

# PCA

Minería de Datos (CC3074) - 2026

---

## Principal Component Analysis (PCA)

---

Semestre 01, 2026

### El problema de la dimensionalidad

---

- Más variables → más dimensiones
- Más dimensiones → más complejidad
- Más complejidad → más difícil interpretar

### Dimensionalidad

---

- Cada variable añade una dimensión
- 2 variables → plano
- 3 variables → espacio
- $p$  variables → espacio  $p$ -dimensional
- Alta dimensionalidad complica el análisis

- Visualización imposible en >3 dimensiones
- Ruido y redundancia entre variables

## Repaso

---

take this

- **Varianza:** mide qué tan dispersos están los datos respecto a la media.
- **Covarianza:** mide cómo dos variables varían juntas.
- Positiva → se mueven en la misma dirección.
- Negativa → se mueven en direcciones opuestas.

## PCA

---

- Principal Component Analysis
- Técnica para reducir dimensionalidad
- Conserva la mayor cantidad de información posible

### Componente Principal (PC)

- Nueva variable creada a partir de las originales
- Combinación lineal de todas las variables

- Nueva dimensión del espacio
- Capturar la mayor varianza posible
- Varianza = Información
- Mayor varianza → mayor dispersión
- Mayor dispersión → mayor contenido informativo
- PCA busca la dirección de proyección que maximice la varianza
- Conserve más información

## Proyección

---

- Imaginemos una linterna iluminando los datos
- La sombra proyectada representa una nueva dimensión
- Buscamos la dirección donde la sombra tenga mayor dispersión
- Esa dirección es el primer componente principal

proy

proy

## Álgebra Lineal

---

- Combinación lineal de variables
- Eigenvectores y Eigenvalores

## **Combinación lineal**

- Cada componente principal es una combinación lineal
- $PC_1 = a_1X_1 + a_2X_2 + \dots + a_pX_p$
- Los coeficientes determinan la dirección óptima

## **Eigenvectores**

- Representan direcciones en el espacio
- Solo tienen dirección, no escala ( $\|v\|=1$ )
- Definen los nuevos ejes del sistema
- vamos a rotar

## **Eigenvalores**

- Miden cuánta varianza explica cada componente
- Mayor eigenvalor → mayor información
- Permiten ordenar los componentes por importancia
- La suma de estos valores = varianza total
- Entre más alto sea el eigenvalor, mayor es la varianza que explica ese componente.

# Pasos

---

## Paso 0: Análisis de viabilidad

---

- Determinar si vale la pena
- Coeficiente KMO
- Prueba de esfericidad de Bartlett

### Coeficiente KMO

- Kaiser-Meyer-Olkin
- Mide la adecuación de la muestra
- Valores entre 0 y 1
- Correlaciones parciales pequeñas → buen PCA

### Escala KMO (1974)

- 0.9 – 1.0 → Excelente
- 0.8 – 0.9 → Buena
- 0.7 – 0.8 → Aceptable
- 0.6 – 0.7 → Regular
- 0.5 – 0.6 → Mala
- < 0.5 → Inaceptable

### Prueba de Bartlett

Hipótesis nula ( $H_0$ ):

- Matriz de correlaciones = identidad

- No hay correlación

Hipótesis alternativa ( $H_1$ ):

- Existen correlaciones

## Interpretación

☐☐☐ QUEREMOS RECHAZAR!!!

- $p < 0.05 \rightarrow$  Rechazar  $H_0$
- Sí hay correlación
- PCA es adecuado
- $p > 0.05 \rightarrow$  No adecuado

## Paso 1: Estandarización

---

- Variables en escalas distintas
- Mayor escala  $\rightarrow$  dominio sobre varianza
- PCA sensible a la magnitud de variables
- Estandarizar evita sesgo por escala

## Proceso

$$Z = \frac{X - \bar{X}}{\sigma}$$

## Resultado

- Media = 0
- Varianza = 1
- Contribución equitativa
- Matriz de correlación

## Paso 2: Matriz de covarianza

---

- Resume cómo se relacionan las variables entre sí
- Cada elemento mide cómo varían dos variables juntas

$$\text{Cov}(X_i, X_j)$$

- Positiva → relación directa
- Negativa → relación inversa
- Cero → independencia lineal aproximada

## Paso 3: Eigenvalores y Eigenvectores

---

- Con la matriz de covarianza
- Se busca la dirección donde la varianza sea máxima

$$\Sigma v = \lambda v$$

$$v \rightarrow \text{dirección} \quad \lambda \rightarrow \text{varianza}$$

Si proyectamos los datos sobre  $v$ , la varianza obtenida es  $\lambda$ .

- Eigenvectores → nuevas direcciones del espacio
- Eigenvalores → cuánta varianza hay en cada dirección

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

El eigenvector asociado a  $\lambda_1$  es la dirección de máxima varianza.

## Paso 4: Proyección de datos

---

- Rotar el sistema de coordenadas
- Proyectar los datos sobre los nuevos ejes

proy

proy

- Cada componente es una nueva variable
- Se reducen dimensiones
- Se conserva la mayor varianza posible

El resultado: nuevas variables no correlacionadas y ordenadas por varianza explicada (importancia).