Análisis de Componentes Principales

Universidad Nacional de Colombia

Yudy Vanesa Puerres Rosero (ypuerresr@unal.edu.co)
Camila Andrea Ayala Camargo (Caayala@unal.edu.co)
Karen Liliana Barrantes Quiroga (kbarrantes@unal.edu.co)
Laura Katherine Martinez Castiblanco (laumartinezca@unal.edu.co)
Fredy Arley Urrea Cifuentes (furreac@unal.edu.co)

Introducción

En la investigación aplicada, especialmente en ciencias sociales, biología, economía e ingeniería es común trabajar con conjuntos de datos de alta dimensionalidad. Aunque un mayor número de variables puede aportar información valiosa, su análisis simultáneo presenta dificultades como el aumento de parámetros a estimar, el riesgo de multicolinealidad y la complejidad para interpretar la estructura subyacente de los datos. Esto hace necesario emplear métodos que resuman la información esencial sin perder las características relevantes del conjunto original.

El Análisis de Componentes Principales (ACP) es una técnica multivariante diseñada precisamente para reducir la dimensionalidad. Transforma las variables originales en un nuevo conjunto menor de variables no correlacionadas, denominadas componentes principales. Estas componentes son combinaciones lineales de las variables originales, construidas de modo que capturen la máxima varianza posible, concentrando así la mayor cantidad de información.

Definición:

El Análisis de Componentes Principales (ACP) es una técnica estadística multivariante cuyo objetivo es transformar un conjunto de variables originales $X_1, X_2, \dots, X_p$ en un nuevo conjunto de variables no correlacionadas llamadas **componentes principales**. Estas nuevas variables son combinaciones lineales de las variables originales y se construyen de forma que capturen la máxima varianza posible, es decir, la mayor cantidad de información contenida en los datos.

Formalmente, el k-ésimo componente principal se define como:

$$Y_k = a_{1k}X_1 + a_{2k}X_2 + \cdots + a_{pk}X_p,$$

donde el vector

$$\mathbf{a}_k = (a_{1k}, a_{2k}, \dots, a_{pk})'$$

es el **autovector** correspondiente al k-ésimo **autovalor** λ_k de la matriz de covarianzas (o correlaciones) del conjunto de variables originales.

Los autovalores cumplen:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0,$$

y cada λ_k representa la **varianza explicada** por el componente principal Y_k.
El primer componente principal captura la mayor varianza posible; el segundo captura la mayor varianza restante bajo la condición de ser **ortogonal** al primero, y así sucesivamente.

Descripción general del método

Existen dos enfoques principales para comprender el método del Análisis de Componentes Principales (ACP). El primero, tradicionalmente utilizado en estadística, consiste en construir los componentes en las direcciones donde la matriz de datos X presenta la **máxima varianza**. Bajo este enfoque, el primer componente principal es la combinación lineal de las variables que captura la mayor variabilidad posible; el segundo componente captura la mayor variabilidad restante bajo la condición de ser ortogonal al primero, y así sucesivamente. Para ello se emplean los autovalores y autovectores de la matriz de covarianzas o correlaciones, lo que garantiza que los componentes sean no correlacionados y estén ordenados según la varianza explicada.

El segundo enfoque proviene del aprendizaje estadístico moderno, donde el ACP se interpreta como un problema de optimización que busca la mejor aproximación de la matriz de datos con menor dimensión. Esta aproximación se obtiene mediante la descomposición en valores singulares (SVD), que produce las mismas direcciones de variabilidad máxima que el enfoque clásico.

Ambos enfoques producen la misma solución: los componentes principales corresponden a las direcciones de máxima varianza de X y simultáneamente a las direcciones que generan la mejor aproximación de rango reducido de la matriz de datos. Esta equivalencia explica la solidez del ACP y su importancia tanto en la estadística clásica como en el aprendizaje automático.

Supuestos del método

El Análisis de Componentes Principales (ACP) se basa en varios supuestos que garantizan la validez de su interpretación y de sus resultados:

1. **Relación lineal entre variables:**
   El ACP asume que las relaciones entre las variables pueden describirse adecuadamente mediante combinaciones lineales. No es adecuado para capturar relaciones no lineales.

2. **Escalas comparables entre variables:**
   Dado que la varianza es sensible a la escala de medición, las variables deben estar estandarizadas cuando poseen unidades o magnitudes muy diferentes.

3. **Varianza significativa en las variables:**
   Las variables deben presentar variabilidad suficiente; variables casi constantes no aportan información y distorsionan los componentes.

4. **Número adecuado de observaciones:**
   Se recomienda contar con un número de observaciones considerablemente mayor que el número de variables para obtener estimaciones estables de la matriz de covarianzas o correlaciones.

5. **Ausencia de multicolinealidad perfecta:**
   Aunque el ACP maneja bien la colinealidad, no puede aplicarse cuando algunas variables son combinaciones lineales exactas de otras, pues la matriz de covarianzas sería singular.

6. **Normalidad multivariada (deseable pero no estrictamente necesaria):**
   El ACP no requiere que las variables sigan una distribución normal multivariada para ser calculado; sin embargo, este supuesto es útil cuando se desea realizar inferencias estadísticas o interpretar los componentes dentro de un modelo probabilístico.

En conjunto, estos supuestos aseguran que el ACP proporcione componentes interpretables y representativos de la estructura interna de los datos. El cumplimiento de estos principios contribuye a mejorar la estabilidad y la calidad de los resultados obtenidos.

Reconstrucción Exacta de una Matriz de Datos usando Valores y Vectores Propios

En el análisis multivariado, y en particular en métodos como el Análisis de Componentes Principales (ACP), es posible reconstruir exactamente una matriz de datos utilizando sus valores y vectores propios gracias a la Descomposición en Valores Singulares (DVS).

Este procedimiento descompone una matriz en el producto de tres matrices con características específicas, lo que no solo facilita su interpretación geométrica, sino que también resulta muy útil para realizar reducción de dimensionalidad.

La Descomposición en Valores Singulares (DVS)

La DVS muestra que cualquier matriz C de tamaño $n \times p$ y rango r puede factorizarse de la siguiente forma:

$$C = VLU'$$

Donde:

- L: matriz diagonal $r \times r$ que contiene los **valores propios** de C, es decir, $\sqrt{\lambda_j}$, donde λ_j es el j-ésimo valor propio de $C'C$
- U: matriz $p \times r$ cuyas columnas son los **vectores propios de** $C'C$
- V: matriz $n \times r$ cuyas columnas son los **vectores propios de** CC'
- Ambas matrices U y V son **ortonormales**: $U'U = V'V = I_r$

Luego, cuando aplicamos la DVS a la matriz X centrada y estandarizada, obtenemos que:

$$X = VLU'$$

Donde:

- L contiene $\sqrt{\lambda_\alpha}$, con λ_α siendo los valores propios de $X'X$
- Las columnas de U son los vectores propios de $X'X$

- Las columnas de V son los vectores propios de XX'

Ahora bien, Se puede escribir X en su forma expandida

$$X \;=\; [\, v_1 \; v_2 \; \ldots \; v_p \,] \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_p} \end{bmatrix} \begin{bmatrix} u_1' \\ u_2' \\ \vdots \\ u_p' \end{bmatrix}$$

Luego, al multiplicar ((V)) por ((L)) tenemos que:

$$VL = [\sqrt{\lambda_1} v_1, \sqrt{\lambda_2} v_2, \ldots, \sqrt{\lambda_p} v_p]$$

y al multiplicar por ((U')):

$$X \;=\; [\, \sqrt{\lambda_1}\, v_1,\; \sqrt{\lambda_2}\, v_2,\; \ldots,\; \sqrt{\lambda_p}\, v_p \,] \begin{bmatrix} u_1' \\ u_2' \\ \vdots \\ u_p' \end{bmatrix}$$

$$X \;=\; \sqrt{\lambda_1}\, v_1 u_1' \;+\; \sqrt{\lambda_2}\, v_2 u_2' \;+\; \cdots \;+\; \sqrt{\lambda_p}\, v_p u_p'$$

$$X \;=\; \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\, v_\alpha u_\alpha'$$

Esto nos muestra que toda la información de la matriz original está contenida en sus valores y vectores propios.

Por otro lado, un resultado clave que surge de la DVS es la **equivalencia entre los valores propios de** (X'X) **y** (XX'). Ambos comparten los mismos valores propios no nulos (λ_α). Esto es:

- Si (u_α) es un vector propio de (X'X), entonces $\frac{1}{\sqrt{\lambda_\alpha}} X u_\alpha$ es un vector propio de (XX').
- Análogamente, si (v_α) es un vector propio de (XX'), entonces $\frac{1}{\sqrt{\lambda_\alpha}} X' v_\alpha$ es un vector propio de (X'X).

Esto implica que **solo es necesario calcular los valores y vectores propios de una de estas matrices** para obtener también los de la otra.

Interpretación y utilidad en ACP

Así, la factorización (X = V L U') muestra una transformación diferente de los datos:

Las Componentes Principales (Scores de los Individuos):
La matriz (Z = V L) contiene las coordenadas de los individuos; es decir, cada columna de (Z) representa la proyección de todas las observaciones sobre la α-ésima componente. Esto muestra cómo se ubican las muestras en el nuevo espacio de características reducido.

Las Cargas Factoriales (Contribuciones de las Variables):
La matriz (U) o, en su versión escalada, $U\sqrt{L}$ reúne las cargas factoriales, que muestran el peso y la dirección que tiene cada variable original en la construcción de las componentes principales. Cuando una carga es alta en valor absoluto, significa que esa variable tiene una influencia importante sobre la componente correspondiente. Es decir, la matriz (U) nos muestra qué variables originales son las que realmente impulsan cada componente.

En general, la reconstrucción exacta de la matriz (X) a través de la DVS no solo es un resultado teórico interesante, sino que también constituye el fundamento computacional del ACP. Gracias a este enfoque es posible reducir la dimensionalidad sin perder la estructura esencial de los datos, facilitar la visualización de las relaciones entre individuos y variables, y comprender mejor la geometría que hay detrás del conjunto de datos.

Varianza de las Componentes Principales

La varianza de cada componente principal puede obtenerse utilizando el Teorema de la Descomposición Espectral (TDE). Si Y es la matriz de datos centrados y estandarizados, la matriz de correlaciones se define como:

$$R = \frac{1}{n} Y'Y.$$

El TDE garantiza que esta matriz puede descomponerse como:

$$R = U\Lambda U',$$

donde:

- U es una matriz ortogonal cuyas columnas son los **vectores propios estandarizados** de R,
- $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ contiene los **valores propios** ordenados de mayor a menor.

Cada valor propio λ_α corresponde a la **varianza** del α-ésimo componente principal:

$$\text{var}(z_\alpha) = \lambda_\alpha.$$

A partir de esto, el **porcentaje de varianza explicada** por el α-ésimo componente se define como:

$$\tau_\alpha = \frac{\lambda_\alpha}{\sum_{i=1}^{p} \lambda_i}.$$

Asimismo, la **varianza explicada acumulada** por los primeros q componentes es:

$$\tau_q = \frac{\sum_{\alpha=1}^{q} \lambda_\alpha}{\sum_{i=1}^{p} \lambda_i}.$$

Estas cantidades permiten evaluar cuánto de la información original está siendo representada por los componentes retenidos. En general, se seleccionan los primeros componentes que explican un porcentaje adecuado de la varianza total, lo cual asegura una representación eficiente del conjunto de datos en menos dimensiones.

Elementos para la interpretación de un ACP

La interpretación de un Análisis de Componentes Principales (ACP) requiere integrar diversos elementos que describen la estructura interna de los datos, la contribución de las variables y la organización de los objetos en el espacio factorial. Aunque cada indicador puede analizarse por separado, la interpretación final debe ser conjunta para obtener una lectura coherente del fenómeno estudiado. A continuación, se presentan los principales elementos utilizados en la interpretación de un ACP

Calidad de la representación

La calidad de la representación evalúa qué tanto los componentes principales logran resumir la información contenida en las variables originales. Se basa en la **varianza explicada acumulada**, que indica el porcentaje de variabilidad capturada por los primeros componentes.

Los criterios más comunes para decidir cuántos componentes conservar son:

- Alcanzar un porcentaje deseado de varianza acumulada (70%–90%).
- Conservar los componentes con autovalor mayor que 1 (regla de Kaiser).
- Analizar el gráfico de sedimentación (*scree plot*).

Una buena selección asegura una reducción de la dimensionalidad sin pérdida significativa de información.

Correlaciones variable–factor

Las correlaciones variable–factor, también llamadas *cargas* o *loadings*, indican el grado en que cada variable original está asociada a un componente principal. Estas correlaciones permiten evaluar cuánto aporta una variable a la construcción de un componente y qué tan bien queda representada en el espacio reducido.

Para la variable Y_j y el componente α, la correlación variable–factor está dada por:

$$w_{j\alpha} = \sqrt{\lambda_\alpha}\, u_{j\alpha},$$

donde $u_{j\alpha}$ es el elemento del vector propio correspondiente al componente α, y λ_α es su autovalor. Esta cantidad corresponde exactamente a la correlación entre la variable original y el componente principal.

Valores altos de $w_{j\alpha}$ indican que la variable está bien explicada por ese componente y contribuye de manera importante a su interpretación.

Típificación de los factores por las variables

Para interpretar los factores (componentes) se utilizan diversos indicadores que relacionan las variables originales con los ejes factoriales obtenidos en el ACP.

Coordenadas de las variables (scores)

Las coordenadas de las variables se obtienen como $w_\alpha = Y'v_\alpha$, donde v_α es el vector propio asociado al componente α. Estas coordenadas coinciden con las correlaciones variable–factor, por lo que permiten identificar qué variables influyen más en cada componente principal.

Contribuciones de las variables

La contribución de la variable j al componente α indica qué proporción de la varianza del componente es explicada por esa variable. Se calcula mediante:

$$\text{contribución}_{j\alpha} = \frac{z_{\alpha j}^2}{\lambda_\alpha},$$

donde $z_{\alpha j}$ es la coordenada de la variable en el componente y λ_α su valor propio. Contribuciones altas indican que la variable participa de manera importante en la construcción del componente.

Cosenos cuadrados (cos²)

Los cosenos cuadrados representan la cantidad de varianza de la variable que es explicada por el componente α. Corresponden a la correlación variable–factor elevada al cuadrado. Si $u_{j\alpha}$ es la carga del vector propio, entonces:

$$\cos^2(\theta_{j,\alpha}) = \lambda_\alpha u_{j\alpha}^2.$$

Valores grandes de $\cos^2$ indican que la variable está bien representada por el componente y que su posición en el plano factorial es confiable.

Planos factoriales

Los planos factoriales representan variables y/u objetos mediante pares de componentes principales. Las variables aparecen como vectores desde el origen y los objetos como puntos. La interpretación de estos planos permite identificar asociaciones entre variables, agrupamientos entre objetos y la calidad de su representación.

A partir de la posición de las variables en el plano, es posible analizar su relación con los componentes, identificar similitudes o asociaciones entre ellas y evaluar su representación mediante las coordenadas, contribuciones y cosenos cuadrados. La interpretación de los planos factoriales complementa la interpretación de los factores y facilita la comprensión de la estructura multivariante de los datos.

Distancia entre variables

La distancia entre dos variables en el ACP puede interpretarse mediante el coseno del ángulo que forman sus vectores en el espacio factorial. Si $\theta_{jj'}$ es el ángulo entre las variables Y_j y $Y_{j'}$, la distancia euclidiana entre ellas puede expresarse como:

$$d^2(Y_j, Y_{j'}) = 2\,(1 - \cos(\theta_{jj'})).$$

Esto permite interpretar la relación entre las variables en función de su correlación:

- Si $\cos(\theta_{jj'}) \to 1$ (variables muy correlacionadas), entonces $d(Y_j, Y_{j'}) \to 0$.

- Si $\cos(\theta_{jj'}) \to 0$ (variables no correlacionadas), entonces $d(Y_j, Y_{j'}) \to \sqrt{2}$.
- Si $\cos(\theta_{jj'}) \to -1$ (variables inversamente correlacionadas), entonces $d(Y_j, Y_{j'}) \to 2$.

Cuando el ACP se realiza a partir de la matriz de covarianzas, la distancia entre las variables X_j y $X_{j'}$ se expresa como:

$$d^2(X_j, X_{j'}) = s_j + s_{j'} - 2s_{jj'},$$

donde s_j y $s_{j'}$ son las varianzas de las variables y $s_{jj'}$ su covarianza. En ambos casos, distancias pequeñas indican variables similares, mientras que distancias grandes reflejan relaciones débiles o inversas entre ellas.

Tipificación de los objetos

Además de interpretar las variables, el ACP permite analizar las posiciones de los objetos u observaciones (como países, universidades, empresas o individuos) en el espacio factorial. A partir de las coordenadas de los objetos en los componentes principales, así como de sus contribuciones y cosenos cuadrados, es posible identificar patrones, agrupamientos o tendencias presentes en los datos.

Este análisis permite asociar características a los objetos según su ubicación en los planos factoriales y según la influencia que ejercen las variables originales en cada componente. De esta manera, la tipificación de los objetos complementa la interpretación global del ACP al revelar similitudes y diferencias entre las observaciones.

Distancia entre objetos

La distancia entre objetos en los planos factoriales del ACP permite interpretar la similitud o diferencia entre las observaciones cuando estas no son anónimas (por ejemplo, ciudades, regiones, universidades, empresas o individuos). Su análisis es análogo al utilizado para las variables y se basa en los mismos indicadores: coordenadas, contribuciones y cosenos cuadrados.

En la interpretación gráfica es importante considerar que:

- Si dos objetos aparecen cerca en cualquier parte del plano factorial, significa que comparten características similares y, por tanto, presentan perfiles parecidos.
- Si dos objetos se encuentran lejos en el plano, esto indica que difieren notablemente en las variables analizadas.
- Si los objetos se ubican en cuadrantes opuestos o muestran coordenadas con signos contrarios en los factores, esto sugiere que presentan características opuestas y, por lo tanto, representan perfiles antagónicos.

De esta forma, la distancia entre objetos ayuda a identificar agrupamientos, contrastes y patrones relevantes dentro del conjunto de observaciones.

Relaciones entre objetos y variables: biplots

Un biplot es una representación gráfica simultánea de objetos y variables sobre el mismo plano factorial. Este tipo de gráfico permite analizar de forma conjunta cómo se relacionan las observaciones con las variables originales del estudio.

La interpretación se basa en la proximidad entre los objetos y los vectores que representan a las variables:

- Los objetos situados cerca de la punta de un vector presentan valores altos en esa variable.
- Los objetos alejados del vector o en dirección opuesta presentan valores bajos o contrarios.
- La relación se evalúa mediante los mismos indicadores utilizados previamente: coordenadas, contribuciones y cosenos cuadrados.

En conjunto, los biplots permiten identificar qué variables explican a cada grupo de objetos y facilitan la interpretación de la estructura global del ACP.

Variables suplementarias

Las variables suplementarias no intervienen en el cálculo del ACP, pero se proyectan sobre los componentes obtenidos con las variables activas. Su interpretación se basa en su proximidad con las variables activas en el plano factorial.

Objetos suplementarios

Los objetos suplementarios son observaciones que tampoco participan en la construcción del ACP. Su proyección en el espacio factorial permite evaluar su similitud con los objetos activos sin alterar el análisis principal.