



**FACULTAD de  
CIENCIAS ECONÓMICAS**

# **ANÁLISIS DE COMPONENTES PRINCIPALES II**

ALAN REYES-FIGUEROA

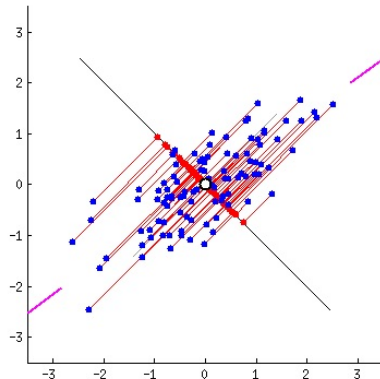
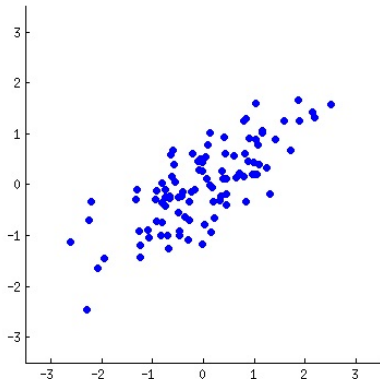
ELEMENTS OF MACHINE LEARNING

(AULA 07) 07.FEBRERO.2023

# Componentes principales

Objetivo: encontrar una estructura subyacente en los datos.

- Proyectar a un subespacio adecuado.



# Componentes principales

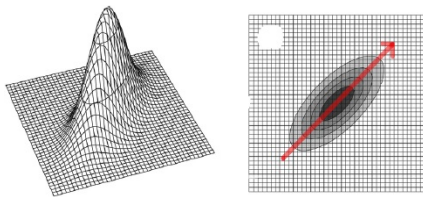
Caso particular 1D: (proyectamos a un subespacio 1-dimensional).

Suponga que proyectamos a un subespacio  $\langle \ell \rangle \Rightarrow \langle \ell, X \rangle = \ell^T X$ .

Buscamos maximizar

$$\max_{\|\ell\|=1} \text{Var}(\ell^T X) = \max_{\ell \neq 0} \frac{\text{Var}(\ell^T X)}{\ell^T \ell} = \max_{\ell \neq 0} \frac{\ell^T \text{Var}(X) \ell}{\ell^T \ell} = \max_{\ell \neq 0} \frac{\ell^T (\mathbb{X}^T \mathbb{X}) \ell}{\ell^T \ell}.$$

(cociente de Rayleigh).



## Teorema (Descomposición en valores singulares (SVD))

Sea  $A \in \mathbb{R}^{n \times d}$  una matriz de rango  $k$ . Para todo  $1 \leq r \leq k$ , existen matrices  $U \in \mathbb{R}^{n \times r}$ ,  $S \in \mathbb{R}^{r \times r}$ ,  $V \in \mathbb{R}^{d \times r}$ , tales que

$$A = USV^T = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

con

- las columnas  $\mathbf{u}_1, \dots, \mathbf{u}_r \in \mathbb{R}^n$  de  $U$  son los autovectores de  $AA^T$ ,
- las columnas  $\mathbf{v}_1, \dots, \mathbf{v}_r \in \mathbb{R}^d$  de  $V$  son los autovectores de  $A^T A$ ,  $S = \text{diag}(\sigma_1, \dots, \sigma_r)$ ,  $\sigma_i^2 = \lambda_i$ , con  $\lambda_i$  los autovalores de  $\mathbf{u}_i$  y de  $\mathbf{v}_i$ ,
- Además,  $\sigma_i \mathbf{u}_i = A \mathbf{v}_i$  y  $\sigma_i \mathbf{v}_i = A^T \mathbf{u}_i$ , para  $i = 1, 2, \dots, r$ .

# Descomposición SVD

El teorema de descomposición espectral ocurre como un caso particular de la descomposición SVD:

Caso especial:  $A$  simétrica

$$A = USU^T = U\Lambda U^T = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{u}_i^T.$$

En este caso los autovectores de  $A$  y  $A^T A = A^2 = A A^T$  coinciden, y los autovalores de  $A$  al cuadrado son los autovalores de  $A^T A$ .

# Direcciones Principales

The diagram illustrates the SVD decomposition of a matrix  $M$  into three components:  $U$ ,  $\Sigma$ , and  $V^*$ .

- Matrix  $M$ :** A gray 4x4 grid representing a matrix of size  $m \times n$ .
- Matrix  $U$ :** A 4x4 grid with columns colored teal, green, blue, and green, representing a matrix of size  $m \times m$ .
- Matrix  $\Sigma$ :** A 4x4 grid with a diagonal of colored squares (orange, yellow, yellow, light blue) and zeros elsewhere, representing a matrix of size  $m \times n$ .
- Matrix  $V^*$ :** A 4x4 grid with rows colored light blue, purple, purple, and pink, representing a matrix of size  $n \times n$ .

The equation  $M = U \Sigma V^*$  is shown below the grids, with the dimensions  $m \times n$ ,  $m \times m$ ,  $m \times n$ , and  $n \times n$  indicated under each matrix respectively.

Diagram illustrating the decomposition of a matrix  $U$  into  $U^*$  and  $I_m$ .

Matrix  $U$  is a 4x4 grid with columns of colors: light blue, green, light blue, green.

Matrix  $U^*$  is a 4x4 grid with rows of colors: light blue, green, light blue, green.

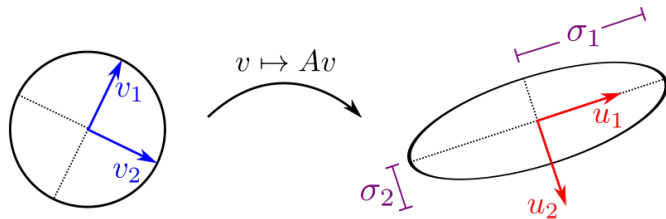
Matrix  $I_m$  is a 4x4 identity matrix with 1s on the diagonal and 0s elsewhere.

The equation shown is  $U = U^* I_m$ .

Diagram illustrating the relationship between a matrix  $V$ , its adjoint  $V^*$ , and the identity matrix  $I_n$ .

Matrix  $V$  is a 3x3 grid with columns of blue, green, and red. Matrix  $V^*$  is a 3x3 grid with rows of blue, green, and red. The identity matrix  $I_n$  is a 3x3 grid with 1s on the diagonal and 0s elsewhere.

# Direcciones Principales



$$A = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \begin{bmatrix} - & v_1^T & - \\ - & v_2^T & - \\ & \vdots & \\ - & v_n^T & - \end{bmatrix}$$

# Direcciones Principales

En general, si queremos proyectar los datos  $\mathbb{X}$  en  $\mathbb{R}^d$  a un subespacio de dimensión  $r$ ,  $0 < r < d$ , debemos encontrar las primeras  $r$  direcciones principales.

Estas direcciones pueden encontrarse de dos formas:

- las direcciones principales  $\mathbf{v}_i$  son los autovectores asociados a los primeros  $r$  autovectores de  $\text{Cov}(\mathbb{X})$ .
- también pueden calcularse como los primeros  $r$  vectores, asociados a los  $r$  mayores valores singulares de  $\mathbb{X}$ .



## Teorema (Eckart-Young)

Sea  $A \in \mathbb{R}^{n \times d}$ ,  $n \geq d$ , una matriz cuya descomposición SVD está dada por

$$A = USV^T = \sum_{i=1}^d \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

Entonces, la matriz  $\hat{A}_r$  de rango  $r$ ,  $1 \leq r \leq d$ , que mejor aproxima  $A$  en el sentido de minimizar

$$\min_{\text{rank } \hat{A}_r \leq r} \|A - \hat{A}_r\|_F^2$$

se obtiene de truncar la descomposición en valores singulares de  $A$ :

$$\hat{A}_r = U_r S_r V_r^T = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

## Teorema (Eckart-Young)

donde

$$U_r = [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_r], \quad S_r = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r), \quad V_r = [\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_r].$$

En ese caso, el error de aproximación está dado por

$$\|A - \hat{A}_r\|_F^2 = \sum_{i=r+1}^d \lambda_i,$$

o

$$\|A - \hat{A}_r\|_2^2 = \lambda_{r+1}.$$

# Aproximaciones de bajo rango

## Obs!

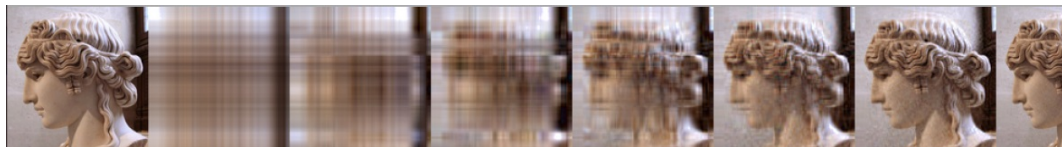
- Las direcciones  $\mathbf{u}_i$  se llaman las **componentes principales** de  $\mathbb{X}$ .
- La descomposición SVD proporciona un mecanismo para proyectar los datos al “mejor” subespacio de dimensión  $r \leq d$ . Dicha proyección se obtiene haciendo

$$\mathbb{X}_{proj} = \mathbb{X} V_r^T.$$

- Los autovalores  $\lambda_i$  de  $\mathbb{X}^T \mathbb{X}$  nos proporcionan un mecanismo para medir el error, vía  $\|A - \hat{A}_r\|_F^2 = \sum_{i=r+1}^d \lambda_i$ .
- El cociente  $\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^d \lambda_i}$ ,  $r = 1, 2, \dots, d$ , se interpreta como el porcentaje de variabilidad de los datos  $\mathbb{X}$  que es explicada por las primeras  $r$  componentes principales.

# Ejemplos

Compresión de imágenes usando PCA.

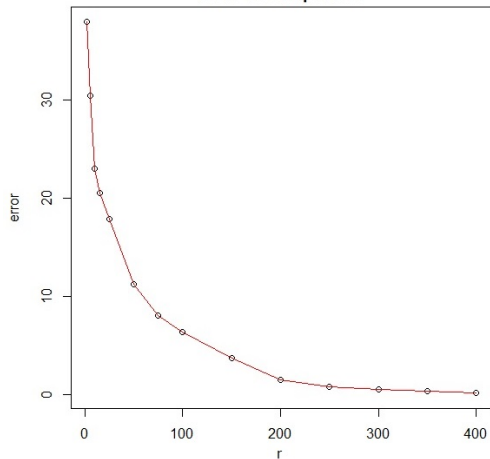


Original     $r = 1$      $r = 2$      $r = 4$      $r = 8$      $r = 16$      $r = 32$      $r = 64$

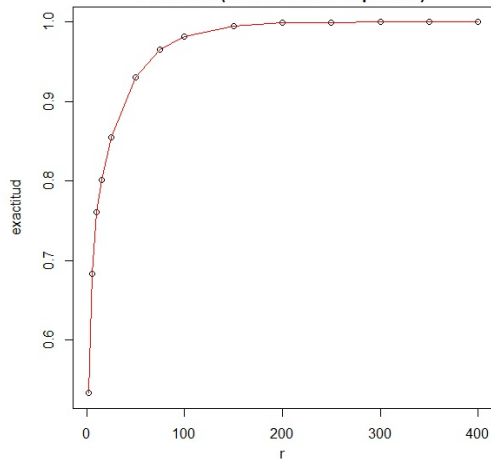
Imagen Original ( $256 \times 256$ ), aproximaciones con rango = 1, 2, 4, 8, 16, 32, 64.

# Ejemplos

Error de compresión

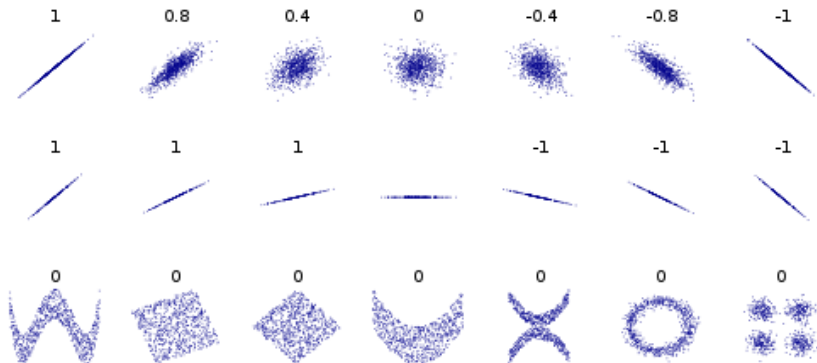


Exactitud (% variabilidad explicada)

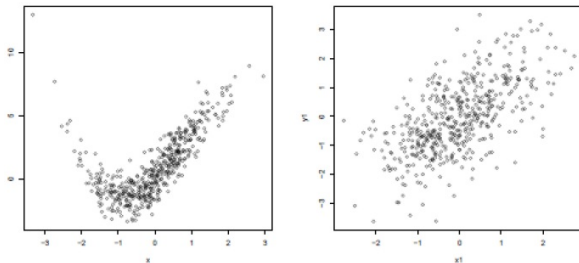


# Ejemplos

En PCA la estructura de los datos se capta solamente a través de las matrices  $\text{Cov}(X)$  o  $\text{Corr}(X)$ .



# Ejemplos



Dos veces misma correlación. =(

## Obs.

- Cuidado con desviaciones fuertes de normalidad.
- Lo ideal es investigar la normalidad de los datos en la práctica, al menos ver si escala es continua, distribución unimodal, simétrica, nubes parecidas a elipses, ...

# Contraejemplos

