Revista INTEGRACIÓN
Universidad Industrial de Santander
Escuela de Matemáticas
Vol. 16, No 1, p. 33–44, enero–junio de 1998

Comparación de dos métodos factoriales cuando los datos son longitudinales

NANCY LACOURLY[‖]
LYDIA LERA[**]

Resumen

Se hace un estudio comparativo entre el método STATIS de Escoufier y el método de Procusto cuando se analizan dos o más configuraciones evolutivas de n puntos en un espacio euclidiano, a través de un análisis factorial sobre tablas de distancias y de la búsqueda de un compromiso. También se encuentra una relación entre R^2 y otros índices dados en la literatura. Los resultados se aplican a un problema meteorológico.

Abstract

We have done a comparative study between the STATIS method of Escoufier and Procrustes analysis across the coefficients RV and R^2 (normalized) to compare two or more evolutive configurations of n points in an euclidian space across factorial analysis on distance tables and the search of a compromise. Also it is found a relation between R^2 and other indexes given in the literature. The results are applied to a meteorological problem.

Key Words: Método STATIS, método de Procusto, datos longitudinales, análisis factorial sobre tablas de distancias, compromiso, INDSCAL, coeficientes de similitud y disimilitud.

1. Introducción

En los últimos tiempos han surgido nuevos métodos de tipo factorial cuyo objetivo principal es la búsqueda de estructuras comunes entre diferentes conjuntos de datos. Estos conjuntos pueden estar formados por:

‖Departamento de Ingeniería Matemática, Universidad de Chile, e-mail: nlacourl@dim.uchile.cl
**Centro de Matemática y Física Teórica, ICIMAF, CITMA, e-mail: lydia@cidet.icmf.

- Los mismos individuos medidos en diferentes ocasiones para las mismas variables.
- Los mismos individuos medidos en diferentes ocasiones para diferentes variables.
- Diferentes conjuntos de individuos a los que se les miden las mismas variables.

Los primeros forman parte de los llamados datos longitudinales, y son aquellos a los cuales nos referiremos en el trabajo, datos pertenecientes a fenómenos que evolucionan en el tiempo.

Los estudios longitudinales se caracterizan por observaciones repetidas en el tiempo para los mismos individuos. Cuando se miden más de 2 variables el problema se complica, por lo que se recomienda hacer previamente un estudio exploratorio para reducir dimensión y plantear un modelo (Lacourly, 1996). Estos datos se organizan en matrices cúbicas o triples (tablas de 3 entradas donde una o más de las de las entradas puede ser el tiempo).

Sea X una matriz cúbica o triple, cuyo elemento x_{ijk}, $(i,j,k) \in I \times J \times K$, representa al individuo i, para la variable j en la ocasión k. Se tienen k tablas formadas por los mismos individuos con mediciones en diferentes momentos (Lavit, 1988).

Si $p_1,...,p_n$ son los pesos no nulos de los individuos tal que $p_1+...+p_n=1$, entonces la matriz X contiene los valores de las variables centradas para los pesos de los individuos:

$$\sum_{i=1}^{n} piX_{ijk} = 0 \text{ para todo } j = 1,...,q \text{ y } k = 1,...,k.$$

Se comparan los métodos y se aplican a un problema meteorológico de comparación de series de temperaturas de la superficie del mar de 49 años en 15 puntos del Pacífico Sur.

2. Descripción de los métodos

2.1. El método STATIS

El método STATIS es un método exploratorio del análisis de datos cuya idea esencial es la búsqueda de estructuras comunes entre diferentes tablas o conjuntos de datos, que pueden estar conformados por mediciones recogidas en diferentes ocasiones sobre los mismos individuos $(I \times J_k \times K)$ o por diferentes individuos a los que se les miden las mismas variables $(I_k \times J \times K)$.

La estructura de los individuos en las tablas está descrita por las distancias mutuas entre individuos; es decir: sea X_k la matriz de datos formada por n filas y m_k columnas, donde n es el número de individuos, m_k es e l número de variables en la k-ésima matriz de datos; entonces $W_k = X_k X_k^T$ va a describir la estructura de los individuos en el interior de las matrices de datos. A esto se le llama intraestructura.

La relación entre las matrices de datos (interestructura) estará descrita por las distancias entre las matrices W_k y se deduce del producto escalar de Hilbert-Schmidt, que equivale a la suma de los cuadrados de las covarianzas entre las variables en diferentes ocasiones y está dado por

$$\langle W_{k1}, W_{k2}\rangle = tr\,(DW_{k1}DW_{k2}),$$

donde D es la matriz diagonal obtenida de los pesos de los individuos. Si los pesos de los individuos son iguales entonces

$$D = \left(\frac{1}{n}\right) I_n.$$

Los elementos de la matriz S son los coeficientes de covarianza vectorial de los productos escalares entre las matrices W_k, es decir,

$$S_{kikj} = \left\langle \frac{W_{ki}}{W_{kj}} \right\rangle = tr(DW_{ki}DW_{kj}) = (1/n^2)tr(X_{ki}X_{ki}^T X_{kj}X_{kj}^T)$$

$$S_{kiki} = \left\langle \frac{W_{ki}}{W_{ki}} \right\rangle = tr(DW_{ki}DW_{ki}) = \left(\frac{1}{n^2}\right) tr\left(X_{ki}X_{ki}^T X_{ki}X_{ki}^T\right)$$
$$= (\frac{1}{n^2})tr(X_{ki}X_{ki})^2;$$

luego

$$\|W_k\| = S_{kk}^{\frac{1}{2}}.$$

A partir de la matriz S se calculan los coeficientes RV (coeficientes de Rayleigh) (Escoufier, 1983) como

$$RV(k_i, k_j) = \frac{S_{k_ik_j}}{S_{k_ik_i}^{\frac{1}{2}} S_{k_jk_j}^{\frac{1}{2}}}. \tag{1}$$

Los coeficientes RV son llamados coeficientes de correlación vectorial entre las matrices W_k. Estos se interpretan como la evolución promedio de los individuos en las diferentes ocasiones.

Si se decide caracterizar los elementos de la tabla por $\frac{W_k}{\|W_k\|}$, entonces la matriz S no es más que la matriz de coeficientes RV.

A partir de esto se construye el compromiso W, que se puede considerar como el valor promedio de las tablas W_k. Cada estructura está descrita por los productos escalares entre individuos dadas por el compromiso W.

Asociado a los productos escalares entre W_k se halla una imagen euclidiana de las matrices X_k . Las coordenadas de los puntos M_k de las tablas se calculan a partir de los valores y vectores propios de $S\triangle$, donde $\triangle$ es la matriz diagonal de los pesos de las tablas. Si los pesos son iguales, basta diagonalizar S para obtener las coordenadas de los puntos M_k .

Los coeficientes a_k del compromiso $W = \sum_{k=1}^{p} a_k W_k$ son los componentes del vector

$$\alpha = \lambda_1^{\frac{-1}{2}} (\sum_{k=1}^{p} \pi_l S_{ll}^{\frac{1}{2}}) \Delta \gamma_1,$$

donde γ_1 son los elementos del vector propio de $S\Delta$ asociado al mayor valor propio λ_1 (todos son positivos).

En la imagen euclidiana del compromiso se sitúan los individuos para cada ocasión, y a esa gráfica continua se la llama trayectoria.

Las coordenadas de las posiciones de compromiso de los individuos sobre el eje 1 son los elementos del vector

$$\lambda_1^{\frac{1}{2}} (WD) \varepsilon_1 = \lambda_1^{-\frac{1}{2}} (WD) WD \varepsilon_1,$$

donde λ_1 y ε_1 son los valores y vectores propios respectivos de WD.

2.2. El método de Procusto[1]

El método Procusto (*Cox* y *Cox*, 1994) se utiliza para comparar dos configuraciones de n puntos en un espacio euclidiano de dimensiones q y p. Sean estas configuraciones o matrices $X_{n \times q}$ y $Y_{n \times p}$. Se supone que $p \geq q$ y que el $r-$ésimo punto de la primera matriz está en correspondencia uno a uno con el r-ésimo punto de la segunda matriz de datos.

Lo primero que se hace es completar con ceros las $p - q$ columnas de la matriz $X_{n \times q}$, y después se calcula la suma de las distancias entre los puntos de Y y los correspondientes de X,

$$R^2 = \sum_{r=1}^{n} (Y_r - X_r)^T (Y_r - X_r). \tag{2}$$

Los puntos en el espacio X serán transformados por dilatación, traslación, rotación y reflexión a las nuevas coordenadas X_r^*,

$$X_r^* = \rho A^T X_r + b,$$

donde A es una matriz ortogonal (da una rotación rígida), b un vector de traslación rígida y r una dilatación.

El método consiste en minimizar la nueva suma de distancias entre puntos,

$$R^2 = \sum_{r=1}^{n} (Y_r - \rho A^T X_r - b)^T (Y_r - \rho A^T X_r - b). \tag{3}$$

Resumiendo, lo que se hace en la práctica es:

[1]En la mitología griega Procusto era un bandido del Ática, quien, tras robar a los viajeros, los adaptaba al tamaño de un lecho de hierro, mutilando o descoyuntando a sus víctimas.

1. Se restan los vectores medios de cada uno de los respectivos puntos de las matrices, para tener los centroides en el origen.
2. Se obtiene la matriz de rotación
$$A = (X^T YY^T X)^{1/2}(Y^T X)^{-1}$$
y se rota la configuración X a XA.
3. Se multiplica cada coordenada por
$$\rho = \frac{tr(X^T YY^T X)^{1/2}}{t^r(X^T X)}$$
para escalar la matriz X.
4. Se calcula el valor minimizado y escalado
$$R^2 = 1 - \frac{[tr(X^T YY^T X)^{1/2}]^2}{[t^r(X^T X)tr(Y^T Y)]}, \qquad (4)$$
y a R^2 se lo llama el estadístico de Procusto.

Lo anterior se conoce como rotación ortogonal no ponderada de Procusto.

2.3. Propiedades de los índices RV y R^2

Coeficiente RV (Escoufier, 1973)

El coeficiente RV es una medida de similitud entre dos matrices de datos.

Sean X_{nxp} y Y_{nxq} las matrices de datos analizadas; de (2) se tiene que el coeficiente RV para estas matrices toma la forma

$$RV(1,2) = \frac{S_{12}}{S_{11}^{1/2} S_{22}^{1/2}},$$

donde

$$S_{12} = (\frac{1}{n^2})tr(XX^T YY^T),$$
$$S_{11} = (\frac{1}{n^2})tr(XX^T)^2,$$
$$S_{22} = (\frac{1}{n^2})tr(YY^T)^2;$$

sustituyendo tenemos

$$RV(1,2) = \frac{tr(XX^T YY^T)}{[tr(XX^T)^2 tr(YY^T)^2]^{1/2}}.$$

Algunas de las propiedades de RV son las siguientes (ver Escoufier, 1973):

1. Si $p = q = 1, RV = r^2$, donde r es el coeficiente de correlación simple entre las variables X y Y.
2. $0 \leq RV \leq 1$ y $RV = 0$ si y solamente si $S_{12} = 0$.
3. Simetría: $RV(X,Y) = RV(Y,X)$.

Coeficiente R^2

El coeficiente escalado R^2 (*Cox* y *Cox*, 1994) es una medida de disimilitud para comparar 2 matrices de datos. Sean X_{nxp} y Y_{nxq} las matrices; entonces se tiene (4).

Este coeficiente cumple las siguientes propiedades:

1. Simetría: $R^2(X,Y) = R^2(Y,X)$:

$$R^2(X,Y) = 1 - \frac{[tr(X^TYY^TX)^{1/2}]^2}{[tr(X^TX)tr(Y^TY)]} =$$

$$= 1 - \frac{[tr(XX^TYY^T)^{1/2}]^2}{[tr(X^TX)tr(Y^TY)]} =$$

$$= 1 - \frac{[tr(Y^TXX^TY)^{1/2}]^2}{[tr(Y^TY)tr(X^TX)]} = R^2(Y,X).$$

2. $R^2(X,Y) \geq 0$.

Para que R^2 sea no negativo es necesario que

$$\frac{[tr(X^TYY^TX)^{1/2}]^2}{[tr(X^TX)tr(Y^TY)]} \leq 1.$$

Utilizando la desigualdad de Cauchy- Schwartz

$$(tr(E^TF))^2 \leq (tr(E^TE))(tr(F^TF))$$

se tiene que en efecto

$$[tr(X^TYY^TX)^{1/2}]^2 = [tr(XX^TYY^T)^{1/2}]^2 \leq tr[(XX^T)^{1/2}(XX^T)^{1/2}]tr[(YY^T)(YY^T)^{1/2}]$$

$$= tr(XX^T)tr(YY^T)^{1/2} = tr(X^TX)tr(Y^TY).$$

3. $R^2(X,X) = 0$:

$$R^2(X,X) = 1 - \frac{[tr(X^TXX^TX)^{1/2}]^2}{[tr(X^TX)tr(X^TX)]} =$$

$$= 1 - \frac{\{tr[(X^TX)^2]^{1/2}\}^2}{[tr(X^TX)]^2} =$$

$$= 1 - 1 = 0.$$

4. $R^2(X,Y) = 0$ si y solo si
$$tr(X^TYY^TX)^{1/2} = 0.$$
Lazraq y Cléroux (1988) y Lazraq et al. (1992) hacen un estudio comparativo entre diferentes medidas de asociación y redundancia entre vectores aleatorios. A continuación enunciaremos algunos de los índices dados por ellos y los compararemos con el coeficiente R^2 .

3. Índices de asociación y redundancia

En este parágrafo se estudian 4 medidas de similitud y disimilitud entre 2 vectores. Los 3 primeros son de similitud y el último de disimilitud. Se comparan con RV y R^2 y se encuentran relaciones entre ellas.

1. **Coeficiente de Stewart y Love (1968):**
$$RV_5 = \frac{tr(Y^TX(X^TX)^{-1}X^TY)}{tr(Y^TY)}.$$
2. **Coeficiente de Robert y Escoufier (1976):**
$$RV_9 = \left[\frac{tr(Y^TX(X^TX)^{-1}X^TY)^2}{tr(Y^TY)^2}\right]^{1/2}.$$
3. **Coeficiente de Lingnes y Schonemann (1974):**
$$RLS = \frac{tr(X^TYY^TX)^{1/2}}{[tr(X^TX)tr(Y^TY)]^{1/2}}.$$
4. **Coeficiente de Rao (1965):**
$$RV_{10} = [tr(Y^TY)](1 - RV_5).$$

Se hizo un programa en MATLAB para calcular los coeficientes anteriores (Lacourly y Lera, 1997).

Estos índices cumplen las siguientes propiedades :

1.
$$0 \leq RV_i,\ RLS \leq 1,\ i = 5, 9. \tag{5}$$
Si $RV_5 = 0$ se tiene que $RV_{10} = tr(Y^TY)$, y si $RV_5 = 1$ resulta $RV_{10} = 0$. Si se escala RV_{10} y se divide por $tr(Y^TY)$, entonces
$$RV_{10} = 1 - RV_5,$$
de tal suerte que (5) se cumple para todos .

2. Si $p = q = 1, RV_i = RLS = r$ para $i = 5, 9$, el coeficiente de correlación simple entre 2 variables.

3. Simetría: $RLS(X,Y) = RLS(Y,X)$. Los otros coeficientes no son simétricos.

4. Si $X = Y$ se cumple que

$$RV_5 = RV_9 = RLS = 1 \quad \text{y} \quad RV_{10} = 0.$$

Lazraq y Cléroux (1988) y Lazraq et al. (1992) encontraron las siguientes relaciones :

1.
$$RV \leq RV_9.$$

2.
$$\frac{1}{(pq)^{1/2}} RLS^2 \leq RV \leq (pq)^{1/2} RLS^2.$$

3.
$$\frac{1}{(p)^{1/2}} RV_5 \leq RV_9 \leq (p)^{1/2} RV_5.$$

Por la forma en que se define RLS se obtiene la relación

$$R^2 = 1 - RLS^2.$$

Utilizando la desigualdad de Cauchy- Schwartz se prueba que se cumple la relación

$$\frac{1}{(pq)^{1/2}}(1 - R^2) \leq RV \leq (pq)^{1/2}(1 - R^2),$$

y que también se cumple que

$$\frac{1}{(pq)^{1/2}}(1 - R^2) \leq RV^2 \leq RV_9^2 \leq (pq)^{1/2} RV_5.$$

4. Comparación de los métodos

Se hizo un estudio comparativo entre los métodos STATIS y de Procusto mediante dos técnicas diferentes:

1. Análisis Factorial de Tablas de Distancias.
2. Búsqueda de un compromiso y representación de trayectorias.

4.1. Análisis Factorial de Tablas de Distancias

El objetivo del Análisis Factorial de Tablas de Distancias (AFTD) es construir una imagen euclidiana simple (I, d_I). Se basa en el esquema de dualidad (Cailliez y Pages, 1976)

$$W = (\frac{1}{2})(d_i^2 + d_j^2 - d_{ij}^2 - d^2),$$

donde d_{ij} es el índice de disimilitud R^2 de (5) simétrico y escalado de Procusto.

Se confeccionó un programa en MATLAB que realiza un AFTD a través del índice de Procusto (Lacourly y Lera, 1997). Se obtiene la matriz de disimilitud D, la matriz de rotación A y una representación gráfica de las tablas a comparar.

Esta técnica tiene el inconveniente de que representa solo las tablas.

4.2. Búsqueda de un compromiso y representación de trayectorias

Se busca un compromiso C que minimiza la suma de las distancias al cuadrado de k tablas con respecto a C. Aquí se utiliza el modelo INDSCAL (Cailliez y Pages, 1976).

Si $k = T$ entonces el compromiso C es tal que

$$min \sum_{j=1}^{T} \| X_j - \rho_j C P_j \|^2 ,$$

donde ρ_j es el coeficiente de dilatación, y P_j es una matriz de rotación ortogonal tal que $P_j P_j T = I$.

Si

$$Q = \sum_{t=1}^{T} [(\rho_t C P_t - X_t)(\rho_t C P_t - X_t)^T] - \sum_{t=1}^{T} \Lambda_t (P_t P_t^T - I),$$

se calcula el compromiso C tal que

$$\frac{\partial Q}{\partial C} = 0 \Rightarrow C = \frac{\sum_t \rho_t X_t P_t^T}{\sum_t \rho_t},$$

$$\frac{\partial Q}{\partial \rho_t} = 0 \Rightarrow \rho_t = \frac{tr(X_t P_t^T C)}{tr(CC^T)},$$

$$\frac{\partial Q}{\partial P_t} = 0 \Rightarrow \Lambda_t = \rho_t^2 C^T C - \rho_t C^T P_t.$$

Se supone que

$$\sum_{t=1}^{T} \rho_t^2 = 1.$$

Caso particular $t = 2$:

$$C1 = \frac{\rho_1 X_1 P_1^T + \rho_2 X_2 P_2^T}{\rho_1^2 + \rho_2^2},$$

$$\rho_1 = \frac{tr(X_1 P_1^T C)}{tr(CC^T)},$$

$$\rho_2 = \frac{tr(X_2 P_2^T C)}{tr(CC^T)},$$

$$\Lambda_1 = \rho_1^2 C^T C - \rho_1 C^T P_1, \tag{6}$$

$$\Lambda_2 = \rho_2^2 C^T C - \rho_2 C^T P_2. \tag{7}$$

Poniendo la restricción

$$\rho_1^2 + \rho_2^2 = 1,$$

y sumando (6) y (7) se tiene que

$$\begin{aligned}\Lambda_1 + \Lambda_2 &= \rho_1^2 C^T C - \rho_1 C^T X_1 P_1^T + \rho_2^2 C^T C - \rho_1 C^T X_2 P_2^T \\ &= C^T C - \rho_1 C^T X_1 P_1^T - \rho_1 C^T X_2 P_2^T \\ &= C^T C - C^T C = 0;\end{aligned}$$

por lo tanto $\Lambda_1 = -\Lambda_2$.

Se hizo un programa en MATLAB (Lacourly y Lera, 1997) para calcular el óptimo. A continuación se describe brevemente el algoritmo.

Etapa 0:

- Con compromiso inicial centrado

$$\rho_{k0} = \frac{tr(X_k P_k^T C_0^T)}{tr(C_0 C_0^T)}$$

se normaliza

$$\rho_{k0} = \frac{\rho_{k0}}{(\sum \rho_{k0}^2)^{1/2}},$$

para que se cumpla

$$\sum \rho_{k0}^2 = 1.$$

Etapa (m):

$$P_{k0} = (C_0^T X_k X_k^T C_0)^{1/2} (X_k^T C_0)^{-1}.$$

- C_{km}
- ρ_{km}
- P_{km}.

Etapa $(m+1)$:

- C_{m+1}
- ρ_{m+1}
- P_{m+1}.

5. Aplicación a datos meteorológicos

Se aplicaron los métodos analizados a un estudio meteorológico de comparación de series de temperaturas mensuales de la superficie del mar durante 49 años (1946-1994) en 15 puntos del Océano Pacífico Sur (Lacourly y Lera, 1997). Los datos pertenecen al proyecto FODECYT 1961110.

Se hizo un análisis STATIS para el coeficiente de disimilitud $1 - R^2$ utilizando un programa confeccionado en MATLAB (Lacourly y Lera, 1997), y se hizo un AFTD para R^2. Se analizó el caso en que las tablas eran los meses, las variables los puntos de Grilla y los individuos los años.

Se ve en las gráficas que las representaciones son similares. Los años 1957, 1968, 1982, 1974, 1982 y 1987 son años cuyas trayectorias se diferencian. Los años 1982 y 1987 son años con temperaturas muy elevadas, siendo años con el fenómeno de El Niño. El año 1968 fue un año muy seco.

Agradecimientos

Deseamos agradecer a la TWAS (Third World Academy of Sciences) y al CONICYT (Comisión Nacional de Ciencias y Tecnología) de Chile su apoyo financiero para la realización de este proyecto de investigación. También al Departamento de Ingeniería Matemática de la Universidad de Chile por su apoyo al desarrollo de este trabajo.

Referencias

[1] BORE, G., DI CIACCIO, A. (1989): "Comparisons among three factorial methods for analysing three-mode data". In: *Multiway Data Analysis*, eds: Coppy, R. y Bolasco, S. Elsevier Science Publishers B.V.

[2] CAILLIEZ, F. Y PAGES, J.P. (1976): *Introduction a L'Analyse des Données*. SMASH, Paris.

[3] CARLIER, A., LAVIT, C., PAGES, M., PERNIN, M.O., TURLOT, J.C. (1989): "A comparative review of methods which handle a set of methods which handle a set of indexed data tables". In: *Multiway Data Analysis*, eds: Coppy, R. y Bolasco, S. Elsevier Science Publishers B.V.

[4] COX, T.F. Y COX, M.A.A. (1994): *Multidimensional Scaling*. Monographs on Statistics and Applied Probability 59. Chapman & Hall .

[5] GREEN, B.F. (1952): "The orthogonal approximation of an oblique structure in factor analysis". *Psychometrika*, 17, 429 - 440.

[6] JACKSON, J.E. (1991): *A User's Guide to Principal Components.* John Wiley & Sons, New York.

[7] LACOURLY, N. (1996): "Panorama de métodos para el análisis de datos longitudinales". *Memorias del Seminario de Capacitación e Investigación Recolección y Análisis de Datos Longitudinales.*

[8] LAVIT, CH. (1988): *Analyse Conjointe de Tableaux Quantitatifs.* Masson, Paris.

[9] LAZRAQ, A. Y CLÉROUX, R. (1988): "Etude comparative de differentes measures de liaison entre deux vecteurs aleatoires et tests d'independance". *Statistique et analyse des données.* Vol 13, no. 1, p. 15-38.

[10] LAZRAQ, A. , CLÉROUX, R. Y KIERS, H.A.L. (1992): "Measures de liaison vectorielle et généralisation de L'Analyse Canonique". *Rev. Statistique Appliquée* XXXIX (1), p. 23-35 .

[11] LECHEVELIER, F. (1990): L'Analyse en Composantes conJointes d'une famille de triplets indexés. *Statistique et Analyse des Données,* vol. 15, no. 2, p. 35-75.

[12] MARDIA, K.V. , KENT, J.T. Y BIBBY. J.M. (1979): *Multivariate Analysis.* Academic Press, London.

[13] PROYECTO FONDECYT 1961110: *Predictibilidad estacional de anomalías pluviométricas y térmicas en las regiones norte y central de Chile.*