Análisis Multivariado 2 - Práctica 1

1 Componentes principales y autoconsistencia

1. Sea $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ una matriz de covarianza definida positiva. Considere el problema de aproximar $\mathbf{\Sigma}$ por una matriz $\mathbf{\Gamma} \in \mathbb{R}^{p\times p}$ de rango r minimizando

$$\|\mathbf{\Sigma} - \mathbf{\Gamma}\| = \left[\sum_{i=1}^{m}\sum_{i=1}^{m}(\sigma_{ij} - \gamma_{ij})\right]$$

   (a) Muestre que

$$\|\mathbf{\Sigma} - \mathbf{\Gamma}\|^2 = \text{TR}(\mathbf{\Lambda} - \mathbf{P})(\mathbf{\Lambda} - \mathbf{P})^{\text{T}}$$

   donde $\mathbf{H}^{\text{T}}\mathbf{\Sigma}\mathbf{H} = \mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_m)$, $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$, $\mathbf{H}$ una matriz ortogonal y $\mathbf{P} = \mathbf{H}^{\text{T}}\mathbf{\Gamma}\mathbf{H}$.

   (b) Muestre que la matriz $\mathbf{\Gamma}$ de rango r que minimiza $\|\mathbf{\Sigma} - \mathbf{\Gamma}\|$ es

$$\mathbf{\Gamma} = \sum_{j=1}^{r}\lambda_j \mathbf{h}_j \mathbf{h}_j^{\text{T}}\,,$$

   donde $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_p]$.

2. Sea $\mathbf{x}$ un vector aleatorio de dimensión d con media $\mathbf{0}$, y matriz de dispersión $\mathbf{\Sigma} = (\sigma_{jk})$. Sea $v_j = \boldsymbol{\gamma}_j^{\text{T}}\mathbf{x}$ la j-ésima componente principal de $\mathbf{x}$. Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ los autovalores de $\mathbf{\Sigma}$.

   (a) Verifique que las y_j son no correlacionadas y que $\text{VAR}(y_j) = \lambda_j$

   (b) Muestre que la correlación entre x_j y v_ℓ es $\gamma_{\ell j}\sqrt{\lambda_\ell/\sigma_{jj}}$, donde $\gamma_{\ell,j}$ es el j-ésimo elemento de $\boldsymbol{\gamma}_\ell$.

3. Implemente en R el test del cociente de máxima verosimilitud para la hipótesis nula

$$H_{0,(r,h)} : \lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_{r+h}$$

contra la alternativa

$$H_{1,(r,h)} : \lambda_{r+1} > \lambda_{r+2} > \cdots > \lambda_{r+h}$$

4. Implemente en R el test bilateral para porcentajes para la hipótesis nula

$$H_{0,(r,h)} : \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} = p_0$$

contra la alternativa

$$H_{0,(r,h)} : \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \neq p_0 \, .$$

5. Considere el conjunto de datos `turtles`.

   (a) Calcule las rectas de regresión $X_2 = aX_1 + b$ y $X_2 = cX_1 + d$.

   (b) Grafíquelas junto con las observaciones.

   (c) Superponga la recta de cuadrados mínimos ortogonales.

   (d) Verifique que las tres rectas pasan por la media muestral y que la recta de cuadrados mínimos ortogonales se encuentra entre ambas rectas de regresión.

6. Considere la siguiente matriz de correlación obtenida por MacDonnell (1902) a partir de observaciones de siete características físicas de 3000 criminales.

$$R = \begin{pmatrix} \text{Head length} & 1.000 & & & & & \\ \text{Head breadth} & 0.402 & 1.000 & & & & \\ \text{Face breadth} & 0.396 & 0.618 & 1.000 & & & \\ \text{Left finger length} & 0.301 & 0.150 & 0.321 & 1.000 & & \\ \text{Left forearm length} & 0.305 & 0.135 & 0.289 & 0.846 & 1.000 & \\ \text{Left foot length} & 0.339 & 0.206 & 0.363 & 0.759 & 0.797 & 1.000 \\ \text{Height} & 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1.000 \end{pmatrix}$$

Halle las componentes principales. Cómo interpretaría las componentes obtenidas?

7. Sea $\mathbf{x} \in \mathbb{R}^p$ y $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$ 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 $i = 1, \ldots, k$ 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 $\mathbf{x}$.

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

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

Concluya que $\mathbf{\Sigma_x} - \mathbf{\Sigma_y}$ 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

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

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

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

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 $MSE(\mathbf{x}, \mathbf{z})$?

10. Sea $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$ donde

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

(a) Obtenga los autovalores $\lambda_1 > \lambda_2$ y autovectores $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ de $\mathbf{\Sigma}$.

(b) Muestre que $\mathbf{y}_{(1)} = \boldsymbol{\beta}_1 \boldsymbol{\beta}_1^{\mathrm{T}} \mathbf{x}$ y $\mathbf{y}_{(2)} = \boldsymbol{\beta}_2 \boldsymbol{\beta}_2^{\mathrm{T}} \mathbf{x}$ 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 la 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 $\mathbf{c} = (c, c/3)^{\mathrm{T}}$ 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

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

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

$$\begin{aligned} \mathbf{S}_{\text{POOLED}} &= \frac{1}{n}(\mathbf{Q}_1 + \mathbf{Q}_2 + \mathbf{Q}_3) \\ \mathbf{S} &= \frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}})^{\mathrm{T}} \qquad \overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{n_i}\mathbf{x}_{ij} \end{aligned}$$

   (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 $\widehat{\boldsymbol{\gamma}}_\ell$, $1 \leq \ell \leq p = 4$, contruya las componentes principales como

$$\mathbf{v}_{ij} = \widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i) \quad 1 \leq j \leq n_i \quad 1 \leq i \leq 3$$

  donde $\widehat{\boldsymbol{\Gamma}} = (\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_p)^{\mathrm{T}}$.
- 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 `microtus` de la `library(Flury)` consiste en las mediciones de huesos y dientes de ratones campestres (de la especie *Microtus*). Considere solamente las 43 ratas del grupo 1.

(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, medidos en mm./1000, para las 43 ratas del grupo 1.

i. Halle las componentes principales muestrales y los autovalores de S.

ii. 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.

iii. Realice un biplot.

(b) Realice un biplot con todas las variables.

(c) Repita el análisis anterior agregando a los datos las mediciones del grupo 2. Qué observa?

3. Considere el conjunto de datos que se encuentran en el archivo `Datos Paises.xlx`, correpondiente 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.