# Fundamentos para el Análisis de Datos y la Investigación
# Tema 3: Fundamentos de matemáticas para la ciencia de datos
## *Análisis de Componentes Principales*


**José María Sarabia**
**Máster Universitario en Ciencia de Datos**
**CUNEF Universidad**
**Curso 2023-24**

# *Indice*

1. Introducción e hipótesis del PCA
2. Cálculo de las componentes principales
3. Selección del número de componentes e Interpretación
4. Variantes y rotaciones
5. Aplicaciones del PCA
6. Ejemplo
7. Ejercicios

# Análisis de Componentes Principales: Términos Clave

- Análisis de Componentes Principales (PCA, *Principal Component Analysis*)
- Reducción de la dimensión
- Componentes o factores
- Autovalores (varianza de los componentes) y autovectores de una matriz:
- Diagrama de codo o de sedimentación elegir el numero de factores
- Test de Bartlett e índice KMO
- Cargas factoriales las puntuaciones
- Puntuaciones factoriales
- Biplot
- Rotaciones

# Introducción

**Análisis de componentes principales (PCA), Análisis Factorial y Análisis de Correspondencias.**

autovalores varianza de las variables que cogemos, explican todas las variables, correlación lineal

- Son técnicas de análisis de datos que permiten la reducción de la dimensión, es decir, se pasa de un número de variables *p* a otro número de variables *m<p* que recojan la mayor parte de la información posible.
- Desde el punto de vista del machine learning se trata de una técnica de aprendizaje no supervisado, es decir, extraen información de los datos sin necesidad de entrenar un modelo con datos etiquetados.
- El aprendizaje no supervisado construye un modelo de los datos sin distinguir entre la variable respuesta y las variables predictoras.
- Se trata por tanto de resumir las variables observadas en un conjunto de variables más pequeño (denominadas factores y que son combinaciones lineales de las variables iniciales), con la menor pérdida de información posible.
- El análisis de componentes principales PCA no requiere la hipótesis de normalidad de los datos.

# Hipótesis del PCA

COMPONENTES PRINCIPALES

Analisis junta

- El PCA tiene sentido, cuando las variables de estudio presentan correlaciones entre sí. Las variables en PCA deben ser numéricas.
- Por tanto, si las variables originales $x_1,\ldots,x_p$ están **incorreladas**, el PCA no tiene sentido.

Si V son categoricas CORRESPONDENCIA
y el V mix FACTORIAL MIXTO

- Para conocer si las variables están correlacionadas entre si globalmente, se realiza el llamado **Test de esfericidad de Bartlett**.
- La hipótesis nula del contraste es que el determinante de la matriz de correlaciones es 1, frente a que es distinta de 1.
- Tener en cuenta que las componentes principales de $x_1,\ldots,x_p$ dependen de las unidades de medida. Por tanto, si se cambian las unidades cambian las variables.
- El **Test KMO** (Kaiser, Meyer y Olkin) es una medida de adecuación muestral. Su valor está entre 0 y 1, de modo que si su valor es próximo a 1, el PCA resulta más conveniente.
- **Determinante de la matriz de correlaciones**: el determinante de la matriz de correlaciones debe ser positivo para proceder con el PCA

primero hay que escalar

se hace el test de Bartlett

las variables deben de estar correlacionadas

# Cálculo de los Componentes Principales

- Disponemos de un conjunto *p* de variables y se busca un conjunto *m* de variables (llamadas componentes principales o factores), de modo que p>>m, que verifiquen tres características:
    1. Sean combinaciones lineales de las variables originales.
    2. Recojan la mayor parte de la información o variabilidad posible.
    3. Sean ortogonales entre sí, es decir, la información de una componente no está contenida en la otra.

mayor varianza

de mayor importancia a menor

M1, M2, M3 (Entre ellas con estan correlacionadas), mayor varianza y mayor autovalor

la varianza nos da dispersión

# Cálculo de los Componentes Principales

- La matriz de datos es *n*x*p*: *n* individuos u objetos medidos en *p* variables. Los datos son $x_{ij}$, i=1,2,…,*n*; j=1,2,…,*p*, con matriz de covarianzas $\Sigma$.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

- Las variables son

$$(x_1, \ldots, x_p)$$

medidas sobre un conjunto de objetos o individuos y tenemos que obtener un nuevo conjunto de variables

$$(y_1, \ldots, y_p)$$

que sean combinaciones lineales de las anteriores, y que además estén incorreladas y con la mayor varianza posible. La varianza de cada nueva variable irá decreciendo.

# Cálculo de los Componentes Principales

- Las nuevas variable $y_j$ cumplirán,

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p = \mathbf{a_j'x}$$

- Trabajamos con transformaciones ortogonales, de modo que el módulo del vector $\mathbf{a_j'}$ sea la unidad,

$$\mathbf{a_j'a_j} = a_{1j}^2 + \cdots + a_{pj}^2 = 1$$

- La primera componente es una combinación lineal de todas las variables, y maximiza la varianza, con la condición de que su módulo se 1.
- La varianza de la primera componente se corresponde con el mayor autovalor y sus coordenadas con el correspondiente autovector
- La segunda componente maximiza la varianza, tiene módulo 1 y es ortogonal a la primera componente (no tienen información en común), y su varianza se corresponde con el segundo autovector. Y así sucesivamente

# Cálculo de componentes principales desde la matriz de correlaciones

- Normalmente, los componentes se calculan sobre las variables originales estandarizadas.
- Esto equivale a calcular los componentes principales **desde la matriz de correlaciones**.
- De este modo, los componentes son los **autovectores** de la matriz de correlaciones (son diferentes de los de la matriz de covarianzas).
- Así, se da la misma importancia a todas las variables.
- Si las variables originales están tipificadas, las matrices de covarianzas y de correlaciones coinciden y la variabilidad total (la traza de la matriz) es igual al número de variables de la muestra.
- La suma de los autovalores es p y la proporción de varianza recogida por el autovector i-ésimo es $\lambda_i/p$, según lo visto antes.

# Selección del número de componentes e Interpretación

- Un aspecto importante del PCA es decidir cuántos factores se deben elegir.
- Se eligen los componentes que recogen **la mayor varianza posible**.
- Uno de los principales criterios es elegir aquellos componentes cuyos autovalores son mayores que 1.
- Para ello resulta de ayuda el **gráfico de sedimentación o de codo**.
- Una vez elegido el número de componentes, un aspecto importante es identificar los componentes y conseguir una **interpretación** de cada uno de los seleccionados.
- Para ello, hay que identificar las variables que aparecen en cada factor (*que se saturan en el factor*).
- Un gráfico **Biplot** combina simultáneamente los datos y las variables.
- El aspecto más importante en la interpretación del Biplot son las direcciones. Lo más sencillo es considerar los 4 cuadrantes, y las observaciones que se se sitúan en cada uno de ellos.

# Variantes: PCA Robusto y Rotaciones

- Variante del PCA basado en estimaciones robustas que en ocasiones ayuda a una mejor interpretación de los componentes.
- Las rotaciones de los factores permiten una mejor interpretación del análisis de componentes principales. De este modo, se consiguen Identificar con claridad las variables que definen cada factor.
- El Objetivo de las rotaciones es conseguir que la correlación de cada una de las variables sea cercana a 1 en un solo factor, y próxima a 0 en el resto.

# Aplicaciones del ACP

1) Encontrar un conjunto de **Factores Subyacentes** (no directamente observables) y que sean combinación lineal de algunas variables.

2) Construir **Indicadores Sintéticos**, que tengan en cuenta todas las variables. Por ejemplo, construir un índice de desarrollo humano, a partir de diversos indicadores relacionados con la salud, la educación y la renta.

3) En regresión aparece el problema de la **Multicolinealidad**. Por estas técnicas se pueden sustituir las variables de la regresión por otras nuevas que sean combinaciones lineales de algunas y estén incorreladas.

# Aplicaciones del ACP

***Aplicación 1.*** Encuesta sobre situación laboral en España. Se realiza un total de 9 preguntas. Tras realizar el correspondiente análisis se obtienen los siguientes resultados.

**Matriz de componentes rotados[a]**

| | Componente | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Ganas de trabajar | ,815 | -,118 | ,076 |
| Comodidad | ,762 | -,065 | ,083 |
| Preparación | ,715 | ,080 | -,074 |
| Búsqueda | ,682 | -,064 | ,224 |
| Crisis | -,081 | ,796 | -,053 |
| Política de empleo | -,061 | ,785 | ,017 |
| Empresarios | ,024 | ,509 | ,263 |
| Reparto | -,008 | ,120 | ,823 |
| Pluriempleo | ,205 | ,023 | ,777 |

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

***Aplicación 2.*** Disponemos de las notas de un conjunto de 200 estudiantes en 8 asignaturas: Lengua; Matemáticas; Física; Inglés; Filosofía; Historia; Química y E. Física

Se trata de resumir las 8 variables en 2 “Factores”, que contengas toda la información posible, y que sean independientes entre si:

**F1=a11*LENGUA+a12*INGLÉS+a13*HISTORIA +a14*FILOSOFÍA, (46,4%)**

**F2=a21*FÍSICA+a22*QUÍMICA+a23*MATEMÁTICAS, (35,8%)**

# Ejemplo de ACP: *Valoración de 8 marcas de coches*

- Un grupo de conductores valoran 8 marcas de coches (A, B,…,H) según 3 características: Precio, prestaciones, y diseño (escala de 1 a 10).

| Marca | Precio | Prestaciones | Diseño |
|---|---|---|---|
| A | 3 | 5 | 5 |
| B | 2 | 3 | 6 |
| C | 3 | 2 | 4 |
| D | 4 | 5 | 4 |
| E | 5 | 5 | 4 |
| F | 8 | 9 | 6 |
| G | 9 | 7 | 7 |
| H | 9 | 9 | 8 |

- Nos preguntamos sobre los **factores que determinan la elección de un vehículo**.
- Se busca una fórmula para la **valoración de la Marca** en función de las variables precio, prestaciones y diseño.
- **Matriz de correlaciones:**

| | Precio | Prestaciones | Diseño |
|---|---|---|---|
| Precio | 1.0000 | 0.8952 | 0.7065 |
| Prestaciones | 0.8952 | 1.0000 | 0.6828 |
| Diseño | 0.7065 | 0.6828 | 1.0000 |

- **Test de esfericidad de Bartlett**

```
Bartlett's sphericity test

chi.square =  12.056 , df =  3 , p-value =  0.007193606
```

la hipotesis nula se rechaza cuando el pvalor es menor a 0,05

- **Estadístico KMO**

```
$overall
[1] 0.7024613

$report
[1] "The KMO test yields a degree of common variance middling."

$individual
      Precio Prestaciones       Diseño
   0.6427508    0.6566676    0.8967921
```

- **Determinante matriz de correlaciones**

```
> round(cor(pm), 3)
             Precio Prestaciones Diseño
Precio        1.000        0.895  0.707
Prestaciones  0.895        1.000  0.683
Diseño        0.707        0.683  1.000
> det(cor(pm))
[1] 0.09696271
```

- **Resumen información componentes**

```
Importance of components:
                          Comp.1    Comp.2    Comp.3
Standard deviation     1.5897197 0.6072408 0.3225677
Proportion of Variance 0.8424029 0.1229138 0.0346833
Cumulative Proportion  0.8424029 0.9653167 1.0000000
```

- **Varianzas de las componentes**

```
Varianzas:
> val.propios
   Comp.1    Comp.2    Comp.3
2.5272087 0.3687414 0.1040499
```

- **Vectores propios (cargas):**

```
Loadings:
           Comp.1 Comp.2 Comp.3
Precio      0.597  0.344  0.724
Prestaciones 0.592  0.420 -0.688
Diseño      0.541 -0.840

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
```

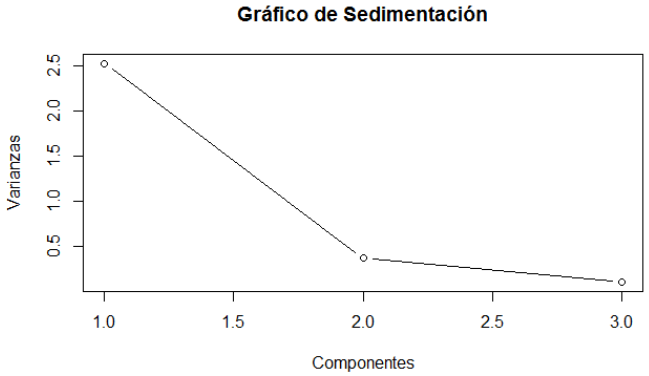
- **Primera y segunda componente** coger los valores absolutos

F1=0.597*__Precio__+0.592*__Prestaciones__+0.541*Diseño

F2=0.344*Precio+0.420*Prestaciones-0.840*__Diseño__

- **Interpretación:** Primera componente: <u>Precio y Prestaciones</u>; segunda: <u>Diseño</u>.

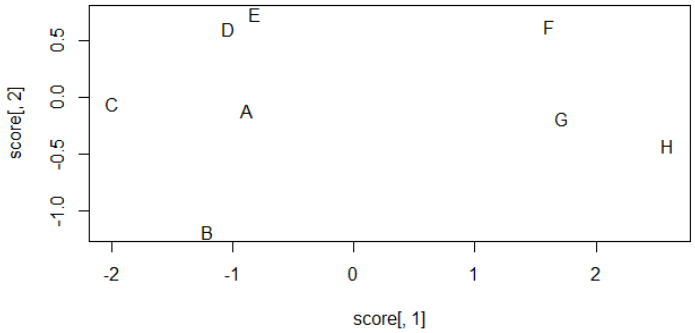
- **Diagrama de codo**




- **Puntuaciones**

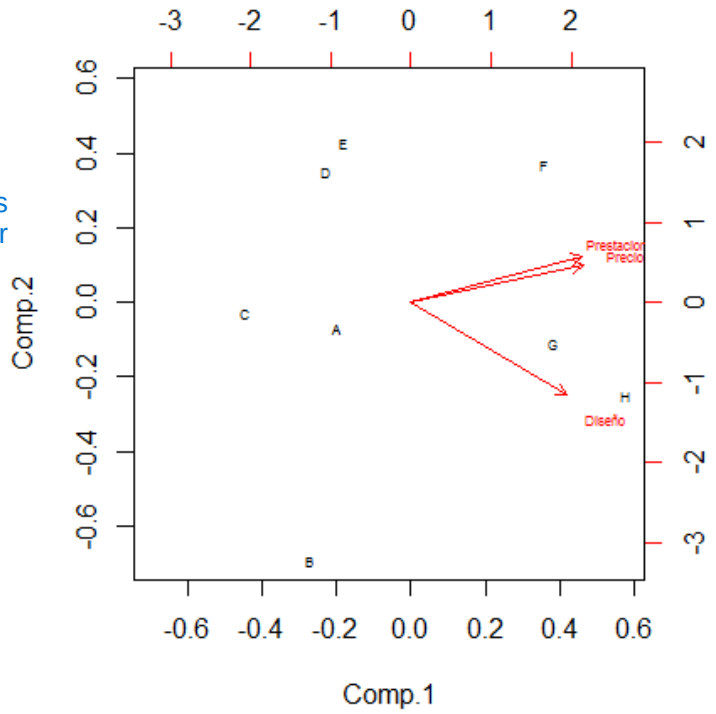
```
> score
      Comp.1      Comp.2      Comp.3
A -0.8732342 -0.11686843 -0.44341994
B -1.2071896 -1.18933268 -0.17126024
C -1.9974885 -0.04908146  0.45153233
D -1.0337324  0.60495321 -0.14102482
E -0.8116564  0.73302397  0.12825027
F  1.6086267  0.63101971 -0.27927417
G  1.7188234 -0.18530302  0.53143572
H  2.5958509 -0.42841130 -0.07623914
```

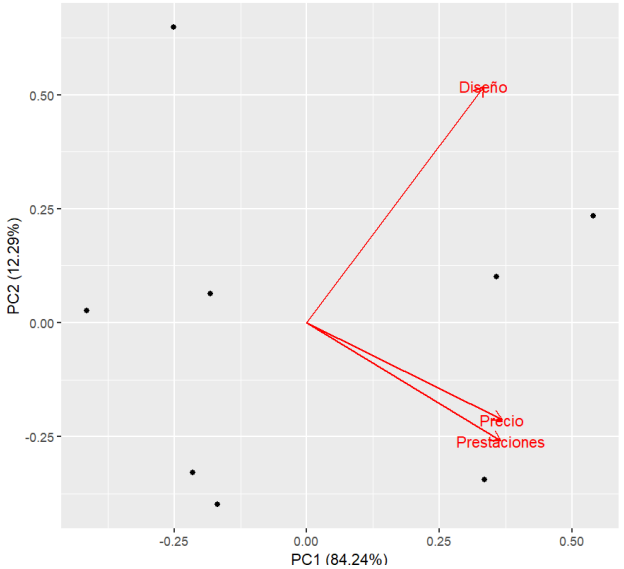
- **Gráfico de puntuaciones**



- **Biplot**

las flechas juntas son las que representan al factor



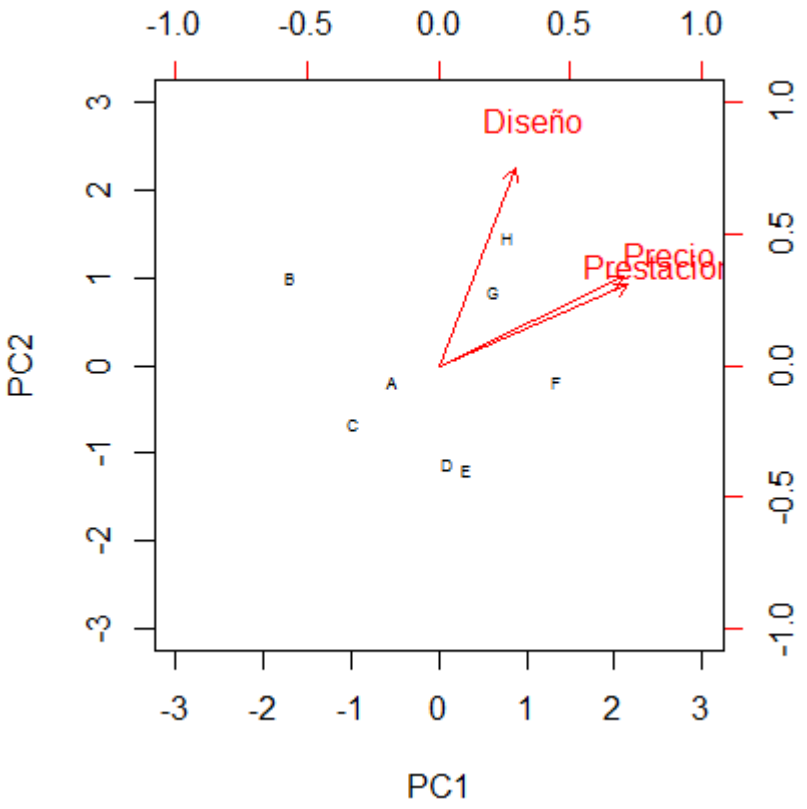

- **Rotación Varimax (2 componentes)**

```
Loadings:
             PC1   PC2
Precio       0.876 0.422
Prestaciones 0.897 0.381
Diseño       0.361 0.932

                PC1   PC2
SS loadings   1.703 1.193
Proportion Var 0.568 0.398
Cumulative Var 0.568 0.965
```

- **Interpretación:** Primera componente: Precio y Prestaciones; segunda: Diseño (las variables se *saturan* en cada factor).

- **Biplot basado en rotación Varimax**



# Ejercicios

***Ejercicio 1***. Con los datos de Happiness de 2017:

a) Seleccionar variables y realizar un PCA completo
b) Objetivo: Construir un índice de felicidad usando las componentes

***Ejercicio 2***. Realizar un análisis de componentes principales sobre los datos de R: USArrests.

***Ejercicio 3***. Con datos de los municipios de la Comunidad de Madrid, construir un índice turístico municipal a partir de un conjunto de variables relacionadas con la actividad turística.