Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 7: 26/08/24 – 31/08/24

Introducción

- Al inicio de cualquier investigación suele haber una escasa teoría sobre el campo que se va abordar, y en consecuencia, **el investigador colecta información sobre un número grande de variables, que a su juicio considera relevantes**.
- Vimos la **dificultad** para **visualizar padrones de asociación** entre ellas.
- Si se tienen **8 variables**, es **necesario estimar** $\binom{8}{2} = 28$ correlaciones (**parámetros**).

Introducción

- El **Análisis de Componentes Principales** (**ACP** o PCA, *Principal Component Analysis*) **transforma linealmente** un conjunto de p **variables correlacionadas** en un conjunto de $k < p$ **variables no correlacionadas**, que contiene la **mayor parte de la variabilidad** presente en el **conjunto original**.
- **ACP** es un método que suministra información sobre la interdependencia entre las variables.
- Si NO EXISTE CORRELACIÓN entre las variables originales, NO TIENE SENTIDO HACER ACP, pues las componentes se corresponderían con cada variable por orden de magnitud en la varianza.

Introducción

- Si las variables originales siguen una distribución **normal multivariada** [1], **las componentes principales también serán normales multivariadas y serán independientes**.
- Con frecuencia, el ACP **revela relaciones** de las que no se sospechaba inicialmente, y por tanto este análisis **permite interpretaciones** de los datos **que no podrían ser derivadas directamente** de las variables originales.
- **ACP** fue **propuesta** por **Karl Pearson** en 1901 y fue **estudiada** y llamada ACP por **Harold Hotelling** en 1933.

[1] Este supuesto NO ES PREREQUISITO en el ACP

Objetivos del ACP

Objetivos principales del ACP:

- ✓ **Reducir** la dimensión de los datos
- ✓ Obtener **combinaciones** de las variables que sean **interpretables**.
- ✓ **Eliminar** las variables que **aporten poco** al estudio.
- ✓ **Describir** y comprender la **estructura de correlación** de las variables originales.

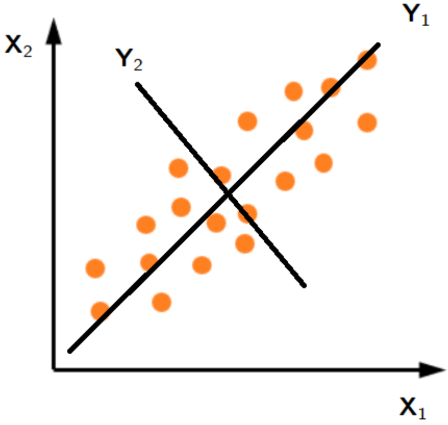
Generación de Componentes Principales

Antes de definir/construir/generar las componentes principales, tenga en cuenta que:

- **Componentes principales exactas**: son obtenidas a partir de la **matriz de varianzas y covarianzas poblacional $\boldsymbol{\Sigma}$** cuando ésta es **conocida**.
- **Componentes principales estimadas:** se extraen de la **matriz de varianzas y covarianzas muestrales S**, cuando $\boldsymbol{\Sigma}$ es desconocida.

Generación de Componentes Principales

Considere que se tienen $p = 2$ **variables correlacionadas** positivamente:



Generación de Componentes Principales

Observe que:

- Y_1 es la **dirección de mayor variabilidad** de X_1 y X_2,
- Y_2 es la **segunda dirección de mayor variabilidad** de X_1 y X_2,
- Y_1 y Y_2 son perpendiculares (**no correlacionados**).

Podemos hacer una **transformación en los ejes**

$$X_1 \qquad \text{e} \qquad X_2$$

para los ejes de

$$Y_1 \qquad \text{e} \qquad Y_2$$

Generación de Componentes Principales

Sea

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}_{p\times 1}$$

un vector $p-$ variado con **vector de medias** y **matriz de varianzas y covarianzas poblacionales**

Generación de Componentes Principales

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}_{p\times 1} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}_{p\times p}$$

Pontificia Universidad JAVERIANA Bogotá

Generación de Componentes Principales

Aquí el **interés** es que algunas de las variables $X_1, X_2, \cdots, X_p$ estén **correlacionadas**, es decir $\sigma_{ik} \neq 0$ para $i \neq k$, $i, k = 1, 2, \ldots, p$. En este caso, existe **redundancia entre dimensiones** y el interés es **reducir la dimensionalidad** del problema costruyendo **nuevas variables no correlacionadas** entre que sean **combinaciones lineales de las X_i's**.

Las $k < p$ **nuevas variables** pueden **explicar gran parte de la variabilidad existente** en las p **variables originales**.

Generación de Componentes Principales

Sea

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

los **autovalores** de $\mathbf{\Sigma}$, y

$$\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$$

sus correspondientes **autovectores**, donde

$$\mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{bmatrix}$$

Generación de Componentes Principales

satisface

i. $\mathbf{e}_i^\top \mathbf{e}_j = 0$ para todo $i \neq j$
ii. $\mathbf{e}_i^\top \mathbf{e}_i = 1$ para todo $i = 1, 2, \ldots, p$.
iii. $\boldsymbol{\Sigma}_{p\times p}\,\mathbf{e}_i = \lambda\,\mathbf{e}_i$ para todo $i = 1, 2, \ldots, p$.

Considere la **matriz ortogonal**

$$\mathbf{O} = \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{p1} \\ e_{12} & e_{22} & \cdots & e_{p2} \\ \vdots & \vdots & & \vdots \\ e_{1p} & e_{2p} & \cdots & e_{pp} \end{bmatrix} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \ldots \quad \mathbf{e}_p]$$

El **vector de componentes principales** de $\boldsymbol{\Sigma}$ está dado por

$$\mathbf{Y}_{p\times 1} = \mathbf{O}^\top \mathbf{X}$$

$$= \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & & \vdots \\ e_{p1} & e_{p2} & \cdots & e_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

Generación de Componentes Principales

Así

- La **primera componente** (Y_1) va ser la traspuesta del primer (mayor) vector propio ($\mathbf{e}_1$) correspondiente al primer (mayor) valor propio (λ_1) multiplicado por el vector $\mathbf{X}$. Esto es

$$\begin{aligned} Y_1 &= \mathbf{e}_1^\top \mathbf{X} \\ &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \end{aligned}$$

- La **segunda componente** (Y_2) va ser la traspuesta del segundo (mayor) vector propio ($\mathbf{e}_2$) correspondiente al segundo (mayor) valor propio (λ_2) multiplicado por el vector $\mathbf{X}$. Esto es

$$\begin{aligned} Y_2 &= \mathbf{e}_2^\top \mathbf{X} \\ &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \end{aligned}$$

Generación de Componentes Principales

Y_2 **no está correlacionada** con Y_1 y reúne la máxima variabilidad restante de la variación total que no esta contenida en Y_1.

El proceso se realiza hasta encontrar los p-vectores propios.

Algunas **propiedades**:

- La $i-$**ésima componente principal** de $\boldsymbol{\Sigma}$ está dada por

$$Y_i = \mathbf{e}_i^\top \mathbf{X}$$

- $\mathrm{E}[Y_i] = \mathbf{e}_i^\top \boldsymbol{\mu}$
- $\mathrm{Var}[Y_i] = \mathbf{e}_i^\top \boldsymbol{\Sigma} \mathbf{e}_i = \lambda_i$

Pontificia Universidad JAVERIANA Bogotá

Generación de Componentes Principales

- $\text{Cov}[Y_i, Y_k] = 0 \qquad i \neq k = 1, 2, \ldots, p.$
- La **proporción de varianza total** de **X** que es **explicada** por la $i-$**ésima componente principal** está dada por

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

- La $k-$ésima componente del autovector i

$$\mathbf{e}_i^\top = [e_{i1} \quad e_{i2} \quad \ldots \quad e_{ik} \quad \ldots \quad e_{ip}]$$

mide la **importancia** de la $k-$**ésima variable** sobre la i-**ésima componente principal**, independientemente de las demás variables.

Generación de Componentes Principales

- El **coeficiente de correlación** entre Y_i e X_k está dado por

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

Generación de Componentes Principales

En la práctica no conocemos $\mathbf{\Sigma}$, entonces **estimamos** las **componentes principales** de $\mathbf{\Sigma}$ usando los **autovalores y autovectores** de $\mathbf{S}$, la matriz de varianzas y covarianzas muestrales.

Sean $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ los autovalores de $\mathbf{S}$, y $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_p$ sus correspondientes autovectores. Entonces, la $i-$**ésima componente principal muestral** está dada por

$$\hat{Y}_i = \hat{\mathbf{e}}_i^\top \mathbf{X}$$

y

- $\text{Var}[\widehat{Y}_j] = \hat{\lambda}_j$,
- $\text{Cor}[\widehat{Y}_i, \widehat{Y}_k] = 0 \qquad i \neq k = 1, 2, \ldots, p$.

Generación de Componentes Principales

- La **proporción de varianza total** de $\mathbf{X}$ que es **explicada por la** $j-$**ésima componente principal muestral** está dada por

$$\frac{\hat{\lambda}_j}{\sum_{j=1}^{p} \hat{\lambda}_j}$$

- El **coeficiente de correlación** entre $\hat{Y}_i$ e X_k está dado por

$$r_{\widehat{Y}_i, \mathbf{X}_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}} \qquad i, k = 1, 2, \ldots, p$$

Pontificia Universidad JAVERIANA Bogotá

Generación de Componentes Principales

Example

Suponga que

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim \left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix} \right)$$

Obtenga las componentes principales.

Example 8.1 de Jhonson and Wichern (2013), Applied Multivariate Statistical Analysis, pp. 434.

Observación: $\sigma_{31} = \sigma_{13} = 0$ *y* $\sigma_{32} = \sigma_{23} = 0$*, es decir* X_2 ***no se correlaciona con las demás variables***.

Generación de Componentes Principales

Solución:
Debemos encontrar los autovectores de **S**.

```
# ----- ACP ----- #
Sigma<-matrix(c(1,-2,0,-2,5,0,0,0,2),ncol=3)
> Sigma
     [,1] [,2] [,3]
[1,]    1   -2    0
[2,]   -2    5    0
[3,]    0    0    2
auto<-eigen(Sigma)
auto$values
[1] 5.8284271 2.0000000 0.1715729
```

Pontificia Universidad JAVERIANA Bogotá

Generación de Componentes Principales

```
auto$vectors
           [,1] [,2]      [,3]
[1,] -0.3826834    0 0.9238795
[2,]  0.9238795    0 0.3826834
[3,]  0.0000000    1 0.0000000
```

De ahí tenemos

$$\lambda_1 = 5.83 \qquad \lambda_2 = 2 \qquad \lambda_3 = 0.17$$

y

$$\mathbf{e}_1 = \begin{bmatrix} -0.383 \\ 0.924 \\ 0 \end{bmatrix} \qquad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \mathbf{e}_3 = \begin{bmatrix} 0.924 \\ 0.383 \\ 0 \end{bmatrix}$$

Pontificia Universidad JAVERIANA Bogotá

Generación de Componentes Principales

Por lo tanto las **componentes principales** son:

$$Y_1 = \mathbf{e}_1^\top \mathbf{X} = -0.383X_1 + 0.924X_2$$

$$Y_2 = \mathbf{e}_2^\top \mathbf{X} = X_3$$

$$Y_3 = \mathbf{e}_3^\top \mathbf{X} = 0.924X_1 + 0.383X_2$$

*Observación: X_3 es una de las componentes porque **no está correlacionada** con ninguna de las otras variables. Además, porque su información no es llevada al nuevo sistema por ninguna de las otras componentes.*

Generación de Componentes Principales

La **proporción de la varianza total explicada por** Y_1 es

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{5.83}{8} \approx 0.73$$

Esto significa que el **73 % de la varianza total es explicada por la primera componente principal**.

Generación de Componentes Principales

La **proporción de la varianza total explicada por las dos primeras componentes principales** es

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{5.83 + 2}{8} \approx 0.98$$

Esto significa que el **98 % de la varianza total es explicada por las dos primeras componentes principales**.

Generación de Componentes Principales

Por otro lado,

$$\rho_{Y_1,X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = 0.925$$

$$\rho_{Y_1,X_2} = \frac{e_{21}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = 0.998$$

- En la primera componente principal, la variable X_2 tiene la mayor ponderación y ella también tiene la mayor correlación con Y_1.
- La correlación de X_1 con Y_1 es casi tan grande, en magnitud, como la de X_2 con Y_1, lo que indica que las dos variables son casi igualmente importantes para la primera componente principal.

Generación de Componentes Principales

- Los tamaños relativos de los coeficientes de X_1 y X_2 sugieren que X_2 contribuye más a la determinación de Y_1 que X_1.

También

$$\rho_{Y_2,X_1} = \rho_{Y_2,X_2} = 0$$

$$\rho_{Y_2,X_3} = \frac{\sqrt{\lambda_2}}{\sigma_{33}} = \frac{\sqrt{2}}{\sqrt{2}} = 1$$

Las demás correlaciones puede ser despreciadas puesto que la tercera componente principal no es importante.

Generación de Componentes Principales

Cuando las variables $X_i's$ son de magnitudes diferentes, las variabilidades son diferentes. Esa diferencia va influenciar en el análisis de componentes principales. En esos casos podemos recurrir a la estandarización de las variables y en esos casos el análisis se conoce como Análisis de Componentes Principales vía Matriz de Correlación.

Generación de Componentes Principales

Si

$$\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

y si las σ_{ii} son muy diferentes, defina

$$\mathbf{Z} = [Z_1 \quad Z_2 \quad \ldots \quad Z_p]^\top$$

donde

$$Z_i = \frac{X_i - \mu_i}{\sigma_{ii}} \qquad \mu_i = \mathrm{E}[X_i] \qquad \sigma_{ii} = \mathrm{Var}[X_i]$$

Aquí

$$\mathrm{Cov}[\mathbf{Z}] = \mathrm{Cor}[\mathbf{X}] = \mathbf{P}$$

apartir de la cuál se obtienen las componentes principales

$$\mathbf{Y}_{p\times 1} = \mathbf{O}^\top \mathbf{Z}$$

Generación de Componentes Principales

donde

$$\mathbf{O} = \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{p1} \\ e_{12} & e_{22} & \cdots & e_{p2} \\ \vdots & \vdots & & \vdots \\ e_{1p} & e_{2p} & \cdots & e_{pp} \end{bmatrix} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \ldots \quad \mathbf{e}_p]$$

es la **matriz ortogonal** cuyas son los **autovalores** de **P**.

Generación de Componentes Principales

La **proporción de varianza total** de $\mathbf{Z}$ que es **explicada por la** $i-$**ésima componente principal** está dada por

$$\frac{\lambda_i}{p}$$

donde

$$\lambda_i = \text{Var}[Y_i] = \mathbf{e}_i^\top \, \mathbf{P} \, \mathbf{e}$$

La **correlación entre** Y_i **e** X_k está dada por

$$\rho_{Y_i, X_k} = e_{ik}\sqrt{\lambda_i} \qquad\qquad i, k = 1, 2, \ldots, p$$

Generación de Componentes Principales

En la práctica **P** es **deconocida**, así que se utiliza la **matriz de correlación muestral R**, a partir de la cuál se obtienen los **autovalores y autovectores estandarizados** y realizando el mismo procedimiento, se tiene que la **proporción de varianza total** de **Z** que es **explicada por la** $i-$**ésima componente principal muestral** es

$$\frac{\hat{\lambda}_i}{p}$$

donde

$$\hat{\lambda}_i = \mathrm{Var}[Y_i] = \hat{\mathbf{e}}_i^\top \, \mathbf{R} \, \hat{\mathbf{e}}$$

Generación de Componentes Principales

y la **correlación entre** $\hat{Y}_i$ e X_k está dada por

$$r_{\widehat{Y}_i, X_k} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i} \qquad i, k = 1, 2, \ldots, p$$

Generación de Componentes Principales

Example

Considere un vector bivariado cuyas matrices de covarianzas y de correlación están dadas por

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix} \qquad \mathbf{P} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$$

Determine las componentes principales en ambos casos.

Example 8.2 de de Jhonson and Wichern (2013), Applied Multivariate Statistical Analysis, pp. 437.

Observación: Note que las ***varianzas*** *son muy* ***diferentes****.*

Solución:

- Componentes derivadas de $\mathbf{\Sigma}$:

$$Y_1 = \mathbf{e}_1^\top \mathbf{X} = 0.040X_1 + 0.999X_2$$
$$Y_2 = \mathbf{e}_2^\top \mathbf{X} = -0.999X_1 + 0.040X_2$$

Debido a que X_2 tiene una **gran varianza, ella domina completamente la primera componente principal**. Esta componente explica una proporción de

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101} = 0.992$$

de la varianza total.

Generación de Componentes Principales

- Componentes derivadas de $\mathbf{P}$:

$$Y_1 = \mathbf{e}_1^\top \mathbf{Z} = 0.707Z_1 + 0.707Z_2$$
$$Y_2 = \mathbf{e}_2^\top \mathbf{Z} = -0.707Z_1 + 0.707Z_2$$

Observe que **las variables contribuyen igualmente a la primera componente principal**. Esta componente explica una proporción de

$$\frac{\lambda_1}{p} = 0.70$$

de la varianza total.

Generación de Componentes Principales

Además

$$\rho_{Y_1,Z_1} = e_{11}\sqrt{\sigma_1} = 0.707\sqrt{1.4} = 0.873$$
$$\rho_{Y_1,Z_2} = e_{21}\sqrt{\sigma_1} = 0.707\sqrt{1.4} = 0.873$$

las variables estandarizadas tienen la misma correlación con la primera componente principal.

Conclusión: la estandarización afecta bastante los resultados, y que las componentes principales derivadas de $\boldsymbol{\Sigma}$ son diferentes de las derivadas de $\mathbf{P}$.

Generación de Componentes Principales

Notas

- En la interpretación de las componentes principales se deben tener en cuenta los coeficientes e_{ik} de las componentes y las correlaciones ρ_{Y_i,X_k}.
- Las correlaciones permiten analizar la importancia de las variables aunque tengan diferentes varianzas. Sin embargo, miden solamente la importancia de una sola X_k sin tener en cuenta las otras variables presentes en la componente.