TRABAJO BUSCAR ANALISIS DISCRIMINANTE

El trabajo es extraido del trabajo de (Cabedo Nebot 2021) y (Fernández Avilés and Montero 2024) en el que se considero las partes esenciales para el desarrollo de la presente monografía

Resumen

El análisis múltiple de correspondencia (AMC) es una técnica exploratoria utilizada para visualizar la relación entre más de dos variables categóricas, el caso de solo dos variables es el análisis de correspondencias o análisis de correspondencia simple. Para usar en R se necesita la librería **FactoMineR** y **ca**. El AMC es útil para identificar coocurrencias entre categorías sin necesidad de tener una variable independiente y puede complementarse con el análisis de clúster para agrupar elementos derivados del análisis.

Metodología

Generalmente parte de una tabla de contingencia que se muestra a continuación

	B_1	B_2	...	B_C	**Total**
A_1	n_{11}	n_{12}	...	n_{1C}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2C}	$n_{2.}$
...	...	...	...	...	...
A_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.C}$	N

Tabla 35.1: Ejemplo de tabla de contingencia $R \times C$

Dada una tabla de contingencia, a partir de las frecuencias observadas n_{ij}, se definen las distancias entre perfiles

- Para los perfiles fila, $d_{ii'} = \sum_{j=1}^{C} \frac{1}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2$
- Para los perfiles columna, $d_{jj'} = \sum_{i=1}^{R} \frac{1}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}} \right)^2$

Cuanto más se diferencien unos perfiles de otros, más grandes serás las diferencias anteriores. El análisis de correspondencias busca construir **dimensiones** (habitualmente, de dos) y obtener las coordenadas de los niveles de ambos factores en dichas dimensiones:

$$\mathbf{A} = (\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_R)$$

Con $\mathbf{a}_i = \begin{pmatrix} a_{i1} \\ a_{i2} \end{pmatrix}$ y

$$\mathbf{B} = (\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_C)$$

Con $\mathbf{a}_j = \begin{pmatrix} b_{j1} \\ b_{j2} \end{pmatrix}$

siendo $\mathbf{a}_i$ las coordenadas del nivel fila $\mathbf{A}_i$ y $\mathbf{b}_j$ las del nivel columna $\mathbf{B}_j$ en el plano, de forma que reproduzcan las distancias entre perfiles y columna y los residuos estandarizados (asociaciones):

$$d(\mathbf{a}_i, \mathbf{a}_{i'}) = \sqrt{(a_{i1} - a_{i'1})^2 + (a_{i2} - a_{i'2})^2} \approx d_{ii'},$$

$$d(\mathbf{b}_j, \mathbf{b}_{j'}) = \sqrt{(b_{j1} - b_{j'1})^2 + (b_{j2} - b_{j'2})^2} \approx d_{jj'},$$

$$\mathbf{a}_i' \mathbf{b}_j \approx r_{ij}$$

Una vez en disposición de las coordenadas contenidas en las matrices $\mathbf{A}$ y $\mathbf{B}$ es posible "visualizar" la posición relativa de cada factor en las nuevas dimensiones. Esta estructura permite ver tanto las "distancias" que hay entre los niveles de cada factor (mediante la distancia de representación en el plano) como las "asociaciones" entre niveles de ambos factores (ya que mientras más asociación haya, más cerca se representarán en el plano).

Para resolver el problema de la estimación de las matrices $\mathbf{A}$ y $\mathbf{B}$ se lleva a cabo una descomposición de la matriz $\mathbf{R} = (r_{ij})$ en valores singulares.

Según la importancia que se dé al ajuste de uno de los perfiles o a la matriz de residuos, se tienen diferentes métodos de selección, llamados **normalizaciones**.

Proyecciones fila, columna y simétrica

El punto de partida es la matriz de frecuencias relativas $\mathbf{F}$, cuyas entradas son $f_{ij} = n_{ij}/N$, también llamada matriz de correspondencias. Definiendo el vector de unos, $\mathbf{1}$, con la dimensión adecuada, las masas, o frecuencias marginales, de filas y columnas, $r_i = f_i = \sum_{j=1}^{C} f_{ij}$ y $c_{.j} = f_{.j} = \sum_{i=1}^{R} f_{ij}$, respectivamente, se pueden expresar matricialmente como $\mathbf{r} = \mathbf{F1}$ y $\mathbf{c} = \mathbf{F}'\mathbf{1}$ o, en forma de matrices diagonales, como:

$$\mathbf{D}_R = diag(r) \equiv diag(r_1, \cdots, r_R) \text{ y } \mathbf{D}_C = diag(c) \equiv diag(c_1, \cdots, c_C)$$

Se calcula la ***matriz de residuos estandarizados*** como:

$$\mathbf{R}_{est} = \mathbf{D}_R^{-\frac{1}{2}} (\mathbf{F} - \mathbf{r}\mathbf{c}') \mathbf{D}_C^{-\frac{1}{2}}$$

La matriz $\mathbf{R}_{est}$ se descompone en valores singulares, calculando las matrices $\mathbf{U}$, $\mathbf{D}$ y $\mathbf{V}$ tales que:

$$\mathbf{R} = \mathbf{U}\mathbf{D}\mathbf{V}',$$

$$\mathbf{U}\mathbf{U}' = \mathbf{V}'\mathbf{V} = \mathbf{I}, \quad \mathbf{U}_{(RxK)}, \quad \mathbf{V}_{(CxK)}$$

$$K = min(R - 1, C - 1)$$

$$\mathbf{D} = diag(\mu_1, \cdots, \mu_K),$$

Donde los μ_i son los ***valores singulares*** (autovectores), estando ordenados de forma decreciente $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. A partir de la descomposición se pueden obtener:

- Las coordenadas estándar de las filas, $\Phi = \mathbf{D}_R^{-\frac{1}{2}}\mathbf{U}$ y sus coordenadas principales, $\mathbf{H} = \Phi\mathbf{D}$.
- Las coordenadas estándar de las columnas, $\Gamma = \mathbf{D}_C^{-\frac{1}{2}}\mathbf{V}$ y sus coordenadas principales, $\mathbf{G} = \Gamma\mathbf{D}$.
- Las inercias principales, $\lambda_i = \mu_i^2$.

Las coordenadas estándar permiten representar los perfiles en un plano, pero no permiten una comparación fácil entre perfiles fila columna. Para evitar este efecto, se escalan, dando lugar a las coordenadas principales, utilizadas para definir las proyecciones fila y proyecciones columna, que representan los correspondientes perfiles, formando los llamados mapas asimétricos. Las inercias principales indican el grado de variabilidad entre los perfiles fila o columna y los respectivos vectores de medias, por lo que tienen una interpretación equivalente a la variabilidad explicada por cada componente principal en el análisis de componentes principales.

Por último, las matrices $\mathbf{A} = \mathbf{D}_R^{-\frac{1}{2}}\mathbf{UD}$ y $\mathbf{B} = \mathbf{D}_C^{-\frac{1}{2}}\mathbf{VD}$ representan las coordenadas de ambos perfiles en un espacio común, llamado ***mapa simétrico.***

Referencias

Cabedo Nebot, Adrián. 2021. *Estadística Aplicada Con R: Visualización y Validación de Datos Poblacionales Pragmáticos y Fonéticos*. España: Universitat de Valéncia.

Fernández Avilés, Gema, and José María Montero. 2024. *Fundamentos de Ciencia de Datos Con R*. McGraw-Hill.