Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 273 a 293

Análisis de correspondencias a partir de una muestra probabilística

Analysis of Correspondence from a Probabilistic Sample

Javier Ramírez[a], Guillermo Martínez[b]

Departamento de Matemáticas y Estadística, Facultad de Ciencias Básicas e Ingenierías, Universidad de Córdoba, Montería, Colombia

Resumen

A partir del análisis de correspondencias clásico aplicado a las tablas denominadas de correspondencias, se desarrolla la teoría para dicho análisis a partir de una muestra probabilística. El enfoque de esta teoría se encamina a la estimación de los valores y vectores propios asociados a las matrices por diagonalizar, ya sea en el análisis simple o en el múltiple, para luego establecer las estimaciones de los vectores propios que conducen a los ejes factoriales, permitiéndose una representación gráfica para mejorar la interpretación en el análisis. Se realizan además estimaciones de las medidas de calidad asociadas a la representación, como son: inercia, contribuciones y cosenos cuadrados.

Palabras clave: análisis de correspondencias, muestreo probabilístico, *Bootstrap*, *Jackknife*.

Abstract

From the classic analysis of correspondences applied to the denominated tables of correspondences, the theory for this analysis from a probabilistic sample is developed. The approach of this theory directs to the estimation of eigenvalues and eigenvectors associated to the matrices to be diagonalized, either in a simple analysis or in the multiple one, to establish estimations of the eigenvectors that lead to the factorial axes, allowing a graphical representation to improve performance in the analysis. Estimates of quality measures associated to the representation are made, such as inertia, contributions and squares cosines.

Key words: Correspondence analysis, Probability sampling, *Bootstrap*, *Jackknife*.

[a]Profesor asistente. E-mail: javierramirez@sinu.unicordoba.edu.co
[b]Profesor asociado. E-mail: gmartinez@sinu.unicordoba.edu.co

1. Introducción

Una de las técnicas de los métodos factoriales que analiza la asociación entre dos o más variables categóricas es el denominado análisis de correspondencias. A través del análisis de correspondencias simples (ACS) aplicado a las tablas de contingencia, se construyen las representaciones de las asociaciones entre filas y columnas de estas tablas, basados en la distancia χ^2. Se trata de tablas de efectivos obtenidos cruzando las modalidades de dos variables cualitativas definidas sobre una misma población de n individuos Escofier & Pagés (1992). Por otra parte, con el análisis de correspondencias múltiples el cual es una extensión del dominio de aplicación del ACS, se describen grandes tablas de variables categóricas, representando las categorías de las variables como puntos en un espacio de pocas dimensiones Clausen (1998).

Ahora bien, un requisito fundamental para este tipo de análisis es la obtención de los valores y vectores propios, y por ende las coordenadas sobre los ejes factoriales que permiten la interpretación de las asociaciones entre las variables categóricas. En este trabajo se presenta una metodología de estimación de los valores y vectores propios de las matrices por diagonalizar en los análisis de correspondencias simples y múltiples, a partir de una muestra probabilística. Con ellos se obtienen los ejes, las coordenadas factoriales y las relaciones de transición entre los espacios, la estimación de la inercia, las contribuciones y los cosenos cuadrados. Lo que se tiene entonces es una complementación entre los diseños de muestreo probabilístico y el análisis de correspondencias, lo que permite describir no solo el comportamiento o la asociación entre variables categóricas obtenidas a través de una muestra probabilística tomada de alguna población bajo estudio, sino también inferir acerca de dicho comportamiento y el grado de asociación entre las variables de estudio, siguiendo la metodología dada por Martínez (1998).

En la sección 2 se presenta la propuesta de estimación de los elementos de base en el análisis de correspondencias simples y múltiples, al igual que las demás medidas que intervienen en el análisis; por otra parte, se propone el cálculo de la varianza de los valores propios estimados mediante las técnica *Jackknife* y *Bootstrap*, donde en la sección 3 se muestra un ejemplo de aplicación, en el que se comparan estas dos técnicas y se llega a discusiones importantes. Por último en la sección 4, se presentan los métodos computacionales utilizados, y en la sección 5 se dan a conocer las conclusiones del trabajo.

2. Resultados y discusión

2.1. Análisis general

El procedimiento para efectuar un análisis factorial para métricas y matrices de peso cualesquiera es diagonalizar la matriz $A = X'LXM$, donde M corresponde a la métrica y L a la matriz, de masa o peso, para encontrar los q valores propios más grandes de dicha matriz y a partir de estos obtener las coordenadas factoriales necesarias para llevar a cabo el análisis.

En general, si se desea efectuar un procedimiento de análisis factorial como el de correspondencias a partir de una muestra probabilística el interés se centra en la diagonalización de la matriz estimada, $\widehat{A} = X'\widehat{L}X\widehat{M}$ a partir del diseño muestral empleado, para así obtener los valores propios estimados de esta matriz (Lebart, Morineau & Piron 2000).

El polinomio característico dado por la ecuación (1)

$$|A - \lambda I| = 0 \tag{1}$$

resulta ser de la forma

$$p(\lambda) = (-1)^r \left(\lambda^r + b_{r-1}\lambda^{r-1} + \cdots + b_1\lambda + b_0\right) = 0 \tag{2}$$

el cual es posible estimarlo con la expresión

$$p\left(\widehat{\lambda}\right) = \left|\widehat{A} - \widehat{\lambda}I\right| = (-1)^r \left(\widehat{\lambda}^r + b_{r-1}\widehat{\lambda}^{r-1} + \cdots + b_1\widehat{\lambda} + b_0\right) = 0 \tag{3}$$

donde r es el número de modalidades de estudio y $b_{r-1}, \ldots, b_0$ son valores numéricos que se pueden escribir como funciones de totales poblacionales estimados. De esta manera se puede establecer que los valores estimados de λ son de la forma

$$\widehat{\lambda} = f\left(\widehat{t}_1, \widehat{t}_2, \ldots, \widehat{t}_r\right) \tag{4}$$

donde $\widehat{t}_i$ son los estimadores de los totales poblacionales que conforman la matriz A.

Así, obtenidos estos valores, es posible efectuar por completo el análisis factorial a partir de la información de la muestra, dado que con los valores y vectores propios estimados se pueden construir las demás componentes del análisis (Lebart et al. 2000).

2.2. Análisis de correspondencias simples a partir de una muestra probabilística

Dada la población $U = \{1, \ldots, N\}$, suponga que a los elementos de U se les miden dos variables, digamos Z_1 y Z_2 con p_1 y p_2 modalidades, respectivamente. La matriz de datos, resultado de la medición de las variables sobre los N individuos, es como sigue

$$Z = \left[Z_1 \vdots Z_2\right]$$

donde

$$Z_m = \begin{bmatrix} z_{11} & \cdots & z_{1i} & \cdots & z_{1p_m} \\ \vdots & & \vdots & & \vdots \\ z_{l1} & \cdots & z_{li} & \cdots & z_{lp_m} \\ \vdots & & \vdots & & \vdots \\ z_{N1} & \cdots & z_{Ni} & \cdots & z_{Np_m} \end{bmatrix}$$

con

$$z_{li} = \begin{cases} 1, & \text{si el sujeto } l \text{ seleccionó la modalidad } i \text{ de la pregunta } Z_m \\ 0, & \text{si el sujeto } l \text{ no seleccionó la modalidad } i \text{ de la pregunta } Z_m \end{cases}$$

así, para $m = 1, 2$, $z_{li} = 1$ ó $z_{li} = 0$ para $l = 1, 2, \ldots, N$ e $i = 1, 2, \ldots, p_m$.

2.2.1. Tabla de contingencia

La tabla de contingencia a partir de las matrices Z_1 y Z_2 es

$$\mathbf{C} = \mathbf{Z}_1^T \mathbf{Z}_2$$

es decir

$$\mathbf{C} = \begin{bmatrix} k_{11} & \ldots & k_{1j} & \ldots & k_{1p_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ k_{i1} & \ldots & k_{ij} & \ldots & k_{ip_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ k_{p_1 1} & \ldots & k_{p_1 j} & \ldots & k_{p_1 p_2} \end{bmatrix}$$

donde

$$k_{ij} = \sum_{l=1}^{N} z_{ijl} \tag{5}$$

corresponde a un total de un dominio, en este caso el total de individuos que respondieron la modalidad i de la pregunta Z_1 y la modalidad j de la pregunta Z_2 simultáneamente con:

$$z_{ijl} = z_{il} \times z_{jl} = \begin{cases} 1, & \text{si} \quad z_{il} = 1 \quad \text{y} \quad z_{jl} = 1 \\ 0, & \text{si} \quad z_{il} = 0 \quad \text{ó} \quad z_{jl} = 0 \end{cases}$$

2.2.2. Tabla de contingencia estimada

Basados en una muestra probabilística S obtenida a través del diseño $p(\cdot)$ (ver Särndal, Swensson & Wretman 1992), con probabilidades de inclusión π_l para los elementos de U, podemos estimar cada total en la ecuación (5), a través de un π estimador de la siguiente forma:

$$\widehat{k}_{ij\pi} = \sum_{l \in s} \frac{z_{ijl}}{\pi_l} \tag{6}$$

donde los π_l son las probabilidades de inclusión de cada individuo; así, la matriz de correspondencias estimadas es:

$$\widehat{C} = Z_n^T \Pi^{-1} Z_n \tag{7}$$

donde n corresponde a los individuos de la muestra, ahora la matriz de código binario asociado a las dos variables es

$$Z_n \begin{bmatrix} z_{11} & \cdots & z_{1i} & \cdots & z_{1p_m} \\ \vdots & & \vdots & & \vdots \\ z_{l1} & \cdots & z_{li} & \cdots & z_{lp_m} \\ \vdots & & \vdots & & \vdots \\ z_{n1} & \cdots & z_{ni} & \cdots & z_{np_m} \end{bmatrix}$$

para $m = 1, 2$ con matriz de probabilidades de inclusión

$$\Pi = \begin{bmatrix} \pi_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & \pi_l & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \pi_{n_s} \end{bmatrix}$$

entonces la matriz de correspondencias estimada (7), tendrá la siguiente forma

$$\widehat{C} = \begin{bmatrix} \widehat{k}_{11\pi} & \ldots & \widehat{k}_{1j\pi} & \ldots & \widehat{k}_{1p_2\pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{k}_{i1\pi} & \ldots & \widehat{k}_{ij\pi} & \ldots & \widehat{k}_{ip_2\pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{k}_{p_1 1\pi} & \ldots & \widehat{k}_{p_1 j\pi} & \ldots & \widehat{k}_{p_1 p_2\pi} \end{bmatrix}$$

con $\widehat{k}_{ij\pi}$ representando la estimación de cada total de la matriz de correspondencias, definida en la ecuación (6), donde

$$\widehat{k}_{\pi} = \sum_{l \in s} \frac{1}{\pi_l} = \widehat{N} \tag{8}$$

2.2.3. Criterio por maximizar y matriz por diagonalizar

En el espacio de las columnas $\mathbb{R}^{p_2}$, el interés es maximizar la suma ponderada de los cuadrados de las proyecciones sobre el eje, es decir, dada la muestra s, maximizar la ecuación

$$\text{Máx}_{\widehat{u}} \left\{ \sum_i \widehat{f}_i \widehat{d}^2(i, O) \right\} \tag{9}$$

lo que es equivalente a maximizar la expresión

$$\widehat{u}' \widehat{D}_{p2}^{-1} \widehat{F}' \widehat{D}_{p1}^{-1} \widehat{F} \widehat{D}_{p2}^{-1} \widehat{u} \tag{10}$$

con la restricción

$$\widehat{u}' \widehat{D}_{p2}^{-1} \widehat{u} = 1 \tag{11}$$

donde $\widehat{u}$ es el vector propio de la matriz estimada

$$\widehat{S} = \widehat{F}'\widehat{D}_{p1}^{-1}\widehat{F}\widehat{D}_{p2}^{-1} \tag{12}$$

asociado al valor propio estimado $\widehat{\lambda}$ más grande diferente de 1. La matriz $\widehat{F}$ corresponde a la matriz de frecuencias relativas estimadas de término general $\widehat{f}_{ij}$, es decir

$$\widehat{F} = \{\widehat{f}_{ij}\} \tag{13}$$

donde

$$\widehat{f}_{ij} = \frac{\widehat{k}_{ij\pi}}{\widehat{k}} \tag{14}$$

y las matrices de márgenes filas $\widehat{f}_{i\cdot}$ y columna $\widehat{f}_{\cdot j}$ están dadas por:

$$\widehat{D}_{p1} = \text{diag}\{\widehat{f}_{i\cdot}\} \tag{15}$$

para $i = 1, 2, \ldots, p_1$ y $j = 1, 2, \ldots, p_2$

$$\widehat{D}_{p2} = \text{diag}\{\widehat{f}_{\cdot j}\} \tag{16}$$

respectivamente, donde:

$$\widehat{f}_{i\cdot} = \sum_{j=1}^{p_2} \frac{\widehat{k}_{i\cdot\pi}}{\widehat{k}} \qquad \widehat{f}_{\cdot j} = \sum_{i=1}^{p_1} \frac{\widehat{k}_{\cdot j\pi}}{\widehat{k}}$$

con

$$\widehat{k}_{i\cdot\pi} = \sum_{j=1}^{p_2} \widehat{k}_{ij\pi} \qquad \widehat{k}_{\cdot j\pi} = \sum_{i=1}^{p_1} \widehat{k}_{ij\pi}$$

De esta forma se tiene en $\mathbb{R}^{p_2}$ que la métrica $\widehat{M}$ es $\widehat{D}_{p2}^{-1}$ y la matriz de pesos $\widehat{N}$ es $\widehat{D}_{p1}^{-1}$; así, la matriz por diagonalizar es:

$$\widehat{S} = \{\widehat{S}_{jj'}\} \tag{17}$$

de término general

$$\widehat{S}_{jj'} = \sum_{i=1}^{p1} \frac{\widehat{f}_{ij}}{\widehat{f}_{i\cdot}} \frac{\widehat{f}_{ij'}}{\widehat{f}_{\cdot j'}} = \sum_{i=1}^{p1} \frac{\widehat{k}_{ij}}{\widehat{k}_{i\cdot}} \frac{\widehat{k}_{ij'}}{\widehat{k}_{\cdot j'}} \quad \text{para} \quad j, j' = 1, 2, \ldots, p_2 \tag{18}$$

De la misma forma, en el espacio de las filas estimadas $\mathbb{R}^{p1}$, se busca maximizar la cantidad

$$\widehat{v}'\widehat{D}_{p1}^{-1}\widehat{F}\widehat{D}_{p2}^{-1}\widehat{F}'\widehat{D}_{p1}^{-1}\widehat{v} \tag{19}$$

con la restricción

$$\widehat{v}'\widehat{D}_{p1}^{-1}\widehat{v} = 1 \tag{20}$$

donde $\widehat{v}$ es el vector propio asociado a un valor propio de la siguiente matriz:

$$\widehat{T} = \widehat{F}\widehat{D}_{p2}^{-1}\widehat{F}'\widehat{D}_{p1}^{-1} \tag{21}$$

Así, de acuerdo con Lebart et al. (2000), la métrica $\widehat{M}$ en $\mathbb{R}^{p1}$ es $\widehat{D}_{p1}^{-1}$ y la matriz de pesos $\widehat{N}$ es $\widehat{D}_{p2}^{-1}$. Para la diagonalización de la matriz $\widehat{S}$ a partir de una muestra probabilística, definida en (12), podemos obtener un estimador de λ según la metodología definida en Särndal et al. (1992), sección 5. Entonces, siguiendo a Särndal et al. (1992), λ se puede escribir como una función de totales estimados de la forma

$$\widehat{\lambda} = f\left(\widehat{k}_{11\pi}, \widehat{k}_{12\pi}, \ldots, \widehat{k}_{1p2\pi}, \widehat{k}_{21\pi}, \widehat{k}_{22\pi}, \ldots, \widehat{k}_{2p2\pi}, \ldots, \widehat{k}_{p_1 1\pi}, \widehat{k}_{p_1 2\pi}, \ldots, \widehat{k}_{p_1 p_2 \pi}\right)$$

El estimador anterior se encuentra resolviendo el polinomio característico

$$\left|\widehat{S} - \widehat{\lambda} I_{p2}\right| = (-1)^{p2}\left(\widehat{\lambda}^{p2} + \widehat{b}_{p-1}\widehat{\lambda}^{p2-1} + \cdots + \widehat{b}_1\widehat{\lambda} + \widehat{b}_0\right) = 0 \tag{22}$$

donde $\widehat{k}_{ij\pi}$ es definido en la ecuación (6) y

$$\widehat{b}_r = f\left(\widehat{k}_{11}, \widehat{k}_{12}, \ldots, \widehat{k}_{1p2}, \widehat{k}_{21}, \widehat{k}_{22}, \ldots, \widehat{k}_{2p2}, \ldots, \widehat{k}_{p_1 1}, \widehat{k}_{p_1 2}, \ldots, \widehat{k}_{p_1 p_2}\right) \tag{23}$$

2.3. Estimación de la varianza (ACS)

De acuerdo con Särndal et al. (1992, cap. 5) el π estimador propuesto

$$\widehat{\lambda} = f\left(\widehat{k}_{11\pi}, \widehat{k}_{12\pi}, \ldots, \widehat{k}_{1p2\pi}, \widehat{k}_{21\pi}, \widehat{k}_{22\pi}, \ldots, \widehat{k}_{2p2\pi}, \ldots, \widehat{k}_{p_1 1\pi}, \widehat{k}_{p_1 2\pi}, \ldots, \widehat{k}_{p_1 p_2 \pi}\right)$$

es un estimador aproximadamente insesgado para λ.

Esto se puede demostrar a través de la técnica de linealización de primer orden de la serie de Taylor alrededor del punto k_{ij}, para $i = 1, 2, \ldots, p_1$ y $j = 1, 2, \ldots, p_2$; procediendo como en Martínez (1998), se tiene por aproximación de primer orden de Taylor una aproximación al estimador $\widehat{\lambda}$ por un pseudo estimador $\widehat{\lambda}_0$ de la forma

$$\widehat{\lambda} \doteq \widehat{\lambda}_0 = \lambda + \sum_{i,j} a_{ij}\left(\widehat{k}_{ij\pi} - k_{ij}\right) \tag{24}$$

donde

$$a_{ij} = \left.\frac{\partial f}{\partial \widehat{k}_{ij\pi}}\right|_{\left\{\widehat{k}_{ij\pi}\right\} = \{k_{ij}\}_{i=1,\ldots,p_1;\, i=1,\ldots,p_2}}$$

Como $\widehat{k}_{ij\pi}$ es el total de un dominio de estudio, entonces

$$E\left(\widehat{k}_{ij\pi}\right) = E\left(\sum_{l\in s}\frac{z_{ijl}}{\pi_l}\right) = \sum_{l\in U}\frac{z_{ijl}}{\pi_l}\pi_l = \sum_{l\in U} z_{ijl} = k_{ij} \tag{25}$$

luego, en vez de (24), se tiene que

$$\begin{aligned} E\left(\widehat{\lambda}\right) &\doteq E\left(\widehat{\lambda}_0\right) \\ &\doteq \lambda + \sum_{i,j} a_{ij}(E(\widehat{k}_{ij\pi}) - k_{ij}) \\ &\doteq \lambda + \sum_{i,j} a_{ij}(k_{ij} - k_{ij}) \\ &\doteq \lambda \end{aligned}$$

Por tanto, $E\left(\widehat{\lambda}\right) \approx E\left(\widehat{\lambda}\right)$.

Nuestro objetivo ahora es obtener una medida de la calidad de la estimación de λ. Entonces, siguiendo el método de funciones de totales dado en Särndal et al. (1992), definimos $u_l = \sum_{i,j} a_{ij} z_{ijl}$ y $\check{u}_l = u_l/\pi_l$ donde a_{ij} está dado en (24); así, la varianza y un estimador de la varianza Horvith-Thompson para funciones de totales es

$$AV\left(\widehat{\lambda}\right) = \sum_{l\in U}\sum_{l'\in U} \Delta_{ll'}\check{u}_l\check{u}_{l'} \tag{26}$$

Dado que las cantidades u_l dependen de valores desconocidos, el estimador de la aproximación de la varianza de Horvitz-Thompson $AV\left(\widehat{\lambda}\right)$ bajo la técnica es

$$\widehat{V}\left(\widehat{\lambda}\right) = \sum_{l\in s}\sum_{l'\in s} \check{\Delta}_{ll'}\frac{\widehat{u}_l}{\pi_l}\frac{\widehat{u}_{l'}}{\pi_{l'}} \tag{27}$$

donde

$$\widehat{u}_l = \sum_{i,j} \widehat{a}_{ij} z_{ijl}$$

y los coeficientes $\widehat{a}_{ij}$ se obtienen como

$$\widehat{a}_{ij} = \frac{\partial f}{\partial \widehat{k}_{ij\pi}}. \tag{28}$$

2.3.1. El estimador *Jackknife*

Dada la cantidad de parámetros por calcular para estimar la varianza de Horvitz-Thompson, se estudian los métodos de *Jackknife* y *Bootstrap* para la estimación de la varianza, los cuales son usados para este tipo de situaciones dada su simplicidad de cálculo y los supuestos para su aplicación; por tanto, la estimación para la varianza de $\widehat{\lambda}$, según el método *Jackknife* presentado en Wolter (1985), se define para este estimador como

$$v_{jk} = \frac{n-1}{n}\sum_{l=1}^{n}\left(\widehat{\lambda}_{n-1,l} - \frac{1}{n}\sum_{l'=1}^{n}\widehat{\lambda}_{n-1,l'}\right)^2 \tag{29}$$

Luego este estimador se conoce como el estimador *Jackknife* (delete-1) de $V\left(\widehat{\lambda}_n\right)$, donde

$$\widehat{\lambda}_{n-1,l} = f\left(\widehat{k}_{11\pi_{(n-1)}}, \widehat{k}_{12\pi_{(n-1)}}, \ldots, \widehat{k}_{p_1 1\pi_{(n-1)}}, \widehat{k}_{p_1 2\pi_{(n-1)}}, \ldots, \widehat{k}_{p_1 p_2 \pi_{(n-1)}}\right) \quad (30)$$

y

$$\widehat{k}_{ij\pi(n-1,l)} = \sum_{l' \in S-\{l\}} \frac{z_{ijl'}}{\pi_{l'}} \quad (31)$$

Es decir, $\widehat{\lambda}_{n-1,l}$ es el estimador del valor propio correspondiente, basado en la muestra de tamaño $n-1$ que resulta luego de eliminar el individuo l-ésimo de la muestra, y $\widehat{k}_{ij\pi(n-1,l)}$ es la estimación de un dominio eliminando la misma observación.

2.3.2. El estimador *Bootstrap*

Teniendo en cuenta la importancia de utilizar el método de remuestreo *Bootstrap* para estimar valores propios, Milan & Whittaker (1995) realizan una aplicación del *Bootstrap* paramétrico a modelos que incorporan valores singulares, donde se desarrollan discusiones importantes sobre el efecto de la variación de muestreo en las estimaciones.

Para calcular los valores propios mediante el método el remuestreo *Bootstrap*, se realizan los siguientes pasos:

1. Dada la muestra de tamaño n, calcular $\widehat{\lambda}$. La distribución de esta muestra se considera equivalente a la distribución de la población y $\widehat{\lambda}$ es el estimador muestral del parámetro poblacional λ.

2. Generar B muestras *Bootstrap* de tamaño n mediante muestreo con remplazo de la muestra original, y calcular los correspondientes valores $\widehat{\lambda}^{*1}, \widehat{\lambda}^{*2}, \ldots, \widehat{\lambda}^{*B}$ para cada una de las B muestras *Bootstrap*.

3. Estimar el error estándar del parámetro estimado $\widehat{\lambda}$ calculando la desviación estándar de las B réplicas *Bootstrap*.

Así, obtenemos que el error estándar es

$$\sigma^*_{\lambda} = \sqrt{\frac{\sum_{b=1}^{B}\left(\lambda^{b^*} - \bar{\widehat{\lambda}}^*\right)^2}{(B-1)}} = \sigma_{BOOT} \quad (32)$$

donde

$$\bar{\widehat{\lambda}}^* = \frac{1}{B}\sum_{b=1}^{B}\lambda^{b^*} \quad (33)$$

corresponde al promedio de los valores propios calculados en cada remuestra.

2.4. Estimación de las coordenadas factoriales

Siguiendo la metodología definida en Lebart et al. (2000), para análisis factorial ponderado, es decir el caso general, podemos obtener los cosenos cuadrados, inercia, para el análisis de correspondencias a partir de una muestra probabilística. Las coordenadas factoriales estimadas, analizando los perfiles fila $\frac{\widehat{f}_{ij}}{\widehat{f}_{i.}}$ en el espacio de las columnas $\mathbb{R}^{p2}$ y los perfiles columna $\frac{\widehat{f}_{ij}}{\widehat{f}_{.j}}$ en el espacio de las filas $\mathbb{R}^{p1}$ vendrán dados respectivamente por:

$$\widehat{\psi}_{\alpha} = \widehat{D}_{p1}^{-1}\widehat{F}\widehat{D}_{p2}^{-1}\widehat{u}_{\alpha} \tag{34}$$

$$\widehat{\varphi}_{\alpha} = \widehat{D}_{p2}^{-1}\widehat{F}'\widehat{D}_{p2}^{-1}\widehat{v}_{\alpha} \tag{35}$$

con términos generales

$$\widehat{\psi}_{\alpha i} = \sum_{j=1}^{p_2} \frac{\widehat{f}_{ij}}{\widehat{f}_{i.}\widehat{f}_{.j}}\widehat{\mathbf{u}}_{\alpha j} \tag{36}$$

$$\widehat{\varphi}_{\alpha j} = \sum_{i=1}^{p_1} \frac{\widehat{f}_{ij}}{\widehat{f}_{i.}\widehat{f}_{.j}}\widehat{\mathbf{v}}_{\alpha i} \tag{37}$$

respectivamente. Ahora se presentan las relaciones entre los espacios, fruto de la estimación de los vectores propios asociados a los valores propios estimados.

$$\widehat{\mathbf{v}}_{\alpha} = \frac{1}{\sqrt{\lambda_{\alpha}}}\widehat{\mathbf{F}}\widehat{\mathbf{D}}_{p_2}^{-1}\widehat{\mathbf{u}}_{\alpha}$$

$$\widehat{\mathbf{u}}_{\alpha} = \frac{1}{\sqrt{\lambda_{\alpha}}}\widehat{\mathbf{F}}'\widehat{\mathbf{D}}_{p_1}^{-1}\widehat{\mathbf{v}}_{\alpha}$$

2.4.1. Relaciones de transición

Las relaciones fundamentales existentes entre los puntos fila y puntos columna sobre el eje α son llamadas relaciones de transición, y se calculan así:

$$\widehat{\psi}_{\alpha i} = \frac{1}{\sqrt{\widehat{\lambda}_{\alpha}}}\sum_{j=1}^{p_2} \frac{\widehat{f}_{ij}}{\widehat{f}_{i.}}\widehat{\varphi}_{\alpha j} \tag{38}$$

$$\widehat{\varphi}_{\alpha j} = \frac{1}{\sqrt{\widehat{\lambda}_{\alpha}}}\sum_{i=1}^{p_1} \frac{\widehat{f}_{ij}}{\widehat{f}_{.j}}\widehat{\psi}_{\alpha i} \tag{39}$$

donde $\frac{1}{\sqrt{\widehat{\lambda}_{\alpha}}}$ es el coeficiente de dilatación estimado.

2.4.2. Estimación de la inercia, contribuciones y cosenos cuadrados

Siguiendo a Lebart et al. (2000), se presentan a continuación los estimadores muestrales de la inercia, las contribuciones a los ejes factoriales y los cosenos cuadrados estimados en el análisis de correspondencias.

Inercia. La inercia total estimada para el caso de análisis de correspondencias simple es

$$\widehat{I} = \sum_{\alpha=1}^{p-1} \widehat{\lambda}_\alpha \tag{40}$$

donde $\widehat{\lambda}_\alpha$ es valor propio estimado definido en la sección 2.3.

Contribuciones. La estimación de las contribuciones mediante los perfiles fila, es decir en el espacio $\mathbb{R}^{p_2}$, es:

$$\widehat{Cr}_\alpha(i) = \frac{\widehat{f}_{i\cdot}\widehat{\psi}_{\alpha i}^2}{\widehat{\lambda}_\alpha} = \frac{\widehat{k}\left(\sum_{j=1}^{p_2} \frac{\widehat{k}_{ij}}{\widehat{k}_{\cdot j}}\widehat{\mathbf{u}}_{\alpha j}\right)^2}{\widehat{k}_{i\cdot}\widehat{\lambda}_\alpha} \tag{41}$$

y la estimación para los perfiles columna en el espacio $\mathbb{R}^{p_1}$

$$\widehat{Cr}_\alpha(j) = \frac{\widehat{f}_{\cdot j}\widehat{\varphi}_{\alpha j}^2}{\widehat{\lambda}_\alpha} = \frac{\widehat{k}\left(\sum_{j=1}^{p_1} \frac{\widehat{k}_{ij}}{\widehat{k}_{i\cdot}}\widehat{\mathbf{v}}_{\alpha i}\right)^2}{\widehat{k}_{\cdot j}\widehat{\lambda}_\alpha} \tag{42}$$

Cosenos cuadrados. Los cosenos cuadrados estimados para los perfiles fila son:

$$\widehat{Cos}_\alpha^2(i) = \frac{\widehat{\psi}_{\alpha i}^2}{\widehat{d}^2(i,G)} = \frac{\widehat{k}^2\left(\sum_{j=1}^{p_2} \frac{\widehat{k}_{ij}}{\widehat{k}_{\cdot j}}\widehat{\mathbf{u}}_{\alpha j}\right)^2}{\widehat{k}_{i\cdot}^2\widehat{d}^2(i,G)} \tag{43}$$

y para los perfiles columna

$$\widehat{Cos}_\alpha^2(j) = \frac{\widehat{\varphi}_{\alpha j}^2}{\widehat{d}^2(j,G)} = \frac{\widehat{k}^2\left(\sum_{j=1}^{p_1} \frac{\widehat{k}_{ij}}{\widehat{k}_{i\cdot}}\widehat{\mathbf{v}}_{\alpha i}\right)}{\widehat{k}_{\cdot j}^2\widehat{d}^2(j,G)} \tag{44}$$

donde en los perfiles fila la distancia de un punto i al centro de gravedad tendrá la siguiente estimación

$$\widehat{d}^2(i,G) = \sum_{j=1}^{p_2} \frac{1}{\widehat{f}_{\cdot j}}\left(\frac{\widehat{f}_{ij}}{\widehat{f}_{i\cdot}} - \widehat{f}_{\cdot j}\right)^2 = \sum_{j=1}^{p_2} \frac{\widehat{k}}{\widehat{k}_{\cdot j}}\left(\frac{\widehat{k}_{ij}}{\widehat{k}_{i\cdot}} - \frac{\widehat{k}_{\cdot j}}{\widehat{k}}\right)^2 \tag{45}$$

y para los perfiles columna

$$\widehat{d}^2(j,G) = \sum_{i=1}^{p_1} \frac{1}{\widehat{f}_{i\cdot}}\left(\frac{\widehat{f}_{ij}}{\widehat{f}_{\cdot j}} - \widehat{f}_{i\cdot}\right)^2 = \sum_{i=1}^{p_1} \frac{\widehat{k}}{\widehat{k}_{i\cdot}}\left(\frac{\widehat{k}_{ij}}{\widehat{k}_{\cdot j}} - \frac{\widehat{k}_{i\cdot}}{\widehat{k}}\right)^2 \tag{46}$$

2.5. Análisis de correspondencia múltiple a partir de una muestra probabilística

Extendiendo el análisis de correspondencias a partir de una muestra probabilística del espacio $U = \{1, \ldots, N\}$ al caso de m variables con $p_1, \ldots, p_m$ modalidades respectivamente, se obtienen las expresiones a este caso más general a través de un diseño $p(\cdot)$; luego la matriz S definida en caso simple puede ser extendida al caso múltiple como:

$$S = F' D_N^{-1} F D_p^{-1} = \frac{1}{m} Z' Z D^{-1} = \frac{1}{m} B D^{-1} \tag{47}$$

donde

$$\mathbf{Z} = \begin{bmatrix} z_{11} & \ldots & z_{1j} & \ldots & z_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{l1} & \ldots & z_{lj} & \ldots & z_{lp} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{N1} & \ldots & z_{Nj} & \ldots & z_{Np} \end{bmatrix}$$

con

$$z_{lj} = \begin{cases} 1, & \text{si el sujeto } l \text{ seleccionó la modalidad } j \\ 0, & \text{si el sujeto } l \text{ no seleccionó la modalidad } j \end{cases}$$

Entonces $z_{lj} = 1$ ó $z_{lj} = 0$ con $l = 1, 2, \ldots, N$ y $j = 1, 2, \ldots, p$. La matriz $\mathbf{D}$ es diagonal de orden (p, p) obtenida a partir de la matriz de Burt, $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$, sin pérdida de generalidad, definimos $p = \sum_{j=1}^{m} p_j$. Así, la matriz de Burt está dada por:

$$B = Z'Z = \begin{bmatrix} \sum_{i=1}^N z_{i1} z_{i1} & \sum_{i=1}^N z_{i1} z_{i2} & \cdots & \sum_{i=1}^N z_{i1} z_{ip} \\ \sum_{i=1}^N z_{i2} z_{i1} & \sum_{i=1}^N z_{i2} z_{i2} & \cdots & \sum_{i=1}^N z_{i2} z_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^N z_{ip} z_{i1} & \sum_{i=1}^N z_{ip} z_{i2} & \cdots & \sum_{i=1}^N z_{ip} z_{ip} \end{bmatrix}$$

Entonces los elementos de S son de la forma

$$s_{jj'} = \frac{1}{m z_{.j'}} \sum_{i=1}^{N} z_{ij} z_{ij'} \tag{48}$$

donde

$$z_{.j'} = \sum_{i=1}^{N} z_{ij'}$$

Así, $s_{jj'}$ es una función de totales poblacionales de las variables $z_{j'}$ y $z_{jj'} = z_j z_{j'}$. Entonces el polinomio característico de S de acuerdo con la ecuación (1) viene dado por

$$p(\lambda) = \mid \mathbf{S} - \lambda \mathbf{I} \mid = (-1)^p \left(\lambda^p + b_{p-1} \lambda^{p-1} + \cdots + b_1 \lambda + b_0 \right) \tag{49}$$

donde cada b_j en el polinomio característico es función de los valores $s_{jj'}$ que a su vez son funciones de totales poblacionales, así

$$b_r = f\left(t_{z_1}, \ldots, t_{z_{j'}}, \ldots, t_{z_p}, \ldots, t_{z_{11}}, \ldots, t_{z_{jj'}}, \ldots, t_{z_{pp}}\right) \tag{50}$$

donde

$$t_{z_{j'}} = \sum_{i=1}^{N} z_{ij'}$$

y

$$t_{z_{jj'}} = \sum_{i=1}^{N} z_{ij} z_{ij'}$$

Por tanto se puede asumir que λ es también función de totales poblacionales

$$\lambda = f\left(t_{z_1}, \ldots, t_{z_{j'}}, \ldots, t_{z_p}, \ldots, t_{z_{11}}, \ldots, t_{z_{jj'}}, \ldots, t_{z_{pp}}\right) \tag{51}$$

Así, t_{zi} puede ser estimado a través de un π estimador basado en una muestra probabilística s de tamaño n, como sigue

$$\widehat{t}_{zj'} = \sum_{i \in s} \frac{z_{ij'}}{\pi_i}$$

y

$$\widehat{t}_{zjj'} = \sum_{i \in s} \frac{z_{ij} z_{ij'}}{\pi_i}$$

De esta manera, se puede establecer que un estimador aproximado para λ es de la forma

$$\widehat{\lambda} = f\left(\widehat{t}_{z_1}, \ldots, \widehat{t}_{z_{j'}}, \ldots, \widehat{t}_{z_p}, \ldots, \widehat{t}_{z_{11}}, \ldots, \widehat{t}_{z_{jj'}}, \ldots, \widehat{t}_{z_{pp}}\right) \tag{52}$$

resultado de resolver el polinomio característico estimado a partir de la muestra s de la matriz

$$\widehat{S} = Z_n' \Pi^{-1} Z_n \widehat{D}^{-1} = \widehat{B} \widehat{D}^{-1} \tag{53}$$

dado por

$$p(\lambda) = \left|\widehat{S} - \widehat{\lambda} I\right| = (-1)^p \left(\widehat{\lambda}^p + \widehat{b}_{p-1} \widehat{\lambda}^{p-1} + \cdots \widehat{b}_1 \widehat{\lambda} + \widehat{b}_0\right) \tag{54}$$

donde los b_r están definidos en la ecuación (50), la matriz de Burt estimada en la ecuación (53) es

$$\widehat{B} = Z_n' \Pi^{-1} Z_n \tag{55}$$

la matriz diagonal estimada obtenida a partir de la matriz de Burt, corresponde a una matriz de orden (p, p), dada por

$$\widehat{D} = \text{diag}\{Z_n' \Pi^{-1} Z_n\} \tag{56}$$

y Π es la matriz diagonal de probabilidades de inclusión

$$\Pi = \text{diag}\{\pi_1, \ldots, \pi_n\} \tag{57}$$

entonces

$$\mathbf{Z}_n = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

donde

$$\widehat{\mathbf{B}} = \mathbf{Z}_n' \Pi^{-1} \mathbf{Z}_n = \begin{bmatrix} \sum_{i=1}^n \frac{z_{i1} z_{i1}}{\pi_i} & \sum_{i=1}^n \frac{z_{i1} z_{i2}}{\pi_i} & \cdots & \sum_{i=1}^n \frac{z_{i1} z_{ip}}{\pi_i} \\ \sum_{i=1}^n \frac{z_{i2} z_{i1}}{\pi_i} & \sum_{i=1}^n \frac{z_{i2} z_{i2}}{\pi_i} & \cdots & \sum_{i=1}^n \frac{z_{i2} z_{ip}}{\pi_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \frac{z_{ip} z_{i1}}{\pi_i} & \sum_{i=1}^n \frac{z_{ip} z_{i2}}{\pi_i} & \cdots & \sum_{i=1}^n \frac{z_{ip} z_{ip}}{\pi_i} \end{bmatrix}$$

y la matriz diagonal estimada

$$\widehat{\mathbf{D}} = diag\left\{\mathbf{Z}_n' \Pi^{-1} \mathbf{Z}_n\right\} = \begin{bmatrix} \sum_{i=1}^n \frac{z_{i1}^2}{\pi_i} & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^n \frac{z_{i2}^2}{\pi_i} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{i=1}^n \frac{z_{ip}^2}{\pi_i} \end{bmatrix}$$

2.6. Estimación de la varianza (ACM)

El π estimador propuesto

$$\widehat{\lambda} = f\left(\widehat{t}_{z_1}, \ldots, \widehat{t}_{z_{j'}}, \ldots, \widehat{t}_{z_p}, \ldots, \widehat{t}_{z_{11}}, \ldots, \widehat{t}_{z_{jj'}}, \ldots, \widehat{t}_{z_{pp}}\right)$$

es un estimador aproximadamente insesgado para λ.

Esta prueba es análoga como en (27); entonces para el caso múltiple

$$\widehat{u}_i = \sum_r \widehat{a}_r z_{ri}$$

donde

$$\widehat{a}_r = \frac{\partial f}{\partial \widehat{t}_r}$$

Nuevamente las formas explícitas para $\widehat{u}_i$ son muy complicadas de calcular en la práctica; ende tanto se hace necesario establecer un estimador para la varianza del π estimador $\widehat{\lambda}$, a través del método *Jackknife* o *Bootstrap*, como sigue.

2.6.1. El estimador *Jackknife*

La expresión del estimador para la varianza es como en (29), donde

$$\widehat{\lambda}_{n-1,i} = f\left(\widehat{t}_{z_1(n-1,i)}, \ldots, \widehat{t}_{z_p(n-1,i)}, \ldots \widehat{t}_{z_{11}(n-1,i)}, \ldots, \widehat{t}_{z_{jj'}(n-1,i)} \ldots, \widehat{t}_{z_{pp}(n-1,i)}\right)$$

con

$$\widehat{t}_{zp\pi(n-1,i)} = \sum_{l \in S-\{i\}} \frac{z_{lj}}{\pi_l} \tag{58}$$

También es posible utilizar aproximaciones "apropiadas" de v_{JK} que requieran menos cálculos.

Shao & Tu (1995) introducen dos métodos computacionales para desarrollar el *Jackknife* delete-1 en la estimación de varianza, el método de agrupamiento y el método de submuestreo aleatorio; sin embargo en este trabajo se compara *Jackknife* con *Bootstrap*, mostrando que el remuestreo *Bootstrap* resulta ser más eficiente.

2.6.2. El estimador *Bootstrap*

El estimador de la varianza es similar como en la ecuación (32), donde se define la desviación *Bootstrap*; el procedimiento para la estimación es análogo.

2.7. Estimación de los elementos de base en el análisis de correspondencias múltiples

Construcción de las nubes. Suponga el conjunto de las coordenadas fila una nube de N puntos en el espacio de las p columnas. Cada punto i tiene por coordenadas en $\mathbb{R}^p$ $\left\{\frac{z_{ij}}{\widehat{z}_{i\cdot}}; j = 1, 2, \ldots, p\right\}$ y cada punto j tiene por coordenadas en $\mathbb{R}^N$ $\left\{\frac{z_{ij}}{\widehat{z}_{\cdot j}}; i = 1, 2, \ldots, N\right\}$

Selección de distancias. En $\mathbb{R}^N$ la distancia estimada χ^2 entre modalidades es

$$\widehat{d}^2(j, j') = \sum_{i=1}^{N} \widehat{N}\left(\frac{z_{ij}}{\widehat{z}_{\cdot j}} - \frac{z_{ij'}}{\widehat{z}_{\cdot j'}}\right)^2 \tag{59}$$

y en $\mathbb{R}^p$ la distancia estimada entre dos individuos es

$$\widehat{d}^2(i, i') = \frac{1}{m} \sum_{i=1}^{N} \frac{\widehat{N}}{\widehat{z}_{\cdot j}} \left(z_{ij} - z_{i'j}\right)^2 \tag{60}$$

con $\widehat{z}_{\cdot j} = \widehat{t}_{z_j} = \sum_{i \in s} \frac{z_{ij}}{\pi_i}$ y $\widehat{N} = \sum_{j=1}^{p} \sum_{i \in s} \frac{z_{ij}}{\pi_i}$

2.7.1. Coordenadas factoriales estimadas

Retomando los resultados del análisis de correspondencias se tiene

$$\begin{aligned}
\widehat{\mathbf{F}} &= \frac{1}{\widehat{N}m}\mathbf{Z} \text{ de término general } \widehat{f}_{ij} = \frac{z_{ij}}{\widehat{N}p} \\
\widehat{\mathbf{D}}_p &= \frac{1}{\widehat{N}m}\widehat{\mathbf{D}} \text{ de término general } \widehat{f}_{.j} = \delta_{ij}\frac{\widehat{z}_{.j}}{\widehat{N}m} \\
\widehat{\mathbf{D}}_N &= \frac{1}{\widehat{N}}\mathbf{I}_N \text{ de término general } \widehat{f}_{i.} = \frac{\delta_{ij}}{\widehat{N}}
\end{aligned}$$

donde $\mathbf{I}_{\widehat{N}}$ es la matriz identidad de orden $(\widehat{N}, \widehat{N})$ y $\delta_{ij} = 1$ si $i = j$ y cero si no.

Para encontrar los ejes factoriales $\widehat{u}_\alpha$ se diagonaliza la matriz

$$\widehat{\mathbf{S}}^+ = \frac{1}{m^2}\widehat{\mathbf{D}}^{-\mathbf{1}}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^{-\mathbf{1}}\widehat{\mathbf{B}}$$

donde $\widehat{\mathbf{D}}$ es la matriz diagonal de orden (p, p) de la matriz $\widehat{\mathbf{B}} = \mathbf{Z}'\mathbf{Z}$.

En $\mathbb{R}^p$, la ecuación del α-ésimo eje factorial estimado es

$$\frac{1}{m}\mathbf{Z}'_n\Pi^{-1}\mathbf{Z}_n\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{u}}_\alpha = \widehat{\lambda}_\alpha\widehat{\mathbf{u}}_\alpha \tag{61}$$

la ecuación del α-ésimo factor estimado es

$$\frac{1}{m}\widehat{\mathbf{D}}^{-1}\mathbf{Z}'_n\Pi^{-1}\mathbf{Z}_n\widehat{\varphi}_\alpha = \widehat{\lambda}_\alpha\widehat{\varphi}_\alpha \tag{62}$$

del mismo modo se escribe el α-ésimo factor estimado en $\mathbb{R}^N$ como

$$\frac{1}{m}\left[\mathbf{Z}_n\Pi^{-1}\widehat{\mathbf{D}}^{-1}\mathbf{Z}'_n\right]\widehat{\psi}_\alpha = \widehat{\lambda}_\alpha\widehat{\psi}_\alpha \tag{63}$$

Las coordenadas factoriales estimadas de un individuo i sobre el eje α están dadas por

$$\widehat{\psi}_{\alpha i} = \frac{1}{m\sqrt{\widehat{\lambda}_\alpha}} = \sum_{j\in p(i)}\widehat{\varphi}_{\alpha i} \tag{64}$$

$$\widehat{\varphi}_{\alpha i} = \frac{1}{\sqrt{\widehat{\lambda}_\alpha}} = \sum_{i\in I(j)}\widehat{\psi}_\alpha \tag{65}$$

donde $p(i)$ designa al conjunto de modalidades seleccionadas por el individuo i, y por otra parte $I(j)$ designa al conjunto de los individuos que seleccionaron la modalidad j en la muestra.

2.7.2. Relaciones de transición

Los factores estimados $\widehat{\varphi}_\alpha$ y $\widehat{\psi}_\alpha$ de norma $\widehat{\lambda}_\alpha$ representan las coordenadas estimadas de los puntos fila y los puntos columna sobre el eje factorial α; luego las

relaciones de transición vienen dadas por:

$$\widehat{\varphi}_\alpha = \frac{1}{\sqrt{\widehat{\lambda}_\alpha}} \widehat{\mathbf{D}}^{-1} \mathbf{Z}_n' \Pi^{-1} \widehat{\psi}_\alpha \tag{66}$$

y

$$\widehat{\psi}_\alpha = \frac{1}{m\sqrt{\widehat{\lambda}_\alpha}} \left[\mathbf{Z}_n \Pi^{-1}\right]' \widehat{\varphi}_\alpha \tag{67}$$

respectivamente.

2.7.3. Estimación de la inercia, contribuciones y cosenos cuadrados

Por otra parte, la estimación de los elementos de base en el análisis, como la inercia, las contribuciones y los cosenos cuadrados, se presenta a continuación:

Inercia. Para el caso del análisis de correspondencias multiple, la inercia $I(j)$ de una modalidad j se puede estimar a través de la expresión:

$$\widehat{I}(j) = \frac{1}{m}\left(1 - \frac{\sum_{i\in s} \frac{z_{ij}}{\pi_i}}{\widehat{N}}\right) \tag{68}$$

donde

$$\widehat{N} = \sum_{j_q=1}^{p_q} \sum_{i\in S} \frac{z_{ij_q}}{\pi_i}$$

Con z_{ij_q} variable dicótoma, así, es igual a uno, si el individuo i respondió la modalidad j de la pregunta q, y cero en otra modalidad de la misma pregunta. La inercia $I(q)$ de una pregunta q esta dada por:

$$\widehat{I}(q) = \sum_{j=1}^{p_q} \widehat{I}(j)$$

Entonces la inercia total estimada es:

$$I = \sum_q \widehat{I}(q)$$

Contribuciones. Las contribuciones estimadas en el análisis de correspondencias para los puntos fila ($\mathbb{R}^p$) y columna ($\mathbb{R}^n$), respectivamente, son:

$$\widehat{C}r_\alpha(i) = \frac{\widehat{\psi}_{\alpha i}^2}{\widehat{N}\widehat{\lambda}_\alpha} \quad y \quad \widehat{C}r_\alpha(j) = \frac{\widehat{z}_{\cdot j}\widehat{\varphi}_{\alpha j}^2}{\widehat{N}m\widehat{\lambda}_\alpha} \tag{69}$$

Cosenos cuadrados. Los cosenos cuadrados para los puntos fila ($\mathbb{R}^p$) y columna ($\mathbb{R}^n$) son:

$$\widehat{Cos}^2_\alpha(i) = \frac{\widehat{\psi}^2_{\alpha i}}{\widehat{d}^2(i,G)} \quad y \quad \widehat{Cos}^2_\alpha(j) = \frac{\widehat{\varphi}^2_{\alpha j}}{\widehat{d}^2(j,G)} \tag{70}$$

respectivamente, donde en los puntos fila la distancia de un punto i al centro de gravedad tendrá la siguiente estimación:

$$\widehat{d}^2(i,G) = \sum_{j=1}^{p_2} \frac{1}{\widehat{f}_{\cdot j}} \left(\frac{\widehat{f}_{ij}}{\widehat{f}_{i\cdot}} - \widehat{f}_{\cdot j} \right)^2 = \frac{\widehat{N}}{\widehat{z}_{\cdot j}} - 1 \tag{71}$$

3. Ejemplo de aplicación

Esta aplicación se refiere a las condiciones de vida de una población que se encuentra descrita por 150 UPM (unidades primarias de muestreo), las cuales se emplean para realizar un estudio por muestreo probabilístico a través de una muestra correspondiente al 30 % de las UPM, seleccionada bajo un diseño de muestreo bietápico con MAS en la primera etapa y MAS en la segunda etapa. En la siguiente tabla figuran las etiquetas de las modalidades de las cuatro preguntas:

Tabla 1: Descripción de variables.

Id	Preguntas	Modalidades
1	Se siente bien en el hogar	FA01. "Sí" FA02. "No"
2	Los gastos de vivienda son	DL01. "Despreciable" DL02. "Sin problema" DL03. "Gran carga" DL04. "Carga muy pesada"
3	Ha sufrido de dolor de espalda	MA01. "Sí" MA02. "No"
4	Se impone restricciones	RE01. 'Sí" RE02. "No"

El objetivo de este ejemplo es comparar los resultados de aplicar un análisis de correspondencias múltiple a la población descrita anteriormente y a una muestra aleatoria seleccionada de dicha población a través de un diseño de MAS-MAS. Se verifica luego que hay seis valores propios poblacionales y estimados no nulos ($6 = 10 - 4 = p - m$) que se muestran a continuación:

Tabla 2: Comparación de valores propios poblacionales y estimados.

Valores poblacionales				Valores estimados *Jackknife*				
No	V.P.	% Ine.	% Acum.	V.P.	% Ine.	% Acum.	ECM_{JK}	cve
1	0,11531	29,04	29,04	0,11927	30,14	30,14	0,0000417	0,0442
2	0,09634	24,26	53,31	0,08895	22,47	52,61	0,0001527	0,1028
3	0,06356	16,01	69,31	0,06331	16,00	68,61	0,0000071	0,0409
4	0,05298	13,34	82,66	0,05327	13,46	82,07	0,0000351	0,1114
5	0,04353	10,96	93,62	0,04402	11,12	93,19	0,0000142	0,0850
6	0,02532	6,38	100	0,02694	6,81	100	0,0000076	0,0869

En la tabla 2 notamos que el primer plano factorial poblacional, conformado por los dos primeros valores propios diferentes de cero, proporcionan un porcentaje

de inercia del 53,31 %, mientras que el primer plano factorial estimado presenta un porcentaje de variación del 52,61 % y coeficiente de variación del 10 %, lo cual es un buen indicio de estimación; además se puede observar la precisión en las estimaciones de manera puntual a cada uno de los valores propios, y finalmente es de resaltar la calidad de las estimaciones, a través de los coeficientes de variación, ya que resultan bastante efectivos.

Tabla 3: Comparación de valores propios poblacionales y estimados.

Valores poblacionales				Valores estimados *Bootstrap*				
No	V.P.	% Ine.	% Acum.	V.P.	% Ine.	% Acum.	ECM_{Boot}	cve
1	0,11531	29,04	29,04	0,11666	29,25	29,25	0,0000158	0,0442
2	0,09634	24,26	53,31	0,10082	25,27	54,52	0,0000421	0,1028
3	0,06356	16,01	69,31	0,06351	15,92	70,44	0,0000090	0,0409
4	0,05298	13,34	82,66	0,04982	12,49	82,93	0,0000480	0,1114
5	0,04353	10,96	93,62	0,04307	10,80	93,73	0,0000132	0,0850
6	0,02532	6,38	100	0,02501	6,27	100	0,0000061	0,0869

Se aprecia en la tabla 3 que utilizando el método *Bootstrap*, las estimaciones resultan ser un poco más eficientes que mediante el método *Jackknife* teniendo en cuenta la simulación realizada, en particular para los valores propios más grandes basándose en el error cuadrático medio y demás estimaciones. Cabe destacar que se realizaron 2000 simulaciones con $B = 500$ remuestras *Bootstrap*; estos valores se consideraron teniendo en cuenta que no se presentaron cambios significativos para un número mayor de simulaciones en las estimaciones.

Las estimaciones de los elementos de base en el análisis de correspondencias se presentan en el anexo.

4. Métodos computacionales

Una vez seleccionada la muestra probabilistica de la población de interés, y determinados los factores de expansión a partir del diseño de muestreo seleccionado, es posible aplicar algunas rutinas de programas estadísticos como SAS, SPAD o XLSTAT para estimar los elementos de base del análisis de correspondencias. Esto se logra incluyendo la variable "Peso" en el análisis correspondiente (en nuestro contexto, a los factores de expansión del diseño). Más explícitamente: mediante SAS, consiste en cargar el paquete PROC CORRESP con la opción Weigth p, donde p es la columna que contiene los factores de expansión del análisis por muestreo probabilístico. Mediante SPAD, consiste en establecer una columna en la base de datos de la muestra como los factores de expansión, la cual se colocará como "ponderación" en los parámetros del software. Mediante XLSTAT, consiste en insertar la matriz de Burt, Disyuntiva Completa, o Individuos Variable, en la opción que especifica la matriz con la cual se realizará el análisis, en nuestro caso la matriz de Burt, y después en la opción Peso se insertan los factores de expansión.

5. Conclusiones

- Dado que los valores propios por estimar se pueden escribir como funciones de las variables observadas en la muestra, como totales, razones y dominios poblacionales, es posible determinar medidas de calidad para las estimaciones. La varianza de estos estimadores se puede obtener a través de técnicas de aproximación de varianza, como es el caso de la técnica de linealización de primer orden de Taylor, y su estimación se puede realizar mediante métodos robustos como el *Jackknife* y *Bootstrap*, siendo este último más eficiente teniendo en cuenta la simulación realizada.

- Los diseños de muestreo probabilístico pueden ser utilizados para estimar los elementos de base en el análisis de correspondencias dando resultados confiables.

- Es posible utilizar diseños de muestreo más complejos, a través del muestreo en dos fases, utilizando como estimador de la varianza el estimador *Jackknife*, presentado por Pacheco & Martínez (2007).

Agradecimientos

Los autores del presente trabajo agradecen de manera muy especial a todas aquellas personas que contribuyeron en la elaboración y corrección del mismo. En particular agradece al Departamento de Matemáticas y Estadística de la Universidad de Córdoba.

[Recibido: abril de 2009 — Aceptado: octubre de 2010]

Referencias

Clausen, S. E., ed. (1998), *Applied Correspondence Analylis: An Introducction*, number 121 *in* 'Series on Quantitative Applications in the Social Sciences', Sage University Papers, Thousand Oaks, California.

Escofier, B. & Pagés, J. (1992), *Análisis factoriales simples y múltiples: objetivos, métodos e interpretación*, Universidad del País Vasco, Bilbao.

Lebart, L., Morineau, A. & Piron, M. (2000), *Statistique Exploratoire Multidimensionnelle*, Dunod, Francia.

Martínez, G. (1998), Estimación de los coeficientes de un análisis en componentes principales a partir de una muestra probabilística, Trabajo final, Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia.

Milan, L. & Whittaker, R. J. (1995), 'Application of the Parametric Bootstrap to Models that Incorporate a Singular Value Decomposition', *Applied Statistics* **44**(1), 31–49.

Pacheco, M. & Martínez, G. (2007), 'Un estimador jackknife de varianza en muestreo en dos fases con probabilidades desiguales', *Revista Colombiana de Estadística* **30**(2), 203–212.

Shao, J. & Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag, New York.

Särndal, C. E., Swensson, B. & Wretman, J. H. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

Apéndice A. Tablas

Tabla 4: Comparación de las coordenadas poblacionales y estimadas.

Modalidad	Coordenadas poblacionales					Coordenadas estimadas				
	1	2	3	4	5	1	2	3	4	5
FA01	−0,05	0,25	0,06	−0,15	−0,16	0,12	0,25	0,05	−0,11	−0,21
FA02	0,11	−0,54	−0,12	0,33	0,36	−0,22	−0,48	−0,09	0,21	0,40
DL01	−0,60	0,81	0,25	1,15	−0,07	−0,03	1,04	0,03	1,18	0,09
DL02	−0,30	−0,23	−0,06	−0,21	0,09	−0,37	−0,03	−0,04	−0,21	−0,02
DL03	0,62	0,19	−0,36	0,06	−0,20	0,65	−0,24	−0,42	0,05	−0,07
DL04	0,66	−0,11	1,71	−0,14	0,23	0,50	−0,21	1,53	−0,05	0,25
MA01	0,05	0,40	−0,07	−0,10	0,32	0,26	0,33	−0,08	−0,20	0,32
MA02	−0,04	−0,37	0,07	0,09	−0,29	−0,21	−0,26	0,07	0,16	−0,26
RE01	0,40	−0,02	0,01	0,02	0,01	0,33	−0,13	0,01	0,03	−0,04
RE02	−0,60	0,03	−0,01	−0,04	−0,02	−0,59	0,24	−0,02	−0,06	0,07

Tabla 5: Comparación de los cosenos cuadrados poblacionales y estimados.

Modalidad	Coordenadas poblacionales					Coordenadas estimadas				
	1	2	3	4	5	1	2	3	4	5
FA01	0,02	0,52	0,03	0,19	0,23	0,10	0,47	0,02	0,09	0,33
FA02	0,02	0,52	0,03	0,19	0,23	0,10	0,47	0,02	0,09	0,33
DL01	0,14	0,27	0,02	0,54	0,01	0,01	0,43	0,01	0,56	0,01
DL02	0,41	0,23	0,02	0,20	0,04	0,64	0,01	0,01	0,21	0,01
DL03	0,57	0,05	0,19	0,05	0,06	0,54	0,07	0,23	0,01	0,01
DL04	0,12	0,01	0,83	0,05	0,02	0,09	0,02	0,84	0,01	0,02
MA01	0,01	0,57	0,02	0,04	0,40	0,20	0,33	0,02	0,13	0,31
MA02	0,01	0,57	0,02	0,04	0,37	0,20	0,33	0,02	0,13	0,31
RE01	0,82	0,01	0,01	0,01	0,01	0,68	0,11	0,01	0,01	0,01
RE02	0,82	0,01	0,01	0,01	0,01	0,68	0,11	0,01	0,01	0,01