1 Análisis Multivariado - Práctica 4

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

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

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. En el ejercicio 1 de la Sección 1.1 de la Práctica 3 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 1 de la Sección 1.3 de la práctica 3 decida si es razonable hacer un plot de la primera coordenada discriminante.

3. En el ejercicio 2 de la Sección 1.1 de la Práctica 3 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. Que 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 $\widehat{\mathbf{a}} = \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$.

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 = 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)
   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 = 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).

1.2 Componentes principales

1. Sea $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ una muestra aleatoria bivariada, con $Var(X_i, Y_i) = \Sigma \in \mathbb{R}^2$. Deducir las ecuaciones de la recta que minimiza la distancia a todos los puntos (es decir, la recta de mínimos cuadrados ortogonales), y compararla con la recta de regresión de X en Y y la de Y en X.

   (a) Suponiendo Σ conocida (encontrar los autovalores).

   (b) Estimando a Σ.

   (c) En la tabla 4.1 figuran los datos correspondientes a mediciones realizadas sobre los caparazones de 24 tortugas macho. Sean $X_1 = 10\ ln(longitud)$ y $X_2 = 10\ ln(ancho)$. (La transformación logarítmica es frecuente en estudios morfométricos, multiplicamos por 10 los datos para que la escala en la que trabajamos sea conveniente numéricamente). Hacer un gráfico de X_1 vs. X_2 y ajustarle las 3 rectas propuestas en este ejercicio.

2. Sea $\mathbf{x}$ un vector aleatorio bivariado, con distribución uniforme en el semicírculo $\mathbf{x}\acute{\mathbf{x}} \leq 1,\ x_2 > 0$. Hallar las componentes principales de $\mathbf{x}$.

3. Sea $\mathbf{x}$ un vector aleatorio con matriz de dispersión Σ equicorrelacionada, es decir

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

con $\sigma^2 > 0$ y $\rho \in (-1, 1)$. Hallar los autovalores y autovectores de Σ (o sea, las componentes principales).

(a) Probar que las componentes principales son equivariantes por transformaciones ortogonales (es decir que si uno le aplica una transformación ortogonal a los datos originales las componentes principales cambian de igual modo).

(b) Veamos, con un ejemplo, que las componentes principales no son invariantes por cambios de escala. Sea $\mathbf{x}$ un vector aleatorio bivariado con esperanza $\mathbf{0}$ y matriz de dispersión Σ

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}$$

Calcular las componentes principales cuando $\sigma > 1$, hallar la aproximación unidimensional de la primer componente principal y calcular porcentaje de la variabilidad total explicado por ella. Luego, cambiar la escala de las mediciones del siguiente modo: sean $U_1 = aX_1$, y $U_2 = X_2$, repetir la cuenta anterior y ver que el porcentaje de la variabilidad total explicado por la la aproximación unidimensional de la primer componente principal es tan grande como se quiere con tal que a sea suficientemente grande; y que tomando $a < 1/\sigma$ la primer componente principal cambia y nuevamente cuando $a \to 0$, el porcentaje de la variabilidad total explicado por la la aproximación unidimensional de la primer componente principal es tan grande como se quiere.

4. En la tabla 4.2 figuran los datos de las mediciones de huesos y dientes de ratones campestres (de la especie *Microtus*). Consideremos solamente las primeras 3 variables (Y_1, Y_2, Y_3) que son el ancho de los molares superiores izquierdos 1, 2 y 3 respectivamente, medidos en mm./1000, para las 43 ratas del grupo 1.

   (a) Hallar las componentes principales muestrales y los autovalores de S.

   (b) Hallar los porcentajes de la variabilidad total explicados por la primera y por las dos primeras componentes, e interpretarlas en función de las variables originales.

5. Repetir el análisis anterior pero agregar a los datos las mediciones del grupo 2.

6. Sea $\mathbf{x}$ un vector aleatorio de dimensión d con media $\mathbf{0}$, y matriz de dispersión $\Sigma = [(\sigma_{jk})]$, con todas las covarianzas positivas. Sea $\mathbf{t}_1$ un autovector de norma 1 correspondiente al mayor autovalor, mostrar que todos sus coeficientes son o bien positivos o bien negativos.

7. Sea $\mathbf{x}$ un vector aleatorio de dimensión d con media $\mathbf{0}$, y matriz de dispersión $\Sigma = [(\sigma_{jk})]$, y sea $y_j = \mathbf{t}_j'\mathbf{x}$ la j-ésima componente principal de $\mathbf{x}$. Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ los autovalores de Σ.

   (a) Verificar que las y_j son no correlacionadas y que $var(y_j) = \lambda_j$

   (b) Mostrar que la correlación entre x_i y y_j es $t_{ij}\sqrt{\lambda_j/\sigma_{ii}}$, donde t_{ij} es el i-ésimo elemento de $\mathbf{t}_j$.

8. En base a la distribución asintótica de la suma de autovalores,

(a) ¿Cuándo se puede asegurar que la primer componente principal explica el 80% de la variabilidad total con un $(1-\alpha)$ 100% de confianza?

(b) ¿Cuándo se puede asegurar que las dos primeras componentes principales explican el 80% de la variabilidad total con un $(1-\alpha)$ 100% de confianza?

(c) Hallar cuál es el número de componentes principales que se necesitan para asegurar que se ha explicado el 80% de la variabilidad total con un $(1-\alpha)$ 100% de confianza.