INTRODUCCIÓN A LA ASIGNATURA

Estadística Multivariada

GIEN
GRUPO DE INVESTIGACIÓN EN ECOLOGÍA NEOTROPICAL
UNIMAGDALENA

https://grupogien.jimdofree.com/

Docente: Javier Rodríguez Barrios
jrodriguez@unimagdalena.edu.co

PROGRAMACIÓN

1. Introducción a la estadística multivariada.
2. Algebra lineal aplicada a datos multivariados
3. Análisis exploratorio de datos multivariados
4. Análisis de Componentes Principales – PCA
5. Escalamiento Multidimensional no Métrico – nMDS
6. Análisis de Correspondencias – CA, DCA, CCA y RDA
7. Distancias y Coeficientes de Similitud
8. Análisis de Clúster – CLA
9. Análisis Discriminante – DA
10. Introducción a las pruebas de hipótesis multivariadas
    - Paramétricas (T^2, MANOVA)
    - No paramétricas (npMANOVA, perMANOVA)

INTRODUCCIÓN

La mayoría de datos multivariantes consisten en una matriz de datos, las filas que corresponden a observaciones, y las columnas se relacionan a las variables medidas.

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

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

A continuación se describen algunas formas matriciales de uso común en la estadística Multivariada

Tipos de Matrices

- Cuadradas. $A_{n \times n}$
- Diagonal.
- Triangular.
- Identidad. I
- Transpuesta. $A'_{m \times n'}, A^T$
- Simétrica. $S = S'$
- Determinante. $|A|$ o $det(A)$
- Inversa. A^{-1}
- Autovalores. μ_i
- Autovectores. λ_i
- Correlación. δ
- Covarianza. Σ
- Distancia. Di

Operaciones con Matrices

Producto de matrices:
Si denominados a las n filas de la matriz $\boldsymbol{A}_{nm}$ como: $\boldsymbol{a_1}, \ldots, \boldsymbol{a_n}$ y a las columnas de $\boldsymbol{B}_{mp}$ como $\boldsymbol{b_1}, \ldots, \boldsymbol{b_p}$

$$\boldsymbol{AB} = (\boldsymbol{AB})_{np} = \begin{bmatrix} \boldsymbol{a_1} \cdot \boldsymbol{b_1} & \ldots & \boldsymbol{a_1} \cdot \boldsymbol{b_p} \\ \vdots & \vdots & \vdots \\ \boldsymbol{a_n} \cdot \boldsymbol{b_1} & \ldots & \boldsymbol{a_n} \cdot \boldsymbol{b_p} \end{bmatrix}$$

Determinante de una matrix 2×2

$$|\boldsymbol{B}| = \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} = b_{11}b_{22} - b_{12}b_{21}$$

Inversa de una Matriz

Si a la matriz cuadrada A_{nn} se le puede hallar otra matriz B_{nn} tal que $AB = BA = I$ entonces a B se le denomina la inversa de A y se denota por A^{-1}

Uso de la inversa para despejar una ecuación matricial:

$$A_{nn}x_n = b_n$$

$$Ax = b$$

Se despeja x como:

$$x = A^{-1}b$$

Valores y Vectores Propios

Los valores propios λ_i de una matriz A son tales que para u_i cumplen la siguiente propiedad:

$$\boldsymbol{Au_i = \lambda_i u_i}$$

Donde λ_i son los valores propios y u_i son los vectores propios.

Ecuación característica:

$$|A - \lambda i I| = 0$$

La solución a esta ecuación característica para λ_i son los valores propios.

Estadísticos multivariados

Covarianza (S_{ik}): relación entre dos variables. En términos absolutos

$$S_{ik} = \mathrm{cov}(i,k) = \frac{1}{n}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)(X_{kj} - \bar{X}_k)$$

$$S_{ii} = \mathrm{cov}(i,i) = S_i^2$$

La matriz de varianza – covarianza $\boldsymbol{\Sigma}$ resalta variaciones absolutas y lineales entre variables de igual unidad (ej. abundancia de especies).

Una covarianza (S_{ik}) de cero (0) indica que no existe relación lineal entre dos variables, pero puede haber relación de otro tipo.

Estadísticos multivariados

Coeficiente de Correlación (***r***): asociación entre dos variables. Estandarización de los datos.

$$r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}} = \sum \frac{(X_{ij} - \bar{X}_i)(X_{kj} - \bar{X}_k)}{\sqrt{S_{ii}}\sqrt{S_{kk}}}$$

j=1, …, n
i=1, …, p

La ***estandarización*** vuelve la matriz (***cov***) en un rango de (-1 0 1).

Si ***r***= 0, entonces ***cov***= 0 no hay correlación ni ***cov*** Lineal entre las variables (falta ajuste lineal).

Valores y Vectores Propios

Fuimos a campo y reportamos los siguientes datos:

	Sp 1	Sp 2	Sp 3
Sitio 1	5	4	1
Sitio 2	6	1	4
Sitio 3	9	2	10

```
> A<-matrix(c(5,4,1,6,1,4,9,2,10),3,3,byrow=T)
> A
     [,1] [,2] [,3]
[1,]    5    4    1
[2,]    6    1    4
[3,]    9    2   10
```

```
> eigen(A)
$values
[1] 13.830860  3.763631 -1.594490

$vectors
           [,1]        [,2]       [,3]
[1,] -0.2811639 -0.56449641  0.5288879
[2,] -0.4029939 -0.03172489 -0.8039722
[3,] -0.8709436  0.82482564 -0.2718573
```

$$Au_i = \lambda_i u_i$$

Probar esta igualdad en R

Valores y Vectores Propios

2.19 Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & -2 \\ -1 & 2 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$

(a) Find the eigenvalues and associated normalized eigenvectors.

```
> A<-matrix(c(1,1,-2,-1,2,1,0,1,-1),3,3,byrow=T)
> eigen(A)
```

```
$values
  [1]    2   1   -1

$vectors
           [,1]      [,2]      [,3]
 [1,]     0,302    -0,802   7,1E+05
 [2,]     0,905    -0,535   4,6E-11
 [3,]     0,302    -0,267   7,1E+05
```

$$Au_i = \lambda_i u_i$$

Probar esta igualdad en R

Estadísticos multivariados

Si a esos datos queremos medirle la **relación entre especies** que hacemos?

	Sp 1	Sp 2	Sp 3
Sitio 1	5	5	1
Sitio 2	6	6	4
Sitio 3	9	8	1

```
> A<-matrix(c(5,5,1,6,6,4,9,8,1),3,3,byrow=T)
> A
     [,1] [,2] [,3]
[1,]    5    5    1
[2,]    6    6    4
[3,]    9    8    1
```

```
> cov(A)
          [,1]      [,2] [,3]
[1,]  4.333333  3.166667 -1.0
[2,]  3.166667  2.333333 -0.5
[3,] -1.000000 -0.500000  3.0
```

```
> cor(A)
           [,1]       [,2]       [,3]
[1,]  1.0000000  0.9958706 -0.2773501
[2,]  0.9958706  1.0000000 -0.1889822
[3,] -0.2773501 -0.1889822  1.0000000
```

Estadísticos multivariados

Si intentamos medir el nivel de **relación entre sitios** que hacemos?

	Sp 1	Sp 2	Sp 3
Sitio 1	5	5	1
Sitio 2	6	6	4
Sitio 3	9	8	1

```
> t(A)
     [,1] [,2] [,3]
[1,]    5    6    9
[2,]    5    6    8
[3,]    1    4    1
```

```
> cov(t(A))
          [,1]     [,2] [,3]
[1,]  5.333333 2.666667   10
[2,]  2.666667 1.333333    5
[3,] 10.000000 5.000000   19
```

```
> cor(t(A))
          [,1]      [,2]      [,3]
[1,] 1.0000000 1.0000000 0.9933993
[2,] 1.0000000 1.0000000 0.9933993
[3,] 0.9933993 0.9933993 1.0000000
```

Estadísticos multivariados

Si a esos **sitios** les medimos **variables fisicoquímicas**?

	pH	DOC mg$C.m^{-3}$	PO4 µ$MOL.L^{-1}$
Sitio 1	5	200	7
Sitio 2	6	210	7
Sitio 3	3	300	17

	Sitio 1	Sitio 2	Sitio 3
pH	5	6	3
DOC	200	210	300
PO4	7	7	17

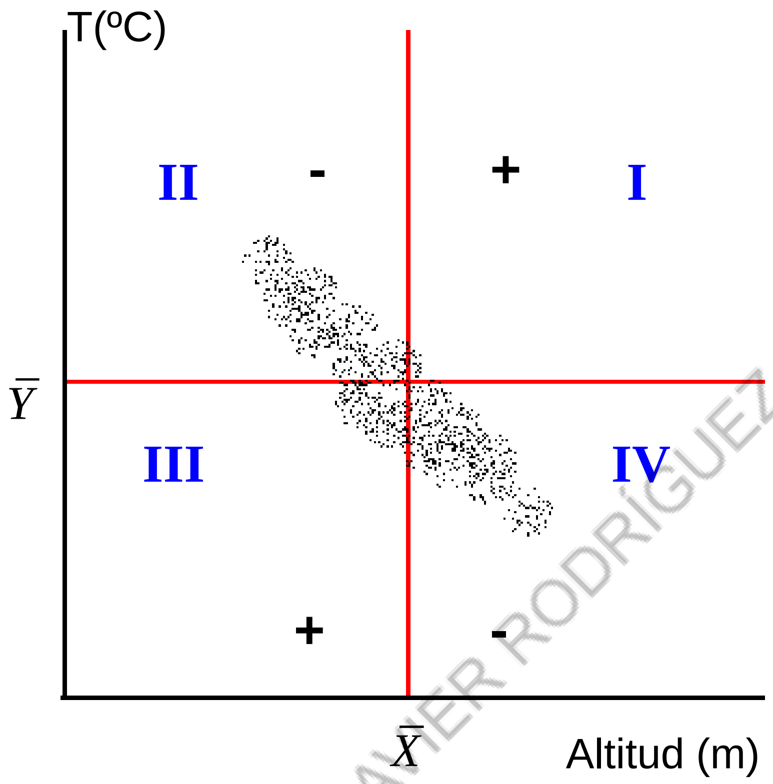
Relación entre variables

```
> cov(A)
            [,1]       [,2]       [,3]
[1,]   2.333333  -76.66667  -8.333333
[2,] -76.666667 3033.33333 316.666667
[3,]  -8.333333  316.66667  33.333333

> cor(A)
            [,1]       [,2]       [,3]
[1,]  1.0000000 -0.9112932 -0.9449112
[2,] -0.9112932  1.0000000  0.9958706
[3,] -0.9449112  0.9958706  1.0000000
```

Relación entre sitios

```
> cov(t(A))
         [,1]     [,2]     [,3]
[1,] 12546.33 13160.17 18760.33
[2,] 13160.17 13804.33 19675.17
[3,] 18760.33 19675.17 28082.33

> cor(t(A))
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.999989 0.9994603
[2,] 0.9999891 1.000000 0.9992959
[3,] 0.9994603 0.999296 1.0000000
```

Estadísticos multivariados



I, **X** es (+) y **Y** (+)

II, **X** es (-) y **Y** (+)

III, **X** es ( ) y **Y** ( )

III, **X** es ( ) y **Y** ( )

r (correlación) y ***S*** (cov) puede ser (+) o (-).

Estadísticos multivariados