# Análisis de Componentes Principales (ACP)

Teoría y Práctica

Dr. Bartolo Villar

November 28, 2021

TICs, ENES-UNAM

# Introducción

- El **Análisis de Componentes Principales (ACP)** es una técnica matemática de **aprendizaje no supervisado**.
- Consiste en explicar la estructura de **varianza-covarianza** de un conjunto de variables a través de **pocas combinaciones lineales** de las variables originales.
- Generalmente se utiliza para
    - reducción de la dimensionalidad de los datos,
    - interpretación,
    - sirve como un paso intermedio en investigación.
- Los $k$ componentes principales reemplazan a las $p$ variables originales.
- El ACP únicamente depende de la matriz de varianzas-covarianza, $\boldsymbol{\Sigma}$, o en su defecto de la matriz de correlaciones, $\boldsymbol{\rho}$.

# Formulación matemática

Sea $\boldsymbol{X}' = (X_1, X_2, \ldots, X_p)$ un vector aleatorio con matriz de covarianzas $\boldsymbol{\Sigma}$ a su vez con eigenvalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

Considere las siguientes combinaciones lineales

$$
\begin{aligned}
Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

$$var(Y_i) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_i \qquad\qquad i = 1, 2, ..., p$$

$$covar(Y_i, Y_k) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_k \qquad\qquad i, k = 1, 2, ..., p$$

# ACP

- Los componentes principales son aquellas combinaciones lineales no correlacionadas (ortogonales), $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_p$ cuyas varianzas son las más grandes.
- El primer componente principal (CP1) es aquellas combinación lineal con máxima varianza

$$var(\boldsymbol{Y}_1) = \boldsymbol{a_1'} \Sigma \boldsymbol{a_1}$$

- Note que la varianza puede incrementarse si se multiplica a $\boldsymbol{a}_1$ por una constante, para evitar esto se impone la siguiente reestricción

$$\boldsymbol{a_1'} \boldsymbol{a_1} = 1$$

# ACP...

$CP_1$ $\boldsymbol{Y}_1$
Combinación lineal $\boldsymbol{a}_1'\boldsymbol{X}$ que maximiza la $var(\boldsymbol{a}_1'\boldsymbol{X})$ sujeto a que $\boldsymbol{a_1'a_1} = 1$.

$CP_2$ $\boldsymbol{Y}_2$
Combinación lineal $\boldsymbol{a}_2'\boldsymbol{X}$ que maximiza la $var(\boldsymbol{a}_2'\boldsymbol{X})$ sujeto a que $\boldsymbol{a_2'a_2} = 1$, y la $cov(\boldsymbol{a_1X}, \boldsymbol{a_2X}) = 0$.

$\vdots$

$CP_i$ $\boldsymbol{Y}_i$
Combinación lineal $\boldsymbol{a}_i'\boldsymbol{X}$ que maximiza la $var(\boldsymbol{a}_i'\boldsymbol{X})$ sujeto a que $\boldsymbol{a_i'a_i} = 1$, y la $cov(\boldsymbol{a_iX}, \boldsymbol{a_kX}) = 0$ para $i < k$.

# ACP

**Resultado**
Sea $\mathbf{\Sigma}$ la matriz de covarianzas asociada con el vector aleatorio $\boldsymbol{X}' = (X_1, X_2, \ldots, X_p)$. Sea $\mathbf{\Sigma}$ tal que tenga pares de eigenvalores-eigenvectores $(\lambda_1, \boldsymbol{e}_1) \ldots (\lambda_p, \boldsymbol{e}_p)$ con $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Entonces el $i-$ésimo componente principal está dado por

$$\begin{aligned} \boldsymbol{Y}_i &= \boldsymbol{e}_i \boldsymbol{X} \\ &= e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p \end{aligned}$$

$$\begin{aligned} var(\boldsymbol{Y}_i) &= \boldsymbol{e}_i' \mathbf{\Sigma} \boldsymbol{e}_i = \lambda_i \qquad & i = 1, 2, \ldots, p \\ cov(\boldsymbol{Y}_i, \boldsymbol{Y}_k) &= \boldsymbol{e}_i' \mathbf{\Sigma} \boldsymbol{e}_k = \lambda_i \qquad & i \neq k. \end{aligned}$$

# Varianzas e eigenvalores

**Resultado**
Sea $\mathbf{\Sigma}$ la matriz de covarianzas asociada con el vector aleatorio $\boldsymbol{X}' = (X_1, X_2, \ldots, X_p)$. Sea $\mathbf{\Sigma}$ tal que tenga pares de eigenvalores-eigenvectores $(\lambda_1, \boldsymbol{e}_1) \ldots (\lambda_p, \boldsymbol{e}_p)$ con $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Sean $\boldsymbol{Y}_1 = \boldsymbol{e}_1\boldsymbol{X} \ldots \boldsymbol{Y}_p = \boldsymbol{e}_p\boldsymbol{X}$ sus componentes principales. Entonces

$$\begin{aligned}
\sigma_{11} + \ldots + \sigma_{pp} &= \sum_{i=1}^{p} var(X_i) \\
&= \lambda_1 + \ldots + \lambda_p \\
&= \sum_{i=1}^{p} var(Y_i)
\end{aligned}$$

# Varianza del k-ésimo componente

- Del resultado anterior podemos obtener la proporción de varianza explicada por el k-ésimo componente principal:

$$\left( \begin{array}{c} \text{proporción} \\ \text{de varianza del k-ésimo} \\ \text{componente} \end{array} \right) = \frac{\lambda_k}{\lambda_1 + \ldots + \lambda_p} \quad \forall k = 1, \ldots, p$$

- En la práctica, generalmente con los primeros 3 o 4 componentes principales se explica al menos el 80% de la varianza total de $X$.
- Por otra parte, cada componente de $\boldsymbol{e}$ merece inspeccionarlos. La magnitud de $e_{ik}$ mide la importancia de la k-ésima variable al i-ésimo componente

$$e_{ik} \propto \rho(Y_i, X_k)$$

## Correlación entre $Y_i$ y $X_k$

**Resultado**
Si $\boldsymbol{Y}_1 = \boldsymbol{e}_1\boldsymbol{X} \dots \boldsymbol{Y}_p = \boldsymbol{e}_p\boldsymbol{X}$ son los componentes principales, con $\boldsymbol{\Sigma}$ siendo la matriz de covarianzas, entonces

$$\rho(Y_i, X_k) = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \qquad i, k = 1, 2, \dots, p,$$

son los coeficientes de correlación entre $Y_i$ y $X_k$.

# Ejemplo

La matriz de covarianzas para $\boldsymbol{X} = (X_1, X_2, X3)$ es

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Se puede verificar que

$$\begin{aligned} \lambda_1 &= 5.828 \\ \lambda_2 &= 2 \\ \lambda_3 &= 0.172 \end{aligned}$$

$$\begin{aligned} \boldsymbol{e}_1' &= \begin{bmatrix} -0.383 & 0.924 & 0 \end{bmatrix} \\ \boldsymbol{e}_2' &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \\ \boldsymbol{e}_3' &= \begin{bmatrix} 0.924 & 0.383 & 0 \end{bmatrix} \end{aligned}$$

## Ejemplo. . .

Los componentes principales son:

$$\begin{aligned} Y_1 &= \mathbf{e}_1' X = \\ Y_2 &= \mathbf{e}_2' X = \\ Y_3 &= \mathbf{e}_3' X = \end{aligned}$$

$$var(Y_1) = var(-0.383X_1 - 0.924X_2)$$

$$cov(Y_1, Y_2) = cov(-0.383X_1 - 0.924X_2, X_3)$$

# Ejemplo. . .

$$\begin{aligned}\sigma_{1,1}+\sigma_{2,2}+\sigma_{3,3} &= 1+5+2\\ &= \lambda_1+\lambda_2+\lambda_3\\ &= 5.828+2+0.172\end{aligned}$$

Note que los dos primeros componentes principales explican el 98% de la varianza contenida en $\boldsymbol{X}$:

$\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3} = 0.98$

$\rho(Y_1, X_1) = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} =$

$\rho(Y_1, X_2) = \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{12}}} =$

# Representación gráfica del ACP

# ACP con variables estandarizadas

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}$$
$$\vdots$$
$$Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$$

En notación matricial

$$\boldsymbol{Z} = (\boldsymbol{V}^{1/2})^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \text{ donde } \boldsymbol{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{\sigma_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Con $E(\boldsymbol{Z}) = \boldsymbol{0}$

$$cov(\boldsymbol{Z}) = \boldsymbol{V}^{1/2})^{-1}\Sigma\boldsymbol{V}^{1/2})^{-1} = \boldsymbol{\rho}.$$