Análisis de Componentes Principales

Nombre del autor

March 15, 2023

1 Introducción

El Análisis de Componentes Principales (ACP) es una técnica de reducción de dimensionalidad que se utiliza para explorar y visualizar la estructura de los datos. El ACP se utiliza comúnmente en campos como la estadística, la econometría, la psicometría y la ingeniería, entre otros.

El objetivo del ACP es transformar un conjunto de variables originales en un conjunto de variables no correlacionadas llamadas componentes principales. Estas componentes principales explican la mayor cantidad posible de la varianza de los datos originales.

2 Definiciones básicas

Antes de introducir el algoritmo del ACP, se presentan algunas definiciones básicas que se utilizan en el análisis de componentes principales.

- **Matriz de datos**: Una matriz de datos es una matriz de n filas y p columnas, donde n es el número de observaciones y p es el número de variables.
- **Vector de medias**: El vector de medias es un vector de p elementos que contiene la media de cada variable en la matriz de datos. Este vector se utiliza para centrar los datos antes de realizar el ACP.
- **Matriz de covarianza**: La matriz de covarianza es una matriz simétrica de $p \times p$ que muestra la covarianza entre todas las variables en la matriz de datos. La covarianza entre dos variables es una medida de la relación lineal entre ellas.
- **Valores propios y vectores propios**: Los valores propios y vectores propios son conceptos relacionados con las matrices cuadradas. Si A

Objetivo

Dado un conjunto de n observaciones de p variables $X = [X_1, X_2, \ldots, X_p]$, el objetivo del análisis de componentes principales (ACP) es encontrar un nuevo conjunto de p variables no correlacionadas llamadas componentes principales $Y = [Y_1, Y_2, \ldots, Y_p]$ que expliquen la mayor cantidad posible de la varianza de los datos originales.

Procedimiento

El procedimiento para encontrar los componentes principales es el siguiente:

1. Calcular la matriz de covarianza de los datos centrados:

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T$$

   donde $\bar{X}$ es el vector de medias de las variables en X.

2. Calcular los valores propios $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ y los vectores propios $v_1, v_2, \ldots, v_p$ de la matriz de covarianza S. Los vectores propios son las direcciones en las que la varianza de los datos es máxima, y los valores propios son las varianzas correspondientes en esas direcciones.

3. Ordenar los valores propios y los vectores propios de forma descendente según los valores propios.

4. Tomar los k vectores propios con los valores propios más grandes y construir la matriz de transformación $V_k = [v_1, v_2, \ldots, v_k]$. Esta matriz se utiliza para transformar los datos originales X en los componentes principales Y:

$$Y = XV_k$$

   donde Y es una matriz de n filas y k columnas.

Interpretación

La interpretación de los componentes principales es la siguiente:

- Cada componente principal Y_j es una combinación lineal de las variables originales $X_1, X_2, \ldots, X_p$.
- Los componentes principales están ordenados por su capacidad para explicar la varianza de los datos originales.

- El porcentaje de varianza explicado por cada componente principal se puede calcular como:

  Porcentaje de varianza explicado por el componente principal $Y_j = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \times 100\%$

  donde λ_j es el valor propio correspondiente al componente principal Y_j.
- Los componentes principales son no correlacionados entre sí, lo que significa que la matriz de correlación de los componentes principales es una matriz diagonal. Esto simplifica el análisis y la interpretación de los datos.

Ventajas y desventajas

Las ventajas del ACP son:

- Permite reducir la dimensionalidad de los datos sin perder información importante.
- Ayuda a identificar patrones y relaciones en los datos que no son evidentes en la representación original.
- Proporciona una forma de resumir la información de múltiples variables en un número reducido de componentes principales.
- Es una técnica no supervisada, lo que significa que no se necesita información previa sobre las categorías o grupos de las observaciones.

Las desventajas del ACP son:

- Puede ser difícil de interpretar los componentes principales, especialmente si tienen una combinación compleja de variables originales.
- No es adecuado para datos categóricos o nominales, ya que la matriz de covarianza se basa en la varianza de las variables continuas.
- El ACP supone una distribución normal multivariante de los datos, por lo que puede no ser adecuado para datos que no siguen esta distribución.
- La elección del número de componentes principales puede ser subjetiva y puede tener un impacto en la interpretación de los resultados.

Implementación en R

El siguiente código en R muestra cómo realizar un análisis de componentes principales en un conjunto de datos:

[language=R]

Carga de datos data ¡- read.csv("datos.csv")

Análisis de componentes principales acp ¡- prcomp(data, scale = TRUE)

Varianza explicada $\text{var}_explicada < -acp\text{sdev}^2/sum(acp\text{sdev}^2)$

Gráfico de varianza explicada $\text{plot}(\text{var}_explicada, xlab = "Componentesprincipales", ylab = "Varianzaexplicada", type = "b", main = "Gráficodevarianzaexplicada")$

Gráfico de componentes principales biplot(acp, main = "Gráfico de componentes principales")

En este ejemplo, se cargan los datos desde un archivo CSV y se realiza un análisis de componentes principales utilizando la función `prcomp()` de R. La opción `scale = TRUE` se utiliza para estandarizar los datos antes del análisis. A continuación, se calcula el porcentaje de varianza explicado por cada componente principal y se muestra en un gráfico. También se muestra un gráfico de los componentes principales utilizando la función `biplot()` de R. Este gráfico muestra las variables originales en el espacio de los componentes principales, lo que ayuda a interpretar la relación entre las variables y los componentes.

Ejemplo numérico

Para ilustrar el proceso del ACP, consideremos un ejemplo con un conjunto de datos de 6 observaciones y 3 variables:

Observación	**Variable 1**	**Variable 2** height1
3	6 height2	5
3 height3	1	7 height4
4	2 height5	2
8 height6	6	1 height

Table 1: Conjunto de datos de ejemplo.

El objetivo es reducir la dimensionalidad de los datos, manteniendo la mayor cantidad de información posible. Para ello, se realiza el siguiente análisis de componentes principales:

1. Se calcula la matriz de covarianza de las variables originales:

$$\mathbf{S} = \begin{bmatrix} 3.2 & -2.2 \\ -2.2 & 9.2 \end{bmatrix}$$

2. Se calculan los autovalores y autovectores de la matriz de covarianza:

$$\lambda_1 = 10.929 \qquad \mathbf{v}_1 = \begin{bmatrix} 0.658 \\ -0.753 \end{bmatrix}$$

$$\lambda_2 = 1.271 \qquad \mathbf{v}_2 = \begin{bmatrix} 0.753 \\ 0.658 \end{bmatrix}$$

Los autovalores representan la varianza explicada por cada componente principal, mientras que los autovectores representan la dirección de los componentes principales en el espacio de las variables originales.

3. Se ordenan los autovalores de mayor a menor y se eligen los componentes principales correspondientes a los autovalores más grandes. En este caso, se elegirían los dos componentes principales.
4. Se calcula la matriz de componentes principales:

$$\mathbf{T} = \begin{bmatrix} -0.626 & -1.388 \\ 0.343 & -0.283 \\ -1.095 & 0.789 \\ 0.002 & -0.284 \\ -0.760 & 1.012 \\ 1.735 & -0.847 \end{bmatrix}$$

donde cada fila representa una observación en el espacio de los componentes principales. Cada columna representa un componente principal.

5. Se calcula la varianza explicada por cada componente principal:

$$\text{Varianza explicada por CP1} = \frac{\lambda_1}{\lambda_1 + \lambda_2} = 0.896$$

$$\text{Varianza explicada por CP2} = \frac{\lambda_2}{\lambda_1 + \lambda_2} = 0.104$$

Podemos ver que la mayoría de la variación se encuentra en la dirección del primer componente principal.

6. Finalmente, se puede utilizar la matriz de componentes principales $\mathbf{T}$ para realizar análisis posteriores, como regresión o clustering, en el espacio de los componentes principales en lugar del espacio original de las variables.

Conclusión

El análisis de componentes principales es una técnica útil para reducir la dimensionalidad de un conjunto de datos, manteniendo la

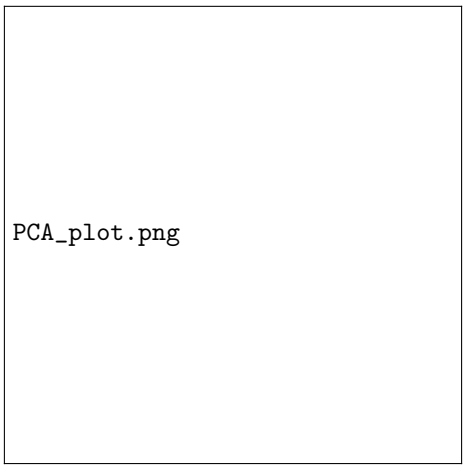

Figure 1: Representación gráfica de los datos en el espacio de los componentes principales.

mayor cantidad de información posible. El proceso implica la transformación de las variables originales en un conjunto de variables no correlacionadas llamadas componentes principales, que se ordenan en función de su varianza explicada. Aunque esta técnica no siempre es apropiada para todos los conjuntos de datos, puede ser una herramienta valiosa en la exploración y visualización de datos de alta dimensionalidad.