3. Nociones de Algebra

Javier Rodríguez Barrios

Contents

Introducción

Este capítulo se orienta a una revisión muy básica del algebra lineal asociada a datos multivariados, por lo que se requiere de la revisión de otros documentos con un enfoque matemático, para aquellos que no estén familiarizados con esta temática. El énfasis de este capítulo se orienta, además, a la aplicación del programa RStudio (ver capítulo 2), para la construcción de algunas matrices de uso común en el análisis de datos ecológicos y ambientales (ANDEA), las operaciones matriciales deberán ser consultadas para dar solución a las interrogantes que se consignan al final, en los ejercicios propuestos.

La mayoría de los conjuntos de datos multivariantes tienen una forma común, y consisten en una matriz, en donde las filas corresponden a las observaciones (en algunos casos pueden ser las variables), y las columnas se refieren a las variables medidas. Las letras mayúsculas suelen representar a las matrices (cuadradas o rectangulares) y en este documento, los valores relacionados en estas matrices corresponderán a números reales ubicados en medio de paréntesis. Simbólicamente un conjunto de datos multivariantes puede ser representados por una matriz A, dada por la siguiente notación:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nq} \end{pmatrix}$$

Donde, n es el número de observaciones de la muestra, q es el número de variables medidas en la muestra y a_{nq} indica el valor de la variable q-ésima de la unidad n-ésima.

Table 1: Datos hipotéticos de abundancia de órdenes de insectos en diferentes localidades ubicadas en dos tipos de ecosistemas

Localidades	Ecosistema	Coleop	Dipter	Himeno	Hemip	Lepid
S1	E1	3	4	4	6	1
S2	E1	5	1	1	7	3
S3	E1	6	2	0	2	6
S4	E1	1	1	1	0	3
S5	E1	4	7	3	6	2
S6	E2	2	2	5	1	0
S7	E2	0	4	1	1	1
S8	E2	0	6	4	3	5
S9	E2	7	6	5	1	4
S10	E2	2	1	4	3	1

Las unidades en un conjunto de datos multivariantes con frecuencia son observaciones individuales, por ejemplo, cada uno de los peces en una investigación ictiológica, o de plantas en un estudio de investigación de vegetación. También pueden ser bahías, localidades, municipios o departamentos. En todos los casos las unidades son a menudo conocidas simplemente como "observaciones", un término que generalmente es usado en este texto, para representar a las filas de matrices y bases de datos.

Un ejemplo de una matriz de datos multivariados es dado en la Tabla 3. Aquí n= 10, q= 7 (filas y columnas) y por ejemplo x_{44}= 1 (elemento en la posición, fila 4 y columna 4). En esta base de datos las variables son de diferente naturaleza. Cuatro niveles de medición son usualmente distinguidos en la tipología de variables:

1. Nominal: Variables categóricas sin ordenar. Ejemplos de estas variables corresponden al tipo de ecosistema (Tabla 1: E1, E2), sexo o color de los ojos de los individuos, presencia o ausencia de depredadores, entre otros.

2. Ordinal: Se otorga un orden, sin implicar la misma distancia de los puntos en una escala prediseñada por el usuario. Se incluyen ejemplos como las capas del suelo, niveles de contaminación de un ecosistema, o niveles de temperatura en diferentes comunidades.

3. Intervalo: Las distancias entre los puntos de una escala son iguales, pero la posición del cero en la escala es arbitraria. El ejemplo clásico es la escala de temperatura en grados Celsius o Fahrenheit.

4. Razón: Se analiza la magnitud relativa de las mediciones, así como las diferencias entre estas. La posición del cero es fija. Como ejemplos se incluye a la medida absoluta de la variable temperatura en grados Celsius, la abundancia, el peso o la edad de individuos en una o varias poblaciones biológicas.

La información cualitativa de la Tabla 3, puede ser presentada en términos de códigos numéricos (requeridos por diferentes programas estadísticos), como las localidades S1, . . . , S10, como 1, . . . , 10, o el ecosistema E1 = 1 y el ecosistema E2 = 2. El análisis de datos nominales se puede limitar al resumen de resultados estadísticos como la moda. En el análisis de datos ordinales, la media y la desviación estándar no son recomendables.

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 9 > 1' in coercion to 'logical(1)'
```

3.1 Resumen estadístico de datos con múltiples variables.

El objeto de cualquier análisis multivariado consiste en iniciar con un resumen de cada una de las variables por separado, para poder entender su comportamiento general en el grupo de datos. Para este propósito se suele utilizar medidas de tendencia central y de variación (asumiendo que las variables son continuas o discretas). Adicionalmente se resume la información del conjunto de variables mediante sus correlaciones o covarianzas.

3.1.1 Medias.

Para las q variables, el vector de medias de una población es usualmente representado por $\mu^t = [\mu_1, \mu_2, ..., \mu_q]$, donde $\mu_i = E(x_i)$)

mu_i es la media de la población (o el valor esperado, denotado por el operador E) de la i-ésima variable. Un estimador de µ´, se basa en n, de las observaciones con q-dimensiones. De esta manera $\bar{x}^t = [\bar{x}_1, \bar{x}_2, ..., \bar{x}_q]$, donde $\bar{x}^t$ es la media de la variable $\bar{x}_i$. Para ilustrar el cálculo de un vector, tomaremos las columnas de la Tabla 3 (archivo: Insectos.csv), en donde se exponen las abundancias de 5 órdenes de insectos acuáticos de una muestra con 10 localidades. El vector de medias se puede extraer directamente de R, mediante la siguiente función:

```
insectos = read.csv2("Tabla1.csv",row.names=1)
medias = colMeans(insectos [,c(3:7)])
medias
```

```
## Coleop Dipter Himeno  Hemip  Lepid
##    3.0    3.4    2.8    3.0    2.6
```

3.1.2 Varianzas.

El vector de la varianza de una población puede representarse por $\sigma^t = [\sigma_1^2, \sigma_2^2, ..., \sigma_q^2]$, donde $\sigma_i^2 = E(x_i - \mu_i)^2]$

Un estimador de σ^t basado en n, de las observaciones con q-dimensiones es $s_i^2 = [s_1^2, s_2^2, ..., s_q^2]$, donde s_i^2 es la varianza muestral de x_i. El cálculo de las varianzas de la Tabla 3, es realizado en el programa R mediante la función *sd*, de la siguiente forma:

```
insectos = read.csv2("Tabla1.csv",row.names=1)
varianzas = var(insectos [,c(3:7)])
round(varianzas,2)
```

```
##        Coleop Dipter Himeno Hemip Lepid
## Coleop   6.00   0.44  -0.11  1.56  1.89
## Dipter   0.44   5.38   1.76  0.89  0.73
## Himeno  -0.11   1.76   3.51  0.00 -1.31
## Hemip    1.56   0.89   0.00  6.22 -0.44
## Lepid    1.89   0.73  -1.31 -0.44  3.82
```

Los valores resultantes de la desviación estándar son, se calculan con el comando: `sqrt(varianzas)`

3.1.3 Covarianzas.

La covarianza para dos variables: x_i y x_j, es definida por la siguiente función $\Sigma\sigma_i^2 = Cov(x_i, x_j) = (x_i - \mu_i)(x_j - \mu_j)$, cuando i = j, la covarianza de la variable es su varianza. La covarianza de x_i y x_j usualmente se denota por σ_{ij}. La matriz de covarianza es simétrica y de dimensión qxq, se denota como Σ, donde:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nq} \end{pmatrix}$$

Nótese que $\sigma_{ij} = \sigma_{ji}$. Esta matriz es conocida como varianza-covarianza o simplemente covarianza. La matriz es estimada por la matriz `S` de la siguiente manera:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_j - \bar{x})^t$$

Donde $x_i = [x_{i1}, x_{i2}, ..., x_{iq}]$ es el vector de observaciones para el `i_ésimo` objeto. La diagonal principal de S contiene a la varianza de cada variable. La matriz de covarianza se obtiene mediante el comando `var` de R, generando la matriz de varianzas (en la diagonal principal) y de covarianzas (elementos por fuera de la diagonal).

```
cov = var(insectos [,c(3:7)])
round(cov,2)
```

```
##        Coleop Dipter Himeno Hemip Lepid
## Coleop   6.00   0.44  -0.11  1.56  1.89
## Dipter   0.44   5.38   1.76  0.89  0.73
## Himeno  -0.11   1.76   3.51  0.00 -1.31
## Hemip    1.56   0.89   0.00  6.22 -0.44
## Lepid    1.89   0.73  -1.31 -0.44  3.82
```

3.1.4 Resumen estadístico.

Una forma abreviada de reportar los resultados estadísticos de una base de datos se obtiene en el programa R mediante la función summary, de la siguiente manera:

```
summary(insectos[,c(3:7)])
```

```
##      Coleop          Dipter          Himeno        Hemip           Lepid     
##  Min.   :0.00   Min.   :1.00   Min.   :0.0   Min.   :0.00   Min.   :0.00  
##  1st Qu.:1.25   1st Qu.:1.25   1st Qu.:1.0   1st Qu.:1.00   1st Qu.:1.00  
##  Median :2.50   Median :3.00   Median :3.5   Median :2.50   Median :2.50  
##  Mean   :3.00   Mean   :3.40   Mean   :2.8   Mean   :3.00   Mean   :2.60  
##  3rd Qu.:4.75   3rd Qu.:5.50   3rd Qu.:4.0   3rd Qu.:5.25   3rd Qu.:3.75  
##  Max.   :7.00   Max.   :7.00   Max.   :5.0   Max.   :7.00   Max.   :6.00  
```

Donde para cada variable, `Min`: mínimo valor, `1st Qu.`: primer cuartil, `Median`: mediana, `Mean`: media o valor promedio, `3rd Qu.`: tercer cuartil, `Max.`: valor máximo.

3.2 Correlaciones.

Es una forma de estandarizar a las covarianzas en variables que presentan unidades disimiles (ej. Variables físico-químicas). Esta estandarización se realiza dividiendo por el producto de las desviaciones estándar de dos variables, mediante un coeficiente de correlación de Pearson ρ_{ij}, donde:

$$S = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{ii}}}$$

El coeficiente de correlación ρ, presenta sus límites entre -1 y +1, dando una medida del nivel de asociación o correspondencia lineal entre dos variables x_i y x_j. Cuando la asociación es positiva, el coeficiente tiende a +1 y si la asociación es negativa se tiende a -1. Si no existe asociación entre las variables el coeficiente tiende a 0. La matriz de correlación se obtiene mediante la siguiente función:

```
round(cor(insectos[,c(3:7)]),2)
```

```
##        Coleop Dipter Himeno Hemip Lepid
## Coleop   1.00   0.08  -0.02  0.25  0.39
## Dipter   0.08   1.00   0.40  0.15  0.16
## Himeno  -0.02   0.40   1.00  0.00 -0.36
## Hemip    0.25   0.15   0.00  1.00 -0.09
## Lepid    0.39   0.16  -0.36 -0.09  1.00
```

3.3 Distancias.

Este concepto es detallado en el capítulo de clasificación de este libro, pues puede ser categorizado en múltiples tipos de distancias (`Legendre y Legendre 1998`). Una visión general demuestra que las distancias entre observaciones son atributos de mucha importancia para las técnicas multivariadas. La distancia de mayor uso es la Euclidea, la cual se deriva del Teorema de Pitágoras, en donde cada par observaciones corresponden a dos puntos del plano cartesiano y su distancia se determina mediante la ecuación de triángulos (distancia métrica que cumple con la desigualdad triangular). De esta manera la distancia euclidea para dos vectores (dos filas `i` e `j`, representadas por dos observaciones), se calcula mediante la siguiente función:

$$D = \sqrt{\sum_{i=1}^{q} (x_{ik} - x_{jk})^2}$$

Para el cálculo de la Distancia Euclídea se utiliza la función dist de la siguiente manera:

```
round(dist(insectos[,c(3:7)]),2)
```

```
##       1    2    3    4    5    6    7    8    9
## 2  5.20
## 3  8.37 6.08
## 4  7.87 8.06 6.32
## 5  3.46 6.56 8.37 9.27
## 6  5.66 8.43 8.83 5.29 7.87
## 7  6.56 8.60 8.19 3.87 7.42 5.00
## 8  6.16 8.89 8.37 6.93 6.00 7.07 5.74
## 9  7.42 9.06 6.86 8.89 6.56 7.55 8.83 7.42
## 10 4.36 6.16 7.68 4.80 7.14 2.65 5.10 6.71 8.00
```

El cálculo de la matriz de distancias euclídeas ignora las diferencias en las unidades o escalas de las variables (ej. En el caso de variables fisicoquímicas), por lo que se sugiere hacer una estandarización, previa al cálculo de las distancias. Una buena opción para estandarizar las variables consiste en utilizar la función scale de R, la cual toma cada dato y le resta su media y le divide su desviación estándar. Luego se calcula la distancia euclídea sobre los datos estandarizados, de la siguiente manera:

```
insectos.estand <- scale (insectos[,c(3:7)])
round(dist(insectos.estand),2)
```

```
##       1    2    3    4    5    6    7    8    9
## 2  2.47
## 3  3.99 2.65
## 4  3.43 3.25 2.76
## 5  1.54 2.90 3.83 3.92
## 6  2.34 3.79 4.40 2.72 3.39
## 7  2.84 3.56 3.71 1.75 3.12 2.50
## 8  2.81 3.87 3.74 3.14 2.63 3.33 2.85
## 9  3.17 3.99 3.39 3.95 2.81 3.37 3.98 3.06
## 10 1.81 2.77 3.76 2.29 3.06 1.17 2.36 3.08 3.48
```

3.4 Valores y Vectores propios.

También conocido como autovalores - λ_i (eigenvalues) y autovectores - μ_i (eigenvectors). Los valores propios de una matriz cuadrada A, son tales que para µi cumplen la siguiente propiedad:

$$A.\mu_i = \lambda_i.\mu_i$$

De esta manera se observa que los valores propios - λ_i son escalares que mediante un valor único resumen la información contenida en una matriz A, siempre y cuando cumplan con la igualdad de la función anterior. Para el cálculo de λ_i y μ_i se debe realizar el siguiente determinante, que hace parte de la ecuación característica:

$$|A.\lambda_i.I| = 0$$

Donde `I` corresponde a la matriz identidad de `A`. de esta manera se obtiene la solución a la ecuación característica de autovalores y autovectores de `A`. revisemos el siguiente ejemplo, en donde tenemos una matriz cuadrada `A` de dimensión 2 x 2.

$$A = \begin{pmatrix} 1 & 2 \\ -1 & 4 \end{pmatrix}$$

El cálculo de λ_i y μ_i se realiza en R mediante la función eigen, luego se hace una descomposición mediante los siguientes comandos:

```
A <- matrix(c(1,2,-1,4),2,2,byrow=T)
A
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]   -1    4
```

Con el comando `eigen` se obtienen los `autovectores` (dos vectores) y `autovalores` (dos escalares):

```
eigen(A)
```

```
## eigen() decomposition
## $values
## [1] 3 2
##
## $vectors
##            [,1]       [,2]
## [1,] -0.7071068 -0.8944272
## [2,] -0.7071068 -0.4472136
```

Para probar la igualdad ($A.\mu_i = \lambda_i.\mu_i$) primero se extraen los autovalores y los autovectores mediante las dos funciones siguientes:

```
autovalor <- eigen(A)$values[1]
autovector <- eigen(A)$vectors[,1]
```

Posteriormente se prueba la igualdad `(==)` con la siguiente operación:

```
A %*% autovector == autovalor * autovector
```

```
##      [,1]
## [1,] TRUE
## [2,] TRUE
```

R reporta que la igualdad es verdadera, similar a lo realizado con la matriz `A` de 2 x 2, se puede obtener los valores y vectores propios de bases de datos más complejas siempre y cuando sean cuadradas. En el caso de la matriz de distancias euclídeas `(dist)` obtenidas de la Tabla 1, esta operación puede ser realizada, debido a que es una matriz cuadrada y simétrica.

Ejemplo 1. MATRICES EN RSTUDIO

En el presente ejemplo, se realizarán algunas operaciones matriciales básicas y de uso común en el análisis de datos ecológicos y ambientales.

```
# Ejercicio 1
# A (2,1,1,3)
# B (1,4,2,5,0,3)
# Calcular: (1) B'.A' (2) (A.B)' (3) Demostrar que B'.A' = (A.B)'

# R./
A = matrix(c(2,1,1,3),2,2,byrow=T)      # Estructura de comandos para las matrices
A                        # 2,2: dos filas y dos columnas.
```

```
##      [,1] [,2]
## [1,]    2    1
## [2,]    1    3
```

```
B = matrix(c(1,4,2,5,0,3),2,3,byrow=T)  # byrow: ordenar valores por filas de la matriz.
B                        # 2,3: dos filas y tres columnas
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    2
## [2,]    5    0    3
```

```
tA= t(A)        # t(A): Transpuesta de la matriz A
tA
```

```
##      [,1] [,2]
## [1,]    2    1
## [2,]    1    3
```

```
tB= t(B)        # t(B): Transpuesta de la matriz B
tB
```

```
##      [,1] [,2]
## [1,]    1    5
## [2,]    4    0
## [3,]    2    3
```

```
tB%*%tA == t(A%*%B)    # %*%: para multiplicar matrices.
```

```
##      [,1] [,2]
## [1,] TRUE TRUE
## [2,] TRUE TRUE
## [3,] TRUE TRUE
```

```r
                        #==: demostrar una igualdad.

#-------------
# Ejercicio 2
# A (2,3,3,2)
# B (1,4,2,5,0,3)
# Calcular: Determinante de A y de B

#R./
A = matrix(c(2,3,3,2),2,2,byrow=T)        # Matriz A
A
```

```
##      [,1] [,2]
## [1,]    2    3
## [2,]    3    2
```

```r
B = matrix(c(1,4,2,5,0,3),3,3,byrow=T)  # Matriz B
```

```
## Warning in matrix(c(1, 4, 2, 5, 0, 3), 3, 3, byrow = T): data length differs
## from size of matrix: [6 != 3 x 3]
```

```r
B
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    2
## [2,]    5    0    3
## [3,]    1    4    2
```

```r
det(A)      # det(A): Determinante de la matriz A
```

```
## [1] -5
```

```r
det(B)      # det(b): Determinante de la matriz B
```

```
## [1] 0
```

```r
#------------
# Ejercicio 3
# A (5,2,2,2)
# Calcular inversa de A

# R./
A = matrix(c(5,2,2,2),2,2,byrow=T)        # Matriz A
A
```

```
##      [,1] [,2]
## [1,]    5    2
## [2,]    2    2
```

```r
solve(A)        # solve(A): Determinante de la matriz A
```

```
##            [,1]       [,2]
## [1,]  0.3333333 -0.3333333
## [2,] -0.3333333  0.8333333
```

```r
#------------
# Ejercicio 4
# A (5,2,2,2)
# Calcular valores y vectores propios de A

# R./
A = matrix(c(5,2,2,2),2,2,byrow=T) # Matriz A
A
```

```
##      [,1] [,2]
## [1,]    5    2
## [2,]    2    2
```

```r
eigen(A)    # Valores y vectores propios
```

```
## eigen() decomposition
## $values
## [1] 6 1
##
## $vectors
##            [,1]       [,2]
## [1,] -0.8944272  0.4472136
## [2,] -0.4472136 -0.8944272
```

```r
# Ejercicios:
# 1. Extraer los valores y vectores propios por separado.
# 2. Extraer el primer vector fila y el primer vector columna, por separado del resto de insumos.


# R./ ***
eigen(A)$vectors
```

```
##            [,1]       [,2]
## [1,] -0.8944272  0.4472136
## [2,] -0.4472136 -0.8944272
```

```r
eigen(A)$vectors[,1]
```

```
## [1] -0.8944272 -0.4472136
```

```r
eigen(A)$vectors[1,]
```

```
## [1] -0.8944272  0.4472136
```

```r
#------------
# Ejercicio 5
# Fila (observación) 1: 5,2
# Fila (observación) 2: 2,2
# Calcular la matriz de distancia euclidea entre las filas

# R./
A = matrix(c(5,2,2,2),2,2,byrow=T)  # Matriz A
A
```

```
##      [,1] [,2]
## [1,]    5    2
## [2,]    2    2
```

```r
dist(A)   # dist(A): Matriz de distancia euclídea
```

```
##   1
## 2 3
```

```r
#------------
# Ejercicio 6
# Fila (observación) 1: 5,2
# Fila (observación) 2: 2,2
# Fila (observación) 3: 2,5
# Calcular la matriz de distancia euclídea entre las filas

# R./
A = matrix(c(5,2,2,2,2,5),3,2,byrow=T)   # Matriz A
A
```

```
##      [,1] [,2]
## [1,]    5    2
## [2,]    2    2
## [3,]    2    5
```

```r
dist(A)          # dist(A): Matriz de distancia euclídea
```

```
##          1        2
## 2 3.000000
## 3 4.242641 3.000000
```

```r
round(dist(A),1)  # dound(): Disminuye decimales del resultado
```

```
##     1   2
## 2 3.0
## 3 4.2 3.0
```

```r
#------------
# Ejercicio 7
# Variable 1: 2,1,2
```

```r
# Variable 2: 4,3,5
# Calcular las matrices de Covarianza y de Correlación

# R./
A = matrix(c(2,1,2,4,3,5),3,2,byrow=F)
A
```

```
##      [,1] [,2]
## [1,]    2    4
## [2,]    1    3
## [3,]    2    5
```

```r
cov(A)
```

```
##           [,1] [,2]
## [1,] 0.3333333  0.5
## [2,] 0.5000000  1.0
```

```r
var(A)
```

```
##           [,1] [,2]
## [1,] 0.3333333  0.5
## [2,] 0.5000000  1.0
```

```r
cor(A)
```

```
##           [,1]      [,2]
## [1,] 1.0000000 0.8660254
## [2,] 0.8660254 1.0000000
```

```r
round(cov(A),1)
```

```
##      [,1] [,2]
## [1,]  0.3  0.5
## [2,]  0.5  1.0
```

```r
round(cor(A),1)
```

```
##      [,1] [,2]
## [1,]  1.0  0.9
## [2,]  0.9  1.0
```

```r
# Calcular la matriz de distancia y analizar el resultado.
dist(A)
```

```
##          1        2
## 2 1.414214
## 3 1.000000 2.236068
```

```r
round(dist(A),1)
```

```
##     1   2
## 2 1.4    
## 3 1.0 2.2
```

3.5 Ejercicios propuestos

1. Descargar la base de datos de lirios `data(iris)`, de forma resumida se requiere explicar a los elementos de esta base, apoyado del comando `help(iris)` y Realizar los procedimientos descritos en los ejemplos 2 y 3, para las especies setosa y versicolor.

2. Descargar la base de datos ambiental `data(airquality)`, de forma resumida se requiere explicar a los elementos de esta base, apoyado del comando help(airquality) y Realizar los procedimientos descritos en los ejemplos 2 y 3, para los meses 5 y 6.

3. Para el siguiente procedimiento, se requiere descargar las siguientes librerías o paquetes: `gcookbook` y `ggplot2`. Posteriormente se debe descargar la base de datos de estudiantes `data(heightweight)`, de forma resumida se requiere explicar a los elementos de esta base, apoyado del comando `help(heightweight)` y Realizar los procedimientos descritos en los ejemplos 2 y 3.

Nota: para cada procedimiento se debe realizar el respectivo análisis biológico o ambiental. Las respuestas a los tres ejercicios no deben superar 9 páginas.