1 Análisis Multivariado - Práctica 4 - Parte 1

Los ejercicios marcados en rojo no son para elegir para exponer, aunque deben hacerse.

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

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

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

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

Consideremos la siguiente medida de separación:

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

   (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 $\mathbf{a}_1$.

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

3. En el ejercicio 1 de la Práctica 3 Parte 1 se estudiaba el costo de transporte de la leche desde las granjas hasta las lecherías para $n_1 = 36$ camiones nafteros y $n_2 = 23$ camiones a diesel. En base a los resultados obtenidos en el ejercicio 8 de la Práctica 3 Parte 2 decida si es razonable hacer un gráfico de la primera coordenada discriminante.

4. En el ejercicio 2 de la Práctica 3 Parte 1 se estudiaba longitud de las antenas y de las alas de nueve insectos *Amerohelea fasciata (Af)* y seis *A. pseudofasciata (Apf)*. Consideremos las variables x_1=*longitud de las antenas + longitud de las alas* y x_2=*longitud de las alas*.

   (a) Testee si las dos poblaciones tienen igual matriz de covarianza. Tomar $\alpha = 0.01$. En base al resultado, decida si es razonable hacer un plot de la primera coordenada discriminante.

   (b) Haga un plot de las dos primeras coordenadas discriminantes. ¿Qué observa en la segunda coordenada?

   (c) Haga un plot de los puntos originales y grafique la recta $\widehat{\mathbf{a}}^{\mathrm{T}} \{\mathbf{x} - (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)/2\} = 0$ donde $\widehat{\mathbf{a}} = \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$. Qué observa?

   (d) Comprobar que la distancia euclídea entre los promedios de cada grupo expresados en la primer coordenada canónica coincide con la distancia de Mahalanobis entre los promedios $\overline{\mathbf{x}}_1$ y $\overline{\mathbf{x}}_2$ expresados en las variables originales, es decir, coincide con $\sqrt{(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}} \mathbf{S}^{-1} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)}$.

5. En el ejercicio 5 de la Práctica 3 Parte 1 se estudiaba las medidas de cinco variables biométricas sobre gorriones hembra, recogidos después de una tormenta de los cuales $n_1 = 21$ sobrevivieron y $n_2 = 28$ murieron.

   (a) En base a los resultados obtenidos en el ejercicio 11 de la Práctica 3 Parte 2 decida si es razonable hacer un gráfico de la primera coordenada discriminante.

   (b) Si es razonable hacer el análisis interprete los coeficientes de la primer coordenada canónica, ubique la proyección del promedio de cada grupo. Qué observa?

6. Consideremos los datos "iris" del R. Es un conjunto de datos analizados por Fisher que consisten en 4 mediciones realizadas en 50 flores iris de cada una de 3 especies distintas (Setosa, Versicolor y Virginica). Las 4 variables, medidas en centímetros, son

   - x_1 = Longitud de los sépalos (sepal length)
   - x_2 = Ancho de los sépalos (sepal width)
   - x_3 = Longitud de los pétalos (petal length)
   - x_4 = Ancho de los pétalos (petal width)

   (a) Realizar un scatterplot de las primeras 2 coordenadas discriminantes, indicando cada grupo con un color diferente.

(b) Realice un gráfico con las círcunferencias de confianza de nivel 95% de las proyecciones $\boldsymbol{\nu}_i$, $1 \leq i \leq 3$, de los valores esperados $\boldsymbol{\mu}_i$ de cada población en el plano de las coordenadas discriminantes. Ubique la media muestral proyectada de cada grupo. Qué observa?

(c) Analice si los supuestos para realizar este gráfico se cumplen, suponiendo que los datos son normales.

7. Del conjunto de datos "iris" consideremos solamente las variables

- x_2 = Ancho de los sépalos
- 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 $\boldsymbol{\Sigma}$, 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).