# Tema 5: Análisis Canónico de Poblaciones (MANOVA)

Aurea Grané
Departamento de Estadística
Universidad Carlos III de Madrid



## Muestras de $g$ poblaciones

Sean $\Omega_1, \Omega_2, \ldots, \Omega_g$ $g$ poblaciones sobre las que se han tomado $g$ muestras de tamaños $n_1, n_2, \ldots, n_g$, respectivamente, correspondientes a $p$ variables $X_1, X_2, \ldots, X_p$. La matriz de datos es:

$$\begin{array}{c} \Omega_1 \\ \\ \\ \Omega_2 \\ \\ \\ \vdots \\ \\ \\ \Omega_g \\ \\ \\ \end{array} \left(\begin{array}{cccc} x_{1,11} & x_{1,12} & \ldots & x_{1,1p} \\ \vdots & \vdots & & \vdots \\ x_{1,n_1 1} & x_{1,n_1 2} & \ldots & x_{1,n_1 p} \\ \hline x_{2,11} & x_{2,12} & \ldots & x_{2,1p} \\ \vdots & \vdots & & \vdots \\ x_{2,n_2 1} & x_{2,n_2 2} & \ldots & x_{2,n_2 p} \\ \hline \vdots & \vdots & & \vdots \\ \hline x_{g,11} & x_{g,12} & \ldots & x_{g,1p} \\ \vdots & \vdots & & \vdots \\ x_{g,n_g 1} & x_{g,n_g 2} & \ldots & x_{g,n_g p} \end{array}\right) = \mathbf{X}$$



## Objetivo del análisis canónico de poblaciones

Representar $g$ grupos de individuos (o poblaciones) de forma óptima a lo largo de unos ejes ortogonales, de manera que la dispersión entre estos grupos sea máxima con relación a la dispersión dentro de los grupos.

La represenatción en estos ejes ortogonales (*ejes canónicos* va a permitir estudiar mejor las relaciones entre los distintos grupos (o poblaciones).

En la representación canónica, la distancia euclídea entre dos individuos expresados en función de los ejes canónicos coincide con la distancia de Mahalanobis entre estos individuos expresados en función de las variables originales.



La matriz de datos $\mathbf{X}$ está dividida en $g$ cajas de dimensiones $n_\alpha \times p$, para $\alpha = 1, 2, \ldots, g$, es decir:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_g \end{pmatrix},$$

donde cada $\mathbf{X}_\alpha$ es la submatriz de datos correspondiente a la población $\Omega_\alpha$.
Para cada población $\Omega_\alpha$ ($\alpha = 1, 2, \ldots, g$) podemos calcular:

- el vector de medias muestral, es decir, el *individuo medio*:

$$\overline{\mathbf{x}}_\alpha = \frac{1}{n_\alpha} \mathbf{X}'_\alpha \mathbf{1}_{n_\alpha},$$

donde $\mathbf{1}_{n_\alpha}$ es el vector de unos de dimensión $n_\alpha$.

El vector $\overline{\mathbf{x}}_\alpha$ será el *representante* de la población $\Omega_\alpha$.

- y la matriz de sumas de productos:

$$\mathbf{C}_\alpha = \mathbf{X}'_\alpha \, \mathbf{H}_{n_\alpha} \, \mathbf{X}_\alpha = n_\alpha \, \mathbf{S}_\alpha,$$

donde $\mathbf{H}_{n_\alpha}$ es la matriz de centrado de dimensión $n_\alpha$ y $\mathbf{S}_\alpha$ es la matriz de covarianzas muestral de la submatriz $\mathbf{X}_\alpha$.

La suma de las $g$ matrices $\mathbf{C}_\alpha$ se denomina **matriz de dispersión dentro de los grupos** (*within groups matrix*):

$$\mathbf{W} = \sum_{\alpha=1}^{g} \mathbf{C}_\alpha = \sum_{\alpha=1}^{g} n_\alpha \, \mathbf{S}_\alpha$$

A partir de la matriz $\mathbf{W}$ se obtiene la **matriz de covarianzas ponderada dentro de los grupos** (*pooled within matrix*):

$$\mathbf{S}_P = \frac{1}{n-g} \, \mathbf{W} = \frac{1}{n-g} \sum_{\alpha=1}^{g} n_\alpha \, \mathbf{S}_\alpha = \frac{1}{n-g} \sum_{\alpha=1}^{g} (n_\alpha - 1) \, \tilde{\mathbf{S}}_\alpha$$



Si ahora consideramos las $g$ cajas juntas:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_g \end{pmatrix}$$

tenemos una matriz $n \times p$, donde $n = n_1 + n_2 + \ldots + n_g$, de la que podemos calcular:

- el vector de medias muestral: $\overline{\mathbf{x}} = \frac{1}{n} \, \mathbf{X}' \, \mathbf{1}_n$, que se denomina **vector de medias global**,
- y la matriz de sumas de cuadrados: $\mathbf{T} = \mathbf{X}' \, \mathbf{H}_n \, \mathbf{X}$, que se denomina **matriz total de sumas de cuadrados**.



La dispersión entre las poblaciones viene dada por la matriz

$$\mathbf{B} = \sum_{\alpha=1}^{g} n_\alpha \, (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}}) \, (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})',$$

denominada **matriz de dispersión entre los grupos** (*between groups matrix*).

**Descomposición de la variabilidad total**:
Se verifica que $\mathbf{T} = \mathbf{B} + \mathbf{W}$, es decir,

$$\sum_{\alpha=1}^{g} \sum_{i=1}^{n_\alpha} (\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}) \, (\mathbf{x}_{\alpha i} - \overline{\mathbf{x}})' = \sum_{\alpha=1}^{g} n_\alpha (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}}) \, (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})' + \sum_{\alpha=1}^{g} \sum_{i=1}^{n_\alpha} (\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha) \, (\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha)'.$$



*Demostración:*
Teniendo en cuenta la identidad $\mathbf{x}_{\alpha i} - \overline{\mathbf{x}} = (\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha) + (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})$:

$$\mathbf{T} = \sum_{\alpha=1}^{g} \sum_{i=1}^{n_\alpha} (\mathbf{x}_{\alpha i} - \overline{\mathbf{x}})(\mathbf{x}_{\alpha i} - \overline{\mathbf{x}})' = \sum_{\alpha=1}^{g} \sum_{i=1}^{n_\alpha} [(\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha) + (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})] \, [(\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha) + (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})]'$$

$$\begin{aligned}
&= \sum_{\alpha=1}^{g} \sum_{i=1}^{n_\alpha} (\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha)(\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha)' + \sum_{\alpha=1}^{g} \sum_{i=1}^{n_\alpha} (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})(\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})' \\
&+ \sum_{\alpha=1}^{g} \sum_{i=1}^{n_\alpha} (\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha)(\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})' + \sum_{\alpha=1}^{g} \sum_{i=1}^{n_\alpha} (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})(\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha)' \\
&= \mathbf{W} + \underbrace{\sum_{\alpha=1}^{g} n_\alpha \, (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})(\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})'}_{\mathbf{B}} + 2 \sum_{\alpha=1}^{g} \underbrace{\left[ \sum_{i=1}^{n_\alpha} (\mathbf{x}_{\alpha i} - \overline{\mathbf{x}}_\alpha) \right]}_{n_\alpha \overline{\mathbf{x}}_\alpha - n_\alpha \overline{\mathbf{x}}_\alpha = 0} (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})' \\
&= \mathbf{W} + \mathbf{B}
\end{aligned}$$

En resumen:

Tenemos $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_g$ matrices de datos correspondientes a $g$ muestras de tamaños $n_1, n_2, \ldots, n_g$, provinentes de $g$ poblaciones $\Omega_1, \Omega_2, \ldots, \Omega_g$.

Para cada población $\Omega_\alpha$ calculamos: $\overline{\mathbf{x}}_\alpha$ y $\mathbf{S}_\alpha$ a partir de cada $\mathbf{X}_\alpha$.

Para el total de grupos, calculamos el vector de medias global, $\overline{\mathbf{x}}$, y la matriz $\mathbf{T} = \mathbf{X}' \mathbf{H}_n \mathbf{X}$, que mide la dispersión total de los individuos.

La matriz $\mathbf{W} = \sum_{\alpha=1}^{g} n_\alpha \mathbf{S}_\alpha$ mide la dispersión dentro de los grupos.

La matriz $\mathbf{B} = \sum_{\alpha=1}^{g} n_\alpha (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}}) (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}})'$ mide la dispersión entre los grupos.



## Distancia de Mahalanobis entre dos poblaciones

La versión muestral de la distancia de Mahalanobis entre dos poblaciones $\Omega_\alpha$ y $\Omega_\beta$, para $\alpha, \beta = 1, 2, \ldots, g$, se define como

$$\delta^2_{\alpha,\beta} = (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}}_\beta)' \, \mathbf{S}_P^{-1} \, (\overline{\mathbf{x}}_\alpha - \overline{\mathbf{x}}_\beta),$$

donde $\mathbf{S}_P = \frac{1}{n-g}\mathbf{W}$ es la matriz de covarianzas ponderada, que expresa la dispersión dentro de los grupos.

**Inconveniente:** La representación de las $g$ poblaciones (o de los $g$ grupos) mediante la métrica de Mahalanobis es bastante complicada, porque la dimensión puede ser muy grande y los ejes de representación son oblícuos, es decir, no ortogonales.



## Idea del análisis canónico

Obtener unos ejes de representación ortogonales (*ejes canónicos*), $Y_1, \ldots, Y_k$, que tengan la siguiente propiedad:

$$\delta^2_{\alpha,\beta} = (\overline{\mathbf{y}}_\alpha - \overline{\mathbf{y}}_\beta)' \, (\overline{\mathbf{y}}_\alpha - \overline{\mathbf{y}}_\beta),$$

es decir, que la distancia euclídea entre dos individuos expresados en las nuevas coordenadas coincida con la distancia de Mahalanobis entre estos dos individuos.



## Obtención de las variables canónicas

Puesto que se quiere obtener la máxima dispersión entre grupos con respecto de la dispersión interna de los grupos, se trata de una diagonalización relativa de $\mathbf{B}$ respecto de $\mathbf{S}_P$.

Hay que encontrar los valores propios $\lambda_i$ y los vectores propios $\mathbf{v}_i$ de $\mathbf{B}$ respecto de $\mathbf{S}_P$, es decir:

$$\mathbf{B}\,\mathbf{v}_i = \lambda_i \, \mathbf{S}_P \, \mathbf{v}_i, \quad i = 1, 2, \ldots, p,$$

con la condición de normalización

$$\mathbf{v}_i' \, \mathbf{S}_P \, \mathbf{v}_j = \begin{cases} 1, & \text{si } i = j, \\ 0, & \text{si } i \neq j. \end{cases}$$

En notación matricial:

Sean $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ y $\mathbf{\Lambda} = \mathsf{diag}(\lambda_1, \ldots, \lambda_p)$, con $\lambda_1 > \ldots > \lambda_p$, entonces:

$$\mathbf{B}\,\mathbf{V} = \mathbf{S}_P\,\mathbf{V}\,\mathbf{\Lambda}, \quad \mathbf{V}'\,\mathbf{S}_P\,\mathbf{V} = \mathbf{I}.$$

Se verifica que el número máximo de valores propios no nulos, es decir, la dimensión máxima del espacio canónico, es $m = \min(g-1, p)$.

Se definen las **variables canónicas** como las siguientes combinaciones lineales de las variables originales $X_1, X_2, \ldots, X_p$:

$$Y_j = \mathbf{v}_j' \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}, \quad \text{para } j = 1, \ldots, m,$$



Si $\mathbf{X}$ es la matriz de datos:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}_{n \times p}$$

las coordenadas de los individuos en función de las nuevas variables canónicas son:

$$\mathbf{Y}_{n\times m} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = \mathbf{X}\,\mathbf{V}_{p\times m}$$

Análogamente, los vectores de medias de cada población en las nuevas coordenadas son (expresados en fila):

$$\overline{\mathbf{y}}_\alpha' = \overline{\mathbf{x}}_\alpha'\,\mathbf{V}, \quad \text{para } \alpha = 1, 2, \ldots, g.$$



## Ejemplo 1: Problema 8.1

Se dispone de las medidas de cinco variables biométricas sobre gorriones hembra, recogidos casi moribundos después de una tormenta. Los primeros 21 sobrevivieron mientras que los 28 restantes no lo consiguieron.

$X_1$ = longitud total
$X_2$ = extensión del ala
$X_3$ = longitud del pico y de la cabeza
$X_4$ = longitud del húmero
$X_5$ = longitud del esternón



| gorriones supervivientes | | | | | gorriones no supervivientes | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
| 156 | 245 | 31.6 | 18.5 | 20.5 | 155 | 240 | 31.4 | 18.0 | 20.7 |
| 154 | 240 | 30.4 | 17.9 | 19.6 | 156 | 240 | 31.5 | 18.2 | 20.6 |
| 153 | 240 | 31.0 | 18.4 | 20.6 | 160 | 242 | 32.6 | 18.8 | 21.7 |
| 153 | 236 | 30.9 | 17.7 | 20.2 | 152 | 232 | 30.3 | 17.2 | 19.8 |
| 155 | 243 | 31.5 | 18.6 | 20.3 | 160 | 250 | 31.7 | 18.8 | 22.5 |
| 163 | 247 | 32.0 | 19.0 | 20.9 | 155 | 237 | 31.0 | 18.5 | 20.0 |
| 157 | 238 | 30.9 | 18.4 | 20.2 | 157 | 245 | 32.2 | 19.5 | 21.4 |
| 155 | 239 | 32.8 | 18.6 | 21.2 | 165 | 245 | 33.1 | 19.8 | 22.7 |
| 164 | 248 | 32.7 | 19.1 | 21.1 | 153 | 231 | 30.1 | 17.3 | 19.8 |
| 158 | 238 | 31.0 | 18.8 | 22.0 | 162 | 239 | 30.3 | 18.0 | 23.1 |
| 158 | 240 | 31.3 | 18.6 | 22.0 | 162 | 243 | 31.6 | 18.8 | 21.3 |
| 160 | 244 | 31.1 | 18.6 | 20.5 | 159 | 245 | 31.8 | 18.5 | 21.7 |
| 161 | 246 | 32.3 | 19.3 | 21.8 | 159 | 247 | 30.9 | 18.1 | 19.0 |
| 157 | 245 | 32.0 | 19.1 | 20.0 | 155 | 243 | 30.9 | 18.5 | 21.3 |
| 157 | 235 | 31.5 | 18.1 | 19.8 | 162 | 252 | 31.9 | 19.1 | 22.2 |
| 156 | 237 | 30.9 | 18.0 | 20.3 | 152 | 230 | 30.4 | 17.3 | 18.6 |
| 158 | 244 | 31.4 | 18.5 | 21.6 | 159 | 242 | 30.8 | 18.2 | 20.5 |
| 153 | 238 | 30.5 | 18.2 | 20.9 | 155 | 238 | 31.2 | 17.9 | 19.3 |
| 155 | 236 | 30.3 | 18.5 | 20.1 | 163 | 249 | 33.4 | 19.5 | 22.8 |
| 163 | 246 | 32.5 | 18.6 | 21.9 | 163 | 242 | 31.0 | 18.1 | 20.7 |
| 159 | 236 | 31.5 | 18.0 | 21.5 | 156 | 237 | 31.7 | 18.2 | 20.3 |
| | | | | | 159 | 238 | 31.5 | 18.4 | 20.3 |
| | | | | | 161 | 245 | 32.1 | 19.1 | 20.8 |
| | | | | | 155 | 235 | 30.7 | 17.7 | 19.6 |
| | | | | | 162 | 247 | 31.9 | 19.1 | 20.4 |
| | | | | | 153 | 237 | 30.6 | 18.6 | 20.4 |
| | | | | | 162 | 245 | 32.5 | 18.5 | 21.1 |
| | | | | | 164 | 248 | 32.3 | 18.8 | 20.9 |

a) Construir las matrices de dispersión dentro de los grupos, $\mathbf{W}$, y de dispersión entre los grupos, $\mathbf{B}$.

b) Encontrar el primer eje canónico y estandarizarlo respecto de la matriz de covarianzas común.

c) Obtener las coordenadas de los individuos medios en función del primer eje canónico estandarizado.

d) Comprobar que la distancia euclídea entre los individuos medios expresados en las coordenadas canónicas coincide con la distancia de Mahalanobis entre los individuos medios expresados en las variables originales.



a) Llamamos $\mathbf{X}$ e $\mathbf{Y}$ a las matrices de datos del grupo de supervivientes y del de no supervivientes, respectivamente. Los tamaños muestrales de cada grupo son: $n_X = 21$, $n_Y = 28$. Los individuos medios de cada grupo (vectores de medias) son:

$$\mathbf{m}_X = \frac{1}{n_X}\mathbf{X}'\mathbf{1}_{n_X} = \begin{pmatrix} 157.3810 \\ 241.0000 \\ 31.4333 \\ 18.5000 \\ 20.8095 \end{pmatrix} \qquad \mathbf{m}_Y = \frac{1}{n_Y}\mathbf{Y}'\mathbf{1}_{n_Y} = \begin{pmatrix} 158.4286 \\ 241.5714 \\ 31.4786 \\ 18.4464 \\ 20.8393 \end{pmatrix}$$

y el vector de medias global (o centroide) es

$$\begin{aligned} \mathbf{m} &= (n_X\,\mathbf{m}_X + n_Y\,\mathbf{m}_Y)/(n_X + n_Y) \\ &= (157.9796, 241.3265, 31.4592, 18.4694, 20.8265)' \end{aligned}$$



Las matrices de covarianzas de cada grupo son:

$$\mathbf{S}_X = \frac{1}{n_X}\mathbf{X}'\mathbf{H}_{n_X}\mathbf{X} = \begin{pmatrix} 10.5215 & 8.6667 & 1.4825 & 0.8286 & 1.2249 \\ 8.6667 & 16.6667 & 1.8190 & 1.2476 & 0.8381 \\ 1.4825 & 1.8190 & 0.5060 & 0.1800 & 0.2283 \\ 0.8286 & 1.2476 & 0.1800 & 0.1676 & 0.1262 \\ 1.2249 & 0.8381 & 0.2283 & 0.1262 & 0.5475 \end{pmatrix}$$

$$\mathbf{S}_Y = \frac{1}{n_Y}\mathbf{Y}'\mathbf{H}_{n_Y}\mathbf{Y} = \begin{pmatrix} 14.5306 & 16.5765 & 2.1628 & 1.6837 & 2.8260 \\ 16.5765 & 31.3878 & 3.2765 & 2.8449 & 3.9204 \\ 2.1628 & 3.2765 & 0.7024 & 0.4528 & 0.5391 \\ 1.6837 & 2.8449 & 0.4528 & 0.4189 & 0.4878 \\ 2.8260 & 3.9204 & 0.5391 & 0.4878 & 1.2738 \end{pmatrix}$$



La matriz de dispersión dentro de los grupos es

$$\mathbf{W} = n_X\,\mathbf{S}_X + n_Y\,\mathbf{S}_Y = 10^3 \begin{pmatrix} 0.6278 & 0.6461 & 0.0917 & 0.0645 & 0.1049 \\ 0.6461 & 1.2289 & 0.1299 & 0.1059 & 0.1274 \\ 0.0917 & 0.1299 & 0.0303 & 0.0165 & 0.0199 \\ 0.0645 & 0.1059 & 0.0165 & 0.0152 & 0.0163 \\ 0.1049 & 0.1274 & 0.0199 & 0.0163 & 0.0472 \end{pmatrix}.$$

y la matriz de dispersión entre los grupos es

$$\begin{aligned} \mathbf{B} &= n_X\,(\mathbf{m}_X - \mathbf{m})\,(\mathbf{m}_X - \mathbf{m})' + n_Y\,(\mathbf{m}_Y - \mathbf{m})\,(\mathbf{m}_Y - \mathbf{m})' \\ &= \begin{pmatrix} 13.1696 & 7.1832 & 0.5695 & -0.6738 & 0.3746 \\ 7.1832 & 3.9180 & 0.3106 & -0.3675 & 0.2043 \\ 0.5695 & 0.3106 & 0.0246 & -0.0291 & 0.0162 \\ -0.6738 & -0.3675 & -0.0291 & 0.0345 & -0.0192 \\ 0.3746 & 0.2043 & 0.0162 & -0.0192 & 0.0107 \end{pmatrix}. \end{aligned}$$

b) La matriz de covarianzas común es

$$\mathbf{S}_P = \frac{1}{n_X + n_Y - 2}\,\mathbf{W} = \begin{pmatrix} 13.3576 & 13.7477 & 1.9509 & 1.3733 & 2.2309 \\ 13.7477 & 26.1459 & 2.7647 & 2.2523 & 2.7100 \\ 1.9509 & 2.7647 & 0.6445 & 0.3502 & 0.4232 \\ 1.3733 & 2.2523 & 0.3502 & 0.3244 & 0.3470 \\ 2.2309 & 2.7100 & 0.4232 & 0.3470 & 1.0035 \end{pmatrix}.$$

Los ejes canónicos se obtienen a partir de la diagonalización de $\mathbf{B}$ respecto de $\mathbf{S}_P$:

```
[V,L] = eig(B,Sp);
diag(L) = 2.8248   -0.0000   0.0000  -0.0000   0.0000
V(:,1)' = -0.3201  -0.0546  -0.1924   2.1298   0.1426
```

Solamente obtenemos un eje canónico, puesto que sólo hay un valor propio relativo no nulo ($\min(p, g-1) = \min(5,1) = 1$). Si llamamos $\mathbf{v}$ a este eje canónico, puede comprobarse que $\mathbf{v}'\,\mathbf{S}_P\,\mathbf{v} = 1$.



$Y_1 = -0.3201\,X_1 - 0.0546\,X_2 - 0.1924\,X_3 + 2.1298\,X_4 + 0.1426\,X_5$ contrapone medidas de longitud (signo negativo) frente a medidas de robustez (signo positivo). A la derecha del eje, se sitúan los gorriones más robustos, mientras que a la izquierda están los más estilizados.

c) Las coordenadas de los individuos medios en función del primer eje canónico estandarizado son

$$\widetilde{\mathbf{m}}_X = \mathbf{m}_X'\,\mathbf{v} = -27.2238, \quad \widetilde{\mathbf{m}}_Y = \mathbf{m}_Y'\,\mathbf{v} = -27.7090.$$

Observad que el grupo de supervivientes corresponde a los gorriones más robustos.

d) La distancia de Mahalanobis entre los individuos medios es

$$(\mathbf{m}_X - \mathbf{m}_Y)'\,\mathbf{S}_P^{-1}\,(\mathbf{m}_X - \mathbf{m}_Y) = 0.2354,$$

y la distancia euclídea entre los individuos medios en función de las coordenadas canónicas es $(\widetilde{\mathbf{m}}_X - \widetilde{\mathbf{m}}_Y)^2 = 0.2354$.



## Porcentaje de variabilidad explicada

El porcentaje de variabilidad explicada por los $q$ primeros ejes canónicos, es

$$P_q = \frac{\lambda_1 + \ldots + \lambda_q}{\lambda_1 + \ldots + \lambda_m} \times 100\%, \quad q \leq m.$$

## Aspectos inferenciales

Supongamos que las $g$ matrices de datos $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_g$ provienen de $g$ poblaciones normales $N_p(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$ para $\alpha = 1, 2, \ldots, g$.

Para poder aplicar correctamente el análisis canónico de poblaciones es conveniente que los vectores de medias $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_g$ sean diferentes y que las matrices de covarianzas $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_g$ sean iguales.



## Comparación de medias (suponiendo covarianzas iguales)

$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g$

Consideremos las matrices $\mathbf{B}$, $\mathbf{W}$ y $\mathbf{T}$.

Si $H_0$ es cierta, entonces $\mathbf{B} \sim W_p(\boldsymbol{\Sigma}, g-1)$, $\mathbf{W} \sim W_p(\boldsymbol{\Sigma}, n-g)$, $\mathbf{T} \sim W_p(\boldsymbol{\Sigma}, n-1)$ y además $\mathbf{B}$ y $\mathbf{W}$ son independientes.

El estadístico de contraste

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|} \sim \Lambda(p, n-g, g-1) \text{ Lambda de Wilks.}$$

No rechazaremos $H_0$ cuando el valor del estadestadístico $\Lambda$ sea pequeño y significativo.

Interpretación: cuando $H_0$ es cierta, las $g$ poblaciones se confunden en una sola y la representación canónica se reduce a un único punto, exceptuando las diferencias en el muestreo.

**Comparación de covarianzas**

$H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g$

Se resuelve mediante el test de la razón de verosimilitudes:

$$\lambda_R = \frac{|\mathbf{S}_1|^{n_1/2}\,|\mathbf{S}_2|^{n_2/2}\,\ldots\,|\mathbf{S}_g|^{n_g/2}}{|\mathbf{S}|^{n/2}},$$

donde $\mathbf{S}_\alpha$ es la matriz de covarianzas muestral de la población $\Omega_\alpha$ y $\mathbf{S} = \frac{1}{n}\,\mathbf{W}$.

Rechazaremos $H_0$ si el estadístico

$$-2\log\lambda_R = n\,\log|\mathbf{S}| - \sum_{\alpha=1}^{g} n_\alpha\,\log|\mathbf{S}_\alpha| \sim \chi^2_q,$$

donde $q = (g-1)p(p+1)/2$, es significativo.



Interpretación: cuando $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g$ no es cierta, los elipsoides de concentración de los distintos grupos están orientados de forma distinta y, por tanto, no se pueden determinar unos ejes comunes de representación.

Esta hipótesis de igualdad de covarianzas raramente se cumple en las aplicaciones. A pesar de ello, si los signos de los elementos de las matrices de covarianzas muestrales de cada grupo no cambian de un grupo a otro, la orientación de los elipsoides no es demasiado distinta y todavía es posible realizar este análisis.



**Regiones de confianza para los individuos medios**

Cuando el número de individuos en cada grupo es muy grande o bien cuando se dispone de muchos grupos, la representación de individuos da lugar a un gráfico poco claro. En estos casos es habitual representar los grupos mediante su *individuo medio* y construir una región de confianza alrededor de él (para ello, hay que suponer normalidad multivariante con la misma matriz de covarianzas para todos los grupos).

Fijado un nivel de significación $\varepsilon > 0$, los radios de las esferas multidimensionales que definen estas regiones de confianza son $R_\varepsilon/\sqrt{n_\alpha}$, donde

$$R^2_\varepsilon = F_\epsilon\,\frac{(n-g)\,p}{n-g-p+1},$$

y $F_\epsilon$ es el percentil $(1-\varepsilon)\,100\%$ de la ley $F$ de Fisher con $p$ y $n-g-p+1$ grados de libertad.



**Ejemplo 2: Problema 8.2**

Se dispone de cuatro variables numéricas $X_1$ =longitud del sépalo, $X_2$ =anchura del sépalo, $X_3$ =longitud del pétalo, $X_4$ =anchura del pétalo medidas sobre tres especies de flores del género Iris: *Iris setosa*, *Iris versicolor* e *Iris virginica* (Fuente: Fisher 1936).

*Iris setosa*

*Iris versicolor*

*Iris virginica*

| *I. setosa* | | | | *I. versicolor* | | | | *I. virginica* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| 5.1 | 3.5 | 1.4 | 0.2 | 7.0 | 3.2 | 4.7 | 1.4 | 6.3 | 3.3 | 6.0 | 2.5 |
| 4.9 | 3.0 | 1.4 | 0.2 | 6.4 | 3.2 | 4.5 | 1.5 | 5.8 | 2.7 | 5.1 | 1.9 |
| 4.7 | 3.2 | 1.3 | 0.2 | 6.9 | 3.1 | 4.9 | 1.5 | 7.1 | 3.0 | 5.9 | 2.1 |
| 4.6 | 3.1 | 1.5 | 0.2 | 5.5 | 2.3 | 4.0 | 1.3 | 6.3 | 2.9 | 5.6 | 1.8 |
| 5.0 | 3.6 | 1.4 | 0.2 | 6.5 | 2.8 | 4.6 | 1.5 | 6.5 | 3.0 | 5.8 | 2.2 |
| 5.4 | 3.9 | 1.7 | 0.4 | 5.7 | 2.8 | 4.5 | 1.3 | 7.6 | 3.0 | 6.6 | 2.1 |
| 4.6 | 3.4 | 1.4 | 0.3 | 6.3 | 3.3 | 4.7 | 1.6 | 4.9 | 2.5 | 4.5 | 1.7 |
| 5.0 | 3.4 | 1.5 | 0.2 | 4.9 | 2.4 | 3.3 | 1.0 | 7.3 | 2.9 | 6.3 | 1.8 |
| 4.4 | 2.9 | 1.4 | 0.2 | 6.6 | 2.9 | 4.6 | 1.3 | 6.7 | 2.5 | 5.8 | 1.8 |
| 4.9 | 3.1 | 1.5 | 0.1 | 5.2 | 2.7 | 3.9 | 1.4 | 7.2 | 3.6 | 6.1 | 2.5 |
| 5.4 | 3.7 | 1.5 | 0.2 | 5.0 | 2.0 | 3.5 | 1.0 | 6.5 | 3.2 | 5.1 | 2.0 |
| 4.8 | 3.4 | 1.6 | 0.2 | 5.9 | 3.0 | 4.2 | 1.5 | 6.4 | 2.7 | 5.3 | 1.9 |
| 4.8 | 3.0 | 1.4 | 0.1 | 6.0 | 2.2 | 4.0 | 1.0 | 6.8 | 3.0 | 5.5 | 2.1 |
| 4.3 | 3.0 | 1.1 | 0.1 | 6.1 | 2.9 | 4.7 | 1.4 | 5.7 | 2.5 | 5.0 | 2.0 |
| 5.8 | 4.0 | 1.2 | 0.2 | 5.6 | 2.9 | 3.6 | 1.3 | 5.8 | 2.8 | 5.1 | 2.4 |
| 5.7 | 4.4 | 1.5 | 0.4 | 6.7 | 3.1 | 4.4 | 1.4 | 6.4 | 3.2 | 5.3 | 2.3 |
| | ... | ... | | | ... | ... | | | ... | ... | |
| 4.6 | 3.2 | 1.4 | 0.2 | 6.2 | 2.9 | 4.3 | 1.3 | 6.5 | 3.0 | 5.2 | 2.0 |
| 5.3 | 3.7 | 1.5 | 0.2 | 5.1 | 2.5 | 3.0 | 1.1 | 6.2 | 3.4 | 5.4 | 2.3 |
| 5.0 | 3.3 | 1.4 | 0.2 | 5.7 | 2.8 | 4.1 | 1.3 | 5.9 | 3.0 | 5.1 | 1.8 |

a) Realizar la representación canónica de las tres especies, especificando los porcentajes de variabilidad explicados por cada eje canónico.
b) Suponiendo normalidad multivariante, construir las regiones confidenciales para los individuos medios de cada grupo.



Para poder utilizar la función de Matlab `canp.m` debemos escribir los datos de la tabla anterior en una matriz `X=[X1;X2;X3]` de dimensión $N \times p$, donde $p$ es el número de variables observadas, y $N$ es el número total de individuos. En este caso $p = 4$ y $N = 150$. Las matrices `X1`, `X2` y `X3` contienen a los individuos de cada uno de los tres grupos. Debemos introducir también un vector `n=[n1 n2 n3]` que contenga el número de individuos de cada grupo. Consideramos como grupo 1 a la especie *Iris setosa*, como grupo 2 a la especie *Iris versicolor* y como grupo 3 a la especie *Iris virginica*.

```
n = [50 50 50];
[mY,V,B,W,percent,Test1,texto1,Test2,texto2] = canp(X,n)
```



El número de ejes canónicos es $\min(p, g-1) = \min(4, 2) = 2$. Por tanto, la representación en dimensión 2 expresa el 100% de la variabilidad. Las coordenadas de los individuos medios en función de los ejes canónicos son:

```
mY =[-5.5025   -6.8766
      3.9302   -5.9336
      7.8877   -7.1742]
```

Las columnas de `V` son los coeficientes que definen los ejes canónicos:

```
V =[-0.8294   -0.0241
    -1.5345   -2.1645
     2.2012    0.9319
     2.8105   -2.8392]
```

$$Y_1 = -0.8294\, X_1 - 1.5345\, X_2 + 2.2012\, X_3 + 2.8105\, X_4, \quad \rightarrow 99.1213\%$$
$$Y_2 = -0.0241\, X_1 - 2.1645\, X_2 + 0.9319\, X_3 - 2.8392\, X_4, \quad \rightarrow 0.8787\%$$



`B` y `W` son las matrices de dispersión entre grupos (*between*) y de dispersión dentro de los grupos (*within*), respectivamente:

```
B =[63.2121  -19.9527  165.2484   71.2793
   -19.9527   11.3449  -57.2396  -22.9327
   165.2484  -57.2396  437.1028  186.7740
    71.2793  -22.9327  186.7740   80.4133]

W =[38.9562   13.6300   24.6246    5.6450
    13.6300   16.9620    8.1208    4.8084
    24.6246    8.1208   27.2226    6.2718
     5.6450    4.8084    6.2718    6.1566]
```

Contraste de comparación de medias:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{B}|} \sim \Lambda(4, 147, 2) \approx F(8, 288) = 199.1453$$

Se infiere que las medias son distintas.

Contraste de comparación de covarianzas (test de Bartlett):

$$-2\log\lambda_R = 150\,\log|\mathbf{S}| - \sum_{\alpha=1}^{3} 50\,\log|\mathbf{S}_\alpha| = 149.6564 \sim \chi^2_{20}$$

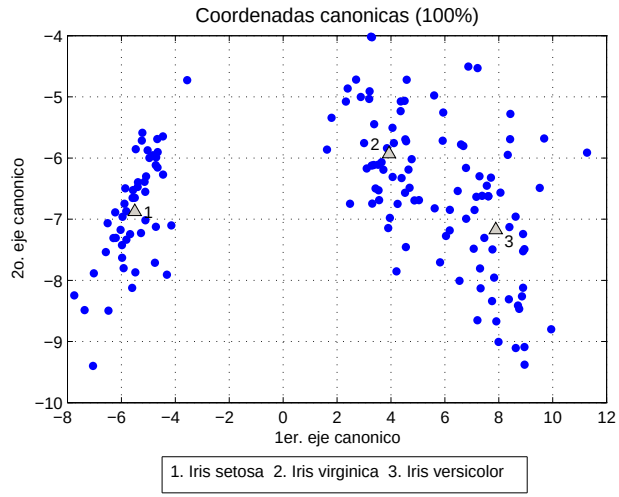
donde $\mathbf{S} = \frac{1}{150}\,\mathbf{W}$. Se rechaza la hipótesis nula de igualdad de covarianzas (!)

Observemos cómo son las matrices $\mathbf{S}_1$, $\mathbf{S}_2$ y $\mathbf{S}_3$:

```
S1=[0.1218  0.0972  0.0160  0.0101     S2=[0.2611  0.0835  0.1792  0.0547
    0.0972  0.1408  0.0115  0.0091         0.0835  0.0965  0.0810  0.0404
    0.0160  0.0115  0.0296  0.0059         0.1792  0.0810  0.2164  0.0716
    0.0101  0.0091  0.0059  0.0109]        0.0547  0.0404  0.0716  0.0383]

S3=[0.3963  0.0919  0.2972  0.0481
    0.0919  0.1019  0.0700  0.0467
    0.2972  0.0700  0.2985  0.0478
    0.0481  0.0467  0.0478  0.0739]
```








Bajo el supuesto de normalidad multivariante y en el caso de la representación en dimensión 2, las regiones confidenciales son círculos centrados en los individuos medios y de radio

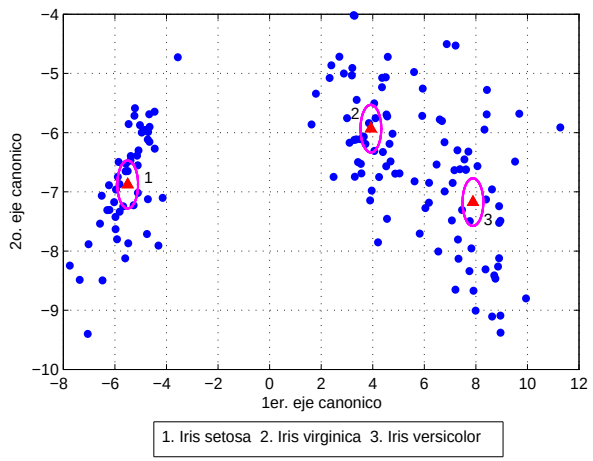
$$r_i = \sqrt{F_\varepsilon\,\frac{(150-3)\,4}{(150-3-4+1)\,50}}, \quad \text{para } i = 1,2,3,$$

donde $F_\varepsilon$ es el percentil $(1-\varepsilon)\,100\%$ de la ley $F$ de Fisher con 4 y $150-3-4+1$ grados de libertad.

Para representar las regiones confidenciales al $(1-\varepsilon)\,100\%$ para los individuos medios utilizaremos la función de Matlab `regconf.m`. Por ejemplo, para un nivel de confianza del 90%, obtenemos:

```
r = regconf(mY,n,4,0.90)
r =[0.4026    0.4026    0.4026]
```

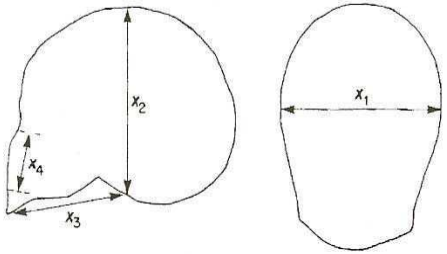





$Y_1 = -0.8294\,X_1 - 1.5345\,X_2 + 2.2012\,X_3 + 2.8105\,X_4$ ordena las 3 especies de flores según el tamaño de sus sépalos y pétalos: pétalos grandes y sépalos pequeños a la derecha del eje, frente a pétalos pequeños y sépalos grandes a la izquierda del eje.

## Ejemplo 3: Problema 8.3

Se dispone de cuatro medidas sobre cráneos de varones egipcios de cinco períodos históricos distintos (*Grupo 1*: 4000 aC, *Grupo 2*: 3300 aC, *Grupo 3*: 1850 aC, *Grupo 4*: 200 aC, *Grupo 5*: 150 dC). Para cada período temporal se midieron 30 cráneos. Las variables observadas son: $X_1$ =anchura máxima, $X_2$ =altura basibregmática, $X_3$ =longitud basialveolar, $X_4$ =longitud de la nariz.
(Fuente: http://lib.stat.cmu.edu/DASL/DataArchive.html).






| 4000 a.C. | | | | 3300 a.C. | | | | 1850 a.C. | | | | 200 a.C. | | | | 150 d.C. | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| 131 | 138 | 89 | 49 | 124 | 138 | 101 | 48 | 137 | 141 | 96 | 52 | 137 | 134 | 107 | 54 | 137 | 123 | 91 | 50 |
| 125 | 131 | 92 | 48 | 133 | 134 | 97 | 48 | 129 | 133 | 93 | 47 | 141 | 128 | 95 | 53 | 136 | 131 | 95 | 49 |
| 131 | 132 | 99 | 50 | 138 | 134 | 98 | 45 | 132 | 138 | 87 | 48 | 141 | 130 | 87 | 49 | 128 | 126 | 91 | 57 |
| 119 | 132 | 96 | 44 | 148 | 129 | 104 | 51 | 130 | 134 | 106 | 50 | 135 | 131 | 99 | 51 | 130 | 134 | 92 | 52 |
| 136 | 143 | 100 | 54 | 126 | 124 | 95 | 45 | 134 | 134 | 96 | 45 | 133 | 120 | 91 | 46 | 138 | 127 | 86 | 47 |
| 138 | 137 | 89 | 56 | 135 | 136 | 98 | 52 | 140 | 133 | 98 | 50 | 131 | 135 | 90 | 50 | 126 | 138 | 101 | 52 |
| 139 | 130 | 108 | 48 | 132 | 145 | 100 | 54 | 138 | 138 | 95 | 47 | 140 | 137 | 94 | 60 | 136 | 138 | 97 | 58 |
| 125 | 136 | 93 | 48 | 133 | 130 | 102 | 48 | 136 | 145 | 99 | 55 | 139 | 130 | 90 | 48 | 126 | 126 | 92 | 45 |
| 131 | 134 | 102 | 51 | 131 | 134 | 96 | 50 | 136 | 131 | 92 | 46 | 140 | 134 | 90 | 51 | 132 | 132 | 99 | 55 |
| 134 | 134 | 99 | 51 | 133 | 125 | 94 | 46 | 126 | 136 | 95 | 56 | 138 | 140 | 100 | 52 | 139 | 135 | 92 | 54 |
| 129 | 138 | 95 | 50 | 133 | 136 | 103 | 53 | 137 | 129 | 100 | 53 | 132 | 133 | 90 | 53 | 143 | 120 | 95 | 51 |
| 134 | 121 | 95 | 53 | 131 | 139 | 98 | 51 | 137 | 139 | 97 | 50 | 134 | 134 | 97 | 54 | 141 | 136 | 101 | 54 |
| 126 | 129 | 109 | 51 | 131 | 136 | 99 | 56 | 136 | 126 | 101 | 50 | 135 | 135 | 99 | 50 | 135 | 135 | 95 | 56 |
| 132 | 136 | 100 | 50 | 138 | 134 | 98 | 49 | 137 | 133 | 90 | 49 | 133 | 136 | 95 | 52 | 137 | 134 | 93 | 53 |
| 141 | 140 | 100 | 51 | 130 | 136 | 104 | 53 | 129 | 142 | 104 | 47 | 136 | 130 | 99 | 55 | 142 | 135 | 96 | 52 |
| 131 | 134 | 97 | 54 | 131 | 128 | 98 | 45 | 135 | 138 | 102 | 55 | 134 | 137 | 93 | 52 | 139 | 134 | 95 | 47 |
| 135 | 137 | 103 | 50 | 138 | 129 | 107 | 53 | 129 | 135 | 92 | 50 | 131 | 141 | 99 | 55 | 138 | 125 | 99 | 51 |
| 132 | 133 | 93 | 53 | 123 | 131 | 101 | 51 | 134 | 125 | 90 | 60 | 129 | 135 | 95 | 47 | 137 | 135 | 96 | 54 |
| 139 | 136 | 96 | 50 | 130 | 129 | 105 | 47 | 138 | 134 | 96 | 51 | 136 | 128 | 93 | 54 | 133 | 125 | 92 | 50 |
| 132 | 131 | 101 | 49 | 134 | 130 | 93 | 54 | 136 | 135 | 94 | 53 | 131 | 125 | 88 | 48 | 145 | 129 | 89 | 47 |
| 126 | 133 | 102 | 51 | 137 | 136 | 106 | 49 | 132 | 130 | 91 | 52 | 139 | 130 | 94 | 53 | 138 | 136 | 92 | 46 |
| 135 | 135 | 103 | 47 | 126 | 131 | 100 | 48 | 133 | 131 | 100 | 50 | 144 | 124 | 86 | 50 | 131 | 129 | 97 | 44 |
| 134 | 124 | 93 | 53 | 135 | 136 | 97 | 52 | 138 | 137 | 94 | 51 | 141 | 131 | 97 | 53 | 143 | 126 | 88 | 54 |
| 128 | 134 | 103 | 50 | 129 | 126 | 91 | 50 | 130 | 127 | 99 | 45 | 130 | 131 | 98 | 53 | 134 | 124 | 91 | 55 |
| 130 | 130 | 104 | 49 | 134 | 139 | 101 | 49 | 136 | 133 | 91 | 49 | 133 | 128 | 92 | 51 | 132 | 127 | 97 | 52 |
| 138 | 135 | 100 | 55 | 131 | 134 | 90 | 53 | 134 | 123 | 95 | 52 | 138 | 126 | 97 | 54 | 137 | 125 | 85 | 57 |
| 128 | 132 | 93 | 53 | 132 | 130 | 104 | 50 | 136 | 137 | 101 | 54 | 131 | 142 | 95 | 53 | 129 | 128 | 81 | 52 |
| 127 | 129 | 106 | 48 | 130 | 132 | 93 | 52 | 133 | 131 | 96 | 49 | 136 | 138 | 94 | 55 | 140 | 135 | 103 | 48 |
| 131 | 136 | 114 | 54 | 135 | 132 | 98 | 54 | 138 | 133 | 100 | 55 | 132 | 136 | 92 | 52 | 147 | 129 | 87 | 48 |
| 124 | 138 | 101 | 46 | 130 | 128 | 101 | 51 | 138 | 133 | 91 | 46 | 135 | 130 | 100 | 51 | 136 | 133 | 97 | 51 |



a) Realizar la representación canónica de los cinco grupos, especificando los porcentajes de variabilidad explicados por los ejes canónicos.
b) Representar las regiones confidenciales para un nivel de confianza del 90%.
c) Interpretar el primer eje canónico.
d) Obtener la matriz de distancias entre los cinco grupos.



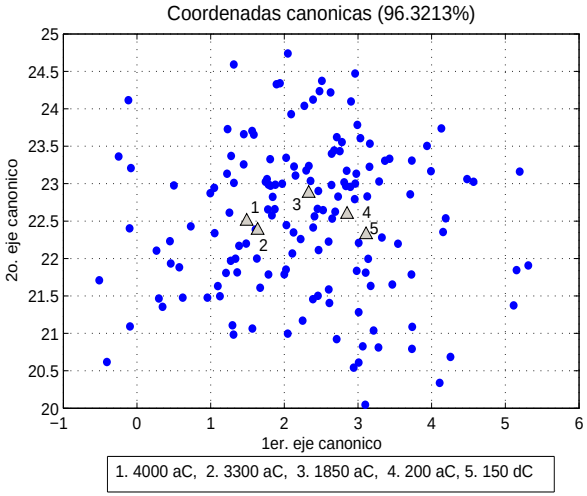
a) Sea `X=[X1;X2;X3;X4;X5]` la matriz $150 \times 4$ que contiene los datos de la tabla anterior. Desde Matlab hacemos:

```
n = [30 30 30 30 30];
[mY,V,B,W,percent,Test1,texto1,Test2,texto2] = canp(X,n)
```

y obtenemos:

```
percent = 88.2272    8.0941    3.2594    0.4193
Test1 =  3.8968   16.0000  434.0000    0.0000
texto1 = Test1: Igualdad de medias (Lambda de Wilks): p-valor=7.1776e-007
Test2 = 50.2206   40.0000    0.1291
texto2 = Test2: Igualdad de covarianzas (test Bartlett): p-valor=0.12905
```

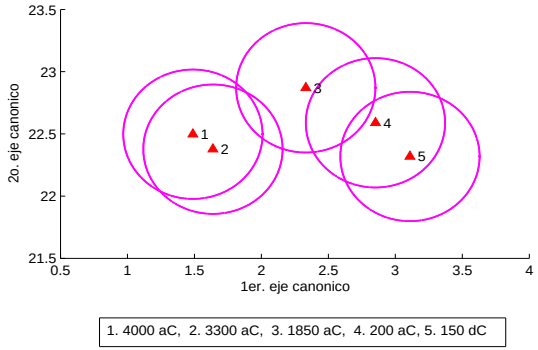
Se rechaza la igualdad de medias con un p-valor menor que $10^{-6}$ y no se rechaza la igualdad de covarianzas, puesto que el p-valor asociado es de 0.12905. Estos resultados confirman que el análisis canónico de poblaciones es aplicable para este conjunto de datos.






b) Las regiones confidenciales son:

```
r = regconf(mY,n,4,0.90)
r =[0.5198   0.5198   0.5198   0.5198   0.5198]
```






c) Los ejes canónicos son las columnas de la matriz `V`:

```
V=[ 0.1267    0.0387    0.0928    0.1488
   -0.0370    0.2101   -0.0246   -0.0004
   -0.1451   -0.0681    0.0147    0.1325
    0.0829   -0.0773   -0.2946    0.0669]
```

El primer eje canónico, que explica el 88.2272% de la variabilidad, es

$$Y_1 = 0.1267\,X_1 - 0.0370\,X_2 - 0.1451\,X_3 + 0.0829\,X_4.$$

En la figura anterior puede apreciarse la ordenación temporal de los cinco períodos históricos a lo largo del primer eje canónico. Por tanto, este primer eje puede interpretarse como la evolución del cráneo a lo largo de la historia, con una tendencia hacia cráneos más anchos y algo achatados, con mandíbulas pequeñas y narices relativamente largas.



d) La matriz de distancias entre los cinco grupos puede obtenerse a partir de las distancias euclídeas entre las filas de la matriz `mY`, que contiene las coordenadas de los individuos medios en función de las coordenadas canónicas:

```
squareform(pdist(mY))=[     0   0.1920   0.9216   1.3660   1.6303
                       0.1920        0   0.8507   1.2317   1.4719
                       0.9216   0.8507        0   0.5913   0.9535
                       1.3660   1.2317   0.5913        0   0.3736
                       1.6303   1.4719   0.9535   0.3736        0]
```

Las distancias que se observan en la representación canónica (figura anterior) coinciden con las distancias de Mahalanobis entre los individuos medios en función de las variables originales.

Por tanto, para estudiar posibles relaciones entre los distintos grupos será más cómodo observar el gráfico de la representación canónica que la matriz de distancias de Mahalanobis.

**Ejercicio:**

Con el fin de ver si existen diferencias entre la composición química de una serie de piezas de cerámica romana encontradas en cuatro yacimientos de Reino Unido (1. Llanederyn, 2. Caldicot, 3. Island Thorns y 4. Ashley Rails), se midieron los porcentajes de óxidos de varios metales sometidos a espectroscopia de absorción atómica: $X_1 =$"porcentaje de óxido de aluminio", $X_2 =$"porcentaje de óxido de hierro", $X_3 =$"porcentaje de óxido de magnesio", $X_4 =$"porcentaje de óxido de calcio", $X_5 =$"porcentaje de óxido de sodio".



Por alguna causa que se desconoce, solamente se ha conservado la siguiente información:

- El grupo 1 consta de 14 individuos, mientras que el grupo 2 sólo tiene 2 individuos.
- Los vectores de medias de los grupos 3 y 4 son, respectivamente:

| | | | | |
|---|---|---|---|---|
| 18.1800 | 1.7120 | 0.6740 | 0.0260 | 0.0540 |
| 17.3200 | 1.5120 | 0.6060 | 0.0520 | 0.0480 |

- Los coeficientes del primer eje canónico son:

| | | | | |
|---|---|---|---|---|
| 0.4485 | -1.2350 | -0.2489 | -8.8171 | -2.2078 |



Identifica los cuatro grupos.