

Componentes Principales

Benjamín Vallejos

1 Introducción

El **Análisis de Componentes Principales (PCA)** es una técnica fundamental en el análisis multivariante que permite reducir la dimensionalidad de un conjunto de datos, conservando la mayor cantidad de información posible. Este método es especialmente útil cuando se trabaja con un gran número de variables correlacionadas.

2 Planteamiento del Problema

Dado un conjunto de datos multivariantes representado por una matriz X de dimensiones $n \times p$, el PCA busca proyectar estos datos en un espacio de menor dimensión $r < p$.

- **Matriz de datos:** X con n filas (observaciones) y p columnas (variables).
- **Matriz de covarianzas:** $S = \frac{1}{n}X'X$, donde X tiene media cero.

3 Cálculo de los Componentes Principales

Los componentes principales se obtienen a través de la descomposición en valores y vectores propios de la matriz de covarianzas S .

3.1 Primer Componente Principal

El primer componente principal z_1 es la combinación lineal de las variables originales que maximiza la varianza. Se obtiene resolviendo el problema de optimización:

$$\max a_1' S a_1 \quad \text{sujeto a} \quad a_1' a_1 = 1.$$

La solución es el vector propio a_1 asociado al mayor valor propio λ_1 de la matriz S :

$$S a_1 = \lambda_1 a_1.$$

3.2 Segundo Componente Principal

El segundo componente principal z_2 se obtiene de manera similar, pero con la restricción adicional de que z_2 debe ser ortogonal a z_1 :

$$S a_2 = \lambda_2 a_2, \quad a_1' a_2 = 0.$$

3.3 Generalización a r Componentes

El espacio de dimensión r está definido por los r vectores propios asociados a los r mayores valores propios de S . La matriz de componentes principales es:

$$Z = XA,$$

donde A es la matriz de vectores propios.

4 Propiedades de los Componentes Principales

1. **Conservación de la variabilidad:** La suma de las varianzas de los componentes principales es igual a la suma de las varianzas de las variables originales.

$$\sum_{i=1}^p \lambda_i = \text{traza}(S).$$

2. **Proporción de variabilidad explicada:** Para el componente h :

$$\frac{\lambda_h}{\sum_{i=1}^p \lambda_i}.$$

3. **Correlaciones con las variables originales:**

$$\text{Corr}(z_i, x_j) = \frac{a_{ij} \sqrt{\lambda_i}}{s_j}.$$

5 Análisis Normado (PCA con Correlaciones)

Cuando las variables tienen unidades distintas, se recomienda estandarizarlas y trabajar con la matriz de correlaciones R en lugar de la matriz de covarianzas S .

6 Selección del Número de Componentes

Existen varias reglas para seleccionar el número de componentes:

- **Gráfico de codo:** Se seleccionan los componentes hasta que los valores propios restantes son aproximadamente iguales.
- **Proporción de varianza explicada:** Se eligen los componentes que explican al menos un 80% de la varianza.
- **Regla de valores propios mayores que 1:** Se retienen componentes con valores propios mayores que 1.

7 Representación Gráfica

Los componentes principales pueden representarse en un espacio de dos dimensiones definido por los dos primeros componentes, facilitando la interpretación de la estructura de los datos.

8 Datos Atípicos

Los datos atípicos pueden distorsionar el PCA, ya que afectan la matriz de covarianzas. Es importante detectarlos antes de aplicar el análisis.

9 Conclusión

El **Análisis de Componentes Principales (PCA)** permite reducir la dimensionalidad de los datos, facilitando su interpretación y visualización. A través de la descomposición en valores y vectores propios, se identifican las direcciones de máxima variabilidad en los datos.