

1 Análisis Multivariado - Práctica 2

1.1 Coordenadas discriminantes

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

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

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

$$\mathbf{B} = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$$
$$\mathbf{S} = \frac{1}{n-k} \sum_{i=1}^k \mathbf{Q}_i$$

Consideremos la siguiente medida de separación:

$$\Delta_s^2 = \sum_{i=1}^k (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})$$

- (a) Mostrar que $\Delta_s^2 = \lambda_1 + \lambda_2 + \dots + \lambda_p = \lambda_1 + \lambda_2 + \dots + \lambda_s$, donde $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_s$ 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 $\lambda_1 + \lambda_2 + \dots + \lambda_r$ 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. 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 2.1 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.
3. Grogan y Smith (1981) describen dos especies de mosquitos (midges): *Amerohelea fasciata* (*Af*) y *A. pseudofasciata* (*Apf*) recientemente descubiertas. En la tabla 2.2 aparecen los datos correspondientes a las mediciones de la longitud de las antenas y de las alas de nueve insectos *Af* y seis *Apf*. Consideremos las variables $x_1 = \text{longitud de las antenas} + \text{longitud de las alas}$ y $x_2 = \text{longitud de las alas}$.
- (a) Testee si las dos poblaciones tienen igual matriz de covarianza. Tomar $\alpha = 0.01$. En base al resultado obtenido, 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 $\hat{\mathbf{a}}^T(\mathbf{x} - (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2) = 0$ donde $\hat{\mathbf{a}} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.
4. 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
 $X_1 = \text{Longitud de los sépalos (sepal length)}$
 $X_2 = \text{Ancho de los sépalos (sepal width)}$
 $X_3 = \text{Longitud de los pétalos (petal length)}$
 $X_4 = \text{Ancho de los pétalos (petal width)}$
 Realizar un scatterplot de las primeras 2 coordenadas discriminantes. Analice si los supuestos para realizar este gráfico se cumplen, suponiendo que los datos son normales.
5. Del conjunto de datos “iris” consideremos las variables $X_2 = \text{Ancho de los sépalos}$ y $X_4 = \text{Ancho de los pétalos}$ 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 Σ , testear a nivel $\alpha = 0.05$, la hipótesis $H_0 : \mu_1 = \mu_2 = \mu_3$, versus $H_1 : \text{al menos una de las } \mu_i \text{ es distinta de las otras}$. ¿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) .