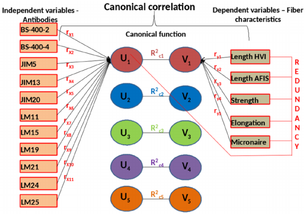
Estadística Aplicada III

Análisis de Correlación Canónica

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Semana 11



Introducción

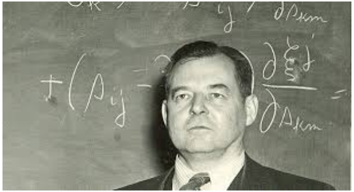
Origen y propósito del Análisis de Correlación Canónica (ACC) I

- El propósito del **Análisis de Correlación Canónica (CCA)** es identificar y cuantificar las asociaciones entre dos conjuntos de variables $\{X_1, \ldots, X_p\}$ y $\{Y_1, \ldots, Y_q\}$.
- Geométricamente, responde a la pregunta: ¿qué direcciones (combinaciones lineales) son las que explican mejor la variabilidad conjunta de dos grupos de variables?
- Las cc's miden las fuerzas de asociación lineal entre dos conjuntos de variables, a través de *combinaciones lineales* en cada conjunto de variables.
- 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, es decir, en que un conjunto de variables *causa* al otro conjunto.
  - **Asimétrica**: Se supone una posible relación causal: en donde unas variables explican a otras pero esta explicación no es bidireccional. Un ejemplo de este tipo de relaciones es la regresión lineal, en donde una variable de respuesta y se relaciona con otro grupo de variables $x_1, \ldots, x_p$ (los predictores). El caso asimétrico no será desarrollado aquí.

Ejemplo. [1. Relación entre habilidades literarias y matemáticas]

Ejemplo simétrico. Harold Hotelling (1935) desarrolló la teoría con el siguiente problema: El propósito es evaluar la relación lineal entre los *constructos* de lectura y de aritmética. 140 niños de 7° grado, recibieron 4 pruebas y sus evaluaciones se modelaron con las siguientes variables:

- $X_1 =$ velocidad de lectura
- $X_2 =$ comprensión de lectura
- $Y_1 =$ velocidad aritmética
- $Y_2 =$ capacidad aritmética

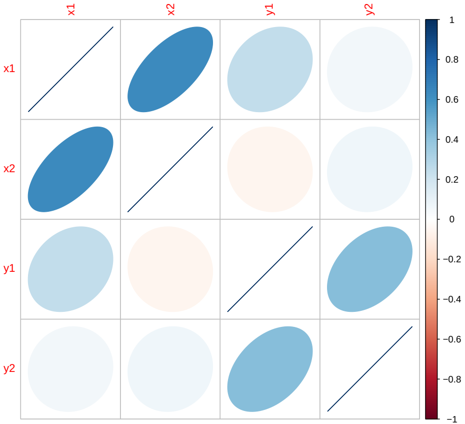


Hotelling consideró la pregunta: la velocidad y comprensión de lectura, consideradas *como un conjunto*, ¿están linealmente relacionadas a la velocidad y capacidad aritmética, consideradas también de manera conjunta?
Las correlaciones entre las variables observadas en su desempeño:

$$\mathbf{R} = \begin{array}{c c} & \begin{array}{cccc} x_1 & x_2 & y_1 & y_2 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ y_1 \\ y_2 \end{array} & \begin{bmatrix} 1 & 0.6328 & 0.2412 & 0.0586 \\ 0.6328 & 1 & -0.0553 & 0.0655 \\ 0.2412 & -0.0553 & 1 & 0.4248 \\ 0.0586 & 0.0655 & 0.4248 & 1 \end{bmatrix} \end{array} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

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

La correlación se ve de la siguiente forma:



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

- La correlación entre x_1 y y_1 es relativamente alta comparada con el cruce de las otras variables, pero en general no son valores altos.
- ¿Existe alguna combinación lineal con valores a_1, a_2, b_1, b_2 tales que los constructos $X = a_1X_1 + a_2X_2$ tenga una alta correlación con $Y = b_1Y_1 + b_2Y_2$? ¿Podemos encontrar dos combinaciones lineales de las variables X's y Y's, es decir dos combinaciones $\mathbf{a}'\mathbf{x}$ y $\mathbf{b}'\mathbf{y}$, tales que maximicen su correlación $\text{cor}(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}) = \mathbf{a}'\mathbf{R}\mathbf{b}$? La respuesta es sí, pero imponiendo algunas restricciones.

□

Ejemplo. [2. Relación entre datos médicos y demográficos.]

Se quiere establecer una relación entre variables médicas y demográficas. Caso simétrico.

Conjunto X:	Conjunto Y:
X_1 = Disposición a comprar medicamentos	Y_1 = Nivel educacional
X_2 = Número de visitas médicas anuales	Y_2 = Ingreso
X_3 = Horas de ejercicio a la semana	Y_3 = Edad
X_4 = Dosis de medicamento A semanal	Y_4 = Cuenta con seguro médico
	Y_5 = Género
	Y_6 = Tipo de empleo

□

Ejemplo. [3. El análisis de regresión es un caso particular de ACC.]

El primer conjunto con una variable (la respuesta) y el otro conjunto corresponde a 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:

- En educación, análisis entre rendimiento escolar y el tiempo de ocio. Un caso asimétrico es el rendimiento en la preparatoria relacionado con el rendimiento en la Universidad. El primero explica el segundo, pero no al revés.
- 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*.

- Los vectores $\mathbf{a}$ y $\mathbf{b}$ se llaman *direcciones canónicas*.
- Las combinaciones lineales $\{\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}\}$ son las *variables canónicas*.
- El valor máximo que toma la correlación de las dos combinaciones lineales, es la *correlación canónica*.
- En un conjunto de $p + q$ variables, podemos tener un número $r = \text{mín}\{p, q\}$ correlaciones canónicas, pero se espera que un número $l < r$ mucho menor sean suficientes para capturar las asociaciones lineales más relevantes.
- Usualmente son de interés las dos o tres primeras correlaciones canónicas, aunque la interpretación siempre es complicada.

Modelo

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

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

- El vector agrupado con los dos grupos de variables es $\mathbf{W} = \begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix}$, entonces sabemos que

$\mathrm{E}(\mathbf{W}) = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}$ y $\mathrm{Var}(\mathbf{W}) = \begin{pmatrix} \boldsymbol{\Sigma}_y & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_x \end{pmatrix}$

Ejemplo. [datos de Hotelling]

Con los datos de Hotelling, $p = q = 2$, $\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ $\mathbf{x} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$.

Aquí $\mathbf{w} = (Y_1, Y_2, X_1, X_2)'$, $\boldsymbol{\Sigma}_y = \begin{pmatrix} 1 & 0.4248 \\ 0.4248 & 1 \end{pmatrix}$, $\boldsymbol{\Sigma}_x = \begin{pmatrix} 1 & 0.6328 \\ 0.6328 & 1 \end{pmatrix}$, $\boldsymbol{\Sigma}_{xy} = \begin{pmatrix} 0.2412 & 0.0586 \\ -0.0553 & 0.0655 \end{pmatrix}$, y $\boldsymbol{\Sigma}_{yx} = \begin{pmatrix} 0.2412 & -0.0553 \\ 0.0586 & 0.0655 \end{pmatrix}$.

$\square$

- Noten que los pq elementos de $\mathbf{\Sigma}_{xy}$ o equivalentemente $\mathbf{\Sigma}_{yx}$, 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.

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

$$\rho(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{xy}\mathbf{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_x\mathbf{a}}\sqrt{\mathbf{b}'\boldsymbol{\Sigma}_y\mathbf{b}}}$$

- Para tener solución única, se impone la restricción de que las variables canónicas tengan varianza unitaria: $\mathbf{a}'\boldsymbol{\Sigma}_x\mathbf{a} = 1$ y $\mathbf{b}'\boldsymbol{\Sigma}_y\mathbf{b} = 1$
- El problema de optimización[1] a resolver, es el siguiente:

$$\max_{\mathbf{a},\mathbf{b}} \rho(\mathbf{a},\mathbf{b}) = (\mathbf{a}'\boldsymbol{\Sigma}_{xy}\mathbf{b})$$

sujeto a:

$$\mathbf{a}'\boldsymbol{\Sigma}_x\mathbf{a} = 1$$
$$\mathbf{b}'\boldsymbol{\Sigma}_y\mathbf{b} = 1$$

[1] Noten 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

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

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

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{a}} &= \boldsymbol{\Sigma}_{xy}\mathbf{b} - \kappa_1\boldsymbol{\Sigma}_x\mathbf{a} = \mathbf{0} \\ \frac{\partial f}{\partial \mathbf{b}} &= \boldsymbol{\Sigma}_{yx}\mathbf{a} - \kappa_2\boldsymbol{\Sigma}_y\mathbf{b} = \mathbf{0} \end{aligned}$$

Se obtienen las dos siguientes ecuaciones:

$$\begin{aligned} \boldsymbol{\Sigma}_{xy}\mathbf{b} &= \kappa_1\boldsymbol{\Sigma}_x\mathbf{a} \quad (1) \\ \boldsymbol{\Sigma}_{yx}\mathbf{a} &= \kappa_2\boldsymbol{\Sigma}_y\mathbf{b} \quad (2) \end{aligned}$$

Multiplicando (1) por $\mathbf{a}'$ y (2) por $\mathbf{b}'$ y aplicando las restricciones obtenemos:

Solución correlaciones canónicas II

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

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

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

Despejando $\mathbf{b}$ de la segunda ecuación, $\mathbf{b} = \kappa_1^{-1}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}\mathbf{a}$ y sustituyendo en la primera ecuación, se obtienen las identidades:

$$\begin{aligned}\boldsymbol{\Sigma}_{xy}(\kappa_1^{-1}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx})\mathbf{a} &= \kappa_1\boldsymbol{\Sigma}_x\mathbf{a}\\ (\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx})\mathbf{a} &= \kappa_1^2\mathbf{a}\end{aligned}$$

Entonces las correlaciones canónicas también son eigenvectores, de la matriz $\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}$.

Solución correlaciones canónicas III

Solución de correlación canónica

- La dirección canónica **a** resulta ser un vector propio de la matriz cuadrada
$$\underset{q\times q}{\mathbf{A}} = \mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}$$
con valor propio $\lambda = \kappa_1^2 = (\mathbf{a}'\mathbf{\Sigma}_{xy}\mathbf{b})^2 = (\mathbf{b}'\mathbf{\Sigma}_{yx}\mathbf{a})^2$.
- Del mismo modo se puede obtener que **b** es el vector propio ligado a κ_2^2 de la matriz
$$\underset{p\times p}{\mathbf{B}} = \mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}$$
- En conclusión, obtenemos que $\rho^* = \lambda$ es el cuadrado de la correlación entre las variables canónicas óptimas, $\mathbf{a}'\mathbf{x}$ y $\mathbf{b}'\mathbf{y}$, y para maximizar tomamos el valor propio λ_1 correspondiente más grande.
- 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 $r = \text{mín}\{p, q\}$ pares, y corresponderán justamente a los primeros r eigenvalores de las matrices que se obtuvieron arriba.
- De hecho, las matrices **A** y **B** definidas anteriormente tienen los mismos valores propios, y son no negativos.

Con los datos de Hotelling:

```
 Sx <- matrix(c(1,0.6328,0.6328,1), nrow = 2)
 Sy <- matrix(c(1,0.4248,0.4248,1), nrow = 2)
Sxy <- matrix(c(0.2412,-0.0553,0.0586,0.0655), nrow=2)
Syx <- matrix(c(0.2412,0.0586,-0.0553,0.0655), nrow=2)
(A <- solve(Sy) %*% Syx %*% solve(Sx) %*% Sxy)

            [,1]        [,2]
[1,]  0.15675924 0.002742911
[2,] -0.06231246 0.003615713

eigen(A) #obten los vectores y valores propios de A

eigen() decomposition
$values
[1] 0.155634923 0.004740029

$vectors
           [,1]        [,2]
[1,]  0.9252848 -0.01804025
[2,] -0.3792729  0.99983726
```

En este ejemplo, $\kappa_1^2 = 0.1556$ y la dirección canónica **a** $= (0.9252848, -0.3792729)$.
Ahora encontramos el valor de **B**:

```
(B <- solve(Sx) %*% Sxy %*% solve(Sy) %*% Syx)

            [,1]        [,2]
[1,]  0.12005273 -0.04361705
[2,] -0.09407052  0.04032222

eigen(B) #obten los vectores y valores propios de B

eigen() decomposition
$values
[1] 0.155634923 0.004740029

$vectors
           [,1]      [,2]
[1,]  0.7748662 0.3537872
[2,] -0.6321252 0.9353259
```

el vector $\mathbf{b} = (0.7748662, -0.6321252)$. Las combinaciones lineales con correlación $\kappa_1 = 0.1556349$ están dadas por:

$$\mathbf{a}'\mathbf{x} = 0.9252848X_1 - 0.3792729X_2 \text{ y } \mathbf{b}'\mathbf{y} = 0.7748662Y_1 - 0.6321252Y_2$$

En este ejemplo no podemos generar la gráfica de las combinaciones lineales porque no tenemos los datos, sólo la matriz de correlaciones.

Subsecuentes correlaciones canónicas I

- Hemos dicho que las matrices **A** y **B** cuyos vectores propios son la solución al problema de correlación canónica, tienen los mismos valores propios y son no negativos. La importancia de esta afirmación es que nos permitirá encontrar r pares de variables canónicas, que cumplirán las siguientes propiedades:
  1. tienen correlación máxima cuando vienen del mismo valor propio.
  2. son no correlacionadas en cada grupo.
  3. son no correlacionadas si corresponden a distintos vectores propios.
- Para comprobar que las matrices **A** y **B** tienen valores propios reales no negativos, se puede utilizar el siguiente lema que prueba que **A** y **B** tienen los mismos valores propios que una matriz semidefinida positiva.

Lema

Sea **L** una matriz definida positiva y **M** una matriz conforme a las dimensiones. Entonces las matrices $\mathbf{L}^{-1}\mathbf{M}$ y $\mathbf{L}^{-1/2}\mathbf{M}\mathbf{L}^{-1/2}$ tienen los mismos valores propios.
Además, si $\mathbf{v}$ es un eigenvector de $\mathbf{L}^{-1}\mathbf{M}$, el vector $\mathbf{w} = \mathbf{L}^{1/2}\mathbf{v}$ es un eigenvector de la segunda.

Demostración.

Subsecuentes correlaciones canónicas II

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:

$$\mathbf{L}^{-1/2}\mathbf{M}\mathbf{v} = \lambda\mathbf{L}^{1/2}\mathbf{v},$$

lo que prueba la segunda afirmación. Por otra parte, podemos escribir:

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

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

$\square$

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

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

  y

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

  tienen los mismos valores propios, porque en el lema podemos hacer $\mathbf{L} = \boldsymbol{\Sigma}_x$ y $\mathbf{M} = \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}$.
- Comprueben entonces que **A** tiene los valores propios de $\mathbf{HH}'$ y que **B** tiene los mismos valores propios de $\mathbf{H}'\mathbf{H}$. Como las matrices son semidefinidas positivas, los eigenvalores de **A** y **B** son reales y no negativos.

- Noten que en el desarrollo previo, nosotros consideramos la solución de correlación canónica sobre las matrices

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

mientras que Johnson y Wichern consideran las alternativas

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

que por el lema anterior, dan los mismos eigenvalores, pero no los mismos eigenvectores. Adicionalmente, $\mathbf{H}'\mathbf{H}$ y $\mathbf{HH}'$ son simétricas, mientras que $\mathbf{A}$ y $\mathbf{B}$ en general no lo son.

Ejemplo. [datos de Hotelling (cont.)]

```
# obtenemos la raiz cuadrada de Sx y Sy:
Sx.eig <- eigen(Sx)
Sx.sqrt <- Sx.eig$vectors %*% diag(sqrt(Sx.eig$values)) %*% solve(Sx.eig$vectors)
Sy.eig <- eigen(Sy)
Sy.sqrt <- Sy.eig$vectors %*% diag(sqrt(Sy.eig$values)) %*% solve(Sy.eig$vectors)
(H <- solve(Sy.sqrt) %*% Syx %*% solve(Sx.sqrt))

            [,1]        [,2]
[1,]  0.33189006 -0.19838322
[2,] -0.03007891  0.09981631

H %*% t(H) # Notar que las matrices son simétricas.

            [,1]        [,2]
[1,]  0.14950691 -0.02978477
[2,] -0.02978477  0.01086804

t(H) %*% H

            [,1]        [,2]
[1,]  0.11105575 -0.06884379
[2,] -0.06884379  0.04931920
```

Ejemplo II

```
eigen(t(H) %*% H)

eigen() decomposition
$values
[1] 0.155634923 0.004740029

$vectors
           [,1]       [,2]
[1,] -0.8393855 -0.5435365
[2,]  0.5435365 -0.8393855

(A <- solve(Sy) %*% Syx %*% solve(Sx) %*% Sxy)

            [,1]        [,2]
[1,]  0.15675924 0.002742911
[2,] -0.06231246 0.003615713

eigen(A) #mismos eigenvalues, pero diferentes eigenvectores...

eigen() decomposition
$values
[1] 0.155634923 0.004740029

$vectors
           [,1]        [,2]
[1,]  0.9252848 -0.01804025
[2,] -0.3792729  0.99983726
```

□

Ortogonalidad de las variables canónicas

- 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,

$$\text{cor}\left(\mathbf{a}_i'\mathbf{x}, \mathbf{a}_j'\mathbf{x}\right) = \mathbf{a}_i'\boldsymbol{\Sigma}_x\mathbf{a}_j = \boldsymbol{\alpha}_i'\boldsymbol{\alpha}_j = \delta_{ij}$$

donde $\delta_{ij} = 1$ si $i = j$ y 0 en otro caso, y del mismo modo

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

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

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

- Entonces las variables canónicas tienen correlación 0 ya sea entre las variables del mismo grupo y también tienen correlación 0 entre grupos.

Invarianza de las variables canónicas

- Si se definen variables $\mathbf{y}^* = \mathbf{U}\mathbf{y} + \mathbf{u}$ y $\mathbf{x}^* = \mathbf{V}\mathbf{x} + \mathbf{v}$, donde $\underset{p\times p}{\mathbf{U}}$ y $\underset{q\times q}{\mathbf{V}}$ son matrices no singulares y $\mathbf{u}, \mathbf{v}$ son vectores fijos. Entonces se cumplen las siguientes condiciones:
  - Las correlaciones canónicas entre $\mathbf{y}^*$ y $\mathbf{x}^*$ son las mismas que para $\mathbf{y}$ y $\mathbf{x}$.
  - Las direcciones canónicas para $\mathbf{y}^*$ y $\mathbf{x}^*$ están dados por $\mathbf{a}_i^* = \mathbf{U}^{-1}\mathbf{a}$ y $\mathbf{b}_i^* = \mathbf{V}^{-1}\mathbf{b}$.
- Por lo anterior, las correlaciones canónicas no cambian con la estandarización, pero la elección de las direcciones canónicas puede no ser única.
- Recuerden por ejemplo, que esta propiedad de invarianza no se cumple en componentes principales.

- 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` y en `python` se puede importar `pyrcca` y usar la función `pyrcca.CCA`. En `Julia` se puede usar el paquete `MultivariateStats.jl` y llamar con `fit(CCA, X, Y; ...)`.

Ejemplos y 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:
  1. gasto en alimento
  2. gasto en ropa
  3. gasto en menaje
  4. gasto en transporte
  5. gasto en educación

  y los últimos 4 se refieren a la estructura del hogar:
  1. numero de personas
  2. numero de personas menores a 14 años
  3. nivel educativo
  4. número de personas que aportan al gasto

  En este ejemplo, $p = 5$ y $q = 4$, y el número máximo de variables canónicas que podemos encontrar es $r = \text{mín}\{4, 5\} = 4$.
- Se quiere asociar el gasto con algún índice demográfico para reducir la dimensión del problema.

- En este caso haremos el ejercicio utilizando las fórmulas y después utilizaremos los paquetes disponibles para hacer comparaciones.

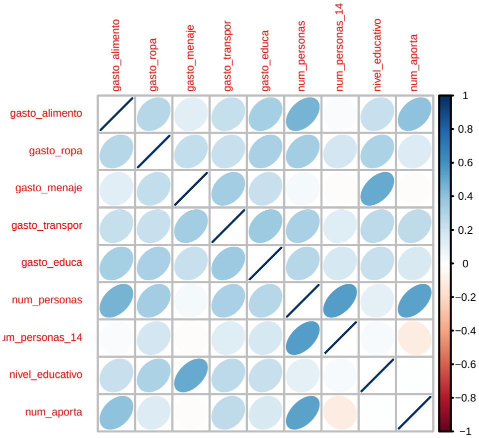
```
options(width=150)
W <- read.table("https://raw.githubusercontent.com/jvega68/EA3/master/datos/hogares.dat",header=T,sep="")
head(W)
```

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

Podemos primero darnos una idea de las correlaciones entre los dos conjuntos de variables

```
corrplot(R, tl.cex = 0.3, cl.cex =0.3, method ="ellipse")
```



Hacemos los cálculos de la matriz **A** (asociada a las variables de estructura, las últimas. Noten que `R11` está asociada a las últimas variables):

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

```
#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
```

- 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

$$\mathbf{a}'\mathbf{y} = 0.78 * \texttt{num\_personas} - 0.27 * \texttt{num\_personas14} + 0.56 * \texttt{nivel\_edu} + 0.07 * \texttt{num\_aporta}$$

  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))
```

```
eigen() decomposition
$values
[1]  4.399732e-01  2.092477e-01  5.364298e-02  1.192928e-02 -2.048037e-18

$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:

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

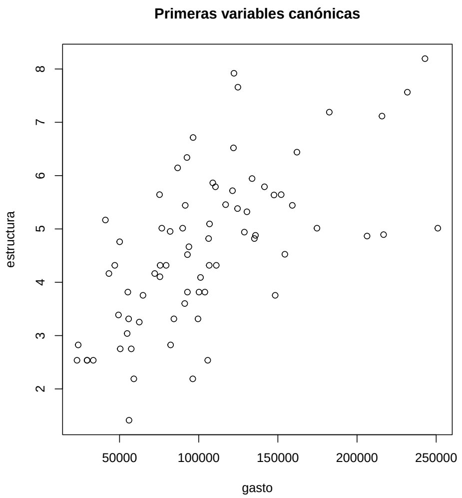
  3. La variabilidad explicada por las primeras variables canónicas es $100 \times \frac{\lambda_1}{\sum_{i=1}^{4} \lambda_i} = 61.55\,\%$.

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 = "gasto", ylab = "estructura", main="Primeras variables canónicas")
```


Primeras variables canónicas
estructura
gasto
2
3
4
5
6
7
8
50000
100000
150000
200000
250000

Datos de hogares con `cancor` 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 = F, ycenter = F) #default es centrar datos, pero no usa datos estandarizados.
#Correlaciones canónicas:
u$cor
```

```
[1] 0.9423165 0.4457288 0.1948461 0.1059006
```

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

```
                       [,1]          [,2]          [,3]        [,4]
num_personas    -0.018411678 -0.0351325365  1.836236e-02  0.10356447
num_personas_14  0.011817520  0.0009024039 -1.090058e-01 -0.09752172
nivel_educativo -0.014504152  0.0671593818  4.707156e-06 -0.02113311
num_aporta      -0.002508877 -0.0338618371  1.096842e-02 -0.14189240
```

```
#coeficientes para gasto
u$ycoef
```

```
                       [,1]          [,2]          [,3]          [,4]          [,5]
gasto_alimento -7.606911e-07 -6.284380e-07  1.065907e-06  2.530369e-07 -3.133897e-07
gasto_ropa     -4.254178e-07  9.978801e-08 -3.331778e-06  3.607008e-06  2.101454e-06
gasto_menaje   -4.551304e-07  4.716540e-06  9.560689e-07 -4.697890e-07 -2.515273e-07
gasto_transpor -3.071238e-07 -1.380674e-06 -8.113091e-07 -2.935941e-06  3.126251e-06
gasto_educa     6.912484e-08 -6.404162e-08 -2.038980e-06 -1.192897e-06 -3.091129e-06
```

- 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,] 0.6447877
```

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

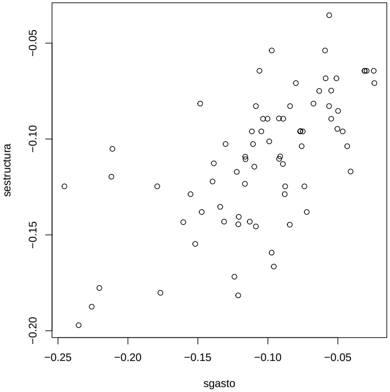
```
            [,1]
[1,] 0.002591279
```

```
var(sestructura)
```

```
          [,1]
[1,] 0.0012354
```

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

Datos de hogares con `cancor` III



- Consideremos 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`}.
- Queremos ver si la habilidad de un estudiante en los exámenes a libro abierto, está correlacionada con su habilidad en exámenes a libro cerrado. Estos conjuntos de variables están relacionados, como se puede ver en la matriz de correlaciones. O podemos tratar de usar los resultados de los exámenes a libro abierto para predecir los resultados da libro cerrado (o viceversa).

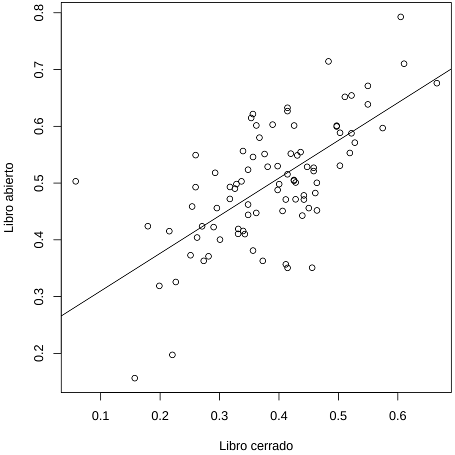
```
E <- read.csv("https://raw.githubusercontent.com/jvega68/EA3/master/datos/score.txt",header=T,sep="",row.names = 1)
colnames(E) <- c("mec","vec","alg","ana","sta")
cor(E)

          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.

- Las segundas direcciones canónicas no son significativas

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

$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

(sqrt(cancor(E[,1:2],E[,3:5])$cor)) # estas son las verdaderas correlaciones canónicas, las raíces cuadradas. El letrero puede ser confuso

[1] 0.8142801 0.2023510

x1 <- as.matrix(E[,1:2]) %*% u$xcoef[,1]
y1 <- as.matrix(E[,3:5]) %*% u$ycoef[,1]
par(pty="s"); plot(x1, y1, xlab = "Libro cerrado", ylab = "Libro abierto")
abline(coef(lm(y1 ~ x1)))
```

Ejemplo 2: scores (Mardia, 1979) III



Ejemplo 2: scores (Mardia, 1979) IV

- Una vez encontradas las variables canónicas, podemos hacer una regresión entre ellas para hacer predicción de una variable a partir de la otra, por ejemplo.

```
summary(lm(y1 ~ x1))  # Modelo para pronóstico

Call:
lm(formula = y1 ~ x1)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.194930 -0.051129 -0.009936  0.051502  0.221081 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.24358    0.03240   7.517 4.96e-11 ***
x1           0.66305    0.08072   8.214 1.95e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08072 on 86 degrees of freedom
Multiple R-squared:  0.4396, Adjusted R-squared:  0.4331 
F-statistic: 67.47 on 1 and 86 DF,  p-value: 1.95e-12
```

- 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 categórica o factor, y el otro conjunto son las 8 variables.
- Como la variable `region` es categórica, 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("https://raw.githubusercontent.com/jvega68/EA3/master/datos/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
```

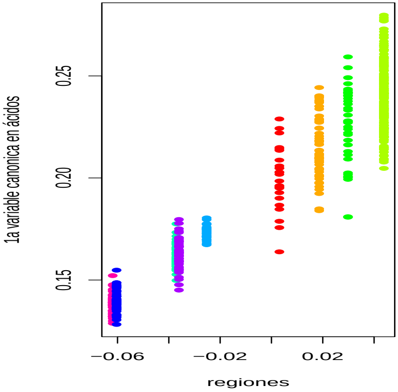
```
as.matind <- function(z) { #crea una matriz de indicadoras, z es categorica
                                z <- as.factor(z)
                                l <- levels(z)
                                b <- as.numeric(z==rep(l,each=length(z)))
                                return(matrix(b,length(z)))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

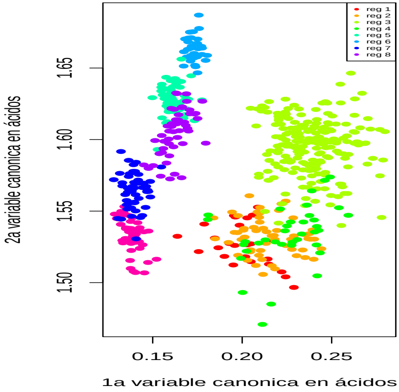
- Una vez establecida la estructura necesaria, ejecutamos la correlación canónica sobre los conjuntos de variables **y** y **x**

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

```
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)
par(mfrow=c(1,2))
plot(regiones[,1],acidos[,1], col = colores[W[,1]], pch=16,
        xlab = "regiones",
        ylab = "1a variable canonica en ácidos")

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



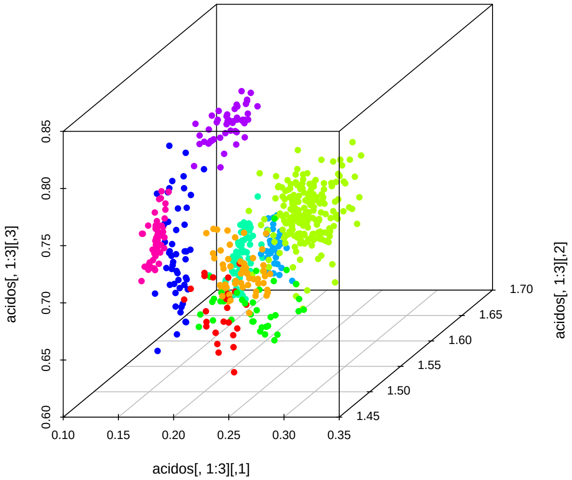


- En este ejemplo, el análisis de correlación canónica nos sirve para hacer una *clasificación* de los datos. El *análisis discriminante lineal* hace exactamente lo mismo que correlación canónica. Más adelante cubriremos este tema.
- Podemos todavía considerar una tercera variable para ver si las direcciones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

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

Opción estática:

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


0.60
0.65
0.70
0.75
0.80
0.85
acidos[, 1:3][,3]
0.10
0.15
0.20
0.25
0.30
0.35
acidos[, 1:3][,1]
1.45
1.50
1.55
1.60
1.65
1.70
acidos[, 1:3][,2]

Inferencia

- Las distribuciones muestrales asociadas con el análisis de correlación canónica son muy complicadas. Para referencia, consultar Kshirsagar (1972 pp. 261-277). Aquí consideraremos sólo un caso de prueba para la matriz de correlaciones.
- Cuando se puede asumir normalidad en la matriz de datos se tiene que $\mathbf{W} \sim \mathcal{N}_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y se puede evaluar si tiene sentido realizar un análisis de correlación canónica, probando primero si $\boldsymbol{\Sigma}_{yx} = \mathbf{0}$.

Prueba para $\boldsymbol{\Sigma}_{xy} = \mathbf{0}$

La prueba de verosimilitud para:

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

rechaza H_0 para valores grandes de la estadística:

$$\lambda = n \log\left(\frac{|\mathbf{S}_x||\mathbf{S}_y|}{|\mathbf{S}|}\right) = -n \log \prod_{i=1}^{r} \left(1 - \hat{\rho}_i^{*2}\right) \underset{n\to\infty}{\sim} \chi^2_{pq}$$

donde $\mathbf{S} = \begin{pmatrix} \mathbf{S}_x & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_y \end{pmatrix}$ 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

$$\begin{aligned} |\mathbf{S}| &= |\mathbf{S}_x||\mathbf{S}_y - \mathbf{S}_{yx}\mathbf{S}_x^{-1}\mathbf{S}_{xy}| && (3) \\ &= |\mathbf{S}_x||\mathbf{S}_y||\mathbf{I} - \mathbf{S}_y^{-1}\mathbf{S}_{yx}\mathbf{S}_x^{-1}\mathbf{S}_{xy}| && (4) \\ & && (5) \end{aligned}$$

Entonces

$$\lambda = -n \log(|\mathbf{I} - \mathbf{S}_y^{-1}\mathbf{S}_{yx}\mathbf{S}_x^{-1}\mathbf{S}_{xy}|) = -n \log(\prod_{i=1}^{r}(1 - \lambda_i^2))$$

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

Ejemplo. [Datos de Hotelling]

En los datos de Hotelling, ya habíamos calculado los coeficientes de correlación canónica, recordando se obtuvo $\rho_1^2 = 0.1556$ y $\rho_2^2 = 0.004740$, tenemos $n = 140$ y $p = q = 2$. Entonces la estadística para la prueba $H_0 : \mathbf{S}_{xy} = \mathbf{0}$ nos da

```r
n <- 140; p <- 2; q <- 2
lambda <- -n*log((1-0.15556)*(1-0.004740))
1- pchisq(lambda, df = p*q) # p-valor
```

```
[1] 6.837605e-05
```

```r
# Con corrección de Bartlett:
m <- n-0.5*(p+q+3)
(lambdaB <- -m*log((1-0.15556)*(1-0.004740)))
```

```
[1] 23.72819
```

```r
1- pchisq(lambdaB, df = p*q)
```

```
[1] 9.054596e-05
```

En ambos casos, se rechaza la hipótesis de que las variables no están relacionadas, aunque la correlación no es tan alta como podría pensarse.

□

- La prueba anterior puede extenderse para hacer el contraste de hipótesis sobre las correlaciones canónicas de la cola:

$$H_0 : \rho^*_{s+1} = \cdots = \rho^*_p \quad \text{vs.} \quad H_a : \rho^*_k > 0 \text{ para al menos una} k \in \{s+1, \dots, p\}$$

- La prueba de verosimilitud LRT es ahora:

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

Observaciones a las pruebas de hipótesis

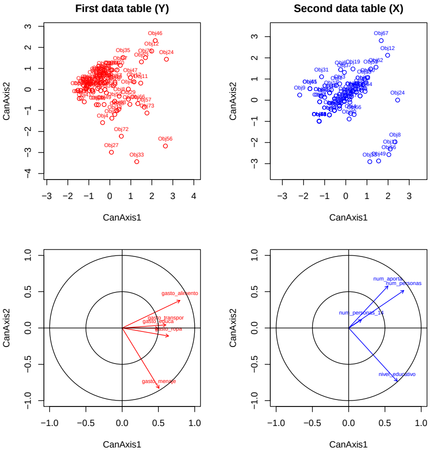
- En la práctica usualmente se aplica esta prueba secuencialmente con pruebas parciales de nivel α, pero de esa manera el nivel de significancia global **no** será α.
- Es importante notar que el análisis de Correlación canónica es sensible a la normalidad de los datos y a la presencia de valores atípicos,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

Ejemplo. [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.

Ejemplo II

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

W <- read.table("https://raw.githubusercontent.com/jvega68/EA3/master/datos/hogares.dat", header = T, sep = "")
cc1 <- CCorA(W[,1:5], W[,6:9])
cc1


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(cc1, cex = c(0.6,0.6), pch = 16)
```



First data table (Y)
Second data table (X)
CanAxis1
CanAxis2
gasto_alimento
gasto_transpor
gasto_educa
gasto_ropa
gasto_menaje
num_aporta
num_personas
num_personas_14
nivel_educativo

Ejemplo IV

```
# Pruebas de significancia para las correlaciones obtenidas:
# forman parte del paquete CCP
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:  0.4399732 10.84168   5  69 1.017949e-07

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