Revista de Matemática: Teoría y Aplicaciones 2000 **7**(1-2) : 77–90
cimpa – ucr – ccss issn: 1409-2433

# ANÁLISIS FACTORIAL DE CORRESPONDENCIAS INTRACLASES PONDERADO [*]

Belén Castro Iñigo[**] – Miguel Angel García Montoya [***]
Amaya Zárraga Castro [****]

*Recibido: 23 Mayo 2000*

**Resumen**

En este artículo se presenta una técnica de Análisis Factorial útil para el estudio de tablas de datos en las que existe una partición sobre el conjunto de individuos, definida de forma natural. Dicha técnica permite poner de manifiesto las relaciones existentes entre los individuos, pertenecientes a una misma clase, discriminándolos, a su vez, de los componentes de otras clases por los pesos relativos que se asignen a las mismas. Con una selección apropiada de la ponderación asignada a cada clase, la nueva metodología obtiene las variables locales del Análisis Parcial, contando con la ventaja sobre éste de proporcionar una representación de los individuos, como elementos activos, en los ejes locales.

**Palabras clave:** Análisis Factorial de Correspondencias condicionado a un grafo, Grafo de Partición, Inercia Intraclases, Análisis de Diferencias Locales.

**Abstract**

We present a Factor Analysis technique useful for the study of data tables which contain a partition of the individuals set, defined in a natural way. Such a technique allows to exhibit the relations that exist between individuals belonging to the same class, discriminating them from the members of other classes using the relative weights assigned to these classes. By an appropriate selection of the class weights, the new methodology obtains local variables of the Partial Analysis, with the advantage of a representation of the individuals, as active elements, of the local axes.

[*]Este trabajo está financiado por el proyecto UPV 036.351-HA062/99.

[**]Dpto. Economía Aplicada IV, Facultad de Ciencias Económicas y Empresariales, Universidad del País Vasco (Euskal Herriko Unibertsitatra), Avda. Lehendakari Aguirre 83, 48015 Bilbao, España. E-Mail: eupcainb@sdx01.bs.ehu.es

[***]Dpto. Economía Aplicada V, misma dirección. E-Mail: eupgamom@lg.ehu.es

[****]Dpto.Economía Aplicada III, misma dirección. E-Mail: az@alcib.bs.ehv.es

**Keywords:** Correspondence Factor Analysis conditionned by a graph, Graph of Partition, Intraclasses Inertia, Local Differences Analysis.

**Mathematics Subject Classification:** 62H25

# 1. Introducción

El presente trabajo se enmarca dentro de una línea de investigación en la que se pretende avanzar en el desarrollo de las técnicas factoriales descriptivas de análisis de datos condicionadas a una estructura de grafo. Más concretamente, en este artículo se presenta un nuevo modelo de ajuste al Análisis Factorial de Correspondencias [3], en sustitución del clásico modelo de independencia, cuyo objetivo es poder analizar las relaciones existentes en una tabla de datos entre un conjunto de individuos sobre los que existe definida una partición.

Este nuevo análisis, que denominamos Intraclases Ponderado, explica las relaciones existentes entre los individuos de una misma clase, asociando, además, una ponderación a cada uno de ellos, que permite representar el peso relativo que tiene la clase a la que pertenece sobre las demás.

Esta nueva metodología que aquí presentamos, tiene una estrecha relación con el Análisis Local o Parcial desarrollado por T. Aluja y L. Lebart [1] y el Análisis de Diferencias Locales de H. Benali y B. Escofier [2].

Cabe destacar la gran aplicabilidad del método propuesto a cualquier análisis de conjuntos de datos, que sean susceptibles de ser analizados mediante un Análisis Factorial de Correspondencias, en los que se deseen comprender las relaciones existentes entre grupos de individuos relacionados entre sí. Como ejemplo, dentro del campo de la Economía, podríamos nombrar el análisis de las tablas Input Output condicionado a la existencia de diversas clases basadas en las relaciones intersectoriales o, dentro del campo del Marketing, el posicionamiento de un producto teniendo en cuenta diversos nucleos de consumidores ligados bien geográficamente o bien por aspectos socio-económicos, políticos, lingüísticos, etc.

La estructura del artículo que presentamos es la siguiente, en la primera sección se muestra la generalización del Análisis Factorial de Correspondencias cuando los datos se ajustan a un modelo, diferente del de independencia, y se define el nuevo modelo propuesto para el Análisis Intraclases Ponderado junto con sus características más relevantes. En la siguiente sección se muestra tanto el Análisis de Diferencias Locales como el Análisis Local-Parcial y las relaciones que existen entre éstos y el método propuesto. Para finalizar el trabajo, se desarrolla un pequeño ejemplo de aplicación cuyo objetivo es servir de ayuda para comprender mejor los resultados obtenidos y las conclusiones más relevantes del mismo.

# 2. El análisis factorial intraclases ponderado

## 2.1. Análisis factorial de correspondencias ajustado a un modelo. Generalización del AFC

Se considera una tabla de datos $K_{IJ}$ formada por $n$ filas y $p$ columnas cuyo término general $k_{ij}$ representa el número de individuos de la población que poseen simultáneamente la modalidad $i$ de $I$ y $j$ de $J$, $i \in \{1, \cdots, n\}$, $j \in \{1, \cdots, p\}$.

Se denotará por $F_{IJ}$ a la tabla $\{f_{ij}, i \in I, j \in J\}$ con las frecuencias relativas de la tabla $K_{IJ}$. Si $\forall i \in I, f_{i\cdot} = \sum_{j \in J} f_{ij}$ y $\forall j \in J, f_{\cdot j} = \sum_{i \in I} f_{ij}$, en $\mathcal{R}^J$, el Análisis de Correspondencias es un análisis factorial sobre $\left\{ \frac{f_{ij}}{f_{i\cdot}}, i \in I, j \in J \right\}$, cuyas filas están afectadas por los pesos $\{f_{i\cdot}, i \in I\}$ y a la que se le aplica como métrica la métrica $\chi^2$ asociada a la matriz diagonal $\left\{ \frac{1}{f_{\cdot j}}, j \in J \right\}$. La distancia entre las diferentes filas de la tabla se mide a través de la distancia $\chi^2$. Mediante dicha distancia es posible evaluar las desviaciones entre la tabla de frecuencias $F_{IJ}$ y una tabla modelo, correspondiente a la hipótesis de independencia, cuyas frecuencias teóricas vienen dadas por el producto de las marginales $\{f_{i\cdot} f_{\cdot j}, i \in I, j \in J\}$.

Este análisis clásico se puede generalizar a un modelo diferente del de independencia bajo el supuesto de que las dos marginales de la nueva tabla sean iguales a los de la tabla estudiada. Los puntos resultantes en las nuevas nubes tienen por coordenadas las diferencias entre perfiles (fila o columna) de la tabla original y los de la tabla correspondiente al modelo estudiado.

Como el análisis de correspondencias está elaborado con referencia al modelo de independencia, es preciso, para poder hacer uso de su metodología en la nueva tabla creada, eliminar el modelo de independencia implícito en el mismo. Así, a partir de la tabla $F_{IJ}$ y del nuevo modelo que se desea ajustar, se puede elaborar la nueva tabla de datos de la siguiente manera:

$$\text{datos - modelo nuevo + modelo de independencia}$$

Si se denota por $m_{ij}$ el término general del modelo $M_{IJ,}$ la matriz que se analiza al ajustar los datos a un modelo es la que tiene por término general:

$$f_{ij} - m_{ij} + f_{i\cdot} f_{\cdot j}$$

Si se verifica que $\forall i \in I$, $m_{i\cdot} = f_{i\cdot}$ y $\forall j \in J, m_{\cdot j} = f_{\cdot j}$ con $m_{i\cdot} = \sum_{j \in J} m_{ij}$ y $m_{\cdot j} = \sum_{i \in I} m_{ij}$, los pesos y las métricas del análisis se mantienen respecto a los del análisis de la tabla inicial, aunque cabe la posibilidad de generalizar el resultado eliminando este supuesto.

En el caso particular del modelo de independencia se toma, $\forall i \in I$, $\forall j \in J$, $m_{ij} = f_{i\cdot} f_{\cdot j}$

## 2.2. Definición del modelo para el Análisis Intraclases Ponderado

Supongamos que sobre el conjunto de individuos $I$ se conoce, de forma natural, una partición $\{I_p, p \in P\}$ en $|P|$ clases ($|P|$ cardinal del conjunto $P$).

En este trabajo se construye un modelo nuevo para la tabla $F_{IJ}$, que tiene por objeto estudiar las relaciones existentes entre los individuos pertenecientes a una misma clase, a la vez que pone de manifiesto la importancia relativa de cada clase con respecto a las demás. El modelo que debe sustraerse de la tabla de datos deberá contener, tanto las relaciones entre las diferentes clases (relación inter clases) y que se representa mediante los baricentros de sus entornos, como las relaciones entre los individuos pertenecientes a una misma clase que quedan sin explicar por la ponderación asignada a su clase.

De esta forma, si se denota $\forall j \in J$, por $f_{pj} = \sum_{i \in I_p} f_{ij}$ y por $f_{p\cdot} = \sum_{i \in I_p} f_{i\cdot}$ $\forall p \in P$, la tabla de datos objeto de análisis será:

$$w_{ij} = f_{ij} - m_{ij} + f_{i\cdot} f_{\cdot j}$$

donde, si $i \in I_p$, y $h_p$ representa el peso asignado a la clase $p$:

$$m_{ij} = \frac{f_{pj}}{f_{p\cdot}} f_{i\cdot} + (1 - h_p) \left( f_{ij} - \frac{f_{pj}}{f_{p\cdot}} f_{i\cdot} \right)$$

La nueva tabla de datos, $W_{IJ}$, se construirá como:

$$w_{ij} = h_p (f_{ij} - \frac{f_{pj}}{f_{p\cdot}} f_{i\cdot}) + f_{i\cdot} f_{\cdot j}$$

Se denotará por $N_W(I)$ (respectivamente $N_W(J)$) a la nube de perfiles fila (respectivamente columna) de la tabla $W_{IJ}$.

### 2.2.1. Los centros de gravedad de las nubes $N_W(I)$ y $N_W(J)$

Si se denota, $\forall j \in J$, por $w_{\cdot j} = \sum_{i \in I} w_{ij}$ y $\forall i \in I$, por $w_{i\cdot} = \sum_{j \in J} w_{ij}$ las coordenadas de los centros de gravedad de las nubes transformadas $N_W(I)$ y $N_W(J)$, se verifica:

$$w_{i\cdot} = 2 f_{i\cdot} - m_i; \qquad w_{\cdot j} = 2 f_{\cdot j} - m_{\cdot j}$$

Ahora bien, el nuevo modelo definido cumple que, independientemente de los valores asignados a $\{h_p, p \in P\}$, sus perfiles son iguales a los de la tabla original por lo que se cumple que $\forall i \in I$, $w_{i\cdot} = f_{i\cdot} = m_{i\cdot}$ y $\forall j \in J, w_{\cdot j} = f_{\cdot j} = m_{\cdot j}$.

Si se considera $0 \leq h_p \leq 1$, el perfil de la fila $i$ de la tabla $m_{IJ}$ es una combinación lineal convexa del perfil de la fila $i$ de $F_{IJ}$ y del centro de gravedad de la clase a la que pertenece, $G_p$.

En $\mathcal{R}^J$, el análisis intraclase ponderado se define como un Análisis de Correspondencias sobre la tabla $\left\{ \frac{w_{ij}}{w_{i\cdot}}, i \in I, j \in J \right\}$ a la que se le asocian pesos $\{f_{i\cdot}, i \in I\}$ y métrica $\left\{ \frac{1}{f_{\cdot j}}, j \in J \right\}$.

### 2.2.2. La distancia $\chi^2$ en la nube $N_W(I)$

La distancia $\chi^2$ entre dos individuos de la nube $N_W(I)$ de perfiles fila de la tabla $W_{IJ}$ adquiere una interpretación importante si éstos pertenecen a la misma clase. En general,

para $i, l \in I$ :

$$d_w^2(i,l) = \sum_{j \in J} \frac{1}{w_{\cdot j}} \left( \frac{w_{ij}}{w_{i\cdot}} - \frac{w_{lj}}{w_{l\cdot}} \right)^2$$

Si $i, l \in I_p$ entonces:

$$\begin{aligned} d_w^2(i,l) &= \sum_{j \in J} \frac{1}{f_{\cdot j}} \left[ h_p \left( \frac{f_{ij}}{f_{i\cdot}} - \frac{f_{pj}}{f_{p\cdot}} \right) - h_p \left( \frac{f_{lj}}{f_{l\cdot}} - \frac{f_{pj}}{f_{p\cdot}} \right) \right]^2 = h_p^2 \sum_{j \in J} \frac{1}{f_{\cdot j}} \left( \frac{f_{ij}}{f_{i\cdot}} - \frac{f_{lj}}{f_{l\cdot}} \right)^2 = \\ &= r_p d_f^2(i,l) \end{aligned}$$

donde $r_p = h_p^2$ y $d_f^2(i,l)$ denota la distancia $\chi^2$ entre los individuos $i$ y $l$ en la nube original.

De la misma forma, la distancia $\chi^2$ de un individuo $i \in I$ al centro de gravedad de la nube $N_W(I)$, $G_I$, es:

$$d_w^2(i, G_I) = \sum_{j \in J} \frac{1}{w_{\cdot j}} \left( \frac{w_{ij}}{w_{i\cdot}} - w_{\cdot j} \right)^2$$

Si $i$ pertenece a la clase $p$ entonces:

$$d_w^2(i, G_I) = \sum_{j \in J} \frac{1}{f_{\cdot j}} \left[ h_p \left( \frac{f_{ij}}{f_{i\cdot}} - \frac{f_{pj}}{f_{p\cdot}} \right) \right]^2 = r_p d_f^2(i, G_p)$$

es decir, es la distancia del individuo $i$ al centro de gravedad de la clase a la que pertenece, ponderada por el cuadrado del peso de la clase (*distancia intraclase ponderada)*, en la nube $N(I)$.

### 2.2.3. Inercia de la nube $N_W(I)$

La inercia total estudiada en este nuevo análisis es:

$$V_T = Inercia\ Total = Inercia(N_W(I)) = \sum_{j \in J} V_j$$

con $V_j$ inercia de la variable $j$.

Teniendo en cuenta que se cumple que las marginales son idénticas $V_j$ adquiere la expresión:

$$\begin{aligned} V_j &= \frac{1}{f_{\cdot j}} \sum_{i \in I} f_{i\cdot} \left( \frac{w_{ij}}{w_{i\cdot}} - f_{\cdot j} \right)^2 = \frac{1}{f_{\cdot j}} \sum_{p \in P} \sum_{i \in I_p} f_{i\cdot} \left[ h_p \left( \frac{f_{ij}}{f_{i\cdot}} - \frac{f_{pj}}{f_{p\cdot}} \right) \right]^2 = \\ &= \sum_{p \in P} r_p \left[ \frac{1}{f_{\cdot j}} \sum_{i \in I_p} f_{i\cdot} \left( \frac{f_{ij}}{f_{i\cdot}} - \frac{f_{pj}}{f_{p\cdot}} \right)^2 \right] = \sum_{p \in P} r_p V I_j^p \end{aligned}$$

donde $VI_j^p$ representa la inercia intraclase de la variable $j$ en la nube $N(I)$. Esta inercia explica las diferencias de comportamiento en el interior de la clase, para la variable $j$.

# 3. Relación con otras técnicas de Análisis Factorial

En esta sección se efectúa una comparación de la técnica de análisis de datos expuesta con técnicas de Análisis de Correspondencias, que tienen en cuenta una estructura de grafo sobre los individuos estudiados.

## 3.1. Análisis Factorial de Correspondencias sobre Grafos de Partición

### 3.1.1. Análisis Factorial de Diferencias Locales

Se considera un grafo no dirigido, $G$, de orden $n$ cuyos vértices son el conjunto de individuos $I$ y cuyas $n_a$ aristas son un subconjunto $A \subset I \times I$.

El Análisis Factorial de Diferencias Locales [2] analiza las diferencias existentes entre los individuos, eliminando las tendencias generales que surgen de la contigüidad de los vértices del grafo. Para ello, compara el perfil de un individuo con el perfil medio de la vecindad a la que pertenece.

Este análisis es un análisis intraclase en el cual, se elimina de la tabla la frecuencia teórica en el caso en que un individuo se comportara como el baricentro de los de su vecindad. En definitiva, se elimina del análisis la dispersión interclase de las diferentes vecindades y se analiza la dispersión intraclase de las mismas.

Esta técnica supone que $G$ es un grafo valorado cuyos valores, $g_{ii'} \geq 0$, $(i, i') \in A$, cumplen la restricción $g_{i\cdot} = \sum_{i' \in I} g_{ii'} = cte; \forall i \in I$.

Si se denota por $\{l_{ij}, i \in I, j \in J\}$ las medias de cada vecindad ponderadas por el grafo, es decir, $l_{ij} = \sum_{i' \in I} g_{ii'} f_{i'j}$, la tabla, $W^e_{IJ}$, que se analiza es la que tiene por término general:

$$w^e_{ij} = f_{ij} - f_{i\cdot}\frac{l_{ij}}{l_{i\cdot}} + f_{i\cdot}\sum_{i \in I} f_{i\cdot}\frac{l_{ij}}{l_{i\cdot}}$$

con $l_{i\cdot} = \sum_{j \in J} l_{ij}$.

El Análisis de Diferencias Locales de la tabla original $F_{IJ}$ asociada al grafo $G$, se define como el Análisis de Correspondencias de la tabla $W^e_{IJ}$.

Si se considera una partición sobre el conjunto de los individuos, $\{I_p, p \in P\}$, y se definen para cada $I_p$ y cada $i \in I_p$ los valores de $G$ como:

$$g_{ii'} = \begin{cases} \frac{1}{|I_p|} & i' \in I_p \\ 0 & \text{en caso contrario} \end{cases}$$

entonces, si $i \in I_p$, $l_{ij}$ adquiere la expresión:

$$l_{ij} = \frac{f_{pj}}{|I_p|}, \quad \forall j \in J$$

y la tabla objeto de análisis coincide con la del intraclase:

$$w^e_{ij} = f_{ij} - f_{i\cdot}\frac{f_{pj}}{f_{p\cdot}} + f_{i\cdot}f_{\cdot j}$$

La inercia de cada variable $j, V_j^e$, explicada por este análisis es:

$$V_j^e = \frac{1}{f_{\cdot j}} \sum_{p \in P} \sum_{i \in I_p} f_{i\cdot} \left[ \frac{f_{ij}}{f_{i\cdot}} - \frac{f_{pj}}{f_{p\cdot}} \right]^2 = \sum_{p \in P} VI_j^p$$

### 3.1.2. Análisis Local-Parcial

Este análisis, al igual que el anterior, pretende explicar las relaciones entre los individuos y las variables basadas en la estructura definida por el grafo.

El Análisis Local-Parcial [1] aplicado a la estructura del grafo $G$, en $\mathcal{R}^J$, es el Análisis Factorial de la tabla $\left\{ \frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}}, (i,i') \in A, j \in J \right\}$ con peso $\left\{ \frac{f_{i\cdot} f_{i'\cdot}}{2 \sum_{(i,i') \in A} f_{i\cdot} f_{i'\cdot}}, (i,i') \in A \right\}$ y con métrica $\left\{ \frac{1}{f_{\cdot j}}, j \in J \right\}$.

En este análisis los individuos no son los vértices del grafo sino las aristas. Se denominan variables globales a las columnas de la matriz original de datos y variables locales a las columnas de la matriz analizada. En definitiva el análisis se realiza sobre las variables locales.

La inercia de la variable $j$, $V_j^l$, explicada por este análisis es:

$$V_j^l = \frac{1}{2m} \frac{1}{f_{\cdot j}} \sum_{(i,i') \in A} f_{i\cdot} f_{i'\cdot} \left( \frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

donde $m = \sum_{(i,i') \in A} f_{i\cdot} f_{i'\cdot}$.

Para el caso particular en el que $G$ sea un grafo de partición, la expresión anterior se puede escribir como:

$$V_j^l = \frac{1}{2m} \frac{1}{f_{\cdot j}} \sum_{p \in P} \sum_{i \in I_p} \sum_{i' \in I_p} f_{i\cdot} f_{i'\cdot} \left( \frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

siendo $m$ en este caso $\sum_{p \in P} f_{p\cdot}^2$.

En aquellos casos en los que el grafo recoja relaciones de similitud el análisis se denomina Análisis Parcial y si el mismo representa proximidades geográficas se denomina Análisis Local.

## 3.2. Relación con el análisis propuesto

Como paso previo a la comparación de las técnicas de Análisis de Correspondencias expuestas en la sección anterior con el Análisis Intraclases Ponderado, pasamos a realizar una comparación entre el Análisis Local-Parcial y el Análisis de Diferencias Locales para el caso particular de un grafo de partición.

Consideremos ambas técnicas sobre una estructura de grafo de partición $G$. Desarrollando, en este caso particular, las expresiones de $V_j^e$ y $V_j^l$ se obtiene:

$$V_j^e = \frac{1}{f_{\cdot j}} \sum_{p \in P} \left[ \sum_{i \in I_p} \frac{f_{ij}^2}{f_{i\cdot}} + \frac{f_{pj}^2}{f_{p\cdot}^2} \sum_{i \in I_p} f_{i\cdot} - 2 \frac{f_{pj}}{f_{p\cdot}} \sum_{i \in I_p} f_{ij} \right] = \frac{1}{f_{\cdot j}} \sum_{p \in P} \left[ \sum_{i \in I_p} \frac{f_{ij}^2}{f_{i\cdot}} - \frac{f_{pj}^2}{f_{p\cdot}} \right]$$

$$V_j^l = \frac{1}{m}\frac{1}{f_{.j}}\sum_{p\in P}\left[\sum_{i\in I_p}\frac{f_{ij}^2}{f_{i.}}f_{p.} - f_{pj}\sum_{i\in I_p}f_{ij}\right] = \frac{1}{mf_{.j}}\sum_{p\in P}f_{p.}\left[\sum_{i\in I_p}\frac{f_{ij}^2}{f_{i.}} - \frac{f_{pj}^2}{f_{p.}}\right]$$

De una comparación de las mismas se puede decir que el Análisis Local, en este caso particular, es un análisis intraclase en el que, a diferencia del anterior, cada una de las clases está ponderada por el peso que la misma tiene en el análisis de la nube global. Es decir, es un análisis en el que cada individuo se analiza respecto al centro de gravedad de la clase a la que pertenece, pero en el que se da más importacia a los individuos que se encuentran en clases con mayor peso.

Si en una tabla de datos la partición se realiza sobre clases homogéneas, es decir, $f_{p.} = \frac{1}{|P|}, \forall p \in P$, entonces $m = \sum_{p\in P} f_{p.}^2 = \frac{1}{|P|}$ y $V_j^l = V_j^e$ por lo que ambos análisis coinciden.

Sin embargo, es preciso tener en cuenta que en el Análisis Local no se obtiene una representación exacta de los individuos dado que la tabla de datos objeto de análisis cruza aristas con variables. Es decir, únicamente se pueden obtener los individuos representados como ilustrativos en los ejes obtenidos para las variables locales, lo que limita considerablemente la interpretación de las relaciones entre los mismos.

El Análisis de Correspondencias Intraclases Ponderado coincide con el de Diferencias Locales en el caso particular en el que $r_p = 1$, $\forall p \in P$, y con el Análisis Local si $\forall p \in P, \quad r_p = \frac{f_{p.}}{m}$. En este último caso $V_j = V_j^l, \forall j \in J$.

Por lo tanto, si $\forall p \in P, \quad r_p = \frac{f_{p.}}{m}$, el análisis de la tabla $W_{IJ}$ conduce al análisis de las variables locales, pero cuenta con la ventaja de obtener una representación de los individuos como activos dentro del análisis. De esta forma, los individuos pueden analizarse directamente y se dispone, no sólo de su calidad de representación, sino de su contribución a la formación de los ejes y a la inercia local.

En este caso en el que $\forall p \in P, \quad r_p = \frac{f_{p.}}{m}$ la diferencia entre el análisis propuesto y el Análisis de Diferencias Locales radica en la matización de los individuos de acuerdo al peso relativo de la clase a la que pertenecen. Dicha diferencia será más notable cuanto mayor sea el peso de la clase a la que pertenece el individuo en el análisis global.

La mayor diferencia en la contribución a la inercia de una variable en una clase $p$, en los Análisis de Diferencias Locales y el propuesto se obtiene cuando la tabla cumple que $r_p = \frac{1+\sqrt{|P|}}{2}$ y $r_q = \frac{1}{2} \quad \forall q \neq p; q \in P$.

Para llegar a este resultado supongamos que , $\forall p \in P, \quad r_p = \frac{f_{p.}}{m}$ y que deseamos dar el máximo peso a la clase 1. Sea el peso de dicha clase $f_{p_1\cdot} = n_q f_{q\cdot}$con $n_q \geq 0$, $\forall q \in P$ y $n_1 = 1$. Entonces, como se verifica que $\sum_{p\in P} f_{p.} = 1$ se llega a:

$$r_1 = \frac{\sum_{p\in P}\frac{1}{n_p}}{\sum_{p\in P}\left(\frac{1}{n_p}\right)^2}$$

que tiene su valor máximo $r_1 = \frac{\rho}{2}$ en $n_q = \rho, \forall q \neq 1, q \in P$ con $\rho$ raíz de $Z^2 - 2Z - (|P| - 1)$.

Del resultado anterior se deduce que las máximas diferencias en la contribución a la inercia de una clase $p$ entre el Análisis de Diferencias Locales y el Análisis Intraclases

Ponderado con $r_p = \frac{f_{p.}}{m}$, $\forall p \in P$, se dan en el caso en el que la tabla $F_{IJ}$ cumpla $f_{p.} = \frac{1}{\sqrt{|P|}}$ y $f_{q.} = \frac{1}{\sqrt{|P|}\left(1+\sqrt{|P|}\right)}$ $\quad \forall q \neq p, q \in P$.

### 3.3. Descomposición de la inercia total

Siguiendo la fórmula de *Huygens*, la inercia total de la nube se puede descomponer en inercia interclases, que explica la variabilidad entre las diferentes clases e inercia intraclases que explica la existente dentro de cada clase.

Es decir, $\forall j \in J$ :

$$V_j^T = Vi_j + \sum_{p \in P} VI_j^p$$

donde con $Vi_j$ representamos la inercia inter de la variable $j$.

Cuando se realiza el análisis intraclases (Análisis de Diferencias Locales sobre un grafo de partición) se analizan los datos $w_{ij}^e$ que permiten explicar el segundo término de la inercia. Si se analizan $l_{ij} = f_{i.}\frac{f_{pj}}{f_{p.}}$ se explica la inercia inter de la variable (Análisis Alisado [2]).

Desde el punto de vista del Análisis Parcial sobre un grafo no valorado, $G$, la inercia de la variable $j$ se descompone en:

$$V_j^T = V_j^l m^l + V_j^d m^d$$

es decir, inercia local descrita por las diferencias recogidas por el grafo

$$V_j^l m^l = \frac{1}{2f_{.j}} \sum_{(i,i') \in A} f_{i.} f_{i'.} \left( \frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

e inercia diferida que describe las diferencias entre los individuos no unidos por aristas del grafo:

$$V_j^d m^d = \frac{1}{2f_{.j}} \sum_{(i,i') \notin A} f_{i.} f_{i'.} \left( \frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

$m^l = \sum_{(i,i') \in A} f_{i.} f_{i'.}$ y $m^d = \sum_{(i,i') \notin A} f_{i.} f_{i'.}$.

Desde el punto de vista del Análisis Intraclase Ponderado la inercia total de una variable $j$ se descompone en:

$$V_j^T = V_j + V_j^m + 2 \sum_{p \in P} h_p (1 - h_p) VI_j^p$$

donde $V_j^m$ representa la inercia de la variable $j$ explicada por el modelo definido en el análisis. En el caso particular en el que $r_p = f_{p.}$, $V_j$ coincide con la inercia local $V_j^l m^l$, mientras que:

$$V_j^d m^d = V_j^m + 2 \sum_{p \in P} h_p (1 - h_p) VI_j^p$$

El segundo término de la expresión explica la covariación entre las dos partes que resumen la inercia intraclase debidas a la ponderación.

Ahora bien, es posible ajustar los datos a un nuevo modelo:

$$m'_{ij} = \sqrt{1-r_p}\left(f_{ij} - \frac{f_{pj}}{f_{p\cdot}}f_{i\cdot}\right) + \frac{f_{pj}}{f_{p\cdot}}f_{i\cdot} \quad \forall i \in I_p, j \in J$$

de cuyo análisis se obtiene la parte de inercia total que deja sin explicar $W_{IJ}$. En relación con el Análisis Local la inercia explicada por $m'_{ij}$, $V_j^{m'}$ sería la inercia diferida del grafo y en relación al Análisis de Diferencias Locales los datos del modelo explicarían la inercia interclases y la parte de inercia intra que deja sin explicar la ponderación:

$$V_j^{m'} = V_j^d m^d = Vi_j + \sum_{p\in P}(1-r_p)VI_j^p$$

## 4. Ejemplo de ilustración

Como ejemplo de aplicación y teniendo como objetivo el poner de relieve los distintos resultados presentados en las secciones anteriores, se ha simulado una pequeña matriz de datos de dimensiones $(9 \times 4)$. Para ello hemos utilizado un algoritmo Montecarlo que genera tablas de contingencia uniformes con marginales fijas [6]. A los nueve individuos o filas que la componen los hemos identificado por **i1**, **i2**,...., **i9** y a las cuatro variables las que hemos denominado **v1**, **v2**, **v3** y **v4.**. Además, se ha particionado el conjunto de individuos en tres clases tales que la primera de ellas está formada por los cuatro primeros individuos, la segunda por el quinto, sexto y séptimo y la tercera por los individuos ocho y nueve.

Con el objetivo de presentar la diferencia entre las tres metodologías, se aplica el Análisis Intraclases Ponderado al caso en el que la inercia intraclase $p$ aparece ponderada por $r_p = \frac{f_{p\cdot}}{m}$, $\forall p \in P$. En este caso particular los tres análisis son comparables.

El grafo $G$ construido para la aplicación del Análisis Parcial es no valorado y sus aristas relacionan a los individuos pertenecientes a una misma clase, es decir es un grafo de partición. Los valores del grafo asignados para el Análisis de Diferencias Locales son los ya comentados en las secciones precedentes.

En base a todo lo expuesto, se han simulado los datos obteniendo la siguiente matriz de frecuencias relativas:

| | v1 | v2 | v3 | v4 | $f_{i\cdot}$ |
|---|---|---|---|---|---|
| i1 | 0.0010 | 0.1589 | 0.1567 | 0.0298 | 0.3464 |
| i2 | 0.0450 | 0.0125 | 0.0096 | 0.0484 | 0.1155 |
| i3 | 0.0039 | 0.0017 | 0.0306 | 0.0215 | 0.0577 |
| i4 | 0.0001 | 0.0003 | 0.0409 | 0.0164 | 0.0577 |
| i5 | 0.0200 | 0.0467 | 0.0097 | 0.0504 | 0.1268 |
| i6 | 0.0500 | 0.0020 | 0.0007 | 0.0001 | 0.0528 |
| i7 | 0.0300 | 0.0013 | 0.0003 | 0.0001 | 0.0317 |
| i8 | 0.0600 | 0.0238 | 0.0005 | 0.0426 | 0.1268 |
| i9 | 0.0800 | 0.0028 | 0.0011 | 0.0006 | 0.0845 |
| $f_{\cdot j}$ | 0.29 | 0.25 | 0.25 | 0.21 | 1 |

$$\begin{cases} f_{p_1\cdot} = 0{,}5774 & r_1 = 1{,}3666 \\ f_{p_2\cdot} = 0{,}2113 & r_2 = 0{,}5 \\ f_{p_3\cdot} = 0{,}2113 & r_3 = 0{,}5 \end{cases}$$

Posteriormente se ha transformado esta matriz obteniéndose tanto los valores correspondientes para realizar el Análisis de Diferencias Locales como aquéllos que hacen posible aplicar la nueva metodología expuesta en este artículo. Con todo ello, se ha realizado el Análisis de Correspondencias sobre las tres matrices de datos así como el de la matriz original. Los resultados de los tres análisis se han superpuesto en un mismo plano factorial con el único objetivo de que resulte más sencilla su comparación, mostrándose los resultados de las distintas proyecciones, tanto de los individuos como de las variables, en el gráfico adjunto (ejes factoriales uno y dos).

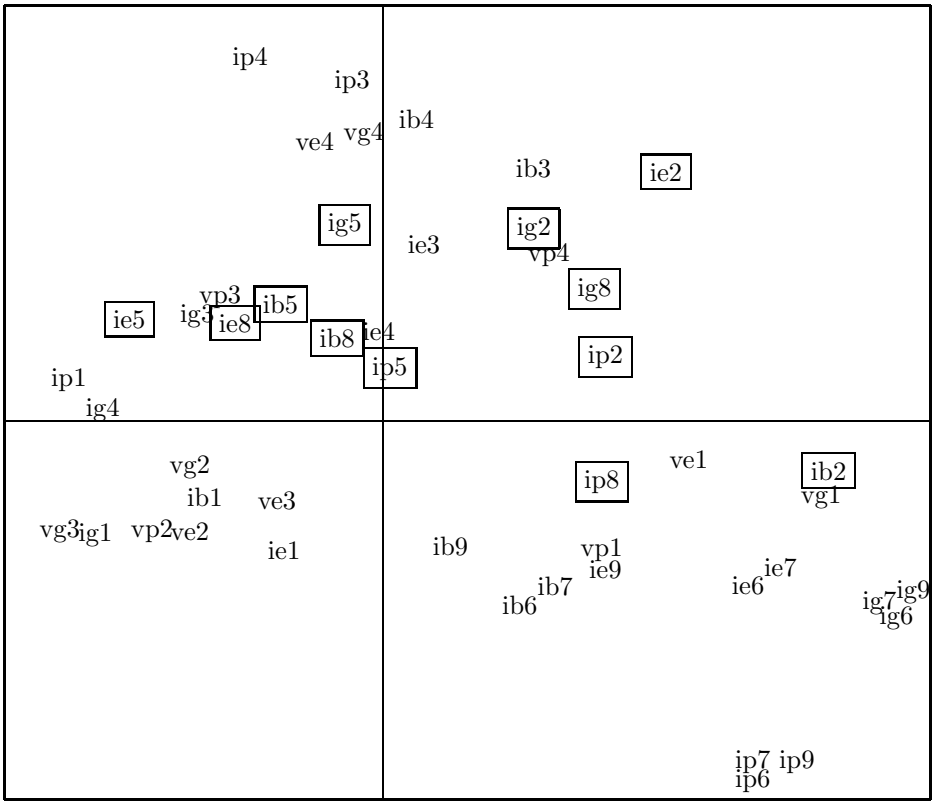


Figura 1: Superposición de los planos 1-2 del Análisis Global (*vg,ig*), Local-Parcial (*vp,ip*), de Diferencias Locales (*vd,id*), e Intraclases Ponderado (*ib*). La representación de las variables del Análisis Intraclases Ponderado no aparecen por coincidir con los del Análisis Local-Parcial.

Atendiendo en primer lugar a los resultados del Análisis de Correspondencias obtenidos en el análisis de la matriz original destaca, en el primer eje, la variable 1 (**vg1**) en el lado positivo del mismo opuesta, fundamentalmente, a las variables 2 y 3 (**vg2** y **vg3**), mientras que eje 2 está formado casi exclusivamente por la variable 4 (**vg4**). Respecto a los individuos, están fuertemente asociados a **vg1** el 6 (**ig6**), 7 (**ig7**) y 9 (**ig9**) (cuarto cuadrante) que a su vez tienen niveles muy bajos en el resto de las variables. Por el lado negativo destacan, en el tercer cuadrante, el individuo 1 (**ig1**) con valores altos en **vg3** y

**vg2** y muy bajos en **vg4** y sobre todo en **vg1**, y en el segundo cuadrante, **ig4** fuertemente asociado a **vg3**, **ig3** con valores altos tanto en **vg3** como en **vg4** e **ig5** con proporciones por encima de la media en la variables **vg2** y **vg4**. Por último, en el primer cuadrante aparecen los individuos 2 (**ig2**) y 8 (**ig8**) ligados a las variables 1 (**vg1**) y 4 (**vg4**). La inercia total de esta nube de puntos es de 0.88045.

Teniendo en cuenta la partición efectuada sobre los individuos en diferentes clases, el Análisis de Diferencias Locales nos presenta unos resultados en los que los perfiles de los individuos son comparados, no con el perfil medio de la nube completa, sino con los perfiles medios de las clases a las que pertenece cada uno. Por ejemplo, en la clase 1, si nos fijamos en el individuo 2 destacan más las diferencias positivas que hay respecto al perfil medio de las variables 4 y, sobre todo, 1 y las negativas que hay con respecto a las variables 2 y 3, cuando las comparamos con el perfil medio de la clase que cuando la comparación es con el perfil medio global. Esto hace que **ie2** (individuo 2 en el Análisis de Diferencias Locales) esté más alejado que el correspondiente **ig2 (**la distancia al origen de este punto es de 0.43 en el Análisis de la nube original frente a 1.12 en el Análisis de Diferencias Locales).Otro buen ejemplo de cómo este análisis elimina las tendencias generales poniendo de relieve las diferencias intraclases, es el individuo 5 (**ie5**), en el que se remarcan más diferencias que tiene en la variable 1 si lo comparamos con sus vecinos, los individuos 6 y 7. Si nos fijamos en las variables, la interpretación que podemos hacer de ellas es similar a la que hacemos de los individuos. Por último, la inercia explicada por este análisis, que como dijimos coincide con la inercia intraclase, es 0.50868 lo que supone un 57.8 % de la inercia global.

Por lo que respecta al Análisis Parcial sabemos, como ya lo hemos puesto de manifiesto anteriormente, que la diferencia que existe con el Análisis de Diferencias Locales, en el caso de que el grafo utilizado sea de partición, es unicamente la ponderación que se le da a las distintas clases que componen la partición. Mientras que el Análisis de Diferencias Locales pondera por igual a todas las clases, el Parcial las pondera dependiendo del peso de cada una de ellas en el reparto de las frecuencias. Esto hace que los resultados obtenidos pongan más de manifiesto las diferencias que se presenten en las clases con un mayor peso. Por ejemplo, si nos fijamos en la variable 3 de nuestros datos vemos cómo cambia radicalmente su posición en plano cuando aplicamos a los datos el Análisis Parcial situándose en el lado positivo del eje 2 (**vp3**) frente a donde se situaba tanto en el análisis de la tabla inicial (**vg3**) como en el Análisis de Diferencias Locales (**ve3**). Como ya se ha indicado, el inconveniente de este análisis es que la proyección de los individuos se hace como elementos suplementarios sobre los ejes obtenidos al analizar las aristas del grafo empleado. Esto hace que los resultados, a la hora de analizar estos elementos, no sean los deseados teniéndonos que ceñir, en muchas ocasiones, a la interpretación únicamente de las variables locales. Justamente, la aportación fundamental del Análisis Intraclases Ponderado, en este caso particular, consiste en poder representar, de una manera eficaz, los individuos sobre el plano parcial obteniéndose, como vamos a ver a continuación, muy buenos resultados ya que se tiene en cuenta la importacia de las clases que componen el grafo de partición. Este análisis explica una inercia de 0.48734 lo que representa un 55.35 % del total.

Pongamos como ejemplos a los individuos antes comentados, el individuo 2 del Intraclases Ponderado (**ib2**) y al individuo 5 (**ib5**). Como ya advertíamos antes, la mayor

diferencia que presenta el individuo 2 si comparamos su perfil con el global o con el de su clase es en la variable 1. Dado que este individuo pertenece a la clase 1, que es la de mayor peso, esta diferencia se ve resaltada en el plano hasta el punto de colocar a este individuo en el cuarto cuadrante y más alejado del origen que lo que estaba en el Análisis de Diferencias Locales (la distancia al origen de este individuo es de 1.53 frente al 1.16 del análisis anterior).

Por el contrario, si nos fijamos en el individuo 5 es menor la distancia al origen que tiene en este análisis (0.27) que la que tiene en el Análisis de Diferencias Locales (0.53) debido a la menor importancia de la clase 2. Además este punto sirve para ver la deficiente representación del Análisis Parcial ya que sitúa al individuo en el origen del primer eje (**ip5**) cuando está claramente próximo a la variable 2 y alejado de la variable 1 si lo comparamos con los individuos de su clase (basta con comparar los perfiles del individuo analizado).

Otro punto que resulta especialmente indicado para comprender el objetivo y la eficacia del método propuesto es el individuo 8 (**ib8**). Como se puede observar en el gráfico, pasa de ser considerado en el análisis de la tabla inicial como un individuo asociado a la variable 1 a estar por debajo de la media de esta variable cuando lo comparamos con su vecino -el individuo 9-. De ahí que al realizar tanto el Análisis Intraclases Ponderado como el Análisis de Diferencias Locales se sitúe en la parte negativa del eje 1 (**ie8**, **ib8**). También podemos observar como la distancia al origen es menor en el Análisis Intraclase Ponderado (0.11) que en Análisis de Diferencias Locales (0.50) debido, al igual que en el individuo 5, al menor peso de la clase a la que pertenece, que es la tercera. Por último, las coordenadas obtenidas como un elemento ilustrativo del Análisis Parcial (**ip8**) otra vez ponen de manifiesto lo erróneo del método ya que lo sitúa no sólo en la parte positiva del eje 1 (cuando su aportación a la variable 1 está por debajo de la media como ya hemos dicho) sino que lo coloca en la parte negativa del segundo eje, siendo el valor del perfil con respecto a la variable 4 (**vp4**) superior al de la media de su clase.

En definitiva, con este ejemplo hemos confirmado la validez del método del Análisis Intraclase Ponderado para analizar las tendencias de los datos teniendo en cuenta las relaciones de vecindad de un grafo de partición a través de las variables locales y, sobre todo, la eficacia del método de representación de los individuos dentro de este análisis, que hace de él un método factorial más completo de lo que era hasta ahora en el caso particular analizado.

## 5. Conclusiones

En este trabajo hemos presentado un modelo de ajuste de datos para el Análisis de Correspondencias que permite explicar las relaciones existentes entre los individuos pertenecientes a una misma clase dentro de una partición. La nueva técnica, conceptualmente análoga al Análisis Intraclases, permite resaltar la importancia de los individuos pertenecientes a las clases más dominantes en el análisis. Esta importancia relativa de cada una de las clases se modeliza asignando ponderaciones a cada elemento de la partición, constantes dentro de la misma.

Si en el método propuesto la ponderación elegida sobre cada clase es el perfil que dicha clase tiene en la tabla de datos original, el análisis coincide con el Análisis Parcial definido sobre un grafo de partición no ponderado. Sin embargo, la ventaja que el Análisis Intraclases Ponderado tiene sobre el anterior es que los individuos contribuyen a la formación de los ejes y pueden, por lo tanto, ser explicados con respecto a las variables locales que definen la inercia del análisis. En este caso particular, en el Análisis Intraclases Ponderado, la proyección de un individuo sobre un eje es la proyección del individuo inicial, como ilustrativo sobre el eje (coordenada del Análisis Parcial), trasladada por la del centro de gravedad de la clase a la que pertenece y afectada por el peso de su clase. Es decir, si $F_\alpha^{IP}(i)$ representa la coordenada sobre el eje $\alpha$ de un individuo $i \in I_p$ en el Análisis Intraclases Ponderado (*individuo local*) y $F_\alpha^P(i)$ es la coordenada del individuo en el Análisis Parcial y $F_\alpha(G_p)$ la coordenada del centro de gravedad de la clase $p$:

$$F_\alpha^{IP}(i) = h_p \left[F_\alpha^P(i) - F_\alpha(G_p)\right]$$

El análisis propuesto es un Análisis de Diferencias Locales sobre un grafo de partición si la ponderación elegida se selecciona idéntica para todas las clases.

En el caso particular en el que la ponderación asignada al Análisis Intraclases Ponderado sea el perfil que dicha clase tiene en la tabla inicial, este análisis será tanto más parecido al Análisis en Diferencias Locales sobre un grafo de partición cuanto más homogéneo sea el reparto del perfil medio de la nube entre las diferentes clases de la partición.

En este caso, además, las coordenadas de un individuo en el Análisis Intraclases Ponderado sobre un eje local es la coordenada del individuo del Análisis en Diferencias Locales sobre el mismo eje afectada por la ponderación. Es decir,

$$F_\alpha^{IP}(i) = h_p F_\alpha^{DL}(i)$$

con $F_\alpha^{DL}(i)$ la coordenada del individuo $i$ en la tabla $W_{IJ}^e$ proyectada sobre el eje $\alpha$ obtenido del Análisis Intraclases Ponderado.

## Referencias

[1] Aluja, T., Lebart, L. (1984) "Local and partial principal components analysis and correspondence analysis", *Proceedings COMPSTAT-84*, Physica Verlag, Viena: 113-118.

[2] Benali, H.; Escofier, B. (1990) "Analyse factorielle lissée et analyse factorielle des différences locales", *Revue Statistique Appliquée* **38**(2): 55–76.

[3] Escofier, B. (1984) "Analyse factorielle en référence à un modèle. Application au traitement des tableaux d'échanges", *Revue de Statistique Appliquée* **32**(4): 25–36.

[4] Escofier, B. (1989) "Multiple correspondence analysis and neighboring relation", *Actes Du Colloque de L'INRIA*, Antibes: 55–62.

[5] García-Montoya, M.A. (1998) *Tratamiento Factorial de Estructuras definidas mediante Grafos. Aplicación al Estudio de las Tablas Input-Output.* Tesis Doctoral, U.P.V./E.H.U., Bilbao.

[6] Holmes, R.B.; Jones, L.K. (1996) "On uniform generation of two-way tables with fixed margins and conditional volume test of Diaconis and Efron", *The Annals of Statistics* **24**(1): 64–68.