Maestría en Computo Estadístico
Estadística Multivariada
Tarea 5
28 de mayo de 2021
Enrique Santibáñez Cortés
Repositorio de Git: Tarea 5, EM.

Ejercicio 1.

Otra forma de derivar los resultados del análisis de correspondencia simple es encontrando una matriz $\hat{\mathbf{P}}$ de dimensión $r \times s$ con rango reducido $t < \text{mín}(r, s)$ que aproxime $\mathbf{P}$ minimizando el criterio de mínimos cuadrados ponderados:

$$\text{tr}\{\mathbf{D}_r^{-1/2}(\mathbf{P} - \hat{\mathbf{P}})\mathbf{D}_c^{-1}(\mathbf{P} - \hat{\mathbf{P}})'\mathbf{D}_r^{-1/2}\}.$$

Usando el teorema de Eckart-Young, encuentre la matrix $\hat{\mathbf{P}}$ que arroje la mejor aprroximación de rango reducido de $\mathbf{P}$ en este sentido. Muestre que la mejor aproximación de rango 1 de $\mathbf{P}$ es la solución trivial $\hat{\mathbf{P}} = \mathbf{r}\mathbf{c}'$.

RESPUESTA:

Teorema: 1 ***Teorema de Eckart-Young*** *. Sea A una matriz de tamaño $m \times k$ con $m \geq k$ y con descomposición de valores singulares $U\Lambda V'$. Sea $s < k = rank(A)$. Entonces*

$$B = \sum_{i=1}^{s} \lambda_i u_i v_i,$$

es la aproximación de mínimos cuadrados de rango s de A. Y minimiza

$$tr\left[(A - B)(A - B)'\right]$$

de todas la matrices de tamaño $m \times k$ que tienen un rango no mayor que s.

Recordemos las definiciones algunas matrices a utilizar. Si n es el número total de frecuencias en la matriz de datos X, y sea $P = \{p_{ij}\} = \{\frac{x_{ij}}{n}\}$ la matriz de correspondencias. Ahora podemos definir los vectores de las sumas de los renglón y columna como $\mathbf{r}$ y $\mathbf{c}$ respectivamente. Y las matrices diagonales $\mathbf{D_c}$ y $\mathbf{D_r}$ con los elementos de los vectores $\mathbf{r}$ y $\mathbf{c}$ en la diagonal respectivamente. Entonces,

$$r_i = \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n}, \quad i = 1, 2, \cdots, I, \quad \Leftrightarrow \quad r_{(I\times 1)} = P_{(I\times J)}1_{(J\times 1)}$$
$$c_j = \sum_{i=1}^{I} p_{ij} = sum_{i=1}^{I} \frac{x_{ij}}{n}, \quad j = 1, 2, \cdots, J, \quad \Leftrightarrow \quad c_{(J\times 1)} = P'_{(J\times I)}1_{(I\times 1)}$$

donde $\mathbf{1}$ es una vector de unos y

$$\mathbf{D}_r = diag(r_1, r_2, \cdots, r_I), \quad y \quad \mathbf{D}_c = diag(c_1, c_2, \cdots, c_J). \tag{1}$$

Y definimos la raiz cuadrada de las matrices como

$$\mathbf{D}_r^{1/2} = diag(\sqrt{r_1}, \sqrt{r_2}, \cdots, \sqrt{r_I}), \quad y \quad \mathbf{D}_r^{-1/2} = diag(1/\sqrt{r_1}, 1/\sqrt{r_2}, \cdots, 1/\sqrt{r_I}) \tag{2}$$
$$\mathbf{D}_c^{1/2} = diag(\sqrt{c_1}, \sqrt{c_2}, \cdots, \sqrt{c_J}) \quad y \quad \mathbf{D}_c^{-1/2} = diag(1/\sqrt{c_1}, 1/\sqrt{c_2}, \cdots, 1/\sqrt{c_J}). \tag{3}$$

Con todo lo anterior, primero definamos a la matriz $\mathbf{B} = \mathbf{D_r^{-1/2} P D_c^{-1/2}}$. Ahora, del **Teorema 1** (Johnson and Wichern 2007), sabemos que la mejor $\hat{\mathbf{B}}$ aproximación de menor rango s de $\mathbf{B}$ esta dada por los primeros s terminos de la descomposición de valores singulares

$$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} = \sum_{k=1}^{J} \lambda_k^* u_k^* v_k'^*,$$

donde

$$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} v_k^* = \lambda_k^* u_k^*, \quad y \quad u_k'^* \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} = \lambda_k^* v_k'^* \tag{4}$$

y

$$|(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2})(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2})' - \lambda_k^{2*} I| = 0, \quad \text{para} k = 1, \cdots, J.$$

Es decir, la mejor aproximación de B con rango menor s es

$$\hat{\mathbf{B}} = \mathbf{D}_r^{-1/2}\hat{\mathbf{P}}\mathbf{D}_c^{-1/2} = \sum_{k=1}^{s} \lambda_k^* u_k^* v_k'^*.$$

Y por lo tanto, podemos encontrar la aproximación para P como

$$\hat{\mathbf{P}} = \mathbf{D}_r^{1/2}\hat{\mathbf{B}}\mathbf{D}_c^{1/2} = \sum_{k=1}^{s} \lambda_k^* (\mathbf{D}_r^{1/2} u_k^*)(\mathbf{D}_c^{-1/2} v_k^*)'. \tag{5}$$

Veamos que si consideramos a $u_1^* = \mathbf{D}_r^{1/2} 1_I$ y $v_1^* = \mathbf{D}_c^{1/2} 1_J$, comprobemos primero que cumplan (4)

$$\begin{aligned}
\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} v_k^* &= \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}(\mathbf{D}_c^{1/2} 1_J) \\
&= \mathbf{D}_r^{-1/2}\mathbf{P} 1_J \\
&= \mathbf{D}_r^{-1/2}\mathbf{r} \\
&= \mathbf{D}_r^{1/2} 1_I = u_1^*
\end{aligned}$$

y

$$\begin{aligned}
u_k'^* \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} &= (\mathbf{D}_r^{1/2} 1_I)' \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} \\
&= 1_I' \mathbf{P}\mathbf{D}_c^{-1/2} \\
&= c' \mathbf{D}_c^{-1/2} \\
&= (\mathbf{D}^{1/2} I_J)' = v_k'^*.
\end{aligned}$$

Entonces podemos decir, que $(u_1^*, v_1^*) = (\mathbf{D}_r^{1/2} 1_I, \mathbf{D}_c^{1/2} 1_J)$ son vectores singulares asociados con el valor singular $\lambda_1^* = 1$. Y entonces podemos ver que

$$\lambda_1^* (\mathbf{D}_r^{1/2} u_1^*)(\mathbf{D}_c^{-1/2} v_1^*)' = \lambda_1^* (\mathbf{D}_r^{1/2}\mathbf{D}_r^{1/2} 1_I)(\mathbf{D}_c^{-1/2}\mathbf{D}_c^{1/2} 1_J)' = D_r 1_I 1_J' D_c = \mathbf{r}\mathbf{c}'. \tag{6}$$

Por lo tanto, podemos reescribir la expresión (5) como cuando $s \geq 2$.

$$\hat{\mathbf{P}} = \mathbf{r}\mathbf{c}' + \sum_{\mathbf{k=2}}^{\mathbf{s}} \lambda_{\mathbf{k}}^* (\mathbf{D_r^{1/2} u_k^*})(\mathbf{D_c^{-1/2} v_k^*})'$$

Entonces, debido a que $(u_1^*, v_1^*) = (\mathbf{D}_r^{1/2} 1_I, \mathbf{D}_c^{1/2} 1_J)$ son vectores singulares asociados con el valor singular $\lambda_1^* = 1$ y esto implica la expresión (6), **podemos concluir la mejor aproximación de rango 1 de P es la solución $\hat{\mathbf{P}} = \mathbf{r}\mathbf{c}'$.** $\blacksquare$.

Ejercicio 2.

El conjunto de datos **mundodes** representa 91 países en los que se han observado 6 variables, Razón de natalidad, Razón de mortalidad, mortalidad infantil, esperanza de vida en hombres, esperanza de vida en mujeres y PNB per cápita. Del conjunto de datos se ha tomado la esperanza de vida de hombres y de mujeres. Se han formado cuatro categorías tanto para la mujer como para el hombre. Se denotan por M1 y H1 a las esperanzas entre menos de 41 años a 50 años, M2 y H2, de 51 a 60 años, M3 y H3, de 61 a 70 años, y M4 y H4, para entre 71 a más de 80. La siguiente tabla de contingencia muestra las frecuencias de cada grupo

Mujer / Hombre	H1	H2	H3	H4
M1	10	0	0	0
M2	7	12	0	0
M3	0	5	15	0
M4	0	0	23	19

Cuadro 1: Conjunto de datos mundodes

Realiza proyecciones por filas, por columnas y conjuntas de filas y columnas. Comprobar que en la proyección por filas las categorías están claramente separadas y que en el caso del hombre, las dos últimas categorías están muy cercanas. Comprobar en la proyección conjunta la cercanía de las categorías H3 con M3 y M4.

RESPUESTA:

Primero evaluemos la independencia los grupos de mujeres y hombres para diferentes rangos de edad, para ello cuparemos la prueba de χ^2.

```r
# datos del problema.
mundodes <- matrix(c(10,0,0,0,
                     7, 12 ,0,0,
                     0, 5, 15, 0,
                     0, 0, 23, 19), nrow=4, byrow=T)
n <- sum(mundodes) # número de registros

# prueba de chi cuadrada para probar independecia.
chisq.test(mundodes)
```

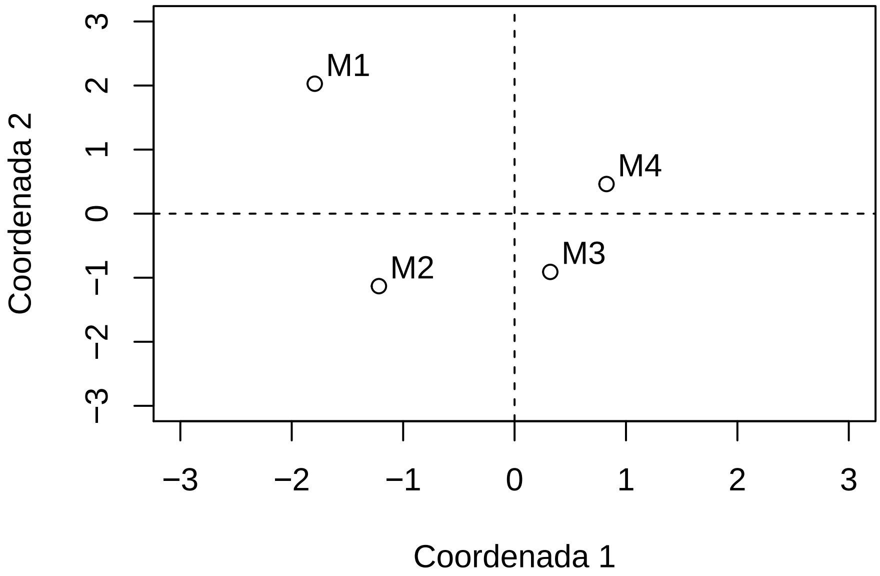
```
##
##  Pearson's Chi-squared test
##
## data:  mundodes
## X-squared = 121.86, df = 9, p-value < 2.2e-16
```

Cómo el p-value es menor a 0.05, podemos rechazar la hipótesis nula de independencia entre los grupos. Y por lo tanto podemos concluir que existe algún tipo de asociación entre ellos. Ahora, procedemos a calcular las proyecciones por filas y columnas

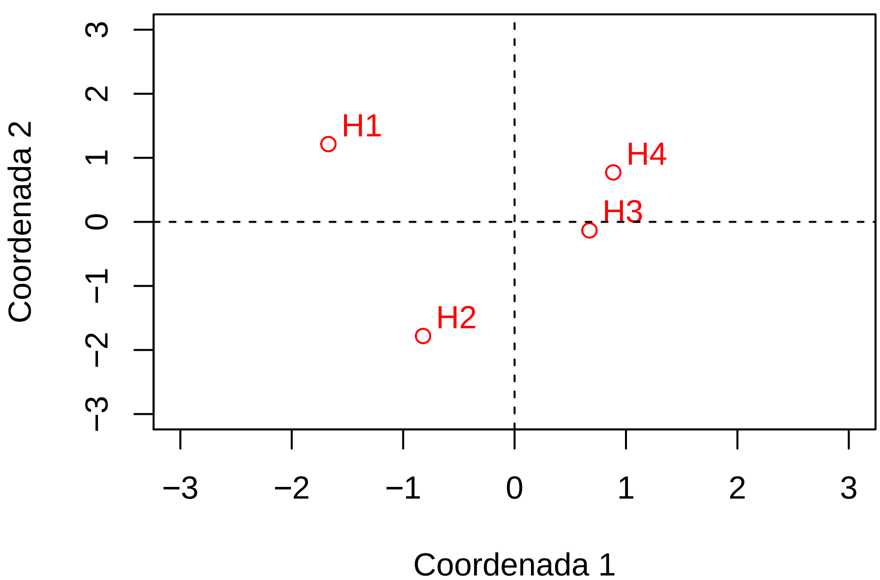
```r
library(ca)
corres<-ca(mundodes, nd = 2)

Cr<-corres$rowcoord #coordenadas de las filas
```

```r
plot(Cr,xlim=range(-3,3),ylim=range(-3,3),
     xlab="Coordenada 1",ylab="Coordenada 2",lwd=1)
text(Cr+0.3,labels=c("M1","M2","M3","M4"),col=1,lwd=2)
abline(h=0,lty=2)
abline(v=0,lty=2)
```




```r
Cc<-corres$colcoord  #coordenadas de las columnas
plot(Cc,xlim=range(-3,3),ylim=range(-3,3),
     xlab="Coordenada 1",ylab="Coordenada 2",lwd=1, col=2)
text(Cc+0.3,labels=c("H1","H2","H3","H4"),col=2,lwd=4)
abline(h=0,lty=2)
abline(v=0,lty=2)
```

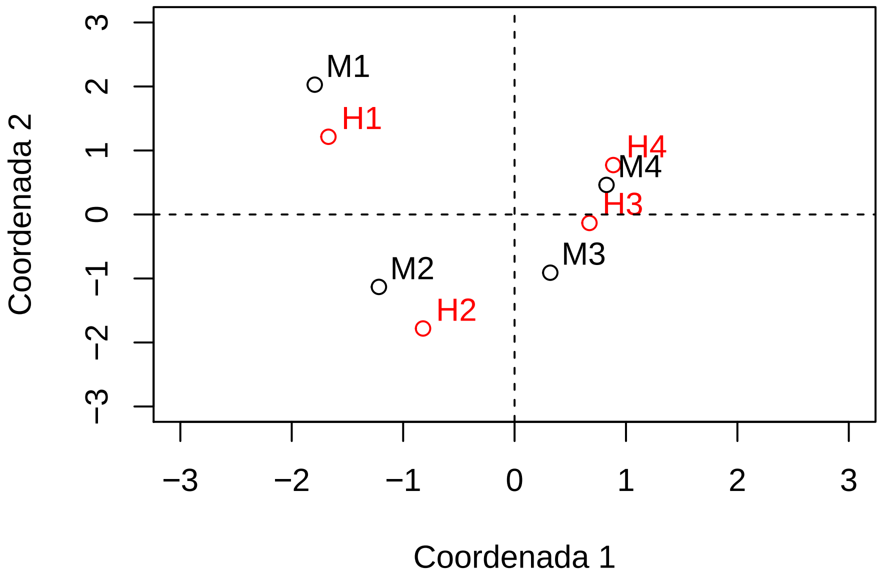


De las gráficas anteriores no podemos indicar que las filas tengan un perfil similar a través de las columnas, ni que las columnas tienen un perfil similar a través de las filas. Pero podemos interpretar las componentes 1 como la esperanzas de las personas menores de 60 años.

Ahora, se obtiene la representacion conjunta de los renglones y columnas en el espacio de dos dimensiones

```
plot(Cr,xlim=range(-3,3),ylim=range(-3,3),
     xlab="Coordenada 1",ylab="Coordenada 2",lwd=1)
points(Cc,col=2)
text(Cr+0.3,labels=c("M1","M2","M3","M4"),col=1,lwd=2)
text(Cc+0.3,labels=c("H1","H2","H3","H4"),col=2,lwd=4)

abline(h=0,lty=2)
abline(v=0,lty=2)
```



Entocnes como las puntos de las mujeres están cercanos a los hombres con el mismo rango de edad, es decir no hay independencia entre los diferentes rangos de esperanza de vida entre hombre y mujeres. En esta representación se observa la cercaní de las categorías H3 con M3 y M4.

Bibliografía

Johnson, Richard Arnold, and Dean W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. 6. ed. Upper Saddle River, NJ: Prentice Hall. http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+330798693&sourceid=fbw_bibsonomy.