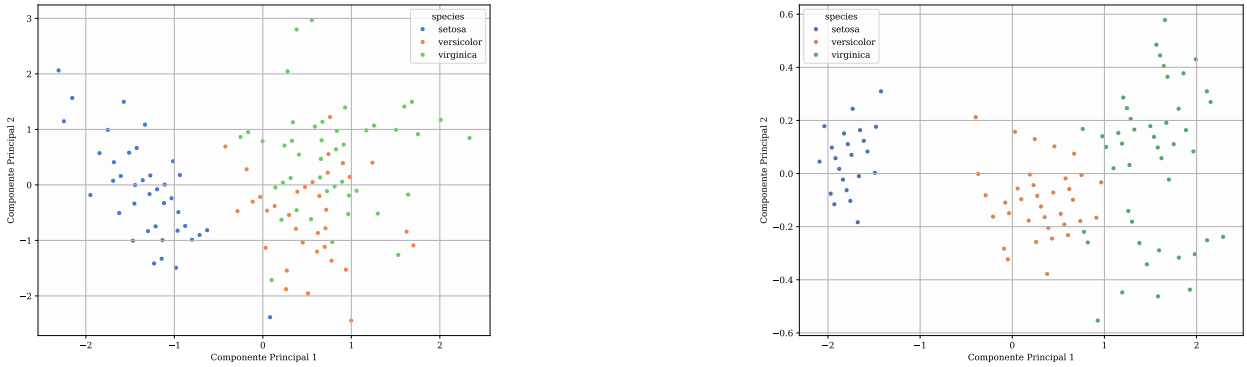
Podemos notar que, en esta representación inicial, resulta difícil identificar si existe alguna correlación entre las variables que nos permita clasificar las flores según su especie. Sin embargo, al aplicar Análisis de Componentes Principales (PCA) sobre las variables del sépalo y del pétalo, obtenemos representaciones más informativas como las mostradas en la figura 5.3, donde la separación entre especies es más evidente.



(a) PCA de Longitudes de Sépalo **(b)** PCA Longitudes de Pétalo

Figura 5.2: Diagramas de dispersión tras aplicar PCA a las categorías de datos

Esto se aprecia de mejor manera al agregar colores que representan la especie a la que corresponde cada flor, mostrados en la figura



(a) PCA de Longitudes de Sépalo

(b) PCA Longitudes de Pétalo

Figura 5.3: Diagramas de dispersión tras aplicar PCA a las categorías de datos con especies

Referencias

[1] Christopher Torrence and Gilbert P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998.

[2] R. A. Fisher. Iris. UCI Machine Learning Repository, 1936. DOI: https://doi.org/10.24432/C56C76.