Análisis Multivariado

Freddy Hernández Barajas

11/1/23

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1 Introducción

En este capítulo se presentan algunos aspectos introductorios del análisis multivariado.

1.1 ¿En qué consiste el análisis multivariado?

Figure 1.1: Ilustración

1.2 ¿En cuáles situaciones se usa el análisis multivariado?

Los conceptos de análisis multivariado se usan en muchas áreas, a continuación algunas de ellas con ejemplos ilustrativos.

1.2.1 Mercadeo

Se estudian seis características acerca de un producto percibidas por un grupo de consumidores, éstas son: calidad del producto, nivel de precio, velocidad de despacho o entrega, servicio, nivel

de uso comparado con otros porductos sustitutos, nivel de satisfacción. Se quiere saber acerca de la incidencia, tanto individual como conjunta, de las variables anteriores en la decisión de compra del producto.

Figure 1.2: Ilustración

1.2.2 Geología

A lo largo de líneas transversales (en inglés “transects”) toman varias muestras del suelo para estudiar los contenidos (en porcentaje) de arena, azufre, magnesio, arcilla, materia orgánica y pH. También se miden otras variables físicas tales como estructura, humedad, conductividad eléctrica y permeabilidad. El objetivo es determinar las características más relevantes del suelo y hacer una clasificación de éstos.

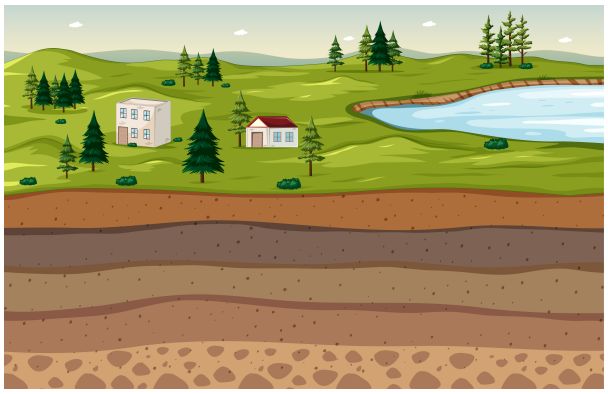


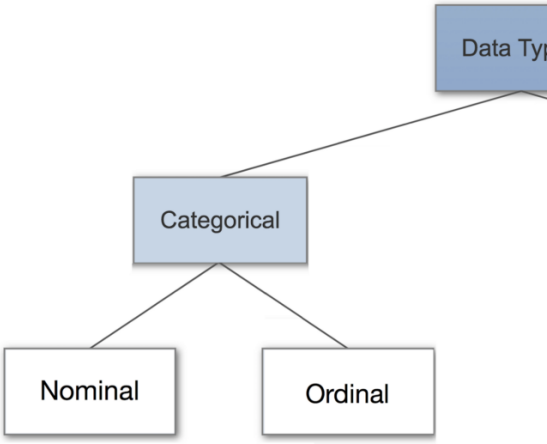
Figure 1.3: Ilustración

1.2.3 Arqueología

Se realizan varias excavaciones en tres regiones donde se tiene la evidencia que habitaron comunidades indígeneas diferentes. Sobre los cráneos conseguidos se midió: la circunferencia, ancho máximo, altura máxima, altura nasal y longitud basialveolar. Esta información permitirá hacer comparaciones entre estas comunidades.

Figure 1.4: Ilustración

1.3 Tipos de variables



En la siguiente figura se muestran los tipos de variables básicos.

1.4 Matriz de diseño

La matriz de diseño se denota por X y es un arreglo de n filas que representan los objetos o sujetos analizados con p colmunas que representan las variables observadas.

$$\mathbb{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Figure 1.5: Ilustración

1.5 Clasificación de los métodos multivariados

En la siguiente figura se presentan los métodos multivariados tradicionales y una clasificación.

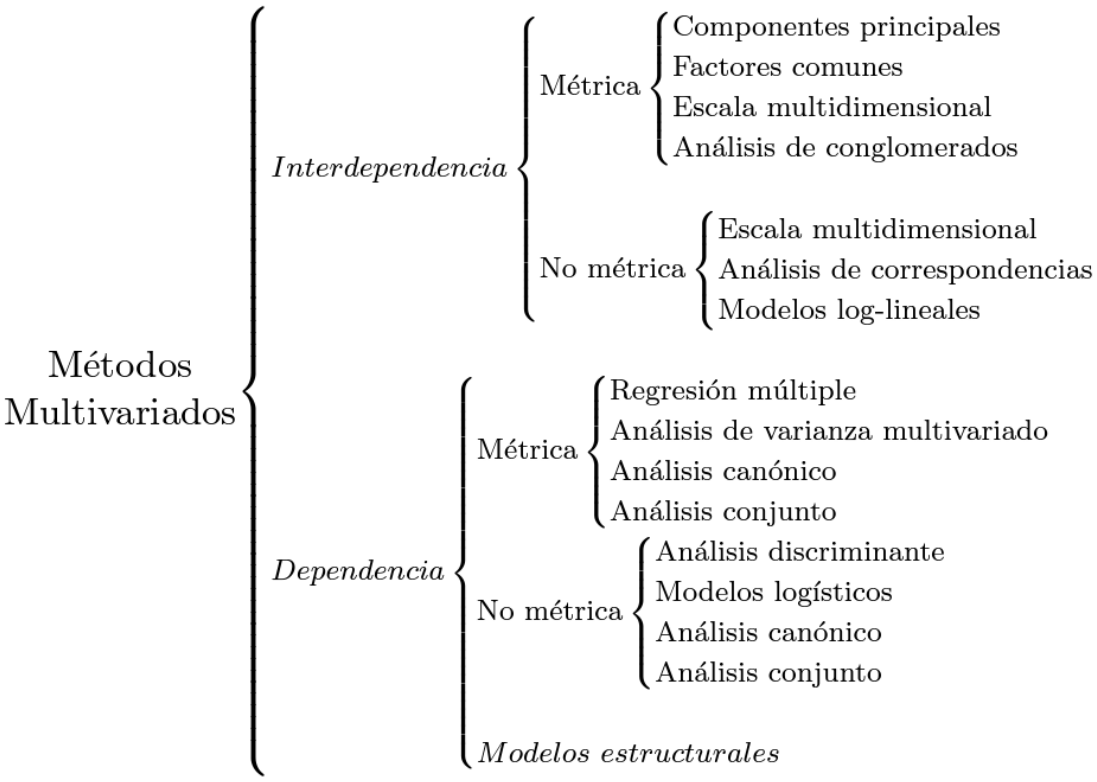


Figure 1.6: Ilustración

2 Componentes principales

En los tiempos modernos es usual tener gran cantidad de datos y es necesario contar con herramientas para manejarlos.

Figure 2.1: Ilustración

El análisis de componentes principales es una herramienta para reducir el numero de variables originales por nuevas variables o componentes "incorrelacionadas".

En el caso de dos variables X_1 y X_2, las componentes principales (PC_1 y PC_2) corresponden a dos nuevas variables que son independientes (perpendiculares) entre ellas. A continuación una ilustración.

2.1 Details

- Suppose $[X_1, X_2, \dots, X_p] = X^\top$ is a set of p random variables, with mean vector μ and variance-covariance matrix Σ.

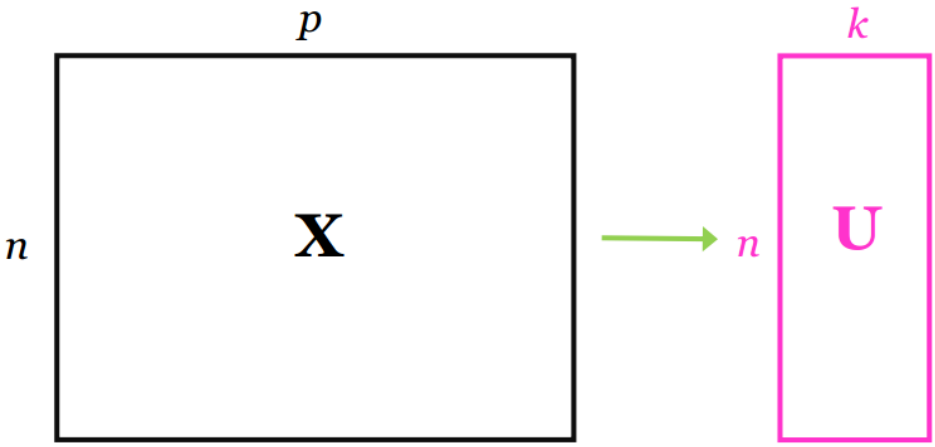


Figure 2.2: Ilustración

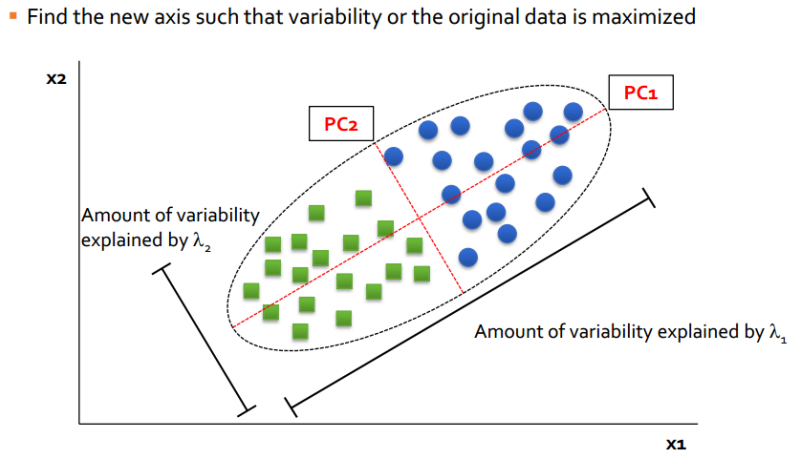


Figure 2.3: Ilustración

- We want to define p linear combinations of $X^\top$ that represent the information in $X^\top$ more parsimoniously.

- Specifically, find $a_1, \dots, a_p$ such that $a_1^\top X, \dots, a_p^\top X$ gives the same information as $X^\top$, but the new random variables, $a_1^\top X, \dots, a_p^\top X$, are 'nicer'.
- Suppose $[X_1, X_2, \dots, X_p] = X^\top$ is a set of p random variables, with mean vector μ and variance-covariance matrix Σ.

- We want to define p linear combinations of $X^\top$ that represent the information in $X^\top$ more parsimoniously.

- Specifically, find $a_1, \dots, a_p$ such that $a_1^\top X, \dots, a_p^\top X$ gives the same information as $X^\top$, but the new random variables, $a_1^\top X, \dots, a_p^\top X$, are 'nicer'.

2.2 Propiedades

1) $Var(a_i^\top X) = a_i^\top \Sigma a_i = \lambda_i$.

2) a_i and a_j are orthogonal, i.e., $a_i^\top a_j = 0$.

3) $Cov(a_i^\top X, a_j^\top X) = a_i^\top \Sigma a_j = a_i^\top \lambda_j a_j = \lambda_j a_i^\top a_j = 0$.

4) $Tr(\Sigma) = \lambda_1 + \cdots + \lambda_p =$ sum of variances for all p principal components, and for $X_1, \dots, X_p$.

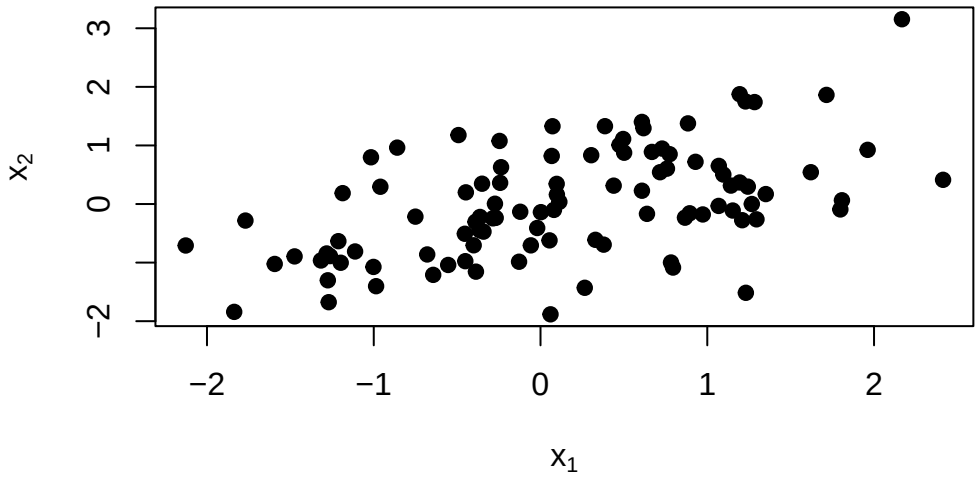
5) The importance of the i^{th} principal component is $\lambda_i / Tr(\Sigma)$.

2.3 Ejemplo

Supongamos que tenemos dos variables cuantitativas X_1 y X_2 como se muestra a continuacion. Queremos encontrar un eje sobre el cual proyectar los puntos de tal manera que las sombras tengan la mayor variabilidad.

```
mu <- c(0,0)                          # Mean
Sigma <- matrix(c(1, 0.5,
                  0.5, 1), ncol=2)  # Covariance matrix

# Generate sample from N(mu, Sigma)
library(MASS)
```

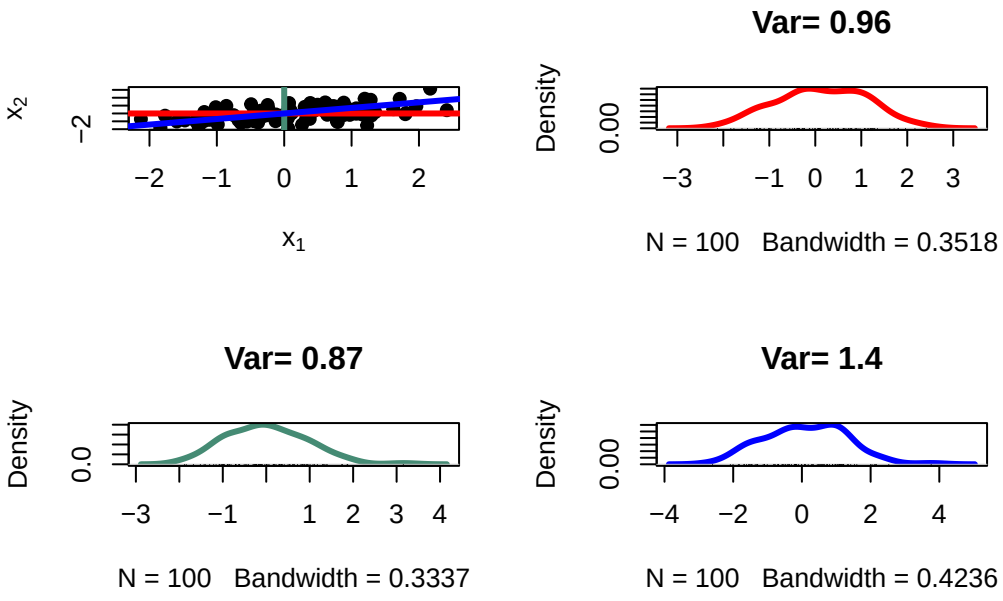
```r
dt <- mvrnorm(100, mu=mu, Sigma=Sigma)
plot(dt, xlab=expression(x[1]), pch=19,
     ylab=expression(x[2]))
```



Proyectando los puntos sobre los vectores $(1,0)^\top$, $(0,1)^\top$ y $(1/\sqrt{2}, 1/\sqrt{2})^\top$.

```r
par(mfrow=c(2, 2))

plot(dt, xlab=expression(x[1]), ylab=expression(x[2]), pch=19)
abline(h=0, col='red', lwd=3)
abline(v=0, col='aquamarine4', lwd=3)
abline(a=0, b=1/sqrt(2), col='blue', lwd=3)

y <- dt %*% matrix(c(1, 0), nrow=2)
plot(density(y), lwd=3, col='red',
     main=paste('Var=', round(var(y),2)))
rug(y)

y <- dt %*% matrix(c(0, 1), nrow=2)
plot(density(y), lwd=3, col='aquamarine4',
     main=paste('Var=', round(var(y),2)))
```

```
rug(y)

y <- dt %*% matrix(c(1/sqrt(2), 1/sqrt(2)), nrow=2)
plot(density(y), lwd=3, col='blue',
     main=paste('Var=', round(var(y),2)))
rug(y)
```



2.4 Ejemplo

A continuacion una base de datos sobre medidas corporales a 36 estudiantes de la universidad el año pasado.

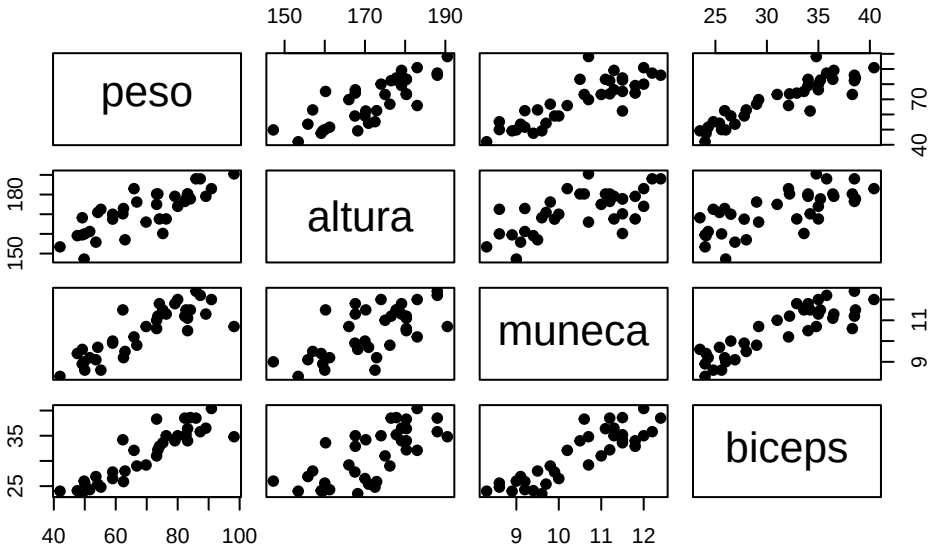
```
myurl <- 'https://raw.githubusercontent.com/fhernanb/datos/master/medidas_cuerpo'
datos <- read.table(file=myurl, header=T, sep='')
head(datos)
```

```
  edad peso altura   sexo muneca biceps
1   43 87.3  188.0 Hombre   12.2   35.8
2   65 80.0  174.0 Hombre   12.0   35.0
3   45 82.3  176.5 Hombre   11.2   38.5
```

```
4   37 73.6  180.3 Hombre   11.2   32.2
5   55 74.1  167.6 Hombre   11.8   32.9
6   33 85.9  188.0 Hombre   12.4   38.5
```

Vamos a crear varios diagramas de dispersión para mostrar la relación entre las variables.

```
pairs(datos[, c('peso', 'altura', 'muneca', 'biceps')],
      pch=19)
```



Vamos a calcular la matriz de varianzas y covarianzas sin incluir la variable sexo.

```
dt <- datos[, c('peso', 'altura', 'muneca', 'biceps')]
Sigma <- var(dt)
Sigma
```

```
           peso     altura    muneca    biceps
peso   221.08713 124.728698 14.844667 70.738381
altura 124.72870 110.673968  8.156476 39.021048
muneca  14.84467   8.156476  1.381714  5.400571
biceps  70.73838  39.021048  5.400571 27.398857
```

```r
sum(diag(Sigma))
```

```
[1] 360.5417
```

Eigenvalores e eigenvectores de los datos.

```r
ei <- eigen(Sigma)
ei
```

```
eigen() decomposition
$values
[1] 325.1349702  30.8091070   4.3076215   0.2899759

$vectors
           [,1]        [,2]        [,3]         [,4]
[1,] 0.81049522  0.47697293  0.33927752  0.022024447
[2,] 0.52109937 -0.85221951 -0.04660443 -0.002319266
[3,] 0.05465892  0.04305245 -0.12686657 -0.989476509
[4,] 0.26184984  0.21063052 -0.93092624  0.142988748
```

```r
sum(ei$values)
```

```
[1] 360.5417
```

Eigenvalores e eigenvectores de los datos escalados.

```r
dt.s <- scale(dt)  # Datos escalados
sum(apply(dt.s, MARGIN=2, FUN=var))
```

```
[1] 4
```

```r
ei <- eigen(var(dt.s))
ei
```

```
eigen() decomposition
$values
[1] 3.40781664 0.37981249 0.13515565 0.07721522

$vectors
           [,1]          [,2]       [,3]       [,4]
[1,] -0.5229877 -0.0001370885  0.4515021  0.7229313
[2,] -0.4613002 -0.8346529032 -0.2363538 -0.1862619
[3,] -0.4985496  0.4607270199 -0.7282171  0.0942266
[4,] -0.5149118  0.3018031239  0.4582385 -0.6586336
```

```r
sum(ei$values)
```

```
[1] 4
```

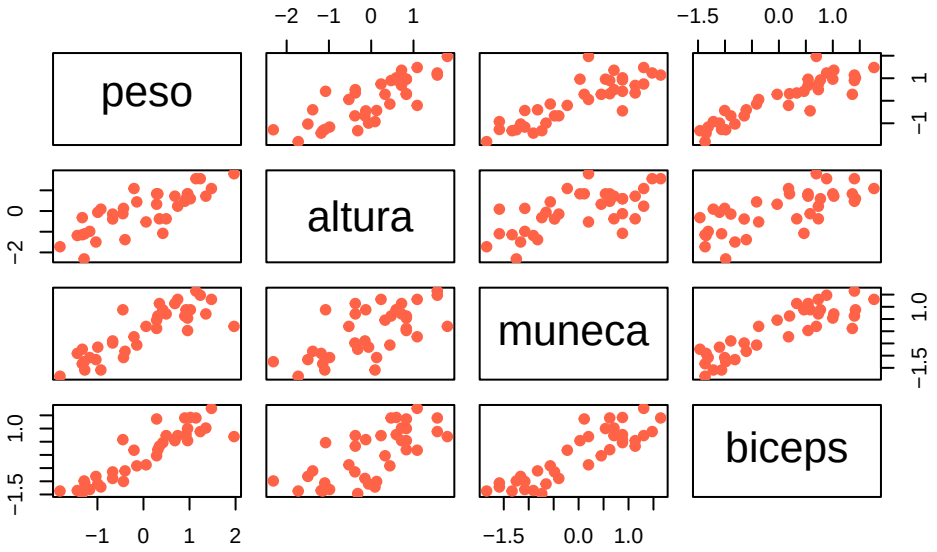
PCA usando la funcion `princomp` de stats.

```r
mod <- prcomp(~ peso + altura + muneca + biceps,
              data=datos, scale=TRUE)
mod
```

```
Standard deviations (1, .., p=4):
[1] 1.8460273 0.6162893 0.3676352 0.2778763

Rotation (n x k) = (4 x 4):
              PC1           PC2        PC3        PC4
peso   -0.5229877  0.0001370885 -0.4515021  0.7229313
altura -0.4613002  0.8346529032  0.2363538 -0.1862619
muneca -0.4985496 -0.4607270199  0.7282171  0.0942266
biceps -0.5149118 -0.3018031239 -0.4582385 -0.6586336
```

Vamos a crear varios diagramas de dispersión para las variables escaladas.

```r
pairs(dt.s[, c('peso', 'altura', 'muneca', 'biceps')],
      pch=19, col='tomato')
```



A continuación una tabla de resumen de la aplicación de las componentes principales.

```
summary(mod)
```

```
Importance of components:
                          PC1     PC2     PC3    PC4
Standard deviation     1.8460 0.61629 0.36764 0.2779
Proportion of Variance 0.8519 0.09495 0.03379 0.0193
Cumulative Proportion  0.8519 0.94691 0.98070 1.0000
```

3 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

References

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.