QÜESTIIÓ, vol. 23, 3, p. 465-488, 1999

INDEPENDENCIA ENTRE LAS CUESTIONES EN EL ANÁLISIS FACTORIAL DE TABLAS DISYUNTIVAS INCOMPLETAS CON PREGUNTAS CONDICIONADAS

A. ZÁRRAGA
B. GOITISOLO
Universidad del País Vasco-Euskal Herriko Unibertsitatea*

El análisis de correspondencias múltiples (ACM) estudia la relación entre varias variables cualitativas definidas sobre una misma población. Sin embargo, una de las principales fuentes de información son las encuestas donde es frecuente encontrar cierto número de datos ausentes y de preguntas condicionadas. Escofier (Escofier 1981) propone analizar la tabla disyuntiva incompleta sustituyendo la marginal real de la tabla sobre los individuos por una marginal impuesta constante. El análisis de la tabla disyuntiva incompleta está asociado a unas tasas de inercia pequeñas que no deben ser interpretadas como partes de información explicada por los ejes. Se estudiará el caso en el cual las cuestiones son independientes dos a dos y se propondrá una corrección a estas tasas de inercia en el análisis con marginal modificada de una tabla disyuntiva incompleta.

Independence between questions in the factor analysis of incomplete disjunctive tables with conditioned questions

Palabras clave: Análisis de correspondencias múltiples, tabla disyuntiva incompleta, independencia entre variables cualitativas, valores propios, tasas de inercia

Clasificación AMS (MSC 2000): 62H25

* Universidad del País Vasco-Euskal Herriko Unibertsitatea. Dpto. de Economía Aplicada III. Fac. CCEE y EE. Avda. Lehendakari Aguirre, 83. 48015 Bilbao. E-mails: az@alcib.bs.ehu.es y bg@alcib.bs.ehu.es.
Este trabajo ha sido financiado por el Proyecto de Investigación PB98-0149 de la Dirección General de Enseñanza Superior del Ministerio Español de Educación y Ciencia y el Proyecto UPV 038.321-HA041/99 de la Universidad del País Vasco (UPV/EHU).

– Recibido en febrero de 1998.
– Aceptado en julio de 1999.

1. INTRODUCCIÓN

Las razones por las que existen datos ausentes pueden ser diversas, revelando problemas que deberán ser tratados también de diferentes formas según se indica en §4 y en (Escofier & Pagès 1992).

- Las no respuestas pueden estar causadas por un olvido involuntario del individuo y no tener un significado especial. Suelen representar una proporción muy pequeña de los datos y afectar por igual a todas las cuestiones e individuos.
- Una segunda razón para la existencia de la no respuesta corresponde a una actitud particular del entrevistado, deseo de no revelar cierta información (por ejemplo, los ingresos, la ideología política, etc). Este tipo de no respuesta no se reparte de forma aleatoria en la tabla de datos sino que afecta más a determinadas cuestiones y grupos de individuos.
- Otra razón por la que aparecen las tablas disyuntivas incompletas −definidas en §2− muy frecuente en las encuestas se debe a la existencia de preguntas condicionadas, es decir, aquéllas a las cuales un individuo debe contestar o no dependiendo de cual haya sido su respuesta a una cuestión anterior. Por ejemplo, se le pregunta si sabe o no inglés; a continuación, y sólo si ha respondido saber inglés, se le pregunta su nivel de inglés en determinados aspectos. En una encuesta pueden existir varios grupos de preguntas condicionadas (los que saben inglés, francés, sólo los que tienen familiares y se relacionan con ellos contestarán la frecuencia con que lo hacen, etc). En este caso, la no respuesta se agrupa en un determinado número de cuestiones (aquellas cuya respuesta está condicionada por una pregunta anterior) y caracteriza a determinados grupos de individuos.

Tras la definición de las tablas disyuntivas incompletas y la notación básica utilizada −§2−, se verá el problema que plantea la aplicación del análisis de correspondencias múltiples clásico a este tipo de tablas −§3− y algunas posibles alternativas dependiendo de las razones de la ausencia −§4−. La solución propuesta para el caso de preguntas condicionadas (el análisis con marginal modificada) se desarrolla −§5 a §9− insistiendo en las diferencias con el análisis de correspondencias clásico (por ejemplo, en la elección del origen −§6− y número de ejes −§9−) y en los diferentes resultados según se posea una proporción pequeña de los datos ausentes o esta proporción sea elevada (relaciones entre los factores −§8−).

En §10 se estudia el caso de independencia entre las cuestiones y a partir de los resultados obtenidos se propone una corrección a las tasas de inercia calculadas en el análisis general con marginal modificada de la tabla disyuntiva incompleta.

2. DEFINICIÓN DE LAS TABLAS DISYUNTIVAS INCOMPLETAS Y NOTACIÓN

Se considera la tabla de datos que recoge en forma lógica y disyuntiva las respuestas de un conjunto de individuos a un conjunto de preguntas o cuestiones, poseyendo cada una de ellas un conjunto finito de modalidades de respuesta. En el análisis de correspondencias múltiples clásico se impone a todos los individuos la obligación de pertenecer a alguna de las modalidades de cada cuestión y se denomina a la tabla así obtenida tabla disyuntiva completa (Z). Se dirá que tal tabla de datos es disyuntiva incompleta (Z^*) cuando los individuos no dan respuesta a una o más de las cuestiones preguntadas.

Tabla 1

Tabla disyuntiva completa (Z) o Tabla disyuntiva incompleta (Z^*)

	$q=1$			$\ldots$	q	$\ldots$	$q=Q$
	$j=1$	$\ldots$	$j=J_1$		j		
1	1		0				
2	0		0				
3	0		1				
$\vdots$							
i					z_{ij}		
$\vdots$							
n							

donde:

$\mathcal{Q} = \{1,\ldots,q,\ldots,Q\}$ es el conjunto de variables a las cuales debe responder el individuo

$\mathcal{J}_q = \{1,\ldots,j,\ldots,J_q\}$ es el conjunto de modalidades de la variable $q \in \mathcal{Q}$

$\mathcal{J} = \{1,\ldots,j,\ldots,J\} = \cup_{q=1}^{Q}\mathcal{J}_q$ es el conjunto de modalidades de todas las variables

$\mathcal{I} = \{1,\ldots,i,\ldots,n\}$ es el conjunto de individuos

$$z_{ij} = \begin{cases} 1 & \text{si el individuo } i \in \mathcal{I} \text{ responde la modalidad } j \in \mathcal{J} \\ 0 & \text{en otro caso} \end{cases}$$

$z_{i.} = \sum_{j\in\mathcal{J}} z_{ij}$ es el número de cuestiones a las que responde el individuo $i \in \mathcal{I}$

$z_{.j} = \sum_{i\in\mathcal{I}} z_{ij}$ es el número de individuos que eligen la modalidad $j \in \mathcal{J}$

Se denotará $z_{.j}^{q}$ cuando interese dejar constancia de la variable $q \in \mathcal{Q}$ a la que pertenece dicha modalidad

$z = \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} z_{ij}$ es el total de la tabla

En las tablas disyuntivas incompletas −al igual que en las completas analizadas mediante el ACM clásico− las variables siguen estando definidas a través de un conjunto de modalidades a las que el individuo debe responder sobre su pertenencia ($z_{ij} = 1$) o no ($z_{ij} = 0$).

Pero, en ocasiones, esas modalidades correspondientes a una misma variable no están definidas en forma completa, es decir, el individuo puede no pertenecer a ninguna de ellas; en otras ocasiones a pesar de estar definidas en forma completa el individuo puede no revelar a que modalidad pertenece; en ambos casos:

$$z_{ij} = 0 \qquad i \in \mathcal{I} \qquad \forall j \in \mathcal{J}_q \quad q \in \mathcal{Q}$$

Por ello será necesario definir también una variable que toma el valor 1 si el individuo $i \in \mathcal{I}$ responde a la cuestión $q \in \mathcal{Q}$ y 0 en caso contrario:

$$z_{i.}^q = \sum_{j \in \mathcal{J}_q} z_{ij} \qquad \forall q \in \mathcal{Q} \qquad \forall i \in \mathcal{I}$$

Y se denotará por z_q el número de individuos que han respondido a la cuestión $q \in \mathcal{Q}$:

$$z_q = \sum_{j \in \mathcal{J}_q} z_{.j} \qquad \forall q \in \mathcal{Q}$$

En resumen, las tablas disyuntivas incompletas se caracterizan porque en ellas dejan de cumplirse algunas de las relaciones que se dan en las completas:

$$\begin{aligned} z_{i.}^q &= 1 && \forall q \in \mathcal{Q} && \forall i \in \mathcal{I} \\ z_q &= n && \forall q \in \mathcal{Q} \\ z_{i.} &= Q && \forall i \in \mathcal{I} \\ z &= nQ \end{aligned}$$

Como ya es sabido, en todo análisis de correspondencias se definen ((Escofier & Pagès 1992), (Lebart, Morineau & Tabard 1977) y (Abascal & Grande 1989) entre otros) las frecuencias relativas conjuntas y marginales como:

$$\begin{aligned} f_{ij} &= \frac{z_{ij}}{z} && \forall i \in \mathcal{I} && \forall j \in \mathcal{J} \\ f_{i.} &= \frac{z_{i.}}{z} = \sum_{j \in \mathcal{J}} f_{ij} && \forall i \in \mathcal{I} \\ f_{.j} &= \frac{z_{.j}}{z} = \sum_{i \in \mathcal{I}} f_{ij} && \forall j \in \mathcal{J} \end{aligned}$$

y los perfiles fila $i, i \in \mathcal{I}$:

$$\frac{z_{ij}}{z_{i.}} \qquad \forall j \in \mathcal{J}$$

y perfiles columna $j, j \in \mathcal{J}$:

$$\frac{z_{ij}}{z_{.j}} \qquad \forall i \in \mathcal{I}$$

que forman las nubes $\mathcal{N}(\mathcal{I}) \in \Re^J$ y $\mathcal{N}(\mathcal{J}) \in \Re^n$ respectivamente.

3. PROBLEMA QUE PLANTEA LA APLICACIÓN DEL ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES HABITUAL A UNA TABLA DISYUNTIVA INCOMPLETA

El problema que plantea este tipo de tablas es el mismo independientemente de la razón para esa ausencia de datos: la marginal sobre $\mathcal{I}$ ya no es constante.

Podría pensarse en aplicar directamente el análisis de correspondencias simples a esta tabla, (notar que originalmente se creó para el estudio de tablas de contingencia donde las marginales no son constantes). Sin embargo, cuando se posee una tabla disyuntiva incompleta, la distancia χ^2 y los pesos definidos en el análisis clásico no se ajustan a los objetivos, ya conocidos de un análisis de correspondencias.

La aplicación de la distancia χ^2 entre dos perfiles fila i e $i' \in \mathcal{I}$ sería:

$$\begin{aligned} d^2(i,i') &= \sum_{j\in\mathcal{J}} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 = \\ &= \sum_{j\in\mathcal{J}} \frac{z}{z_{.j}} \left(\frac{z_{ij}}{z_{i.}} - \frac{z_{i'j}}{z_{i'.}} \right)^2 \end{aligned} \tag{1}$$

Si los individuos i e $i' \in \mathcal{I}$ no contestan al mismo número de preguntas, entonces $z_{i.}, i \in \mathcal{I}$ difiere de $z_{i'.}, i' \in \mathcal{I}$ y por tanto la distancia χ^2 aumenta también con las respuestas comunes. Este es, por tanto, un concepto de distancia no deseable puesto que no reflejaría la similaridad entre individuos $-$en términos de modalidades comunes elegidas$-$ buscada en un análisis de correspondencias.

La distancia χ^2 entre dos perfiles columna j y $j' \in \mathcal{J}$ sería:

$$\begin{aligned} d^2(j,j') &= \sum_{i\in\mathcal{I}} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 = \\ &= \sum_{i\in\mathcal{I}} \frac{z}{z_{i.}} \left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}} \right)^2 \end{aligned} \tag{2}$$

En esta distancia χ^2 cada individuo tendría una ponderación distinta dependiendo del número de respuestas elegidas. No parece lógico asignar menos importancia a aquellos individuos que responden a la totalidad de las preguntas frente a quienes no lo hacen.

Por tanto, la aplicación directa del análisis de correspondencias clásico no es adecuada al estudio de las tablas disyuntivas incompletas.

4. POSIBLES ALTERNATIVAS AL PROBLEMA DE LAS TABLAS DISYUNTIVAS INCOMPLETAS: ANÁLISIS FACTORIAL CON MARGINAL MODIFICADA

Una vez descartada la aplicación del análisis clásico, se debe buscar otra alternativa para el estudio de las tablas disyuntivas incompletas que se adapte mejor. Se busca un método que minimice la influencia de la no respuesta sobre el análisis.

Una solución evidente desde el punto de vista analítico sería eliminar aquellos individuos que no responden a todas las cuestiones, obteniendo de esta forma una tabla disyuntiva completa. Con esta alternativa perderíamos la información referente a esos individuos en el resto de las cuestiones; información que puede no ser importante cuando la no respuesta se debe a un descuido y por tanto representa una pequeña proporción sobre el total, pero que alteraría los resultados cuando revela una actitud particular o implica a un gran número de individuos (caso de las preguntas condicionadas).

Una práctica habitual es crear para cada variable con datos ausentes una modalidad de no respuesta, obteniendo de esta forma una tabla disyuntiva completa a la que se puede aplicar el análisis clásico. Esta solución puede ser adecuada cuando la no respuesta se debe a una actitud particular del individuo (deseo de no revelar determinada información por ejemplo); sin embargo, cuando la no respuesta se debe a un descuido involuntario la modalidad de no respuesta no tendría una interpretación adecuada y perturbaría los resultados.

En los casos en los que la no respuesta se debe a la existencia de preguntas condicionadas ya se ha indicado en la introducción que caracteriza a un grupo de individuos. A pesar de ello, la inclusión de una modalidad de no respuesta en cada cuestión no sería adecuada, puesto que se estaría creando una serie de modalidades todas ellas con el mismo perfil e idéntico a una de las modalidades de la pregunta condicionante (no saben inglés) o a una combinación lineal de ellas («no tienen familiares» y «tienen pero no se relacionan»). Esto podría perturbar los resultados hasta el punto de llegar a crear uno de los primeros ejes del análisis, como así ocurre en la aplicación a la Encuesta de Condiciones de Vida de 1989 de la Comunidad Autónoma de Euskadi presentada en (Goitisolo & Zárraga 1998a).

Escofier (Escofier 1981) propone sustituir la marginal real de la tabla ($f_{i.} = z_{i.}/z, i \in \mathcal{I}$ que no es constante), por una marginal constante $g_{i.} = 1/n, i \in \mathcal{I}$ en todo el análisis.

Posteriormente (Benali 1985), (Benali 1988), (Benali & Escofier 1987) y (Escofier 1990) utilizan también esta técnica, pero siempre para el caso de tablas disyuntivas

incompletas donde la no respuesta se debe a una omisión involuntaria y el caso de modalidades de efectivo débil. En el cálculo de la distancia entre dos individuos (y por tanto, en la inercia de la nube), las modalidades tienen una ponderación inversa a su efectivo. En consecuencia, las modalidades muy raras pueden influir demasiado; Benali y Escofier proponen eliminarlas y tratar a esos individuos como si no hubieran dado respuesta a la cuestión.

Se analizará en detalle lo adecuado de esta sustitución y las consecuencias sobre el análisis, tanto para los casos de omisión involuntaria y de modalidades raras como para el caso de preguntas condicionantes, en el que centraremos nuestro interés.

5. NUBE DE INDIVIDUOS: $\mathcal{N}(\mathcal{I})$

El punto $i \in \mathcal{I}$ se representa en $\Re^J$ por el perfil $\dfrac{f_{ij}}{g_{i.}} = n\dfrac{z_{ij}}{z}, i \in \mathcal{I}, j \in \mathcal{J}$. Este perfil es diferente del perfil obtenido en el análisis de correspondencias múltiples clásico ($z_{ij}/Q,\ i \in \mathcal{I}, j \in \mathcal{J}$) al ser el efectivo total de la tabla (z) distinto de nQ.

La distancia cuadrática propuesta entre dos individuos i e $i' \in \mathcal{I}$ es:

$$
\begin{aligned}
d^2(i,i') &= \textstyle\sum_{j\in\mathcal{J}} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{g_{i.}} - \frac{f_{i'j}}{g_{i'.}} \right)^2 = \\
&= \textstyle\frac{n^2}{z} \sum_{j\in\mathcal{J}} \frac{1}{z_{.j}} \left(z_{ij} - z_{i'j} \right)^2
\end{aligned}
\tag{3}
$$

Se comprueba que únicamente las respuestas diferentes hacen aumentar la distancia, sin tener en cuenta si ambos individuos responden al mismo número de cuestiones o no.

La ponderación de cada modalidad es, al igual que en correspondencias múltiples con datos completos, el inverso de su efectivo (la distancia aumenta en mayor proporción cuando la modalidad poseída por sólo uno de los individuos es rara).

Considerar la distancia anterior entre los puntos i e $i' \in \mathcal{I}$ es equivalente a buscar la distancia euclídea habitual en un espacio dotado de métrica $1/f_{.j},\ j \in \mathcal{J}$.

Cada punto $i \in \mathcal{I}$ está dotado de un peso $g_{i.} = 1/n,\ i \in \mathcal{I},$ que a pesar de no venir representado por la marginal $f_{i.},\ i \in \mathcal{I}$ −del ACM clásico− coincide con el peso que se asigna a los individuos en correspondencias múltiples habitual. Este peso constante significa que todos los individuos tienen la misma importancia, independientemente del número de cuestiones que han respondido. Es por tanto, más adecuado que $f_{i.}, i \in \mathcal{I}$.

La coordenada j-ésima del centro de gravedad de la nube es:

$$
G_I(j) = f_{.j} = \frac{z_{.j}}{z} \qquad \forall j \in \mathcal{J}
$$

que coincide con la correspondiente al centro de gravedad de la nube de individuos en correspondencias múltiples habitual.

Este será también el origen de los ejes de máxima inercia que se han de buscar.

6. NUBE DE MODALIDADES: $\mathcal{N}(\mathcal{J})$

El punto $j \in \mathcal{J}$ se representa en $\Re^n$ por el perfil $\dfrac{f_{ij}}{f_{.j}} = \dfrac{z_{ij}}{z_{.j}},\ i \in \mathcal{I},\ j \in \mathcal{J}$, es decir, el mismo que en el A.C.M. clásico.

La distancia cuadrática propuesta entre dos modalidades j y $j' \in \mathcal{J}$ es:

$$
\begin{aligned}
d^2(j,j') &= \sum_{i\in\mathcal{I}} \frac{1}{g_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 = \\
&= n \sum_{i\in\mathcal{I}} \left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}} \right)^2
\end{aligned}
\tag{4}
$$

Esta distancia es semejante a la utilizada cuando la tabla es completa y apropiada también para el caso de preguntas condicionadas. Equivale a considerar la distancia euclídea en un espacio dotado de métrica $1/g_{i.}, i \in \mathcal{I}$.

Cada punto $j \in \mathcal{J}$ está dotado de un peso $f_{.j} = z_{.j}/z,\ j \in \mathcal{J}$, coincide con el asignado en correspondencias múltiples de tablas completas y supone que cada modalidad tiene una importancia proporcional a la población que representa. Las modalidades tienen un peso en la construcción de los ejes tanto menor cuanto menor sea su efectivo.

La coordenada i-ésima del centro de gravedad de la nube es:

$$G_J(i) = f_{i.} = \frac{z_{i.}}{z} \qquad \forall i \in \mathcal{I}$$

Sin embargo, en (Goitisolo & Zárraga 1998b) se justifica la elección como origen del punto O_J (cuya coordenada $i-$ésima es $g_{i.}, i \in \mathcal{I}$). La diferencia entre los puntos O_J y G_J conlleva que en el análisis las nubes han de ser siempre centradas, no existiendo la equivalencia conocida en correspondencias múltiples con tablas completas entre los análisis de las nubes centradas y no centradas.

7. OBTENCIÓN DE LOS FACTORES DE AMBAS NUBES

Calcular la sucesión de ejes ($u_s, s \in \mathcal{S} = \{1, \ldots, s, \ldots, S\}, S \leq J$) que maximizan la inercia proyectada de la nube $\mathcal{N}(\mathcal{I})$ equivale a:

$$
\begin{array}{lll}
\text{maximizar:} & {u_s}^T M X^T P X M u_s & \\
\text{sujeto a:} & {u_s}^T M u_s = 1 & \\
 & {u_s}^T M u_t = 0 & \forall t < s
\end{array} \tag{5}
$$

donde:

- X es una matriz $(n \times J)$ de término general:

$$x_{ij} = \frac{f_{ij}}{g_{i.} f_{.j}} - 1 \qquad i \in \mathcal{I} \qquad j \in \mathcal{J} \tag{6}$$

  Al igual que en análisis de correspondencias clásico cada elemento de esta matriz contiene las desviaciones entre la tabla de datos $f_{ij}, i \in \mathcal{I}, j \in \mathcal{J}$ y una tabla de término general que corresponde a la hipótesis de independencia. La diferencia con el análisis clásico radica en que la frecuencia relativa marginal correspondiente a las filas es impuesta en función del número de filas en lugar de obtenida a partir de los datos.

- M es una matriz diagonal correspondiente a la métrica del espacio:

$$m_j = f_{.j} \qquad j \in \mathcal{J} \tag{7}$$

- P es la matriz (también diagonal) de pesos:

$$p_i = g_{i.} \qquad i \in \mathcal{I} \tag{8}$$

Se puede demostrar (Escofier & Pagès 1992) que la nube definida por las filas de la matriz X, con la métrica y los pesos considerados, es isomorfa de la definida en §5 como objetivo del estudio. Ambas nubes mantienen las mismas distancias entre dos puntos cualesquiera.

La resolución de este problema lleva a la diagonalización de la matriz $X^T P X M$ de orden $(J \times J)$ cuyo término general es:

$$a_{jj'} = n \sum_{i \in \mathcal{I}} \frac{f_{ij}\, f_{ij'}}{f_{.j}} - f_{.j'} \qquad j, j' \in \mathcal{J} \tag{9}$$

Las proyecciones de la nube de individuos sobre los ejes de máxima inercia resultantes son:

$$F_s \;=\; X M u_s \qquad s \in \mathcal{S}$$

Su i-ésima coordenada adopta la expresión:

$$
\begin{array}{rcll}
F_s(i) & = & \sum_{j \in \mathcal{J}} \left(\frac{n\, f_{ij}}{f_{.j}} - 1 \right) f_{.j}\, u_{sj} = & \\
 & = & n \sum_{j \in \mathcal{J}} f_{ij}\, u_{sj} - \sum_{j \in \mathcal{J}} f_{.j}\, u_{sj} & i \in \mathcal{I}, \quad s \in \mathcal{S}
\end{array} \tag{10}
$$

Al diagonalizar la matriz XMX^TP cuyo término general es:

$$d_{ii'} = n\sum_{j\in\mathcal{J}} \frac{f_{ij}\, f_{i'j}}{f_{.j}} - f_{i'.} - f_{i.} + \frac{1}{n} \qquad i, i' \in \mathcal{I}$$

se obtienen los ejes $v_s, s \in \mathcal{S}$[1] que maximizan la inercia proyectada de la nube $\mathcal{N}(\mathcal{J})$ y tras premultiplicar dicha matriz por X^TP las proyecciones $G_s, s \in \mathcal{S}$ de dicha nube cuya j-ésima coordenada puede expresarse:

$$\begin{aligned} G_s(j) &= \sum_{i\in\mathcal{I}} \left(\frac{n\, f_{ij}}{f_{.j}} - 1\right) \frac{1}{n}\, v_{si} = \\ &= \sum_{i\in\mathcal{I}} \frac{f_{ij}}{f_{.j}}\, v_{si} - \frac{1}{n}\sum_{i\in\mathcal{I}} v_{si} \qquad j \in \mathcal{J}, \quad s \in \mathcal{S} \end{aligned} \tag{11}$$

8. RELACIONES ENTRE LOS FACTORES

Los factores de ambas nubes se relacionan a través de las expresiones:

$$F_s = \frac{1}{\sqrt{\lambda_s}}\, XMG_s \qquad s \in \mathcal{S} \tag{12}$$

$$G_s = \frac{1}{\sqrt{\lambda_s}}\, X^TPF_s \qquad s \in \mathcal{S} \tag{13}$$

En el análisis de correspondencias con marginal modificada, igual que ocurre en el análisis clásico para la cantidad $f_{i.}, i \in \mathcal{I}$, los factores $F_s, s \in \mathcal{S}$ están centrados para la cantidad $g_{i.}, i \in \mathcal{I}$:

$$\sum_{i\in\mathcal{I}} g_{i.}\, F_s(i) = 0 \qquad s \in \mathcal{S}$$

Aplicando la fórmula de transición (13):

$$\begin{aligned} G_s(j) &= \frac{1}{\sqrt{\lambda_s}} \sum_{i\in\mathcal{I}} \left(\frac{f_{ij}}{f_{.j}\, g_{i.}} - 1\right) g_{i.}\, F_s(i) = \\ &= \frac{1}{\sqrt{\lambda_s}} \sum_{i\in\mathcal{I}} \frac{f_{ij}}{f_{.j}}\, F_s(i) \qquad j \in \mathcal{J} \qquad s \in \mathcal{S} \end{aligned} \tag{14}$$

que coincide con la relación baricéntrica del análisis de correspondencias.

[1] Las relaciones de dualidad entre ambos espacios, que se verifican en todo análisis de correspondencias, permiten establecer que los subespacios de ajuste, asociados a valores propios no nulos, son de idéntica dimensión.

Sin embargo, a diferencia del análisis de correspondencias clásico los factores $G_s, s \in \mathcal{S}$ no están centrados por la cantidad $f_{.j}, j \in \mathcal{J}$ porque el análisis se hace tomando como origen un punto diferente al centro de gravedad.

$$\sum_{j\in\mathcal{J}} f_{.j}\, G_s(j) \quad = \quad \sum_{i\in\mathcal{I}} v_{si}\, f_{i.} - \sum_{i\in\mathcal{I}} g_{i.}\, v_{si} \qquad\qquad s \in \mathcal{S}$$

Si la marginal impuesta difiere de la marginal propia de la tabla esta cantidad es distinta de cero.

Por ello los factores $F_s, s \in \mathcal{S}$ no pueden interpretarse como el baricentro de los $G_s, s \in \mathcal{S}$ como en análisis clásico. Según la fórmula de transición (12):

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j\in\mathcal{J}} \left(\frac{f_{ij}}{f_{.j}\, g_{i.}} - 1 \right) f_{.j}\, G_s(j) =$$

$$= \frac{1}{\sqrt{\lambda_s}} \left\{ \sum_{j\in\mathcal{J}} \frac{f_{ij}}{g_{i.}}\, G_s(j) - \sum_{j\in\mathcal{J}} f_{.j}\, G_s(j) \right\} \qquad i \in \mathcal{I} \qquad s \in \mathcal{S} \tag{15}$$

El segundo sumatorio corresponde a la proyección del centro de gravedad de $\mathcal{N}(\mathcal{J})$, que al no haber sido tomado como origen de los ejes es diferente de 0.

Benali y Escofier (Benali & Escofier 1987) afirman que «Este término, en la práctica es casi nulo, lo que permite interpretar como en correspondencias múltiples clásico la abscisa de un individuo como el baricentro de las modalidades que ha elegido». Hacen referencia a tablas disyuntivas incompletas en las que el efectivo de datos ausentes representa una proporción reducida en relación al total (caso de datos ausentes por olvido distribuidos de forma aleatoria a lo largo de la tabla). Sin embargo, puede alterar los resultados y su interpretación cuando se considera nulo en una tabla de datos en la cual la proporción de no respuesta es grande o corresponde a ciertos grupos de individuos (caso de tablas de datos con preguntas condicionadas) como se ha podido comprobar en la aplicación a la Encuesta de Condiciones de Vida de 1989 de la Comunidad Autónoma de Euskadi presentada en (Goitisolo & Zárraga 1998a).

Lo cierto es que este segundo sumatorio es el mismo para todos los individuos, aunque difiere para los distintos ejes, por lo que se podría trasladar los factores $F_s(i), s \in \mathcal{S}, i \in \mathcal{I}$ de tal forma que en la representación superpuesta de ambas nubes un individuo siga estando representado en el baricentro de las modalidades que posee:

$$F_s^*(i) \quad = \quad \tfrac{1}{\sqrt{\lambda_s}}\, n \sum_{j\in\mathcal{J}} f_{ij}\, G_s(j) \qquad\qquad i \in \mathcal{I} \qquad s \in \mathcal{S} \tag{16}$$

9. NÚMERO DE EJES

En el análisis de correspondencias de una tabla disyuntiva completa el número de ejes S es igual al número de modalidades activas menos el número de variables, porque todas

las modalidades correspondientes a cada una de las variables se encuentran restringidas al mismo hiperplano. En el análisis de correspondencias con marginal modificada de una tabla disyuntiva incompleta, las modalidades de una misma cuestión no cumplen ningún tipo de restricción por lo que pueden existir tantos ejes como número de modalidades activas exista en el análisis. Si existen cuestiones con datos completos, sus modalidades mantendrán la misma restricción que en el análisis clásico por lo que la cantidad de ejes disminuirá en ese número de cuestiones completas.

La existencia de preguntas condicionadas en el análisis también reduce la cantidad de ejes, puesto que los individuos que han de responder a una pregunta condicionada vienen determinados por la respuesta a una modalidad (o combinación de ellas) anterior.

10. ESTUDIO DE LA ASOCIACIÓN ENTRE LAS CUESTIONES

El estudio de la independencia entre dos variables cualitativas q y $q' \in \mathcal{Q}$ con $\mathcal{J}_q$ y $\mathcal{J}_{q'}$ modalidades respectivamente se realiza habitualmente a través de su tabla de contingencia donde han de ser clasificados todos los individuos. Para ello, en el caso de que alguna de las cuestiones no haya sido respondida por todos los individuos, será necesario tener presente las modalidades de no respuesta. La tabla de contingencia tendrá la forma:

Tabla 2

	1	$\ldots$	j'	$\ldots$	$J_{q'}$	a'
1						
$\vdots$						
j			$b_{jj'}^{qq'}$			
$\vdots$						
J_q						
a						

donde a y a' representan las modalidades de no respuesta de las cuestiones q y $q' \in \mathcal{Q}$ respectivamente.

Un elemento $b_{jj'}^{qq'}; j \in \mathcal{J}_q; j' \in \mathcal{J}_{q'}; q, q' \in \mathcal{Q}$ de esta tabla indica el número de individuos que pertenecen simultáneamente a las modalidades j y j' de las cuestiones q y $q' \in \mathcal{Q}$ respectivamente. Es decir:

$$b_{jj'}^{qq'} = \sum_{i \in \mathcal{I}} z_{ij} z_{ij'} = \text{Card}\{i : z_{ij} = z_{ij'} = 1 | i \in \mathcal{I}\} \quad j \in \mathcal{J}_q \quad j' \in \mathcal{J}_{q'} \quad q, q' \in \mathcal{Q}$$

Analizar la independencia entre ambas cuestiones lleva a comparar los términos:

(17) $\frac{b_{jj'}^{qq'}}{n}$ y $\frac{b_{jj}^{qq}\, b_{j'j'}^{q'q'}}{n^2}$ $\quad j \in \mathcal{J}_q \quad j' \in \mathcal{J}_{q'} \quad q, q' \in \mathcal{Q}$

que, en función de la tabla disyuntiva incompleta, equivale a comparar:

(18)
$$\frac{\sum_{i\in\mathcal{I}} z_{ij} z_{ij'}}{n} \quad \text{y} \quad \frac{\sum_{i\in\mathcal{I}} z_{ij}^2 \sum_{i\in\mathcal{I}} z_{ij'}^2}{n^2} \qquad j \in \mathcal{J}_q \quad j' \in \mathcal{J}_{q'} \quad q, q' \in \mathcal{Q}$$

Sin embargo, la no respuesta en el caso de cuestionarios con preguntas condicionadas es determinada por una cuestión diferente. Si, por ejemplo, las dos cuestiones han sido condicionadas por la misma pregunta −de tal forma que ambas tienen en común el total de las respuestas ausentes− no existe independencia. Puesto que el interés se centra en la asociación entre las modalidades respondidas, parece más apropiado no tener en consideración la relación entre las modalidades de no respuesta a la hora de analizar la independencia entre ambas cuestiones.

En ACM clásico el análisis simultáneo de la dependencia entre las Q cuestiones se realiza a través de tabla de Burt definida en función de la tabla disyuntiva completa (Z) de la siguiente forma:

$$B = Z^T Z$$

La tabla B de término general $b_{jj'}^{qq'}$ está formada por Q^2 bloques. El bloque (q, q'), de orden $(J_q, J_{q'})$ es la tabla de contingencia que cruza las respuestas a las cuestiones q y $q' \in \mathcal{Q}$.

El elemento $b_{jj'}^{qq'},\ j, j' \in \mathcal{J}$ es nulo si las dos modalidades pertenecen a la misma cuestión y representa el número de individuos que eligen una determinada modalidad (denominado también por $z_{.j}$ o $z_{.j}^{q},\ j \in \mathcal{J}, q \in \mathcal{Q}$ -en §2-) si las modalidades j y $j' \in \mathcal{J}$ coinciden.

Por analogía se define la matriz B^* de orden $(J \times J)$ que será denominada pseudo-tabla de Burt y puede expresarse en función de la tabla disyuntiva incompleta:

$$B^* = {Z^*}^T Z^*$$

La diferencia entre B y B^* radica en que en la matriz B^* el total de cada una de las subtablas que la forman no es constante, sino la cantidad de individuos que responden a las cuestiones q y $q' \in \mathcal{Q}$ (no necesariamente coincidente con el número total de individuos encuestados).

En (Goitisolo & Zárraga 1998b) se demuestra la equivalencia entre los análisis factoriales de la tabla disyuntiva incompleta y la pseudo-tabla de Burt asociada, obteniéndose

factores proporcionales asociados a valores propios que se relacionan mediante:

$$\lambda_s^{B^*} = \left(\tfrac{z}{n}\lambda_s\right)^2 \qquad s \in \mathcal{S} \tag{19}$$

donde $\lambda_s^{B^*}$ son los valores propios correspondientes al análisis de la pseudo-tabla de Burt y λ_s los del análisis de la tabla disyuntiva incompleta.

10.1. Independencia entre las cuestiones

Al igual que en el análisis de correspondencias múltiples con datos completos, el análisis de la tabla disyuntiva incompleta está asociado a unas tasas de inercia pequeñas que no deben ser interpretadas como partes de información explicada por los ejes. A continuación se estudia el caso en el cual las cuestiones son independientes dos a dos y se propone una corrección a estas tasas de inercia en el análisis con marginal modificada de una tabla disyuntiva incompleta.

La matriz de diagonalización en el análisis con marginal modificada de la tabla disyuntiva incompleta, tiene un término general (recogido en la ecuación (9)) que puede ser expresado en función de los elementos de la tabla disyuntiva incompleta de la siguiente forma:

$$a_{jj'} = \tfrac{n}{z}\left(\tfrac{1}{z_{.j}}\textstyle\sum_{i\in\mathcal{I}} z_{ij}\, z_{ij'} - \tfrac{1}{n}\, z_{.j'}\right) \qquad j, j' \in \mathcal{J} \tag{20}$$

Notar que:

- $\displaystyle\sum_{i\in\mathcal{I}} z_{ij}\, z_{ij'} = 0$ si: $j, j' \in \mathcal{J}_q$ y $j \neq j'$ $q \in \mathcal{Q}$
  porque un individuo no puede pertenecer a dos modalidades de la misma cuestión
- $\displaystyle\sum_{i\in\mathcal{I}} z_{ij}\, z_{ij'} = z_{.j}$ si: $j, j' \in \mathcal{J}_q$ y $j = j'$ $q \in \mathcal{Q}$
  número de individuos que pertenecen a una determinada modalidad
- $\displaystyle\sum_{i\in\mathcal{I}} z_{ij}\, z_{ij'} = \frac{z_{.j}\, z_{.j'}}{n}$ si: $j \in \mathcal{J}_q\,,\quad j' \in \mathcal{J}_{q'}\,,\quad q \neq q'\quad q, q' \in \mathcal{Q}$ y además existe independencia entre ambas cuestiones

y que por tanto:

- $a_{jj'} = -\dfrac{z_{.j'}}{z}$ si: $j, j' \in \mathcal{J}_q$ y $j \neq j'$ $q \in \mathcal{Q}$
- $a_{jj'} = \dfrac{n - z_{.j'}}{z}$ si: $j, j' \in \mathcal{J}_q$ y $j = j'$ $q \in \mathcal{Q}$

- $a_{jj'} = 0$ si: $j \in \mathcal{J}_q\,, \quad j' \in \mathcal{J}_{q'}\,, \quad q \neq q' \quad q, q' \in \mathcal{Q}$ y existe independencia entre ambas cuestiones.

En consecuencia, la matriz A_{Z^*} que se ha de diagonalizar tiene la forma:

$$A_{Z^*} = \begin{bmatrix} A^1_{Z^*} & & & & 0 \\ & \ddots & & & \\ & & A^q_{Z^*} & & \\ & & & \ddots & \\ 0 & & & & A^Q_{Z^*} \end{bmatrix}$$

donde cada submatriz $A^q_{Z^*}$ es:

$$A^q_{Z^*} = \begin{bmatrix} \frac{n - z^q_{,1}}{z} & \frac{-z^q_{,2}}{z} & \frac{-z^q_{,3}}{z} & \cdots & \frac{-z^q_{.J_q}}{z} \\ \frac{-z^q_{,1}}{z} & \frac{n - z^q_{,2}}{z} & \frac{-z^q_{,3}}{z} & \cdots & \frac{-z^q_{.J_q}}{z} \\ \frac{-z^q_{,1}}{z} & \frac{-z^q_{,2}}{z} & \frac{n - z^q_{,3}}{z} & \cdots & \frac{-z^q_{.J_q}}{z} \\ & & & \ddots & \\ \frac{-z^q_{,1}}{z} & \frac{-z^q_{,2}}{z} & \frac{-z^q_{,3}}{z} & \cdots & \frac{n - z^q_{.J_q}}{z} \end{bmatrix}$$

Siendo $z^q_{.j}$ equivalente al $z_{.j}$ anterior.

Para diagonalizar la matriz A_{Z^*} se ha de resolver la ecuación:

$$|A_{Z^*} - \lambda_s I| \quad = \quad 0 \qquad s = 1, \ldots, J$$

Debido a la estructura diagonal por bloques de la matriz A_{Z^*}, los valores λ_s (incluyendo los valores nulos) que solucionan esa ecuación son los resultantes de resolver el sistema de Q ecuaciones siguiente:

$$|A^q_{Z^*} - \lambda^q_s I_{J_q}| \quad = \quad 0 \qquad\qquad \forall q \in \mathcal{Q} \qquad s = 1, \ldots, J_q$$

siendo I_{J_q} la matriz identidad de orden $J_q, q \in \mathcal{Q}$.

Las matrices $A^q_{Z^*}, q \in \mathcal{Q}$ pueden expresarse como:

$$A^q_{Z^*} \quad = \quad \tfrac{n}{z}\, I_{J_q} - E^q$$

siendo:

$$E^q = \begin{bmatrix} \frac{z^q_{,1}}{z} & \frac{z^q_{,2}}{z} & \frac{z^q_{,3}}{z} & \cdots & \frac{z^q_{.J_q}}{z} \\ \vdots & & & \ddots & \vdots \\ \frac{z^q_{,1}}{z} & \frac{z^q_{,2}}{z} & \frac{z^q_{,3}}{z} & \cdots & \frac{z^q_{.J_q}}{z} \end{bmatrix}$$

Como se puede comprobar fácilmente, esta matriz (de orden J_q) es de rango 1 y su traza es z_q/z (número de individuos que responden la cuestión $q \in \mathcal{Q}$ entre el número total de respuestas de los n individuos a las Q cuestiones), por ello tiene un valor propio $\mu^q = z_q/z$ y $(J_q - 1)$ valores propios nulos.

A través de la relación entre los valores propios de $A^q_{Z^*}$ y de esta matriz $E^q,\ q \in \mathcal{Q}$:

$$\begin{aligned} A^q_{Z^*}\, u_s &= \lambda^q_s\, u_s \\ \tfrac{n}{z}\, I_{J_q}\, u_s - E^q\, u_s &= \lambda^q_s\, u_s \\ E^q\, u_s &= \left(\tfrac{n}{z} - \lambda^q_s\right)\, u_s \\ E^q\, u_s &= \mu^q_s\, u_s \end{aligned}$$

donde $s = 1, \ldots, J_q$; se obtienen los valores propios de cada matriz $A^q_{Z^*}$:

$$\lambda^q_s = \begin{cases} \dfrac{n}{z} & \text{si } s = 1, \ldots, J_q - 1 \\ \dfrac{n}{z} - \dfrac{z_q}{z} & \text{si } s = J_q \end{cases} \qquad \forall q \in \mathcal{Q}$$

y de la matriz A_{Z^*} que resultan ser:

$$\lambda_s = \begin{cases} \dfrac{n}{z} & \text{si } s = 1, \ldots, J - Q \\ \dfrac{n}{z} - \dfrac{z_q}{z} & \text{si } s = (J - Q + 1), \ldots, J \end{cases} \qquad \forall q \in \mathcal{Q} \tag{21}$$

Donde existen tantos valores propios nulos como cuestiones a las que responden todos los individuos ($z_q = n,\ q \in \mathcal{Q}$).

Al igual que en el caso de tablas disyuntivas completas, la independencia entre cuestiones no se refleja en una inercia nula y no se debe, obviamente, a asociaciones entre las cuestiones sino a un efecto de estructura o de construcción de la tabla disyuntiva

incompleta. Cada uno de los J ejes estaría recogiendo una inercia trivial que evidentemente engorda los valores propios del análisis de la tabla disyuntiva incompleta cuando no existe la independencia.

10.2. Tasas de inercia

Cuando no existen datos ausentes, la aplicación del análisis de correspondencias simples a la tabla de Burt lleva a la obtención de los mismos factores del análisis de la tabla disyuntiva completa. En el caso particular de dos cuestiones, ambos análisis producen los mismos factores que el análisis de correspondencias simples de la tabla de contingencia. Sin embargo, los valores propios que se obtienen en los tres análisis son diferentes y dan lugar a distintas tasas de inercia proyectada sobre cada uno de los ejes.

Se pone de manifiesto de esta forma que el análisis de correspondencias múltiples (bien se realice a través de la tabla disyuntiva completa o bien a partir de la tabla de Burt) estudia las relaciones de dependencia entre cada par de cuestiones y se revela el escaso interés de los valores propios como medida de la información explicada por cada uno de los factores.

Benzécri (Benzécri 1979) propone por ello, basándose en la equivalencia entre el análisis de correspondencias de la tabla disyuntiva completa y de la tabla de contingencia cuando el número de cuestiones es dos, la siguiente corrección de los valores propios:

$$\lambda_s^* = \left(\frac{Q}{Q-1}\right)^2 \left(\lambda_s - \frac{1}{Q}\right)^2 \qquad s \in \mathcal{S}$$

para aquellos valores λ_s superiores a $1/Q$, siendo λ_s los valores propios resultantes del análisis de la tabla disyuntiva completa. Al estar los valores propios λ_s comprendidos entre 0 y 1, el término $\left(\frac{Q}{Q-1}\right)^2$ permite obtener unos valores propios corregidos comprendidos también entre 0 y 1, que hacen posible su comparación con otros análisis. Esta modificación de los valores propios lleva a definir las tasas de inercia proyectada:

$$\tau_s = \frac{\lambda_s^*}{\sum_s \lambda_s^*} \qquad s \in \mathcal{S}$$

donde la suma del denominador se extiende a los valores λ_s superiores a $1/Q$.

Esta corrección encuentra su justificación, para el caso de más de dos cuestiones, en la equivalencia entre los análisis de la tabla disyuntiva completa y de la tabla de Burt. En éste último análisis se introducen en la diagonal principal tablas que cruzan una cuestión consigo misma haciendo incrementar la inercia total y donde el resto de las tablas de contingencia aparecen dos veces (Greenacre 1993).

Una razón alternativa para calcular las tasas de inercia corregidas se encuentra en el estudio del caso particular donde las Q cuestiones son independientes dos a dos. A pesar del nulo interes del análisis de correspondencias clásico cuando se conoce (en ocasiones únicamente tras los resultados del análisis) la independencia de las cuestiones, su aplicación proporciona una inercia total no nula y $(J-Q)$ factores con valores propios asociados iguales a $1/Q$ (Zárraga 1989), demostrando que los valores propios del análisis (exista o no independencia) recogen una inercia trivial debida a un efecto de estructura o de construcción de la tabla disyuntiva completa.

Cuando existen datos ausentes, el análisis del caso en el cual las cuestiones son independientes, en la forma definida, revela que las tasas de inercia calculadas como los valores propios entre la inercia total, no son una buena medida de la asociación entre las cuestiones recogida por cada eje, por ello se propone, en el caso general en que las cuestiones no son independientes, calcular los valores propios y las tasas de inercia de la siguiente forma:

$$\lambda_s^* = \left(\frac{z}{z-n}\right)^2\left(\lambda_s - \frac{n}{z}\right)^2 \qquad \forall \lambda_s > \frac{n}{z} \qquad s = 1, \ldots, J$$

$$\tau_s^* = \frac{\lambda_s^*}{\sum_{\lambda_s^*>0} \lambda_s^*} \qquad s = 1, \ldots, J$$

donde λ_s es el valor propio obtenido en el análisis con marginal modificada de la tabla disyuntiva incompleta cuando no existe independencia. Estos nuevos valores propios y tasas de inercia coinciden, si la tabla es disyuntiva completa, en la que $z = nQ$, con los propuestos por Benzécri (1979).

REFERENCIAS

Abascal, E. & Grande, I. (1989). *Métodos multivariantes para la investigación comercial. Teoría, aplicaciones y programación BASIC*, Ariel economía.

Benali, H. (1985). *Stabilité de l' analyse en composantes principales et de l' analyse des correspondances multiples en présence de certains types de perturbations. Méthodes de dépouillement d' enquêtes. Thèse de troisième cycle*, Université de Rennes I.

Benali, H. (1988). «Données Manquantes et Modalités à Faible Effectif en Analyse des Correspondances Multiples et Conditionnelle», *Data Analysis and Informatics V*, 311-318.

Benali, H. & Escofier, B. (1987). «Stabilité de l' analyse factorielle des correspondances multiples en cas de données manquantes et de modalités à faibles effectifs». *Revue de Statistique Appliquée*, XXXV (1), 41-51.

Benzécri, J.P. (1979). «Sur le calcul des taux d' inertie dans l' analyse d' un questionnaire, addendum et erratum à», *Les Cahiers de l' Analyse des Données*, IV (3), 377-378.

Escofier, B. (1981). «Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte», *INRIA*.

Escofier, B. (1990). «Traitement des Variables Incomplètes en Analyse des Correspondances Multiples», *Revue de Modulad*, 5, 13-27.

Escofier, B. & Pagès, J. (1992). *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*, Servicio Editorial Universidad del País Vasco.

Goitisolo, B. & Zárraga, A. (1998a). «Application of the Incomplete Disjunctive Tables Study to the C.A.V. Living Conditions Survey, in *Analyses Multidimensionnelles des donnes. IV^{éme} Congrès International NGUS'*97, Fernandez-Aguirre, K. and Morineau, A., 301-313.

Goitisolo, B. & Zárraga, A. (1998b). «Equivalence between the Incomplete Disjunctive Table and the Associated Burt Pseudo-table Analysis», in *Analyses Multidimensionnelles des donnes. IV^{éme} Congrès International NGUS'*97, Fernandez-Aguirre, K. and Morineau, A., 227-238.

Greenacre, M.J. (1990). «Some Limitations of Multiple Correspondence Analysis», *Computational Statistics Quartely*, 3, 249-256.

Greenacre, M. (1993). *Correspondence Analysis in Practice*, Academic Press.

Lebart, L., Morineau, A. & Tabard, N. (1977). *Techniques de la description statistique*, Dunod.

Zárraga, A. (1989). *Análisis de correspondencias múltiples por bandas. Aplicación al estudio de una gran encuesta*. Tesis Doctoral, Universidad del País Vasco.

ENGLISH SUMMARY

INDEPENDENCE BETWEEN QUESTIONS IN THE FACTOR ANALYSIS OF INCOMPLETE DISJUNCTIVE TABLES WITH CONDITIONED QUESTIONS

A. ZÁRRAGA
B. GOITISOLO
Universidad del País Vasco-Euskal Herriko Unibertsitatea*

Multiple Correspondence Analysis (MCA) studies the relationship between several categorical variables defined with respect to a certain population. However, one of the main sources of information are those surveys in which it is usual to find a certain number of absent data and conditioned questions that do not need to be answered by the whole population. In these cases, the data codification in a complete disjunctive table requires the inclusion of non-answer categories that can alter the results.
Escofier (Escofie 1981) suggests the analysis of the incomplete disjunctive table (IDT) by substituting the real marginal of the table about the individuals for a constant imposed marginal. As in the analysis of multiple correspondences with complete data, the analysis of the incomplete disjunctive table is related to small percentages of inertia that cannot be regarded as a part of the information expounded by the axes. This paper will examine the case in which the questions are independent two by two and will propose the correction to these percentages of inertia in a modified marginal analysis of an incomplete disjunctive table.

Keywords: Multiple correspondence analysis, incomplete disjunctive table, independence between categorical variables, eigenvalues, percentages of inertia

AMS Classification (MSC 2000): 62H25

* Universidad del País Vasco-Euskal Herriko Unibertsitatea. Dpto. de Economía Aplicada III. Fac. CCEE y EE. Avda. Lehendakari Aguirre, 83. 48015 Bilbao. E-mails: az@alcib.bs.ehu.es y bg@alcib.bs.ehu.es.
This work was supported by Dirección General de Enseñanza Superior del Ministerio Español de Educación y Ciencia and Universidad del País Vasco (UPV/EHU) under research grants PB98-0149 and UPV 038.321-HA041/99

– Received February 1998.
– Accepted July 1999.

1. INTRODUCTION

In surveys, there can be absent answers for either unwilling slips by the surveyed, a will to hide certain information, the existence of conditioned questions answered by that person or not, depending on his answer to a previous question.

2. DEFINITION OF INCOMPLETE DISJUNCTIVE TABLES AND THEIR NOTATION

Table 1 gathers the answers by a certain colectivity of individuals $\mathcal{I}$ into a group of questions $\mathcal{Q}$, each of them with a finite group of answer categories $\mathcal{J}_q,\ q \in \mathcal{Q}$. Each element z_{ij} of the table assumes value 1 if individual i answers category j and 0 otherwise. It is said that it is an incomplete disjunctive table - I.D.T.- when for some i and $j\ \ z_{ij} = 0,\ \ i = \{1, \ldots, i, \ldots, n\}, j = \{1, \ldots, j, \ldots, J\}$.

A variable $z_{i.}^{q}$ is defined with value 1 if the individual i answers q and with value 0 otherwise, and z_q as the number of individuals responding to question $q,\ q \in \mathcal{Q}$.

3. PROBLEM BROUGHT ABOUT BY THE APPLICATION OF STANDARD M.C.A. TO AN I.D.T.

The distance χ^2 between two row profiles i and i' (equation 1) also increases with the common answers, when individuals i and i' do not answer the same number of questions.

In the distance χ^2 between two column profiles j and j' (equation 2) each member could have a different mass according to the number of answers previously chosen.

Therefore, the direct application of the standard M.C.A. is not appropriate to the study of an I.D.T.

4. POSSIBLE ALTERNATIVES TO THE I.D.T. PROBLEM: A FACTOR ANALYSIS WITH A MODIFIED MARGINAL

Either the elimination of those people not answering all the questions, or the creation for each absent data variable of a non-answer category would allow for a complete disjunctive table, but that could alter the results as well. As for the I.D.T. cases with a number of absent data, Escofier (1981) proposes the substitution of the real marginal of the table ($f_{i.}$ which is not constant) for a constant marginal $g_{i.} = 1/n$ in the whole analysis.

5. CLOUD OF INDIVIDUALS

Point i, provided with a weight $g_{i.}$, is represented in $\Re^J$ by the profile $f_{ij}/g_{i.}$. The distance between both i and i' (equation 3) increases only with different answers. This is equivalent to looking for the Euclidean distance in a space provided with a metric $1/f_{.j}$.

The weighting of each category is the inverse to its mass.

The j-th gravity centre coordinate of the cloud is $f_{.j}$. That will also originate the axes of maxima inertia to be searched for.

6. CLOUD OF CATEGORIES

Point j, provided with a weight $f_{.j}$, is represented in $\Re^n$ by the profile $f_{ij}/f_{.j}$.

The distance between categories j and j' (equation 4) is similar to that used when the table is complete. It is equivalent to the Euclidean distance in a space provided with a metric $1/g_{i.}$.

G_J, the i-th gravity centre coordinate of the cloud is: $f_{i.}$. However, (Goitisolo & Zárraga 1998*b*) justifies the election of point O_J (whose i-th coordinate is $g_{i.}$) as origin. The difference between the points O_J and G_J implies that the clouds must always be centered during the analysis.

7. COMPUTATION OF FACTORS IN BOTH CLOUDS

Calculating the succession of the axes (u_s) that maximize the inertia projected on the $\mathcal{N}(\mathcal{I})$ cloud is equivalent to maximizing the expression (5) with the matrices X, M, and P, whose general terms appear in the equations (6, 7 and 8).

The solution to this problem leads to the diagonalization of the matrix $X^T PXM$. The projection of the cloud of individuals on the maxima inertia resulting axes are: $F_s = XMu_s$. Its i-th coordinate assumes the expression (10).

Once the matrix $XMX^T P$ is diagonalized, the axes v_s that maximize the projected inertia about the cloud $\mathcal{N}(\mathcal{J})$ are obtained. And the projetions G_s, whose j-th coordinates are gathered in (11), are also obtained, just after premultiplication of the matrix by $X^T P$.

8. RELATIONSHIPS BETWEEN THE FACTORS

The factors of both clouds are related either with the expressions 12 and 13 or for every coordinate, according to the expressions 14 and 15.

The second summatory of equation 15 corresponds to the gravity center projection of $\mathcal{N}(\mathcal{J})$, which, not having been regarded as the origin of the axes, is different from 0.

Benali and Escofier (1987) suggest that «This term is nearly invalid in practice, which, like in a classic M.C.A., allows one to consider the abscissa of an individual to be the baricentre of the categories chosen». However, that can alter the results and its interpretation when regarded as invalid in a table of data, in which the non-answer proportion is in fact very large. That is exactly what happened in the application to the 1989 Life Conditions in the Basque Country Survey presented in (Goitisolo & Zárraga 1998*a*).

The truth is that this second summatory is similar to all the individuals for which the factors $F_s(i)$ could be shifted so that an individual could still be represented in the baricentre of all his categories during the superposed representation in both clouds (equation 16).

9. NUMBER OF AXES

The categories of the same question do not feature any kind of restriction and there can be as many axes as numbers of active categories in the analysis.

10. STUDY OF THE ASSOCIATION BETWEEN THE QUESTIONS

The study of the independence between two categorical variables q and $q'(q, q' \in \mathcal{Q})$ usually takes place with its contingency table (Table 2, where a and a' represent the non-answer categories of the questions q and q' respectively).

Analyzing the independence between both questions leads to comparing the terms in either equation 17 or 18.

When there are conditioned questions, the association between the categories answered has to be searched for.

In analogy with Burt's table, the matrix B^* is defined and will be named Burt's pseudo-table. It can be expressed according to the incomplete disjunctive table $B^* = {Z^*}^T Z^*$

(Goitisolo & Zárraga 1998*b*) shows the equivalence between the factor analysis Z^* and B^* and they obtain the proportional factors associated to eigenvalues related to equation 19, where $\lambda_s^{B^*}$ are the eigenvalues corresponding to Burt's pseudo-table analysis and λ_s those in the I.D.T.

10.1. Independence between the questions

The I.D.T. analysis is related to small percentages of inertia that cannot be regarded as parts of the information expounded by all the axes.

When there is independence between the questions, the diagonalization matrix in the analysis with an I.D.T. modified marginal does have a general term gathered in equation 9. Its expression according to I.D.T. elements appears in equation 20. Equation 21 gathers the eigenvalues of the analysis. There are as many invalid eigenvalues as questions answered by all the individuals.

10.2. Percentages of inertia

The relationship between the eigenvalues of the B^* and Z^* analyses, and the analysis of the case when the questions are independent two by two reveals that the percentages of inertia calculated as the ratio between the eigenvalues and the total inertia are not an appropriate measure of the association between the questions gathered by each axis. That is why we propose to calculate the eigenvalues and the percentages of inertia as follows:

$$\begin{aligned} \lambda_s^* &= \left(\frac{z}{z-n}\right)^2 \left(\lambda_s - \frac{n}{z}\right)^2 \qquad \forall \lambda_s > \frac{n}{z} \qquad s = 1, \ldots, J \\ \tau_s^* &= \frac{\lambda_s^*}{\sum_{\lambda_s^* > 0} \lambda_s^*} \qquad s = 1, \ldots, J \end{aligned}$$

λ_s is the eigenvalue attained in the analysis with a modified I.D.T. marginal, when there is no independence. These new eigenvalues and percentages of inertia do agree if the table is a complete disjunctive one, in which $z = nQ$, with those proposed by Benzécri (1979).