LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI

El análisis de correspondencias

Les cahiers de l'analyse des données, tome 2, n° 2 (1977), p. 125-142

<http://www.numdam.org/item?id=CAD_1977__2_2_125_0>

Les Cahiers de l'Analyse des Données
Vol. II - 1977 - nº 2 - p. 125-142

EL ANÁLISIS DE CORRESPONDENCIAS

par J.-P. Benzécri
d'après S. Carreiro

Según nosotros, el dato de base de la estadística descriptiva es la matriz rectangular de números positivos. A partir de los cálculos clásicos de la prueba de χ^2 (test de χ^2) (§ 1), asociamos a la matriz dos nubes de puntos que representan respectivamente : el conjunto de las filas y el conjunto de las columnas en un espacio métrico de dimensión elevada (§ 2). Se impone esta representación geométrica particular, por una exigencia de estabilidad : obtener resultados que no cambien cuando en una matriz se suman dos filas (o dos columnas) proporcionales entre sí (§ 3).

Para pasar de espacios de dimensión elevada a representaciones planas accesibles a la intuición, es necesaria la ayuda de una computadora. No intentaremos aquí dar la teoría de los cálculos, pero sí presentaremos exactamente las propiedades de los resultados (§ 4) que condicionan la interpretación (§ 5) y diremos brevemente como conclusión qué clases de matrices han sido analizadas con éxito (§ 6).

1. La métrica de χ^2

Sea un punto $\{x\} = \{x_1, x_2, \ldots, x_n\}$ perteneciente a R^n con una distribución normal esférica, ley para la cual las diversas coordenadas x_i son variables normales independientes de varianza 1, dicha ley es :

$$p(x_1, x_2, \ldots, x_n) = (2\pi)^{-n/2} \exp[-(x_1^2 + x_2^2 + \ldots + x_n^2)/2]$$

El cuadrado de la distancia del punto $\{x\}$ al origen,

$$\|\{x\}\|^2 = x_1^2 + x_2^2 + \ldots + x_n^2$$

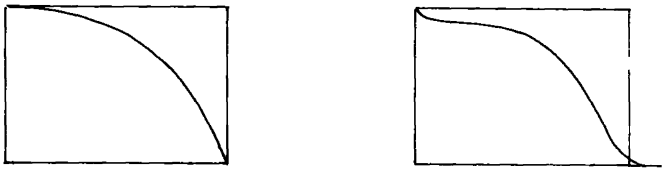
es en sí mismo un valor aleatorio para el cual la ley es llamada ley de χ^2 a n dimensiones. Se pueden demostrar fácilmente, diversas propiedades de esta ley : por ejemplo, que la esperanza matemática (o media) de $\|\{x\}\|^2$ es n : además esta ley está tabulada dado que ella da a los estadísticos una prueba de validez para la hipótesis de que una muestra experimental se ajusta a una ley de probabilidad : es el conocido test de χ^2 del cual recordaremos su uso mediante un ejemplo.

Partamos de la hipótesis que el cociente entre el número de franceses con edades comprendidas entre a, b y el total de la población es :

$$p(a,b) = (4/270) \int_a^b (1-(x(90)^3)dx = \int_a^b p(x)dx$$

Según esta hipótesis las zonas de edades x decrecen como $(1-(x/90)^3)$: no hay sobrevivientes de más de 90 años ; el coeficiente 4/270 ha sido calculado para dar el valor 1 a la p[0,90]. En un régimen estacionario, la tasa de mortandad anual. $(p'(x)/p(x)) = x^2/(243000(1-(x/90)^3))$, parte de cero en el nacimiento para llegar a infinito para los 90 años. La forma de la pirámide de edades es bastante realista : haría falta, solamente, introducir dos correcciones:una para la mortalidad infantil,

la otra para los nonagenarios ; no es éste nuestro fín dado que hemos tomado esta fórmula solamente para nuestro ejemplo



nuestra fórmula pirámide corregida

Fig 1 : dos modelos de pirámide de edades.

Partiendo de este supuesto, para comparar los resultados de un censo a este modelo, dividimos el eje del tiempo en intervalos sucesivos $T_1, T_2, \ldots, T_j, \ldots$ iguales o no

$$T_1 = [0,a_1] \; ; \; T_2 = [a_1, a_2] \; ; \ldots ; \; T_j = [a_{j-1}, a_j] \; ;$$

la notación empleada es :

$p_j = p(\, a_{j-1}, a_j])$

$k(j)$ = efectivo de intervalo de edades T_j según el censo ;

k = población total ;

$f_j = k(j)/k$; frecuencia relativa del intervalo T_j según el censo.

La diferencia entre la ley empírica (las frecuencias f_j) y la ley teórica (las probabilidades p_j) se puede estimar calculando la suma :

$$\|p_J - f_J\|^2_{p_J} = (p_1 - f_1)^2/p_1 + \ldots + (p_j - f_j)^2/p_j + \ldots$$

$$= \Sigma\{(p_j - f_j)^2/p_j \mid j \in J\} \; ;$$

Se calcula la probabilidad de que χ^2 con dimensión igual al número de clases menos 1 sea superior al valor $k\|p_J - f_J\|^2_{p_J}$; si esta probabilidad no es muy pequeña (por ejemplo superior a 0,05) la hipótesis se mantiene, dado que las fluctuaciones sobre una muestra de k componentes, pueden dar en el 5% de los casos, diferencias entre ley y frecuencia de valores superiores a los observados aquí.

Para nosotros, no es esta prueba de validez lo importante, sino la fórmula misma del cálculo de la desviación : será ella la que con el nombre de distancia de χ^2 utilizaremos en el análisis de datos. Examinemos esta fórmula, precisando al mismo tiempo notaciones que conservaremos de ahora en adelante.

J = conjunto que recorre el índice j ; en nuestro ejemplo, consideramos clases T_j consecutivas, y los índices j son naturalmente los que sirven para enumerar las

clases : ej. los números enteros del 1 a 10 si tenemos 10 clases ; pero en general J es un conjunto cualquiera, no un conjunto de números : por ejemplo es conjunto de profesiones, de especies animales o vegetales, etc . y j designa un punto arbitrario del conjunto J.

$p_J = \{p_j \mid j \in J\}$: a cada punto j del conjunto J, le corresponde un número p_j ; la expresión $\mid p_j \mid j \in J \mid$ se lee de la siguiente forma : conjunto de los p_j para j en J, o para j variando en J ; la notación p_J es cómoda por su brevedad.

$\Sigma\{p_j \mid j \in J\} = 1$: esta fórmula se lee : la suma de los p_j para j en J es 1 ; efectivamente los p_j son números positivos, es decir las probabilidades de los eventos j de los cuales siempre una y sólo una se realiza ; también decimos que p_J es una ley de probabilidad o un perfil, sobre el conjunto J.

$f_J = \{f_j \mid j \in J\} = \{k(j)/k \mid j \in J\}$: f_J tiene las mismas propiedades que p_J ; pero como f_J ha sido calculada a partir de datos empíricos, la ley de probabilidad f_J es más precisamente una ley de frecuencia.

$\Sigma\{(p_j - f_j)^2/p_j \mid j \in J\}$: a cada punto j del conjunto J, le corresponde un número $(p_j - f_j)^2/p_j$, función de j : hacemos la suma de todos esos números.

$\|p_J - f_J\|^2_{p_J}$ = cuadrado de la distancia entre los perfiles p_J y f_J en la métrica de χ^2 de centro p_J. Esta fórmula es parecida a la de la geometría analítica usual en R^n :

$$d(\{x\},\{y\})^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 + .. + (x_n - y_n)^2 ;$$

pero aquí, en vez de hacer la suma de los cuadrados de las diferencias de coordenadas, multiplicamos cada cuadrado por un término particular $(1/p_j)$ que está dado por el perfil p_J : por este motivo, es necesario aclarar : métrica (o distancia) de χ^2 de centro p_J.

$\|f'_J - f_J\|^2_{p_J}$ = cuadrado de la distancia entre los perfiles f_J y f'_J en la métrica de χ^2 de centro p_J. Para el test de χ^2 tenemos solamente que calcular la diferencia $\|p_J - f_J\|^2_{p_J}$ entre dos perfiles de los cuales uno, p_J, es tomado como centro (da los coeficientes de los cuadrados en la suma) ; pero en análisis de datos se calcula generalmente la distancia entre dos perfiles cualesquiera que sean.

$$\|f'_J - f_J\|^2_{p_J} = \Sigma\{(f'_j - f_j)^2/p_j \mid j \in J\}.$$

Como vemos, el uso de la métrica de χ^2 presupone la elección de un centro : es la matriz de datos la que nos lo da, como lo explicaremos ahora.

2. Matriz de números positivos y nubes asociadas

Como introducción al presente punto tomemos un ejemplo también de demografía. Para estudiar la población de Herault, J.M. Calvet, y M. Volle del I.N.S.E.E. construyeron la siguiente matriz rectangular : en fila los distintos distritos del departamento considerado ; en 32 columnas la repartición de la población, hombres y mujeres, por intervalos de edades consecutivas de 5 años :

columna 1 M : varones de 0 à 4 años ;

columna 1 F : niñas de 0 à 4 años ;

columna 2 M : varones de 5 à 9 años ; etc...

hasta la columna 16 F, mujeres de más de 74 anos.

Cada fila da, para un distrito una pirámide de edades distinguiendo los sexos ; una columna da para una clase de edad determinada, su distribución geográfica sobre el conjunto de distritos.

En la elaboración estadística de esta matriz y de muchas otras parecidas que se pueden construir, particularmente en el dominio de las ciencias naturales o humanas, conviene acabar las notaciones matemáticas que continuarían las del punto 1.

Notaremos entonces

I : un conjunto finito, el primer conjunto, o conjunto de filas : en este caso será el conjunto de los distritos de Herault ; un elemento arbitrario del conjunto I es llamado i, i', i". Para designar un elemento determinado, comunmente se fijan abreviaciones, por ejemplo en 3 letras : notaremos LØD el distrito de Lodève, LUN el distrito de Lunel etc...

J : otro conjunto finito, el segundo conjunto, o conjunto de columnas : en nuestro ejemplo el conjunto de las 32 clases de edad-sexo ; un elemento arbitrario de J sera notado j,j',j". Como para I, se tomarán las abreviaciones : en nuestro ejemplo se tomó un número seguido de una letra : 1M, 1F etc... La distinción entre el primer y segundo conjunto, entre filas y columnas se hace solamente para fijar una escritura : lógicamente podríamos escribir las mismas informaciones en una matriz en la cual las filas fueran las clases de edad-sexo y las columnas los distritos.Ademas las construcciones matemáticas que efectuaremos de aquí en adelante atribuyen a los dos conjuntos I y J funciones estrictamente simétricas (y eso es una condición importante que no todos los métodos estadísticos satisfacen). En numerosos estudios, lo usual es colocar en filas - el conjunto I - los individuos y en columnas los caracteres - conjunto J (parámetros) , pero esta distinción no es estricta : por ejemplo en una matriz de dato fitosociológicos es cierto que las parcelas (donde son efectuados los sondeos) son individuos caracterizados por la presencia (o la ausencia) de determinada especie (carácter) : pero a su vez les especies son también individuos que se caracterizan por su distribución en el espacio. En el fondo hay mas simetría en los roles de los datos que lo que parece a primera vista.

I x J : conjunto producto de I por J ; es decir el conjunto de los pares (i,j) formados por un punto de I y un punto de J ; concretamente I x J es idéntico al conjunto de casillas de la matriz (casillas sin su contenido numérico)

k(i,j) : el número escrito en la casilla (i,j) de la matriz de datos ; en el ejemplo k(i,j) es la cantidad de habitantes del distrito i pertenecientes a la clase de edad-sexo j ; así k(IØD, 2F) es la cantidad de niñas de 5 a 9 años que viven en el distrito de Lodève. Es necesario para los cálculos que efectuaremos seguidamente que k(i,j) sea un número positivo. Hemos querido partir del ejemplo de una matriz de frecuencia, donde los k(i,j) se obtienen contando ; porque en ese caso, el lenguage probabilístico nos ayuda naturalmente. Pero los cálculos requieren solamente números positivos y no frecuencias ; y el language en sí mismo se extiende por el juego de la analogía a datos de formas muy distintas que hemos analizado con éxito. Sobre los cuales haremos comentarios al final de este capítulo.

$\{k(i,j) \mid i \in I,\ j \in J\}$: el conjunto de números positivos k(i,j) para i perteneciente a I y j perteneciente a J ; es el contenido de la matriz de datos.

$k(i) = \Sigma\{(i,j) \mid j \in J\}$: suma de los elementos de la fila i ; en el ejemplo, es la población total del distrito i sin distinción de edad ni sexo.

$k(j) = \Sigma\{k(i,j) \mid i \in I\}$: suma de los k(i,j) para j fijo e i recorriendo I ; o suma de la columna j ; en el ejemplo es el total de la clase de edad-sexo j, en todo el departamento de Herault sin distinción de distrito.

$k = \Sigma\{k(i,j) \mid i \in I,\ j \in J\}$: suma de todo los k(i,j) de la matriz ; en el ejemplo es la población total de Hérault.

k(i) i I : el conjunto de números k(i) para i en I : es el "margen" (vertical) de la matriz.

{k(j)|j J} : el conjunto de los k(j) para j en J : es el "margen" (horizontal) de la matriz.

Hasta ahora hemos considerado solamente datos brutos o sumas parciales de ellos ; es decir, en el ejemplo, frecuencias absolutas o números enteros dando hechos individuales (ej : existencia de una o más mujeres de más de 75 años, en un distrito dado. En lo sucesivo consideramos proporciones o frecuencias relativas, e introducimos el lenguage de las probabilidades.

$f_{ij} = k(i,j)/k$: frecuencia relativa de la clase (i,j) dada en el ejemplo por una residencia i y una clase de edad-sexo j.

$f_i = k(i)/k$: frecuencia relativa marginal del distrito i en la población total de Hérault ; remarquemos que tenemos :

$f_i = \Sigma\{f_{ij} \mid j \in J\}$ (como k(i), f_i es la suma de una fila).

$f_j = k(j)/k$: frecuencia relativa marginal de j ; en el ejemplo, parte de la clase de edad-sexo j ; remarquemos que :

$f_j = \Sigma\{f_{ij} \mid i \in I\}$.

$f_{IJ} = \{f_{ij} \mid i \in I,\ j \in J\}$: el sistema de frecuencias relativas f_{ij} define en el conjunto producto I x J una ley de probabilidad, más precisamente una ley de frecuencia : f_{ij} es la masa relativa o probabilidad del par (i,j) ; en el ejemplo : es la probabilidad que un habitante de Hérault, tomado al azar, entre dentro de la clase (i,j) ; es claro que tenemos

$\Sigma\{f_{ij} \mid i \in I,\ j \in J\} = 1$;

la suma de los f_{ij} para i en I y j en J es 1.

$f_I = \{f_i \mid i \in I\}$: ley marginal sobre I ; en el ejemplo, se trata de una ley de frecuencia propiamente dicha, pero en el caso de una matriz de números positivos k(i j) cualesquiera que sean, hablaremos más bien de perfil marginal. Es claro que

$\Sigma\{f_i \mid i \in I\} = 1$

$f_J = \{f_j \mid j \in J\}$: ley marginal (o perfil) en J.

$f_i^j = f_{ij}/f_j = k(i,j)/k(j)$: en términos probabilísticos frecuencia condicional de i, dado j ; es el peso relativo de la casilla (i,j) en la columna j ; en el ejemplo, es la parte de la clase de edad-sexo que reside en el distrito i.

$f_I^j = \{f_i^j \mid i \in I\}$ perfil de la columna j ; tenemos que

$\Sigma\{f_i^j \mid i \in I\} = 1$. El perfil nos muestra una realidad más estable y más significativa que la columna tomada tal cual. Supongamos que consideramos inútil distinguir entre las clases de 35 a 39 años (columnas 8M y 8F) y de 40 a 44 años (columnas 9M y 9F), constituiremos nuevas columnas designadas por ejemplo por las siglas 40M y 40F y correspondientes al intervalo de edad 35-44 años : la nueva columna 40M es la suma de las anteriores columnas 8M y 9M, igualmente 40F para 8F y 9F. Es probable que 8M y 9M tengan perfiles vecinos, con lo cual el perfil de la columna 40M diferirá poco, contrariamente el peso relativo de la columna 40M es aproximadamente el doble del de cada una de las columnas 8M o 9M. Ese peso relativo depende evidentemente de la amplitud de los intervalos que el estadístico ha elegido arbitrariamente como cortes ; mientras que el perfil es sólo función del nivel (por los 40 años, o : por los 70 años) de la clase considerada.

$f_j^i = f_{ij}/f_i = k(i,j)/k(i)$: frecuencia condicional de j, dado i ; o : peso relativo de la casilla j en la fila i ; en el ejemplo, parte de la población del distrito i que está comprendida en la classe de edad-sexo j.

$f_J^i = \{f_j^i \mid j \in J\}$: perfil de la fila i ; tenemos que $\Sigma\{f_j^i \mid j \in J\} = 1$

Dos distritos i e i' que tienen el mismo tipo de población tienen el mismo perfil, aun cuando sus pesos relativos, f_i , $f_{i'}$, puedan ser muy diferentes. Puede ocurrir que une reforma administrativa subdivida en dos un distrito suficientemente homogéneo, resultarán entonces dos nuevos distritos en los cuales los perfiles serán parecidos al del distrito del cual provienen ; igualmente fusionando dos distritos parecidos se creará una nueva unidad sin gran variación de perfil.

Las definiciones y notaciones dichas sirven a la codificación o a la traducción geométrica de las informaciones contenidas en la matriz de datos $[k(i,j) \mid i \in I, j \in J]$. Notaremos :

R_I : espacio vectorial de las medidas sobre el conjunto I : una medida (o distribución de masa), $x_I = \{x_i \mid i \in I\}$ está definida por el sistema de masas x_i de los puntos de I ; aquí los números x_i no están sujetos a ninguna condición. El espacio R_I estará provisto (dotado) de la métrica de χ^2 de centro f_I : el cuadrado de la distancia entre x_I e y_I esta dado por la fórmula (§ 1) :

$$\|x_I - y_I\|^2_{f_I} = \Sigma\{(x_i - y_i)^2/f_i \mid i \in I\}$$

P_I = conjunto de los perfiles sobre I ; P_I es una parte de R_I , lo que escribimos : $P_I \subset R_I$; un elemento x_I de R_I define un perfil si se cumplen las dos condiciones siguientes :

$\forall\, i \in I : x_i \geq 0$; para todo i en I, x_i es positivo o nulo.

$\Sigma\{x_i \mid i \in I\} = 1$; la suma de los x_i para i en I es 1.

R_J = espacio vectorial de las medidas sobre el conjunto J ; una medida x_J está dada por $\{x_j \mid j \in J\}$. En el estudio de una matriz de datos, el espacio R_J será provisto de la métrica de χ^2 teniendo por centro la ley marginal f_J : el cuadrado de la distancia entre x_J e y_J está dado por la fórmula :

$$\|x_J - y_J\|^2_{f_J} = \Sigma\{(x_j - y_j)^2/f_j \mid j \in J\}$$

P_J = conjunto de perfiles sobre J ; $P_J \subset R_J$; tenemos $x_J \in P_J$ con las dos condiciones :

$\forall\, j \in J : x_j \geq 0$; $\Sigma\{x_j \mid j \in J\} = 1$

$N(I) = \{(f^i_J , f_i) \mid i \in I\}$: nube asociada al conjunto I : por nube se entiende el conjunto de puntos provistos de masa ; a cado elemento i de I (en el ejemplo, i es un distrito de Hérault) le hacemos corresponder un punto, su perfil f^i_J , que es un punto del espacio R_J , más precisamente de P_J ; y una masa - su frecuencia relativa f_i.

$N(J) = \{(f^j_I , f_j) \mid j \in J\}$: nube asociada al conjunto J, por una construcción igual a la de la nube $N(I)$.

En el espacio R_J , la nube $N(I)$ tiene como centro de gravedad el perfil marginal (o perfil medio) f_J : este resultado es evidente para quienes están familiarizados con los cálculos de centro de gravedad en los espacios vectoriales ; también sugerirá al menos, a todo lector, que en R_J la nube $N(I)$ de los perfiles de las filas está distribuida alrededor de f_J. Lo mismo en R_I , $N(J)$ tiene por centro f_I.

Para nosotros el análisis de datos - taxonomía numérica, determinación de formas (reconocimiento de formas), análisis factorial - consiste esencialmente en el estudio de dos nubes $N(I)$ y $N(J)$ asociadas a una matriz rectangular de números positivos. Por eso es por lo que pareció imposible ahorrar al utilizador este desfile de fórmulas entrecortadas de comentarios : sin corchetes ni índices, temíamos ser tan obscuros como el gran Laplace, describiendo por circunloquios de prosa la ley que lleva su nombre. Como ejercitación calculemos sobre una matriz ficticia dando para 3 cosechas de planctón las cantidades de diatomeas, coccolitoforidos y flagelos : aun cuando en la educación primaria la teoría de conjuntos ha reemplazado la tabla de sumar y la ayuda de la computadora sea accesible a todo científico, es prudente asegurarse sobre un ejemplo numérico que hemos entendido las definiciones generales.

	DIA	COC	FLA	DIN	marge
R1	3	2	0	1	6
R2	2	1	3	4	10
R3	0	2	2	3	7
marge	5	5	5	8	23

cuadro II 1 : matriz de datos y sus marginales

$f_J^{R1} = 0,5 ; 0,33 ; 0 ; 0,16$

$f_J^{R2} = 0,2 ; 0,1 ; 0,3 ; 0,4$

$f_J^{R3} = 0 ; 0,29 ; 0,29; 0,42$

$f_J = 0,22 ; 0,22 ; 0,22; 0,35$

cuadro II 2 : Perfil de cosechas y perfil medio sobre J.

$$d^2 R1,R2) = ((0,5-0,2)^2/0,22) + ((0,33-0,1)^2/0,22) + ((0-0,3)^2/0,22)+((0,16-0,4)^2/0,35$$

$$d^2(R1,R2) = \|f_J^{R1} - f_J^{R2}\|^2_{f_J} = 1,23 \; ; \; d(R1,R2) = 1,11$$

3. Principio de equivalencia distribucional

Frente a la ley de frecuencia f_{IJ} en el conjunto I x J, un dístico prevee generalmente la hipótesis de independencia entre i y j ; es decir la hipótesis según la cual los i y los j asociados al azar, la frecuencia f_{ij} no difiere del producto $f_i \times f_j$ de las frecuencias marginales salvo por las fluctuaciones del muestreo. Mas precisamente esta hipótesis puede ser sometida a la prueba de χ^2 : Calculamos la suma (*) :

$$\|f_{IJ} - f_I f_J\|^2_{f_I f_J} = \Sigma\{(f_{ij} - f_i f_j)^2/f_i f_j) \,|\, i \in I, \; j \in J\} \; ,$$

y miramos (§ 1) si no es demasiado débil la probabilidad de que un χ^2 con dimensión (Card I -1)(Card J -1) (por Card I , Card J designamos el cardinal o número de elementos de los conjuntos I, J) pase el valor $k\|f_{IJ} - f_I f_J\|^2$ (por k, cf § 2, designamos la cantidad total de la muestra; en el ejemplo, la población de Hérault).

Con la hipótesis de independencia, el perfil de toda fila f_J^i es idéntico al perfil marginal f_J ; igualmente el perfil de toda columna f_I^j es igual a f_I : las nubes $N(I)$ y $N(J)$ están concentradas en un punto. Como consecuencia no hay análisis de datos posible ; en la medida en que para nosotros, el objeto de este análisis es el estudio de las nubes $N(I)$ y $N(J)$ o, lo que es equivalente, el estudio estructural de la variación entre f_{IJ} y la ley producto $f_I f_J$. Es interesante el que la medida de esta variación calculada para aplicar la prueba de χ^2 sea igual a la inercia total de una u otra de las nubes $N(I)$ y $N(J)$ (es decir a la suma ponderada por las frecuencias marginales, de los cuadrados de las distancias de los perfiles al perfil medio) : en efecto tenemos :

$$\|f_{IJ} - f_I f_J\|^2_{f_I f_J} = \Sigma\{f_i \|f_J^i - f_J\|^2_{f_J} \,|\, i \in I\} = \text{inercia de } N(I)$$

$$= \Sigma\{f_j \|f_I^j - f_I\|^2_{f_I} \,|\, j \in J\} = \text{inercia de } N(J)$$

(*) otra expresión de la diferencia entre f_{IJ} y $f_I f_J$ está dada por la teoría de la información.

De ahora en adelante, las representaciones esquemáticas de $N(I)$ y $N(J)$ dadas por el análisis factorial y la clasificación automática serán estimadas por el porcentaje de la dispersión total (también llamada : traza (*) $\|f_{IJ} - f_I\, f_J\|^2$, del cual ellas cuenta (en un sentido que podemos precisar).

Al definir los perfiles de filas y columnas, hemos mostrado sobre el ejemplo propuesto, que dichos perfiles parecían poco sensibles a las elecciones arbitrarias en cuanto a la extensión de las circunscripciones territoriales y de las clases de edades. De forma más precisa, tenemos el siguiente resultado :

Sean i_1 e i_2 dos elementos que tienen el mismo perfil : $f_J^{i_1} = f_J^{i_2}$; entonces no cambiamos las nubes $N(I)$ y $N(J)$ si substituimos las dos filas i_1 e i_2 por una fila única, i_s que sea su suma.

De donde el *principio* de *equivalencia distribucional*, principio según el cual dos elementos i_1 , i_2 que tienen la misma distribución (es decir mismo perfil, que se asocian por igual a los elementos del otro conjunto ; el término de distribución proviene de los trabajos del linguista Z.S. Harris) pueden ser distinguidos o confundidos en el análisis indiferentemente.Efectivamente en $N(I) = \{(f_J^i\ ,\ f_i) \mid i \in I\}$ la única modificación es que substituimos los puntos confundidos $f_J^{i_1}$ y $f_J^{i_2}$ que tienen por masa f_{i_1} y f_{i_2} , por un punto único $f_J^{i_s}$, con masa $f_{i_s} = f_{i_1} + f_{i_2}$. Para $N(J)$, a primera vista la modificación es más importante : dado que espacio ambiente de la nube, el espacio R_I ha cambiado pues el conjunto I ha sido modificado (i_s reemplaza i_1 e i_2) ; pero los nombres de las columnas no cambian, y se puede demostrar por cálculos que las distancias tampoco cambian.

Evidentemente, el *principio* vale también para dos elementos j_1 y j_2 que tengan el mismo perfil $f_I^{j_1} = f_I^{j_2}$; no cambiamos $N(I)$ ni $N(J)$ substituyendo las columnas j_1 y j_2 de la matriz por una columna j_s.

Es el principio de la equivalencia distribucional el cual en una fórmula de distancia tal como :

$$d^2(i,i') = \Sigma\{(f_j^i - f_j^{i'})^2\ \alpha_j \mid j \in J\},$$

impone elegir los coeficientes α_j proporcionales a $1/f_j$, es decir impone la elección de la distancia de χ^2. Esta elección está de acuerdo con lo usual en estadísticas para le comparación de las leyes de probabilidad (comparación entre sí de dos perfiles, y comparación de éstos al perfil marginal, o medio) ; pero el principio de equivalencia distribucional lo hace legítimo y útil aun fuera del dominio de las matrices de frecuencia (§ 6). En cuanto a la substitución de las filas (o columnas) brutas por los perfiles, es una condición previa necesaria para el éxito de muchos análisis : sin ella, el efecto de talla (la oposición entre filas pesadas y filas ligeras) predomina e impide distinguir las otras dimensiones. Codificados por sus perfiles, dos individuos de talla muy diferente pero de forma parecida se transforman en puntos vecinos de masa distinta.

(*) "traza" es un vocablo técnico (francés, inglés : trace ; alemán : Spur, cuyo significado se acerca lo más al castellano : huella).

De ahora en adelante, consideramos las matrices rectangulares de números positivos cualquiera que sea su origen aun cuando los datos k(i,j) no sean únicamente frecuencias (es el caso, por ejemplo, de una matriz de medidas tomadas sobre una serie de cráneos) haremos las construcciones matemáticas introducidas en el punto 2 con el lenguage de las probabilidades. Una matriz $\{k(i,j) \mid i \in I, j \in J\}$ de números positivos es llamada *matriz de correspondencias*, porque siempre el número k(i,j) nos da una relación entre los elementos de dos conjuntos distintos, y que el sistema de estas relaciones da sobre los dos conjuntos I y J estructuras que se corresponden. Ciertos linguistas estructuralistas, inspirándose de una filosofia nominalista han afirmado que todo individuo (por ejemplo toda palabra) no es otra cosa, que el conjunto de relaciones que lo unen a los otros. Una tesis tan radical es inaceptable : las cosas existen por sí mismas ; pero es por el inventario ordenado de las relaciones establecidas entre las cosas que la ciencia positiva descubre lo que ellas son. A este inventario ordenado, la estadística ayudada por la computadora, pretende contribuir.

4. El análisis de correspondencias

En este punto describimos un conjunto de resultados que podemos obtener a partir de una matriz de números positivos - o matriz de correspondencias. Se pueden encontrar, en los cursos de estadística, explicaciones más o menos elementales del *análisis factorial de correspondencias* : aquí nos limitamos estrictamente a describir los resultados dados por este análisis, a enunciar algunas propiedades geométricas.

El análisis da simultáneamente sobre los dos conjuntos I y J una serie ordenada de pares de funciones o *factores* F_α , G_α a cada uno de los cuales está asociado un número real λ_α , comprendido entre 0 y 1, el *valor proprio*, que decrece de factor en factor. Así pues tenemos :

$$F_1 = \{F_1(i) \mid i \in I\} \; ; \; G_1 = G_1(j) \mid j \in J\} \; ; \; \lambda_1 \; ;$$

$$F_2 = \{F_2(i) \mid i \in I\} \; ; \; G_2 = G_2(j) \mid j \in J\} \; ; \; \lambda_2 < \lambda_1 \; ;$$

$$\vdots$$

$$F_\alpha = \{F_\alpha(i) \mid i \in I\} \; ; \; G_\alpha = \{G_3(j) \mid j \in J\} \; ; \; \lambda_\alpha < \ldots < \lambda_2 < \lambda_1 \; ;$$

$$\vdots$$

$F_\alpha(i)$ es el valor del factor de rango α , en el punto i del conjunto I ; $G_\alpha(j)$ es el valor del factor de rango α , en el punto j del conjunto J. En un sentido que será precisado, la importancia de los factores es evaluada por el valor propio λ_α ;y a menudo hablamos de *extracción de los factores*, como si el cálculo extrajera sucesivamente la estructura que contiene la matriz analizada dando (F_1 , G_1) después (F_2 , G_2) etc... La serie de los factores se acaba en el rango n que es el mínimo de los números (Card I - 1) y (Card J - 1), (donde Card I = número de elementos de I) ; pero practimente el examen de los factores se acaba antes de su fin teórico : se llega raramente hasta el décimo.

En los conjuntos I y J, los factores son funciones que tienen media nula, varianza λ_α y que no están correlacionados entre sí, que es lo que expresan las fórmulas siguientes (donde α y β son dos indices diferentes) :

$$\Sigma\{f_i\, F_\alpha(i) \mid i \in I\} = 0 \quad ; \Sigma\{f_j\, G_\alpha(j) \mid j \in J\} = 0 \quad ;$$

$$\Sigma\{f_i\, F_\alpha(i)^2 \mid i \in I\} = \lambda_\alpha \quad ; \Sigma\{f_j (G_\alpha(j))^2 \mid j \in j\} = \lambda_\alpha \quad ;$$

$$\Sigma\{f_i\, F_\alpha(i)\, F_\beta(i) \mid i \in I\} = 0 \; ; \Sigma\{f_j\, G_\alpha(j)\, G_\beta(j) \mid j \in J\} = 0 \; ;$$

en estas fórmulas, los elementos i y j reciben por masa sus frecuencias marginales f_i , f_j.

Conocer los perfiles marginales f_I , f_J , y la serie de todos los factores y valores propios (F_α , G_α , λ_α), es conocer la misma matriz f_{IJ} : es la *fórmula de reconstitución* de *los datos* a partir de los factores :

$$f_{ij} = f_i\, f_j\, (1 + \Sigma_\alpha\, \lambda_\alpha^{-1/2}\, F_\alpha(i)\, G_\alpha(j).$$

Tal cual esta fórmula es poco práctica, pues la sumatoria Σ_α se extiende a todos les factores : pero existen fórmulas aproximadas tales como la signiente :

$$f_{ij} \approx f_i\, f_j\, (1 + \lambda_1^{-1/2} F_1(i)\, G_1(j) + \lambda_2^{-1/2} F_2(i)\, G_2(j) + \lambda_3^{-1/2}\, F_3(i)\, G_3(j)) \; ,$$

que considera hasta el 3° factor.

Hay que considerar atentamente los términos sucesivos de la fórmula de reconstitución, pues de ellos depende la interpretación de los resultados del análisis. El primer término, $f_i\, f_j$, sin factor : detenida en el rango cero, la fórmula de aproxima ción da la hipótesis de independencia (cf § 3) ; y los factores sucesivos describen la estructura de la matriz f_{IJ} por las correciones aportadas a esta hipótesis. La parte del factor α es $f_i\, f_j\, \lambda_\alpha^{-1/2}\, F_\alpha(i)\, G_\alpha(j)$: este término es positivo si $F_\alpha(i)$ y $G_\alpha(j)$ tienen igual signo, en caso contrario, es negativo : tener igual signo corresponde por lo tanto a una afinidad entre i y j, es como si el factor α midiera una cierta característica común a los elementos de los dos conjuntos y gobernara sus asociaciones. El coeficiente $\lambda_\alpha^{-1/2}$ es inquietante : hemos dicho que la serie de los λ_α decrece cuando crece el rango α ; por lo tanto $\lambda_\alpha^{-1/2}$ (lo inverso de la raiz α cuadrada de λ_α) crece ; esto parece indicar que la importancia de los términos crece también con el rango. No es así porque $F_\alpha(i)$ y $G_\alpha(j)$ tienen como orden de importancia $\lambda_\alpha^{1/2}$ que es la desviación-tipo (raiz cuadrada de la varianza) del factor α ; finalmente tenemos :

$$\lambda_\alpha^{-1/2}\, F_\alpha(i)\; G_\alpha(j) \approx \lambda_\alpha^{-1/2} \times \lambda_\alpha^{1/2} \times \lambda_\alpha^{1/2} = \lambda_\alpha^{1/2}$$

la importancia del término decrece como la desviación-tipo $\lambda_\alpha^{1/2}$.

Para las nubes $N(I)$ y $N(J)$ en los espacios R_I y R_J provistos de sus métricas, los factores suministran nuevas coordenadas ortonormales ligadas a f_j^i y f_i^j por transformaciones lineales. Es claro que la fórmula de reconstitución permite calcular f_j^i o f_i^j en función de los factores :

$$f_j^i = f_{ij}/f_i = f_j\, (1 + \Sigma\, \lambda_\alpha^{-1/2}\, F_\alpha(i)\; G_\alpha(j))$$

$$f_i^j = f_{ij}/f_j = f_i\, (1 + \Sigma\, \lambda_\alpha^{-1/2}\, F_\alpha(i)\; G_\alpha(j)).$$

Recíprocamente una fórmula da los $F_\alpha(i)$ como combinación lineal de los f_i^j : no se trata más que de una fórmula de geométria analítica que expresa nuevas variables en función de las variables iniciales ; tenemos :

$$F_\alpha(i) = \lambda_\alpha^{-1/2} \Sigma\{f_j^i \; G_\alpha(j) \mid j \in J\} \; ;$$

$$G_\alpha(j) = \lambda_\alpha^{-1/2} \Sigma\{f_i^j \; F_\alpha(i) \mid i \in I\} \; ;$$

los coeficientes que sirven para el cálculo del factor $F_\alpha(i)$ en función de los f_j^i son suministrados por $G_\alpha(j)$ (asimismo $G_\alpha(j)$ se calcula a partir de los f_i^j con los $F_\alpha(i)$ como coeficientes). Por eso esta fórmula suele ser llamada *fórmula de transición* : pues permite pasar de un factor dado sobre uno de los conjuntos al mismo factor sobre el otro (ej. conociendo F_α y los f_i^j se calcula G_α). la fórmula permite además otra lectùra, esencial a la interpretación de los resultados del análisis : el valor $F_\alpha(i)$ del factor de rango α en el punto i de I es si se le quita el multiplicador $\lambda_\alpha^{-1/2}$) el promedio de los valores $G_\alpha(j)$ de ese factor sobre el conjunto J ponderados por las coordenadas f_j^i del perfil f_J^i de i. Es el llamado principio del centro de gravedad o *principio baricéntrico* (baricentro significa, en mecánica, centro de gravedad, o media, de un conjunto de masas puntuales de diferente tamaño : aquí las masas son las f_j^i) : cualitativemente el principio implica que el signo de $F_\alpha(i)$ está dado por el de G_α en los puntos j con los cuales i se asocia más (términos f_j^i predominantes del perfil f_J^i) : repitiendo, los factores son como las cualidades comunes a los elementos de los conjuntos I y J, y determinan sus asociaciones.

En fin los factores son coordenadas ortonormales en el sentido que en función de los factores el cuadrado de la distancia se expresa como una suma de cuadrados con todos los coeficientes iguales a uno ; tenemos :

$$\| f_J^i - f_J^{i'} \|^2_{f_J} = \Sigma_\alpha \; (F_\alpha(i) - F_\alpha(i'))^2 \; ;$$

$$\| f_J^i - f_J \|^2_{f_J} = \Sigma_\alpha F_\alpha(i)^2 \; ;$$

$$\| f_I^j - f_I^{j'} \|^2_{f_I} = \Sigma_\alpha (G_\alpha(j) - G_\alpha(j'))^2$$

$$\| f_I^j - f_I \|^2_{f_I} = \Sigma_\alpha \; G_\alpha(j)^2$$

Así pues la inercia total de la nube $N(I)$ está dada por la suma :

$$\| f_{IJ} - f_I \, f_J \|^2_{f_I f_J} = \Sigma\{f_i \| f_J^i - f_J \|^2 \mid i \in I\}$$

$$= \Sigma\{f_i \; \Sigma_\alpha \; F_\alpha(i)^2 \mid i \in I\}$$

$$= \Sigma_\alpha \, \Sigma\{f_i F_\alpha(i)^2 \mid i \in I \; = \Sigma_\alpha \; \lambda_\alpha \; ;$$

y lo mismo para $N(J)$. La suma de los valores propios $(\lambda_1 + \lambda_2 + ... + \lambda_\alpha + ..) = \Sigma_\alpha \; \lambda_\alpha$ no es sino la inercia total de la nube $N(I)$ o $N(J)$; y el cociente $\lambda_\alpha / \Sigma_\alpha \; \lambda_\alpha$, generalmente expresado como τ_α (y expresado por un porcentage), es la parte relativa, o tasa de inercía referente al factor de rango α.

Hay que imaginar en el espacio R_J (espacio de mediciones sobre el conjunto J) un sistema de ejes ortogonales entre ellos (llamados *ejes factoriales*), cuyo origen está en el centro f_J de la nube $N(I)$: $F_\alpha(i)$ es la coordenada del perfil f_J^i (punto de R_J) sobre el α^o eje de ese sistema. Se puede, por la fórmula de transición calcular en el sistema las coordenadas de un perfil cualquiera $x_J \in P_J \subset R_J$; tenemos :

$$F_\alpha\ (x_J) = \lambda_\alpha^{-1/2} \Sigma\{x_j\ G_\alpha(j) \mid j \in J\}$$

Relativamente a la nube $N(I)$ esos ejes no son más que los ejes principales de inercia, noción usual en mecánica ; y los valores propios λ_α son los momentos principales de inercia. Determinando sucesivamente los ejes y *extrayendo los factores*, el análisis factorial realiza una *reducción de la dimensión* de la nube. En resumen el primer eje es, entre todas las rectas de R_J , aquella de la cual la nube $N(I)$ se separa menos ; los dos primeros ejes definen asimismo el plano de R_J respecto del cual la nube $N(I)$ tiene la desviación mínima ; asimismo los p primeros ejes definen el subespacio de dimensión p de R_J que da la mejor aproximación de $N(I)$: esta noción de subespacio de dimensión p es la generalización algebráica a R_J de las nociones geométricas del espacio ordinario . Para fijar el pensamiento con una imagen dibujaremos la nube $N(I)$ alrededor de su centro f_J , aplanada en el plano de los dos primeros ejes y más estirada en la dirección del primero.

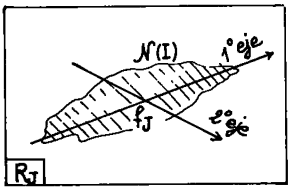


Fig 2 esquema plano de la nube N(I)

Es evidente que en R_I se tiene también un sistema de ejes de origen f_I , que el sistema de los ejes principales de inercia de la nube $N(J)$. Construyendo simultáneamente en R_I y R_J dos sistemas de ejes, el análisis ha ligado esos dos espacios : se superpondrá entonces (cf § 5) por ejemplo, el plan de los dos primeros ejes (o más generalmente el plan de los ejes α y β) de R_I y de R_J ; y con esos planos las proyecciones sobre ellos de las nubes $N(J)$ y $N(I)$: representación tanto más fecunda cuanto que en virtud del principio baricéntrico ésta coloca cada elemento aproximadamente (con o sin el coeficiente $\lambda_\alpha^{-1/2}$) en el centro de aquellos del otro conjunto que se asocian con él.

Hemos hablado hasta aquí de factores y de valores propios como si ellos estuvieran determinados sin ambigüedad por la matriz de datos analizada. Esto es cierto con dos restricciones.

Primeramente : el par (F_α , G_α) no tiene un signo fijo : se puede cambiar simultaneante el signo de F_α y G_α , es decir cambiar simultáneamente la orientación de los ejes factoriales de rango α en R_J y R_I (cambiar simultáneamente es esencial, sino es evidente que las relaciones se destruyen, por ejemplo la fórmula de transición).

En segundo lugar : si dos factores por ejemplo (F_1 , G_1) y (F_2 , G_2) son relativos a un mismo valor propio $\lambda_1 = \lambda_2 = \lambda$ esos factores pueden ser reemplazados indistintamente por otros que son las coordenadas sobre nuevos ejes que se han hecho girar de un ángulo θ cualquiera :

$$F_{1'} = F_1 \cos\theta + F_2 \sin\theta \quad ; \quad F_{2'} = -F_1 \sin\theta + F_2 \cos\theta \quad ;$$

lo mismo para G. En la práctica es excepcional que dos valores propios sean rigurosamente iguales ; pero se debe saber que si dos factores consecutivos, p. e. los factores 2, 3 son relativos a valores propios muy vecinos, las fluctuaciones de la muestra pueden en un nuevo análisis hacer girar los ejes : en la interpretación se considerará pues el plano de los ejes 2 y 3 que es estable, mientras que ni el eje 2, ni el eje 3, ni los planos 1 x 2 y 1 x 3 lo son.

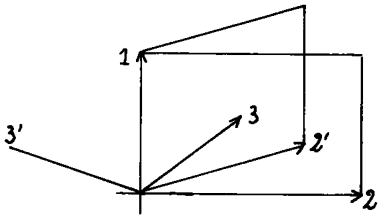


Fig 3 inestabilidad del plano 1x2 cuando $\lambda_2 = \lambda_3$.

En fin, es posible considerar los factores F_α y G_α definidos respectivamente sobre I y J como funciones particulares definidas sobre el conjunto I x J de los pares (i,j) : es suficiente poner :

$F_\alpha(i,j) = F_\alpha(i)$ y $G_\alpha(i,j) = G_\alpha(j)$. Es posible entonces calcular sobre I x J provisto de la ley f_{IJ}, la correlacion entre F_α y G_α (σG_β). Se constata que

$$\text{Corr}(F_\alpha, G_\alpha) = \lambda_\alpha^{-1/2} \text{ y } \text{Corr}(F_\alpha, G_\beta) = 0 \text{ si } \beta \neq \alpha \ ;$$

además, el máximo de la correlación sobre I x J entre una función de i y una función de j es realizada por F_1, G_1 ; después, entre las funciones ortogonales a éstas, por F_2, G_2 etc. No nos extenderemos, sobre esta interpretación limitándonos a señalar que volvemos a encontrarnos con una técnica llamada análisis canónico, es decir, dados dos grupos de variables definidas sobre un mismo conjunto, se busca entre las combinaciones lineales de las variables de uno de los grupos las más correlacionadas con las combinaciones lineales salidas del otro grupo.

5. *Interpretación*

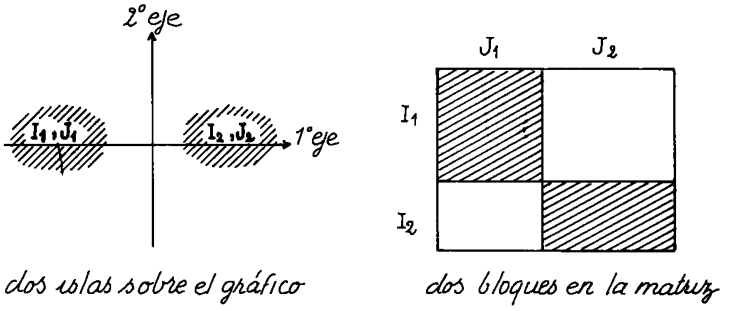
Para conocer los resultados de un análisis, lo más rápido es considerar los diagramas planos ; y en primer lugar las nubes $N(I)$ y $N(J)$ representadas simultáneamente en el plano de los ejes 1 y 2. La computadora imprime un gráfico sobre el cual los puntos están representados por siglas abreviadas de dos o tres caracteres ; el punto i tiene por abscisa $F_1(i)$ y por ordenada $F_2(i)$; el punto j tiene por abscisa $G_1(j)$ y por ordenada $G_2(j)$. Después se consideran los planos 1 x 3, 2 x 3, 1 x 4 etc..

En general, se tienen de los conjuntos I y J ciertas ideas que se confrontan primero a los resultados del análisis. Si por ejemplo I es un conjunto de animales (cráneos, fósiles, etc.) sometidos a un conjunto J de mediciones, se buscará si sobre el gráfico plano los individuos de los cuales se supone que pertenecen a la misma subespecie, están agrupados ; y lo mismo para las mediciones j análogas, por ejemplo los largos, los anchos o las medidas relativas a la cápsula craneana con exclusión de

las de la cara... Si aparecen agrupaciones de puntos inesperadas se busca la característica común que pudo reunirlos. El análisis es más fecundo cuando requiere ideas nuevas, sugiere nuevas colecciones de hechos. Tratando la matriz de valores independientemente de todo modelo matemático e incluso de toda hipótesis *a priori* se pueden encontrar los ejes mismos de la naturaleza. No se debe esperar entonces, por ejemplo que la oposición entre dos problaciones de individuos recogidos en lugares distintos aparezca simplemente sobre un eje α , los individuos de la 1° población teniendo $F_\alpha(i)$ positivo y los de la otra $F_\alpha(i)$ negativo. Las poblaciones no son más que híbridos, escalonados entre grandes tipos y la orientación de los ejes puede justamente sugerir donde buscar los tipos puros, verdaderas subespecies. Además se debe admitir que un fenómeno complejo no está suficientemente representado por un gráfico por bueno que sea : hay estructuras que no son aprehendidas más que por un esfuerzo de imaginación después del examen paciente de varios *gráficos*. Las representaciones geométricas surgidas del cálculo deben sostener al espíritu en su vuelo pero no pueden reemplazar el vuelo.

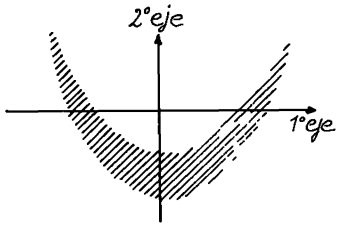
A veces un *gráfico* mudo es suficiente para revelar una estructura ; sería el caso de una nube bien separada en dos sub-nubes : de un lado I_1 con J_1 , del otro I_2 con J_2.

Esta partición es generalmente visible volviendo a escribir la matriz de datos : basta con agrupar las filas en dos grupos sucesivos I_1 , I_2 ; e igualmente para las columnas :

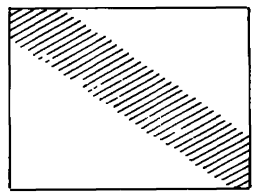


dos islas sobre el gráfico *dos bloques en la matriz*

los números más elevados se agrupan en dos bloques $I_1 \times J_1$ y $I_2 \times J_2$. Otra forma típica : la nube en media luna ; volviendo a ordenar las filas y las columnas en el orden creciente del primer factor, se ve aparecer en la matriz una banda diagonal de números fuertes, encuadrados por dos triángulos de números débiles.



nube en media luna sobre el gráfico



banda diagonal sobre la matriz

Dado que se funda principalmente sobre los gráficos, la interpretación es accesible al utilizador aun si conoce poco de matemáticas. Sin embargo es necesario mirar los números también. Primeramente los valores propios λ_α y las tasas τ_α. Supongamos que tenemos : $\tau_1 + \tau_2 = 87\%$; el *gráfico* 1 x 2 da cuenta de casi toda la inercia (o dispersión) de los puntos de la nube de la cual *él* da una imagen fiel ; los factores siguientes darán sólo más detalles (tal vez muy interesantes). Si tenemos :

$\tau_1 = 33\%$, $\tau_2 = 23\%$, $\tau_3 = 19\%$, $\tau_4 = 7\%$ etc., es importante considerar desde el comienzo de la interpretación, los tres primeros factores. Los valores propios, como hemos dicho, están comprendidos entre 0 y 1. Un valor propio muy elevado (p. ej. 0,8) da generalmente una dicotomía (dos islotes separados). Una matriz de medidas tomadas sobre los vegetales puede dar valores propios del orden de 0,1 ; mientras que una matriz de medidas animales dará $\lambda_1 = 0{,}001$: el análisis no es menos significativo por esto ; la debilidad de los valores propios corresponde a la dispersión débil de los perfiles ; entre animales de una misma especie las diferencias de formas pueden ser reales e interpretables ; pero son de poca amplitud.

El lugar excéntrico de un punto sobre un diagrama puede provocar errores : la importancia del punto i para el eje α no se mide por $F_\alpha(i)$ sino por $f_i\, F_\alpha(i)^2$ pues tenemos :

$$\lambda_\alpha = \Sigma\{f_i\, F_\alpha(i)^2 \mid i \in I\} ;$$

Por esta razón, $f_i\, F_\alpha(i)^2$ es llamado : *contribución absoluta del punto i al factor* α. Igualmente, el que un punto i se aleje sobre el eje α no es suficiente para afirmar que el eje α explica por sí solo toda la originalidad de i , es decir toda la diferencia entre f_J^i y el perfil medio f_J , pues tenemos :

$$\| f_J^i - f_J \|^2 = \Sigma_\alpha\, F_\alpha(i)^2 ;$$

es por esto por lo que consideraremos a veces el cociente $F_\alpha(i)^2 / \| f_J^i - f_J \|^2$ llamado *contribución relativa del factor* α *al elemento i.*

Hemos dicho que la fórmula de transición permite situar sobre los ejes cualquier perfil. A menudo, a las nubes $N(I)$ y $N(J)$ añadiremos sobre los gráficos *elementos suplementarios.* Podrá ser el perfil medio de tal subconjunto señalado *a priori* : es más rápido ver que el perfil medio de los individuos machos tiene un factor F_2

netamente positivo, que hacer el recuento (cosa que se hará a continuación) de los 2/3 de machos (contra solamente 1/3 de hembras) que tienen un factor $F_2 \geq 0$. O también, una vez hecho el estudio, llegan nuevos datos, eventualmente menos seguros que los de la matriz principal : éstos constituirán elementos suplementarios. Finalmente, es frecuente que el primer análisis revele la ubicación muy excéntrica de algunos individuos, o de un carácter (tal vez difícil de medir con precisión) lo que perturba la visión del resto. Sin suprimir completamente los parásitos, construiremos los factores y los ejes como si aquellos no existieran, o tuvieran masa nula, y los volveremos a introducir en elementos suplementarios.

6. Campo de aplicación del análisis de correspondancias

No enumeraremos aquí les disciplinas muy diversas - lingüística, psicología, geología, historía...- que dan material para interesantes análisis. Aquí precisaremos solamente le estructura de las matrices que se pueden tratar con éxito antes de dar algunos ejemplos de análisis de datos lingüísticos.

El caso típico de las matrices de frecuencia nos ha sugerido un lenguage probabilistico y construcciones geométricas : de ahí un método de cálculo que desde el punto de vista de exigencias numéricas, acepta toda matriz de números positivos. Según el principio de equivalencia distribucional el análisis de correspondancias parece particularmente apto a tratar una matriz de medidas cuando hay indecisión para reagrupar las filas y las columnas : es el caso de las filas que son parcelas de terreno que se pueden unir o subdividir ; si las columnas son longitudes medidas sucesivamente sobre lineas entre marcas que se pueden elegir más o menos numerosas (nudos sobre un vegetal ; apofisis o suturas sobre un animal etc.).

Es aún más sorprendente que se puedan tratar matrices llenas no de candidades continuas, sino de 0 y de 1 (Si-No ; Presencia-Ausencia). Sea por ejemplo un cuestionario donde Q es el conjunto de preguntas : supongamos que para la Pregunta q hay un conjunto J_q de respuestas posibles ; y notemos J la unión de los J_q :

$J = \cup \{J_q \mid q \in Q\}$, es decir el conjunto de todas las respuestas posibles a todas las preguntas de Q (con 7 preguntas que no admiten cada una más que la respuesta Si o No, J no tiene más que 14 elementos ; pero en general una pregunta q admitirá más de dos modalidades de respuesta). Las respuestas relativas a un individuo i pueden ser codificadas por una fila de 0 y de 1 :

$k(i,j) = 1$ si i ha dado la respuesta j, y 0 si no ; es lo que llamamos : *Codificación bajo forma disyuntiva completa*. Es importante ver cuan general es este formato de matriz pues comprende no solamente las respuestas del hombre i que completa un cuestionario, y también cualquier ficha descriptiva de un objeto, construida mediante un conjunto de preguntas, tales como presencia o ausencia de escamas o de bracteas ; aptitud de metabolizar el ázoe atmosférico ; una medida continua puede ser también codificada como una pregunta si se subdivide, el intervalo de sus variaciones, en segmentos sucesivos. Por otra parte si el análisis de una matriz en (0,1) resulta provechoso es por que da los mismos factores sobre J que el de la matriz de frecuencia cuadrada $J \times J$ así definida :

k(j,j') = cantitades de i para los que vale simultáneamente k(i,j) = k(i,j') = 1. Por este lado volvemos al tipo inicial.

Dado que el principio de equivalencia distribucional permite esperar, en los resultados, una estabilidad que la experiencia ha confirmado, es importante fundar la recolección de datos sobre una base indiscutible. [Lo que se obtendrá respetando las exigencias de homogeneidad y de exhaustividad. Exhaustividad no quiere decir : medir todas las ratas, sino tomar un muestrario que no se haya elegido arbitrariamente ; ni tomar todas las medidas sino en caso de medir un cráneo determinar unas líneas principales y dividirlas por algunas señales de referencia netas entre las que se medirá]. En cuanto a la homogeneidad, se tratará de no mezlar longitudes con masas, ni tampoco masas de hormonas (ínfimas) con masas de aminoácidos. Pero por una parte la codificación de las medidas más diversas, como la de las respuestas a una pregunta, permite obtener una cierta homogeneidad a datos inaceptables de otra forma ; por otro lado cuando se quiere reunir en una misma matriz dos grupos J_1 y J_2 de columnas en la que cada una es homogénea, lo podemos hacer multiplicando una de las submatrices por un número elegido de tal forma que las contribuciones de J_1 y J_2 a la inercia total de la nube $N(J)$ sean del mismo orden : un cambio tal de escala permitirá por ejemplo juxtaponer una matriz de medidas del esqueleta a una matriz de balance alimentario (de animales alimentados en laboratorio y luego sacrificados). Se utiliza para esto un programa llamado de *ponderación* (A. Hamrouni).

Para finalizar, con el estado actual de las computadores y de los programas, el método de análisis de correspondancias puede tratar en un tiempo de cálculo aceptable los mayores conjuntos de datos que los investigadores normalmente reunen : por ejemplo una matriz 2000 x 150 (2000 individuos x 150 caracteres). El tratamiento de una matriz 10.000 x 1000 supondría ciertos problemas de cálculo (que no son completamente inabordables) ; pero según nosotros, la crítica previa del contenido de una matriz tan grande sería suficiente para desaconsejar el análisis!