Métodos lineales. El análisis de componentes principales

El análisis de componentes principales (*Principal Component Analysis*, PCA) es una técnica que busca sintetizar un conjunto de datos de modo que su información y estructura de dependencia pueda representarse por medio de un nuevo conjunto de variables: las componentes. Estas no son observables y constituyen un nuevo sistema de coordenadas que se construye mediante transformación lineal de las variables originales. Los nuevos ejes se interpretan como los factores responsables de la variación observada (ver Fig. 1).

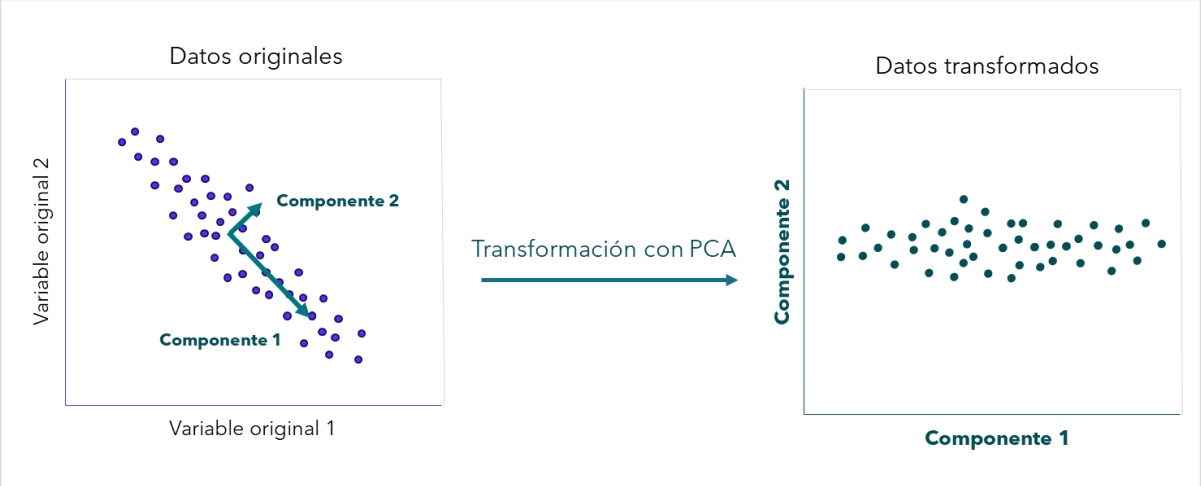

Fig. 1. Ejemplo de la transformación obtenida con análisis de componentes principales.

El PCA se basa en la premisa de que, si entre p variables originales existen relaciones de dependencias, estas pueden aprovecharse para "condensar" la información en m nuevas variables no correlacionadas ($m<p$), que expliquen la mayor variación del conjunto de datos. A diferencia de las técnicas de aprendizaje de *manifold*, el PCA no tiene como objetivo modelar las estructuras subyacentes en los datos. El PCA es un método de proyección a partir de transformaciones lineales para reducir la dimensionalidad y encontrar las direcciones principales en las que los datos tienen la mayor variabilidad.

¿Qué características exhiben las componentes?

- Las componentes principales son un conjunto de nuevas variables no correlacionadas, que son combinaciones lineales de las variables originales.
- En el proceso de generación, la primera componente explica la mayor varianza, la segunda la siguiente mayor varianza y así sucesivamente.

Aplicaciones del PCA

- Reducción de dimensionalidad. Al seleccionar las primeras componentes principales, que explican la mayor variación de los datos, se incluye la parte más significativa de la información original, reduciendo lo redundante y, al mismo tiempo, la complejidad del conjunto de datos.
- Multicolinealidad. El PCA produce componentes no correlacionadas, por lo que resuelve el problema de la multicolinealidad.
- Visualización de datos. Al crear una representación de menor dimensión de un conjunto de datos con muchas variables, se pueden visualizar y comprender mejor las relaciones subyacentes entre ellos. Para este fin, se suelen utilizar las dos o tres primeras componentes.
- Compresión de datos. El PCA se puede utilizar como técnica de compresión de datos (por ejemplo, en imágenes y señales) para reducir la redundancia en la información, lo que a su vez permite ahorrar espacio de almacenamiento.
- Identificación de variables relevantes. El PCA permite identificar los atributos que contribuyen significativamente a la variabilidad de los datos, lo que es útil para seleccionar las características más importantes en un problema dado.

Bibliografía

Prosise, J. (2022). Applied Machine Learning and AI for Engineers_ Solve Business Problems That Can't Be Solved Algorithmically. O'Reilly Media (2022)

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.