

Técnicas de Análisis Multivariado 2 - Práctica 1

1 Componentes principales y autoconsistencia

1. Sea $\mathbf{x} \in \mathbb{R}^p$ y $\{\mathcal{A}_1, \dots, \mathcal{A}_k\}$ una partición de \mathbb{R}^p , i.e., $\cup_{i=1}^k \mathcal{A}_i = \mathbb{R}^p$ y $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ si $i \neq j$. Suponga que $\mathbb{P}(\mathbf{x} \in \mathcal{A}_i) > 0$ para $i = 1, \dots, k$ y defina un vector $\mathbf{y} \in \mathbb{R}^p$ como

$$\mathbf{y} = \mathbb{E}(\mathbf{x} | \mathbf{x} \in \mathcal{A}_i) \quad \text{si } \mathbf{x} \in \mathcal{A}_i$$

Pruebe que \mathbf{y} es auto-consistente para \mathbf{x} .

2. (a) Sean $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ vectores aleatorios con matrices de covarianza $\Sigma_{\mathbf{x}}$ y $\Sigma_{\mathbf{y}}$ respectivamente. Pruebe que si \mathbf{y} es auto-consistente para \mathbf{x} entonces

$$MSE(\mathbf{x}, \mathbf{y}) = \mathbb{E}\|\mathbf{x} - \mathbf{y}\|^2 = \text{TR}(\Sigma_{\mathbf{x}}) - \text{TR}(\Sigma_{\mathbf{y}}) = \sum_{j=1}^p \text{VAR}(x_j) - \text{VAR}(y_j)$$

Concluya que $\Sigma_{\mathbf{x}} - \Sigma_{\mathbf{y}}$ es semidefinida positiva.

- (b) Sea $X \sim N(0, 1)$. Definamos

$$Y = \begin{cases} -\sqrt{\frac{2}{\pi}} & \text{si } X < 0 \\ \sqrt{\frac{2}{\pi}} & \text{si } X \geq 0 \end{cases}$$

entonces, Y es auto-consistente para X y

$$MSE(X, Y) = 1 - \frac{2}{\pi}$$

3. Sea $\mathbf{x} = (X_1, X_2)^T$ un vector aleatorio tal que $\mathbf{x} \sim N_2(\mathbf{0}, \mathbf{I})$. Sea

$$\mathbf{y} = (Y_1, Y_2)^T = \sqrt{\frac{2}{\pi}} \begin{pmatrix} \text{signo}(X_1) \\ \text{signo}(X_2) \end{pmatrix}$$

donde $\text{signo}(t) = 1$ si $t > 0$ y -1 si $t < 0$. Sea $\mathbf{z} = (X_1, 0)$

- (a) Muestre que tanto \mathbf{y} como \mathbf{z} son auto-consistentes para \mathbf{x}
(b) Cuál de las dos aproximaciones auto-consistentes \mathbf{y} y \mathbf{z} tiene menor error cuadrático, es decir, cual es menor entre $MSE(\mathbf{x}, \mathbf{y})$ y $MSE(\mathbf{x}, \mathbf{z})$?
4. Sea $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$ donde

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

- (a) Obtenga los autovalores $\lambda_1 > \lambda_2$ y autovectores β_1, β_2 de Σ .
(b) Muestre que $\mathbf{y}_{(1)} = \beta_1 \beta_1^T \mathbf{x}$ y $\mathbf{y}_{(2)} = \beta_2 \beta_2^T \mathbf{x}$ son auto-consistentes para \mathbf{x} . Calcule $MSE(\mathbf{x}, \mathbf{y}_{(1)})$ y $MSE(\mathbf{x}, \mathbf{y}_{(2)})$.
(c) Encuentre a matriz de proyección \mathbf{P} asociada a la proyección sobre la recta $x_2 = x_1/3$ y muestre que $\mathbf{y} = \mathbf{P}\mathbf{x}$ no es auto-consistente para \mathbf{x} . Es decir, encuentre un vector $\mathbf{c} = (c, c/3)^T$ tal que

$$\mathbb{E}(\mathbf{x} | \mathbf{y} = \mathbf{c}) \neq \mathbf{c}$$

2 Biplot

1. Considere el conjunto de datos **iris**. Consiste en medidas (en cm) de largo y ancho del pétalo y sépalo para 50 flores de 3 especies de iris: versicolor (Grupo 1), virginica (Grupo 2) y setosa (Grupo 3). Indiquemos por \mathbf{x}_{ij} , $1 \leq j \leq n_i = 50$, las observaciones del grupo i y por

$$\mathbf{Q}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T$$

la suma de cuadrados del grupo i . Sean $n = n_1 + n_2 + n_3$

$$\mathbf{S}_{\text{POOLED}} = \frac{1}{n}(\mathbf{Q}_1 + \mathbf{Q}_2 + \mathbf{Q}_3)$$

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^3 \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})^T \quad \bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^3 \sum_{j=1}^{n_i} \mathbf{x}_{ij}$$

- (a) Para cada grupo por separado, realice al siguiente análisis
- Hallar las componentes principales (es decir, las basadas en $\mathbf{S}_i = \mathbf{Q}_i/n_i$) e interpretarlas en función de las variables originales.
 - Realice un biplot.
 - Qué observa?
- (b) Considere ahora todas las observaciones juntas y realice los siguientes dos análisis
- i. Considere todas las observaciones como provenientes de una única población
 - Halle las componentes principales muestrales y los autovalores de \mathbf{S} .
 - Halle los porcentajes de la variabilidad total explicados por la primera y por las dos primeras componentes, e interpretarlas en función de las variables originales.
 - Grafique las dos primeras componentes principales. Qué observa?
 - Realice un biplot.
 - ii. Considere ahora a las observaciones como provenientes de poblaciones con igual matriz de covarianza pero distintas medias
 - Halle los autovectores y los autovalores de $\mathbf{S}_{\text{POOLED}}$. En base a esos autovectores $\hat{\gamma}_\ell$, $1 \leq \ell \leq p = 4$, contruya las componentes principales como

$$\mathbf{v}_{ij} = \hat{\Gamma}^T (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) \quad 1 \leq j \leq n_i \quad 1 \leq i \leq 3$$
 donde $\hat{\Gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_p)^T$.
 - Halle los porcentajes de la variabilidad total explicados por la primera y por las dos primeras componentes, e interpretarlas en función de las variables originales.
 - Grafique las dos primeras componentes principales. Qué observa? Como se compara con el análisis hecho en los puntos (a) y (b)(i).
 - Realice un biplot adaptado a esta situación.
 - iii. Realice un test para la hipótesis nula $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$ y para $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3$. Es razonable considerar todas las especies juntas? Alguno de los analisis hechos es (b)(i) o (ii) es razonable?

- (c) Repita el análisis hecho en b) para los grupos versicolor (Grupo 1) y virginica (Grupo 2).
2. El conjunto de datos del archivo `microtus-data.txt` consiste en las mediciones de huesos y dientes de ratones campestres (de la especie *Microtus*). Considere solamente las 43 ratas del grupo *Multiplex*.
- (a) Utilice 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 (M1Left, M2Left, M3Left), medidos en mm./1000, para las 43 ratas del grupo *Multiplex*.
- Halle las direcciones principales muestrales y los autovalores de S .
 - Halle los porcentajes de la variabilidad total explicados por la primera y por las dos primeras componentes, e interpretarlas en función de las variables originales.
 - Realice un biplot.
- (b) Realice un biplot con todas las variables.
- (c) Repita el análisis anterior agregando a los datos las mediciones del grupo *Subterraneus*. Qué observa?
3. Considere el conjunto de datos que se encuentran en el archivo `Datos Países.xlsx`, correspondiente a las siguientes 6 variables indicadoras de desarrollo de 91 países.
- | | |
|----------------|--|
| tasanat | tasa de natalidad cada 1000 habitantes |
| tasamor | tasa de mortalidad cada 1000 habitantes |
| mortinf | mortalidad infantil (por debajo de un año) |
| esphom | esperanza de vida en hombres |
| espmuj | esperanza de vida en mujeres |
| pnb | producto bruto nacional per cápita |
- (a) Realizar un biplot de los datos.
- (b) Realizar un biplot de los datos transformados por logaritmos.
- (c) Cuál elegiría?
4. Realice un análisis de componentes principales para el conjunto de datos `heptathlon` de la librería `HSAUR` y construya el biplot.