

PCA


---

---

---

---

---



# PCA $\rightarrow$ Combinación lineal de las variables

$$\begin{bmatrix} z_1 \\ \vdots \\ z_p \end{bmatrix} \Rightarrow \begin{bmatrix} \Phi_{11} & \Phi_{12} & \dots & \Phi_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{p1} & \Phi_{p2} & \dots & \Phi_{pp} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

$$z_p \Rightarrow \Phi_{1p} x_1 + \dots + \Phi_{pp} x_p$$

$$\sum_{j=1}^p \Phi_{ji}^2 = 1$$

¿Cómo calculamos esos loadings?

Tenemos que encontrar loadings que maximicen la varianza, una forma de calcularlos es con eigenvector y eigenvalue

Imaginemos que tenemos  $x_1$  y  $x_2$  dos variables

1) Calculamos eigenvectores

$$\begin{bmatrix} 1,1 \\ 2,1 \end{bmatrix} \begin{bmatrix} 1,1 \\ 2,1 \end{bmatrix} \rightarrow \text{la transpuesta}$$
$$\begin{bmatrix} \Phi_{11} \\ \Phi_{21} \end{bmatrix} \quad \begin{bmatrix} \Phi_{12} \\ \Phi_{22} \end{bmatrix}$$

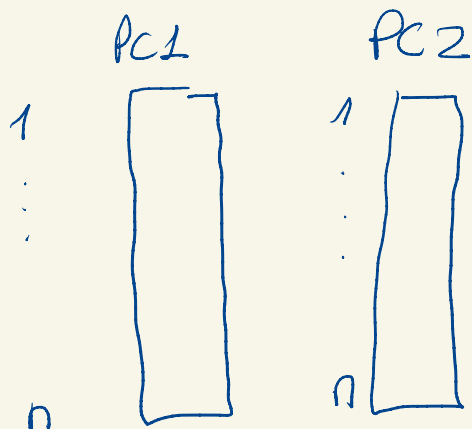
$\rightarrow$  En el caso de que los vectores sean columna

2) Cálculo de la proyección multiplicando por  $X^T$

$$z_1 = \Phi_{11} \cdot x_1 + \Phi_{21} \cdot x_2$$

$$z_2 = \Phi_{12} \cdot x_1 + \Phi_{22} \cdot x_2$$

Entonces nos queda lo siguiente



son los dos  
componentes  
principales  
de los datos

¿Cuánta información es capaz de capturar cada componente?

$$\sum_{j=1}^p \text{Var}(x_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

Varianza total en  
el set de datos  
(desviación media  
0)

y la varianza de la componente  $m$  es:

$$\frac{1}{n} \sum_{i=1}^n z_{im}^2 = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \Phi_{jm} x_{ij} \right)^2$$

la proporción de varianza explicada es:

$$\frac{\sum_{i=1}^n \left( \sum_{j=1}^p \Phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2}$$

Número ótimo de componentes

$N_{\text{comp}} > 95\%$  variância.