**PCA – ANALISIS DE COMPONENTES PRINCIPALES**

**Conceptos y Definiciones Principales**

El Análisis de Componentes Principales (PCA) es una técnica de aprendizaje no supervisado utilizada para reducir la dimensionalidad de un conjunto de datos mientras se conserva la mayor cantidad posible de variación presente. PCA busca transformar las variables originales en un conjunto más pequeño de variables derivadas, llamadas **componentes principales**, que son combinaciones lineales de las originales. Los principales objetivos de PCA son:

1. **Reducción de Dimensionalidad:** Encontrar un conjunto de variables derivadas que expliquen la mayor parte de la varianza en los datos.
2. **Visualización:** Facilitar la interpretación y representación gráfica de los datos en un espacio de menor dimensión.

Cada componente principal está definido por una dirección en el espacio que maximiza la varianza de los datos proyectados sobre esa dirección. El proceso se basa en optimizar combinaciones lineales de variables, sujeto a que las direcciones encontradas sean ortogonales entre sí.

Es **muy importante centrar y, en la mayoría de los casos, estandarizar los datos antes de aplicar PCA**. Esto se debe a que el Análisis de Componentes Principales (PCA) se basa en la varianza de las variables, y estas transformaciones afectan directamente cómo se calculan las componentes principales. Aquí te explico por qué:

**1. Centrado de los Datos**

El centrado implica restar la media de cada variable a sus valores originales, de modo que todas las variables tengan una media de cero. Si X es la matriz de datos y Xj​ es la media de la j-ésima variable, los datos centrados Xc​ se calculan como:

Xc=X−media de las columnas

**¿Por qué es importante?**

1. **PCA trabaja con la matriz de covarianzas o correlaciones**, y ambas dependen de la media de las variables.
2. **La media no centrada introduce un sesgo** en los resultados, ya que la primera componente principal podría reflejar la magnitud de las medias en lugar de la estructura de varianza y covarianza en los datos.

**2. Estandarización de los Datos**

La estandarización implica transformar las variables para que tengan una media de 0 y una varianza de 1. Se realiza dividiendo los datos centrados por la desviación estándar de cada variable:

Xest=Xc/σ​​

Donde σ\sigma es la desviación estándar de cada variable.

**¿Por qué es importante?**

1. **Las escalas diferentes entre variables afectan los resultados del PCA.**
   * Si las variables tienen magnitudes diferentes (por ejemplo, ingresos anuales en miles de dólares frente a edad en años), las variables con mayor varianza dominarán el análisis.
   * PCA busca maximizar la varianza, por lo que las variables con grandes magnitudes tendrán mayor influencia, incluso si no son las más relevantes para explicar los datos.
2. **Estandarizar hace que el análisis sea insensible a las unidades de las variables.**
   * Por ejemplo, cambiar de dólares a euros no debería alterar los resultados del PCA si las variables están estandarizadas.

**¿Cuándo es suficiente centrar y no estandarizar?**

En algunos casos, solo centrar los datos es suficiente:

1. **Si las variables ya están en la misma escala y son comparables.**
   * Por ejemplo, si todas las variables representan porcentajes o tienen rangos similares.
2. **Si te interesa interpretar las componentes principales en términos de las unidades originales de las variables.**
   * En estos casos, la matriz de covarianzas (y no de correlaciones) se utiliza para el PCA.

**¿Qué sucede si no centras o estandarizas?**

1. **Sin centrar:**
   * Las componentes principales podrían no representar las direcciones de mayor varianza correctamente.
   * La primera componente podría alinearse con la media de las variables en lugar de las relaciones entre ellas.
2. **Sin estandarizar (cuando es necesario):**
   * Las variables con mayores magnitudes dominarán el análisis, y las componentes principales podrían no reflejar la estructura subyacente real de los datos.

**Conclusión**

Para aplicar PCA correctamente:

* **Siempre centra los datos** para garantizar que las componentes principales reflejen la varianza y no la media.
* **Estandariza los datos** si las variables tienen escalas o unidades diferentes. Esto asegura que todas las variables tengan la misma influencia en el análisis.

La regla general es:

* **Usa la matriz de correlaciones (centrado + estandarización)** cuando las variables tienen escalas diferentes.
* **Usa la matriz de covarianzas (solo centrado)** si todas las variables están en la misma escala.

**Teorema de la Esfera de la Unidad**

Los coeficientes (o cargas) que definen cada componente principal están normalizados, lo que implica que forman un vector unitario. Es decir, la suma de los cuadrados de las cargas es igual a uno. Este principio evita que los coeficientes se escalen arbitrariamente, lo que podría inflar la varianza explicada de manera artificial.

**Teorema de la Esfera de la Unidad en PCA**

En PCA, buscamos encontrar las direcciones (componentes principales) que maximicen la varianza de los datos proyectados, sujeto a ciertas restricciones. El Teorema de la Esfera de la Unidad establece que los vectores de carga vkv\_kvk​ que definen los componentes principales deben tener longitud unitaria, es decir, deben estar en la superficie de una esfera de radio uno en el espacio de las variables originales.

**Lectura de un Biplot**

Un **biplot** es una representación gráfica en la que se visualizan los datos en el espacio reducido generado por las componentes principales (ejes del gráfico) y las variables originales como vectores. En un biplot:

1. **Observaciones:** Los puntos representan las observaciones transformadas en el espacio de las primeras dos (o más) componentes principales.
2. **Variables:** Los vectores indican la contribución y dirección de las variables originales a las componentes principales. La longitud del vector refleja la importancia de la variable en esa dirección.
3. **Interpretación:** La proyección de los puntos sobre los vectores muestra la relación entre las observaciones y las variables.

**Varianza Explicada**

La proporción de varianza explicada (PVE) mide qué tanto contribuye cada componente principal a la varianza total en los datos. Los pasos principales son:

1. **Cálculo de la varianza total:** Suma de las varianzas de todas las variables originales.
2. **Cálculo de la varianza por componente:** Suma de las varianzas explicadas por cada componente principal.
3. **PVE:** Dividir la varianza explicada por cada componente entre la varianza total.

El **scree plot** es una herramienta común para visualizar la PVE y seleccionar cuántos componentes considerar. En este gráfico, el "codo" indica el punto donde agregar más componentes principales tiene un beneficio marginal reducido​​​​​.

**Conclusión**

PCA es una herramienta poderosa para simplificar y analizar datos de alta dimensionalidad, destacándose por su capacidad para reducir ruido y facilitar interpretaciones visuales significativas.