Correlaciones Canónicas

1

Componentes principales

Datos originales

vector de variables

$$\mathbb{X} = (X_1, ..., X_p)^t$$

matriz de datos:

$$X = [\ \mathbb{X}_1 \mid ..... \mid \mathbb{X}_p\ ]$$

$$X = \begin{bmatrix} x_{11} & x_{12} & ..... & x_{1p} \\ x_{21} & x_{22} & ..... & x_{2p} \\ ..... & ..... & ..... & ..... \\ ..... & ..... & ..... & ..... \\ x_{n1} & x_{12} & ..... & x_{np} \end{bmatrix}$$

$$\Sigma_{\mathbb{X}}$$

Componentes principales

vector de variables

$$\mathbb{Y} = (Y_1, ..., Y_q)^t$$

matriz de datos:

$$Y = [\ \mathbb{Y}_1 \mid ..... \mid \mathbb{Y}_p\ ]$$

$$Y = \begin{bmatrix} y_{11} & y_{12} & ..... & y_{1q} \\ y_{21} & y_{22} & ..... & y_{2q} \\ ..... & ..... & ..... & ..... \\ ..... & ..... & ..... & ..... \\ y_{n1} & y_{12} & ..... & y_{nq} \end{bmatrix}$$

$$\Sigma_{\mathbb{Y}} = diag\{\lambda_1, ..., \lambda_q\}$$

2

Correlaciones canónicas

Grupo de variables 1

vector de variables

$$\mathbb{X} = (X_1, ..., X_p)^t$$

matriz de datos:

$$X = [\ \mathbb{X}_1 \ | \ ..... \ | \ \mathbb{X}_p\ ]$$

$$X = \begin{bmatrix} x_{11} & x_{12} & ..... & x_{1p} \\ x_{21} & x_{22} & ..... & x_{2p} \\ ..... & ..... & ..... & ..... \\ ..... & ..... & ..... & ..... \\ x_{n1} & x_{12} & ..... & x_{np} \end{bmatrix}$$

$$\text{Indice: } V = b^t\mathbb{Y}$$

$$\text{Indice: } U = a^t\mathbb{X}$$

Criterio:

$$\max_{a,b} cor(U, V)$$

Grupo de variables 2

vector de variables

$$\mathbb{Y} = (Y_1, ..., Y_q)^t$$

matriz de datos:

$$Y = [\ \mathbb{Y}_1 \ | \ ..... \ | \ \mathbb{Y}_p\ ]$$

$$Y = \begin{bmatrix} y_{11} & y_{12} & ..... & y_{1q} \\ y_{21} & y_{22} & ..... & y_{2q} \\ ..... & ..... & ..... & ..... \\ ..... & ..... & ..... & ..... \\ y_{n1} & y_{12} & ..... & y_{nq} \end{bmatrix}$$

3

Terminología

Variables canónicas:

$$U = a^t\mathbb{X}$$

$$V = b^t\mathbb{Y}$$

Vectores con los coeficientes de las variables canonicas:

a, b

Correlación canónica

$$\rho = cor(U, V)$$

4

Determinación de las correlaciones canónicas''

5

El problema

Vectores aleatorios (grupos de variables)

$X_{p,1}$, $Y_{q,1}$

Coeficientes

$a \in \mathbb{R}^p$ $b \in \mathbb{R}^q$

Hallar a y b

$$\underset{a,b}{\arg\max}\ \text{cor}(a^t X, b^t Y)$$

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \right)$$

$$\text{Cor}(U,V) = \frac{a^t \Sigma_{XY} b}{(a^t \Sigma_X a)^{1/2} (b^t \Sigma_Y b)^{1/2}}$$

$U = a^t X$

$V = b^t Y$

6

El problema

Vectores aleatorios (grupos de variables)

$X_{p,1}$, $Y_{q,1}$

Coeficientes

$a \in \mathbb{R}^p \quad b \in \mathbb{R}^q$

Hallar a y b

$$\underset{a,b}{\arg\max}\ cor(a^t X, b^t Y)$$

$$U = a^T X \qquad V = b^T Y$$

$$Cor(U,V) = \frac{a^T \Sigma_{XY} b}{(a^T \Sigma_X a)^{1/2} (b^t \Sigma_Y b)^{1/2}}$$

Restricción (extremo condicionado:

$Cor(cU,V) = cor(U,cV) = cor(U,V)$

La correlación es invariante ante cambios de escala, así que el problema es el mismo si se exige:

$$var(U) = a^t \Sigma_{\mathbb{X}} a = 1$$
$$var(V) = b^t \Sigma_{\mathbb{Y}} b = 1$$

7

$$Cor(U,V) = \frac{a^T \Sigma_{XY} b}{(a^T \Sigma_X a)^{1/2} (b^t \Sigma_Y b)^{1/2}} = a^T \Sigma' b$$

Bajo la condición

$$a^T \Sigma_X a = 1$$

$$b^t \Sigma_Y b = 1$$

8

El problema

Vectores aleatorios (grupos de variables)

$X_{p,1}$, $Y_{q,1}$

Coeficientes

$a \in \mathbb{R}^p \quad b \in \mathbb{R}^q$

Hallar a y b

$$\underset{a,b}{\arg\max}\ \text{cor}(a^t X, b^t Y)$$

$U = a^t X \qquad V = b^t Y$

Hallar a y b tales que

$\text{Cor}(a^t X, b^t Y)$

Sea máxima

Condición

$a^t \Sigma_X a = 1$

$b^t \Sigma_Y b = 1$

9

El problema

$U = a^t X \qquad V = b^t Y$

Hallar a y b tales que

$\text{Cor}(a^t X, b^t Y)$

Sea máxima

Condición

$a^t \Sigma_X a = 1$

$b^t \Sigma_Y b = 1$

La solución

Identificar diferentes parejas de índice (variables canónicas)

(U_i , V_i)

Tales que sus correlaciones (correlaciones canónicas) sean (secuencialmente) máximas

$\rho_i = \text{cor}(U_i , V_i)$

10

Punto de partida (algebraico)

$$\begin{matrix} A\ A^t \\ A^t\ A \end{matrix}$$

Tienen los mismos valores propios

$\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_k$

Considerar la matriz

$$A = \Sigma_{xx}^{-1/2}\ \Sigma_{xy}\ \Sigma_{yy}^{-1/2}$$

$A\ A^t$ vectores propios $\gamma_1, \gamma_2, ..., \gamma_k$

$\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_k$

$A^t\ A$ vectores propios $\delta_1, \delta_2, ..., \delta_k$

Vectores propios ortonormales

11

Valores propios: $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_k$
$A\ A^t$ vectores propios $\gamma_1, \gamma_2, ..., \gamma_k$
$A^t\ A$ vectores propios $\delta_1, \delta_2, ..., \delta_k$

1er. par de variables canónicas	$u_1 = a_1^T x$ $\quad v_1 = b_1^T y$ $\quad Var(u_1) = Var(v_1) = 1$ $\rho_1 = cor(u_1, v_1) = \sqrt{\lambda_1}$ $\quad a_1 = \Sigma_{xx}^{-1/2}\delta_1$ $\quad b_1 = \Sigma_{yy}^{-1/2}\delta_1$
2do. par de variables canónicas	$u_2 = a_2^T x$ $\quad v_2 = b_2^T y$ $\quad Var(u_2) = Var(v_2) = 1$ $\rho_2 = cov(u_2, v_2) = \sqrt{\lambda_2}$ $\quad a_2 = \Sigma_{xx}^{-1/2}\delta_2$ $b_2 = \Sigma_{yy}^{-1/2}\delta_2$ $cov(u_2, u_1) = 0 = cov(v_2, v_1)$
.....	

12

Una propiedad muy importante:
Las correlaciones canónicas son invariantes ante traslaciones o cambios de escala, pero las variables canónicas no

Por eso es conveniente trabajar los problemas con variables estandarizadas

13

Para las aplicaciones es importante determinar con cuántas variables canónicas se va a trabajar.

Prueba de hipótesis para ver cuáles (a partir de cuáles) las correlaciones canónicas son 0

14

Prueba de Wilks (normalidad)

$$H_0 : \rho_1 = \rho_2 = ... = \rho_k = 0$$

$$H_0 : \rho_2 = \rho_3 = ... = \rho_k = 0$$

$$H_0 : \rho_3 = \rho_4 = ... = \rho_k = 0$$

Hay diferentes aproximaciones para la prueba basadas en el estadístico de Wilks

$$\Lambda = \prod_{i=s}^{k} (1 - \lambda_i)$$

valores propios de $A\ A^t$

$\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_k$

15

Un ejemplo

16

Ejemplo: Una investigadora está interesada en saber cómo un grupo de variables psicológicas se relaciona con el rendimiento académico de los estudiantes de 1er año de una universidad
En particular está interesada en saber cuántas dimensiones (variables canónicas) son necesarias para entender la asociación entre los dos conjuntos de variables

Variables psicológicas (X):
control
conceptos
motivacion

Variables académicas (Y)
lectura
escritura
mat: resultados en matemáticas
ciencia: resultados en ciencias

Datos
Archivo académico
Instrucciones
instrucciones CC

17

Paso 1
Preparación de los datos

```
#paquetes a utilizar
library(CCA)
library(CCP)

#Paso 1 estandarizar las varibles
ACE=round(scale(AC),2)
```

```
> head(AC)
# A tibble: 6 × 7
  control conceptos motivacion lectura escritura   mat ciencia
    <dbl>     <dbl>      <dbl>   <dbl>     <dbl> <dbl>   <dbl>
1   -0.84     -0.24          1    54.8      64.5  44.5    52.6
2   -0.38     -0.47       0.67    62.7      43.7  44.7    52.6
3    0.89      0.59       0.67    60.6      56.7  70.5    58
4    0.71      0.28       0.67    62.7      56.7  54.7    58
5   -0.64      0.03          1    41.6      46.3  38.4    36.3
6    1.11       0.9       0.33    62.7      64.5  61.4    58
> head(ACE)
     control conceptos motivacion lectura escritura   mat ciencia
[1,]   -1.40     -0.35       0.99     0.29      1.25 -0.78    0.09
[2,]   -0.71     -0.67       0.03     1.07     -0.89 -0.76    0.09
[3,]    1.18      0.83       0.03     0.86      0.44  1.98    0.64
[4,]    0.92      0.39       0.03     1.07      0.44  0.30    0.64
[5,]   -1.10      0.04       0.99    -1.02     -0.63 -1.43   -1.59
[6,]    1.51      1.27      -0.97     1.07      1.25  1.01    0.64
```

18

Paso 2
Crear bases de datos para cada grupo

```
> ACE_X=ACE[ , 1:3]
> ACE_Y=ACE[ , 4:7]
> head(ACE_X,2)
     control conceptos motivacion
[1,]   -1.40     -0.35       0.99
[2,]   -0.71     -0.67       0.03
> head(ACE_Y,2)
     lectura escritura   mat ciencia
[1,]    0.29      1.25 -0.78    0.09
[2,]    1.07     -0.89 -0.76    0.09
```

19

Paso 3
Obtener las correlaciones canónicas

```
> #Paso 3 crear un objeto con los resultados
> #del calculo
> cc_acad=cc(ACE_X,ACE_Y)
>
> #visualizacion de las correlaciones canonicas
> cc_acad$cor
[1] 0.44643583 0.15306011 0.02226664
>
> #visualización de los coeficientes
> #de las variables canonicas
> #$xcoef coeficientes del primer grupo de variables
> #por eso se les llamo ACE_X
> cc_acad$xcoef
                [,1]       [,2]       [,3]
control    -0.8380991  0.5138814  0.3327556
conceptos   0.1676174  0.5940443 -0.8500377
motivacion -0.4276467 -0.9019239 -0.3744708
>
> #$ycoef coeficientes del segundo grupo de variables
> #por eso se les llamo ACE_Y
> cc_acad$ycoef
                [,1]        [,2]       [,3]
lectura    -0.44392139 -0.01667113 -0.9183301
escritura  -0.53679857 -0.87808337  0.9352391
mat        -0.18253658 -0.02843703 -0.7997235
ciencia     0.03694058  1.20683747  0.8597150
```

20

1er par de variables canónicas

U1 = -0.83 control + 0.16 conceptos -0.42 motivacion

V1 = -0.44 lectura – 0.53 escritura -0.18 mat -0.03 ciencia

Correlación canónica: 0.44

```
> cc_acad$cor
[1] 0.44643583 0.15306
>
> #visualización de lo
> #de las variables ca
> #$xcoef coeficientes
> #por eso se les llam
> cc_acad$xcoef
                  [,1]
control     -0.8380991
conceptos    0.1676174
motivacion -0.4276467
>
> #$ycoef coeficientes
> #por eso se les llam
> cc_acad$ycoef
                  [,1]
lectura    -0.44392139
escritura -0.53679857
mat        -0.18253658
ciencia     0.03694058
```

¿Cómo se interpretan los coeficientes?

Ejercicio construya los otros dos pares de variables canónicas

21

```
> #visualizacion de las correlaciones canonicas
> cc_acad$cor
[1] 0.44643583 0.15306011 0.02226664
```

```
> #Paso 4: Prueba de Wilks
> #depositar en un objeto las correlaciones
> r=cc_acad$cor
> #registrar el número de datos (n),
> n=dim(ACE_X)[1]
> n
[1] 600
> #el numero de datos en
> #el grupo X con variables psicologicas (p)
> p=dim(ACE_X)[2]
> p
[1] 3
> #el grupo Y con las variables academicas (q)
> q=dim(ACE_Y)[2]
> q
[1] 4
>
> #la prueba
> p.asym(r,n,p,q,tstat="Wilks")
Wilks' Lambda, using F-approximation (Rao's F):
              stat     approx df1      df2    p.value
1 to 3:  0.7815492 12.7678367  12 1569.222 0.00000000
2 to 3:  0.9760884  2.4105646   6 1188.000 0.02547141
3 to 3:  0.9995042  0.1475746   2  595.000 0.86282962
```

De las correlaciones canónicas veamos cuáles pueden ser consideradas diferentes de 0 y pueden ser utilizadas en el análisis.

Prueba de Wilks.

Error 0.01

22

Interpretación de las dimensiones canónicas (variables canónicas)

Calculemos las correlaciones de la variable canónica (de cada dimensión) con las variables que la constituyen

23

```
#valores de los individuos en las
#variables canonicas (dimensiones)

#variables para el grupo X (psicologia)
#variables canonicas 1
U1=cc_acad$scores$xscores[,1]
```
(valores U1)

Variable canónica 1: U1

Variables que la constituyen

Variables psicológicas (X):
control
conceptos
motivacion

24

```
> #paquete
> library(corrplot)
> #correlaciones
> #variables X
> #correlaciones
> cor(U1,ACE_X)
        control   conceptos motivacion
[1,] -0.9141854 -0.09948917 -0.5853833
>
> #inferencias
> D_X=data.frame(U1,ACE_X)
> I_X=cor.mtest(D_X,conf.level=0.99)
> #valor-p
> round(I_X$p,4)
             U1 control conceptos motivacion
U1         0.0000       0    0.0148          0
control    0.0000       0    0.0000          0
conceptos  0.0148       0    0.0000          0
motivacion 0.0000       0    0.0000          0
> #IC, limite inferior
> round(I_X$lowCI,2)
             U1 control conceptos motivacion
U1         1.00   -0.93     -0.20      -0.65
control   -0.93    1.00      0.07       0.14
conceptos -0.20    0.07      1.00       0.19
motivacion -0.65   0.14      0.19       1.00
> #IC, limite inferior
> round(I_X$uppCI,2)
             U1 control conceptos motivacion
U1         1.00   -0.90      0.01      -0.51
control   -0.90    1.00      0.27       0.34
conceptos  0.01    0.27      1.00       0.38
motivacion -0.51   0.34      0.38       1.00
```

```
> #variables para el grupo Y(academicas)
> #variables canonicas 1
> V1=cc_acad$scores$yscores[,1] (valores V1)
```

Variable canónica 1: V1
Variables que la constituyen
Variables académicas (Y)
lectura
escritura
mat: resultados en matemáticas
ciencia: resultados en ciencias

```
> #variables Y
> cor(V1,ACE_Y)
        lectura  escritura        mat    ciencia
[1,] -0.8800526 -0.9105097 -0.7998427 -0.6938713
> D_Y=data.frame(V1,ACE_Y)
> I_Y=cor.mtest(D_Y,conf.level=0.99)
> #valor-p
> round(I_Y$p,4)
          V1 lectura escritura mat ciencia
V1         0       0         0   0       0
lectura    0       0         0   0       0
escritura  0       0         0   0       0
mat        0       0         0   0       0
ciencia    0       0         0   0       0
> #IC, limite inferior
> round(I_Y$lowCI,2)
            V1 lectura escritura   mat ciencia
V1         1.00   -0.90     -0.93 -0.83   -0.74
lectura   -0.90    1.00      0.56  0.62    0.63
escritura -0.93    0.56      1.00  0.57    0.49
mat       -0.83    0.62      0.57  1.00    0.58
ciencia   -0.74    0.63      0.49  0.58    1.00
> #IC, limite inferior
> round(I_Y$uppCI,2)
            V1 lectura escritura   mat ciencia
V1         1.00   -0.85     -0.89 -0.76   -0.64
lectura   -0.85    1.00      0.69  0.73    0.74
escritura -0.89    0.69      1.00  0.69    0.64
mat       -0.76    0.73      0.69  1.00    0.71
ciencia   -0.64    0.74      0.64  0.71    1.00
```

25

En resumen

26

Variable canónica: U1

Variable Original	coeficientes	correlación	valor-p	IC (99%)	
				LI	LS
control	-0.83	-0.91	0.0000	-0.93	-0.9
conceptos	0.16	-0.09	0.0148	-0.2	0.01
motivacion	-0.42	-0.58	0.0000	-0.65	-0.51

Variable canónica: V1

Variable Original	coeficientes	correlación	valor-p	IC (99%)	
				LI	LS
lectura	-0.44	-0.88	0.0000	-0.90	-0.85
escritura	-0.53	-0.91	0.0000	1.00	-0.89
mat	-0.18	-0.79	0.0000	-0.93	-0.76
ciencias	0.03	-0.69	0.0000	-0.74	-0.64

Correlación canónica: 0.44

27

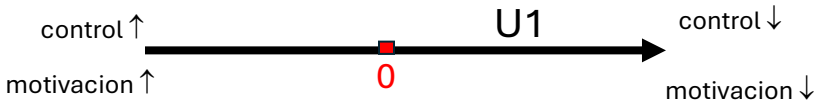
Interpretación

Al igual que en las componentes principales caractericemos las variables canónicas (indicadores)

28

Variable canónica: U1

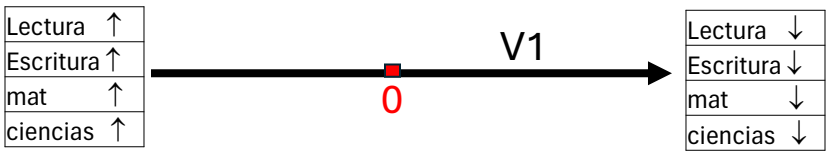
Variable Original	coeficientes	correlación	valor-p	IC (99%)	
				LI	LS
control	-0.83	-0.91	0.0000	-0.93	-0.9
conceptos	0.16	-0.09	0.0148	-0.2	0.01
motivacion	-0.42	-0.58	0.0000	-0.65	-0.51



29

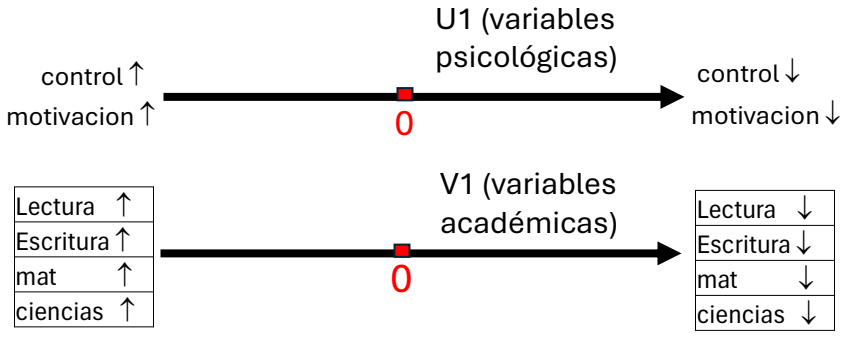
Variable canónica: V1

Variable Original	coeficientes	correlación	valor-p	IC (99%)	
				LI	LS
lectura	-0.44	-0.88	0.0000	-0.90	-0.85
escritura	-0.53	-0.91	0.0000	1.00	-0.89
mat	-0.18	-0.79	0.0000	-0.93	-0.76
ciencias	0.03	-0.69	0.0000	-0.74	-0.64



30

Correlación canónica: + 0.44



(1) Variables más importantes para representar la relación en cada uno de los ejes
(2) Incrementos en control y motivación se relacionan directamente con un mejor desempeño académico

31