Métodos Multivariados: Tarea 5

Aldo Carmona, Diego Arellano, Mateo De La Roche, Victor Contreras

Ejercicio 1

Mostrar que la matriz de covarianzas ρ para las tres variables estandarizadas Z_1, Z_2, Z_3 puede ser generada por el modelo factorial con $m = 1$:

$$\rho = \begin{pmatrix} 1 & 0.63 & 0.45 \\ 0.63 & 1 & 0.35 \\ 0.45 & 0.35 & 1 \end{pmatrix}$$

Las ecuaciones del modelo factorial son:

$$\begin{aligned} Z_1 &= 0.9F_1 + \varepsilon \\ Z_2 &= 0.7F_1 + \varepsilon \\ Z_3 &= 0.5F_1 + \varepsilon \end{aligned}$$

donde $\text{Var}(F_1) = 1$, $\text{Cov}(F_1, \varepsilon) = 0$ y $\Psi =$

$$\begin{pmatrix} 0.19 & 0 & 0 \\ 0 & 0.51 & 0 \\ 0 & 0 & 0.75 \end{pmatrix}$$

Ejercicio 2

Se tiene la siguiente matriz de factores no rotada, obtenida utilizando el método de componentes principales y considerando 4 factores.

```
ej2_factores <- matrix(c( 0.881, 0.828, 0.664, 0.792, 0.731, 0.476,
-0.347, 0.508, -0.711, 0.564, -0.647, 0.804,
-0.165, -0.070,0.154, -0.179, 0.117, 0.329,
0.268, -0.200, -0.031, -0.029, -0.125, 0.135), nrow = 6)
```

```r
row.names(ej2_factores) <- paste0("X",1:6)
colnames(ej2_factores) <- paste0("F",1:4)
ej2_factores
```

```
      F1     F2     F3     F4
X1 0.881 -0.347 -0.165  0.268
X2 0.828  0.508 -0.070 -0.200
X3 0.664 -0.711  0.154 -0.031
X4 0.792  0.564 -0.179 -0.029
X5 0.731 -0.647  0.117 -0.125
X6 0.476  0.804  0.329  0.135
```

a. A partir de estos resultados, ¿se puede saber cómo agrupar las variables? Explicar.

b. ¿Cuál es la variable que más se identifica con las características? Explicar porqué.

Ejercicio 3

Verificar las siguientes identidades:

a. $(I + L'\Psi^{-1}L)^{-1}(I + L'\Psi^{-1}L) = I - (I + L'\Psi^{-1}L)^{-1}$

b. $L'(LL' + \Psi)^{-1} = (I + L'\Psi^{-1}L)^{-1}L'\Psi^{-1}$

Ejercicio 4

El siguiente ejemplo muestra un caso que se conoce como el caso de Heywood. Consideren un modelo factorial con $m = 1$ para la población con matriz de covarianza

$$\Sigma = \begin{pmatrix} 1 & 0.4 & 0.9 \\ 0.4 & 1 & 0.7 \\ 0.9 & 0.7 & 1 \end{pmatrix}$$

Mostrar que hay una solución única para L y Ψ con $\Sigma = LL' + \Psi$, pero que $\psi_3 < 0$, así que la elección no es admisible.

Ejercicio 5

Este ejercicio se basa en los datos de monitoreo atmosférico (REDMA) que se encuentran en GitHub. Los datos son series diarias de los contaminantes que están en el aire medidos en diferentes estaciones de monitoreo. La descripción de los datos la pueden encontrar en: esta liga, y el catálogo de estaciones está aquí. El archivo .zip contiene hojas de Excel con las mediciones diarias de 2019 y por hora, para cada una de las estaciones.

a. Hacer un análisis de estos datos, creando una base de datos con las mediciones de contaminación para cada estación de todos los contaminantes disponibles.

b. Hacer un análisis factorial exploratorio de estos datos. Interpretar y reportar los resultados: ¿Se pueden identificar factores?

c. Calcular los scores por el método de máxima verosimilitud y por el método de componentes principales.

d. ¿Se puede crear un índice de monitoreo ambiental que tome en cuenta todos los contaminantes? Si es así, ¿Cómo se puede interpretar su comportamiento a lo largo del tiempo?

Ejercicio 6

En un estudio sobre pobreza, crimen y disuasión, Parker y Smith (1979) reportan ciertas estadísticas de crimen en varios estados para los años 1970 y 1973. Una porción de su matriz de correlación es de la forma:

$$R = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right] = \left[\begin{array}{cc|cc} 1 & 0.615 & -0.111 & -0.266 \\ 0.615 & 1 & -0.195 & -0.085 \\ \hline -0.111 & -0.195 & 1 & -0.269 \\ -0.266 & -0.085 & -0.269 & 1 \end{array}\right]$$

Las variables son:

- $X_1^{(1)}$ = homicidios no primarios en 1973.
- $X_2^{(1)}$ = homicidios primarios en 1973 (homicidios que involucran familia).
- $X_1^{(2)}$ = severidad de castigo en 1970 (meses promedio de prisión)
- $X_2^{(2)}$ = probabilidad de castigo en 1970 (número de encarcelados entre número de homicidios)

a. Encontrar las correlaciones canónicas muestrales.

b. Determinar el primer par canónico $\hat{U}_1, \hat{V}_1$ e interpretar estas cantidades.

Ejercicio 7

Los datos que se usan en este ejercicio están relacionados con campañas de marketing directas de un banco portugués. Las campañas de marketing están basados en llama- das telefónicas. Con frecuencia, más de un contacto con el mismo cliente fue requerido, para accesar si el producto (depósito bancario a plazo) puede ser o no contratado. El archivo con la información relevante se puede obtener de la siguiente liga: bank.zip

Ejercicio 8

Se tienen tres medidas fisiológicas y tres variables de ejercicios medidas en 20 hombres de 30-40 años en un gimnasio. Los datos están en el archivo `FitnessClubdata.dat`.

Objetivo: determinar si las variables fisiológicas se relacionan de alguna forma con las variables de ejercicio.

a. Analizar la matriz de correlaciones relevantes entre las variables de los dos grupos (dentro y entre grupos de variables).

b. Probar la hipótesis $H_0 : \Sigma_{xy} = 0$.

Ejercicio 9

Una muestra aleatoria de $n = 70$ familias será encuestada para determinar la asociación entre ciertas variables 'demográficas' y ciertas variables de 'consumo'. Sea:

$$\text{Conjunto Criterio} \quad \begin{cases} X_1^{(1)} = \text{frecuencia anual de cena en restaurante} \\ X_2^{(1)} = \text{frecuencia anual ida al cine} \end{cases}$$

$$\text{Conjunto Predictor} \quad \begin{cases} X_1^{(2)} = \text{edad del jefe de familia} \\ X_2^{(2)} = \text{ingreso anual familiar} \\ X_3^{(2)} = \text{nivel de educación del jefe de familia} \end{cases}$$

Supongan que 70 observaciones de las variables precedentes dan una matriz de correlación muestral dada por:

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ 0.8 & 1 & & & \\ 0.26 & 0.33 & 1 & & \\ 0.67 & 0.59 & 0.37 & 1 & \\ 0.34 & 0.34 & 0.21 & 0.35 & 1 \end{bmatrix}$$

a. Determinar las correlaciones canónicas muestrales y probar la hipótesis $H_0 : \Sigma_{12} = 0$ (o equivalente $\rho_{12} = 0$ al nivel de 5%). Si se rechaza H_0, probar la significancia de la primera correlación canónica.

b. Usando las variables estandarizadas, construir las variables canónicas correspondientes a las correlaciones canónicas significativas.

c. Usando los resultados de las partes (a) y (b), preparar una tabla mostrando los coeficientes de las variables canónicas y las correlaciones muestrales de las variables canónicas con sus variables componentes.

d. Dada la información en (c), interpretar las variables canónicas.

e. ¿Tienen las variables demográficas algo que ver con las variables de consumo? ¿Las variables de consumo proveen mucha información sobre las variables demográficas?

Ejercicio 10

(Correlaciones para medidas angulares) Algunas observaciones tales como la dirección del viento, son en forma de ángulos. Un ángulo θ_2 puede ser representado como el par $x = (\cos(\theta_2), \sin(\theta_2))'$.

a. Mostrar que $x = \sqrt{b_1^2 + b_2^2}(\cos(\theta_2 - \beta)$ donde $b_1/\sqrt{b_1^2 + b_2^2} = \cos(\beta)$ y $b_2/\sqrt{b_1^2 + b_2^2} = \sin(\beta)$.

(Hint: $\cos(\theta_2 - \beta) = \cos(\theta_2)\cos(\beta) + \sin(\theta_2)\sin(\beta)$).

b. Sea $X^{(1)}$ con una única componente $X_1^{(1)}$. Mostrar que la correlación canónica simple es

$$\rho_1' = \max_{\beta} \text{Corr}(X_1^{(1)}, \cos(\theta_2 - \beta))$$

Selecciona la variable canónica V_1 tomando en cuenta seleccionar un nuevo origen β para el ángulo θ_2.

c. Sea $X_1^{(1)}$ el ozono (en partes por millón) y θ_2 = dirección de viento medida desde el norte. Se tomaron 19 observaciones en el centro de Milwaukee, Wisconsin, dando la matriz de correlaciones:

$$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.166 & 0.694 \\ 0.166 & 1 & -0.051 \\ 0.694 & -0.051 & 1 \end{pmatrix}$$

Encontrar la correlación canónica muestral r_1' y la variable canónica $\hat{V}_1$, representando el nuevo origen β.