QÜESTIIÓ, vol. 21, 1 i 2, p. 99-108, 1997

ANÁLISIS FACTORIAL DE TABLAS MIXTAS: NUEVAS EQUIVALENCIAS ENTRE ACP NORMADO Y ACM

M. ISABEL LANDALUCE CALVO*
Universidad del País Vasco

En este trabajo se pone de manifiesto que es posible el Análisis Factorial de tablas mixtas sin modificar la naturaleza de ninguno de los dos conjuntos, cualitativo y cuantitativo, que las integran. Se propone codificar de manera apropiada las indicadoras de cada variable cualitativa tratando de respetar, en la medida de lo posible, la estructura inicial de esta última y posteriormente aplicar un Análisis en Componentes Principales (ACP) Normado al conjunto de variables. Los factores obtenidos para el grupo de variables nominales serán iguales a los factores resultantes de un Análisis de Correspondencias Múltiples (ACM) de la Tabla Disyuntiva Completa (TDC).

Factorial Analysis of Mixed Tables: New Equivalences between Weighted PCA and MCA

Keywords: Análisis en Componentes Principales Ponderado, Análisis de Correspondencias Múltiples, Tablas mixtas, Variables indicadoras, Ponderación

Clasificación AMS: 62-07, 62H25

*M. Isabel Landaluce Calvo. Departamento de Economía Aplicada III. (Econometría y Estadística) Facultad de CC.EE. y Empresariales. Universidad del País Vasco. Avda. Lehendakari Aguirre, 83. 48015 BILBAO. **e-mail:**il@alcib.bs.ehu.es

– Article rebut l'abril de 1996.

– Acceptat el gener de 1997.

1. INTRODUCCIÓN

En el campo del Análisis Factorial no es infrecuente encontrarse con tablas de datos que recogen variables cuantitativas y cualitativas conjuntamente, esto es, tablas mixtas. La práctica habitual cuando se analizan este tipo de tablas transforma unas u otras para conseguir un conjunto de variables homogéneas. La técnica más corriente consiste en codificar las variables cuantitativas: se divide el intervalo de R en el que dichas variables toman valores en subintervalos, convirtiéndolas en variables nominales y se aplica al conjunto resultante un Análisis Factorial de Correspondencias. Este método, no obstante, plantea, por una parte, problemas de codificación, elección de la partición, etc, y, por otra parte, conlleva una pérdida de información. Pero, sobre todo pone de manifiesto el hecho de que la variable numérica desaparace para ser reemplazada por el conjunto de todas las funciones correspondientes a la variable codificada.

En este trabajo se propone un tratamiento de tablas mixtas a través de un Análisis en Componentes Principales en el que las variables nominales van a ser debidamente ponderadas.

2. ACP NORMADO Y ACM APLICADOS A VARIABLES CUALITATIVAS

Una variable cualitativa $\vec{x}_k$ se puede considerar como una partición del conjunto I de individuos. Esta variable está representada por el conjunto Q_k de las variables indicadoras de las clases de esta partición o por el subespacio de R^I que engendran. Este subespacio tiene como dimensión el número de modalidades, ya que las variables indicadoras correspondientes a una misma variable son ortogonales entre sí. Es, por tanto, el subespacio de funciones numéricas que toman el mismo valor para los individuos que han elegido la misma modalidad. Así consideradas, un conjunto de K variables cualitativas, con Q variables indicadoras en total, está codificado bajo la forma de una tabla disyuntiva completa, a la que tradicionalmente se aplica un Análisis de Correspondencias Múltiples.

Se puede comprobar que los resultados obtenidos al aplicar un ACM son equivalentes a los obtenidos a partir de un ACP normado de la tabla disyuntiva completa cuando se han ponderado las variables indicadoras a través de la proporción de individuos que no las han elegido. Esta equivalencia la demuestran Brigitte Escofier y Jérôme Pagès, (1990, Cap. 7), siguiendo los siguientes razonamientos:

1. En ACM como consecuencia de la transformación de las columnas, de la métrica en R^I (proporcional a la métrica identidad) y de los pesos de los elementos, las modalidades de las variables poseen las siguientes propiedades cuando se las considera respecto al origen:

- Las modalidades de una misma variable son ortogonales entre sí. La transformación en perfiles no cambia su dirección.
- Todas las modalidades, de peso $\frac{I_q}{IK}$, siendo I_q el número de individuos que han elegido la modalidad q y K el número de variables cualitativas, tienen la misma inercia respecto al origen:

$$\textit{Inercia de la modalidad q respecto al origen} =$$

$$= \frac{I_q}{IK} \sum_i I(\frac{y_{iq}}{I_q})^2 = \frac{1}{K}$$

Siendo y_{iq} el término general de la TDC.

2. Se construye una nube de variables indicadoras con las mismas propiedades inerciales que la nube de modalidades en ACM. Para su posterior tratamiento mediante ACP normado se las considera divididas por su desviación típica pero no centradas. Con este fin, se asigna a cada variable indicadora el peso $\frac{(I-I_q)}{I}$. La nube, así definida, posee las siguientes propiedades inerciales comparables a las mencionadas en el punto anterior:

   - La métrica del espacio R^I es también la métrica indentidad, excepto por el coeficiente $1/I$. La dirección de las variables indicadoras no se modifica por la división entre su desviación típica.
   - Toda variable indicadora posee la misma inercia respecto al origen:

$$\textit{Inercia de la variable indicadora q respecto al origen} =$$

$$= \frac{I-I_q}{I} \sum_i \frac{1}{I} \frac{y_{iq}^2}{I_q \frac{(I-I_q)}{I^2}} = 1$$

3. Existe equivalencia entre las operaciones de centrado del ACM y del ACP cuando se efectúan sobre las nubes definidas anteriormente.

   En ACP, por una parte, el centrado de las variables se interpreta, en el espacio R^I, como una proyección de la nube de variables sobre el hiperplano ortogonal a la primera bisectriz.

   En ACM, por otra parte, considerado como un AFC de la tabla disyuntiva completa, la nube de las variables indicadoras está centrada en otro sentido: el origen está situado en el centro de gravedad de la nube de modalidades N_Q, $(\sum_k Q_k = Q)$. Esta nube, en ACM, presenta las siguientes propiedades:

- El centro de gravedad está situado sobre la primera bisectriz, se confunde con el perfil de la marginal sobre I, estando caracterizado por un perfil perfectamente plano.
- Está contenida en un hiperplano ortogonal a la primera bisectriz. Debido al caracter disyuntivo de la TDC, los vectores que unen el origen a las modalidades de una misma variable son ortogonales entre sí. El conjunto de las modalidades de las variables cualitativas engendran diferentes subespacios que, debido al carácter completo de la TDC, tienen una dirección común: la que une el origen con el centro de gravedad de la nube. Esta dirección se elimina con el centrado. En consecuencia, el centrado en ACM se interpreta como en ACP: una proyección sobre un hiperplano ortogonal a la primera bisectriz.

3. NUEVAS EQUIVALENCIAS ENTRE ACP NORMADO Y ACM

Un estudio profundo y detallado de la equivalencia entre ambos métodos bajo la ponderación señalada nos ha conducido a encontrar relaciones más concretas entre los elementos que intervienen en los análisis, relaciones que se resumen en los puntos siguientes.

3.1. Relación entre las matrices diagonalizadas

Las matrices analizadas en ACM y ACP normado de la TDC son equivalentes, excepto por un coeficiente, el inverso del número de variables cualitativas analizadas.

1. La matriz que se analiza en ACP normado de las variables indicadoras es la matriz de correlación entre las mismas, R. Siendo q_1 y q_2 dos variables indicadoras correspondientes a la misma variable cualitativa, I_{q_1} y I_{q_2} el número de individuos que han elegido la modalidad q_1 y la modalidad q_2 respectivamente, el coeficiente de correlación entre estas dos indicadoras se puede expresar de la siguiente manera:

$$r_{q_1,q_2} = -\sqrt{\frac{I_{q_1}}{(I-I_{q_1})}}\sqrt{\frac{I_{q_2}}{(I-I_{q_2})}}$$

   Se comprueba que el coeficiente de correlación entre las variables indicadoras de una misma variable cualitativa va a tener siempre signo negativo.

   El coeficiente de correlación entre dos variables indicadoras q y h, correspondientes a dos variables cualitativas, siendo y_{iq} y y_{ih} términos generales de la TDC, se puede expresar como sigue:

$$r_{q,h} = \frac{I\sum_{i=1}^{I} y_{iq}y_{ih} - I_q I_h}{\sqrt{I_q(I-I_q)}\sqrt{I_h(I-I_h)}}$$

Teniendo en cuenta que el término $\sum_{i=1}^{I} y_{iq}y_{ih}$ es el número de individuos que han elegido la modalidad q y la modalidad h al mismo tiempo, se comprueba que este coeficiente será positivo para el par de variables indicadoras que hayan sido elegidas mayoritariamente por los mismos individuos y negativo para las que hayan sido elegidas mayoritariamente por diferentes individuos.

2. La matriz que se diagonaliza en ACM es $M^{1/2}X^tDXM^{1/2}$ (al no ser X^tDXM una matriz simétrica) (1990, Cap.4), siendo D y M las matrices diagonales de pesos de individuos y variables, respectivamente. Los términos generales de esta matriz se pueden expresar de la siguiente manera:

– El producto de una modalidad q por sí misma será:

$$\sum_{i=1}^{I}(\frac{Iy_{iq}}{I_q}-1)^2\frac{1}{I}\frac{I_q}{IK} = \frac{1}{K}(\frac{I-I_q}{I})$$

– El producto de dos modalidades q_1 y q_2 de la misma variable será:

$$\sum_{i=1}^{I}(\frac{Iy_{iq_1}}{I_{q_1}}-1)(\frac{Iy_{iq_2}}{I_{q_2}}-1)\frac{1}{I}\sqrt{\frac{I_{q_1}}{IK}}\sqrt{\frac{I_{q_2}}{IK}} =$$

$$= -\frac{1}{K}\sqrt{\frac{I_{q_1}}{I}}\sqrt{\frac{I_{q_2}}{I}}$$

– El producto de dos modalidades q y h pertenecientes a dos variables cualitativas será:

$$\sum_{i=1}^{I}(\frac{Iy_{iq}}{I_q}-1)(\frac{Iy_{ih}}{I_h}-1)\frac{1}{I}\sqrt{\frac{I_q}{IK}}\sqrt{\frac{I_h}{IK}} =$$

$$= \frac{1}{K}\frac{I\sum_{i=1}^{I} y_{iq}y_{ih} - I_q I_h}{I\sqrt{I_q}\sqrt{I_h}}$$

3. Por otra parte, al igual que en el ACM, en ACP normado de las variables indicadoras al recibir éstas diferente ponderación, $\frac{I-I_q}{I}$, la matriz a diagonalizar, RM (considerando que todos los individuos poseen el mismo peso), no es simétrica. Se procede de igual manera que en aquel análisis y la matriz que se diagonaliza realmente es: $M^{1/2}RM^{1/2}$, esto es, cada elemento correspondiente de la matriz

R queda multiplicado por la raiz cuadrada de la ponderación asignada a cada una de las variables indicadoras que intervienen en su cálculo.

Se comprueba, por tanto, que las matrices diagonalizadas en los análisis contrastados son equivalentes, excepto por el coeficiente K, número de variables cualitativas.

3.2. Relación entre distancias

Las distancias definidas en ACP y en ACM aún siendo diferentes guardan una estrecha relación:

1. Con respecto a la distancia entre individuos, siendo i y l dos individuos cualesquiera, en ACP normado de la TDC ponderada se tiene:

$$\begin{aligned} d^2(i,l) &= \sum_{q=1}^{Q} m_q (y_{iq} - y_{lq})^2 = \\ &= 0, \textit{cuando los individuos han elegido las mismas modalidades} \\ &= \textit{un valor que crece con el número de modalidades que difieren entre los individuos.} \end{aligned}$$

La presencia de una modalidad elegida por pocos individuos, modalidad rara, aleja a sus poseedores del resto de individuos.

En ACM, por otra parte, se tiene:

$$\begin{aligned} d^2(i,l) &= \sum_{q=1}^{Q} \frac{IK}{I_q} (\frac{y_{iq}}{K} - \frac{y_{lq}}{K})^2 = \\ &= \frac{I}{K} \sum_{q} \frac{1}{I_q} (y_{iq} - y_{lq})^2 \end{aligned}$$

Como $(y_{iq} - y_{lq})^2$ vale 0 o 1, esta distancia crece con el número de modalidades que difieren entre los individuos. La presencia de una modalidad rara aleja a sus poseedores de los demás individuos.

2. Con respecto a la distancia entre dos modalidades, q y h, hay que recordar que en ACP se estudia la relación entre las variables indicadoras a través de su coeficiente de correlación. Por tanto, se reproduce aquí el resultado obtenido anteriormente:

$$r_{q,h} = \frac{I\sum_{i=1}^{I} y_{iq}y_{ih} - I_q I_h}{\sqrt{I_q(I-I_q)}\sqrt{I_h(I-I_h)}}$$

Se observa que la relación entre dos modalidades aumenta con el número de individuos que las han elegido a la vez, es decir, su distancia disminuye (hay que recordar que en ACP existe la siguiente equivalencia: $d^2(q,h) = 2(1-r_{q,h})$).

En ACM, por otra parte, se tiene:

$$\begin{aligned} d^2(q,h) &= \sum_i I(\frac{y_{iq}}{I_q} - \frac{y_{ih}}{I_h})^2 = \\ &= \frac{I}{I_q I_h}(I_q + I_h - 2\sum_i y_{iq}y_{ih}) \end{aligned}$$

Se observa que esta distancia decrece con el número de individuos que han elegido las dos modalidades a la vez.

3.3. Otras relaciones

En este apartado quedan reflejadas otras relaciones existentes entre los dos análisis.

1. En ACM la inercia de una modalidad q respecto al centro de gravedad vale: $\frac{1}{K}(1-\frac{I_q}{I})$. Al sumar las inercias de todas las modalidades se obtiene que la inercia total de la nube estudiada vale: $\frac{Q}{K} - 1$.

   En ACP normado la inercia de una modalidad q respecto al centro de gravedad vale: $1 - \frac{I_q}{I}$. Al sumar las inercias de todas las modalidades se obtiene que la inercia total de la nube estudiada vale: $Q - K$.

   Se alcanza, de nuevo, un resultado de gran interés: las inercias de los dos análisis son iguales, excepto por el coeficiente $\frac{1}{K}$.

2. En la representación simultánea obtenida a través de un ACP, un individuo i está situado próximo a las variables en las que posee mayoritariamente valores más altos que la media. Tal y como está definida la matriz X en el análisis

de las variables indicadoras ponderadas, un individuo estará próximo de las indicadoras que ha elegido.

En ACM un individuo i está situado, excepto por el coeficiente $1/\sqrt{\lambda}$, en el baricentro de las modalidades que ha elegido. Una modalidad q está situada, excepto por $1/\sqrt{\lambda}$, en el baricentro de los individuos que la poseen.

3. En ACM, por una parte, se sabe que las modalidades con pocos efectivos pueden contribuir mucho a la formación de los factores. En ACP, por otra parte, tal y como se ha definido la ponderación de las indicadoras, $\frac{I-I_q}{I}$, las modalidades raras tienen más peso que las modalidades con elevado número de efectivos.

En conclusión, se puede afirmar, por tanto, que un ACP normado de las variables indicadoras ponderadas, correspondientes a las modalidades de las variables cualitativas, conduce a los mismos factores sobre I que un ACM.

4. BIBLIOGRAFÍA

[1] **B. Escofier** and **J. Pagès**. (1986). «Le Traitement des Variables Qualitatives et Tableaux Mixtes Par Analyse Factorielle Multiple». *Data Analysis and Informatics*, **IV(2)**, 179–191.

[2] **B. Escofier** and **J. Pagès**. (1990). *Analyses Factorielles Simples et Multiples: Objectifs, Méthodes et Interprétation*. Dunod, Paris, 2éme edition.

ENGLISH SUMMARY

FACTORIAL ANALYSIS OF MIXED TABLES: NEW EQUIVALENCES BETWEEN WEIGHTED PCA AND MCA

M. ISABEL LANDALUCE CALVO*

Universidad del País Vasco

The Factorial Analysis of Mixed Tables, when the nature of the numerical and categorical variables remains unchanged, is possible because the Weighted Principal Components Analysis (WPCA) of qualitative indicator variables is equivalent to the Multiple Correspondence Analysis (MCA). The adequate weighting, i.e. the proportion of individuals that has not chosen the correspondent modality, allows us to analyze the Disjunctive Complete Table through a WPCA. In this way, we obtain the same factors as in a MCA.

Keywords: Weighted Principal Components Analysis, Multiple Correspondence Analysis, Mixed Tables, Indicator Variables, Weighting

AMS Classification: 62-07, 62H25

*M. Isabel Landaluce Calvo. Departamento de Economía Aplicada III. (Econometría y Estadística) Facultad de CC.EE. y Empresariales. Universidad del País Vasco. Avda. Lehendakari Aguirre, 83. 48015 BILBAO. **e-mail:**il@alcib.bs.ehu.es

– Received april 1996.

– Accepted january 1997.

The Factorial Analysis of Mixed Tables, when the nature of the numerical and categorical variables remains unchanged, is possible because the Weighted Principal Components Analysis (WPCA) of qualitative indicator variables is equivalent to the Multiple Correspondence Analysis (MCA). The adequate weighting, i.e. the proportion of individuals that has not chosen the correspondent modality, allows us to analyze the Disjunctive Complete Table through a WPCA. In this way, we obtain the same factors as in a MCA.

Escofier & Pagès (1990, Ch. 7) established this equivalence for a very general setting. In a detailed study of the relationship between these two methods, we have found new and more specific equivalences. These equivalences are as follows:

1. The analyzed matrices in MCA and WPCA of the Disjunctive Complete Table are equivalent, except for a constant given by the inverse of the number of cualitative variables. Thats, with this exception, these analyses have the same inertia.

2. The distances between different elements are highly related in these two methods, in spite of not being the same.

3. In the simultaneous representation that a PCA provides an individual is projected close to those variables with values higher than the mean. In the analysis of the weighted indicator variables an individual is located next to the chosen modalities. In MCA an individual is located, except for a constant, on the baricentre of the chosen modalities. A modality is located, apart from the same constant, on the baricentre of the individuals that have chosen this modality.

4. We know that the modalities that have been rarely chosen in an MCA can have enough contribution to the formation of the factors. In addition, for PCA and due to the way the weights are defined, the indicators for the rare modalities have more weight than the modalities more frequently chosen.

As a conclusion, it is posible to carry out the Factorial Analysis of Mixed Tables without changing the nature of the numerical and categorical variables. It is not necessary to transform the numerical variables but we need to put same adequate weights to the indicators of the modalities associated with the categorical variables.