

Análisis de Correspondencias Múltiples (ACM)

Hamdi Raïssi

IES PUCV

hamdi.raïssi@pucv.cl

Objetivos de esta parte

Objetivo :

- a- Resumir información obtenidas en encuestas.
- b- Analizar los hechos sobresalientes de este tipo de datos.

El cuadro de los datos y anotaciones

Queremos estudiar un conjunto de datos con la estructura siguiente :

\times	Var j	...
1	z_{1j}	...
\vdots
i	z_{ij}	...
\vdots
I	z_{Ij}	...

- Tenemos J variables cualitativas con K_j , $j = \{1, \dots, J\}$, modalidades cada una.
- I individuos.
- Esta tabla es entregada al software para análisis.
- Ejemplos : **Variable consumo** : z_{ij} = marca de un producto, **Salud** : z_{im} = hospital/clinica/centro de salud, **Ocupación** : z_{ip} = cesante, trabaja mitad del tiempo, todo el tiempo,etc

El cuadro de los datos y anotaciones

El software utiliza este tipo de datos :

	Variable j					...
\times	1	...	k	...	K_j	...
1	0	0	1	...	0	...
\vdots
i	0	1	0	...	0	...
\vdots
i'	1	...	0	0	0	...
\vdots
l	0	...	0	0	1	...

- Vamos utilizar la variable que tiene 1 si el individuo i tiene una cierta modalidad y 0 sino
- **Por una variable y un individuo** podemos solo tener un 1 y todo el resto 0.

El cuadro de los datos y objetivos

- Dos individuos que tienen los mismos 1 y 0 son iguales del punto de vista de la encuesta.
- La variabilidad de los individuos viene de su diferencias.

Vamos a trabajar con variables para describir las diferencias :

- El cuadro anterior tiene $\mathbf{K} = \sum_{j=1}^J K_j$ columnas (o modalidades en total).
- Por cada columna k (modalidad), $k = 1, \dots, \mathbf{K}$ y fila i (individuo), $i = 1, \dots, I$, tenemos una variable $y_{ik} = 0, 1$.

El cuadro de los datos y anotaciones

El software utiliza este tipo de datos :

	Variable j					...
\times	1	k
1	0	0	1	...	0	...
\vdots
i	0	1	$y_{ik} = 0$...	0	...
\vdots
j	1	...	0	0	0	...
\vdots
l	0	...	0	0	1	...

Objetivos : Construir perfiles por los individuos y variables

Ejemplos de perfiles de individuos :

- Individuos en redes sociales para difundir publicidad en público determinado.
- Perfiles de consumidores.
- Perfiles de opiniones políticas.
-etc...

Ejemplos de perfiles de modalidades :

- Las modalidades "fuma/enfermo" y "no fuma/sano" podrían tener una cercanía.
-etc...

Vamos a proyectar los individuos y las variables para visualizar los hechos

Medir distancias entre individuos

- Tenemos variables $y_{ik} = 0, 1$ que describen los I individuos.
 - La idea es de **destacar** los individuos que tienen una modalidad que **no es común**.
- ⇒ $\tilde{x}_{ik} = y_{ik}/p_k$ donde p_k es la frecuencia de individuos que tiene 1 por la columna k .

- Tenemos :

$$\bar{\tilde{x}}_k = \frac{1}{I} \sum_{i=1}^I y_{ik}/p_k = 1.$$

- Vamos a trabajar con variables centradas : $x_{ik} = y_{ik}/p_k - 1$.

Medir distancias entre individuos

Tenemos :

- Un nube de puntos de los individuos en R^K .
- Un nube de puntos de las modalidades.

Nube de puntos de los individuos : medir distancias entre individuos

Distancia entre individuos :

$$d^2(i, i') = \sum_{k=1}^K \frac{p_k}{J} (x_{ik} - x_{i'k})^2 = \frac{1}{J} \sum_{k=1}^K \frac{1}{p_k} (y_{ik} - y_{i'k})^2.$$

Dar más peso a las modalidades pocas observadas.

- Proyectar esta nube maximizando su inercia (distancia a su centro de gravedad).
- Buscar los componentes de manera secuencial.

Interpretación de los componentes a la luz de los individuos

- Estudiar la posición de unos **extremos individuos** nos da la interpretación de los componentes.
- **Ejemplo** : A la izquierda individuos sin hobbies, a la derecha con muchos hobbies : *Eje horizontal=tendencia a tener hobbies.*

Interpretación de los componentes a la luz de los individuos

- Los coordenados de los puntos en los ejes son variables cuantitativas
 - Las variables de partida son categoricas.
- ⇒ Utilizamos herramientas de la ANOVA a un factor ("one-way-ANOVA").
- En ANOVA de un factor la razón de correlación mide la relación entre una variable cualitativa y cuantitativa.
 - Vamos a medir así la relación entre los ejes (con el cuadrado de la razón de correlación).
 - Obtenemos el cuadro de las razones de correlaciones.

Proyección de las modalidades

- Tenemos una nube de puntos en \mathbb{R}^I .
- Medición de la distancia entre dos modalidades k y k' :

$$d^2(k, k') = \sum_{i=1}^I \left(\frac{y_{ik}}{p_k} - \frac{y_{ik'}}{p_{k'}} \right)^2$$

- Proyección con maximización de la inercia (o varianza) de las distancias entre las modalidades y el centro de gravedad (el origen).

Proyección de las modalidades

- A muchas veces tenemos muchas modalidades
- ⇒ La inercias de los ejes de un ACM es bastante baja en comparación de un ACP.
- Tal como el ACP o el AC podemos tener variables **activas y pasivas** en la construcción del ACM.
- Ejemplo : Variables activas : *hobbies, trabajo*. Variables pasivas : *sexo, nivel economico, sueldo* (la última es cuantitativa...).