ntic master School

UNIVERSIDAD COMPLUTENSE DE MADRID

Análisis de Componentes Principales

Minería de Datos y Modelización Predictiva

Máster en Big Data y Business Analytics
Universidad Complutense de Madrid
Curso 2023-2024

Cuando se recoge información en una muestra de datos, lo más frecuente es tomar el mayor número posible de variables. Sin embargo, disponer de un amplio número de variables puede ocasionar diferentes problemas. En este contexto complejo y **multidimensional**, surge la necesidad de técnicas que puedan simplificar y resumir la información de manera eficiente. Aquí es donde el **Análisis de Componentes Principales (ACP)** juega un papel fundamental.

Cuando se recoge información en una muestra de datos, lo más frecuente es tomar el mayor número posible de variables. Sin embargo, disponer de un amplio número de variables puede ocasionar diferentes problemas. En este contexto complejo y **multidimensional**, surge la necesidad de técnicas que puedan simplificar y resumir la información de manera eficiente. Aquí es donde el **Análisis de Componentes Principales (ACP)** juega un papel fundamental.

El ACP es una técnica estadística que busca **reducir la dimensionalidad** de un conjunto de datos, manteniendo al mismo tiempo la mayor cantidad de **variabilidad** posible. En un mundo donde los datos masivos son comunes, la posibilidad de **simplificar la información** sin perder sus características esenciales se convierte en una herramienta poderosa.

Uno de los principales problemas que el ACP ayuda a resolver es la **multicolinealidad**, es decir, la **correlación alta** entre diversas variables. Esto puede complicar la interpretación de los datos y afectar la eficacia de otros métodos de análisis. Transformando los datos en **componentes principales**, el ACP elimina este problema, facilitando la interpretación y el análisis.

Uno de los principales problemas que el ACP ayuda a resolver es la **multicolinealidad**, es decir, la **correlación alta** entre diversas variables. Esto puede complicar la interpretación de los datos y afectar la eficacia de otros métodos de análisis. Transformando los datos en **componentes principales**, el ACP elimina este problema, facilitando la interpretación y el análisis.

Además, el ACP permite una **visualización** más clara de los datos. Al reducir la cantidad de variables a unos pocos componentes principales, se pueden representar gráficamente los datos en un espacio de baja dimensión. Esto es invaluable para la comprensión y el descubrimiento de patrones y relaciones en conjuntos de datos complejos.

La **versatilidad** del ACP es otra de sus fortalezas. No se limita a un campo de estudio específico, sino que se aplica en diversas disciplinas como la **biología, economía, psicología**, y más. Su capacidad para detectar estructuras en los datos ha llevado a avances significativos en campos tan diversos como la **genómica y las finanzas**.

La **versatilidad** del ACP es otra de sus fortalezas. No se limita a un campo de estudio específico, sino que se aplica en diversas disciplinas como la **biología, economía, psicología**, y más. Su capacidad para detectar estructuras en los datos ha llevado a avances significativos en campos tan diversos como la **genómica y las finanzas**.

El ACP también contribuye a la mejora de la **calidad de los datos**. Identificar y eliminar las variables redundantes mejora la precisión de los modelos, permitiendo un análisis más ajustado a la realidad. Además, a mayor dimensionalidad de los datos, mayor es la **capacidad computacional** que se precisa. Por lo tanto, reducir la dimensionalidad mediante ACP no solo optimiza los recursos computacionales sino que también facilita el manejo y procesamiento de grandes conjuntos de datos, justificando aún más su aplicación.

¿En que consiste el ACP?

El Análisis de Componentes Principales es una **técnica matemática sofisticada** que se aplica a un conjunto de datos compuesto por **variables** que están potencialmente **correlacionadas** entre sí. El objetivo central del ACP es transformar estas variables originales en un nuevo conjunto de variables, que son **linealmente independientes** y, por lo tanto, **no correlacionadas** entre sí, si la estructura de los datos lo permite. Estas nuevas variables se denominan **componentes principales**.

¿En que consiste el ACP?

El Análisis de Componentes Principales es una **técnica matemática sofisticada** que se aplica a un conjunto de datos compuesto por **variables** que están potencialmente **correlacionadas** entre sí. El objetivo central del ACP es transformar estas variables originales en un nuevo conjunto de variables, que son **linealmente independientes** y, por lo tanto, **no correlacionadas** entre sí, si la estructura de los datos lo permite. Estas nuevas variables se denominan **componentes principales**.

La **transformación** se lleva a cabo mediante una **combinación lineal** de las variables originales, donde se seleccionan los **coeficientes** de manera que el **primer componente principal** capture la mayor parte de la **variabilidad** presente en los datos originales. Esto significa que esta primera componente resume, en la medida de lo posible, la información contenida en las variables originales.

El **segundo componente principal** se calcula de manera similar, pero está restringido a ser **ortogonal** al primer componente, es decir, **no correlacionado** con él. Este segundo componente explica la siguiente mayor parte de la variabilidad que no ha sido capturada por el primer componente.

¿En que consiste el ACP?

El **segundo componente principal** se calcula de manera similar, pero está restringido a ser **ortogonal** al primer componente, es decir, **no correlacionado** con él. Este segundo componente explica la siguiente mayor parte de la variabilidad que no ha sido capturada por el primer componente.

Este proceso continúa de manera **secuencial**, con cada componente principal subsiguiente explicando una **proporción decreciente** de la variabilidad restante, y siempre siendo **ortogonal** a todos los componentes principales anteriores. La transformación se realiza de tal manera que se puede representar la estructura compleja y multidimensional de los datos originales en un número reducido de dimensiones, **sin perder significativamente la información** esencial.

¿En que consiste el ACP?

A continuación, se muestra (a la izquierda) una **nube de puntos** que representa la relación entre las variables originales (X, Y), **altamente correlacionadas**.

A continuación, se muestra (a la izquierda) una **nube de puntos** que representa la relación entre las variables originales (X, Y), **altamente correlacionadas**.

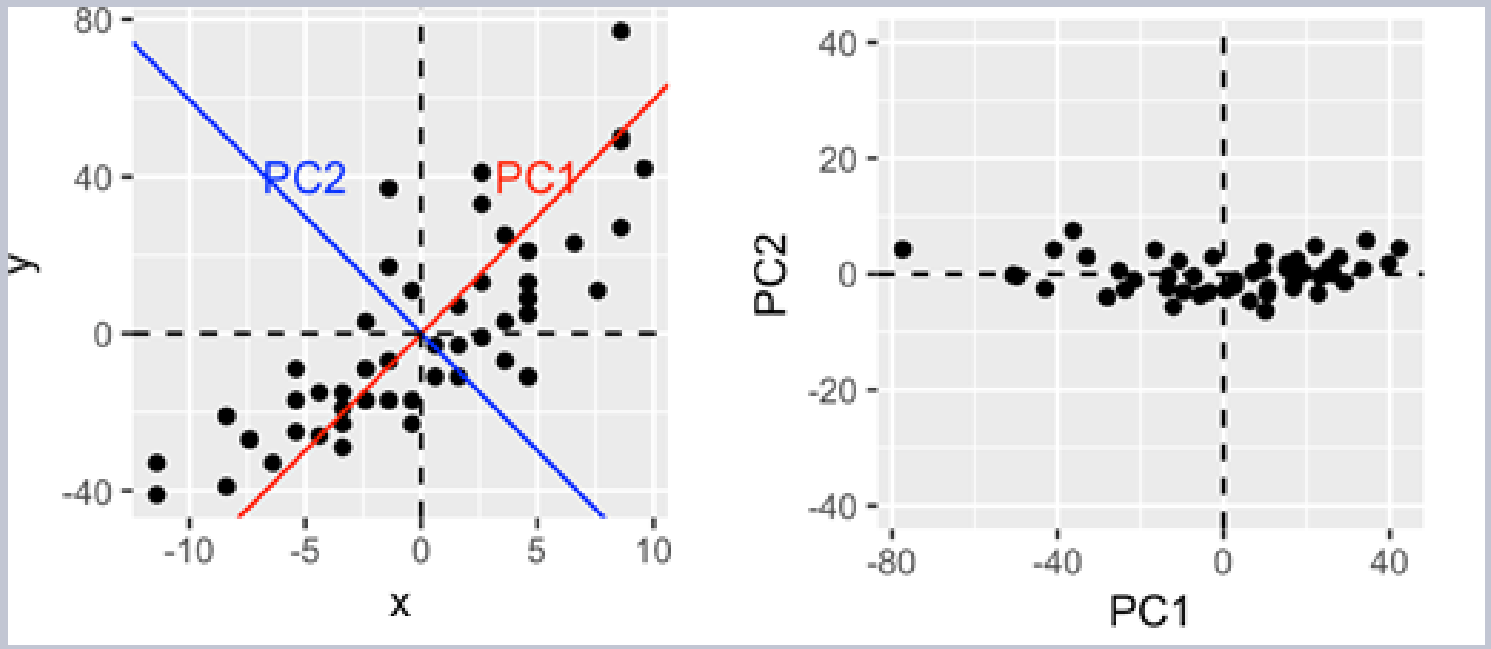

Figura: Representación de las Componentes.

A continuación, se muestra (a la izquierda) una **nube de puntos** que representa la relación entre las variables originales (X, Y), **altamente correlacionadas**.

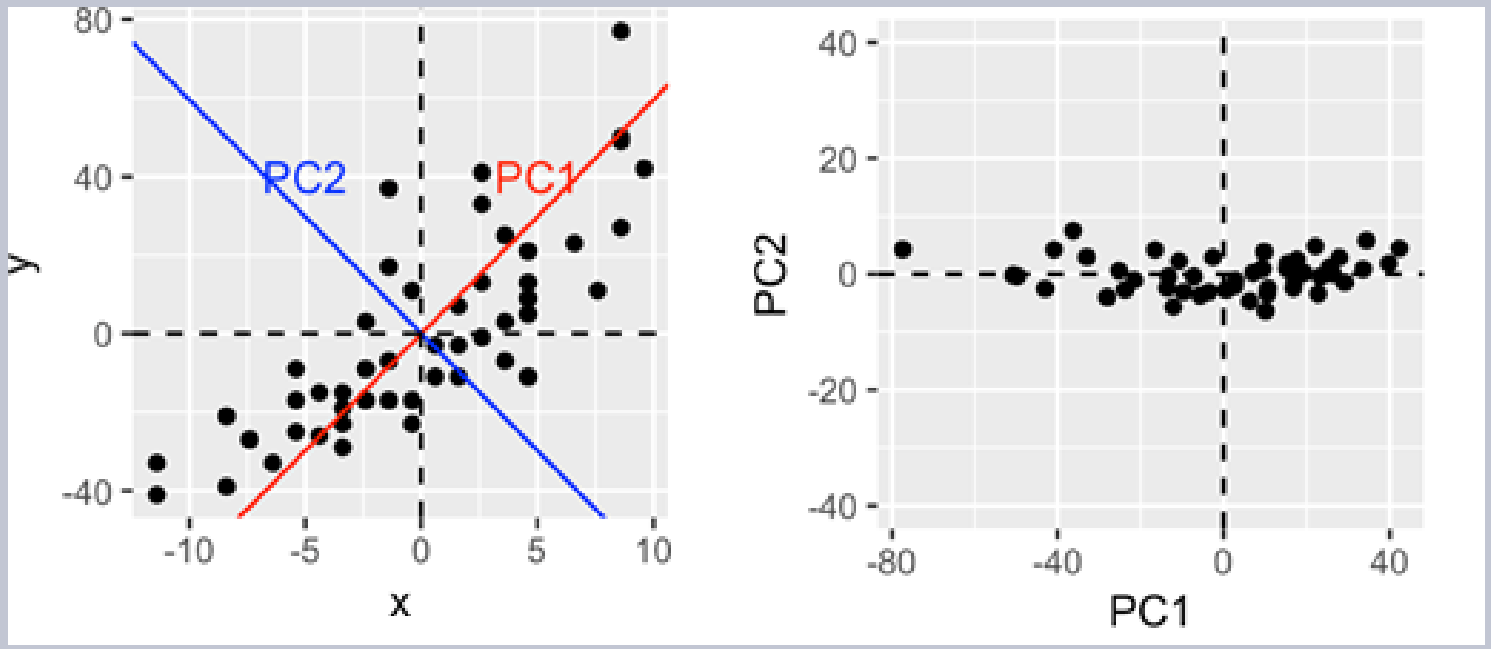

Figura: Representación de las Componentes.

Por otro lado, a la derecha se presenta el resultado del ACP: dos **variables ortogonales** que han sido derivadas de las variables originales (X, Y).

¿En que consiste el ACP?

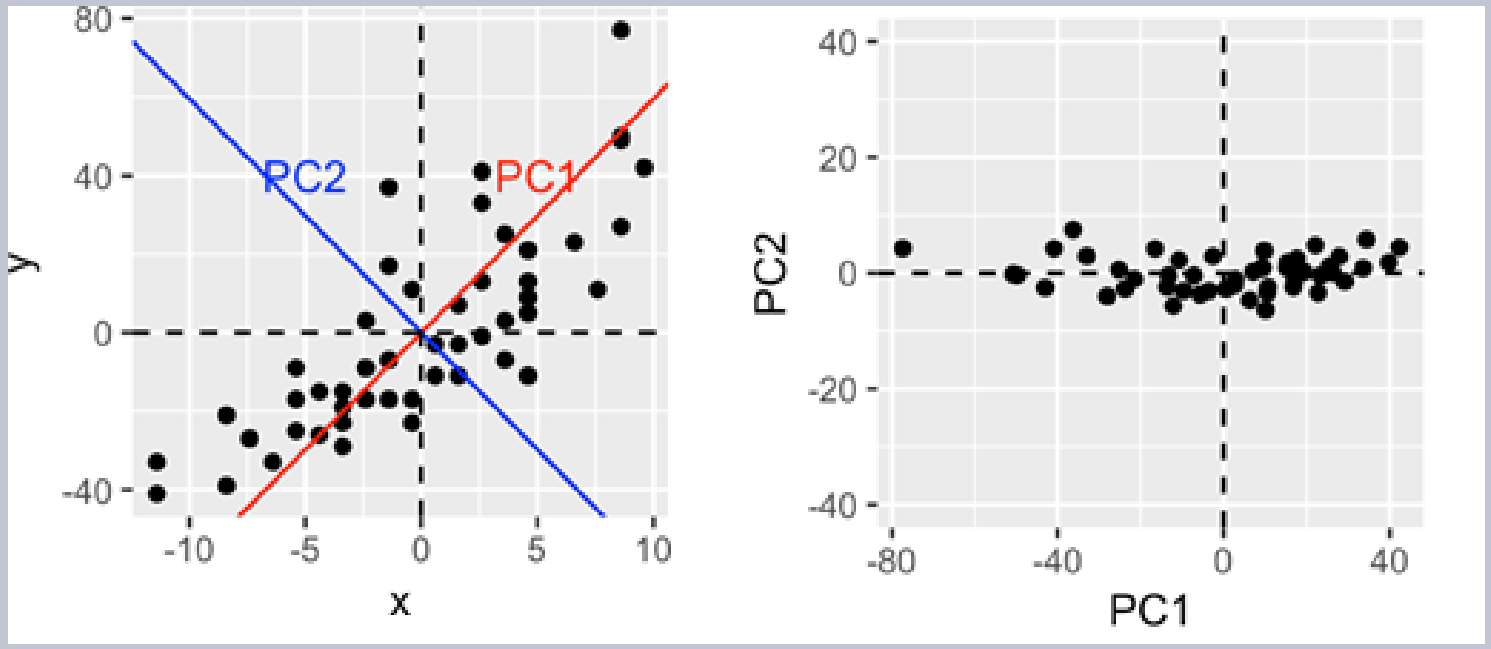

Figura: Representación de las Componentes.

Aunque la **dimensionalidad** no se ha reducido (sigue siendo un espacio bidimensional), lo notable aquí es que la correlación entre las variables originales ha sido eliminada. Las nuevas variables, ahora ortogonales entre sí, ofrecen una vista **descorrelacionada** de los datos, preservando la estructura subyacente pero en un marco donde las variables no están linealmente relacionadas.

Requisitos de los datos

Para que un conjunto de datos sea susceptible de análisis mediante el ACP, es vital cumplir con ciertos requisitos:

Para que un conjunto de datos sea susceptible de análisis mediante el ACP, es vital cumplir con ciertos requisitos:

- **Escala de las Variables**: Las variables de entrada deben estar en la misma escala o ser estandarizadas, para que cada una contribuya de manera equitativa al análisis.

Para que un conjunto de datos sea susceptible de análisis mediante el ACP, es vital cumplir con ciertos requisitos:

- **Escala de las Variables**: Las variables de entrada deben estar en la misma escala o ser estandarizadas, para que cada una contribuya de manera equitativa al análisis.
- **Ausencia de Valores Atípicos**: Los datos extremos pueden dificultar el análisis, y su manejo adecuado es fundamental.

Para que un conjunto de datos sea susceptible de análisis mediante el ACP, es vital cumplir con ciertos requisitos:

- **Escala de las Variables**: Las variables de entrada deben estar en la misma escala o ser estandarizadas, para que cada una contribuya de manera equitativa al análisis.
- **Ausencia de Valores Atípicos**: Los datos extremos pueden dificultar el análisis, y su manejo adecuado es fundamental.
- **Datos perdidos**: La ausencia de datos puede complicar el análisis, por lo que pueden ser necesarias técnicas de imputación.

Para que un conjunto de datos sea susceptible de análisis mediante el ACP, es vital cumplir con ciertos requisitos:

- **Escala de las Variables**: Las variables de entrada deben estar en la misma escala o ser estandarizadas, para que cada una contribuya de manera equitativa al análisis.
- **Ausencia de Valores Atípicos**: Los datos extremos pueden dificultar el análisis, y su manejo adecuado es fundamental.
- **Datos perdidos**: La ausencia de datos puede complicar el análisis, por lo que pueden ser necesarias técnicas de imputación.
- **Distribución de las Variables Originales**: Aunque no es necesario que las variables originales tengan una distribución normal, esto puede mejorar la interpretabilidad de los resultados.

Es fundamental destacar que el ACP presupone correlaciones entre las variables originales. Aunque técnicamente es posible llevar a cabo el análisis en ausencia de correlaciones, el resultado sería trivial y no proporcionaría ninguna comprensión nueva o valiosa acerca de la estructura de los datos. En otras palabras, sin correlaciones entre las variables, el ACP no obtendría ningún resultado útil.

Pasos del Proceso

Supongamos que tenemos un conjunto de datos con n observaciones y p variables donde X_{ij}, es el valor de la observacion i-esima, en la variable j-esima, y queremos reducirlo a k componentes principales, entonces:

Supongamos que tenemos un conjunto de datos con n observaciones y p variables donde X_{ij}, es el valor de la observacion i-esima, en la variable j-esima, y queremos reducirlo a k componentes principales, entonces:

- **Estandarización de los Datos**: Si las escalas de las variables originales difieren, deben ser estandarizadas:

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j}, i \in n, j \in p,$$

donde $\bar{X}_j$ es la media de la variable j-esima, y s_j es su desviación estándar.

Supongamos que tenemos un conjunto de datos con n observaciones y p variables donde X_{ij}, es el valor de la observacion i-esima, en la variable j-esima, y queremos reducirlo a k componentes principales, entonces:

- **Estandarización de los Datos**: Si las escalas de las variables originales difieren, deben ser estandarizadas:

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j}, i \in n, j \in p,$$

donde $\bar{X}_j$ es la media de la variable j-esima, y s_j es su desviación estándar.

- **Cálculo de la Matriz de Covarianzas:** Si los datos originales están en la misma escala, (poco habitual), podremos utilizar la matriz de covarianzas, la cual, es una **matriz cuadrada** $p \times p$ que contiene las covarianzas entre cada par de variables originales. Se calcula como:

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

donde C es la matriz de covarianzas y $\bar{X}$ es el vector de medias de las variables originales.

- **Cálculo de la Matriz de Correlaciones:** Si las variables originales no están medidas en la misma escala, entonces, utilizaremos la matriz de correlaciones, la cual, se calcula a partir de los datos estandarizados y vuelve a ser una matriz cuadrada $p \times p$ que contiene las correlaciones entre cada par de variables estandarizadas. Se calcula como:

$$R = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Z_i - \bar{Z})^T = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T,$$

donde Z_i es la i-ésima fila de la matriz estandarizada Z, y $\bar{Z}$ es el vector de las medias de Z.

- **Cálculo de la Matriz de Correlaciones:** Si las variables originales no están medidas en la misma escala, entonces, utilizaremos la matriz de correlaciones, la cual, se calcula a partir de los datos estandarizados y vuelve a ser una matriz cuadrada $p \times p$ que contiene las correlaciones entre cada par de variables estandarizadas. Se calcula como:

$$R = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Z_i - \bar{Z})^T = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T,$$

donde Z_i es la i-ésima fila de la matriz estandarizada Z, y $\bar{Z}$ es el vector de las medias de Z.

- **Determinación de los Valores Propios:** A partir de la matriz de covarianza (o de correlaciones), se calculan los valores propios, resolviendo la ecuación característica $det(\Sigma - \lambda I) = 0$, donde Σ es la matriz de covarianzas o la matriz de correlaciones, I es la matriz identidad, matriz de dimensión $p \times p$ con unos en la diagonal y ceros en todas las demás posiciones; y λ es el vector que contiene el conjunto de valores propios

- **Determinación de los k primeros Valores Propios:** Los autovalores se ordenan en orden descendente, y se eligen los primeros k, que representarán los componentes principales. Para elegir estos k primeros, hay que tener en cuenta que el valor de cada uno de los autovalores λ_i para $i = 1, ..., p$, es proporcional a la variabilidad explicada por las componentes asociadas. Por ejemplo, estando los autovalores ordenados, tal que $\lambda 1 \geq \lambda_2 \geq ... \geq \lambda_p$, la componente asociada al primer autovalor λ_1, explica un $\frac{\lambda_1}{\sum_{i=1}^{p} \lambda_1} * 100$ por ciento de la variabilidad total de las variables originales.

- **Determinación de los k primeros Valores Propios:** Los autovalores se ordenan en orden descendente, y se eligen los primeros k, que representarán los componentes principales. Para elegir estos k primeros, hay que tener en cuenta que el valor de cada uno de los autovalores λ_i para $i = 1, ..., p$, es proporcional a la variabilidad explicada por las componentes asociadas. Por ejemplo, estando los autovalores ordenados, tal que $\lambda 1 \geq \lambda_2 \geq ... \geq \lambda_p$, la componente asociada al primer autovalor λ_1, explica un $\frac{\lambda_1}{\sum_{i=1}^{p} \lambda_1} * 100$ por ciento de la variabilidad total de las variables originales.
- **Determinación de los vectores propios asociados:** Una vez que tenemos los valores propios λ, podemos encontrar los autovectores asociados, resolviendo la siguiente ecuación lineal para cada λ: $(\Sigma - \lambda I)x = 0$. Donde x es el vector propio que estamos buscando.

- **Construcción de la Matriz de Transformación:** Los primeros k vectores propios forman la matriz de transformación. Es decir, $W = [w_1, w_2, \ldots, w_k]$, donde w_i, para $i = 1, \ldots, k$ son los primeros k vectores propios.

- **Construcción de la Matriz de Transformación:** Los primeros k vectores propios forman la matriz de transformación. Es decir, $W = [w_1, w_2, \ldots, w_k]$, donde w_i, para $i = 1, \ldots, k$ son los primeros k vectores propios.
- **Cálculo de los Componentes Principales:** Una vez que se ha obtenido la matriz de transformación W, las componentes principales CP se pueden calcular como la proyección de los datos originales X o los datos estandarizados Z en el espacio de las nuevas dimensiones (componentes principales). Esto se realiza mediante una multiplicación matricial entre los datos y la matriz de transformación:

$$CP = XW \quad \text{ó} \quad CP = ZW$$

donde CP es una matriz de dimensiones $n \times k$ que contiene los componentes principales de los datos. Cada columna de CP representa un componente principal, y cada fila de CP representa los componentes principales para una observación individual en el conjunto de datos original. Este paso produce las coordenadas de las observaciones en el nuevo espacio definido por los componentes principales.

Pese a que las variables se encuentren en la misma escala, siempre es aconsejable estandarizar los datos, por los siguientes motivos:

Pese a que las variables se encuentren en la misma escala, siempre es aconsejable estandarizar los datos, por los siguientes motivos:

- **Influencia de Variables:** Si no estandarizas, las variables con mayor variabilidad tendrán más peso en la formación de componentes principales. Esto puede ser problemático si la variabilidad es alta debido a ruido o errores de medición, y no porque la variable sea particularmente importante para el fenómeno estudiado.

Pese a que las variables se encuentren en la misma escala, siempre es aconsejable estandarizar los datos, por los siguientes motivos:

- **Influencia de Variables:** Si no estandarizas, las variables con mayor variabilidad tendrán más peso en la formación de componentes principales. Esto puede ser problemático si la variabilidad es alta debido a ruido o errores de medición, y no porque la variable sea particularmente importante para el fenómeno estudiado.
- **Comparabilidad:** Si planeas comparar los resultados de tu ACP con otros estudios que también utilizan ACP pero en diferentes escalas, la estandarización podría ser esencial para hacer que los resultados sean comparables.

Una de las **decisiones** fundamentales en el Análisis de Componentes Principales es la selección del número de componentes principales, denotado como k, que se deben retener.

Selección del Número de Componentes

Una de las **decisiones** fundamentales en el Análisis de Componentes Principales es la selección del número de componentes principales, denotado como k, que se deben retener.

La elección de k es esencial para garantizar que el ACP capture una **cantidad significativa de la variabilidad** presente en los datos originales sin incluir componentes innecesarios. Una técnica común para determinar k es utilizar el ***método del codo***, que se basa en el gráfico de la varianza explicada.

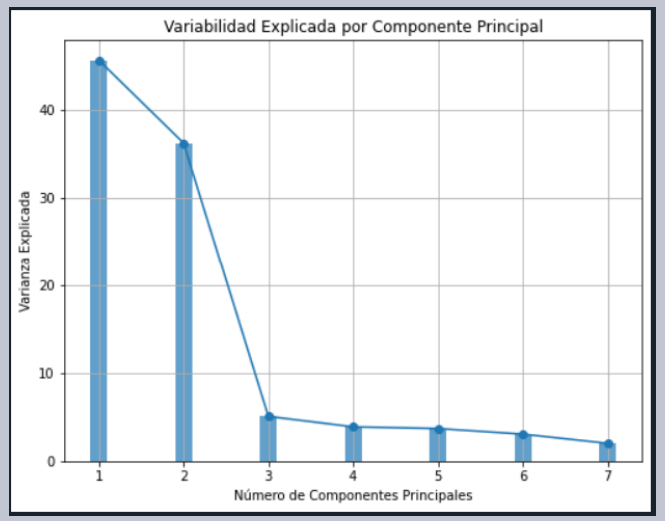

Figura: Representación de la variabilidad explicada

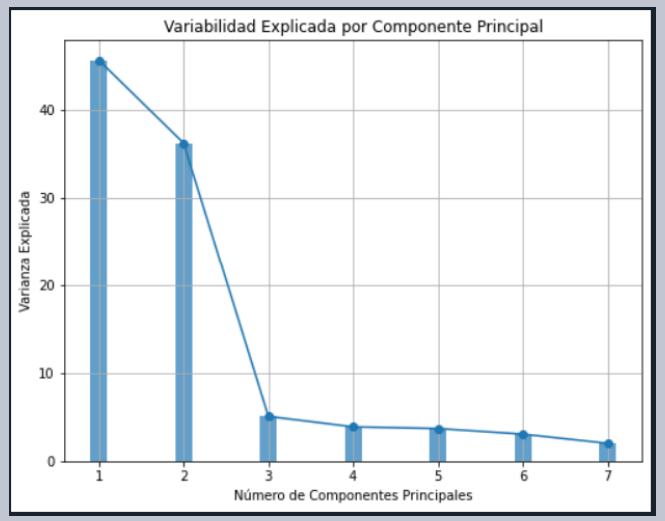

Figura: Representación de la variabilidad explicada

En este gráfico, el eje x representa el número de componentes y el eje y representa la proporción de varianza explicada. Buscamos el punto en el que agregar componentes adicionales ya no proporciona una ganancia significativa en la varianza explicada. Este punto se asemeja a la forma de un 'codoén la curva, y marca el número óptimo de componentes a retener.

Relación entre las Componentes Principales y las Variables

En el contexto del Análisis de Componentes Principales, es fundamental comprender la relación entre las componentes principales y las variables originales. Esta relación se basa en conceptos clave:

En el contexto del Análisis de Componentes Principales, es fundamental comprender la relación entre las componentes principales y las variables originales. Esta relación se basa en conceptos clave:

- λ_i: Representa el autovalor asociado a la i-ésima componente principal (CP_i) obtenido durante el ACP.Los autovalores indican cuánta varianza (variabilidad) se explica por cada CP_i. Un autovalor alto (λ_i grande) implica que CP_i retiene una gran cantidad de información o variabilidad de los datos originales, mientras que un autovalor bajo (λ_i pequeño) indica que CP_i retiene menos información.

- e_{ij}: Representa el coeficiente ubicado en la posición ij del vector propio asociado a la CP_i. Este coeficiente indica la contribución relativa de la variable original X_j en la construcción de la CP_i. En otras palabras, e_{ij} muestra cuánto peso tiene la variable X_j en la definición de los valores de CP_i.

- e_{ij}: Representa el coeficiente ubicado en la posición ij del vector propio asociado a la CP_i. Este coeficiente indica la contribución relativa de la variable original X_j en la construcción de la CP_i. En otras palabras, e_{ij} muestra cuánto peso tiene la variable X_j en la definición de los valores de CP_i.
- $Cov(CP_i, X_j)$: La covarianza entre la i-ésima componente principal (CP_i) y la j-ésima variable original (X_j) se calcula como $\lambda_i \cdot e_{ij}$. Esta covarianza nos proporciona información sobre la relación lineal entre CP_i y X_j. Si el valor resultante es positivo, significa que ambas variables tienden a aumentar o disminuir juntas (correlación positiva). Si es negativo, indica que una variable tiende a aumentar cuando la otra disminuye (correlación negativa). Un valor cercano a cero sugiere que no hay una fuerte relación lineal entre CP_i y X_j.

- $Corr(CP_i, X_j)$: La correlación entre una componente principal CP_i y una variable X_j se calcula como

$$Corr(CP_i, X_j) = \frac{Cov(CP_i, X_j)}{\sqrt{Var(CP_i)Var(X_j)}} = \frac{\lambda_i e_{ij}}{\sqrt{\lambda_i S_j^2}} = e_{ij}\frac{\sqrt{\lambda_i}}{S_j},$$

y estandariza el valor de la covarianza anterior entre $[-1, 1]$. Esta estandarización es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la raíz cuadrada del autovalor λ_i y la desviación estándar de la variable X_j, denotada como S_j.

Contribuciones de las CP a la Variabilidad Explicada de las VO

Para evaluar si las nuevas variables generadas por el ACP capturan eficazmente la información contenida en las variables originales (VO), y para determinar qué variables se explican mejor (o peor) mediante estas nuevas componentes, utilizamos una medida conocida como çosenos al cuadradoçalculados como:

Para evaluar si las nuevas variables generadas por el ACP capturan eficazmente la información contenida en las variables originales (VO), y para determinar qué variables se explican mejor (o peor) mediante estas nuevas componentes, utilizamos una medida conocida como çosenos al cuadradoçalculados como:

$$Cos_{ij}^2 = Corr^2(CP_i, X_j).$$

Contribuciones de las CP a la Variabilidad Explicada de las VO

Para evaluar si las nuevas variables generadas por el ACP capturan eficazmente la información contenida en las variables originales (VO), y para determinar qué variables se explican mejor (o peor) mediante estas nuevas componentes, utilizamos una medida conocida como çosenos al cuadradoçalculados como:

$$Cos_{ij}^2 = Corr^2(CP_i, X_j).$$

Los cosenos al cuadrado, a pesar de su nombre, no están relacionados directamente con la función coseno, sino que se derivan de la geometría subyacente del ACP. Estos valores representan las correlaciones al cuadrado entre cada variable original y las componentes principales. Expresan la proporción de la varianza de cada variable original que se explica mediante cada componente principal.

Contribuciones de las CP a la Variabilidad Explicada de las VO

En otras palabras, los cosenos al cuadrado nos indican cuánto de la variabilidad de una variable original está contenida en cada componente principal. Un valor cercano a 1 implica que la variable se explica en gran medida por la componente, mientras que un valor cercano a 0 indica que la componente tiene poco poder para explicar la variabilidad de esa variable.

Contribuciones de las CP a la Variabilidad Explicada de las VO

En otras palabras, los cosenos al cuadrado nos indican cuánto de la variabilidad de una variable original está contenida en cada componente principal. Un valor cercano a 1 implica que la variable se explica en gran medida por la componente, mientras que un valor cercano a 0 indica que la componente tiene poco poder para explicar la variabilidad de esa variable.

El uso de los cosenos al cuadrado es esencial para evaluar la capacidad del ACP para resumir y retener información relevante de las variables originales, lo que nos permite identificar qué variables son más influyentes en la definición de cada componente principal.

Contribuciones de las VO a las CP

Para comprender cómo cada variable original contribuye a la construcción de las CP en el Análisis de Componentes Principales, podemos calcular su contribución específica. La contribución de una variable en la creación de una CP se calcula mediante la fórmula:

Para comprender cómo cada variable original contribuye a la construcción de las CP en el Análisis de Componentes Principales, podemos calcular su contribución específica. La contribución de una variable en la creación de una CP se calcula mediante la fórmula:

$$Contribucion_{ji} = Cos^2_{ij}\sqrt{\lambda_i}$$ para todo i=1,...,k; j=1,...,p.

Para comprender cómo cada variable original contribuye a la construcción de las CP en el Análisis de Componentes Principales, podemos calcular su contribución específica. La contribución de una variable en la creación de una CP se calcula mediante la fórmula:

$$Contribucion_{ji} = Cos^2_{ij}\sqrt{\lambda_i}$$ para todo i=1,...,k; j=1,...,p.

Esta fórmula nos permite cuantificar la influencia de cada variable en la definición de una CP. Al evaluar las contribuciones de las variables, podemos identificar cuáles son las más relevantes para cada CP y comprender mejor cómo se estructuran las CP en función de las variables originales.

Para comprender cómo cada variable original contribuye a la construcción de las CP en el Análisis de Componentes Principales, podemos calcular su contribución específica. La contribución de una variable en la creación de una CP se calcula mediante la fórmula:

$$Contribucion_{ji} = Cos^2_{ij}\sqrt{\lambda_i}$$ para todo i=1,...,k; j=1,...,p.

Esta fórmula nos permite cuantificar la influencia de cada variable en la definición de una CP. Al evaluar las contribuciones de las variables, podemos identificar cuáles son las más relevantes para cada CP y comprender mejor cómo se estructuran las CP en función de las variables originales.

Estas contribuciones estan en terminos relativos, por lo que se acostumbra a normalizar estas contribuciones entre $[0, 1]$ o entre $[0, 100]$ para que sea más fácilmente comparable la contribución de una variable con respecto al resto en la construcción de una CP.

En el ACP, a menudo se utiliza para resumir y comprender las relaciones entre variables en un conjunto de datos. Sin embargo, en situaciones del mundo real, es común que surjan nuevos elementos o variables que no estaban presentes en el conjunto de datos original. Estos nuevos elementos se conocen como índividuos suplementarios'cuando se refieren a observaciones y 'variables suplementarias'cuando se refieren a características adicionales o medidas.

En el ACP, a menudo se utiliza para resumir y comprender las relaciones entre variables en un conjunto de datos. Sin embargo, en situaciones del mundo real, es común que surjan nuevos elementos o variables que no estaban presentes en el conjunto de datos original. Estos nuevos elementos se conocen como índividuos suplementarios'cuando se refieren a observaciones y 'variables suplementarias'cuando se refieren a características adicionales o medidas.

La inclusión de individuos y variables suplementarias en el ACP nos permite evaluar cómo se sitúan en relación con la estructura existente de las componentes principales y cómo contribuyen a la variabilidad general. Esto es esencial para mantener nuestro análisis actualizado y relevante en entornos en constante cambio.

En el ACP, a menudo se utiliza para resumir y comprender las relaciones entre variables en un conjunto de datos. Sin embargo, en situaciones del mundo real, es común que surjan nuevos elementos o variables que no estaban presentes en el conjunto de datos original. Estos nuevos elementos se conocen como índividuos suplementarios'cuando se refieren a observaciones y 'variables suplementarias'cuando se refieren a características adicionales o medidas.

La inclusión de individuos y variables suplementarias en el ACP nos permite evaluar cómo se sitúan en relación con la estructura existente de las componentes principales y cómo contribuyen a la variabilidad general. Esto es esencial para mantener nuestro análisis actualizado y relevante en entornos en constante cambio.

Para tratar a estas nuevas observaciones, debemos estandarizar sus variables, utilizando la media y la desviación estándar del conjunto de datos original. Posteriormente, los añadimos a nuestra base de datos estandarizada para proceder al cálculo de sus coordenadas en las CP.

Cuando incorporamos una variable categórica a un ACP previamente realizado, estamos enriqueciendo nuestra comprensión de los datos al considerar la influencia de categorías adicionales en la estructura de los datos.

Cuando incorporamos una variable categórica a un ACP previamente realizado, estamos enriqueciendo nuestra comprensión de los datos al considerar la influencia de categorías adicionales en la estructura de los datos.

En este contexto, representar los centroides de estas categorías en los nuevos ejes de las componentes principales se convierte en una estrategia esencial. Los centroides actúan como puntos de referencia que resumen la tendencia central de cada categoría en el espacio de las CP. Este enfoque nos permite visualizar cómo las distintas categorías se relacionan entre sí y con las CP, lo que puede revelar patrones y relaciones más sutiles en los datos que antes no eran evidentes.

Cuando incorporamos una variable categórica a un ACP previamente realizado, estamos enriqueciendo nuestra comprensión de los datos al considerar la influencia de categorías adicionales en la estructura de los datos.

En este contexto, representar los centroides de estas categorías en los nuevos ejes de las componentes principales se convierte en una estrategia esencial. Los centroides actúan como puntos de referencia que resumen la tendencia central de cada categoría en el espacio de las CP. Este enfoque nos permite visualizar cómo las distintas categorías se relacionan entre sí y con las CP, lo que puede revelar patrones y relaciones más sutiles en los datos que antes no eran evidentes.

En esencia, al considerar variables categóricas y representar sus centroides en el ACP, estamos agregando una dimensión adicional a nuestra comprensión de la estructura subyacente de los datos.