Análisis Multivariado - Práctica 4

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. 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 $\mathbf{t}_j$.

5. 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.

6. 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.

   (c) Realizar un biplot.

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

8. 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.