# Conceptos del PCA Desde el Algebra Lineal

Autores.: Univ. Anara Michua Joel Modesto

Univ. Condori Huanquiri Eleazar David

Universidad Mayor de San Andrés

Carrera de Informática

INTELIGENCIA ARTIFICIAL

Resumen

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos, preservando la mayor cantidad posible de su variabilidad original. PCA proyecta los datos originales en un nuevo sistema de coordenadas, donde las nuevas variables, llamadas componentes principales, son combinaciones lineales de las variables originales. Esta técnica es fundamental en álgebra lineal, ya que se basa en la descomposición de matrices y la búsqueda de autovectores y autovalores. PCA se utiliza en diversos campos, desde la compresión de imágenes hasta la mejora de modelos predictivos.

Las palabras clave "Secuencia", "Genoma", "Propagación" y "Modelado" se integran dentro del contexto del resumen, destacando cómo PCA se puede aplicar a diferentes tipos de datos y contextos, como el análisis de genomas y la propagación de características en datos complejos.

1. Introducción

En el contexto de álgebra lineal y análisis de datos, PCA es un método clave para la reducción de la dimensionalidad de conjuntos de datos multivariantes. La idea fundamental es encontrar una representación más eficiente de los datos sin perder mucha información. Esto se logra mediante la identificación de las direcciones (componentes principales) que capturan la mayor parte de la varianza en los datos. PCA tiene aplicaciones en diversos ámbitos, como la compresión de imágenes, el procesamiento de señales y la construcción de modelos predictivos más eficientes.

2. DESARROLLO

#### 1. **Fundamentos Matemáticos de PCA**

El Análisis de Componentes Principales (PCA) es una técnica matemática fundamental en álgebra lineal que permite reducir la dimensionalidad de un conjunto de datos multivariantes mientras conserva la mayor cantidad posible de varianza de los datos originales. Para entender cómo funciona PCA, primero es necesario comprender algunos conceptos básicos de álgebra lineal que se aplican en este proceso.

##### 1.1 **Centro de los Datos**

El primer paso en el proceso de PCA es el **centrado de los datos**, es decir, restar la media de cada variable (o característica) para asegurar que los datos estén alineados con el origen del nuevo sistema de coordenadas. Esto es fundamental porque PCA busca encontrar direcciones de máxima varianza, y la media de las variables podría sesgar esos cálculos. Si los datos no están centrados, las direcciones principales se verían afectadas por el desplazamiento medio de los datos.

La fórmula para centrar los datos es la siguiente:

Xcentrado=X−μX\_{\text{centrado}} = X - \muXcentrado​=X−μ

donde:

* XXX es la matriz de datos originales,
* μ\muμ es el vector de medias de cada columna de XXX.

Una vez centrados, los datos son representados por un conjunto de puntos que se distribuyen alrededor del origen.

##### 1.2 **Matriz de Covarianza**

Una vez centrados los datos, el siguiente paso es calcular la **matriz de covarianza**. Esta matriz captura la relación lineal entre las diferentes variables del conjunto de datos. En términos simples, la covarianza mide cómo varían conjuntamente dos variables. Si dos variables tienen alta covarianza positiva, significa que tienden a aumentar o disminuir juntas. Si la covarianza es negativa, una variable aumenta cuando la otra disminuye.

La fórmula para calcular la matriz de covarianza CCC de los datos centrados es:

C=1n−1XcentradoTXcentradoC = \frac{1}{n-1} X\_{\text{centrado}}^T X\_{\text{centrado}}C=n−11​XcentradoT​Xcentrado​

donde:

* nnn es el número de observaciones,
* XcentradoX\_{\text{centrado}}Xcentrado​ es la matriz de datos centrados.

La matriz de covarianza es simétrica y cuadrada, y sus valores representan la relación entre las variables del conjunto de datos.

##### 1.3 **Autovectores y Autovalores**

El paso siguiente es el cálculo de los **autovectores** y **autovalores** de la matriz de covarianza. Los autovectores representan las direcciones principales en las que los datos tienen mayor varianza, mientras que los autovalores indican la cantidad de varianza en cada una de esas direcciones. De este modo, los autovectores nos indican las direcciones de máxima varianza en los datos, y los autovalores nos dicen cuánta información (varianza) de los datos se explica por cada componente principal.

El cálculo de autovectores y autovalores se realiza resolviendo la siguiente ecuación para la matriz de covarianza CCC:

Cv=λvC v = \lambda vCv=λv

donde:

* vvv es un autovector,
* λ\lambdaλ es el autovalor correspondiente.

El autovector vvv es una dirección en el espacio de los datos, y λ\lambdaλ es la magnitud de la varianza en esa dirección.

##### 1.4 **Selección de Componentes Principales**

Una vez calculados los autovectores y autovalores, el siguiente paso es **ordenar** los autovectores en función de sus autovalores, de mayor a menor. Los primeros kkk autovectores correspondientes a los autovalores más grandes son seleccionados como los **componentes principales**. Estos componentes son las nuevas direcciones de máxima varianza en los datos originales.

Si kkk es menor que el número total de variables originales, entonces se está realizando una **reducción de dimensionalidad**, ya que se está seleccionando solo un subconjunto de los componentes principales. La cantidad de componentes seleccionados depende del porcentaje de varianza que se quiera conservar.

##### 1.5 **Proyección de los Datos**

Finalmente, se proyectan los datos originales sobre los autovectores seleccionados (componentes principales). Esta proyección transforma el conjunto de datos en un nuevo espacio de características, de menor dimensión, pero que preserva la mayor parte de la varianza de los datos originales.

La fórmula para proyectar los datos es:

Xproyectado=Xcentrado⋅VkX\_{\text{proyectado}} = X\_{\text{centrado}} \cdot V\_kXproyectado​=Xcentrado​⋅Vk​

donde:

* XcentradoX\_{\text{centrado}}Xcentrado​ es la matriz de datos centrados,
* VkV\_kVk​ es la matriz que contiene los primeros kkk autovectores.

El conjunto de datos proyectado, XproyectadoX\_{\text{proyectado}}Xproyectado​, tiene kkk dimensiones y contiene las componentes principales de los datos originales.

#### 2. **Propiedades de PCA**

PCA es una técnica que tiene varias propiedades importantes:

* **Reducción de dimensionalidad**: El principal propósito de PCA es reducir la cantidad de variables (dimensiones) de los datos, manteniendo la mayor parte de la información posible. Esto es útil en problemas donde el número de variables es elevado, ya que reduce la complejidad computacional sin perder información relevante.
* **Preservación de la varianza**: PCA garantiza que las nuevas dimensiones (componentes principales) mantengan la mayor cantidad posible de la varianza original del conjunto de datos. Esto significa que los componentes principales capturan las características más importantes del conjunto de datos.
* **Independencia de los componentes**: Los componentes principales obtenidos a través de PCA son ortogonales entre sí, lo que significa que no hay correlación entre ellos. Esto facilita el análisis de los datos y puede mejorar el desempeño de otros algoritmos de aprendizaje automático.

#### 3. **Aplicaciones de PCA**

* **Visualización de datos**: PCA es útil para la visualización de datos de alta dimensión. Al proyectar los datos en 2D o 3D, se puede obtener una representación visual que facilite la interpretación y el análisis.
* **Compresión de imágenes**: En la compresión de imágenes, PCA permite reducir el tamaño de los archivos de imagen al representar la imagen con menos dimensiones, preservando la mayor parte de la información visual.
* **Mejora de modelos predictivos**: PCA puede ser útil en la mejora del desempeño de modelos de aprendizaje automático al reducir la dimensionalidad de los datos y eliminar variables irrelevantes o ruidosas, lo que ayuda a evitar el sobreajuste y mejora la eficiencia computacional.
* **Estudio de secuencias genéticas**: En el campo de la biología y la genética, PCA puede ser utilizado para analizar grandes volúmenes de datos genómicos, como secuencias de ADN. Al reducir la dimensionalidad de estos datos complejos, es posible identificar patrones subyacentes y relaciones entre genes o poblaciones.

Aquí tienes las referencias formateadas según tu solicitud:

**Referencias**

1. Castañeda Naranjo, L. A., & Palacios Neri, J. (2015). Nanotecnología: fuente de nuevos paradigmas. *Mundo Nano. Revista Interdisciplinaria en Nanociencias y Nanotecnología, 7*(12), 45-49.
2. European Commission. (2020). Bioinformatics against COVID-19: €56 million for new research projects.
3. Muñoz Vila, C. (enero-febrero 2012). Lo que se haga por un niño se hace por un pueblo. *Revista Internacional Magisterio, (54)*, 10-17.
4. World Health Organization. (2020). Coronavirus disease (COVID-19) pandemic.