Análisis Multivariado: Tarea 1

Análisis Descriptivo de Datos Multivariados

Fecha de entrega: 28 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 $f : \mathbb{R}^p \to \mathbb{R}$, definido como:

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

Mostrar lo siguiente:

i. La transformación preserva medias, esto es, para una colección de puntos $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$,

$$f_{\bar{\mathbf{x}}}(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}_i}(t).$$

ii. Para dos puntos $\mathbf{x}$, $\mathbf{y}$ en $\mathbb{R}^p$, se cumple que,

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

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

iii. Asumiendo no correlación y varianza constante, σ^2, demuestra que para p impar,

$$\mathsf{Var}(f_{\mathbf{x}}(t)) \;=\; \sigma^2 \left(\frac{p}{2}\right),$$

y para p par, i.e. $p = 2r$, se cumple que,

$$\sigma^2 \left(r - \frac{1}{2}\right) \leq \mathsf{Var}(f_{\mathbf{x}}(t)) \leq \sigma^2 \left(r + \frac{1}{2}\right),$$

¿Qué puedes concluir sobre la varianza con respecto a p? ¿Qué tan adecuados son los supuestos para datos multivariados?

2. (1 punto) 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 de datos $\mathbf{X}_{n\times p}$, la media muestral de $\mathbf{W} = \mathbf{H}_n\mathbf{X}$ es el vector $\mathbf{0}_p$.
   iv. La matriz de varianza y covarianza muestral de $\mathbf{X}$, se puede escribir como,

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

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

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

   Utilizando este resultado justifica y concluye que la matriz de covarianzas muestral y la matriz de correlaciones muestral son semidefinidas positivas.

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

5. (1 punto) Sean $\mathbf{x}$ y $\mathbf{y}$ dos vectores aleatorios independientes.

   i. Demostrar que para constantes reales α y β se tiene que,

$$\mathrm{Var}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha^2\mathrm{Var}(\mathbf{x}) + \beta^2\mathrm{Var}(\mathbf{y}).$$

   ii. Encontrar una fórmula para el caso no independiente y expresarla en términos de$\mathrm{Var}(\mathbf{x})$, $\mathrm{Var}(\mathbf{y})$ y $\mathrm{Cov}(\mathbf{x}, \mathbf{y})$.

6. (1 punto) Sea $\mathbf{X}_{n\times p}$ una matriz de datos. Considera 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

$$\mathbf{S}_\mathbf{Y} = \mathbf{A}\mathbf{S}_\mathbf{X}\mathbf{A}^T.$$

7. (1 punto) Para una matriz de datos, $\mathbf{X}$, considera las siguientes dos transformaciones,

   i. $\mathbf{Y} = \mathbf{H}_n\mathbf{X}\mathbf{D}^{-1}$

   ii. $\mathbf{Z} = \mathbf{H}_n\mathbf{X}\mathbf{S}^{-\frac{1}{2}}$,

   donde $\mathbf{S}$ es la matriz de varianza y covarianza muestral de $\mathbf{X}$ y $\mathbf{D} = \mathrm{diag}(s_1, \ldots, s_p)$. Obtener la media muestra y la matriz de covarianzas de $\mathbf{Y}$ y de $\mathbf{Z}$. ¿En qué difieren estas dos transformaciones?

8. (1 punto) Para un vector aleatorio $\mathbf{x}$, tal que $\mathbb{E}(\mathbf{x}) = \mu$ y $\mathsf{Var}(\mathbf{x}) = \Sigma$, definimos a las medidas de asimetría y curtosis respectivamente como:

$$\begin{aligned} \beta_{1,p} &= \mathbb{E}\left(\left[(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right]^3\right) \\ \beta_{2,p} &= \mathbb{E}\left(\left[(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right]^2\right), \end{aligned}$$

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

9. (1 punto) 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.

10. (1 punto) El archivo *Diabetes.txt* contiene 5 mediciones relacionadas con la diabetes de 145 adultos. Realiza un análisis descriptivo de estos datos. ¿Puedes identificar clusters, outliers y/o variables importantes?