ANALYSIS DE CORRESPONDENCIAS MULTIPLES SOBRE UN GRAFO

Tomás ALUJA BANET
Manuel MARTI RECOBER

RR85/10

Juliol 1985

RESUM

En anàlisi de dades sovint hom analitza matrius de dades formades per variables nominals, correlacionades amb unes altres, anomenades variables "instrumentals", l'efecte de les quals desitgem eliminar de l'anàlisi. En aquests casos realitzem una anàlisi de correspondències múltiples sobre el graf de similitud definit per les variables instrumentals. Els plans factorials obtinguts reflecteixen les oposicions entre variables apart de (o havent fixat) l'efecte de les variables instrumentals. Anomenem a aquesta anàlisi "parcial", pel fet de que és basat en la mateixa idea que l'anàlisi clàssica de correlacions parcials o en les variables insstrumentals de Rao, sense necessitar però, les fortes hipòtesis d'aquestes anàlisis.

ANALISIS DE CORRESPONDENCIAS MULTIPLES SOBRE UN GRAFO

Tomas Aluja Banet, Manuel Martí Recober
Dept. Investigació Operativa i Estadística
Facultat d'Informatica de Barcelona (UPC)

Abstract

In data analysis one must often analyze data matrices formed by categorical variables, correlated with some others variables, called "instrumental" ones, and which effect we wish to keep fixed. In such cases we perform a multiple correspondence analysis upon a similarity graph defined by the instrumental variables. The obtained factorials plans take in to account the oppositions between individuals apart from the instrumental variables. We call such analysis, partial, because it is based on the same idea of partial correlation analysis or on the instrumental variables of Rao, without having the strong hypothesis implied by these analysis.

Palabras clave: Correspondencias múltiples, análisis local, análisis global.

Introducción

El Análisis de Datos consiste en la descripción de grandes matrices numéricas mediantes técnicas de representación visual y de clasificación sin efectuar ninguna suposición sobre la naturaleza probabilista de los datos. Los resultados obtenidos son generalmente una buena aproximación a la realidad "global" (=información) contenida en los datos. El usuario tiene libertad para escoger la codificación de los datos, la métrica del espacio y la ponderación de los individuos.

Nuestro propósito es introducir un grado de libertad más en el análisis, permitiendo la descripción de aspectos parciales de los datos, siempre y cuando estos aspectos parciales admitan una representación en forma de grafo no orientado entre los individuos; hablamos entonces de análisis factoriales sobre un grafo. El origen de estas técnicas remontan a Lebart (1969), el cual lo aplicó sobre grafos de contiguidad geográfica, llamandolo por este motivo análisis factorial local, recientemente ha sido generalizado para grafos de similitud, (Aluja (1985)) y grafos de tipo evolutivo

temporal (Cazier (1985)).

Siendo la codificación nominal muy corriente en numerosas disciplinas, (en particular en la realización de encuestas), y la metrica de Chi-cuadrado la habitual en estos casos para representar las proximidades entre puntos, es interesante generalizar el análisis factorial sobre un grafo y que denominaremos genericamente local, para el caso de una matriz de datos nominales, con métrica de Chi-cuadrado y ponderación uniforme de los individuos, dando lugar al análisis de correspondencias múltiples sobre un grafo.

Notaciones

Sea E un conjunto de n individuos a los cuales se ha interrogado sobre q cuestiones, las cuales contienen en total J modalidades de respuesta. Sea Z la matriz (n,J) de datos formada en codificación disyuntiva completa.

Sea D la matriz diagonal (J,J) de los pesos de las modalidades.

$$d_{jj} = n_{.j} = \sum_i z_{ij}$$

Sea Gj el vector centro de gravedad de los individuos.

$$g_j = n_{.j} / (n * q)$$

Supongamos relacionados los n individuos mediante un grafo simple G(E,T). Sea M la matriz (n,n) asociada al grafo.

$m_{ij} = 1$ si i y j estan unidos por una arista.
$m_{ij} = 0$ en cualquier otro caso.

Sea N la matriz diagonal (n,n) de los grados de los vertices.

$$n_{ii} = \sum_j m_{ij}$$

Sea m dos veces el numero total de aristas del grafo, ($m = \sum_i n_{ii}$).

Sea T la matriz (n(n-1)/2,n) cruzando las aristas con los

vértices. Una arista uniendo los vertices i y j, es codificada mediante una sucesion de 000001000-100, el 1 y el -1 en las posiciones correspondientes a los vertices i y j, siendo la asignacion del signo completamente arbitraria.

Sea I la matriz (n,n) unidad.

Sea U la matriz (n,n) con todos sus terminos igual a la unidad.

Sea B la matriz (n(n-1)/2,n) cruzando todas las aristas del grafo completo con los vertices.

Existe la siguiente relación entre estas matrices:

$$N - M = T'T = B'T = T'B \qquad (1)$$

Notamos por $\sum_{i,i'}^{L}$ el doble sumatorio efectuado solamente sobre las aristas del grafo.

Matriz de Inercia Local

Definimos la matriz de inercia local como:

$$V_L = (n/mq)\ D^{-1/2} Z' (N - M)\ Z\ D^{-1/2} \qquad (2)$$

de termino general:

$$v_{jj'}^{L} = (n/2mq\sqrt{n_{.j}\, n_{.j'}})\ \sum_{i,i'}^{L} (z_{ij} - z_{i'j})\ (z_{ij'} - z_{i'j'})$$

Dicha matriz coincide con la clásica matriz de inercia "global" para el caso de un grafo completo.

$$V_g = (1/q)\ D^{-1/2} Z' Z\ D^{-1/2} - G_j^{1/2} G_j^{1/2\prime} = (1/nq)\ D^{-1/2} Z' (nI - U)\ Z\ D^{-1/2}$$

cuyo termino general se escribe:

$$v_{jj'}^{g} = \sum_{i} z_{ij}\, z_{ij'} / (q\sqrt{n_{.j}\, n_{.j'}}) - \sqrt{n_{.j}\, n_{.j'}} / (nq) =$$

$$= 1/2nq\sqrt{n_{.j}\, n_{.j'}}\ \sum_{i,i'} (z_{ij} - z_{i'j})(z_{ij'} - z_{i'j'})$$

La matriz de inercia local expresa pues la variabilidad "local" entre individuos contiguos segun el grafo.

Análisis de Correspondencias Multiples sobre un grafo. Análisis en R^p.

Tomamos como coordenadas de los n puntos (=vértices) los perfiles-fila de la matriz Z, esto es las filas de la matriz $(1/q)Z$, con pesos $p_i = 1/\sqrt{m}$ y definidos en un espacio de metrica nqD^{-1}. El criterio a maximizar se escribe ahora:

$$\max_{H} \sum_{i,i'} p_i p_{i'} d^2_H(i,i') \qquad (3)$$

Siendo $d^2_H(i,i')$ la distancia proyectada sobre un subespacio H entre los puntos i e i'. Tomando en primer lugar $H = w$ un vector unitario, la proyección de los perfiles-fila sobre w se escribe $\Psi = n Z D w$. Luego el criterio (3) se escribe:

$$\max_{w} 1/m\ \Psi' T' T \Psi = (n^2/m)\ w' D^{-1} Z' T' T Z D^{-1} w = \lambda, \quad (4)$$

con la restriccion $\qquad (nq)\ w' D^{-1} w = 1$

El máximo buscado se encuentra diagonalizando la matriz $(n/mq)\ Z'T'T Z D^{-1}$ la cual haciendo el cambio $u = \sqrt{nq}\, D^{-1/2} w$, se convierte en simétrica, quedando las relaciones (4) en:

$$A u = (n/mq)\ D^{-1/2} Z' T' T Z D^{-1/2} u = \lambda u \qquad (5)$$

con $\qquad u'u = 1$

Luego el máximo buscado corresponde al mayor valor propio de A, siendo u la dirección de proyección de máxima inercia local. Analogamente, el subespacio de dimensión r que maximiza la inercia local proyectada, viene definido por los r vectores propios de A correspondientes a los r mayores valores propios. Notese que la matriz A no es sino la matriz de inercia local definida en (2). El análisis local equivale pues a tomar como individuos, no los vértices (filas de Z), sino las aristas (filas de TZ). Llamamos modalidades locales a las columnas de TZ (de término general $(z_{ij} - z_{i'j})$; mientras que llamamos modalidades globales a las columnas de Z, o bien a las columnas de BZ (modalidades sobre el grafo completo).

La inercia local viene definida por:

$$In_\ell = tr(V_\ell) = \sum_\alpha \lambda_\alpha = \sum_j v_\ell(j)$$

donde $$v_\ell(j) = (n/mqn_{.j}) \sum_{i,\ell}^{L} (z_{ij} - z_{\ell j}) = d^2(j,0)$$

es la contribución de la modalidad j a la inercia local, lo cual significa que una modalidad es tanto mas influyente en el analisis como más dispares sean los valores que toma sobre vertices vecinos en el grafo.

La proyección de los vértices sobre los ejes factoriales locales, se obtiene por la relación:

$$\Psi_\alpha = (\sqrt{n/q})\ Z\, D^{-1/2} u_\alpha \qquad (6)$$

y la proyección de las aristas : $\xi_\alpha = T\Psi_\alpha$. Siendo estas proyecciones difíciles de interpretar, no han sido incluidas en el programa.

Analisis en R^T.

El análisis inducido respecto a las columnas se escribe:

$$(n/mq)\ T\, Z\, D^{-1} Z' T' v = \lambda v \qquad (7)$$

$$\text{con} \qquad v'v = 1$$

Lo cual equivale a tomar por coordenadas de las modalidades las columnas de la matriz $(\sqrt{n/q})\ ZD^{-1}$, ponderadas por D y definidas en un espacio de métrica $(1/m)\ I$.

Las relaciones entre los dos análisis se escriben:

$$v_\alpha = (1/\sqrt{\lambda_\alpha})(\sqrt{n/mq})\ T\, Z\, D^{-1/2} u_\alpha \qquad (8)$$

$$u_\alpha = (1/\sqrt{\lambda_\alpha})(\sqrt{n/mq})\ D^{-1/2} Z' T' v_\alpha \qquad (9)$$

Las proyecciones de las modalidades locales activas sobre los ejes factoriales v, serán:

$$\varphi_\alpha = (\sqrt{n/mq})\ D^{-1} Z' T' v_\alpha = \sqrt{\lambda_\alpha}\, D^{-1/2} u_\alpha . \qquad (10)$$

mientras las proyecciones de las modalidades locales suplementarias:

$$\varphi_\alpha^+ = (1/m)\left(\sqrt{n/q\,\lambda_\alpha}\right) D^{-1} Z_1 T' T\, \Psi_\alpha . \qquad (11)$$

La contribución de una arista se escribe:

$$cont(i,i') = (n/mq) \sum_j 1/n_j (z_{ij} - z_{i'j}) = d^2(i,i';0)$$

luego una arista tiene tanta más importancia en el análisis, cuanto menos coincidan los valores de las modalidades en los vértices que la forman.

Relacion con el analisis global.

El comportamiento de las modalidades puede cambiar de forma ostensible cuando pasamos de la escala global a la escala local, dependiendo de la aleatoriedad de los valores de las modalidades en los vértices del grafo. En este sentido es interesante comparar los dos niveles de las modalidades, global y local.

Desde Lebart (1973,84) sabemos que en el caso de grafo planario y regular, el análisis local es equivalente a un análisis de correlaciones parciales habiendo fijado las coordenadas geográficas de los vértices. Esto es, el análisis local equivale a proyectar las modalidades globales sobre un subespacio ortogonal definido por M_1 y analizar estas modalidades proyectadas.

Por otro lado, con el supuesto de tener más aristas que variables ($m/2 > p$), ambos tipos de variables estan contenidas en el mismo espacio de dimensión p. Luego podremos establecer la relación entre ambas estructuras por la matriz de correlaciones entre los dos conjuntos de variables. La matriz de covariancias se escribe:

$$V_{gl} = (1/q\sqrt{m})\, D^{-1/2} Z' B' T\, \Sigma\, D^{-1/2} = (\sqrt{m}/n)\, V_l$$

y la matriz de correlacion : $C_{gl} = S_g^{-1} V_{gl} S_l^{-1}$.

Particularmente interesante es la correlacion entre una variable global y su homónima local:

$$\text{cor}(j_g, j_\ell) = (\sqrt{m}/n)\ \text{cov}(j_g, j_\ell) / \sqrt{v_g(j)\ v_\ell(j)} =$$

$$= \sqrt{\sum_{i,i'}^{\ell} (z_{ij} - z_{i'j})^2 \ /\ \sum_{i,i'} (z_{ij} - z_{i'j})^2}$$

luego, aunque la correlación entre j y j , no depende del número de aristas del grafo, tiende hacia la unidad conforme operemos con grafos más completos. Generalmente el numero de aristas de un grafo es mucho menor que en el grafo completo (oscila alrededor del 3%), lo cual implica que la variable local es mucho mas pequeña que la variable global, es por este motivo por lo que tomamos como ponderación de los individuos $1/\sqrt{m}$, y no $1/n$ como seria normal, por lo que la variable local queda multiplicada por un factor n^2/m.

Podemos tambien visualizar los cambios operados en las variables al pasar del nivel global al local, proyectando ambos tipos de variables sobre la misma base, mediante la relación:

$$\varphi_{g\ell} = (\sqrt{1/mq\lambda_\alpha})\ D^{-1} Z' T' B\, \psi_{g\alpha}.$$

donde $\psi_{g\alpha}$ son las proyecciones de los individuos sobre la base global.

Este analisis de correspondencias múltiples sobre un grafo ha sido programado en una etapa "VARNO", compatible con el sistema SPAD de analisis de datos.

Bibliografia

T. ALUJA, L. LEBART (1985) "Factorial analysis upon a graph". Bulletin Technique du CESIA. Vol. 3, pp. 4-34.

A. CARLIER (1985) "Analyse des evolutions sur tables de con - tingence: Quelques aspects operationels", INRIA, Paris.

R.C. GEARY (1954) "The contiguity ratio and statistical mapping". The Inc. Statistician. pp. 115-145.

L. LEBART (1969) "Analyse statistique de la contiguite". Publications de l'ISUP, XVIII, pp. 81-112.

L. LEBART (1984) "Correspondece analysis of graph structures". Bulletin Technique du CESIA. Vol 2, n.1-2, pp. 5-19.

C. R. RAO (1964) "The use and interpretation of principal compo- nents analysis in applied research". Sanchya, vol 26, pp. 329-357.