

# Análisis Multivariado: Tarea 1

## Análisis Descriptivo de Datos Multivariados

Fecha de entrega: 24 de febrero.

1. (1 punto) Para un punto  $\mathbf{x}$  en  $\mathbb{R}^p$  con  $p > 1$  considerar para  $t \in [-\pi, \pi]$  el mapeo  $f : \mathbb{R}^p \rightarrow \mathbb{R}$  definido como

$$f(\mathbf{x}) = \begin{cases} \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \cdots + x_n \sin\left(\frac{p}{2}t\right) & \text{si } p \text{ es par} \\ \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \cdots + x_n \cos\left(\frac{(p-1)}{2}t\right) & \text{si } p \text{ es impar} \end{cases}$$

Mostrar que para dos puntos  $\mathbf{x}, \mathbf{y}$  en  $\mathbb{R}^p$ , se cumple que

$$\|f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)\|_{L_2} = \pi \|\mathbf{x} - \mathbf{y}\|^2,$$

donde

$$\|f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)\|_{L_2} = \int_{-\pi}^{\pi} [f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)]^2 dt.$$

¿Cómo se relaciona esta propiedad con la identificación de clusters y outliers?

2. (2 puntos) Mostrar que si  $\mathbf{H}_n$  es la matriz de centrado definida como

$$\mathbf{H}_n = \mathbf{I}_{n \times n} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$$

entonces

- i.  $\mathbf{H}_n$  es simétrica.
- ii.  $\mathbf{H}_n$  es idempotente.
- iii. Para una matriz  $\mathbf{X}_{n \times p}$  la media muestral de  $\mathbf{W} = \mathbf{H}_n \mathbf{X}$  es el vector  $\mathbf{0}_p$ .
- iv. La matriz de varianza y covarianza  $\mathbf{S}$  de  $\mathbf{X}$  se puede escribir como

$$\mathbf{S} = \frac{1}{n-1} (\mathbf{X}^T \mathbf{H}_n \mathbf{X}).$$

3. (1 punto) Sea  $\mathbf{S}$  una matriz cuadrada tal que  $\mathbf{S} = \mathbf{A}^T \mathbf{A}$ , donde  $\mathbf{A}_{n \times p}$  entonces

- i.  $\mathbf{S}$  es simétrica.
- ii.  $\mathbf{S}$  es semidefinida positiva.

Concluir por tanto que la matriz de varianza y covarianza muestral y la matriz de correlación muestral son simétricas y semidefinidas positivas.

4. (1 punto) Mostrar que si  $\mathbf{x}$  es un vector  $p$ -variado donde  $\Sigma = \text{Var}(\mathbf{x})$  entonces  $\text{Det}(\Sigma) \geq 0$ .

5. (1 punto) Sea  $\mathbf{X}_{n \times p}$  una matriz de datos y considerar la transformación

$$\mathbf{Y} = \mathbf{X}\mathbf{A}^T + \mathbf{1}_n \mathbf{b}^T,$$

donde  $\mathbf{A}_{q \times p}$  y  $\mathbf{b}_{q \times 1}$  son constantes. Mostrar que

$$\mathbf{S}_Y = \mathbf{A} \mathbf{S}_X \mathbf{A}^T.$$

6. (1 punto) Para un vector aleatorio  $\mathbf{x}$  tal que  $\mathbb{E}(\mathbf{x}) = \mu$  y  $\text{Var}(\mathbf{x}) = \Sigma$  definimos a las medidas de asimetría y curtosis respectivamente como

$$\beta_{1,p} = \mathbb{E} [(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{y} - \mu)]^3$$

$$\beta_{2,p} = \mathbb{E} [(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)]^2,$$

donde  $\mathbf{x}$  y  $\mathbf{y}$  son independientes e idénticamente distribuidas. Mostrar que estas medidas son invariantes ante transformaciones lineales.

7. (3 puntos) El archivo *wine.txt* contiene 13 variables numéricas derivadas de un análisis químico en vinos de Italia de tres viñedos diferentes. Realizar un análisis descriptivo multivariado de los datos.