

Estadística Aplicada III

Análisis de Correlación Canónica

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Semana 9: 10/12 de octubre de 2018

Introducción

Origen y propósito del Análisis de Correlación Canónica (ACC) I

- El propósito es identificar y cuantificar las asociaciones entre dos conjuntos de variables.
- En términos geométricos, responde a la pregunta: ¿qué direcciones son las que explican mejor la covarianza entre dos conjuntos de datos?
- Las correlaciones canónicas miden las fuerzas de asociación (lineal) entre dos conjuntos de variables, a través de combinaciones lineales de cada conjunto.
- Hay dos posibles situaciones a considerar en este análisis:
 - ▶ **simétrico**: Se consideran a los dos conjuntos de variables del mismo modo, es decir, no hay razón para pensar en una situación causal.
 - ▶ **asimétrica**: Se supone una posible relación causal: en donde unas variables explican a otras pero esta explicación no es bidireccional. El caso asimétrico no será analizado aquí.

Ejemplo. [1]

En este ejemplo se considera la situación simétrica. Harold Hotelling (1935) desarrolló la teoría con el siguiente problema: 140 niños de 7º. grado, recibieron 4 pruebas y sus evaluaciones se modelaron con las siguientes variables:

- X_1 = velocidad de lectura
- X_2 = comprensión de lectura

Origen y propósito del Análisis de Correlación Canónica (ACC) II

- Z_1 = velocidad aritmética
- Z_2 = capacidad aritmética

Las correlaciones observadas en su desempeño fueron las siguientes:

$$\mathbf{R} = \begin{matrix} & x_1 & x_2 & y_1 & y_2 \\ x_1 & 1 & 0.6328 & 0.2412 & 0.0586 \\ x_2 & 0.6328 & 1 & -0.0553 & 0.0655 \\ y_1 & 0.2412 & -0.0553 & 1 & 0.4248 \\ y_2 & 0.0586 & 0.0655 & 0.4248 & 1 \end{matrix} = \begin{pmatrix} \rho_{11} & \phi_{12} \\ \phi_{21} & \rho_{22} \end{pmatrix}$$

Pregunta: ¿podemos encontrar dos combinaciones lineales $\mathbf{a}'\mathbf{x}$ y $\mathbf{b}'\mathbf{y}$ con varianza unitaria tales que maximicen $\text{cor}(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}) = \mathbf{a}'\mathbf{R}\mathbf{b}$?

□

- Los vectores \mathbf{a} y \mathbf{b} son las *direcciones canónicas*.
- Las combinaciones lineales $\{\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}\}$ son las *variables canónicas* y
- el valor máximo que toma la correlación de las dos combinaciones lineales, es la *correlación canónica*.
- Podemos tener un número de $r = \min\{p, q\}$ correlaciones canónicas, pero se espera que un número l mucho menor sean suficientes para capturar la asociación más relevantes.

Origen y propósito del Análisis de Correlación Canónica (ACC) III

Ejemplo. [2]

Se quiere establecer una relación entre variables médicas y demográficas. Este ejemplo muestra una situación simétrica.

Conjunto X

- X_1 = Disposición a comprar medicamentos
- X_2 = Número de visitas médicas anuales
- X_3 = Horas de ejercicio a la semana
- X_4 = Dosis de medicamento A semanal

Conjunto Y

- Y_1 = Nivel educacional
- Y_2 = Ingreso
- Y_3 = Edad
- Y_4 = Cuenta con seguro médico
- Y_5 = Género
- Y_6 = Tipo de empleo



Ejemplo. [3]

En el caso particular de regresión múltiple, tenemos el primer conjunto con una variable (la respuesta) y el otro conjunto con los posibles p predictores. Este es un claro ejemplo de una situación asimétrica.

Del mismo modo, el modelo de regresión múltiple multivariado, es un caso de la situación asimétrica.



Otras aplicaciones relevantes:

- Clasificar y segmentar imágenes y escáneres de resonancia magnética

- Reconstruir modelos tridimensionales de rostros a partir de fotos.
- Índice de eficiencia de una empresa/institución
- Análisis de datos climáticos (temporales) en ciertas regiones geográficas (espaciales).
- Identificación de factores de riesgo en el cáncer de mama

En lo que sigue, se hará principalmente el análisis para la situación de asociación más general, que es el caso simétrico.

Modelo

Planteamiento del problema de Correlación Canónica I

- Sean $\underset{p \times 1}{\mathbf{y}}$ y $\underset{q \times 1}{\mathbf{x}}$ dos vectores de variables aleatorias con $r = \min\{p, q\}$, con la notación usual:

$$E(\mathbf{x}) = \boldsymbol{\mu}_x, E(\mathbf{y}) = \boldsymbol{\mu}_y, \text{Var}(\mathbf{x}) = \boldsymbol{\Sigma}_x, \text{Var}(\mathbf{y}) = \boldsymbol{\Sigma}_y, \text{cov}(\mathbf{x}, \mathbf{y}) = \boldsymbol{\Sigma}_{xy}.$$

- Considerando el vector agrupado $\mathbf{W} = \begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix}$, entonces sabemos que $E(\mathbf{W}) = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}$ y $\text{Var}(\mathbf{W}) = \begin{pmatrix} \boldsymbol{\Sigma}_y & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_x \end{pmatrix}$
- Noten que los pq elementos de $\boldsymbol{\Sigma}_{xy}$ o equivalentemente $\boldsymbol{\Sigma}_{yx}$, contienen la asociación (lineal) entre los dos conjuntos de variables.
- Lo que se logra con la correlación canónica es resumir la información de pq términos en muchas menos dimensiones.

Construcción de las correlaciones canónicas I

- La correlación entre dos combinaciones lineales $\mathbf{a}'\mathbf{y}$ y $\mathbf{b}\mathbf{x}$ está dada por

$$\rho(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}'\Sigma_{xy}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_y\mathbf{a}}\sqrt{\mathbf{a}'\Sigma_x\mathbf{b}}}$$

- Como nos interesa sólo la magnitud y no el signo de la correlación, se maximizará el cuadrado de la correlación, y para tener solución única, se impondrá la restricción de tener combinaciones lineales con varianza unitaria. Entonces el problema de optimización¹ es el siguiente:

$$\max_{\mathbf{a}, \mathbf{b}} \rho^2(\mathbf{a}, \mathbf{b}) = \frac{(\mathbf{a}'\Sigma_{yx}\mathbf{b})^2}{(\mathbf{a}'\Sigma_y\mathbf{a})(\mathbf{b}'\Sigma_x\mathbf{b})}$$

sujeto a:

$$\mathbf{a}'\Sigma_y\mathbf{a} = 1$$

$$\mathbf{b}'\Sigma_x\mathbf{b} = 1$$

¹Noten cómo difiere este problema de CP: ahí maximizamos $\mathbf{a}'\Sigma\mathbf{a}$ sujeto a la restricción $\mathbf{a}'\mathbf{a} = 1$.

Solución correlaciones canónicas I

- Se puede resolver el problema de muchas formas, una de las más fáciles es usar los multiplicadores de Lagrange. En este caso, la función a maximizar es

$$f(\mathbf{a}, \mathbf{b}) = (\mathbf{a}' \boldsymbol{\Sigma}_{yx} \mathbf{b})^2 - \kappa_1(\mathbf{a}' \boldsymbol{\Sigma}_y \mathbf{a} - 1) - \kappa_2(\mathbf{b}' \boldsymbol{\Sigma}_x \mathbf{b} - 1)$$

- Derivando con respecto a los coeficientes y utilizando el hecho de que $\boldsymbol{\Sigma}'_{yx} = \boldsymbol{\Sigma}_{xy}$:

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{a}} &= 2\boldsymbol{\Sigma}_{yx} \mathbf{b} - 2\kappa_1 \boldsymbol{\Sigma}_y \mathbf{a} \\ \frac{\partial f}{\partial \mathbf{a}} &= 2\boldsymbol{\Sigma}_{xy} \mathbf{a} - 2\kappa_2 \boldsymbol{\Sigma}_x \mathbf{b}\end{aligned}$$

Igualando a cero se obtiene que

$$\begin{aligned}\boldsymbol{\Sigma}_{yx} \mathbf{b} &= \kappa_1 \boldsymbol{\Sigma}_y \mathbf{a} \\ \boldsymbol{\Sigma}_{xy} \mathbf{a} &= \kappa_2 \boldsymbol{\Sigma}_x \mathbf{b}\end{aligned}$$

Multiplicando la primera ecuación por \mathbf{a}' y la segunda por \mathbf{b}' se tiene:

$$\begin{aligned}\mathbf{a}' \boldsymbol{\Sigma}_{yx} \mathbf{b} &= \kappa_1 \mathbf{a}' \boldsymbol{\Sigma}_y \mathbf{a} = \kappa_1 \\ \mathbf{b}' \boldsymbol{\Sigma}_{xy} \mathbf{a} &= \kappa_2 \mathbf{b}' \boldsymbol{\Sigma}_x \mathbf{b} = \kappa_2\end{aligned}$$

Solución correlaciones canónicas II

Entonces $\kappa_1 = \mathbf{a}'\Sigma_{yx}\mathbf{b} = \mathbf{b}'\Sigma_{xy}\mathbf{a} = \kappa_2$, por lo que:

$$\Sigma_{yx}\mathbf{b} = \kappa_1 \Sigma_y \mathbf{a}$$

$$\Sigma_{xy}\mathbf{a} = \kappa_1 \Sigma_x \mathbf{b}$$

Despejando \mathbf{b} de la segunda ecuación, $\mathbf{b} = \kappa_1^{-1} \Sigma_x^{-1} \Sigma_{xy} \mathbf{a}$ y sustituyendo en la primera ecuación,

$$\Sigma_{yx}(\kappa_1^{-1} \Sigma_x^{-1} \Sigma_{xy})\mathbf{a} = \kappa_1 \Sigma_y \mathbf{a} \quad (1)$$

$$(\Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \Sigma_{xy})\mathbf{a} = \kappa_1^2 \mathbf{a} \quad (2)$$

Entonces la dirección canónica \mathbf{a} es un vector propio de la matriz cuadrada

$$\underset{p \times p}{\mathbf{A}} = \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \Sigma_{xy}$$

con valor propio $\lambda = \kappa_1^2 = (\mathbf{a}'\Sigma_{yx}\mathbf{b})^2 = (\mathbf{b}'\Sigma_{xy}\mathbf{a})^2$.

Del mismo modo se puede obtener que \mathbf{b} es el vector propio ligado a κ_2^2 de la matriz

$$\underset{q \times q}{\mathbf{B}} = \Sigma_x^{-1} \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx}$$

En conclusión, obtenemos que $\rho^{*2} = \lambda$ es el cuadrado de la correlación entre las variables canónicas óptimas, $\mathbf{a}'\mathbf{y}$ y $\mathbf{b}'\mathbf{x}$, y para maximizar tomamos el valor propio λ_1 correspondiente más grande.

Solución correlaciones canónicas III

- Las subsecuentes correlaciones canónicas (segunda, tercera, etc.) de dos pares de variables canónicas se obtienen de tal forma que sean ortogonales a las previas, y se pueden obtener hasta $r = \min\{p, q\}$ pares, y corresponderán justamente a los primeros r eigenvalores de las matrices que se obtuvieron arriba.
- De hecho, las matrices \mathbf{A} y \mathbf{B} definidas anteriormente tienen los mismos valores propios², y son no negativos.

²Resultado de álgebra lineal: Si $\underset{n \times p}{\mathbf{A}}$ y $\underset{p \times n}{\mathbf{B}}$, los eigenvalores diferentes de 0 de \mathbf{AB} y \mathbf{BA} son los mismos y tienen la misma multiplicidad. Si \mathbf{x} es un eigenvector no trivial de \mathbf{AB} para $\lambda \neq 0$, entonces $\mathbf{y} = \mathbf{Bx}$ es un eigenvector no trivial para \mathbf{BA} .

Subsecuentes correlaciones canónicas I

Lema

Las matrices $\mathbf{L}^{-1}\mathbf{M}$ y $\mathbf{L}^{-1/2}\mathbf{ML}^{-1/2}$ tienen los mismos valores propios. Además, si \mathbf{v} es un eigenvector de la primera, $\mathbf{L}^{1/2}\mathbf{v}$ lo es de la segunda.

Demostración.

Sea λ un valor propio de $\mathbf{L}^{-1}\mathbf{M}$ y sea \mathbf{v} su vector propio asociado. Entonces $\mathbf{L}^{-1}\mathbf{M}\mathbf{v} = \lambda\mathbf{v}$. Multiplicando ambos lados de la igualdad por $\mathbf{L}^{1/2}$ se obtiene: $\mathbf{L}^{-1/2}\mathbf{M}\mathbf{v} = \lambda\mathbf{L}^{1/2}\mathbf{v}$. Además podemos escribir:

$$\begin{aligned}\lambda\mathbf{L}^{-1/2}\mathbf{v} &= \mathbf{L}^{1/2}\mathbf{M}(\mathbf{I})\mathbf{v} \\ &= \mathbf{L}^{1/2}\mathbf{M}(\mathbf{L}^{-1/2}\mathbf{L}^{1/2})\mathbf{v} \\ &= \mathbf{L}^{1/2}\mathbf{M}\mathbf{L}^{-1/2}(\mathbf{L}^{1/2}\mathbf{v})\end{aligned}$$

Por lo que $\lambda\mathbf{h} = \mathbf{L}^{1/2}\mathbf{M}\mathbf{L}^{-1/2}\mathbf{h}$ donde $\mathbf{h} = \mathbf{L}^{1/2}\mathbf{v}$ y entonces λ es un valor propio común de las matrices originalmente consideradas.

□

Subsecuentes correlaciones canónicas II

- Si hacemos $\mathbf{H} = \boldsymbol{\Sigma}_y^{-1/2} \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_x^{-1/2}$ entonces notemos que

$$\mathbf{A} = \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_{xy}$$

y

$$\mathbf{H}'\mathbf{H} = \boldsymbol{\Sigma}_y^{-1/2} \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_y^{-1/2}$$

tienen los mismos valores propios, porque en el lema podemos hacer
 $\mathbf{L} = \boldsymbol{\Sigma}_y$ y $\mathbf{M} = \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_{xy}$

- Comprueben entonces que \mathbf{A} tiene los valores propios de \mathbf{HH}' y que \mathbf{B} tiene los mismos valores propios de $\mathbf{H}'\mathbf{H}$. Como las matrices son semidefinidas positivas, los eigenvalores de \mathbf{A} y \mathbf{B} son reales y no negativos.

Correlación entre las variables canónicas I

- Debido a que las variables canónicas se forman de los eigenvectores de una matriz simétrica ($\mathbf{H}'\mathbf{H}$), entonces los eigenvectores son ortogonales, por lo que en general,

$$\text{cor}(\mathbf{a}_i'\mathbf{y}, \mathbf{a}_j'\mathbf{y}) = \mathbf{a}_i'\Sigma_y\mathbf{a}_j = \boldsymbol{\alpha}_i'\boldsymbol{\alpha}_j = \delta_{ij}$$

donde $\delta_{ij} = I(i=j)$ es la delta de Kronecker,y del mismo modo

$$\text{cor}(\mathbf{b}_i'\mathbf{x}, \mathbf{b}_j'\mathbf{x}) = \delta_{ij}$$

- Entonces en general, si tomamos $\boldsymbol{\eta} = (\mathbf{a}_1'\mathbf{y}, \dots, \mathbf{a}_r'\mathbf{y})$ y $\boldsymbol{\phi} = (\mathbf{b}_1'\mathbf{x}, \dots, \mathbf{b}_r'\mathbf{x})$, entonces

$$\text{Var} \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\phi} \end{pmatrix} = \text{cor} \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\phi} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_r & \text{diag}(\lambda_i^{1/2}) \\ \text{diag}(\lambda_i^{1/2}) & \mathbf{I}_r \end{pmatrix}$$

- Entonces las variables canónicas tienen correlación 0 en el mismo grupo y entre grupos.

Invarianza de las variables canónicas I

- Si se definen variables $y^* = \mathbf{U}y + \mathbf{u}$ y $x^* = \mathbf{V}x + \mathbf{v}$, donde $\mathbf{U}_{p \times p}$ y $\mathbf{V}_{q \times q}$ son matrices no singulares y \mathbf{u}, \mathbf{v} son vectores fijos. Entonces se cumplen las siguientes condiciones:
 - ▶ Las correlaciones canónicas entre y^* y x^* son las mismas que para y y x .
 - ▶ Los vectores de correlación canónica para y^* y x^* están dados por $a_i^* = \mathbf{U}^{-1}\mathbf{a}_i$ y $b_i^* = \mathbf{V}^{-1}\mathbf{b}_i$.

Demostración.

Tarea.

□

- Recuerden por ejemplo, que esta propiedad de invarianza no se cumple en componentes principales.

Herramientas en R I

- hay métodos de componentes principales en varios paquetes de R:
 - ▶ La función `cancor` es la función por default. Esta función utiliza rotaciones primero para aplicar la descomposición en valor singular en una transformación simple y luego regresa los resultados, en lugar de usar las correlaciones. Puede dar eigenvalores diferentes, pero en la dirección adecuada.
 - ▶ `CCA`: Canonical Correlation Analysis: extiende la función `cancor` con cálculos numéricos y con salidas gráficas. También permite extender el análisis de correlación canónica para trabajar con conjuntos de datos que tienen más variables que observaciones.
 - ▶ `CCP`: Significance Tests for Canonical Correlation Analysis. Incluye pruebas paramétricas, no paramétricas y basadas en Monte Carlo (permutaciones).
 - ▶ `candisc`: Visualizing Generalized Canonical Discriminant and Canonical Collelation Analysis.
 - ▶ `vegan` Ordination methods, diversity analysis and other functions for community and vegetation ecologists. Tiene la función `CCorA`
 - ▶ `yacca` Provides an alternative canonical correlation/redundancy analysis function, with associated print, plot, and summary methods. A method for generating helio plots is also included.
- En Matlab hay una función llamada `canoncorr`.

Aplicaciones

Ejemplo 1. Datos de hogares (Peña & Romo, 1997) I

- Se cuenta con un conjunto de datos de 75 hogares españoles. Las primeras 5 variables se refieren a gastos del hogar en diferentes rubros, y los últimos 4 se refieren a la estructura del hogar. En este ejemplo, $p = 4$ y $q = 5$, y el número máximo de variables canónicas que podemos encontrar es $r = \min\{4, 5\} = 4$.
- En este caso haremos el ejercicio utilizando las fórmulas y después utilizaremos los paquetes disponibles para hacer comparaciones.

Ejemplo 1. Datos de hogares (Peña & Romo, 1997) II

```
W <- read.table("../data/hogares.dat", header=T, sep="")
head(W)

  gasto_alimento gasto_ropa gasto_menaje gasto_transpor gasto_educa num_personas num_personas_14
1      55432       6880        780        4120       2400          4            2
2      63076       2620       4296          0        384          2            0
3      62816       1000       3044        2470          0            6            4
4      80236       7980      52016        3744          0            4            2
5      90636       8080      13128        40801      26560          3            1
6      89752      43100       2392        13474      9656          5            0
  nivel_educativo num_aporta
1                  3           1
2                  2           1
3                  2           1
4                  4           2
5                  3           1
6                  1           1

#Obten las correlaciones de acuerdo a los grupos considerados.
#Consideremos el grupo más chico como la primera opción
R <- cor(W)
R11 <- R[6:9,6:9]; R22 <- R[1:5,1:5]; R12 <- R[6:9,1:5]; R21 <- t(R12)

#Cálculo de A (la matriz chica para estructura)
A <- solve(R11) %*% R12 %*% solve(R22) %*% R21
sA <- eigen(A); sA

eigen() decomposition
$values
[1] 0.43997319 0.20924770 0.05364298 0.01192928

$vectors
      [,1]      [,2]      [,3]      [,4]
[1,]  0.77652418 -0.5077004 -0.282966409  0.6498854
[2,] -0.27449701  0.1022458  0.956112901 -0.3794341
[3,]  0.56235382  0.7984566  0.008387624 -0.1130298
[4,]  0.07361916 -0.3070068  0.075549852 -0.6487704
```

Ejemplo 1. Datos de hogares (Peña & Romo, 1997) I

- Podemos observar lo siguiente:

1 El primer eigenvalor es $\lambda_1 = 0.4399732$ y entonces la correlación entre las dos primeras variables canónicas es $\sqrt{\lambda_1} = 0.6633047$. La combinación lineal correspondiente a las variables del grupo de estructura es

$$\mathbf{a}'\mathbf{y} = 0.78\text{num_personas} - 0.27\text{num_personas14} + 0.56\text{nivel_edu} + 0.07\text{num_aporta}$$

2 Para la segunda variable, resolvemos la matriz B, que tiene de hecho los mismos eigenvalores:

```
# Cálculo de la matriz B
B <- solve(R22) %*% R21 %*% solve(R11) %*% R12
sB <- eigen(B);sB

eigen() decomposition
$values
[1] 4.399732e-01 2.092477e-01 5.364298e-02 1.192928e-02 1.871566e-17

$vectors
[,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.76043894  0.41079216  0.6275495 -0.10692113  0.15257396
[2,] 0.42753263 -0.05132244 -0.4514007 -0.71461908 -0.31372013
[3,] 0.30980756 -0.86926435  0.3376244  0.07884318  0.07875654
[4,] 0.37587938  0.26859957 -0.2379261  0.64987762 -0.55520687
[5,] 0.04101753 -0.02914752 -0.4814769  0.22210418  0.75089520
```

La correspondiente combinación lineal es:

$$\mathbf{b}'\mathbf{x} = 0.76\text{galimento} + 0.43\text{gropa} + 0.31\text{gmenaje} + 0.37\text{gtrans} + 0.04\text{geduca}$$

3 La variabilidad explicada por las primeras variables canónicas es 61.55 %.

Ejemplo 1. Datos de hogares (Peña & Romo, 1997) II

- ④ La variable canónica asociada al gasto es un promedio ponderado de los gastos de la familia, dando mayor peso a la alimentación y menor a la educación y esparcimiento, y se relaciona con el indicador de estructura que pondera el tamaño de la familia y el nivel de educación del ingreso principal.
- ⑤ Nota importante: los eigenvectores tienen norma 1, y la condición que se pide para correlación canonica es que la varianza de la variable canónica sea 1, que no se cumple. Entonces los factores pueden diferir en un factor constante de otros cómputos.
- ⑥ El cálculo de los scores y su correlación,

```
sestructura <- as.matrix(W[, 6:9]) %*% sA$vectors[, 1]
sgasto   <- as.matrix(W[, 1:5]) %*% sB$vectors[, 1]
cor(sestructura, sgasto)

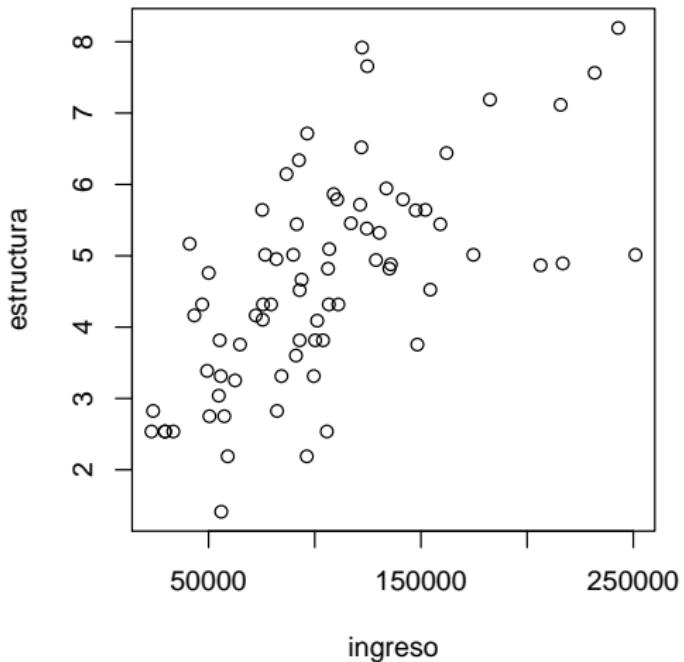
[,1]
[1,] 0.6397437
```

- ⑦ Graficando los scores de las primeras variables canónicas:

```
par(pty="s")
plot(sgasto, sestructura, xlab="ingreso", ylab="estructura", main="Primeras variables canónicas")
```

Ejemplo 1. Datos de hogares (Peña & Romo, 1997) III

Primeras variables canónicas



Datos de hogares con cancer I

- Ahora comparamos los resultados con la función básica `cancor`. La función hace una descomposición diferente algebráicamente (usando la descomposición QR y usando SVD para obtener las variables canónicas, sin usar las correlaciones. Por eso tiene la opción para centrar o no los datos.

```
gasto <- W[,1:5]; estructura <- W[,6:9]
u <- cancor(estructura,gasto,xcenter = T, ycenter=T)
#Correlaciones canónicas:
u$cor

[1] 0.6633047 0.4574360 0.2316095 0.1092213

# coeficientes de la combinación canónica para estructura:
u$xcoef

 [,1]          [,2]          [,3]          [,4]
num_personas -0.056249221 -0.03477786  0.024133772  0.09812361
num_personas_14  0.028766702  0.01013285 -0.117975117 -0.08288272
nivel_educativo -0.053580537  0.07194201 -0.000940947 -0.02244738
num_aporta     -0.007784534 -0.03069895 -0.009405982 -0.14299075

#coeficientes para gasto
u$ycoef

 [,1]          [,2]          [,3]          [,4]          [,5]
gasto_alimento -1.277121e-06 -9.901148e-07  1.633421e-06  2.578713e-07 -4.179282e-07
gasto_ropa      -1.627102e-06  2.803165e-07 -2.662504e-06  3.905637e-06  1.947339e-06
gasto_menaje    -1.099432e-06  4.427146e-06  1.856917e-06 -4.018020e-07 -4.558443e-07
gasto_transpor -1.186430e-06 -1.216731e-06 -1.163906e-06 -2.945757e-06  2.858262e-06
gasto_educa     -1.088922e-07  1.110515e-07 -1.981002e-06 -8.467507e-07 -3.251322e-06
```

Datos de hogares con cancer II

- Calculamos los scores, su correlación y hacemos su gráfica. En este caso cambia la escala, pero no la dirección.

```
sestructura <- as.matrix(estructura) %*% u$xcoef[,1]
sgasto <- as.matrix(gasto) %*% u$ycoef[,1]
cor(sgasto,sestructura)

[,1]
[1,] 0.6633047

#vemos que la forma de estimación pone la misma varianza en las dos variables canónicas:
var(sgasto)

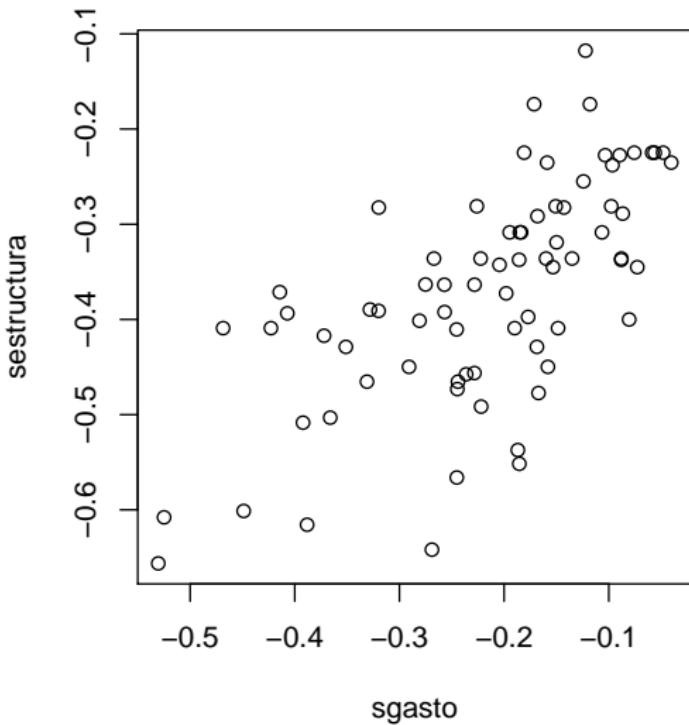
[,1]
[1,] 0.01351351

var(sestructura)

[,1]
[1,] 0.01351351

#gráfica de los scores:
par(pty="s")
plot(sgasto,sestructura)
```

Datos de hogares con cancer III



Ejemplo 2: scores (Mardia, 1979) I

- Consideraremos de nuevo (lo hicimos con PCA) los datos de $n = 88$ estudiantes que toman 5 materias. El score es sobre un máximo de 100 puntos.
- Queremos relacionar el conjunto de variables que fueron a libro cerrado: { mechanics,vectors } con los que fueron a libro abierto: { algebra, analysis,statistics }. Estos conjuntos están relacionados, como se puede ver en la matriz de correlaciones.

```
E <- read.csv("../data/score.txt", header=T, sep="", row.names = 1)
colnames(E) <- c("mec", "vec", "alg", "ana", "sta")
cor(E)
```

	mec	vec	alg	ana	sta
mec	1.0000000	0.5534052	0.5467511	0.4093920	0.3890993
vec	0.5534052	1.0000000	0.6096447	0.4850813	0.4364487
alg	0.5467511	0.6096447	1.0000000	0.7108059	0.6647357
ana	0.4093920	0.4850813	0.7108059	1.0000000	0.6071743
sta	0.3890993	0.4364487	0.6647357	0.6071743	1.0000000

- Aplicando cancor, obtenemos que las primeras direcciones canónicas son (multiplicadas por 1000): $\mathbf{a}_1 = (2.77, 5.517)'$ y $\mathbf{b}_1 = (8.782, 0.86, 0.37)'$, y la primera correlación canónica es 0.663.

Ejemplo 2: scores (Mardia, 1979) II

- Las segundas direcciones canónicas no son significativas

```
u <- cancor(E[,1:2],E[,3:5]); u

$cor
[1] 0.66305211 0.04094594

$xcoef
[,1]      [,2]
mec 0.002769608 0.006820239
vec 0.005517014 -0.008088354

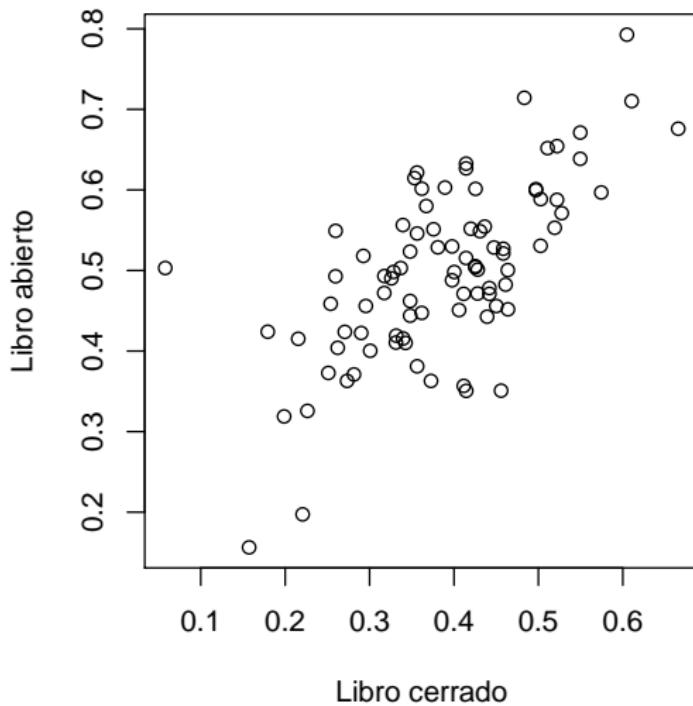
$ycoef
[,1]      [,2]      [,3]
alg 0.0087816197 0.009687244 0.0088433854
ana 0.0008598730 -0.010549747 0.0008906466
sta 0.0003703994 0.001536399 -0.0084575453

$xcenter
mec      vec
38.95455 50.59091

$ycenter
alg      ana      sta
50.60227 46.68182 42.30682

par(pty="s")
plot(as.matrix(E[,1:2]) %*% u$xcoef[,1], as.matrix(E[,3:5]) %*% u$ycoef[,1],
     xlab="Libro cerrado", ylab="Libro abierto")
```

Ejemplo 2: scores (Mardia, 1979) III



Ejemplo 3. Aceites de oliva (Forina, et al, 1983) I

- En este ejemplo, se tienen $n = 572$ aceites de oliva, y cada uno tiene 8 características. La primera variable es una variable indicadora de la región en Italia y las otras 8 variables miden la composición de 8 ácidos grasos.
- El problema consiste en ver la correlación entre las regiones de origen y las medidas de ácidos grasos. En este caso, la variable del primer conjunto son los niveles de una variable categorica o factor, y el otro conjunto son las 8 variables.
- Como la variable `region` es categorica, necesitamos convertirla a una matriz de indicadoras. A continuación se utiliza una función para este fin. No se elimina ninguna categoría porque

Ejemplo 3. Aceites de oliva (Forina, et al, 1983) II

```
W <- read.csv("../data/olive.dat", header=T, sep="")
head(W)

  region palmitic palmitoleic stearic oleic linoleic linolenic arachidic eicosenoic
1       1      1075          75     226    7823      672       36      60      29
2       1      1088          73     224    7709      781       31      61      29
3       1       911          54     246    8113      549       31      63      29
4       1       966          57     240    7952      619       50      78      35
5       1      1051          67     259    7771      672       50      80      46
6       1       911          49     268    7924      678       51      70      44

as.matind <- function(z) { #crea una matriz de indicadoras, z es categorica
z <- as.factor(z)
l <- levels(z)
b <- as.numeric(z==rep(1,each=length(z)))
return(matrix(b,length(z)))
}
y <- as.matind(W[,1])
x <- as.matrix(W[,2:9])
```

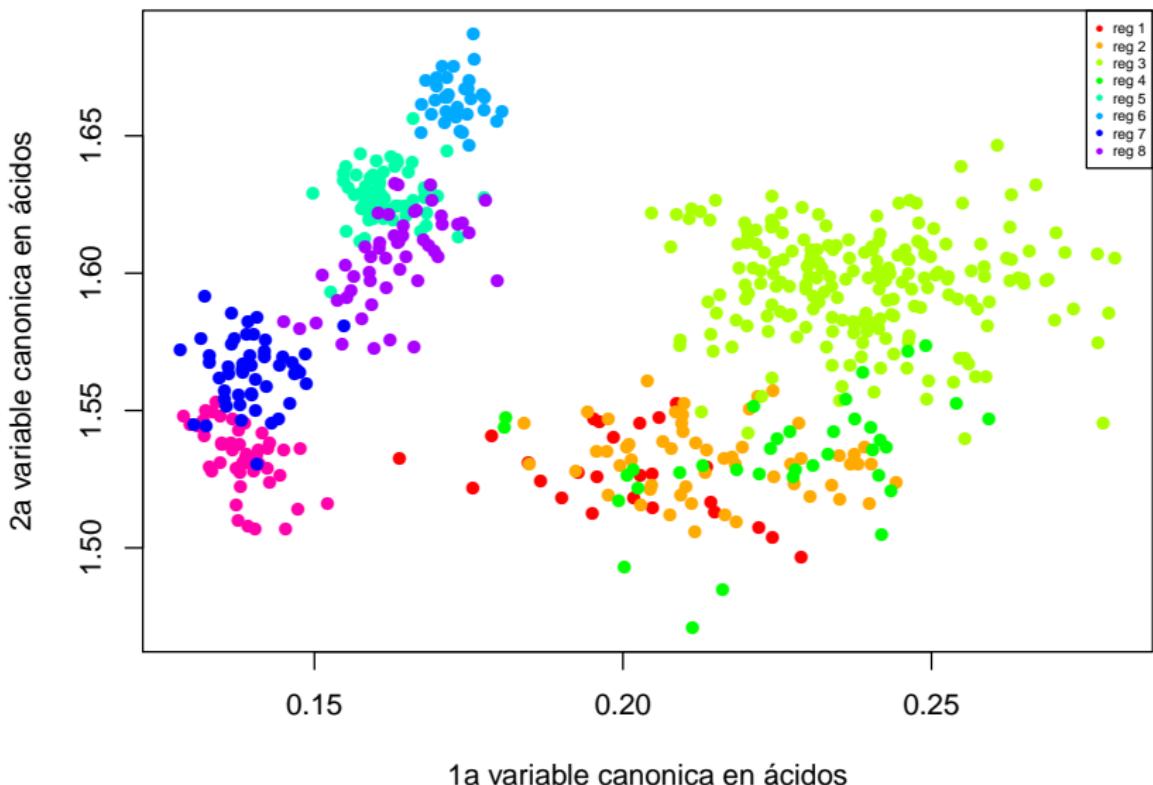
- Una vez establecida la estructura necesaria, ejecutamos la correlación canónica sobre los conjuntos de variables y x

```
u <- cancor(x, y, ycenter = F) #En este caso no tiene sentido centrar y
# Tenemos un total de 8 variables canónicas
acidos   <- x %*% u$xcoef
regiones <- y %*% u$ycoef

colores = rainbow(9)
# Usamos las dos primeras direcciones en el conjunto de los ácidos, y marcamos con y
plot(acidos[,1:2], col = colores[W[,1]], pch=16,
      xlab = "la variable canonica en ácidos",
      ylab = "2a variable canonica en ácidos")

legend("topright", pch=16,col=colores, legend=paste("reg",1:8,sep=" "), cex=0.5)
```

Ejemplo 3. Aceites de oliva (Forina, et al, 1983) III

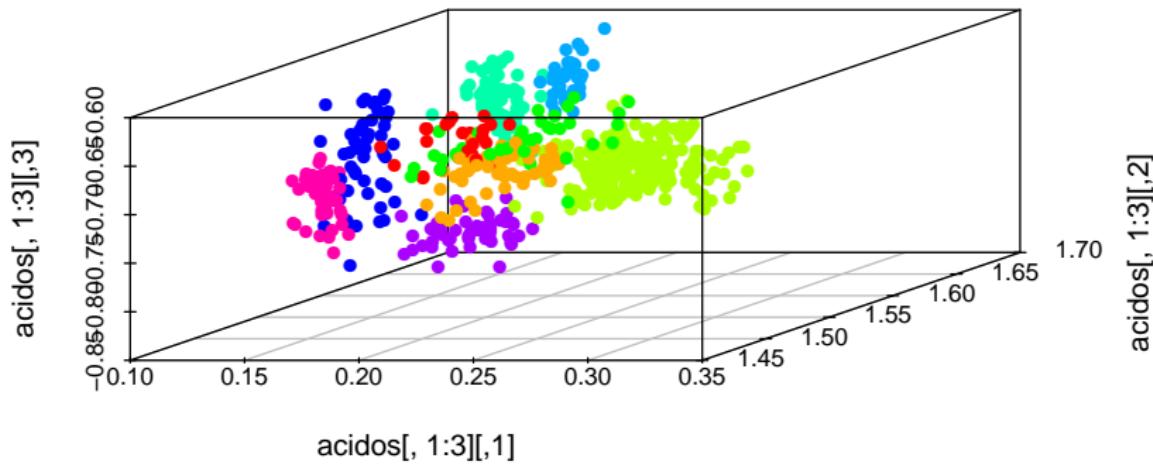


Ejemplo 3. Aceites de oliva (Forina, et al, 1983) I

- En este ejemplo, el análisis de correlación canónica nos sirve para hacer una *clasificación* de los datos. El *análisis discriminante lineal* hace exactamente lo mismo que correlación canónica. Más adelante cubriremos este tema.
- Podemos todavía considerar una tercera variable para ver si las direcciones canónicas ayudan a separar mejor las agrupaciones de las regiones:

```
library(scatterplot3d)
scatterplot3d(acidos[, 1:3], color=colores[W[, 1]], pch=16)
```

Ejemplo 3. Aceites de oliva (Forina, et al, 1983) II



- Una segunda opción interactiva:

```
library(rgl)
plot3d(acidos[,1:3], col=colores[W[,1]])
```

Inferencia

Inferencia bajo supuestos de normalidad

- Cuando se puede asumir normalidad: $\mathbf{W} \sim \mathcal{N}_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces se puede evaluar si tiene sentido realizar un análisis de correlación canónica, probando primero si $\boldsymbol{\Sigma}_{yx} = \mathbf{0}$.
- La prueba de verosimilitud para:

$$H_0 : \boldsymbol{\Sigma}_{yx} = \mathbf{0} \quad \text{vs.} \quad H_a : \boldsymbol{\Sigma}_{yx} \neq \mathbf{0}$$

rechaza H_0 para valores grandes de la estadística:

$$-2 \log \Lambda = n \log \left(\frac{|\mathbf{S}_y||\mathbf{S}_x|}{|\mathbf{S}|} \right) = -n \log \prod_{i=1}^p (1 - \hat{\rho}_i^{*2}) \underset{n \rightarrow \infty}{\sim} \chi_{pq}^2$$

donde $\mathbf{S} = \begin{pmatrix} \mathbf{S}_y & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_x \end{pmatrix}$ es un estimador insesgado de $\boldsymbol{\Sigma}$, y $\hat{\rho}_i^*$ es el estimador de la i -ésima correlación canónica.

- Noten que la prueba anterior compara la varianza generalizada bajo H_0 y bajo H_a . La prueba se deriva del hecho de que

$$|\mathbf{S}| = |\mathbf{S}_y - \mathbf{S}_x^{-1} \mathbf{S}_{xy} \mathbf{S}_y^{-1} \mathbf{S}_{yx}| = \prod_{i=1}^p (1 - \lambda_i^2)$$

- La aproximación mejora si se sustituye n por la corrección de Bartlett, $m = n - \frac{1}{2}(p + q + 3)$.

Prueba parcial de correlación canónica I

- La prueba anterior puede extenderse para hacer el contraste de hipótesis:
 $H_0 : \rho_{s+1}^* = \cdots = \rho_p^*$ vs. $H_a : \rho_k^* > 0$ para al menos un $k \in \{s + 1, \dots, p\}$
- La prueba de verosimilitud LRT es ahora:

$$-2 \log \Lambda = -m \sum_{j=s+1}^p \log(1 - \hat{\rho}_j^{*2}) \sim \chi^2_{(p-s)(q-s)}$$

Observaciones a las pruebas de hipótesis

- En la práctica usualmente se aplica esta prueba secuencialmente con pruebas parciales de nivel α , pero de esa manera el nivel de significancia global **no** será α .
- Es importante notar que el análisis de Correlación canónica es sensible a la normalidad de los datos y a la presencia de valores atípicos, por lo que antes de aplicarlas es necesario verificar normalidad.

Ejemplo I

Ejemplo. [Para datos de Hogares:]

Continuando con el ejemplo de hogares, se muestran los resultados utilizando el paquete vegan y CCP para estimación, gráficas y pruebas de hipótesis.

```
## Hogares
library(vegan)
library(CCP)

W <- read.table("../data/hogares.dat", header=T, sep="")
ccl <- CCorA(W[, 1:5], W[, 6:9])
ccl

Canonical Correlation Analysis

Call:
CCorA(Y = W[, 1:5], X = W[, 6:9])

      Y   X
Matrix Ranks 5 4

Pillai's trace:  0.7147931

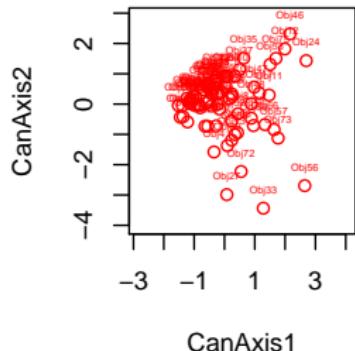
Significance of Pillai's trace:
from F-distribution:  2.8411e-05
                    CanAxis1 CanAxis2 CanAxis3 CanAxis4
Canonical Correlations  0.66330  0.45744  0.23161  0.1092

      Y | X   X | Y
RDA R squares      0.24385  0.2475
adj. RDA R squares 0.20064  0.1929

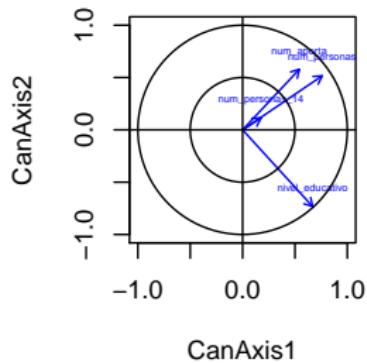
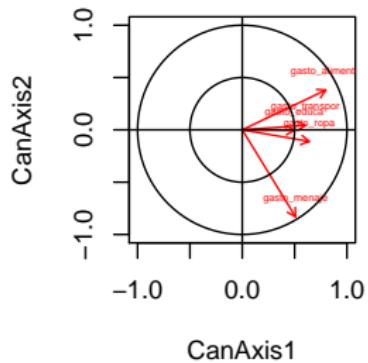
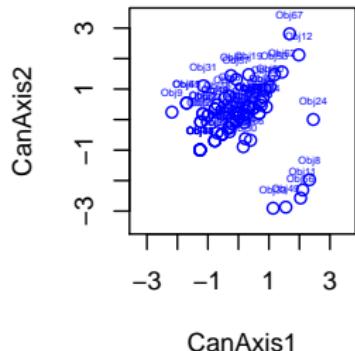
biplot(ccl, cex=c(0.4, 0.4), pch=16)
```

Ejemplo II

First data table (Y)



Second data table (X)



Ejemplo III

Pruebas de significancia para las correlaciones obtenidas:

```
p.asym(cc1$CanCorr, nrow(W), 4, 5, tstat = "Wilks")
```

Wilks' Lambda, using F-approximation (Rao's F):

	stat	approx	df1	df2	p.value
1 to 4:	0.4140877	3.3474009	20	219.8471	5.421572e-06
2 to 4:	0.7394069	1.7885227	12	177.5568	5.304080e-02
3 to 4:	0.9350677	0.7737939	6	136.0000	5.918138e-01
4 to 4:	0.9880707	0.4165291	2	69.0000	6.609780e-01

```
p.asym(cc1$CanCorr, nrow(W), 4, 5, tstat = "Hotelling")
```

Hotelling-Lawley Trace, using F-approximation:

	stat	approx	df1	df2	p.value
1 to 4:	1.11900427	3.6087888	20	258	8.700716e-07
2 to 4:	0.33337548	1.8474558	12	266	4.114689e-02
3 to 4:	0.06875697	0.7849754	6	274	5.823263e-01
4 to 4:	0.01207331	0.4255841	2	282	6.538070e-01

```
p.asym(cc1$CanCorr, nrow(W), 4, 5, tstat = "Pillai")
```

Pillai-Bartlett Trace, using F-approximation:

	stat	approx	df1	df2	p.value
1 to 4:	0.71479315	3.0025949	20	276	2.841113e-05
2 to 4:	0.27481996	1.7459753	12	284	5.708826e-02
3 to 4:	0.06557226	0.8110922	6	292	5.619710e-01
4 to 4:	0.01192928	0.4486862	2	300	6.388942e-01

```
p.asym(cc1$CanCorr, nrow(W), 4, 5, tstat = "Roy")
```

Roy's Largest Root, using F-approximation:

	stat	approx	df1	df2	p.value
1 to 1:	0.43999732	10.84168	5	69	1.017949e-07

F statistic for Roy's Greatest Root is an upper bound.

Ideas sobre Análisis de Correlación asimétrica

Redundancia I

- Una forma corresponde a regresión múltiple multivariada.
- Para medir la capacidad predictiva de un conjunto de variables respecto a otro se introducen los **coeficientes de redundancia**.
- En este contexto, se supone que las variables están estandarizadas

Coeficiente de redundancia

El *coeficiente de redundancia* es el valor promedio del cuadrado de las correlaciones entre \mathbf{y} y $\mathbf{a}'\mathbf{x}$:

$$CR(\mathbf{y}|\mathbf{a}'\mathbf{x}) = \frac{1}{q} \mathbf{a}' \mathbf{R}_{yx} \mathbf{R}_{xy} \mathbf{a}$$

Si $r = \min\{p, q\}$ y se tienen r combinaciones lineales, la *redundancia total* está dada por:

$$R(\mathbf{y}\mathbf{x}) = \sum_{i=1}^r CR(\mathbf{y}|\mathbf{a}_i'\mathbf{x})$$

y es una medida de asociación entre \mathbf{y} y \mathbf{x} .

Redundancia II

- En el caso de correlación canónica,

$$R(\mathbf{y}|\mathbf{x}) = \frac{\text{tr}(\mathbf{R}_{yx}\mathbf{R}_x^{-1}\mathbf{R}_{xy})}{\text{tr}(\mathbf{R}_y)} = \sum_{i=1}^r \frac{1}{q} R_i^2$$

donde R_i^2 es el coeficiente de determinación de la regresión entre y_i y \mathbf{x}