Análisis de Correspondencias

1 Introducción

El objetivo es desplegar en una gráfica en dos dimensiones las categorías de las variables de clasificación de una tabla de contingencia de tamaño $I \times J$, como puntos, tales que la distancia euclideana entre ellos sea la distancia Ji cuadrada entre las categorías.

Cuando se tienen dos criterios o variables de clasificación de las observaciones, al hacer el "cruce" de éstas se genera una tabla de frecuencias como sigue:

$$\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1j} & \cdots & n_{1J} \\ n_{21} & n_{21} & \cdots & n_{2j} & \cdots & n_{2J} \\ \vdots & \vdots & & \vdots & & \vdots \\ n_{i1} & n_{i2} & \cdots & n_{ij} & \cdots & n_{iJ} \\ \vdots & \vdots & & \vdots & & \vdots \\ n_{I1} & n_{I2} & \cdots & n_{Ij} & \cdots & n_{IJ} \end{bmatrix}$$

Donde la variable renglón tiene I categorías y la variables columna tiene J categorías.

2 Notación

Los totales por columna son:

$$\sum_{i=1}^{I} n_{ij} = n_{\cdot j}$$

los totales por renglón:

$$\sum_{j=1}^{J} n_{ij} = n_{i\cdot}$$

el total general :

$$\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} = n$$

y las frecuencias relativas: $r_i = n_{i.}/n$ y $c_j = n_{.j}/n$
Con $i = 1, \cdots, I$y$j = 1, \cdots, J$

3 La χ^2

Para analizar si las variables de clasificación son independientes se puede utilizar la estadística χ^2, dada por

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Cuando esta estadística toma valores "grandes" se rechaza la hipótesis nula de independencia.

4 Geometría

En el siguiente siguiente desarrollo se verá que la χ^2 es una suma ponderada de distancias.

Sea

$$p_{ij} = \frac{n_{ij}}{n}$$

$$\chi^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

Factorizando la r_i que está dentro del cuadrado se puede reescribir como:

$$= n\sum_{i=1}^{I} r_i \underbrace{\sum_{j=1}^{J} \frac{(\frac{p_{ij}}{r_i} - c_j)^2}{c_j}}_{d_i^2}$$

La cantidad d_i es precisamente una distancia *ponderada* entre el *perfil del renglón* f_i:

$$f_i = (\frac{p_{i1}}{r_i}, \frac{p_{i2}}{r_i}, \cdots, \frac{p_{iJ}}{r_i}) = (\frac{n_{i1}}{n_{i.}}, \frac{n_{i2}}{n_{i.}}, \cdots, \frac{n_{iJ}}{n_{i.}})$$

y el punto *centroide* c:

$$c = (c_1, c_2, \cdots, c_J) = (\frac{n_{.1}}{n}, \frac{n_{.2}}{n}, \cdots, \frac{n_{.J}}{n})$$

Esta distancia pondera cada coordenada por $1/c_j$ y se le conoce como distancia "ji cuadrada" entre f_i y c.

A la cantidad

$$\frac{\chi^2}{n} = \sum_{i=1}^{I} r_i \sum_{j=1}^{J} \frac{(\frac{p_{ij}}{r_i} - c_j)^2}{c_j}$$

se le conoce como *inercia total*, lleva este nombre por la fórmula de física que dice masa $\times$ distancia^2 = inercia, en este caso las masas son las r_i.

Vale la pena recordar que las χ^2 grandes llevan a rechazar la hipótesis de independencias, y esto ocurre cuando los perfiles renglón se alejan del centroide, bajo la métrica ji cuadrada. Entonces la estadística χ^2 puede pensarse como una una medida de dispersión de los perfiles hacia el centroide.

Los perfiles renglón y el centroide son pues puntos en un espacio de dimensión $J-1$ (pues cada renglón suma uno), como el objetivo de esta técnica es tener una gráfica, sólo hace falta proyectar estos I puntos en un plano. El plano se selecciona de manera tal que se minimice la suma de distancias de estos I puntos perfil al plano, y que por supuesto contiene al centroide. Pero como las distancias entre los perfiles y el centroide son fijas y por el teorema de pitágoras equivale a maximizar la suma de las p_i los momentos de inercia, que corresponden a las distancias entre los perfiles proyectados y el centroide.

En este curso no se verá con detalle como se llega a las fórmulas para obtener las coordenadas de los puntos perfil proyectados en este plano en especial, sin embargo si les doy las fórmulas.

Sea $P = [p_{ij}]$ de dimensión $I \times J$, se puede escribir a r y a c como $r = P1$ y $c = P'1$ respectivamente. Defino además las matrices diagonales $D_r = diag(r)$ de dimensión $J \times J$ y D_c de dimensión $I \times I$ y la matriz $E = D_r^{-1/2}(P - rc')D_c^{-1/2}$ de dimensión $J \times I$. Note que los elementos de esta matriz E son los sumandos de la estadística χ^2.

De manera análoga a componentes principales se requiere descomponer a E para obtener las nuevas coordenadas. A E se le descompone como $E = UDV$ donde U y V son matrices ortogonales y D es una matriz diagonal, a esto se le llama descomposición de valor singular. Las columnas de U son los eigenvectores de$E'E$ y las de V los eigenvectores de EE', los eigenvalores de las matrices $E'E$ y EE' coinciden y son p, que es el rango de la matriz $[n_{ij}], p = \min[(I-1),(J-1)]$, estos valores singulares $d_1, d_2, \cdots, d_p$ forman la diagonal de la matriz D se conocen como las *inercias principales*, y la suma de sus cuadrados es la *inercia total*.

$$\sum_{i=1}^{p} d_i^2 = \chi^2/n = \text{inercia total}$$

Los perfiles renglón f_i resultan ser los renglones de la matriz

$$F = D_r^{1/2} UD$$

de dimensión $I \times p$.

Entonces para representar los perfiles renglón en el plano tomamos las dos primeras columnas de la matriz U digamos $U_{(2)}$ y la submatriz $D_{(2)}$ con los dos primeros valores singulares de la matriz D.

Finalmente la representación en dos dimensiones está dada por

$$F_{(2)} = D_r^{1/2} U_{(2)} D_{(2)} \text{de dimensión} I \times 2$$

El mismo análisis que se hizo con los perfiles renglón se puede hacer con los perfiles columna g_i que corresponden a los renglones de la matriz $G = D_c^{1/2} VD$de dimensión$J \times p$.

Este análisis dual se da por que existen las siguientes relaciones: $G = D_c^{-1} P' F D^{-1}$ y $F = D_r^{-1} PGD^{-1}$

De manera análoga la representación en dos dimensiones para las columnas se hace tomando las dos primeras columnas de V digamos $V_{(2)}$ y haciendo

$$G_{(2)} = D_c^{1/2} V_{(2)} D_{(2)} \text{de dimensión} J \times 2$$

5 Observaciones

Los puntos renglón y los puntos columna se pueden poner en la misma gráfica pues ambas representaciones tienen las mismas inercias principales (los mismos d_1 y d_2).

Para saber si un punto está bien representado en el plano, nos fijamos en los ángulos que se forman entre el vector que va del centroide al perfil en cuestión y los ejes principales del plano, si estos ángulos son pequeños ($cos(\theta) \approx 1$) el punto estará cerca del plano y su representación será buena.

Es importante señalar que no se puede hablar de distancias entre puntos columna y puntos renglón.

Si se tiene una buena representación, con $(d_1 + d_2)/\sum_{k=1}^{p} d_i \approx 1$, y además los $cos(\theta_s) \approx 1$ y ocurre que un punto renglón aparece muy cerca de un punto columna podemos decir que estas dos categorías están asociadas.

6 Bibliografía

Greenacre,M.J. (1984). *Theory and Applications of Correspondance Analysis.* Academic Press London. QA278 G744
Barnett, V. (1981) *Interpreting Multivariate Data.* John Willey Bath. QA 278 B36
Krzanowski, W.J. y Marriot, F. H. C. (1994). *Multivariate Analysis Part 1.* Distributions, Ordination and Inference. Eduard Arnold London.