Revista de Matemática: Teoría y Aplicaciones 4(1): 51–60 (1997)

ANÁLISIS DE TABLAS MÚLTIPLES CON INDIVIDUOS CAMBIANTES Y VARIABLES FIJAS

William Castillo Elizondo[1] – Jorge González Varela[1]

Resumen

En este artículo se presenta el método Statis cuando se observan las mismas variables en todos los instantes considerados, con individuos posiblemente cambiantes. Se demuestran varias propiedades de la interestructura y del compromiso que justifican eventuales interpretaciones bidimensionales de los datos. Dichas propiedades son también útiles para hacer una implementacin computacional.

Palabras clave: Statis dual, interestructura, compromiso, intraestructura, imagen euclídea.

1. Introducción.

El método Statis (Análisis Estadístico de Tres Indices) se ubica dentro de las técnicas exploratorias de análisis multivariado de datos, comúnmente usados en el estudio simultáneo de varias matrices de datos obtenidas de diferentes observaciones de un mismo fenómeno.

Suponemos que se tienen $X_1, \ldots, X_K$ tablas centradas de individuos por variables que corresponden a K mediciones de un mismo fenómeno. Si además disponemos de las matrices M_k y D_k necesarias para medir distancia entre individuos y las covarianzas entre las variables respectivamente; entonces las K-tablas definen K-configuraciones del tipo $W_k = X_k M_k X_k^t$ o $V_k = X_k^t D_k X_k$.

El método Statis analiza la evolución del fenómeno en estudio comparando estas configuraciones, los W_k, en caso de que los individuos hayan permanecido invariantes en las K-mediciones, o bien los V_k si se han medido las mismas variables en los K-instantes.

La primera de estas situaciones (individuos fijos) se encuentra desarrollada tanto en los aspectos teóricos como computacionales en, por ejemplo, [2], [3] y [4]. Nuestro propósito es abordar ahora el segundo caso, esto es las mismas variables en las K mediciones con individuos posiblemente diferentes al pasar de una tabla a otra. Este método es conocido

[1]Programa de Investigaciones en Modelos y Análisis de Datos (PIMAD), Centro de Investigaciones en Matemáticas Puras y Aplicadas (CIMPA), Escuela de Matemática, Universidad de Costa Rica, 2060 San José, Costa Rica. E-mail: wcastill@cariari.ucr.ac.cr, jgonzale@cariari.ucr.ac.cr

con el nombre de *Statis Dual*. En la literatura especializada ([2], [4] y [8]) no se desarrolla el Statis Dual sino que se menciona únicamente su posible utilización, pese a que en la realidad nos encontramos con muchos casos donde la naturaleza de los problemas sugiere el uso del Statis Dual. Podría citarse a modo de ejemplo el estudio evolutivo de empresas en circunstancias en que unas desaparecen y otras nacen.

Independientemente de cuáles sean las configuraciones a comparar, el método Statis consta de tres etapas:

1. **Interestructura:** Calculando, con la métrica de Hilbert-Schmidt, las distancias entre las K configuraciones, obtenemos información sobre las diferencias y semejanzas entre ellas. Una representación plana óptima de las mismas sirve de ayuda a este propósito.

2. **El Compromiso:** Consiste en calcular una configuración llamada compromiso que es representativa de las K configuraciones y cuyo papel es definir un escenario para la representación de las trayectorias de los individuos y las variables.

3. **Intraestructura:** Es la etapa en la cual analizamos la trayectoria de un individuo particular o una variable, a través de las K observaciones.

Para hacer posible el uso de este método es necesario desarrollar sus propiedades matemáticas útiles tanto para su implementación computacional como para una adecuada interpretación del análisis de los datos. El contenido de este artículo está orientado a la satisfacción de esta necesidad.

2. Statis: las mismas variables en los K instantes.

Se supone que están dados K tripletes (X_k, M, D_k); $k = 1, \ldots, K$ donde X_k es la matriz de datos $n_k \times p$ generada a partir de la medición de p variables cuantitativas sobre n_k individuos, en la ocasión k-ésima. En adelante se supondrá que cada X_k es centrada usando los pesos de la correspondiente matriz diagonal D_k,y que M es la métrica euclídea en el espacio de los individuos (R^p). Considerando la matriz de varianza-covarianza

$V_k = X_k^t D_k X_k$ representativa del triplete (X_k, M, D_k) e introduciendo la métrica de Hilbert-Schmidt, se logra construir representaciones bidimensionales óptimas de los operadores V_k.

2.1. Aproximación óptima de matrices

En esta subsección se presentan los resultados matemáticos a partir de los cuales se deduce la optimalidad de las imágenes euclídeas aproximadas que se construyen en Statis. Hemos notado que normalmente las demostraciones matemáticas de los resultados sobre optimalidad que se incluyen en esta subsección, dependen de un instrumental matemático difícil de introducir en pocas líneas([4] y [5]). Sin embargo es posible una presentación, incluidas las pruebas matemáticas, breve y relativamente sencilla como la que se hace a continuación.

Definición: Sean $(\mathbb{R}^p, M)$ y $(\mathbb{R}^n, N)$ espacios euclídeos, el producto escalar de Hilbert Schmidt se define como $\langle X, Y\rangle_{M,N} = tr\left(X^t NYM\right)$ para todas las matrices X, Y de tamaño $n \times p$.

Teorema 1 *Sea X de rango mayor o igual que q. Una solución del problema*

$$min\left\{\|X - Y\|^2_{M\ ,\ N} \mid \textit{rango de } Y = \rho(Y) = q\right\}$$

es $XMHH^t$ con $H = [v_1 \dots v_q]$, $v_1 \dots v_q$ vectores propios $M-$ ortonormados de $X^t NXM$.

Prueba:

Es claro que expresando $M = L_1^t L_1$ y $N = L_2^t L_2$ se tiene

$$\langle X, Y\rangle_{M,N} = \left\langle L_2 XL_1^t, L_2 YL_1^t\right\rangle_{I_p, I_n} \text{ y } \|X - Y\|_{M,N} = \left\|L_2 XL_1^t - L_2 YL_1^t\right\|_{I_p, I_n}.$$

Sean $x_1, \dots, x_n$ las filas de la matriz $L_2 XL_1^t$. Se sabe que

$$min\left\{\left\|L_2 XL_1^t - L_2 YL_1^t\right\|^2_{I_p,I_n} \mid \rho(Y) = q\right\} = \left\|L_2 XL_1^t - L_2 XL_1^t UU^t\right\|^2_{I_p,I_n}$$

$$= \sum_{i=1}^{n} \|x_i\|^2 - \sum_{k=1}^{q} u_k^t \left(L_2 XL_1^t\right)^t L_2 XL_1^t u_k$$

con $U = [u_1 \dots u_q]$, $u_1, \dots, u_q$ vectores propios I_p - ortonormados de $L_1 X^t NXL_1^t$ asociados a $\lambda_1 \geq \dots \geq \lambda_q > 0$ [7].

Sea $u_j = L_1 v_j$, entonces se deduce que $v_1, \dots, v_q$ son vectores propios M - ortonormados de $X^t NXM$ asociados a $\lambda_1 \geq \dots \geq \lambda_q > 0$.

Por otra parte, como $L_1 H = U$ entonces $L_2 XL_1^t UU^t = L_2 XMHH^t L_1^t$ de donde

$$\left\|L_2 XL_1^t - L_2 XL_1^t UU^t\right\|_{I_p,I_n} = \left\|X - XMHH^t\right\|_{M,N}.$$

Corolario 1 *1.* $min\left\{\|X - Y\|^2_{M,N} \mid \rho(Y) \leq q\right\} = min\left\{\|X - Y\|^2_{M,N} \mid \rho(Y) = q\right\}$.

2. Sea $M = N$ y X simétrica. Entonces $min\left\{\|X - Y\|^2_{M,M} \mid \rho(Y) = q\right\}$ se alcanza en $XMHH^t = \sum_{j=1}^{q} \lambda_j v_j v_j^t$ donde $v_1 \dots v_q$ son vectores propios M - ortonormados de XM asociados a $\lambda_1 \geq \dots \geq \lambda_q > 0$.

3. $min\left\{\|X - Y\|^2_{M,M} \mid \rho(Y) = q\right\} = \sum_{j=q+1}^{r} \lambda_j^2$ *con* $r = \rho(X)$.

Prueba:

1. Sea Y de rango $q_1 \leq q$ y $x_1, \dots, x_n$ las filas de $L_2 XL_1^t$; entonces

$$\begin{array}{rcl} \|X - Y\|^2_{M,N} & \geq & \min\left\{\|X - Z\|^2_{M,N} \mid \rho(Z) = q_1\right\} \\ & = & \sum_{i=1}^{n} \|x_i\|^2 - \sum_{k=1}^{q_1} u_k^t \left(L_2 XL_1^t\right)^t L_2 XL_1^t u_k \\ & \geq & \sum_{i=1}^{n} \|x_i\|^2 - \sum_{k=1}^{q} u_k^t \left(L_2 XL_1^t\right)^t L_2 XL_1^t u_k \\ & = & \min\left\{\|X - Z\|^2_{M,N} \mid \rho(Z) = q\right\} \end{array}$$

2. $X^tNXM = XMXM$. Sean $v_1, \ldots, v_q$ vectores propios M - ortonormados de XM asociados a $\lambda_1 \geq \cdots \geq \lambda_q > 0$, luego también lo son de $XMXM$ asociados a $\lambda_1^2 \geq \cdots \geq \lambda_q^2 > 0$. Por lo tanto $XMHH^t = HD_\lambda H^t = \sum_{j=1}^{q} \lambda_j v_j v_j^t$.

3. Se sabe que $X = \sum_{j=1}^{r} \lambda_j v_j v_j^t$ entonces $X - XMHH^t = \sum_{j=q+1}^{r} \lambda_j v_j v_j^t$. Ahora, como $\left\langle v_i v_i^t, v_j v_j^t \right\rangle_{M,M} = \delta_{ij}$ se deduce que $\left\| X - XMHH^t \right\|^2 = \sum_{j=q+1}^{r} \lambda_j^2$.

2.2. Interestructura

2.2.1. Construcción de una imagen euclídea

La métrica de Hilbert-Schmidt para el caso de los operadores V_k se define y denota por:

$$\langle V_l, V_k \rangle = \text{traza}\,(V_k M V_l M)$$

Sea $\Pi = diag\,(p_k)_{K\times K}$ la matriz diagonal de los pesos dados a los operadores $\{V_1, ..., V_K\}$. La imagen euclídea de los operadores V_k con pesos Π se obtiene diagonalizando la matriz Π- simétrica $S\Pi$; donde $S_{ij} = \langle V_i, V_j \rangle$. Sean $u_1, \ldots, u_r$ los vectores propios $\Pi-$ ortonormados, de $S\Pi$, asociados a los valores propios $\lambda_1 \geq \cdots \geq \lambda_r > 0$.

La matriz S se expresa como $S = \sum_{l=1}^{r} \lambda_l u_l u_l^t = U_r \Delta_\lambda\,(r)\, U_r^t$ donde $U_r = [u_1, \cdots, u_r]_{K\times r}$ y
$\Delta_\lambda\,(r) = diag(\lambda_l)$. Entonces $S = (U_r \Delta_{\sqrt{\lambda}}\,(r))(U_r \Delta_{\sqrt{\lambda}}\,(r))^t$ y las filas de la matriz $U_r \Delta_{\sqrt{\lambda}}\,(r)$ constituyen la imagen euclídea buscada, puesto que

$$\langle V_i, V_j \rangle = S_{ij} = (u_{i1}\sqrt{\lambda_1} \cdots u_{ir}\sqrt{\lambda_r}) \cdot (u_{j1}\sqrt{\lambda_1} \cdots u_{jr}\sqrt{\lambda_r}).$$

2.2.2. Representación bidimensional de la interestructura

Las representaciones bidimensionales de los V_k en tanto que representantes de los tripletes (X_k, M, D_k), se hacen por medio de las filas de $U_r \Delta_{\sqrt{\lambda}}\,(r)$, tomando solamente los dos primeros vectores propios. Esta, representación es óptima:

$$\|S - S(h)\| = \text{mín}\,\{\|S - B\| \;\; B \text{ es de rango} \leq h\}$$

donde $S(h) = \sum_{l=1}^{h} \lambda_l u_l u_l^t$ con $h \leq r$.

La representación bidimensional se obtiene dibujando en un sistema ortogonal, en el plano, los puntos cuyas coordenadas son las filas de $U_2 \Delta_{\sqrt{\lambda}}\,(2)$. La distancia entre dos puntos A_i y A_j (filas i y j de $U_2 \Delta_{\sqrt{\lambda}}\,(2)$) es la que *mejor aproxima la distancia del producto escalar Hilbert-Schmidt* entre V_i y V_j que en este caso se define como $\langle A, B \rangle = tr\left(A^t \Pi B \Pi\right)$. Se tiene la siguiente aproximación:

$$\|A_i - A_j\|^2 = \|A_i\|^2 + \|A_j\|^2 - 2 \cdot A_i \cdot A_j \cong \|V_i\|^2 + \|V_j\|^2 - 2\,\langle V_i, V_j \rangle = \|V_i - V_j\|^2\,.$$

El error en que se incurre por esta aproximación es cuantificado por $\sum_{l=3}^{r} \lambda_l^2$. En efecto: $\|S - S(2)\|^2 = \left\|\sum_{l=3}^{r} \lambda_l u_l u_l^t\right\|^2 = \sum_{l=3}^{r} \left\|\lambda_l u_l u_l^t\right\|^2 = \sum_{l=3}^{r} \lambda_l^2$ puesto que $\left\langle u_l u_l^t, u_s u_s^t \right\rangle = \delta_{ls}$.

2.2.3. Interpretación de la interestructura

Se desarrollan algunos resultados que ayudan a comprender el significado de las proximidades entre los operadores V_k.

Relación entre distancias y correlaciones: Si las tablas X_k son centradas y reducidas, entonces $V_k = R_k$ es la matriz de correlaciones de las columnas de la tabla X_k. Es claro que[2], si $M = I$, $\|R_k\|^2 = \sum_{j=1}^{p} \|R_k(j)\|^2 = \sum_{i=1}^{p}\sum_{s=1}^{p} \left[cor\left(x_k^i, x_k^s\right)\right]^2$ donde $R_k(j)$ es la fila j de R_k. En consecuencia

$$d^2(R_k, R_l) = \sum_{i=1}^{p}\sum_{s=1}^{p} \left[cor\left(x_k^i, x_k^s\right) - cor\left(x_l^i, x_l^s\right)\right]^2.$$

Observaciones:

1. De lo anterior se concluye que la proximidad entre puntos observada en el plano de la interestructura se interpreta como estabilidad en la estructura de correlaciones para las mediciones efectuadas en las ocasiones k y l.

2. Si en la fórmula de $d^2(R_k, R_l)$ sustituimos R_l por αR_k, tenemos:

$$d^2(R_k, \alpha R_k) = \left(\alpha^2 - 1\right)\|R_k\|^2 = \left(\alpha^2 - 1\right)\sum_{i=1}^{p}\sum_{s=1}^{p} \left[cor\left(x_k^i, x_k^s\right)\right]^2.$$

   Por lo tanto la comparación entre dos puntos homotéticos $(\alpha R_k = R_l)$, depende de la magnitud de las correlaciones y de $\alpha^2 - 1$.

3. En caso que las matrices X_k no sean reducidas, se tiene

$$\|V_k\|^2 = \langle V_k, V_k\rangle = \sum_{i=1}^{p}\sum_{s=1}^{p} \left[cor\left(x_k^i, x_k^s\right)\right]^2 var\left(x_k^i\right) var\left(x_k^s\right).$$

   Así entonces, cuando hay estabilidad de las correlaciones entre dos 'instantes' k y l $(k < l)$ y las normas $\|V_k\|$ y $\|V_l\|$ son muy diferentes, se ha producido un aumento o una disminución en las varianzas de las variables de un instante al otro. Dependiendo de la naturaleza del problema analizado, puede ser interesante identificar los factores responsables de dichas variaciones.

Otras propiedades Supongamos que $V_l M = V_r M$ entonces :

1. Los ACP de los tripletes (X_l, M, D_l) y (X_r, M, D_r), tienen los mismos vectores y valores propios y, las componentes principales, en ambos casos, son combinaciones lineales con los mismos pesos, de las mismas variables observadas en los instantes k y l. Es decir, tienen la misma interpretación.

[2]Si $M \neq I$, se expresa M así: $M = L^t L$ y el problema se reduce al presente caso, puesto que $\langle V_k, V_l\rangle_{M,M} = \left\langle LV_k L^t, LV_l L^t\right\rangle_{I_p, I_p}$ (ver la prueba del teorema 1).

2. Si las matrices X_i son reducidas entonces las correlaciones de las variables con las componentes son iguales en ambos casos.

Prueba:

1. Sea C_l la matriz cuyas columnas son las componentes principales D_l - ortonormadas del ACP de (X_l, M, D_l), U la matriz cuyas columnas son los vectores propios $M-$ortonormados de V_lM y Δ_λ la matriz diagonal de los valores propios correspondientes. Se sabe que $C_l = X_lMU\Delta_{\frac{1}{\sqrt{\lambda}}}$. Entonces $c_l^j = \frac{1}{\sqrt{\lambda_j}}X_lMu_j$ y $c_r^j = \frac{1}{\sqrt{\lambda_j}}X_rMu_j$ son las componentes principales j - ésimas de los ACP's de los tripletes (X_l, M, D_l) y (X_r, M, D_r) respectivamente.

2. Las correlaciones de las variables con las componentes son

$$X_l^tD_lC_l = X_l^tD_lX_lMU\Delta_{\frac{1}{\sqrt{\lambda}}} = V_lMU\Delta_{\frac{1}{\sqrt{\lambda}}} = V_rMU\Delta_{\frac{1}{\sqrt{\lambda}}} = X_r^tD_rC_r.$$

Observación: En el caso $V_lM = \alpha V_rM$, entonces $\Delta_\lambda^l = \alpha\Delta_\lambda^r$ donde Δ_λ^l es la matriz diagonal de los valores propios del ACP del triplete (X_l, M, D_l). Además, las correlaciones de las variables definidas por X_l con las componentes, son:

$$X_l^tD_lC_l = U\Delta_l(\sqrt{\lambda})X_r^tD_rC_r = U\Delta_r(\sqrt{\lambda}) = \frac{1}{\sqrt{\alpha}}U\Delta_l(\sqrt{\lambda}) = \frac{1}{\sqrt{\alpha}}X_lD_lC_l.$$

Por lo tanto hay proporcionalidad en la estructura de correlaciones.

3. El compromiso

Suponemos que las matrices X_k son centradas y reducidas, por lo que las configuraciones a estudiar son las matrices de correlaciones.

Definiendo el compromiso como:

$$R = \sum_{i=1}^{K}\beta_iR_i.$$

donde β es vector propio de ΠS asociado al mayor valor propio λ_1, $\sum_{i=1}^k \beta_i = 1$ con $\beta_i \geq 0$ y $\|R_i\| = 1$,se prueba que R verifica las siguientes propiedades:

1. R es el objeto más correlacionado con los R_i, en el sentido que R es el que maximiza el promedio del cuadrado de las correlaciones de R con los R_i. Es decir, R maximiza el cociente

$$\frac{\sum_{i=1}^K p_i \left\langle \sum_{j=1}^K \alpha_jR_j,\ R_i \right\rangle^2}{\left\| \sum_{j=1}^K \alpha_jR_j \right\|^2} \text{ al variar } \alpha \in \mathbb{R}^K$$

2. Si $X^t = \left(X_1^t, X_2^t, \ldots, X_K^t\right)_{p\times n}$ con $n = \sum_{k=1}^K n_k$ y $D_\beta = diag(\beta_j D_j)_{n\times n}$ entonces $R = X^t D_\beta X$. Además las variables definidas por las columnas de X son centradas y reducidas respecto a D_β, por lo que podemos interpretar el compromiso como una matriz de correlaciones.

3. $cor_{D_\beta}(x^s, x^l) = \sum_{k=1}^K \beta_k \; cor_{D_k}(x_k^s, x_k^l)$ donde x^s, x^l son las variables de la matriz X (columnas s-ésima, l-ésima de X) y x_k^s, x_k^l son las correspondientes variables de la matriz X_k (columnas s-ésima y l-ésima de X_k). Puede notarse que la D_β – correlación entre dos variables de X es el promedio de las D_k-correlaciones entre las correspondientes variables de X_k.

4. Si todos los R_i son iguales entonces $\beta_i = p_i$ para todo i.

5. Si algún R_i es muy diferente a los demás ($\langle R_i, R_j\rangle = 0$, para todo $j \neq i$), éste no participa del compromiso ($\beta_i = 0$).

6. Si elegimos todos los pesos de los R_i iguales, esto es $\Pi = \frac{1}{K} I_K$, entonces los mayores β_k corresponden a los R_k que en promedio correlacionan más con el resto de los R_i .

Podemos afirmar de estas propiedades que el compromiso rescata lo que es común a las diferentes configuraciones y descarta las diferencias.

Prueba:

1. Desarrollando el numerador tenemos:

$$\sum_{i=1}^K p_i \left\langle \sum_{j=1}^K \alpha_j R_j, R_i \right\rangle^2 = \sum_{i=1}^K p_i (\sum_{j=1}^K \alpha_j \langle R_j, R_i\rangle)^2 =$$

$$\sum_{i=1}^K p_i (\sum_{j=1}^K \alpha_j S_{ij})^2 = \|S\alpha\|_\Pi^2 = \alpha^t S(\Pi S\alpha) = \langle \Pi S\alpha, \alpha\rangle_S$$

Por otra parte, el denominador se escribe como:

$$\|R\|^2 = \left\langle \sum_{j=1}^K \alpha_j R_j, \sum_{i=1}^K \alpha_i R_i \right\rangle = \sum_{j=1}^K \sum_{i=1}^K \alpha_j \alpha_i S_{ij} = \|\alpha\|_S^2 .$$

Luego

$$\frac{\sum_{i=1}^K p_i \left\langle \sum_{j=1}^K \alpha_j R_j, \; R_i \right\rangle^2}{\|R\|^2} = \frac{\langle \Pi S\alpha, \alpha\beta\rangle_s}{\|\alpha\|_s^2}$$

es un cociente de Rayleigh, por lo que alcanza su máximo cuando $\alpha = \beta$ es un vector propio de ΠS, asociado al mayor valor propio λ_1[4]. Por el teorema de Frobenius [1], el vector β se puede elegir con todas sus entradas no negativas.

2. Como las variables definidas por las matrices X_k son centradas y reducidas respecto de $D_k = diag(d_{ki})_{p\times p}$ se tiene:

$$\|x_k^s\|_{D_k} = 1 \text{ y } \overline{x}_k^s = \sum_{i=1}^{n_k} x_{ki}^s d_{ki} = 0 \text{ para cada } k.$$

Luego se sigue que la varianza y la media de la variable x^sdefinida por la columna s de la matriz X son:

$$\|x^s\|_{D_\beta}^2 = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \beta_k \left(x_{ki}^s\right)^2 d_{ki} = \sum_{k=1}^{K} \beta_k \sum_{i=1}^{n_k} \|x_{ki}^s\|_{D_k}^2 = \sum_{k=1}^{K} \beta_k = 1.$$

$$\overline{x}^s = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \beta_k x_{ki}^s d_{ki} = \sum_{k=1}^{K} \beta_k \sum_{i=1}^{n_k} \overline{x}_{ki}^s d_{ki} = 0$$

3. $cor_{D_\beta}(x^l, x^s) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \beta_k x_{ki}^l x_{ki}^s d_{ki} = \sum_{k=1}^{K} \beta_k cor_{D_k}(x_k^l, x_k^J)$.

4. Si $R_i = R_j$ para todo i, j, entonces como $\|R_i\| = 1$, entonces

$$S_{ij} = 1 \text{ y } \Pi S = \begin{pmatrix} p_1 & \dots & p_1 \\ \vdots & \vdots & \vdots \\ p_K & \dots & p_K \end{pmatrix}$$

Además como el rango de ΠS es 1, $\beta^t = (p_1, \dots, p_K)$ es un vector propio de ΠS asociado al valor propio $\lambda_1 = 1$.

5. Si R_t es ortogonal con todos los demás R_i la matriz ΠS tiene la t-ésima fila nula y el vector propio β tiene su entrada $\beta_t = 0$ por lo que R_t no forma parte del compromiso.

6. Como $\Pi S\beta = \lambda_1\beta$ y los pesos son iguales se tiene $\Pi = \frac{1}{K} I_K$ y $\frac{1}{K}\sum_{i=1}^{K} S_{ki}\beta_i = \lambda_1\beta_k$. Luego $\frac{1}{\lambda_1 K}\sum_{i=1}^{K} cor(R_k, R_i) = \beta_k$

4. Intraestructura

El estudio de la intraestructura involucra la representación bidimensional de las trayectorias (por alusión al tiempo) de las variables y, eventualmente, de los individuos. Ello permite explicar las desviaciones entre tablas de datos observadas en la interestructura, por medio de las desviaciones individuales de las variables en las trayectorias.

Sea $X_\beta^t = \left[\sqrt{\beta_1}X_1^t \dots \sqrt{\beta_K}X_K^t\right]$ y $D = diag\left(D_k\right)_{n\times n}$. Si r es el rango de X_β y $u_1, \dots, u_r$ son los vectores propios M-ortonormados del ACP de (X, M, D_β), asociados a los valores propios $\lambda_1 \geq \dots \geq \lambda_r > 0$, entonces el compromiso es $R = X^t D_\beta X = X_\beta^t D X_\beta$ y los u_i son vectores propios de RM. Se denotan con $c_1, \dots c_r$ las componentes principales correspondientes, de este ACP.

4.1. Representación de las variables

Se consideran representaciones de las variables definidas por las columnas de la tabla X (variables activas) y de las variables definidas por las columnas de las tablas X_k (variables suplementarias).

4.1.1. Variables activas

Por definición $X^t D_\beta c_s = \frac{1}{\sqrt{\lambda_s}} RMu_s$, luego

$$coord_{c_s}(x^j)^t = \left(x^j\right) D_\beta c_s = \frac{1}{\sqrt{\lambda_s}} R^j M u_s = \sum_{k=1}^{K} \frac{\beta_k}{\sqrt{\lambda_s}} R_k\left(j\right) M u_s.$$

donde R^j es la fila j de la matriz R. Esta representación corresponde a una imagen euclídea óptima de rango $q \leq p$, asociada a R.

4.1.2. Variables suplementarias

Se identifica la variable x_k^j con la variable suplementaria $(\widetilde{x}_k^j)^t = [0..,0(x_k^j)^t 0..,0]_{1\times n}$, cuya proyección ortogonal sobre c_s es:

$$coord_{c_s}(\widetilde{x}_k^j) = (\widetilde{x}_k^j)^t D_\beta c_s = \frac{\beta_k}{\sqrt{\lambda_s}} (x_k^j)^t D_k X_k M u_s = \frac{\beta_k}{\sqrt{\lambda_s}} R_k\left(j\right) M u_s$$

donde $R_k\left(j\right)$ es la fila j de R_k.

Nótese que las coordenadas de las variables observadas en el período entero son iguales al promedio de las coordenadas de las variables correspondientes a cada instante, salvo por la constante K:

$$coord_{c_s}(x^j)^t = \sum_{k=1}^{K} coord_{c_s}(\widetilde{x}_k^j)$$

4.2. Representación de los individuos

La representación de un individuo x_i es la usual del ACP, es decir su coordenada en el eje u_j es: $\text{coord}_{u_j}\left(x_i\right) = x_i^t M u_j$.

$$x_i = \sum_{j=1}^{r} \left\langle x_i, u_j \right\rangle_M u_j = \sum_{j=1}^{r} \left(x_i^t M u_j\right) u_j.$$

4.3. Relación entre la interestructura y las trayectorias de las variables

Se trata de identificar las variables que explican las desviaciones observadas en la interestructura. Por la propiedad 6. del compromiso deducida en la sección 3., se sabe los R_k mejor representados en el compromiso corresponden a los β_k aproximadamente iguales y de mayor magnitud. Por lo tanto interesan fundamentalmente las cantidades $\|R_k - R_l\|^2$

donde β_k y β_l son grandes y $\beta_k \approx \beta_l$. Sea $M = I$ entonces $\|\beta_k R_k - \beta_l R_l\|^2 \approx b \|R_k - R_l\|^2$ por lo que se establecerá una relación entre $\|R_k - R_l\|^2$ y las trayectorias de las variables. Por definición

$$\|R_k - R_l\|^2 = \sum_{j=1}^{p} \|R_k(j) - R_l(j)\|^2$$

Pero $R_k(j) = \sum_{h=1}^{p} [R_k(j)\, u_h]\, u_h^t$ donde $u_1, \ldots, u_p$ es una base $I-$ ortonormada de $\mathbb{R}^p$, luego,

$$\|R_k - R_l\|^2 = \sum_{j=1}^{p}\sum_{s=1}^{p} (R_k(j)\, u_s - R_l(j)\, u_s)^2 \approx \sum_{j=1}^{p}\sum_{s=1}^{r} \lambda_s \left(coord_{c_s}(\widetilde{x}_k^j) - coord_{c_s}(\widetilde{x}_l^j) \right)^2 .$$

Se ve que mientras más grande sea el desplazamiento de una variable j entre los instantes k y l, más aporta esta variable a la distancia entre R_k y R_l.

5. Conclusión

Las relaciones matemáticas obtenidas permiten dar interpretaciones estadísticas en Statis Dual. Así por ejemplo, las distancias observadas en el primer plano de la interestructura se pueden relacionar con las varianzas y correlaciones (sección 2.2). En relación con el compromiso se logró deducir su significado estadístico, comprobándose que posee las propiedades que debe tener todo buen "promedio" (sección 3). Por otra parte, las trayectorias de las variables se pueden relacionar con las distancias observadas entre los operadores R_k (sección 4.3). Esto facilita en los análisis de datos, la detección de las variables que más inciden en la magnitud de las distancias observadas en la representación bidimensional de la interestructura.

Finalmente, los resultados y fórmulas deducidos en este artículo hacen posible una posterior implementación computacional del método, lo cual como se sabe, es indispensable en la óptica del análisis de datos. Un artículo en el cual se incluyen los algoritmos necesarios y un ejemplo con resultados numéricos está en proceso de preparación.

Referencias

[1] Acua, O.; Ulate, F. (1994) *Matrices no Negativas.* Editorial de la Universidad de Costa Rica, San Jos.

[2] Glaçon, F. (1981) *Analyse Conjointe de Plusieurs Matrices de Données.*Thèse de 3-ème Cycle, Université Scientifique et Médicale de Grenoble, Grenoble.

[3] González, J.; Rodríguez, O. (1995) "Algoritmo e implementación del método Statis", *IX Simposio Métodos Matemáticos Aplicados a las Ciencias*, J. Trejos (ed.), UCR-ITCR, Turrialba.

[4] Lavit, Ch. (1988) *Analyse Conjointe de Tableaux Quantitatifs.* Méthode+Programmes. Masson, Paris.

[5] Lavit, C.; Escoufier, Y.; Sabatier, R.; Traissac, P. (1994). "The ACT (Statis method)", *Computational Statistics & Data Analysis, North-Holland*, **18, 97–119**.

[6] L'Hermier Des Plantes, H. (1976) *Structuration des Tableaux à Trois Indices de la Statistique.* Thèse de 3ème Cycle, Montpelier.

[7] Diday, E.; Lemaire, J.; Pouget, J.; Testu, F. (1982) *Éléments d'Analyse de Données.* Dunod, París.

[8] Saporta, G.; Lavallard, F.; (1996). *L'Analyse des Données Evolutives: Méthodes et Applications.* Editions Technip, París.