Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 8: 02/09/24 – 07/09/24

Introducción

Recuerde que:

✓

$$\underbrace{\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}}_{\mathbf{X}} \sim \left(\underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}}_{\boldsymbol{\mu}}, \underbrace{\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}}_{\boldsymbol{\Sigma}} \right)$$

✓ Las $X_i's$ $i = 1, 2, \ldots, p$ en **X están correlacionadas**.

Introducción

- ✓ El objetivo del ACP es **explicar la estructura de $\mathbf{\Sigma}_{p\times p}$** por medio de unas pocas $(k < p)$ **combinaciones lineales de las X_i's que sean no correlacionadas**, y son llamadas **componentes principales**.
- ✓ Para explicar toda la variabilidad en $\mathbf{X}$ se necesitan las p componentes principales, pero generalmente, **la mayor parte de esa variabilidad puede ser explicada por un número pequeño de éstas**.
- ✓ En general, el propósito del ACP es **reducir la dimensionalidad de los datos y facilitar la interpretación**.

Introducción

¿Cuántas componentes principales debemos retener?

Para responder esta pregunta se debe considerar:

- la cantidad de la **varianza total** muestral **explicada**,
- los **tamaños** relativos de los **autovalores**,
- las **interpretaciones** de las componentes.

Selección del Número de Componentes Principales

Existen varios **métodos (indicios)** para determinar el **número de componentes principales**:

1. Proporción de la Varianza Total Explicada.
2. Análisis de la Gráfica de la Varianza Explicada.
3. Criterio de Kaiser: Análisis de la matriz **P** o **R**
4. Análisis práctica de las componentes principales.

Selección del Número de Componentes Principales

Proporción de la Varianza Total Explicada.

Bajo este criterio, se debe mantener en el sistema un número de k componentes que conjuntamente representen un porcentaje del $\gamma\,100\,\%$ de la varianza total:

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

No hay un límite definido para el valor de γ y su selección se hará de acuerdo con la naturaleza del fenómeno estudiado.

Selección del Número de Componentes Principales

Proporción de la Varianza Total Explicada.

Bajo este criterio, se debe mantener en el sistema un número de k componentes que conjuntamente representen un porcentaje del $\gamma\,100\,\%$ de la varianza total:

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

No hay un límite definido para el valor de γ y su selección se hará de acuerdo con la naturaleza del fenómeno estudiado. Existen casos donde se vuelve casi que necesario trabajar con **porcentajes debajo del 90 %.** Algunos autores recomiendan trabajar con **porcentajes por encima del 80 %.**

Selección del Número de Componentes Principales

Análisis de la Gráfica de la Varianza Explicada *(Scree plot)*[1]

Una ayuda visual útil para determinar el número de componentes es el gráfico scree, el cual presenta un **gráfico de λ_i contra i, las magnitudes de los autovalores contra su número**. Para determinar el número apropiado de componentes, **buscamos un codo en el gráfico**. El número de componentes que se toman es el determinado por aquel punto para el cual el resto de los autovalores son relativamente pequeños y aproximadamente del mismo tamaño.

[1] Gráfico de Codo

Selección del Número de Componentes Principales

Criterio de Kaiser

Cuando las **componentes principales** son **extraidas** de **P** o **R**, se necesita un **número grande** (en comparación cuando se extraen de $\mathbf{\Sigma}$) de componentes principales para alcanzar un valor de γ. Recuerde que **cuando se usa P** o **R** la **proporción de varianza que es explicada** por la $i-$ésima componente es

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

Entonces, el **criterio para la selección del valor de** k, es de mantener en el sistema apenas las componentes relacionadas a los **autovalores** $\lambda_i \geq 1$.

Selección del Número de Componentes Principales

Análogamente, cuando el análisis esta **basado en la matriz $\boldsymbol{\Sigma}$** o **$\mathbf{S}$**, se puede mantener en el sistema las componentes relacionadas a

$$\lambda_i \geq \widehat{\lambda}_m$$

donde

$$\widehat{\lambda}_m = \frac{\sum_{i=1}^{p} \widehat{\lambda}_i}{p}$$

es la media de los autovalores.

Selección del Número de Componentes Principales

Example

En un estudio del **tamaño y la forma de las tortugas pintadas**, Jolicoeur y Mosimann (1963) midieron la **longitud de la caparazón** (X_1), su **amplitud** (X_2) y su **altura** (X_3). Los datos sugirieron el **análisis en términos de los logaritmos de las variables**. (Jolicoeur, generalmente sugiere el empleo de los logaritmos en los estudios de tamaño y forma (alometría).

Ejemplo 8.4 de Jhonson and Wichern (2013), Applied Multivariate Statistical Analysis, pp. 445

Selección del Número de Componentes Principales

Female			Male		
Length (x_1)	Width (x_2)	Height (x_3)	Length (x_1)	Width (x_2)	Height (x_3)
98	81	38	93	74	37
103	84	38	94	78	35
103	86	42	96	80	35
105	86	42	101	84	39
109	88	44	102	85	38
123	92	50	103	81	37
123	95	46	104	83	39
133	99	51	106	83	39
133	102	51	107	82	38
133	102	51	112	89	40
134	100	48	113	88	40
136	102	49	114	86	40
138	98	51	116	90	43
138	99	51	117	90	41
141	105	53	117	91	41
147	108	57	119	93	41
149	107	55	120	89	40
153	107	56	120	93	44
155	115	63	121	95	42
155	117	60	125	93	45
158	115	62	127	96	45
159	118	63	128	95	45
162	124	61	131	95	46
177	132	67	135	106	47

Selección del Número de Componentes Principales

```
# ----- Tortugas ----- #
require(tidyverse)
Tortugas <- read_excel("D:/Desktop/Tortugas.xlsx")
X<-Tortugas %>% as.data.frame()

Machos<-X[,4:6]          # extraemos a los machos
Log_Machos<-log(Machos) # ln de los machos
```

Determinamos el **vector de medias** y la **matriz de covarianzas** del logaritmo natural de las dimensiones de las 24 tortugas machos:

```
Xbarra<-colMeans(Log_Machos)
S<-cov(Log_Machos)
```

Selección del Número de Componentes Principales

De ahí, sigue que

$$\bar{\mathbf{x}} = \begin{bmatrix} 4.725 \\ 4.478 \\ 3.703 \end{bmatrix} \qquad \mathbf{S} = 10^{-3} \begin{bmatrix} 11.072 & 8.019 & 8.160 \\ 8.019 & 6.417 & 6.005 \\ 8160 & 6.005 & 6.773 \end{bmatrix}$$

El primer paso para aplicar el ACP es comprobar si su uso es válido para nuestros datos. Para ello usamos la **Pueba de esfericidad de Bartlett**, donde se prueba la **hipótesis nula** de que **las variables no están correlacionadas**:

$$H_0 : \mathbf{P} = \mathbf{I}$$
$$H_1 : \mathbf{P} \neq \mathbf{I}$$

Selección del Número de Componentes Principales

```
# --- Esfericidad de Bartlett
require(psych)
R<-cor(Log_Machos)
cortest.bartlett(R)
```

Observación: Evidentemente el uso o aplicación del ACP es válido cuando H_0 *es rechazada.*

En este caso tenemos

$$p - \text{valor} = 8.728986e - 96 < 0.05 = \alpha$$

por lo tanto hay suficiente evidencia en la muestra para rechazar H_0.

Selección del Número de Componentes Principales

Determinamos las componentes

```
# --- Componentes
aa<-eigen(S)
aa<-eigen(S)
lambdai<-aa$values
[1] 0.0233033471 0.0005983049 0.0003598360
e<-aa$vectors
           [,1]       [,2]       [,3]
[1,] 0.6831023 -0.1594791  0.7126974
[2,] 0.5102195 -0.5940118 -0.6219534
[3,] 0.5225392  0.7884900 -0.3244015
```

Selección del Número de Componentes Principales

Luego

$$\widehat{Y}_1 = 0.683\,\ln(\text{longitud}) + 0.510\,\ln(\text{amplitud}) + 0.523\,\ln(\text{altura})$$

$$\widehat{Y}_2 = -1.59\,\ln(\text{longitud}) - 0.594\,\ln(\text{amplitud}) + 0.788\,\ln(\text{altura})$$

$$\widehat{Y}_3 = -0.713\,\ln(\text{longitud}) + 0.622\,\ln(\text{amplitud}) + 0.324\,\ln(\text{altura})$$

Selección del Número de Componentes Principales

La proporción de varianza explicada por cada componente:

- la primera componente explica el

$$\frac{\widehat{\lambda}_1}{\sum_{i=1}^{p} \widehat{\lambda}} = \frac{0.0233033471}{0.02426149} \approx 0.961$$

  96.1 % de la variabilidad total muestral.

- la segunda componente explica el

$$\frac{\widehat{\lambda}_2}{\sum_{i=1}^{p} \widehat{\lambda}} = \frac{0.0005983049}{0.02426149} \approx 0.025$$

  2.5 % de la variabilidad total muestral.

Selección del Número de Componentes Principales

- la tercera componente explica el

$$\frac{\widehat{\lambda}_3}{\sum_{i=1}^{p} \widehat{\lambda}} = \frac{0.0003598360}{0.02426149} \approx 0.015$$

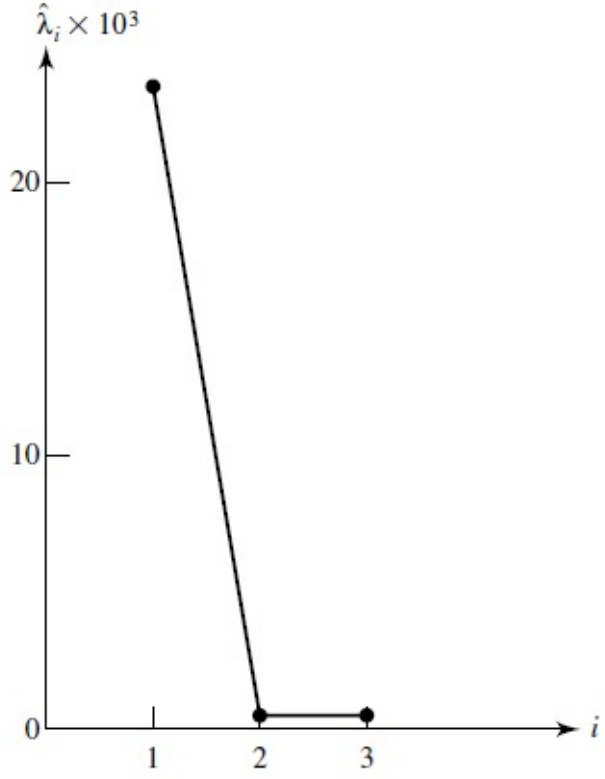
1.5 % de la variabilidad total muestral.

Como la primera componente explica más del 90 % de la variabilidad total muestral, seleccionamos sólo esta componente.

También podemos construir el **scree plot**:

```
# --- Scree plot
plot(lambdai, type = "b")
```

Selección del Número de Componentes Principales



De acuerdo con el gráfico se selecciona $k = 1$ componente.

Selección del Número de Componentes Principales

Por otro lado, la **correlación entre esta componente y cada una de las covariables**:

$$\widehat{\rho}_{Y_1,X_1} = \frac{e_{11}\sqrt{\widehat{\lambda}_1}}{\sqrt{\widehat{\sigma}_{11}}} = \frac{0.6831023\sqrt{0.0233033471}}{\sqrt{0.011072004}} \approx 0.991$$

$$\widehat{\rho}_{Y_1,X_2} = \frac{e_{12}\sqrt{\widehat{\lambda}_1}}{\sqrt{\widehat{\sigma}_{22}}} = \frac{0.5102195\sqrt{0.0233033471}}{\sqrt{0.006416726}} \approx 0.9723$$

$$\widehat{\rho}_{Y_1,X_3} = \frac{e_{13}\sqrt{\widehat{\lambda}_1}}{\sqrt{\widehat{\sigma}_{33}}} = \frac{0.5225392\sqrt{0.0233033471}}{\sqrt{0.006772758}} \approx 0.9693$$

es **alta**.

Selección del Número de Componentes Principales

Resumiendo:

Variable	$\hat{\mathbf{e}}_1\,(\rho_{\hat{Y}_1,X_k})$	$\hat{\mathbf{e}}_2$	$\hat{\mathbf{e}}_3$
$\ln(\text{longitud})$	0.683(0.99)	-0.159	-0.713
$\ln(\text{amplitud})$	0.510(0.97)	-0.594	0.622
$\ln(\text{altura})$	0.523(0.97)	0.788	0.324
$\hat{\lambda}_i$	23.30×10^{-3}	0.60×10^{-3}	0.36×10^{-3}
Prop. Acumulada	96.1	98.5	100

La primera componente principal:

$$\begin{aligned}\hat{Y}_1 &= 0.683\ln(\text{longitud}) + 0.510\ln(\text{amplitud}) + 0.523\ln(\text{altura}) \\ &= \ln\left[\text{longitud}^{0.683} \cdot \text{amplitud}^{0.510} \cdot \text{altura}^{0.523}\right]\end{aligned}$$

Selección del Número de Componentes Principales

tiene una **interpretación interesante**, pues puede ser considerada como **el volumen de una caja con dimensiones ajustadas**. Por ejemplo, la altura ajustada es $\text{altura}^{0.523}$, la cual influye en la forma **redondeada de la caparazón**.

Selección del Número de Componentes Principales

tiene una **interpretación interesante**, pues puede ser considerada como **el volumen de una caja con dimensiones ajustadas**. Por ejemplo, la altura ajustada es $\text{altura}^{0.523}$, la cual influye en la forma **redondeada de la caparazón**.

En **R** podemos realizar el ACP usando el paquete `princomp` del paquete `stats`:

```
# --- Usando princomp
ACP_Turtle<-princomp(Log_Machos,cor = FALSE)
ACP_Turtle$loadings    # autovectores
CP_Turtle$center       # vector de medias
```

Selección del Número de Componentes Principales

```
# --- screeplot
screeplot(ACP_Turtle,
          npcs = 3,     # No. comp.
          type = "lines")
```

Selección del Número de Componentes Principales

```
# --- screeplot
screeplot(ACP_Turtle,
          npcs = 3,     # No. comp.
          type = "lines")
```

Example

Los datos `Protein` del paquete `MultBiplotR` contiene información sobre datos nutricionales de 9 diferentes fuentes de proteínas para los habitantes de 25 países europeos alrededor de 1970:

- **Comunist**: Sí el país es comunista o no
- **Region**: Tres regiones Norte Centro Sur
- **RedMeat**: Consumo de proteínas provenientes de carnes rojas.

Selección del Número de Componentes Principales

Example

- **WhiteMeat**: Consumo de proteínas provenientes de carnes blancas;
- **Eggs**: Consumo de proteínas del huevo;
- **Milk**: Consumo de proteínas de la leche;
- **Fish**: Consumo de proteínas provenientes del pescado;
- **Cereals**:Consumo de proteínas procedentes de cereales;
- **Starch**: Consumo de proteínas provenientes de carbohidratos;
- **Nuts**: Consumo de proteínas procedentes de cereales, frutos secos y semillas oleaginosas;
- **FruitVeg**: Consumo de proteínas procedentes de frutas y verduras.

Selección del Número de Componentes Principales

Example

Estos datos fueron colectados incialmente para entender las diferencias nutricionales entre los paises europeos.

Realice un análisis de componentes principales.

Solución:

Cargamos los datos

```
 # ----- Protein ----- #
require(MultBiplotR)
data("Protein")
X<-Protein[, -c(1,2)]
```

Selección del Número de Componentes Principales

y verificamos si el ACP es válido

```
 # --- Esfericidad de Bartlett
require(psych)
R<-cor(X)
cortest.bartlett(R)
```

Determinamos las componentes principales usando `princomp`

```
ACP_Protein<-princomp(X,cor = FALSE)
ACP_Protein$loadings    # autovectores
```

Para determinar el número de componentes, usamos

Selección del Número de Componentes Principales

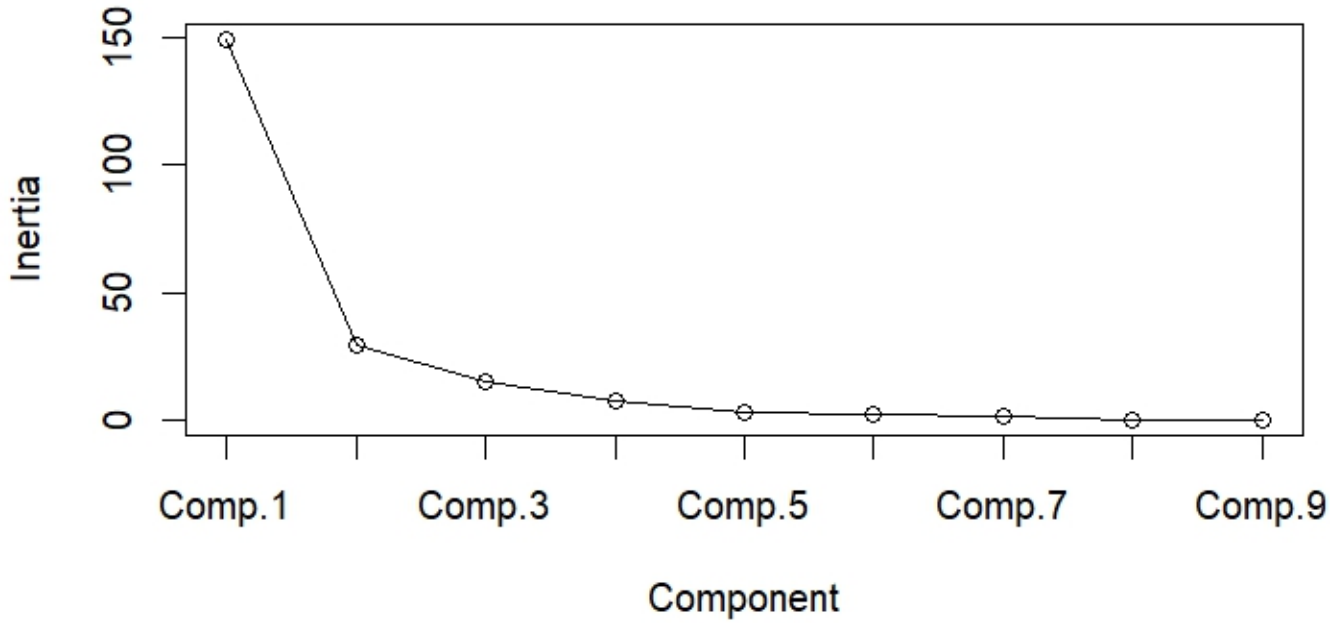
- el criterio de proporción de varianza explicada

```
summary(ACP_Protein)
Importance of components:
                            Comp.1    Comp.2  Comp.3
Standard deviation     12.2075647 5.4287003  3.87527011
Proportion of Variance  0.7105325 0.1405134  0.07160277
Cumulative Proportion   0.7105325 0.8510459  0.92264866
```

- el gráfico screeplot

```
# --- screeplot
screeplot(ACP_Protein,
          npcs = 9, # No. comp.
          type = "lines",
          main="")
```

Selección del Número de Componentes Principales



De acuerdo con el gráfico y con el criterio de la proporción de varianza, se selecciona $k = 2$ componentes:

Selección del Número de Componentes Principales

```
round(ACP_Protein$loadings[,1:2],4)
                  Comp.1  Comp.2
Red_Meat          0.1507  0.1327
White_Meat        0.1295  0.0434
Eggs              0.0673  0.0209
Milk              0.4254  0.8309
Fish              0.1270 -0.2923
Cereal           -0.8609  0.4062
Starch            0.0669 -0.0760
Nuts             -0.1139 -0.0701
Fruits_Vegetables -0.0202 -0.1692
```

Selección del Número de Componentes Principales

Observe que

- el **primer componente principal** tiene **valores positivos** para las carnes **roja** y **blanca**, los **huevos**, la **leche** y el **pescado**, mientras que los valores de **cereales**, **frutos secos** y **frutas y verduras** son **negativos**. Esto indica/sugiere que el consumo de proteinas de procedencia animal aumenta, mientras que el consumo de proteina no animal disminuye.
- Interprete el **segundo componente principal**