Revista de Matemática: Teoría y Aplicaciones 1994 **1**(1) : 47–55
cimpa – ucr – ccss issn: 1409-2433

ANÁLISIS DE TABLA MÚLTIPLES DE DATOS

William Castillo Elizondo* Jorge González Varela**

Resumen

Se presentan los dos métodos más utilizados por la escuela francesa de Análisis de Datos para estudiar las tablas múltiples: el método Statis y el Análisis Factorial Múltiple. Se analizan los tipos de datos que puede procesar cada metodología y la forma en que cada una define el "compromiso" para obtener la representación simultánea de los individuos/variables en un espacio principal. Además, se estudian los aspectos computacionales en vista de una implementación.

Abstract

Two methods for analyzing multiple data tables of the French Data Analysis school are presented: the Statis method and Multiple Factor Analysis. The types of data that can be trated by each methodology are analyzed and how they construct an average simultaneous representation of the individuals/variables in the same principal space. Moreover, computational aspects of an implementation are studied.

1. Introducción

Frecuentemente nos encontramos con tablas de datos de tres índices, un índice para identificar los individuos que son objeto de estudio, un segundo índice para las variables que se han medido sobre esos individuos, y un tercer índice para las diversas situaciones (instantes) en que las mediciones se realizaron.

El objetivo es analizar las semejanzas y diferencias entre las diferentes situaciones a través de las configuraciones de los individuos y de las relaciones entre los diferentes grupos de variables.

Hay, al menos, dos enfoques para este estudio: el de la Escuela Francesa con los métodos Statis (Estructura estadística de tablas de tres índices) [10] y el Análisis Factorial Múltiple) [5], y los denominados Indscal e Idioscal utilizados fundamentalmente en Estados Unidos

*Escuela de Matemática, Universidad de Costa Rica, 2060 San José, Costa Rica. E-Mail: wcastill@cariari.ucr.ac.cr
**Misma dirección. E-Mail: jgonzale@cariari.ucr.ac.cr

y Gran Bretaña [3]. Existen además otros métodos de la escuela francesa, como el método Longi [14] basado en el análisis canónico y los análisis de tablas múltiples de contingencia propuestos por Carlier [2].

Los diferentes métodos difieren en la forma en que se consigue un referencial llamado *compromiso* que permite ubicar en un mismo subespacio los individuos de las diferentes situaciones así como las variables.

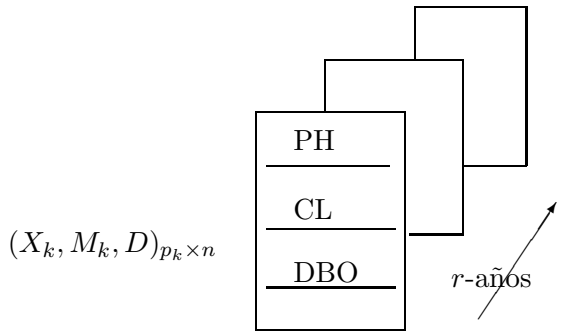
En lo que sigue resumimos los aspectos más notorios del método Statis propuesto por Y. Escoufier y estudiado por L'Hermier des Plantes en [10], y el Análisis Factorial Múltiple propuesto por B. Escofier.

2. Análisis Conjunto de Matrices de Datos Cuantitativas. Método Statis

2.1. Los Datos

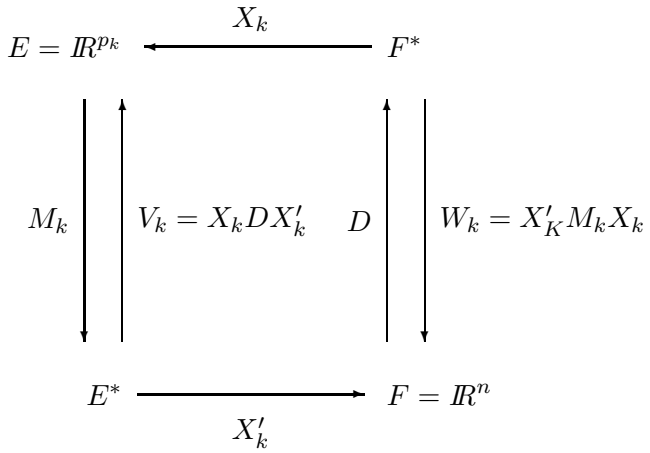
Nos referiremos, como es usual en análisis de datos, a los puntos de muestreo como *individuos* y a los diferentes parámetros medidos como *variables*.

Suponemos que tenemos r tablas de datos $X_1, \ldots, X_r$, donde X_k está provista de las métricas M_k y D, obtenidas de las mediciones durante r años.



Notaciones: X_k : matriz variables $\times$ individuos de tamaño $p_k \times n$, **centrada** respecto de D.
n : número de individuos (puntos de muestreo), el mismo durante los r años.
p_k : número de variables durante el k-ésimo año.
$D = diag(p_1, \ldots, p_n)$: matriz de pesos, invariante durante los r años, métrica en el espacio de las variables $\mathbb{R}^n$.
M_k : métrica utilizada en el k-ésimo año, en el espacio de los individuos $\mathbb{R}^{p_k}$.

El triplete (X_k, M_k, D) resume el siguiente diagrama de dualidad:



2.2. Estructura espacial de los individuos

La estructura espacial de los individuos en el año k-ésimo está determinada por la matriz de productos internos entre individuos $W_k = X_k' M_k X_k$, que son los operadores introducidos por Y. Escoufier (1970, 1975 y 1976) [15].

Comparar las diferentes situaciones expresadas en cada tabla se reduce a comparar las matrices W_k correspondientes. Para tal efecto utilizamos la métrica de Hilberth-Smith y construimos una matriz S de tamaño $m \times n$, de distancia entre tablas definida por:

$$S_{ij} = \langle W_i, W_j \rangle_\Phi = trz(W_i D W_j D)$$

2.3. Imagen euclídea

Obtengamos una imagen euclídea de la nube $\{W_1, \ldots, W_r\}$ a través de la diagonalización de S.

Sea P la matriz cuyas columnas son vectores propios de S, asociados a los valores propios no nulos, y sea $Q = P D_{\sqrt{\lambda}}$ con $D_{\sqrt{\lambda}} = Diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$.

Luego se tiene $S = QQ'$, de donde las filas de Q, forman una imagen euclídea para $(\{W_1, \ldots, W_r\}, \phi)$, la cual satisface las siguientes propiedades:

1. $\langle W_i, W_j \rangle_\phi = \langle Q_i, Q_j \rangle_I = Q_i Q_j'$ Una representación plana óptima, en el sentido de inercia mínima , de los W_i, la obtenemos con las dos primeras filas de Q.

2. Si $W_1 D = W_2 D$ ó $W_1 D = \alpha W_2 D \Longrightarrow Y_1 = Y_2$ ó $Y_1 = \sqrt{\alpha} Y_2$, donde Y_1, Y_2 son las matrices de componentes principales (vectores propios de $W_1 D, W_2 D$), por lo que tendrán la misma configuración de individuos o configuraciones homotéticas.

3. $\langle W_i, W_j \rangle_\Phi = 0 \Longleftrightarrow X_i D X_j' = 0$
   Configuraciones Φ-ortogonales de los años i, j se corresponden con grupos de variables de los mismos años D-ortogonales.

2.4. Evolución de individuos y variables

2.4.1. Obtención del compromiso

Podríamos analizar las posiciones relativas de un individuo (resp. variable) respecto de los otros individuos (resp. variables) al interior de cada una de las tablas, pero esto no permitiría un buen análisis evolutivo, pues es necesario un sistema de referencia común en el cual se representen adecuadamente las diferentes estructuras aportadas por cada tabla. El compromiso propuesto por H. L'Hermier des Plantes [10], en el desarrollo del método Statis, es considerar una expresión de la forma:

$$\mathcal{O} = \sum_{i}^{r} \alpha_i W_i$$

Las siguientes son dos formas de elegir los α_i.

Compromiso 1 Buscamos el punto $\mathcal{O} \in I\!R^{n^2}$ respecto del cual la nube $\{W_1, \ldots, W_r\}$, provista de los pesos $p_1, \ldots, p_r$ tenga inercia mínima:

$$I_O = \sum_{1}^{r} p_i ||W_i - \mathcal{O}||^2_\Phi$$

Como sabemos esto se tiene para $\alpha_i = p_i \;\; i = 1, \ldots, r$, es decir el mínimo se alcanza cuando $\mathcal{O}$ es el centro de gravedad de la nube. Si los pesos son iguales, el compromiso es la media aritmética de la nube.

Compromiso 2 Buscamos un subespacio de dimensión igual a 1 $U = CL\{\mathcal{O}\}$ respecto del cual la inercía I_U de la nube sea mínima o bien su inercía respecto del ortogonal $I_{U^\perp}$ sea máxima. Como es conocido, tal cosa se tiene cuando $\mathcal{O}$ es la primera componente principal del ACP de (X, I, Φ).

$\mathcal{O} = X'v = \sum_{i=1}^{r} \alpha_i W_i$ con $v = (\alpha_1, \ldots, \alpha_r)^t$ vector propio de S , $||v||_I = 1$, asociado al mayor valor propio λ_1

Este compromiso tiene las siguientes propiedades:

1. $||\mathcal{O}||^2_\phi = \lambda_1 \geq ||\sum_{i=1}^{r} \beta_i W_i||^2 \;\; \forall ||(\beta_1, \ldots, \beta_r)^t||_I = 1$
2. $\sum_{i=1}^{r} \langle \sum_{i=1}^{r} \beta_i W_i, W_i \rangle^2_\phi \leq \sum_{i=1}^{r} \langle \mathcal{O}, W_i \rangle_\phi = \lambda_1^2 \;\; \forall ||(\beta_1, \ldots, \beta_r)^t||_I = 1$
3. Si todas las situaciones conducen a configuraciones vecinas de los individuos, entonces todos los α_i serán sensiblemente iguales y el compromiso es próximo a la media aritmética.
4. Si, por el contrario, el grupo es homogéneo pero hay unas pocas situaciones particulares, éstas intervienen poco en la definición del compromiso, es decir el compromiso respeta las mayorías y descarta las eventuales minorías.

2.4.2. Intraestructura

La factorización del compromiso $\sum_{i=1}^{m} \alpha_i W_i$ permite una representación suplementaria de los individuos a través de las filas de las matrices W_i. Si consideramos los primeros ejes obtenemos representaciones planas en las cuales podemos estudiar su trayectoria.

Para las variables, es posible proyectarlas sobre las componentes definidas por el compromiso, obteniendo las representaciones planas correspondientes.

3. Análisis Factorial Múltiple

Con el fin de facilitar la implementación computacional del Análisis Factorial Múltiple (AFM), se presenta una versión algorítmica.

3.1. Los datos

El sistema es alimentado con datos de un archivo que contiene a registros (individuos) y b columnas (variables). Se admiten variables cuantitativas y cualitativas. El usuario selecciona n registros y p columnas para formar un nuevo archivo, notado A, sobre el cual se aplican las operaciones que siguen.

1. Si v es una variable cualitativa con $m_1, \ldots, m_k$ modalidades, se definen sus k indicatrices como sigue: $v_j(i) = 1$ si $v(i) = m_j$ y $v_j(i) = 0$ en otro caso.

2. Una variable v numérica puede ser transformada en cualitativa, así: sean $[a, b]$ el rango de v y $a < c_1 < c_2 < \ldots < c_r = b$ (información suministrada por el usuario). Se guardan las r indicatrices $v_1, \ldots, v_r$ definidas así: $v_j(i) = 1$ si $c_{j-1} \leq v(i) < c_j$ y $v_j(i) = 0$ en otro caso.

3. Para cada una de las variables $0-1$ definidas anteriormente se aplica la transformación $\tilde{v}_s(i) = \sqrt{\frac{n}{c_s}}$ donde $c_s = \sum_{i=1}^{n} v_s(i)$.

4. El usuario elige unas variables cuantitativas del archivo A, excepto las que fueron transformadas, para estructurarlas en q grupos y formar las matrices de datos $X_1, \ldots, X_q$. Las variables cualitativas transformadas como fue indicado, son estructuradas en t grupos para formar las t matrices de datos $Y_1, \ldots, Y_t$.

3.2. Análisis en componentes principales

1. Se realiza un ACP normado (es decir, centrado y reducido) de algunas tablas X_i escogidas por el usuario. Similarmente, se hace un ACP centrado de algunas tablas Y_j escogidas por el usuario. Sean $Z_1, \ldots, Z_r$ dichas tablas centradas y reducidas o solamente centradas, según sean respectivamente de variables cuantitativas o cualitativas. Se graban las primeras d_i componentes principales normadas $c_1^i, \ldots, c_{d_i}^i$ del ACP de cada Z_i, los valores propios $\lambda_{1i} \geq \lambda_{2i} \geq \ldots \geq \lambda_{d_{ii}}$ de cada ACP y se hacen los histogramas de los valores propios.

2. Se forma la tabla
$$Z_p = \left(\lambda_{11}^{-1/2} Z_1 \dots \lambda_{1r}^{-1/2} Z_r \right)$$
y se realiza el ACP de Z_p sin reducción –esta matriz ya es centrada–.

Se graban las primeras componentes principales normadas $C_1, \dots, C_k$, los primeros vectores principales $u_1, \dots, u_k$ y los primeros k valores propios $\lambda_1, \dots, \lambda_k$ de este ACP; k es suministrado por el usuario.

3.3. Representaciones usuales

1. En los planos determinados por los vectores principales del ACP de la tabla Z_p se representan la filas de esta tabla en la forma usual. Las filas a representar deben ser indicadas por el usuario, así como los ejes. La coordenada de la fila x_i de la tabla Z_p, sobre el vector principal u, se expresa por la fórmula
$$\text{coord}_u(x_i) = \sum_{h=1}^{m} x_{ih} u_h$$

2. En el círculo de correlaciones del ACP de Z_p se representan las columnas de Z_p, previa reducción de las variables cuantitativas. En el caso de las indicatrices, el efecto de las ponderaciones se elimina multiplicando la columna j de la matriz $\lambda_{1k}^{-1/2} Z_k$ por el factor $\frac{\lambda_{1k}}{1-c_s/n}^{1}/2$.
Las columnas en Z_p así como las componentes principales, son escogidas por el usuario. La coordenada de la variable x^j en la componente principal C_t se calcula mediante la fórmula
$$\text{coord}_{C_t}(x^j) = \frac{1}{n} \sum_{h=1}^{n} x_h^j C_{th}$$

3.4. Representaciones suplementarias

1. Se pueden representar en suplementario los individuos caracterizados por los diferentes grupos de variables. En este caso un mismo individuo aparece tantas veces como grupos de variables hayan. Los elementos que van a ser representados, así como el número de planos, son especificados por el usuario. La coordenada del individuo i caracterizado por el grupo de variables j, sobre el vector principal u se calcula mediante la fórmula
$$\text{coord}_u(x_i^j) = \sum_{l=1}^{p_j} x_{il}^j u_l^j$$
donde
   - $x_i = (x_i^1 \dots x_i^r)$ es la i-ésima fila de la matriz Z_p y, $x_i^j = (x_{i1}^j, \dots, x_{ip_j}^j)$ el vector de coordenadas del individuo i respecto del grupo de variables j, dadas por las columnas de la tabla $\lambda_{1j}^{-1/2} Z_j$.

- $u = (u^1, \ldots, u^r)$ un vector principal del ACP de Z_p y $u^j = (u^j_1, \ldots, u^j_{p_j})$.
  Junto con esta representación se hace la de los individuos promedio $\frac{1}{r}x$ donde x es una fila de Z_p.

2. Para cada modalidad j –variable cualitativa 0-1–, se representa el centro de gravedad g_j, de los individuos que la poseen, caracterizado por el grupo de variables G_s o por la totalidad de las variables. La coordenada de g_j en el eje u es, para la totalidad del grupo de variables:

$$\text{coord}_u(g_j) = \sum_{h=1}^{m} g_{jh} u_h$$

   y para el grupo de variables G_s es:

$$\text{coord}_u(g_j) = \sum_{h=1}^{p_s} g^s_{jh} u^s_h$$

   donde u_h, u^s_h fueron definidos antes y $g_{jh} = \frac{1}{q_j}\sum_{s\in I_j} x_{sh}$, I_j es el conjunto de filas de Z_p que poseen la modalidad j y q_j su cardinalidad. Por otra parte, g^s_{jh} es g_{jh} referido sólo al grupo de variables G_s dadas por las columnas de la tabla $\lambda_{1s}^{-1/2} Z_s$.

   Se permite hacer una representación simultánea de las filas de Z_p, de los individuos promedio, de los individuos caracterizados parcialmente por los grupos de variables y de los centros de gravedad.

3. El usuario escoge algunas de las componentes principales $c^j_1, \ldots, c^j_{d_j}$ del ACP de Z_i, para ser proyectadas en el círculo de correlaciones del AFM. Sus coordenadas en el eje C_t se obtienen por medio de la fórmula

$$\text{coord}_{C_t}(c^j_f) = \frac{1}{n}\sum_{h=1}^{n} c^j_{fh} C_{th}$$

3.5. La inter-estructura

Siguiendo la misma idea que en el método Statis buscamos un procedimiento para representar la *evolución* de los grupos de variables por medio de configuraciones planas. Se prueba que cada operador $W_j D$ tiene como coordenadas sobre el eje de rango t en el sistema referencial determinado, las cantidades $L(C_t, G_j) = \sum_{v\in G_j}[\sum_h^n C_{th} v_h]^2$ donde v_h y C_{th} son las coordenadas de v y C_t respectivamente. Aquí G_j es el grupo de variables de la matriz Z_j. Se debe advertir que las variables $v \in G_j$ corresponden a las columnas de la matriz $\lambda_{1j}^{-1/2} Z_j$

3.6. Indices de calidad

La calidad de la representación de un elemento –una fila o una columna–, depende de su inercia proyectada sobre un eje o un plano. Si se trata de la calidad de la representación de un conjunto de elementos se hace la suma de las inercias proyectadas por esos elementos. En el caso concreto del AFM tenemos:

1. Para cada fila $x = (x_1, \ldots, x_m)$ de Z_p se calcula la calidad de su representación sobre un vector principal $u = (u_1, \ldots, u_m)$ del ACP de Z_p, por la fórmula

$$\cos^2\theta_{xu} = \frac{(\text{coord}_u(x))^2}{\sum_{t=1}^{m} x_t^2}$$

   Igualmente se calcula la calidad de la representación de los centros de gravedad g_j de una modalidad j, sustituyendo en la fórmula anterior las coordenadas de x por las de g_j.

2. La calidad de la representación de las columnas –las variables–, en la dirección de C_t se calcula a partir de las columnas de Z_p previamente reducidas, lo cual es el cuadrado de la correlación entre la columna y el eje C_t. Esto es, $\text{cor}^2(x^j, C_t) = [\text{coord}_{C_t}(x^j)]^2$.

3. Sea x_i^j la fila i en la matriz Z_p pero caracterizada sólo por el grupo de variables j. La calidad de la representación de la nube $\mathcal{N}_j = \{x_i^j | i = 1, \ldots, n\}$ en el eje u_s es:

$$I_{u_s}(j) = \frac{1}{nI_j} \sum_i (\text{coord}_{u_s}(x_i^j)^2$$

   donde I_j es la inercia total de la nube $\mathcal{N}_j$. Esto es; $I_j = \sum_{s=1}^{d_j} \lambda_{sj}$.

4. La *importancia* del factor común de rango t en el grupo de variables G_j se mide por medio de los siguientes índices:

   - Se compara la importancia de este factor en el grupo G_j en relación con la importancia de los otros factores:

$$A_{jt} = \frac{L(G_j, C_t)}{\sum_{h=1}^{k} L((G_j, C_h)}$$

   - Se compara la importancia de este factor en el grupo G_j en relación con su importancia en los otros grupos:

$$B_{jt} = \frac{L(G_j, C_t)}{\sum_{h=1}^{r} L((G_h, C_t)}$$

5. La presencia del factor de rango t en el grupo de variables G_j también se evalúa por medio de la correlación entre las componentes principales $C_1, \ldots, C_k$ y los vectores $F_u^j = (F_{u1}^j, \ldots, F_{un}^j)$ donde $F_{ui}^j = \text{coord}_u(x_i^j)$.

   Es decir, para cada vector principal u, se calcula

$$\text{corr}(F_u^j, C_t) = \frac{\sum_{i=1}^{n} (F_{ui}^j - \overline{F}_u^j) C_{ti}}{\sqrt{\sum_{i=1}^{n} (F_{ui}^j - \overline{F}_u^j)^2}}$$

Referencias

[1] Cailliez, F.; Pagès, J.P. (1976) *Introduction à l'Analyse des Données*. Smash, París.

[2] Carlier, A. (1985) *Analyse des évolutions sur tables de contingence: quelques aspects opérationnels*, 4èmes Journées Analyse des Données et Informatique, INRIA, tomo 2, E. Diday *et al.* (eds.) pp. 421-428.

[3] Carroll, J.D.; Chang, J.J. (1970) *Analysis of individual differences in multidiemnsional scaling via an N-way generalization of "Eckart-Young"decomposition*, Psychometrika, N^o 35, pp. 283-319.

[4] Chevalier, F. *Analyse en composantes conjointes d'une famille de triplets indexés*, Statistique et Analyse des Données, Vol 2, pp. 35-75.

[5] Escofier, B.; Pagès, J. (1988) *Analyses Factorielles Simples et Multiples; Objectifs, Méthodes et Interprétation*. Dunod, París.

[6] Escofier, B.; Pagès, J. (1986) *Le traitement des variables qualitatives et des tableaux mixtes par analyse factorielle multiple*, 4èmes Journées Analyse des Données et Informatique, INRIA, E.Diday *et al* (eds.).

[7] Glaçon, F. (1981) *Analyse conjointe de plusieurs matrices de données*. Tesis de doctorado, Université Scientifique et Médicale, Grenoble.

[8] Guilbot, A.; Picot-Reboul, B. (1982) *Etude de l'évolution de la qualité des eaux de rivière*, Water Res., vol 16, pp. 1173-1187.

[9] Jambu, M. (1989) *Exploration informatique et statistique des données*. Dunod, París.

[10] L'Hermier des Plantes, H. (1976) *Structuration des tableaux à trois indices de la statistique: théorie et application d'une méthode d'analyse conjointe*. Tesis de doctorado, Université des Sciences et Techniques du Languedoc, Montpellier.

[11] L'Hermier des Plantes, H.; Thiebaut (1987) *Etude de la pluviosité au moyen de la méthode Statis*, Revue de Statistique Appliquee, vol. XXV, No. 2, pp. 57-81.

[12] Lavit, Ch.; Pérez-Hugalde, C. (1985) *The Statis method applied to economic data: multivariate evolution of the spanish provinces*, 4èmes Journées Analyse des Données et Informatique, INRIA, E.Diday *et al* (eds.).

[13] Lavit, Ch. (1988) *Analyse Conjointe de Tableaux Quantitatifs*. Ed. Masson, París.

[14] Pontier, J.; Dufour, A.B.; Normand, M. (1990) *Le Modèle Euclidien en Analyse des Données*. Ed. de l'Université de Bruxelles - Ellipses, Bruselas.

[15] Robert, P.; Escoufier, Y. (1976) *A unifying tool for linear multivariate statistical methods: the RV-coefficient*, Applied Statistics, 25, No. 3, pp. 257-265.