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 Martínez 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.

Fundamento teórico del Análisis de componentes principales

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 identifica direcciones de máxima varianza mediante combinaciones lineales. Por ello, se asume que la estructura subyacente del conjunto de datos puede capturarse adecuadamente a través de relaciones lineales entre las variables. Relaciones no lineales no son detectadas por este método.

2. **Escalas comparables entre variables:** Debido a que el ACP maximiza la varianza, las variables deben tener escalas similares para evitar que aquellas con valores mayores dominen los componentes. Cuando las unidades son heterogéneas, se recomienda aplicar estandarización

3. **Varianza significativa en las variables:** Variables con varianza muy baja o casi constantes no contribuyen información relevante y pueden distorsionar la estructura de los componentes. El ACP requiere que las variables tengan variabilidad apreciable.

4. **Número adecuado de observaciones:** Para obtener estimaciones estables de la matriz de covarianzas o correlaciones, es recomendable que el número de observaciones supere ampliamente al número de variables.

5. **Ausencia de multicolinealidad perfecta:** Aunque el ACP maneja bien la colinealidad, no puede aplicarse cuando algunas variables son combinaciones lineales exactas de otras, ya que la matriz de covarianzas se vuelve singular y no puede descomponerse.

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.

Las componentes principales como direcciones de máxima varianza

Contexto inicial y preparación de datos

Partimos de una matriz de datos original:

$$X_{n\times p} = \{x_{ij}\}$$

donde: - n: número de objetos (individuos, ciudades, etc.) - p: número de variables - x_{ij}: valor de la variable j en el objeto i

Estadísticos básicos por variable:

- Media: $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$
- Varianza: $s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$

Matriz de datos centrados y estandarizados:

$$Y_{n\times p} = \{y_{ij}\} = \left\{\frac{x_{ij} - \bar{x}_j}{s_j}\right\}$$

Propiedades de Y: - Media cero: $\bar{Y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = 0$ - Varianza unitaria: $var(Y_j) = \frac{1}{n}\sum_{i=1}^{n} y_{ij}^2 = 1$

Representación de la matriz Y

Podemos ver Y de dos formas:

1. **Como vectores fila:** $y_i' = (y_{i1}, y_{i2}, \dots, y_{ip})$ para $i = 1, \dots, n$
   $\rightarrow$ Análisis de similitudes entre objetos

2. **Como vectores columna:** $Y_j = (y_{1j}, y_{2j}, \dots, y_{nj})'$ para $j = 1, \dots, p$
   $\rightarrow$ Análisis de asociaciones entre variables

Primera componente principal: dirección de máxima varianza

Objetivo:

Encontrar un vector de ponderaciones $u_1' = (u_{11}, \dots, u_{p1})$ normalizado ($u_1' u_1 = 1$) que maximice la varianza de la combinación lineal:

$$z_{i1} = u_1' y_i = \sum_{j=1}^{p} u_{j1} y_{ij}$$

Formulación del problema de optimización:

La varianza de z_1 es:

$$var(z_1) = \frac{1}{n} \sum_{i=1}^{n} z_{i1}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} u_{j1} y_{ij} \right)^2$$

Problema de optimización:

$$\max_{u_{11},\dots,u_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} u_{j1} y_{ij} \right)^2 \right\}$$

sujeto a: $\sum_{j=1}^{p} u_{j1}^2 = 1$

Resolución mediante multiplicadores de Lagrange

Función de Lagrange:

$$\mathcal{L}(u_1, \lambda) = u_1' R u_1 - \lambda(u_1' u_1 - 1)$$

donde $R = \frac{1}{n} Y'Y$ es la **matriz de correlación** (cuando trabajamos con datos estandarizados).

Derivadas e igualación a cero:

1. **Derivada respecto a** u_1:

$$\frac{\partial \mathcal{L}}{\partial u_1} = 2Ru_1 - 2\lambda u_1 = 0$$
$$\Rightarrow Ru_1 = \lambda u_1$$

2. **Derivada respecto a** λ:

$$\frac{\partial \mathcal{L}}{\partial \lambda} = u_1' u_1 - 1 = 0$$
$$\Rightarrow u_1' u_1 = 1$$

Interpretación del resultado:

La ecuación $Ru_1 = \lambda u_1$ nos dice que: - u_1 es un **vector propio** de la matriz R - λ es el **valor propio** correspondiente

Para **maximizar la varianza**, elegimos el **mayor valor propio** λ_1 de R, y su vector propio correspondiente u_1.

Segunda componente principal

Objetivo:

Encontrar un segundo vector $u_2' = (u_{12}, \dots, u_{p2})$ que: - Sea ortogonal a u_1: $u_1' u_2 = 0$ - Esté normalizado: $u_2' u_2 = 1$ - Maximice la varianza de $z_2 = Y u_2$

Problema de optimización:

$$\max_{u_2} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} u_{j2} y_{ij} \right)^2 \right\}$$

sujeto a: - $u_2' u_2 = 1$ - $u_1' u_2 = 0$

Función de Lagrange ampliada:

$$\mathcal{L}(u_2, \lambda_2, \gamma) = u_2' R u_2 - \lambda_2 (u_2' u_2 - 1) - \gamma u_1' u_2$$

Derivadas:

$$\frac{\partial \mathcal{L}}{\partial u_2} = 2Ru_2 - 2\lambda_2 u_2 - \gamma u_1 = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_2} = u_2' u_2 - 1 = 0$$

$$\frac{\partial \mathcal{L}}{\partial \gamma} = u_1' u_2 = 0$$

Demostración de que $\gamma = 0$:

Multiplicamos la primera ecuación por u_1' por la izquierda:

$$2u_1' R u_2 - 2\lambda_2 u_1' u_2 - \gamma u_1' u_1 = 0$$

Sabemos que: - $u_1' u_2 = 0$ (restricción de ortogonalidad) - $u_1' u_1 = 1$ (normalización) - $u_1' R u_2 = u_1' (\lambda_1 u_1)' u_2 = \lambda_1 u_1' u_2 = 0$

Por tanto:

$$0 - 0 - \gamma(1) = 0 \Rightarrow \gamma = 0$$

Solución:

Con $\gamma = 0$, la ecuación se reduce a:

$$Ru_2 = \lambda_2 u_2$$

Nuevamente, u_2 es un **vector propio** de R, y para maximizar la varianza elegimos el **segundo mayor valor propio** λ_2.

Componentes principales generales

Para el -ésimo componente:

$$z_{i\alpha} = \sum_{j=1}^{p} u_{j\alpha} y_{ij} = y_i' u_\alpha$$

$$z_\alpha = Y u_\alpha$$

donde u_α es el -ésimo vector propio de R asociado al -ésimo mayor valor propio λ_α.

Propiedades:

1. **Media cero:** $m(z_\alpha) = 0$
2. **Varianza:** $var(z_\alpha) = \lambda_\alpha$
3. **Ortogonalidad:** $z_\alpha' z_\beta = 0$ para $\alpha \neq \beta$

Interpretación geométrica

Cada componente principal z_α representa la **proyección** de los datos originales sobre la dirección definida por el vector propio u_α.

La **varianza explicada** por cada componente es exactamente igual al valor propio correspondiente:

$$var(z_\alpha) = \lambda_\alpha$$

El **porcentaje de varianza explicada** por la -ésima componente es:

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

Resumen del procedimiento

1. **Centrar y estandarizar** los datos: $Y = \left\{ \frac{x_{ij} - \bar{x}_j}{s_j} \right\}$
2. **Calcular matriz de correlación:** $R = \frac{1}{n} Y'Y$
3. **Descomposición espectral:** Encontrar valores y vectores propios de R
4. **Ordenar** valores propios en forma descendente: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$
5. **Componentes principales:** $z_\alpha = Y u_\alpha$, donde u_α es el -ésimo vector propio
6. **Seleccionar** las primeras q componentes que acumulen suficiente varianza

Esta aproximación del ACP busca **direcciones de máxima varianza** en los datos, proporcionando una base ortogonal óptima para representar la información contenida en la matriz original con dimensionalidad reducida.

Las componentes principales como aproximación por mínimos cuadrados de la matriz de datos

A diferencia del enfoque de la sección anterior donde las componentes principales se introdujeron como las direcciones en las que los datos presentan la máxima variabilidad aquí se presenta una

perspectiva equivalente pero basada en aproximar la matriz de datos mediante un modelo de dimensiones reducidas, utilizando el criterio de mínimos cuadrados.

El punto de partida es nuevamente la matriz estandarizada:

$$Y_{n\times p} = \{y_{ij}\} = \left\{ \frac{x_{ij} - \bar{x}_j}{s_j} \right\}$$

El objetivo es representar a Y mediante una versión "simplificada", es decir, utilizando solo unas pocas componentes principales. Esta idea se formaliza construyendo una aproximación de rango reducido de la forma:

$$\tilde{\mathbf{Y}}^{(q)} = \mathbf{Z}_q \mathbf{U}_q'$$

donde:

- $\mathbf{U}_q$ es la matriz formada por los primeros q vectores propios de $\mathbf{R}$, de dimensión $p \times q$.
- $\mathbf{Z}_q = \mathbf{Y}\mathbf{U}_q$ contiene los scores o componentes principales, de tamaño $n \times q$.
- $q < p$ es el número de dimensiones deseadas.

Esta representación indica que cualquier fila de $\mathbf{Y}$, originalmente un vector en $\mathbb{R}^p$, se aproxima mediante una combinación lineal de solo q direcciones ortogonales.

Componente principal

Entonces igual que antes se sabe que una componente es una combinación lineal:

$$Z_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p$$

con

$$\sum_{j=1}^{p} u_{j1}^2 = 1$$

Primera componente

$$z_{i1} = \mathbf{u}_1' \mathbf{y}_i = \sum_{j=1}^{p} u_{j1} y_{ij}$$

- y_{ij}: datos centrados y estandarizados.
- $\mathbf{u}_1$: vector de pesos (vector propio).
- z_{i1}: proyección de cada individuo sobre la dirección de la primera componente.

Como cada columna de $\mathbf{Y}$ tiene media 0: - la componente también tiene media cero. - Su varianza es:

$$\text{var}(\mathbf{z}_1) = \frac{1}{n} \sum_{i=1}^{n} z_{i1}^2$$

Problema de optimización

$$\max_{\mathbf{u}_1} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} u_{j1} y_{ij} \right)^2 \quad \text{sujeto a} \quad \sum_{j=1}^{p} u_{j1}^2 = 1$$

Interpretación geométrica

En este caso se tiene que La primera componente es la recta que minimiza las distancias cuadráticas a los puntos. En donde esa recta pasa por el origen y tiene dirección u_1.

Tambien se puede decir que la razón por la que esta recta es óptima es que minimiza la suma de las distancias cuadráticas entre cada punto original y su proyección sobre la recta. Lo cual es equivalente a la definición basada en maximizar la varianza explicada.

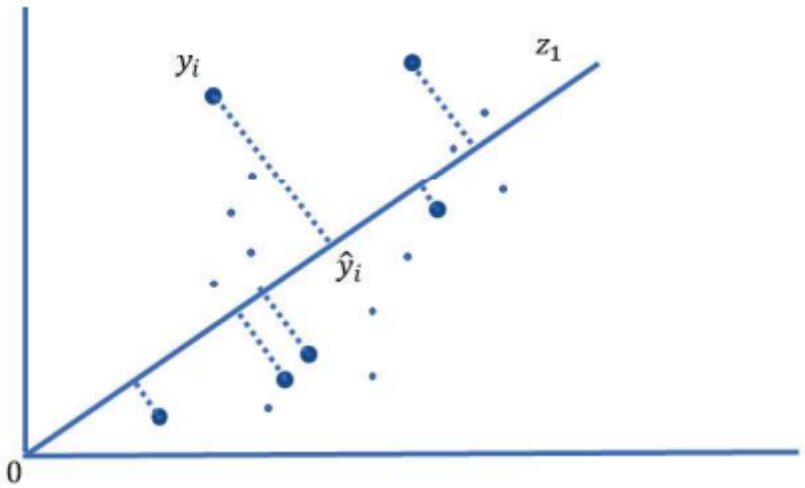


Figura 1: Interpretación geometrica de una componente principal

Si z_{i1} es la coordenada del dato $\mathbf{y}_i$ sobre la primera componente, entonces su proyección sobre la recta es:

$$\hat{\mathbf{y}}_i = z_{i1} \mathbf{u}_1$$

Esta proyección es la mejor aproximación posible del punto original usando solo una dimensión.

Aproximación usando q componentes

De esta manera el ACP permite representar cada observación original mediante una combinación de solo las primeras q componentes principales, reduciendo la dimensión del problema sin perder demasiada información.

Para cada dato original y_{ij}, su aproximación utilizando únicamente q componentes está dada por:

$$\hat{y}_{ij} = \sum_{r=1}^{q} z_{ir} u_{jr},$$

donde:

- z_{ir} son los **scores** de la observación i sobre la componente r, es decir, indican qué tan lejos está el individuo en esa dirección.
- u_{jr} son los **loadings**, que describen cómo contribuye la variable j en la componente r.

Asi, esta descomposición muestra que el ACP reconstruye los datos como la multiplicación:

$$\text{Datos} \approx \text{Scores} \times \text{Loadings}^{\text{T}}.$$

esta aproximación corresponde a proyectar los datos sobre el subespacio generado por los primeros q vectores propios, lo cual constituye la mejor aproximación posible en el sentido de mínimos cuadrados. Esto debido a que el ACP coincide con la descomposición espectral, asi, los vectores propios asociados a los valores propios más grandes generan la mejor aproximación posible de rango q.

Calidad de la aproximación

Para ver que tan buena seria esta aproximación se puede pensar en analizar si las componentes capturan mucha o poca varianza, para esto, la varianza total (SCT) esta dada por:

$$\sum_{j=1}^{p} s_j^2 = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} y_{ij}^2,$$

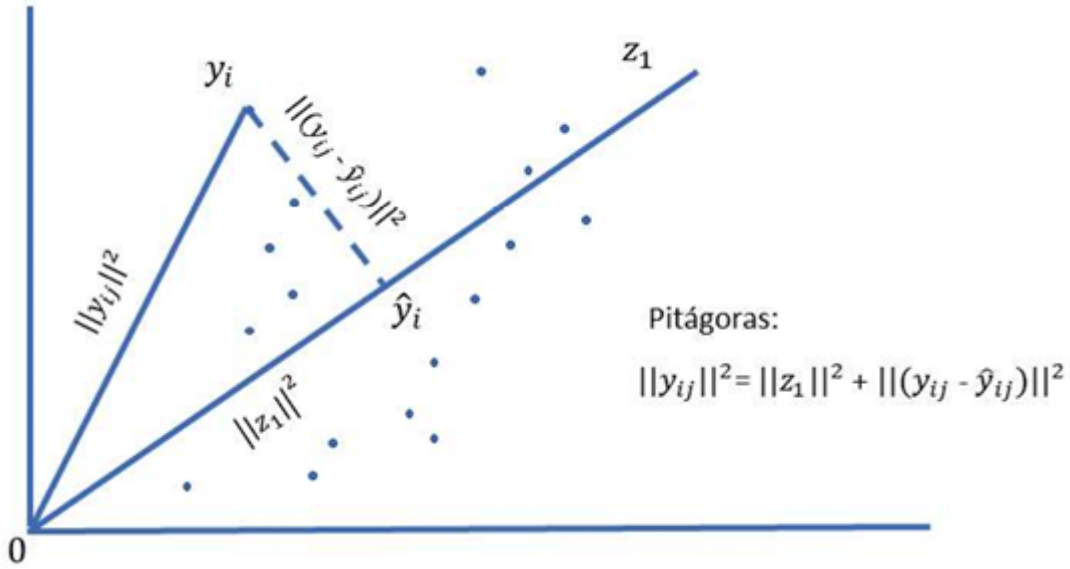
y para calcular la varianza explicada por la r-esima componente, la varianza de esos scores es:

$$\frac{1}{n} \sum_{i=1}^{n} z_{ir}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} u_{jr} y_{ij} \right)^2$$

Entonces se tiene que el porcentaje de varianza explicada por la r-esima componente es:

$$\frac{\sum_{i=1}^{n} z_{ir}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} y_{ij}^2}.$$

Una perspectiva geométrica clave es que la variabilidad total de cada observación puede descomponerse exactamente en dos partes, la varianza explicada en las componentes principales mas la explicada por las diferencias entre los datos y su aproximación que es justamente la que queda sin explicar, de esta manera:



Esto es:

$$\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} y_{ij}^2 = \sum_{r=1}^{q} \frac{1}{n} \sum_{i=1}^{n} z_{ir}^2 + \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} \left(y_{ij} - \sum_{r=1}^{q} z_{ir} u_{jr} \right)^2$$

ahora dividiendo ambas partes de la ecuacion por $\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} y_{ij}^2$ y despejando se puede observar que:

$$\frac{\sum_{r=1}^{q} \sum_{i=1}^{n} z_{ir}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} y_{ij}^2} = 1 - \frac{\sum_{j=1}^{p} \sum_{i=1}^{n} \left(y_{ij} - \sum_{r=1}^{q} z_{ir} u_{jr} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} y_{ij}^2}$$

que esto es justamente la fracción de la variabilidad total que es explicada por los primeros q componentes principales, esto es lo que se conoce como

$$\tau_q = 1 - \frac{SCR}{SCT}$$

donde SCR es la suma de cuadrados residual y SCT es la suma de cuadrados total y se puede decir que si el SCR es pequeño, significa que la aproximación con q componentes es muy buena, en consecuencia, τ_q será grande, indicando que los componentes capturan una alta proporción de la información original.

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.

Equivalencia entre el enfoque de direcciones de máxima varianza y el de aproximación por mínimos cuadrados

Partiendo de la expansión en valores singulares de la matriz estandarizada

$$\mathbf{Y} = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha} \mathbf{v}_\alpha \mathbf{u}'_\alpha,$$

cada entrada de $\mathbf{Y}$ se escribe como

$$y_{ij} = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha} v_{i\alpha} u_{j\alpha}.$$

luego se sabe que la aproximación de rango q se obtiene truncando esta suma en los primeros q términos:

$$\hat{y}_{ij} = \sum_{\alpha=1}^{q} \sqrt{\lambda_\alpha} v_{i\alpha} u_{j\alpha}.$$

Entonces, se puede ver que el residuo de la aproximación es exactamente:

$$y_{ij} - \hat{y}_{ij} = \sum_{\alpha=q+1}^{p} \sqrt{\lambda_\alpha} v_{i\alpha} u_{j\alpha}.$$

Al tomar sumas de cuadrados sobre todas las observaciones y variables y usar la ortonormalidad de los autovectores, se obtiene que la suma de cuadrados residual queda dada por la suma de los autovalores descartados, esto es;

$$SCR = \sum_{j=1}^{p} \sum_{i=1}^{n} (y_{ij} - \hat{y}_{ij})^2 = \sum_{\alpha=q+1}^{p} \lambda_\alpha,$$

mientras que la suma de cuadrados total satisface $SCT = \sum_{\alpha=1}^{p} \lambda_\alpha$. De aquí se puede deduce que minimizar la suma de cuadrados residual equivale a maximizar la varianza explicada por las q

primeras componentes;

$$\tau_q = \frac{\sum_{\alpha=1}^{q} \lambda_\alpha}{\sum_{\alpha=1}^{p} \lambda_\alpha} = 1 - \frac{SCR}{SCT}.$$

En consecuencia, las direcciones que resuelven el problema de máxima varianza y las que resuelven el problema de aproximación por mínimos cuadrados coinciden, asi, ambas son los vectores propios de $\mathbf{Y}'\mathbf{Y}$ asociados a los mayores autovalores, entonces las soluciones son equivalentes.

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 cercaen 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.