1. Análisis Multivariado - Práctica 3

1.1. Coordenadas discriminantes

1. Sean $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n_i}$ observaciones $p-$variadas de la población $i-$ésima, $1 \leq i \leq k$. Sean

$$\overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{i,j} \qquad \overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{\mathbf{x}}_i \qquad \mathbf{Q}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}}$$

donde $n = \sum_{i=1}^{k} n_i$ es el número total de observaciones. Definamos

$$\begin{aligned} \mathbf{B} &= \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}\right) \left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}\right)^{\mathrm{T}} \\ \mathbf{S} &= \frac{1}{n-k} \sum_{i=1}^{k} \mathbf{Q}_i \end{aligned}$$

y consideremos la siguiente medida de separación:

$$\Delta_s^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}\right)^{\mathrm{T}} \mathbf{S}^{-1} \left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}\right)$$

   a) Mostrar que $\Delta_s^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \lambda_1 + \lambda_2 + \cdots + \lambda_s$, donde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s$ son los autovalores no nulos de $\mathbf{S}^{-1}\mathbf{B}$ (o bien de $\mathbf{S}^{-\frac{1}{2}}\mathbf{B}\mathbf{S}^{-\frac{1}{2}}$). También mostrar que $\lambda_1 + \lambda_2 + \cdots \lambda_r$ es la separación resultante cuando se usan sólo las primeras r coordenadas discriminantes.

   b) Deducir que la primer coordenada discriminante produce la principal contribución individual (λ_1) a la medida de separación Δ_s^2 y que en general la $r-$ésima coordenada discriminante contribuye λ_r a la medida de separación Δ_s^2.

2. Supongamos que tenemos dos poblaciones indicadas por 1 y 2 en $\mathbb{R}^2$ con distribuciones $N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y $N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectivamente, donde

$$\boldsymbol{\mu}_1 = (1,2)^{\mathrm{T}}, \quad \boldsymbol{\mu}_2 = (4,1)^{\mathrm{T}}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}.$$

   a) Calcule $\boldsymbol{\Sigma}^{-1}$ y deduzca una expresión para $\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}$.

   b) Calcule la primer coordenada discriminante $\boldsymbol{\alpha}_1$.

   c) En base a $\boldsymbol{\alpha}_1$ como asignaría a un punto $\mathbf{x}_0$ a la población 1?
   ¿Dónde clasificaría $\mathbf{x}_0 = (3,2)^{\mathrm{T}}$?

3. Los cricétidos (Cricetidae) son una familia de roedores miomorfos que forma parte de la gran superfamilia Muroidea. Esta familia incluye a los hamsters, ratas campestres, lemmings y ratones de las Américas.

Con casi 600 especies, entre ellos los *Microtus multiplex* y *Microtus subterraneus* son los más difíciles de distinguir morfológicamente. En este ejercicio, 142 roedores se identificaron a través de un análisis de cromosomas que permitió distinguir 43 *Microtus multiplex*, 46 *Microtus subterraneus* y 53 *Microtus abbreviatus*. Por otra parte, 199 especimenes no se pudieron identificar. Los datos se encuentran en el archivo `microtus-data.txt`.

Para cada roedor se midieron 8 variables

M1Left: Ancho del molar superior izquierdo # 1 (0.001mm)

M2Left: Ancho del molar superior izquierdo # 2 (0.001mm)

M3Left: Ancho del molar superior izquierdo # 3 (0.001mm)

Foramen: Largo de la fosa incisiva (0.001mm)

Pbone: Largo del hueso palatal (0.001mm)

Length Condylo: Largo del cráneo (0.01mm)

Altura: Altura del cráneo sobre bullae (0.01mm)

Rostrum: Ancho del cráneo a través del rostro (0.01mm)

Obteniendo entonces vectores $\mathbf{x}_{i,j} \in \mathbb{R}^8$, $1 \leq j \leq n_i$, $i = 1, 2, 3$ ($k = 3$), $n_1 = 43$, $n_2 = 46$, $n_3 = 53$ donde $i = 1$ indica que el roedor es de la familia *Microtus multiplex*, $i = 2$ que es de la familia *Microtus subterraneus* y $i = 3$ que es *Microtus abbreviatus*. Sean $n = n_1 + n_2 + n_3$, $\mathbf{U} = \mathbf{Q}_1 + \mathbf{Q}_2 + \mathbf{Q}_3$, $\widehat{\mathbf{\Sigma}}_i = \mathbf{Q}_i/n_i$ donde $\mathbf{Q}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}}$. Definamos $\mathbf{H} = \sum_{i=1}^{k} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^{\mathrm{T}}$ donde $\overline{\mathbf{x}} = (1/n) \sum_{i=1}^{k} n_i \overline{\mathbf{x}}_i$. Sean $\mathbf{b}_j$ los autovectores de $\widehat{\mathbf{B}} = (\mathbf{T}^{-1})^{\mathrm{T}} \mathbf{H} \mathbf{T}^{-1}$ tales que $\|\mathbf{b}_j\| = 1$, $\mathbf{b}_j^{\mathrm{T}} \mathbf{b}_\ell = 0$ si $\ell \neq j$ donde $\mathbf{U} = \mathbf{T}^{\mathrm{T}} \mathbf{T}$ y $\mathbf{a}_j = (\mathbf{T}^{-1})^{\mathrm{T}} \mathbf{b}_j \sqrt{n-k}$.

a) Testee con nivel asintótico $\alpha = 0{,}001$ si las tres poblaciones tienen igual matriz de covarianza. Cuál es el $p-$valor?

b) En base a lo obtenido en a) qué estadístico usaría para testear

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$$

Qué distribución tiene bajo H_0. Realice el test con nivel exacto $\alpha = 0.001$.

c) En base a los resultados obtenidos en a) y b), decida si es razonable hacer un plot de las coordenadas discriminantes. ¿Cuántas coordenadas consideraría? Explique claramente porqué.

d) Grafique los datos proyectados en las dos primeras coordenadas discriminantes indicando con cuadrados negros la proyección de los 43 datos *Microtus multiplex*, con circulos rojos la de los 46 datos *Microtus subterraneus* y con triángulos azules la de los 53 datos *Microtus abbreviatus*. Indique además las medias proyectadas de cada grupo y los círculos de confianza para el valor medio esperado de la proyección de cada grupo con nivel 0.95.

Qué puede deducir del gráfico?

Observación: Trate de tener la misma escala en ambos ejes para que los círculos se vean como tales

e) Grafique las semi-rectas que delimitan regiones del espacio que **A**, **B** y **C** que permitan clasificar a cada roedor en una de las tres clases *Microtus multiplex*, *Microtus subterraneus* o *Microtus abbreviatus*.

Indique como asignaría un nuevo roedor a cada una de esas tres clases en base a estas regiones.

f) Grafique luego en el mismo plot con diamantes rellenos verdes los 199 roedores que externamente no se pudieron atribuir a ninguno de los dos grupos. Qué observa respecto de los 199 roedores?

4. En la primer fase de un estudio sobre el costo de transporte de la leche desde las granjas hasta las lecherías, se tomó una muestra de empresas de transporte vinculadas al transporte de lácteos. En la tabla 1 se presentan los datos de costos por milla de

$X_1 =$ combustible

$X_2 =$ reparaciones

$X_3 =$ capital

para $n_1 = 36$ camiones nafteros y $n_2 = 23$ camiones a diesel.

a) Testear si hay diferencias entre los vectores de costos medios. Tomar $\alpha = 0{,}01$.

b) Si la hipótesis de igual vector de costos medios es rechazada en la parte (a), hallar la combinación lineal de las componentes de las medias que es más responsable del rechazo.

c) Construir intervalos de confianza de nivel simultáneo 0.99 para los pares de costos medios. Si los hay ¿Qué costos aparecen como muy distintos?

d) Comentar la validez de los supuestos realizados.

e) En base a los resultados obtenidos decida si es razonable hacer un plot de la primera coordenada discriminante.

5. Del conjunto de datos "iris" consideremos las variables $X_2 =$ Ancho de los sépalos y $X_4 =$ Ancho de los pétalos para las 3 especies de flores.

a) Graficar los pares de datos (X_2, X_4) en el plano. Para cada especie, estos datos ¿tienen aspecto de provenir de una distribución normal bivariada?

b) Asumiendo que las muestras provienen de poblaciones con distribución normal bivariada con matriz de covarianza común $\mathbf{\Sigma}$, testear a nivel $\alpha = 0{,}05$, la hipótesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$, versus H_1 : al menos una de las $\boldsymbol{\mu}_i$ es distinta de las otras. ¿Es razonable el supuesto de igualdad de matrices de covarianza en este caso?

c) Considere ahora solamente las especies Virginica y Versicolor y repita a) y b). Si es razonable el supuesto de igualdad de matrices de covarianza, haga un scatterplot de las primeras coordenadas discriminantes significativas.

d) Repita c) con las variables (X_1, X_2, X_3, X_4).

6. Considere el conjunto de datos `swiss.heads` del paquete `Flury` y realice un análisis de coordenadas discriminantes para los grupos de hombres y mujeres validando los supuestos necesarios.

7. Sean $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,n_i}$ observaciones $p-$variadas de la población $i-$ésima, $1 \leq i \leq k$. Sea $n = \sum_{i=1}^{k} n_i$ y definamos $\widetilde{\mathbf{x}}_1 = \mathbf{x}_{1,1}, \ldots, \widetilde{\mathbf{x}}_{n_1} = \mathbf{x}_{1,n_1}, \ldots, \widetilde{\mathbf{x}}_n = \mathbf{x}_{k,n_k}$, es decir,

$$\begin{aligned} \widetilde{\mathbf{x}}_j &= \mathbf{x}_{1,j} \qquad \text{para } 1 \leq j \leq n_1 \\ \widetilde{\mathbf{x}}_{\sum_{m=1}^{i-1} n_m + j} &= \mathbf{x}_{i,j} \qquad \text{para } 1 \leq j \leq n_i \,, 2 \leq i \leq k \end{aligned}$$

Definamos un vector *dummy*, $\mathbf{y} \in \mathbb{R}^{k-1}$ que indica el grupo de pertenencia de una observación, o sea, para $1 \leq \ell \leq n$

$$\begin{aligned} \mathbf{y}_\ell &= (y_{\ell,1}, \ldots, y_{\ell,k-1})^{\mathrm{T}}) \\ y_{\ell,i} &= \begin{cases} 1 & \text{si } \widetilde{\mathbf{x}}_\ell \text{ pertenece al grupo i,} \\ 0 & \text{en otro caso.} \end{cases} \end{aligned}$$

Llamemos $\mathbf{S}$ a la matriz de varianzas y covarianzas muestral de $\mathbf{z}_\ell = (\widetilde{\mathbf{x}}_\ell^{\mathrm{T}}, \mathbf{y}_\ell^{\mathrm{T}})^{\mathrm{T}}$, $1 \leq \ell \leq n$. Sean

$$\begin{aligned} \mathbf{U} &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}} \\ \mathbf{H} &= \sum_{i=1}^{k} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^{\mathrm{T}} \end{aligned}$$

a) Muestre que $\mathbf{U} + \mathbf{H} = n\mathbf{S}_{11}$.

b) Muestre que $\mathbf{H} = n\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$.

c) Realice un análisis de correlación canónica entre $\widetilde{\mathbf{x}}$ e $\mathbf{y}$.

Deduzca que las variables canónicas asociadas al vector $\widetilde{\mathbf{x}}$ coinciden con las coordenadas discriminantes.

Camiones nafteros			Camiones diesel		
X_1	X_2	X_3	X_1	X_2	X_3
16.44	12.43	11.23	8.50	12.26	9.11
7.19	2.70	3.92	7.42	5.13	17.15
9.92	1.35	9.75	10.28	3.32	11.23
4.24	5.78	7.78	10.16	14.72	5.99
11.20	5.05	10.67	12.79	4.17	29.28
14.25	5.78	9.88	9.60	12.72	11.00
13.50	10.98	10.60	6.47	8.89	19.00
13.32	14.27	9.45	11.35	9.95	14.53
29.11	15.09	3.28	9.15	2.94	13.68
12.68	7.61	10.23	9.70	5.06	20.84
7.51	5.80	8.13	9.77	17.86	35.18
9.90	3.63	9.13	11.61	11.75	17.00
10.25	5.07	10.17	9.09	13.25	20.66
11.11	6.15	7.61	8.53	10.14	17.45
12.17	14.26	14.39	8.29	6.22	16.38
10.24	2.59	6.09	15.90	12.90	19.09
10.18	6.05	12.14	11.94	5.69	14.77
8.88	2.70	12.23	9.54	16.77	22.66
12.34	7.73	11.68	10.43	17.65	10.66
8.51	14.02	12.01	10.87	21.52	28.47
26.16	17.44	16.89	7.13	13.22	19.44
12.95	8.24	7.18	11.88	12.18	21.20
16.93	13.37	17.59	12.03	9.22	23.09
14.70	10.78	14.58			
10.32	5.16	17.00			
8.98	4.49	4.26			
9.70	11.59	6.83			
12.72	8.63	5.59			
9.49	2.16	6.23			
8.22	7.95	6.72			
13.70	11.22	4.91			
8.21	9.85	8.17			
15.86	11.42	13.06			
9.18	9.18	9.49			
12.49	4.67	11.94			
17.32	6.86	4.44			

Cuadro 1: Datos costo por milla de camiones. Corresponde a Johnson, R. A.. y Wichern, D. W.; tabla 6.6