Análisis Multivariado

Producto académico 03

Kevin Heberth Haquehua Apaza

23 de julio del 2025

Table of Contents

# Examen AED multivariado, ACP y AFE

## Ejercicio 1:

Sea un vector aleatorio con matriz de varianza-covarianza dada por

1. Determine las componentes principales y .
2. Calcule la proporción de la varianza total explicada por la primera componente principal.
3. Calcule la matriz de correlaciones a partir de la matriz de covarianzas y determine las componentes principales y a partir de . Calcule la proporción de la varianza total explicada por
4. Calcule la correlación entre las variables y las componentes principales, es decir, calcule

### Solución

a y b)

1. **Calcule la matriz de correlaciones a partir de la matriz de covarianzas y determine las componentes principales y a partir de . Calcule la proporción de la varianza total explicada por**

S <- matrix(c(4, 1, 1, 3), 2, 2) ; S

## [,1] [,2]  
## [1,] 4 1  
## [2,] 1 3

Calcular la matriz de correlaciones

D\_inv <- diag(1 / sqrt(diag(S)))  
R <- D\_inv %\*% S %\*% D\_inv  
R

## [,1] [,2]  
## [1,] 1.0000000 0.2886751  
## [2,] 0.2886751 1.0000000

Ahora hallar los componentes principales

pca <- princomp(covmat = R, cor = TRUE)

Y las varianzas explicadas

explained\_var <- pca$sdev^2 / sum(pca$sdev^2)  
explained\_var

## Comp.1 Comp.2   
## 0.6443376 0.3556624

En el cálculo manual no hya tanta diferencia, debido al cálculo con decimales.

1. **Calcule la correlación entre las variables y las componentes principales, es decir, calcule**

L <- unclass(pca$loadings) ; L

## Comp.1 Comp.2  
## [1,] 0.7071068 0.7071068  
## [2,] 0.7071068 -0.7071068

# Autovalores  
lambda <- pca$sdev^2  
  
# Correlaciones  
cor <- L %\*% diag(sqrt(lambda))  
cor

## [,1] [,2]  
## [1,] 0.8027064 0.5963744  
## [2,] 0.8027064 -0.5963744

## Ejercicio 2:

La siguiente tabla muestra los datos sobre la longitud de huesos registrados de 20 jóvenes a los 8, 8.5, 9 y 9.5 años respectivamente; Verificar si alguno de los individuos es considerado un dato atípico multivariado. Realizar la comprobación paso a paso como se realizó en clase (matricialmente), además tienes que comprobarlo con la función directa en el R

data <- data.frame(X1\_y8 = c(47.8, 46.4, 46.3, 45.1, 47.6, 52.5, 51.2, 49.8,  
 48.1, 45.0, 51.2, 48.5, 52.1, 48.2, 49.6, 50.7,  
 47.2, 53.3, 46.2, 46.3),  
 X2\_y8.5 = c(48.8, 47.3, 46.8, 45.3, 48.5, 53.2, 53.0, 50.0,  
 50.8, 47.0, 51.4, 49.2, 52.8, 48.9, 50.4, 51.7,  
 47.7, 54.6, 47.5, 47.6),  
 X3\_y9 = c(49.0, 47.7, 47.8, 46.1, 48.9, 53.3, 54.3, 50.3,  
 52.3, 47.3, 51.6, 53.0, 53.7, 49.3, 51.2, 52.7,  
 48.4, 55.1, 48.1, 51.3),  
 X3\_y9.5 = c(49.7, 48.4, 48.5, 47.2, 49.3, 53.7, 54.5, 52.7,  
 54.4, 48.3, 51.9, 55.5, 55.0, 49.8, 51.8, 53.3,  
 49.5, 55.3, 48.4, 51.8))

### Solución

Hallemos manualmente

# Calcular vector de medias  
mu <- colMeans(data) ; mu

## X1\_y8 X2\_y8.5 X3\_y9 X3\_y9.5   
## 48.655 49.625 50.570 51.450

data\_centered <- t(apply(data, 1, function(x) x - mu))  
data\_centered <- as.matrix(data\_centered) ; data\_centered

## X1\_y8 X2\_y8.5 X3\_y9 X3\_y9.5  
## [1,] -0.855 -0.825 -1.57 -1.75  
## [2,] -2.255 -2.325 -2.87 -3.05  
## [3,] -2.355 -2.825 -2.77 -2.95  
## [4,] -3.555 -4.325 -4.47 -4.25  
## [5,] -1.055 -1.125 -1.67 -2.15  
## [6,] 3.845 3.575 2.73 2.25  
## [7,] 2.545 3.375 3.73 3.05  
## [8,] 1.145 0.375 -0.27 1.25  
## [9,] -0.555 1.175 1.73 2.95  
## [10,] -3.655 -2.625 -3.27 -3.15  
## [11,] 2.545 1.775 1.03 0.45  
## [12,] -0.155 -0.425 2.43 4.05  
## [13,] 3.445 3.175 3.13 3.55  
## [14,] -0.455 -0.725 -1.27 -1.65  
## [15,] 0.945 0.775 0.63 0.35  
## [16,] 2.045 2.075 2.13 1.85  
## [17,] -1.455 -1.925 -2.17 -1.95  
## [18,] 4.645 4.975 4.53 3.85  
## [19,] -2.455 -2.125 -2.47 -3.05  
## [20,] -2.355 -2.025 0.73 0.35

# Calcular matriz de covarianza e inversa  
S <- cov(data) ; S

## X1\_y8 X2\_y8.5 X3\_y9 X3\_y9.5  
## X1\_y8 6.329974 6.189079 5.777000 5.548158  
## X2\_y8.5 6.189079 6.449342 6.153421 5.923421  
## X3\_y9 5.777000 6.153421 6.918000 6.946316  
## X3\_y9.5 5.548158 5.923421 6.946316 7.464737

S\_inv <- solve(S) ; S\_inv

## X1\_y8 X2\_y8.5 X3\_y9 X3\_y9.5  
## X1\_y8 2.6751033 -2.9161240 0.5007711 -0.1402566  
## X2\_y8.5 -2.9161240 4.3649625 -2.2211535 0.7706152  
## X3\_y9 0.5007711 -2.2211535 4.6620565 -2.9479467  
## X3\_y9.5 -0.1402566 0.7706152 -2.9479467 2.3699236

# Calcular distancia de Mahalanobis manualmente (paso a paso)  
dist\_manual <- (data\_centered) %\*% S\_inv %\*% t(data\_centered)  
dist\_manual<-(diag(dist\_manual)) ; dist\_manual

## [1] 0.7589305 1.2973512 1.7542420 3.8608496 0.8776488 2.8221570  
## [7] 4.0573285 8.1105862 10.9511973 5.8461957 2.8393025 10.5795182  
## [13] 2.5797150 0.6625258 0.3324791 0.8462319 1.1141959 4.4104027  
## [19] 2.1249099 10.1742322

Ahora hallemos con la función directa del R

# Calcular con la función base de R  
dist\_r <- mahalanobis(data, center = mu, cov = S) ; dist\_r

## [1] 0.7589305 1.2973512 1.7542420 3.8608496 0.8776488 2.8221570  
## [7] 4.0573285 8.1105862 10.9511973 5.8461957 2.8393025 10.5795182  
## [13] 2.5797150 0.6625258 0.3324791 0.8462319 1.1141959 4.4104027  
## [19] 2.1249099 10.1742322

Comparemos

# Comparar resultados  
resultado <- data.frame(  
 Observacion = rownames(data),  
 Dist\_Mahal\_Manual = dist\_manual,  
 Dist\_Mahal\_R = dist\_r,  
 Diferencia = abs(dist\_manual - dist\_r)  
)  
  
print(resultado)

## Observacion Dist\_Mahal\_Manual Dist\_Mahal\_R Diferencia  
## 1 1 0.7589305 0.7589305 0.000000e+00  
## 2 2 1.2973512 1.2973512 0.000000e+00  
## 3 3 1.7542420 1.7542420 0.000000e+00  
## 4 4 3.8608496 3.8608496 0.000000e+00  
## 5 5 0.8776488 0.8776488 0.000000e+00  
## 6 6 2.8221570 2.8221570 0.000000e+00  
## 7 7 4.0573285 4.0573285 8.881784e-16  
## 8 8 8.1105862 8.1105862 0.000000e+00  
## 9 9 10.9511973 10.9511973 0.000000e+00  
## 10 10 5.8461957 5.8461957 0.000000e+00  
## 11 11 2.8393025 2.8393025 0.000000e+00  
## 12 12 10.5795182 10.5795182 1.776357e-15  
## 13 13 2.5797150 2.5797150 4.440892e-16  
## 14 14 0.6625258 0.6625258 0.000000e+00  
## 15 15 0.3324791 0.3324791 5.551115e-17  
## 16 16 0.8462319 0.8462319 0.000000e+00  
## 17 17 1.1141959 1.1141959 2.220446e-16  
## 18 18 4.4104027 4.4104027 8.881784e-16  
## 19 19 2.1249099 2.1249099 0.000000e+00  
## 20 20 10.1742322 10.1742322 0.000000e+00

Practimente no se observan diferencias significativas. Ahora veamos los individuos considerados atípicos

# Umbral usando el percentil 97.5 de chi-cuadrado con 4 grados de libertad  
umbral <- qchisq(0.975, df = 4) ; umbral

## [1] 11.14329

# Agregar al dataframe  
resultado$dist\_maha <- dist\_r  
resultado$Atipico <- ifelse(resultado$dist\_maha > umbral, "Sí", "No")  
  
# Mostrar resultado  
print(resultado)

## Observacion Dist\_Mahal\_Manual Dist\_Mahal\_R Diferencia dist\_maha Atipico  
## 1 1 0.7589305 0.7589305 0.000000e+00 0.7589305 No  
## 2 2 1.2973512 1.2973512 0.000000e+00 1.2973512 No  
## 3 3 1.7542420 1.7542420 0.000000e+00 1.7542420 No  
## 4 4 3.8608496 3.8608496 0.000000e+00 3.8608496 No  
## 5 5 0.8776488 0.8776488 0.000000e+00 0.8776488 No  
## 6 6 2.8221570 2.8221570 0.000000e+00 2.8221570 No  
## 7 7 4.0573285 4.0573285 8.881784e-16 4.0573285 No  
## 8 8 8.1105862 8.1105862 0.000000e+00 8.1105862 No  
## 9 9 10.9511973 10.9511973 0.000000e+00 10.9511973 No  
## 10 10 5.8461957 5.8461957 0.000000e+00 5.8461957 No  
## 11 11 2.8393025 2.8393025 0.000000e+00 2.8393025 No  
## 12 12 10.5795182 10.5795182 1.776357e-15 10.5795182 No  
## 13 13 2.5797150 2.5797150 4.440892e-16 2.5797150 No  
## 14 14 0.6625258 0.6625258 0.000000e+00 0.6625258 No  
## 15 15 0.3324791 0.3324791 5.551115e-17 0.3324791 No  
## 16 16 0.8462319 0.8462319 0.000000e+00 0.8462319 No  
## 17 17 1.1141959 1.1141959 2.220446e-16 1.1141959 No  
## 18 18 4.4104027 4.4104027 8.881784e-16 4.4104027 No  
## 19 19 2.1249099 2.1249099 0.000000e+00 2.1249099 No  
## 20 20 10.1742322 10.1742322 0.000000e+00 10.1742322 No

Se observa que no se tiene ningun dato atípicos registrado, sin embargo la observación 9, 11 y 20 están cercas a considerarse outliers multivariados.

## Ejercicio 3:

En el Excel *(pregunta3.xlsx)* se muestran los valores de cinco variables obtenidas en 20 alumnos que quieren entrar a alguna universidad del consejo de rectores. Las variables en estudio son la distancia en kilómetros al lugar del colegio en el que estudiaban (DIST), el promedio de horas que hacían actividad física a la semana (EF), índice de masa corporal (IMC), IQ (coeficiente intelectual) y NEM (promedio de notas con el cual postulan a las universidades). Se quiere determinar las relaciones existentes entre dichas variables intentando reducir la dimensionalidad del problema vía un análisis factorial exploratorio.

1. Sobre el conjunto de datos halle si se cumple o no la normalidad multivariada.
2. Verifique si existe presencia de valores outliers para la data.
3. Realice el análisis factorial exploratorio vía componentes principales e interprete y justifique sus conclusiones.

### Solución

Leamos la data

1. **Sobre el conjunto de datos halle si se cumple o no la normalidad multivariada.**

Veamos si los datos siguen una distribución multivariada con los test de Mardia, Henze-Zirkler y Royston

1. **Verifique si existe presencia de valores outliers para la data.**