Colección de Matemáticas

Análisis de datos con R

François Husson
Sébastien Lê
Jérôme Pagès

Escuela Colombiana de Ingeniería Julio Garavito
EDITORIAL

Prefacio

Qué es el análisis de datos

Tal como se trata el análisis de datos en Francia, y tal como se utiliza en este libro, la terminología «análisis de datos» reagrupa un conjunto de métodos estadísticos que se caracterizan por ser multidimensionales y descriptivas.

El propio término «multidimensional» engloba dos aspectos. En primer lugar, implica que las observaciones (o individuos estadísticos) son descritos por varias variables. En esta introducción nos restringimos a los datos más corrientes, en los que un conjunto de individuos es descrito por un conjunto de variables. Pero más allá de la disponibilidad de numerosas variables para cada individuo estadístico, es la voluntad de estudiarlos simultáneamente lo que caracteriza un enfoque multidimensional. De este modo, recurriremos al análisis de los datos cada vez que la noción de perfil sea pertinente para considerar un individuo ; por ejemplo, el perfil de las respuestas de los encuestados, el perfil biométrico de las plantas, el perfil financiero de las empresas, etc.

Desde un punto de vista dual, si es interesante considerar globalmente los valores de los individuos para un conjunto de variables es porque tales variables están vinculadas entre ellas. Observemos que el estudio sucesivo de todas las relaciones entre las variables tomadas dos a dos no constituye un enfoque multidimensional. Tal enfoque implica la toma en consideración simultánea del conjunto de las relaciones entre las variables tomadas dos a dos. Es lo que se hace, por ejemplo, en la puesta en evidencia de variables sintéticas : tal variable representa varias otras, lo que implica que esté vinculada a cada una de ellas, y ello es posible sólo si estas últimas están vinculadas entre ellas dos a dos. La noción de variable sintética es, pues, intrínsecamente multidimensional y un instrumento potente de descripción de una tabla individuos $\times$ variables. Desde estos dos puntos de vista, es un concepto clave del análisis de métodos multidimensionales y descriptivos.

Hagamos un último comentario sobre el término «análisis de datos» ya que posee por lo menos dos sentidos. El ya precisado y otro, más amplio, de investigación estadística. Este segundo sentido es una opinión del usuario ; es definido por un objetivo (analizar datos) y no estipula nada en cuanto a los métodos estadísticos puestos en marcha. Es lo que engloba el término anglosajón «data analysis». El término «análisis de datos», en el sentido de un conjunto de métodos descriptivos multidimensionales, es más un punto de vista francés en estadística. Jean-Paul Benzécri lo introdujo en Francia en los anos sesenta y su adopción está sin duda vinculada al hecho de que estos métodos multidimensionales son el centro de «data analyses».

A quién está dirigido este libro

Este libro se ha concebido para científicos que no se orientan hacia profesiones de la estadística pero que tendrán que tratar datos por ellos mismos. Está dirigido, pues, a los prácticos confrontados al análisis estadístico de datos. En dicha perspectiva, está orientado hacia las aplicaciones ; el formalismo matemático se ha reducido en lo posible, para dejar sitio a la comprensión a partir del ejemplo y a partir de la intuición. Concretamente, el nivel de una diplomatura científica es suficiente para apropiarse de todos los conceptos introducidos.
Sobre el plano del programa, una iniciación al lenguaje R es suficiente, por lo menos para comenzar. Este programa es gratuito y está disponible en internet en la siguiente dirección : `http://www.r-project.org/`.

Contenido y carácter del libro

El contenido del libra se centra en los cuatro métodos fundamentales del análisis de datos, los que tienen el potencial más vasto de aplicación : el análisis en componentes principales (ACP) cuando las variables son cuantitativas, el análisis factorial de las correspondencias (AFC) y el análisis de correspondencias múltiples (ACM) cuando los datos son cualitativos y la clasificación jerárquica ascendente. El punto de vista geométrico empleado para presentar estos métodos proporciona un contexto único en el sentido de que abastece una visión unificada para el análisis exploratorio de las tablas de datos. En este contexto, presentaremos los principios generales, indicadores, modos de representar y visualizar los objetos (filas y columnas de una tabla de datos) comunes a todos los métodos.
Así, veremos cómo es posible utilizar variables cualitativas en un contexto de ACP donde las variables que hay que analizar son clásicamente cuantitativas, al igual que añadir variables cuantitativas en un contexto de ACM donde las variables son cualitativas. Para cada método, el procedimiento adoptado es el mismo. Un ejemplo permite introducir la problemática y concreta casi paso a paso los elementos teóricos. Esta propuesta es seguida por varios ejemplos, tratados de un modo detallado para ilustrar el aporte del método en las aplicaciones.
A lo largo del texto, cada resultado es acompañado por el comando R que permite obtenerlo. Todos estos comandos son accesibles a partir de FactoMineR, paquete R desarrollado por los autores. El lector que lo desee podrá encontrar los resultados que figuran en este libro, puesto que los juegos de datos (así como las líneas de código) están disponibles en la siguiente dirección : `http://factominer.free.fr/libra`. Así, con esta obra, el lector dispone de un equipo completo (bases teóricas, ejemplos, programas) para analizar datos multidimensionales.

Antes de finalizar este prefacio, nos complace dar las gracias a Inmaculada Calviño Iglesias por la traducción de este libro y a Nuria Durán Adroher por su inestimable colaboración.

Contenido

Chapitre 1

Análisis de componentes principales (ACP)

1.1 Datos, notaciones y ejemplos

El análisis de componentes principales se aplica al cruce de tablas con individuos en fila y variables cuantitativas en columnas. Denotemos para x_{ik} el valor adquirido por el individuo i para la variable k ; i varía de 1 a I y k de 1 a K.
Llamamos $\bar{x}_k$ la media de la variable k, calculada sobre el conjunto I de los individuos :

$$\bar{x}_k = \frac{1}{I} \sum_{i=1}^{I} x_{ik},$$

y s_k la deviación típica de la variable k :

$$s_k = \sqrt{\frac{1}{I} \sum_{i=1}^{I} (x_{ik} - \bar{x}_k)^2}.$$

Los datos pueden ser de naturaleza diversa ; algunos ejemplos se presentan en la tabla 1.1.

Dominio	**Individuos**	**Variables**	x_{ik}
Ecología	Río	Concentración de contaminantes	Concentración del contaminante k en el río i
Economía	Año	Indicadores económicos	Valor del indicador k en el año i
Genética	Paciente	Genes	Expresión del gen k para el paciente i
Marketing	Marca	Índices de satisfacción	Valor del índice k para la marca i
Pedología	Suelo	Composición granulométrica	Índice del componente k para el suelo i
Biología	Animal	Medidas	Medida k para el animal i
Sociología	CSP	Presupuesto del tiempo	El tiempo pasado en la actividad k por los individuos de la CSP i

Tabla 1.1 – Descripción sumaria de algunas tablas de datos.

Ilustramos este capítulo de ACP con el juego de datos de los jugos de naranja, escogido por su sencillez, ya que sólo contiene seis individuos y siete variables. Estos datos se han obtenido en el ámbito de trabajos de estudiantes del Agrocampus. Un jurado integrado por estudiantes evaluó los seis jugos de naranja según siete variables sensoriales (intensidad del olor, tipo de olor, carácter pulposo, intensidad del sabor, carácter ácido, carácter amargo, carácter azucarado). Los promedios de las evaluaciones del jurado aparecen seguidamente (tabla 1.2).

	Intensidad olor	Tipo olor	Pulposo	Intensidad sabor	Carácter ácido	Carácter amargo	Carácter azucarado
Pampryl amb.	2.82	2.53	1.66	3.46	3.15	2.97	2.60
Tropicana amb.	2.76	2.82	1.91	3.23	2.55	2.08	3.32
Fruvita fr.	2.83	2.88	4.00	3.45	2.42	1.76	3.38
Joker amb.	2.76	2.59	1.66	3.37	3.05	2.56	2.80
Tropicana fr.	3.20	3.02	3.69	3.12	2.33	1.97	3.34
Pampryl fr.	3.07	2.73	3.34	3.54	3.31	2.63	2.90

Tabla 1.2 – Datos de los jugos de naranja.

1.2 Objetivos

La tabla de datos puede analizarse por sus filas (los individuos) o a través de sus columnas (las variables), lo que suscita varios tipos de preguntas relacionadas con estos objetos de diferente naturaleza.

1.2.1 Estudio de los individuos

A continuación se ilustra el tipo de preguntas formuladas en el momento del estudio de individuos (figura 1.1). Se representan tres situaciones en las cuales 40 individuos son descritos

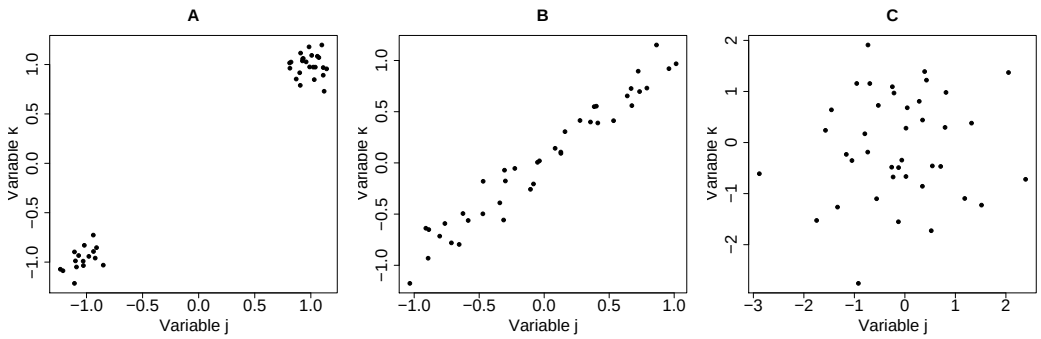

FIGURE 1.1 – Representación de 40 individuos descritos por dos variables (j y k).

por dos variables (j y k). En el gráfico A se ponen en evidencia dos clases muy distintas de individuos, en tanto que en el gráfico B se presenta una dimensión de variabilidad que

opone a individuos extremos (como en el caso del gráfico A) pero esta vez con los individuos intermedios. La forma de la nube de individuos es aquí muy alargada. En el gráfico C se muestra una nube informe (por ejemplo, sin estructura particular).
Es fácil describir los datos en estos ejemplos simples porque están en dos dimensiones. Cuando los individuos son descritos por un gran número de variables, es necesario disponer de un instrumento para explorar el espacio en el cual evolucionan. El estudio de los individuos consiste en aprehender las semejanzas entre individuos desde el punto de vista del conjunto de las variables, es decir establecer una tipología de individuos : ¿cuáles son los individuos más próximos (resp. más alejados) ? ¿Existen grupos de individuos homogéneos desde el punto de vista de sus semejanzas ? Otro aspecto consiste en buscar dimensiones comunes de variabilidad que oponen individuos extremos a individuos intermedios.
En el ejemplo, dos jugos de naranja se han evaluado del mismo modo sobre el conjunto de la descripción sensorial. Decimos entonces que ambos jugos tienen el mismo «perfil» sensorial. De modo más general, nos preguntamos si existen unos grupos de jugos de naranja con perfiles similares, por ejemplo, de dimensiones sensoriales que pueden oponer jugos extremos a jugos intermedios.

1.2.2 Estudio de variables

Paralelamente al estudio de individuos, ¿podemos describir los datos a partir de las variables ? El ACP se centra en las relaciones lineales entre variables. Existen relaciones más complejas, como las relaciones cuadráticas, logarítmicas, exponenciales, etc., pero no se estudian en el ACP. Esto puede parecer restrictivo, pero en la práctica numerosas relaciones pueden considerarse lineales, por lo menos en una primera aproximación.
Tomemos el ejemplo de las cuatro variables (j, k, l y m) de la figura 1.2. Las nubes de puntos construidas con las variables dos a dos muestran que las variables j y k (gráfico A), así como las variables l y m (gráfico F), están muy correlacionadas (positivamente para j y k y negativamente para l y m). En cambio, otros gráficos no muestran ninguna relación. El estudio de estos gráficos sugiere también que las cuatro variables se repartan en dos grupos de dos variables (j, k) y (l, m) tales que, dentro de un grupo, las variables están estrechamente correlacionadas y de un grupo al otro, las variables no están correlacionadas. La construcción de grupos de variables es útil en una óptica de síntesis, exactamente como sucede en la construcción de grupos de individuos ; para los individuos se puede encontrar un continuo con grupos muy particulares de variables y variables intermedias, un poco vinculadas a cada uno de los grupos. En el ejemplo, cada grupo puede estar representado por una sola variable porque las variables del mismo grupo están muy correlacionadas. A estas variables las llamamos variables sintéticas.
Cuando el número de variables es escaso, se puede hacer el balance a partir de las nubes de puntos o de la matriz de correlación que reagrupa el conjunto de los coeficientes de correlación lineal $r(j, k)$ entre las variables tomadas dos a dos ; sin embargo, si el número de variables es importante, la matriz de correlación reagrupa muchos coeficientes de correlación (190 coeficientes para $K = 20$ variables). Es entonces indispensable tener una herramienta que proporcione una visualización sintética de las principales relaciones entre las variables. El objeto del ACP es hacer un balance de las relaciones lineales entre variables, detectando las principales dimensiones de variabilidad. Veremos que este balance se completará con la

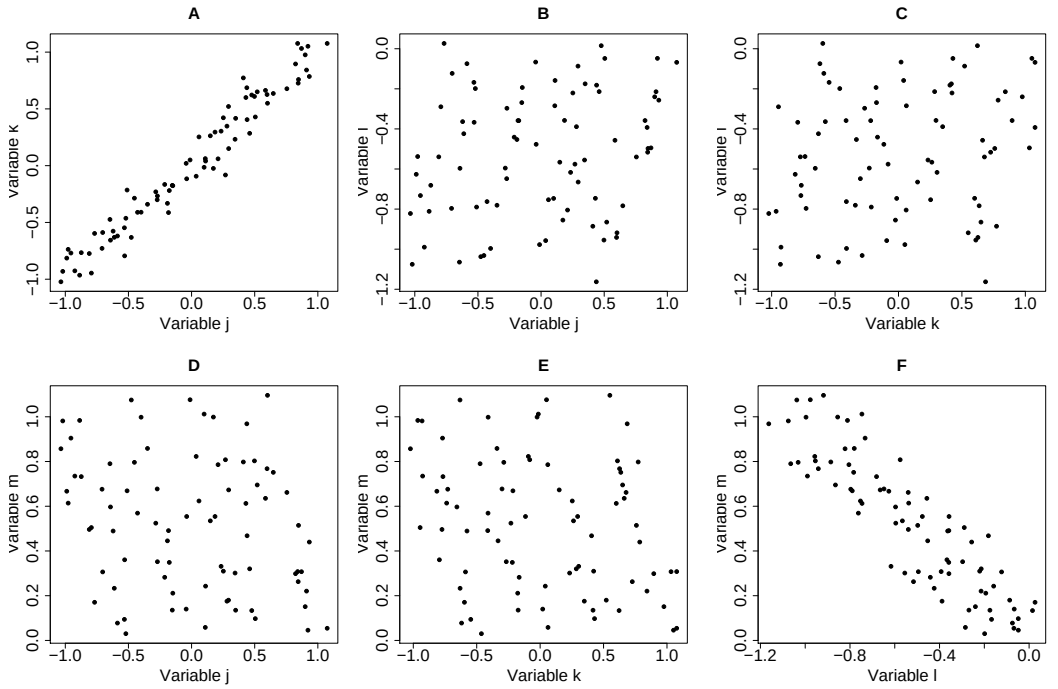

FIGURE 1.2 – Representación de las relaciones entre cuatro variables (j, k, l y m) dos a dos.

definición de variables sintéticas propuestas por el ACP, por lo que será más fácil comentar los datos por algunas variables sintéticas más bien que por el conjunto de las variables.

En el ejemplo de los zumos de naranja, la matriz de correlación (cf. tabla 1.3) reagrupa los 21 coeficientes de correlación. Se pueden reagrupar las variables muy correlacionades por paquetes pero, incluso para este número reducido de variables, esta reagrupación es fastidiosa.

	Intensidad olor	Tipo olor	Pulposo	Intensidad sabor	Carácter ácido	Carácter amargo	Carácter azucarado
Intensidad olor	1,00	0,58	0,66	−0,27	−0,15	−0,15	0,23
Tipo olor	0,58	1,00	0,77	−0,62	−0,84	−0,88	0,92
Carácter pulposo	0,66	0,77	1,00	−0,02	−0,47	−0,64	0,63
Intensidad sabor	−0,27	−0,62	−0,02	1,00	0,73	0,51	−0,57
Carácter acido	−0,15	−0,84	−0,47	0,73	1,00	0,91	−0,90
Carácter amargo	−0,15	−0,88	−0,64	0,51	0,91	1,00	−0,98
Carácter azucarado	0,23	0,92	0,63	−0,57	−0,90	−0,98	1,00

Tabla 1.3 – Datos de los jugos de naranja : matriz de correlación.

1.2.3 Relación entre ambos estudios

El estudio de los individuos y el estudio de las variables están vinculados ya que se realizan sobre la misma tabla de datos. Confrontarlos refuerza su interpretación respectiva.
Si el estudio de individuos permitió distinguir grupos de individuos, se puede poner en una lista los individuos que pertenecen al mismo grupo. Sin embargo, cuando el número de individuos es importante, es preferible caracterizarlos por las variables de la tabla : por ejemplo, precisando que ciertos zumos de naranja son a la vez ácidos, amargos y no pulposos y otros presentan las características inversas.
Del mismo, modo cuando existen grupos de variables, no es fácil interpretar la relación entre las múltiples variables y podemos valernos de individuos-tipos, *i.e.*, de individuos que son extremos desde el punto de vista de estas relaciones. Para ello, hay que conocer bien los individuos. Por ejemplo, la relación entre las variables *ácido-amargo* puede ser ilustrada por la oposición entre dos zumos extremos de naranja : *Pampryl fresco* (percibido a la vez ácido y amargo) contra *Tropicana fresco* (percibido a la vez poco ácido y poco amargo).

1.3 Estudio de individuos

1.3.1 Nube de individuos

Un individuo corresponde a una fila de la tabla, *i.e* un conjunto de K valores numéricos. Los individuos evolucionan pues en el espacio $\mathbb{R}^K$ llamado «espacio de individuos». Si se provee este espacio de la distancia euclidiana usual, la distancia entre dos individuos i y l se escribe :

$$d(i,l) = \sqrt{\sum_{k=1}^{K}(x_{ik} - x_{lk})^2}.$$

Si dos individuos tienen valores próximos en la tabla sobre el conjunto de K variables, entonces son próximos en el espacio $\mathbb{R}^K$. Así, el estudio de la tabla de datos puede ser realizado geométricamente vía el estudio de las distancias entre individuos. Nos interesamos entonces al conjunto de individuos en $\mathbb{R}^K$, *i.e.*, a la nube de individuos (denominado N_I). El análisis de las distancias entre individuos vuelve a estudiar la forma de la nube de puntos. La figura 1.3 ilustra una nube de puntos en el espacio $\mathbb{R}^K$ para $K = 3$.
La forma de la nube N_I no varia aunque se traslade la nube. También centramos los datos, lo que vuelve a considerar $x_{ik} - \bar{x}_k$ en lugar de x_{ik}. Geométricamente, esto vuelve a hacer coincidir el baricentro de la nube G_I (de coordenadas $\bar{x}_k$ para $k = 1, K$) con el origen de la indicación (cf. figura 1.4). El centrado presenta ventajas técnicas y siempre es realizado en el ACP.
La operación de reducción (hablamos también de estandarización), que vuelve a considerar $(x_{ik} - \bar{x}_k)/s_k$ en lugar de x_{ik}, modifica la forma de la nube armonizando su variabilidad en todas las direcciones de los vectores de base. Geométricamente, consiste en escoger la desviación-tipo s_k como unidad de medida en la dirección k. Esta operación es indispensable en el caso en el que las variables no se expresan en las mismas unidades de medida. Incluso fuera de estos casos, se recomienda esta operación ya que vuelve a conceder la misma importancia a cada variable. Más tarde, nos ocupamos de este caso. Hablamos de

FIGURE 1.3 – Vuelo de estorninos que ilustra una nube de puntos en $\mathbb{R}^K$.

ACP normado cuando las variables son centradas y reducidas de ACP no normado cuando las variables son únicamente centradas ; cuando no hay ninguna precisión, es sobreentiende un ACP normado.

Observación sobre el peso de individuos. Hasta aquí, supusimos que todos los individuos tenían el mismo peso. Es el caso de casi la totalidad de las aplicaciones y lo supondremos siempre. Sin embargo, la generalización a individuos con distinta ponderación no plantea ningún problema conceptual (un peso doble es equivalente a dos individuos idénticos) ni práctico (la inmensa mayoría de los programas, incluyendo FactoMineR, preveen esta posibilidad). Puede ser útil destinar un peso diferente a cada individuo después de haber establecido una muestra por ejemplo. De todos modos, es cómodo hacer la suma de los pesos igual a 1. Más tarde, ya que se supondrá que tienen el mismo peso, cada individuo será afectado del peso $1/I$.

1.3.2 Ajuste de la nube de individuos

Mejor representación plana de N_I

La idea del ACP es de representar la nube de puntos en un espacio de dimensión reducida de un modo «óptimo», es decir, deformando lo menos posible las distancias entre individuos. La figura 1.5 proporciona dos representaciones de tres frutas diferentes. Los enfoques utilizados

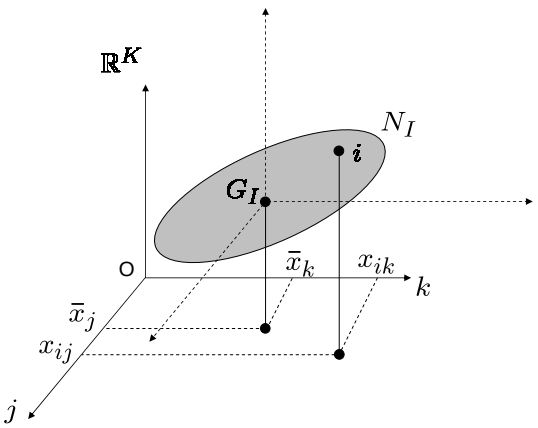

FIGURE 1.4 – La nube de los individuos en $\mathbb{R}^K$.

para fotografiar las frutas en la primera fila permiten difícilmente identificar cada fruta. En la segunda fila, las frutas se reconocen más fácilmente. ¿Qué es lo que diferencia los enfoques de la misma fruta entre la primera fila y la segunda ? Las distancias están menos deformadas en los segundos enfoques y las representaciones ocupan mejor el espacio en la fotografía. La fotografía proyectó un objeto tridimensional en un espacio de dos dimensiones.

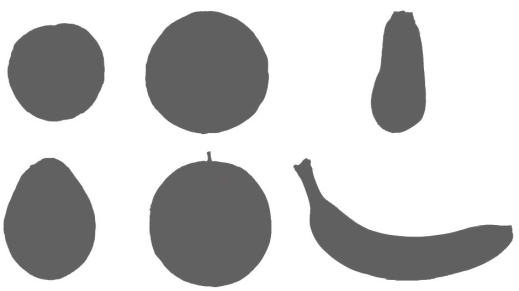

FIGURE 1.5 – Representación de frutas en dos dimensiones : de izquierda a derecha un aguacate, un melón y un plátano ; cada fila corresponde de un tipo de representación.

Una buena representación deberá escoger un buen ángulo de enfoque ; generalmente, el ACP vuelve a buscar el mejor espacio de representación (de dimensión reducida) que permite visualizar lo mejor posible la forma de una nube de K dimensiones. A menudo nos contentamos con una representación plana que puede resultar insuficiente para ciertos juegos de datos particularmente ricos.
Para obtener la mejor representación plana, la nube N_I se proyecta sobre un plano de $\mathbb{R}^K$, escrito P, escogido tal que deforme lo menos posible la nube de puntos. El plano P es de manera que las distancias entre puntos proyectados sean tan próximas como las distancias entre puntos iniciales. Como en proyección una distancia sólo puede disminuir, procuramos hacer las distancias proyectadas lo más grande posible. Escribiendo H_i la proyección del

individuo i sobre el plano P, el problema vuelve a ser encontrar P de manera que :

$$\sum_{i=1}^{I} OH_i^2 \text{ sea máximo.}$$

Este criterio consiste en hacer máximo la varianza de los puntos proyectados. Hablamos así de «varianza explicada». En Francia, utilizamos más bien el lenguaje de la mecánica : O siendo el centro de gravedad de la nube, el criterio puede verse como la inercia de la proyección de N_I.

Observación
Si los individuos tienen pesos diferentes p_i, el criterio a maximizar es $\sum_{i=1}^{I} p_i OH_i^2$.

En ciertos casos, poco frecuentes, puede ser interesante buscar sólo la mejor representación axial de la nube N_I. Este mejor eje se obtiene según el mismo principio : encontrar el eje u_1 tal que $\sum_{i=1}^{I} OH_i^2$ sea máximo (con H_i la proyección de i sobre u_1). Podemos demostrar que el plano P contiene el eje u_1 («mejor» plano contiene el «mejor» eje) : en este sentido, estas dos representaciones encajan. Una ilustración de esta propiedad es presentada en la figura 1.6 : los planetas, que están en un espacio de tres dimensiones, clásicamente se representan sobre un eje, en el cual se sitúan de la mejor manera posible en función de la distancia que separa unos de otros (en términos de inercia de la nube proyectada). También podemos representar los planetas sobre un plano siguiendo el mismo principio : maximizar la inercia de la nube de puntos proyectada (sobre el plano). Esta mejor representación plana contiene la mejor representación axial.

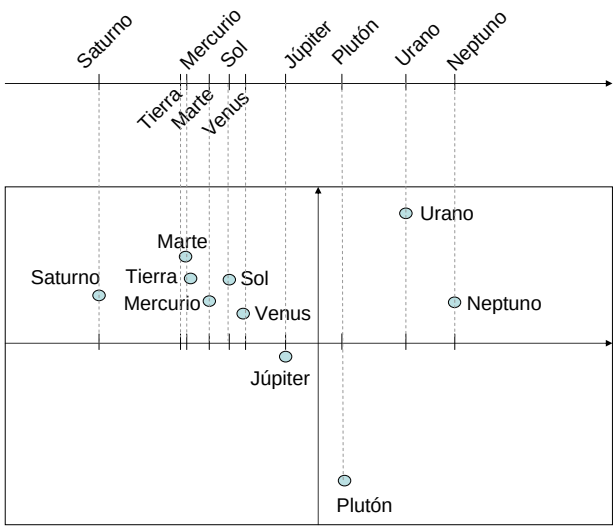

FIGURE 1.6 – La mejor representación axial se incluye en la mejor representación plana. Ejemplo de la posición de los planetas en el sistema solar (el 18 de febrero de 2008).

Definimos el plano P por dos vectores no colineales escogidos del modo siguiente : el vector u_1 que define el mejor eje (y que está incluido en P), el vector u_2 del plano P ortogonal a u_1. El vector u_2 corresponde al vector que expresa más variabilidad de N_I una vez retirada la expresada por u_1. Es decir, la variabilidad expresada por u_2 es el mejor complemento y es independiente de la variabilidad expresada por u_1.

Continuación de ejes de representación de N_I

Generalmente, podemos buscar subespacios encajados por dimensiones de $s = 1$ a S de tal modo que cada subespacio es de inercia máxima para la dimensión s dada. El subespacio de dimensión s se obtiene maximizando $\sum_{i=1}^{I} (OH_i)^2$ (con H_i la proyección de i sobre el subespacio de dimensión s). Como los subespacios están encajados, se puede escoger el vector u_s como el vector del subespacio ortogonal de todos los vectores u_t (con $1 \leq t < s$) que definen los subespacios de dimensión inferior.
El primer plano (definido por u_1, u_2), *i.e.*, la mejor representación plana, es a menudo suficiente para visualizar la nube N_I. Cuando S es superior o igual a 3, tenemos que visualizar la nube N_I en el subespacio de dimensión S con la ayuda de varias representaciones planas : la representación sobre (u_1, u_2) pero también la representación sobre (u_3, u_4) que es el mejor complemento de la representación sobre (u_1, u_2). Sin embargo, en ciertas situaciones, podemos escoger asociar por ejemplo (u_2, u_3) para poner en evidencia un fenómeno particular que aparece en estos dos ejes (cf. el ejemplo sobre las defunciones § 2.10 p. 101).

¿Cómo obtener los ejes ?

Los ejes del ACP se obtienen por la diagonalización de la matriz de correlación que extrae los vectores propios y los valores propios asociados. Los vectores propios corresponden a los vectores u_s asociados cada uno al valor propio de rango s (escrito λ_s), los valores propios ordenados por orden decreciente. El valor propio λ_s se interpreta como la inercia de la nube N_I proyectada sobre el eje de rango s, es decir, como «varianza explicada» por el eje de rango s. Si todos los vectores propios son calculados ($S = K$), entonces el ACP reconstituye una base del espacio $\mathbb{R}^K$. En este sentido, el ACP puede ser visto como un cambio de base en el cual los primeros vectores de la nueva base desempeñan un papel privilegiado.

Observación
Cuando las variables no son reducidas, la matriz diagonalizada es la matriz de varianza-covarianza.

Ejemplo

La distancia entre dos zumos de naranja se calcula tomando como base los siete descriptores sensoriales. Decidimos estandarizar los datos para otorgar la misma influencia a cada descriptor. La figura 1.7 se obtiene a partir de los dos primeros componentes del ACP normado y corresponde pues al el mejor plano de representación de la nube de puntos en el sentido de la inercia proyectada. La inercia proyectada sobre este plano corresponde a la suma de los dos primeros valores propios dividida por la suma de los valores propios, es decir a 86.82% ($= 67.77\ \% + 19.05\ \%$) de la inercia total de la nube de puntos.

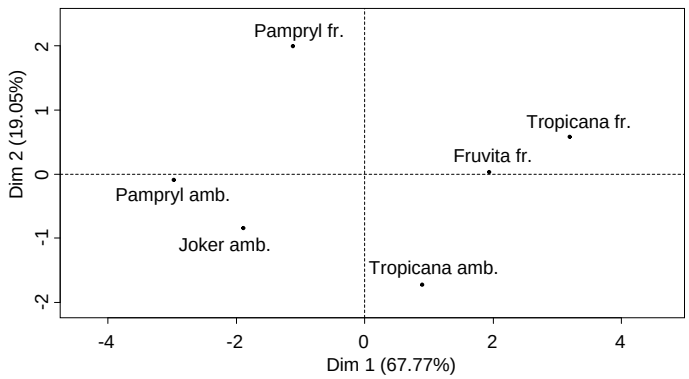

FIGURE 1.7 – Datos del zumo del zumo de naranja : representación plana de la nube de individuos.

El principal eje de variabilidad entre los zumos de naranja, opone *Tropicana fresco* y *Pampryl ambiente*. Según la tabla de datos 1.2, podemos ver que estos zumos de naranja son los más extremos para los descriptores *tipo olor* y *amargo* : *Tropicana fresco* es el zumo de naranja más típico y menos amargo mientras que *Pampryl ambiente* es el menos típico y más amargo. El segundo componente, el que más opone los zumos de naranja una vez el principal eje de variabilidad retirado, separa *Tropicana ambiente*, que es el menos intenso desde el punto de vista del olor, de *Pampryl fresco* entre los más intensos (cf. tabla 1.2).
Esta lectura de los datos es fastidiosa cuando el número de individuos y el número de variables es considerable. Facilitamos la caracterización de los principales ejes con la ayuda de las variables de un modo más directo.

1.3.3 Representación de las variables como ayuda a la interpretación de la nube de individuos

Denotemos $F_s(i)$ la coordenada del individuo i sobre el eje s y F_s el vector de las coordenadas I de los individuos sobre el eje, llamado también componente principal. El vector F_s es de dimensión I y puede así ser asignado a una variable. Para interpretar las posiciones relativas de los individuos sobre el eje del rango s, puede ser interesante calcular los coeficientes de correlación entre el vector F_s y las variables iniciales. Así, cuando el coeficiente de correlación entre F_s y una variable k es positivo (resp. negativo), un individuo que tiene una coordenada positiva sobre el eje F_s, generalmente posee un fuerte (resp. debilidad) valor para la variable k (respecto a la media).
En el ejemplo, F_1 está muy correlada positivamente con las variables *tipo olor* y *azucarado* y muy correlada negativamente con las variables *amargo* y *ácido* (cf. tabla 1.4). Así, *Tropicana fresco*, que tiene la coordenada más fuerte en el eje 1, tiene valores fuertes para el tipo de olor y el azucarado y valores débiles para las variables *ácido* y *amargo*. De la misma manera, podemos interesarnos a las correlaciones entre F_2 y las variables. Podemos anotar que las correlaciones son globalmente más débiles (en valor absoluto) que las correlaciones con el

factor 1. Veremos que esto está relacionado con el porcentaje de inercia asociado a F_2, que por construcción, es inferior al asociado a F_1. El segundo eje puede caracterizarse por las variables *intensidad olor* y *pulposo* (cf. tabla 1.4).

	F_1	F_2
Intensidad.olor	0.46	0.75
Tipo.olor	0.99	0.13
Pulposo	0.72	0.62
Intensidad.sabor	−0.65	0.43
Ácido	−0.91	0.35
Amargo	−0.93	0.19
Azucarado	0.95	−0.16

Tabla 1.4 – Datos zumo de naranja : coeficientes de correlación entre cada una de las variables y los dos primeros componentes principales (las coordenadas de los individuos sobre los dos primeros ejes).

Para facilitar la lectura de estos resultados, particularmente cuando el número de variables es elevado, representamos cada variable sobre un gráfico a partir de sus coeficientes de correlación con F_1 y F_2 que utilizamos como coordenadas (cf. figura 1.8).

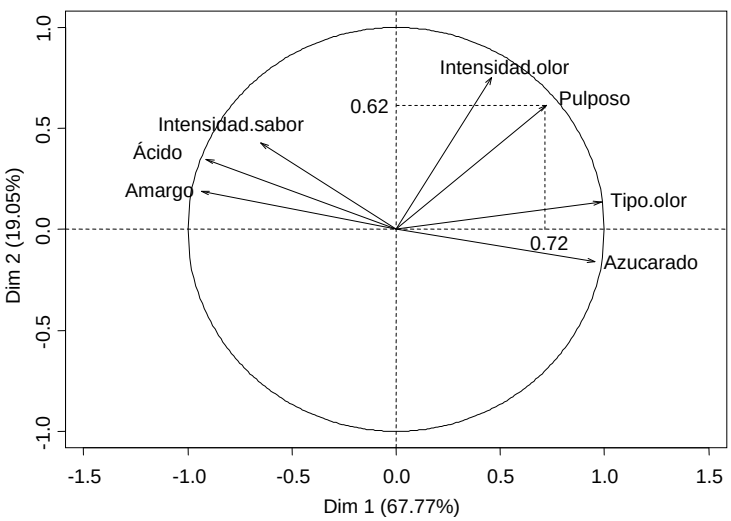

FIGURE 1.8 – Datos zumo de naranja : visualización de coeficientes de correlación entre variables y componentes principales F_1 y F_2.

Podemos ahora interpretar la representación de la nube de individuos con la ayuda de esta representación de variables.

Observación
La representación de una variable está dentro de un círculo de radio 1 (círculo representado sobre la figura 1.8) : en efecto, recordemos que F_1 y F_2 por construcción son ortogonales (*i.e.*,

coeficiente de correlación igual a 0) y que una variable no puede estar vinculada fuertemente y simultáneamente a dos ejes ortogonales. Veremos más precisamente en la sección siguiente por qué la variable está necesariamente dentro del círculo de radio 1.

1.4 Estudio de variables

1.4.1 Nube de variables

Consideremos ahora la tabla de datos a través de variables. Una variable es una columna de la tabla, *i.e.*, un conjunto de I valores numéricos, asimilable a un vector de dimensión I evolucionando en un espacio vectorial de I dimensiones denotado $\mathbb{R}^I$ (y llamado «espacio de variables»). El conjunto de estos vectores constituye la nube de variables y se escribe N_K (cf. figura 1.9).

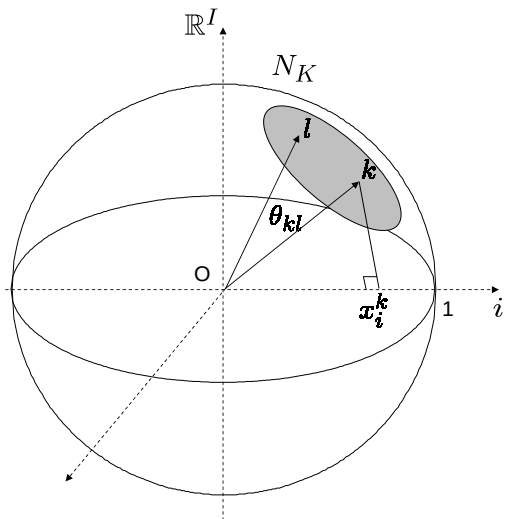

FIGURE 1.9 – La nube N_K de variables en $\mathbb{R}^I$. En ACP normado, las k variables están situadas sobre la hiperesfera de radio 1.

El producto escalar entre dos variables k y l se escribe :

$$\sum_{i=1}^{I} x_{ik} \times x_{il} = \|k\| \times \|l\| \times \cos(\theta_{kl}).$$

con $\|k\|$ (resp. $\|l\|$) la norma de la variable k (resp. l) y θ_{kl} el ángulo formado por los vectores que representan las variables k y l. Como aquí las variables están centradas, la norma de una variable es igual a su desviación-típica multiplicada por la raíz de I y el producto escalar se escribe :

$$\sum_{i=1}^{I} (x_{ik} - \bar{x}_k) \times (x_{il} - \bar{x}_l) = I \times s_k \times s_l \times \cos(\theta_{kl}).$$

Reconocemos en el término de una recta la covarianza entre las variables k y l. También, dividiendo cada término de la ecuación por las desviaciones-típicas s_k y s_l de las variables k y l, obtenemos la relación siguiente :

$$r(k,l) = \cos(\theta_{kl})$$

Esta propiedad es crucial en ACP porque ofrece una interpretación geométrica de la correlación. Así, de la misma manera que la representación de la nube N_I permite visualizar la variabilidad entre los individuos, una representación de la nube N_K permite visualizar el conjunto de las correlaciones (vía los ángulos entre variables), es decir la matriz de correlación. Para poder visualizar más fácilmente los ángulos entre las variables, representamos las variables no por puntos pero por vectores. En general, caso que adoptamos, estando las variables centradas y reducidas, sus longitudes valen 1 (de ahí el nombre de variable normada). Su extremidad se sitúa entonces sobre la esfera (decimos también hiperesfera para recordar que, en general, $I > 3$) de radio 1, que se esquematiza en la figura 1.9.

1.4.2 Ajuste de la nube de variables

Así como para los individuos, la nube de variables N_K está en un espacio $\mathbb{R}^I$ de dimensión elevada y no es posible visualizar la nube en el espacio completo. Por eso es necesario ajustar la nube de variables y para ello podemos utilizar la misma estrategia que para el ajuste de la nube de individuos. Maximizamos un criterio equivalente $\sum_{k=1}^{K} (OH_k)^2$ con H_k la proyección de la variable k sobre el subespacio de dimensión reducida. Aquí, los subespacios están encajados y podemos encontrar una continuación de S ejes ortogonales que definen los subespacios de dimensiones $s = 1$ a S. El vector v_s pertenece al subespacio y es ortogonal a los vectores v_t que componen los subespacios de dimensión inferior. Podemos entonces mostrar que el vector v_s maximiza $\sum_{k=1}^{K} (OH_k^s)^2$ con H_k^s la proyección de la variable k sobre v_s.

Observación
En el espacio de los individuos $\mathbb{R}^K$, el hecho de centrar las variables desplaza el origen de los ejes sobre el punto medio : el criterio maximizado se interpreta entonces como una varianza ; los puntos proyectados deben estar lo más dispersados posible. En $\mathbb{R}^I$, el centrado no tiene el mismo efecto porque el origen no se confunde con el punto medio : los puntos proyectados deben estar lo más lejos posible del origen (y no necesariamente dispersos) con el riesgo de que estén reagrupados e incluso confundidos ; esto significa que la posición de la nube N_K con relación al origen es importante.

Los vectores v_s $(s = 1, ..., S)$ pertenecen al espacio $\mathbb{R}^I$ y por consiguiente pueden ser considerados como nuevas variables. Así, el coeficiente de correlación $r(k, v_s)$ entre la variable k y v_s es igual al coseno del ángulo θ_k^s entre k y v_s si la variable k está centrada-reducida, y entonces normada. La representación de las variables sobre el plano formado por (v_1, v_2) es atractiva porque las coordenadas de una variable k corresponden al coseno del ángulo θ_k^1 y al coseno del ángulo θ_k^2 y como consecuencia,corresponden a los coeficientes de correlación entre la variable k y v_1 y entre la variable k y v_2. Sobre tal representación plana, es fácil ver si una variable k está vinculada a una dimensión de variabilidad.

Por construcción, la variable v_s maximiza el criterio $\sum_{k=1}^{K} (OH_k^s)^2$. Como la proyección de una variable k es igual al coseno del ángulo θ_k^s, el criterio vuelve a maximizar :

$$\sum_{k=1}^{K} \cos^2 \theta_k^s = \sum_{k=1}^{K} r^2(k, v_s).$$

La última expresión muestra que v_s es la nueva variable más correlada al conjunto de K variables iniciales (con la condición de ortogonalidad a los v_t ya encontrados). En este sentido, v_s es una variable sintética. Encontramos aquí el segundo aspecto del estudio de las variables (cf. § 1.2.2).

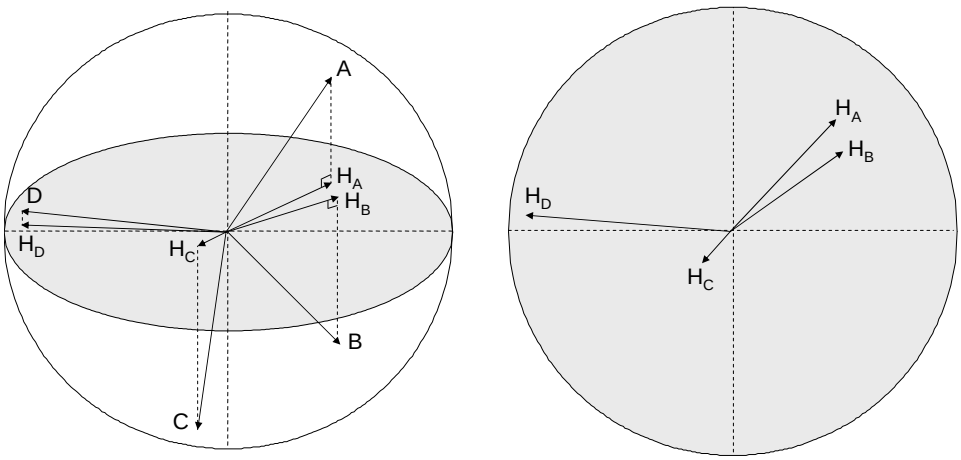

FIGURE 1.10 – Proyección de la nube de variables sobre el primer plano. A la izquierda : visualización en el espacio $\mathbb{R}^I$; a la derecha : visualización de proyecciones en el primer plano.

Observación
Cuando una variable no es normada, su longitud es igual a su desviación-típica. En ACP no normada, el criterio se escribe así para el vector v_s :

$$\sum_{k=1}^{K} (OH_k^s)^2 = \sum_{k=1}^{K} s_k^2 \, r^2 \, (k, v_s) \, .$$

A cada variable k se le asigna un peso igual a su varianza s_k^2.
Podemos mostrar que los ejes de representación de N_K son vectores propios de la matriz de los productos escalares entre individuos. Esta propiedad se utiliza en la práctica únicamente cuando el número de variables es superior al número de individuos. Veremos en el párrafo siguiente que estos vectores propios se deducen de los de la matriz de correlación.

La mejor representación plana de la nube de variables corresponde exactamente al gráfico de la representación de las variables obtenido como ayuda en la interpretación de la representación de los individuos (cf. figura 1.8). Esta propiedad notable no es específica al ejemplo pero vale en cuanto se efectúa un ACP normado, lo que desarrollamos en la sección siguiente.

1.5 Relación entre las representaciones de las nubes N_I y N_K

Las representaciones de las nubes N_I y N_K se obtienen según el mismo principio y a partir de la misma tabla de datos. Por eso es lógico que existan relaciones entre los dos análisis (el de N_I en $\mathbb{R}^K$ y el de N_K en $\mathbb{R}^I$).
Estas relaciones entre ambas nubes N_I y N_K son reagrupadas bajo el término general de relaciones de dualidad haciendo referencia a la doble manera de ver la tabla : teniendo en cuenta las filas o las columnas. Encontramos también el término de «relaciones de transición» (evidentemente de un espacio al otro). Escribiendo $F_s(i)$ la coordenada del individuo i y $G_s(k)$ la coordenada de la variable k sobre el eje de rango s, tenemos las ecuaciones siguientes :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^{K} x_{ik}\ G_s(k),$$

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} (1/I)\ x_{ik}\ F_s(i).$$

Este resultado es esencial para la interpretación y hace del ACP un instrumento de investigación de calidad y sólido. Podemos leerlo del modo siguiente : un individuo está situado del lado de las variables para las cuales toma valores fuertes y en oposición a las variables para las cuales toma pequeños valores. Recordemos que los x_{ik} están centrados, y tienen valores positivos y negativos ; de ahí el alejamiento de un individuo con relación a una variable para la cual tiene un valor débil. F_s es el componente principal de rango s ; λ_s es la varianza de F_s y su raíz es la longitud de F_s en $\mathbb{R}^I$; $v_s = F_s/\sqrt{\lambda_s}$ denominade componente principal normado.
Las inercias totales de ambas nubes son iguales (e igual a K si el ACP es normado). Además, sus descomposiciones eje por eje son idénticas. Esta propiedad es notable : si S dimensiones bastan para representar perfectamente N_I, lo mismo ocurre para N_K. Sino, ¿que podemos decir de una variable sintética suplementaria que no diferenciaría los individuos ?

1.6 Ayudas a la interpretación

1.6.1 Indicadores numéricos

Porcentaje de inercia asociado a un eje

Los primeros indicadores que consultamos dan la inercia proyectada sobre la inercia total. Es decir, para el eje s :

$$\frac{\sum_{i=1}^{I} \frac{1}{I} \left(OH_i^s\right)^2}{\sum_{i=1}^{I} \frac{1}{I} \left(Oi\right)^2} = \frac{\sum_{k=1}^{K} \left(OH_k^s\right)^2}{\sum_{k=1}^{K} Ok^2} = \frac{\lambda_s}{\sum_{s=1}^{K} \lambda_s}.$$

Y si el ACP es normado, $\sum_{s=1}^{K} \lambda_s = K$. Multiplicado por 100, este indicador representa el porcentaje de inercia (de N_I en $\mathbb{R}^K$ o de N_K en $\mathbb{R}^I$) expresado por el eje de rango s. Este porcentaje puede verse de dos maneras :

– como una medidad de la calidad de representación de datos ; en el ejemplo, diremos que el primer eje expresa 67.77 % de la variabilidad de los datos (cf. tabla 1.5). En ACP normado (con $I > K$), a menudo comparamos λ_s a 1, valor por debajo del cual el eje de rango s representa entonces menos datos que una variable aislada y no es digno de interés ;

– como una medida de la importancia relativa de los ejes ; en el ejemplo, diremos que el primer eje expresa tres veces más de variabilidad que el segundo ; en efecto, concierne tres veces más variables pero esta formulación es verdaderamente exacta sólo cuando cada variable está correlada perfectamente a un eje.

A causa de la ortogonalidad de los ejes entre ellos (tanto en $\mathbb{R}^K$ como en $\mathbb{R}^I$), estos porcentajes de inercia se suman para varios ejes.

	Valor propio	Porcentaje de inercia	Porcentaje de inercia acumulada
comp. 1	4.74	67.77	67.77
comp. 2	1.33	19.05	86.81
comp. 3	0.82	11.71	98.53
comp. 4	0.08	1.20	99.73
comp. 5	0.02	0.27	100.00

Tabla 1.5 – Datos de los jugos de naranja : descomposición de la variabilidad por ejes.

Volvamos a la figura 1.5 : las fotografías de las frutas de la primera fila corresponden aproximadamente a una proyección de las frutas sobre el plano formado por los ejes 2 y 3 del ACP mientras que las fotografías de la segunda fila corresponde aproximadamente a una proyección sobre el plano 1-2. Es por esta razón que las frutas son más fáciles de reconocer en la segunda fila : más variabilidad (*i.e.*, más información) es recuperada sobre el plano 1-2 con relación al plano 2-3 y es más fácil aprehender la forma global de la nube. Además, el plátano es más reconocible que el melón sobre el plano 1-2 (en la segunda fila) porque la parte de inercia recuperada por el plano 1-2 es más importante. En efecto, el plátano es una fruta más alargada que el melón, lo que conlleva diferencias de inercia de un eje al otro más marcadas. El melón, al ser casi esférico, los porcentajes de inercia asociados a cada uno de los tres ejes son próximos de 33 % y la parte de inercia recuperada por el plano 1-2 es próxima de 66 % (como la recuperada por el plano 2-3).

Calidad de representación de un individuo o de una variable

La calidad de representación de un individuo i sobre el eje s puede ser medida por la distancia entre el punto en el espacio y la proyección sobre el eje. En realidad, preferimos calcular el porcentaje de inercia del individuo i proyectado sobre el eje s. Así, anotando θ_i^s el ángulo entre Oi y u_s, tenemos :

$$qlt_s\,(i) = \frac{\text{inercia proyectada de } i \text{ sobre } u_s}{\text{inercia total de } i} = \cos^2\theta_i^s.$$

Gracias al teorema de Pitágoras, este indicador se suma para varios ejes y se calcula, la mayoría de las veces, para un plano.

La calidad de representación de una variable k sobre el eje de rango s se escribe :

$$qlt_s\,(k) = \frac{\text{inercia proyectada de } k \text{ sobre } v_s}{\text{inercia total de } k} = \cos^2\theta_k^s.$$

Esta última cantidad es igual a $r^2(k, v_s)$: por esta razón, la calidad de representación de una variable es raramente provista por los programas. En cuanto a la calidad de representación de una variable sobre un plano, esta se evalúa visualmente a partir de la distancia al borde del círculo de radio 1.

Detección de individuos notables

El análisis de la forma de la nube N_I pasa también por la detección de individuos notables o particulares. Un individuo es notable si toma valores extremos sobre varias variables. En la nube N_I, tal individuo está lejos del centro de gravedad de la nube, y podemos evaluar su carácter notable por su distancia al centro de la nube en el espacio completo $\mathbb{R}^K$.
En el ejemplo, ningún individuo es particularmente extremo (cf. tabla 1.6). Los dos individuos más extremos son *Tropicana ambiente* y *Pampryl fresco*.

Pampryl amb.	Tropicana amb.	Fruvita fr.	Joker amb.	Tropicana fr.	Pampryl fr.
3.03	1.98	2.59	2.09	3.51	2.34

Tabla 1.6 – Datos zumo de naranja : distancia de los individuos al centro de la nube.

Contribución de un individuo o de una variable en la construcción de un eje

Los individuos notables influyen en el análisis y es interesante ver cuál es su influencia sobre la construcción de los ejes. Además, ciertos individuos pueden influir en la construcción de ciertos ejes sin ser por eso individuos notables. La detección de los individuos que contribuyen en la construcción de un eje factorial permite evaluar la estabilidad de los ejes. También es interesante evaluar la contribución de una variable en la construcción de un eje (esto, sobre todo, en ACP no normado).
Para ello, descomponemos la inercia de un eje, individuo por individuo (o variable por variable). La parte de inercia explicada por el individuo i sobre el eje s es :

$$\frac{(1/I)\,(OH_i^s)^2}{\lambda_s} \times 100.$$

Las distancias intervienen al cuadrado, lo que acentúa el papel de los individuos alejados del origen. Los individuos más alejados son los más extremos sobre el eje. Estas contribuciones son sobre todo útiles cuando los pesos de los individuos son diferentes.

Observación
Estas contribuciones se suman para varios individuos.

Cuando un individuo contribuye mucho (*i.e.*, mucho más que otros) en la construcción de un eje factorial (por ejemplo *Tropicana ambiente* y *Pampryl fresco*, para el segundo eje cf.

	Dim.1	Dim.2
Pampryl amb.	31.29	0.08
Tropicana amb.	2.76	36.77
Fruvita fr.	13.18	0.02
Joker amb.	12.63	8.69
Tropicana fr.	35.66	4.33
Pampryl fr.	4.48	50.10

Tabla 1.7 – Datos zumo de naranja : contribución de los individuos en la construcción de los ejes.

tabla 1.7), es frecuente que los resultados de un nuevo ACP construido sin este individuo cambien de modo sustancial : los principales factores de variabilidad pueden cambiar y aparecer nuevas oposiciones entre individuos.
Del mismo modo, calculamos la contribución de la variable k en la construcción del eje s, lo que da para el ejemplo, los resultados presentados en la tabla 1.8.

	Dim.1	Dim.2
Intensidad.olor	4.45	42.69
Tipo.olor	20.47	1.35
Pulposo	10.98	28.52
Intensidad.sabor	8.90	13.80
Ácido	17.56	9.10
Amargo	18.42	2.65
Azucarado	19.22	1.89

Tabla 1.8 – Datos zumo de naranja : contribución de las variables en la construcción de los ejes.

1.6.2 Elementos suplementarios

Distinguimos la noción de elementos activos y suplementarios ; hablamos de modo indistinto de elementos suplementarios o ilustrativos. Por definición, un elemento activo contribuye en la construcción de los ejes factoriales, contrariamente a un elemento suplementario. Así, la inercia de la nube de individuos se calcula teniendo como base los individuos activos en un espacio generado por las únicas variables activas ; del mismo modo, la inercia de la nube de variables, en el espacio generado por los únicos individuos activos, se calcula teniendo como base las variables activas. Los elementos suplementarios permiten ilustrar los ejes factoriales, de ahí su nombre de elementos ilustrativos. Al contrario de los elementos activos, que deben ser homogéneos, podemos introducir, de manera ilustrativa, elementos disparatados y numerosos.

Representación de variables cuantitativas suplementarias

Por definición, una variable cuantitativa suplementaria no interviene en el cálculo de las distancias entre individuos. La representamos de la misma manera que las variables activas como ayuda a la interpretación de la nube de individuos (§ 1.3.3) : la coordenada de la

variable suplementaria k' sobre el eje s corresponde al coeficiente de correlación entre k' y F_s el componente principal s ; podemos así representar k' sobre el mismo gráfico que las variables activas.

Más categóricamente, podemos utilizar las fórmulas de transición para calcular la coordenada de la variable suplementaria k' sobre el eje de rango s :

$$G_s(k') = \frac{1}{\sqrt{\lambda_s}} \frac{1}{I} \sum_{i \in \{\text{activo}\}} x_{ik'} F_s(i) = r(k', F_s).$$

con {activo} el conjunto de los individuos activos : el cálculo de esta coordenada se hace únicamente a partir de los individuos activos.

En el ejemplo, disponemos además de los descriptores sensoriales, de las variables fisico-químicas (cf. tabla 1.9). Sin embargo, el punto de vista adoptado sobre los datos no cambia, es decir : describir los zumos de naranja a partir del único perfil sensorial. Esta problemática puede ser enriquecida uniendo las dimensiones sensoriales a las variables fisico-químicas.

	Glucosa	Fructosa	Sacarosa	Capacidad de endulzar	pH	Ácido cítrico	Vitamina C
Pampryl amb.	25.32	27.36	36.45	89.95	3.59	0.84	43.44
Tropicana amb.	17.33	20.00	44.15	82.55	3.89	0.67	32.70
Fruvita fr.	23.65	25.65	52.12	102.22	3.85	0.69	37.00
Joker amb.	32.42	34.54	22.92	90.71	3.60	0.95	36.60
Tropicana fr.	22.70	25.32	45.80	94.87	3.82	0.71	39.50
Pampryl fr.	27.16	29.48	38.94	96.51	3.68	0.74	27.00

Tabla 1.9 – Datos zumo de naranja : variables suplementarias.

El círculo de correlaciones (cf. figura 1.11) permite una representación conjunta de las variables activas y suplementarias. El principal eje de variabilidad opone los zumos de naranja percibidos como ácido-amargo, poco azucarados y poco típicos a los zumos de naranja percibidos como azucarados, típicos, poco ácidos y poco amargos. El análisis de esta percepción sensorial es reforzado por las variables *pH* y *sacarosa*. En efecto, estas dos variables son correladas positivamente al primer eje y se sitúan al lado de los zumos de naranja percibidos como azucarados y poco ácidos (un índice de pH elevado indica una acidez débil). También encontramos la reacción llamada «de inversión (o de hidrólisis) de sacarosa» : sacarosa se descompone en glucosa y fructosa en un medio ácido (los zumos de naranja ácidos contienen más fructosa y glucosa y menos sacarosa que la media).

Observación

Cuando utilizamos el ACP con un fin exploratorio de los datos antes de realizar una regresión múltiple, se aconseja escoger las variables explicativas del modelo de regresión como variables activas del ACP y proyectar en suplementario la variable que hay que explicar. Esto da una idea de las relaciones entre variables explicativas y de la necesidad a seleccionar las variables explicativas en el modelo. Esto también da una idea sobre la calidad de la regresión : si la variable que hay que explicar está bien proyectada, el modelo se ajusta bien los datos.

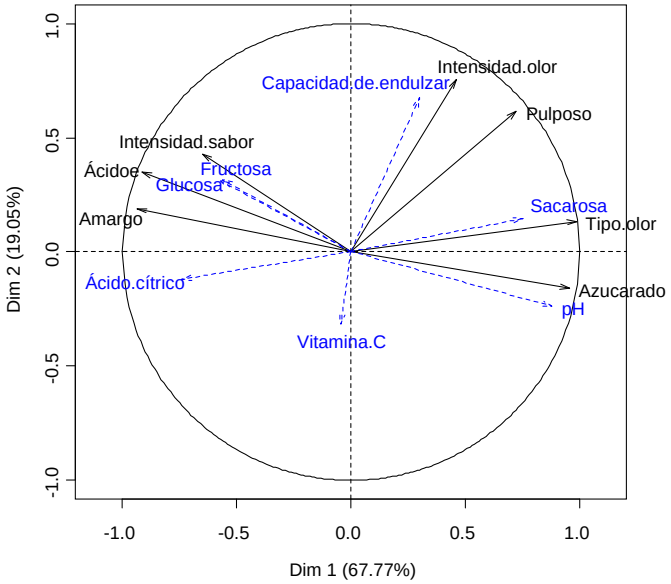

FIGURE 1.11 – Datos zumo de naranja : representación de variables activas y suplementarias.

Representación de variables cualitativas suplementarias

Las variables activas de un ACP son necesariamente cuantitativas pero es posible utilizar la información procedente de variables cualitativas a título ilustrativo =(suplementario), es decir, que no se utilizan en el cálculo de las distancias entre individuos.
Las variables cualitativas no pueden representarse de la misma manera que las variables cuantitativas suplementarias ya que es imposible calcular la correlación entre una variable cualitativa y F_s. La información de una variable cualitativa se sitúa a nivel de las modalidades. Es natural representar una modalidad en el baricentro del conjunto de los individuos que la poseen. Así, como respuesta a la proyección sobre el plano factorial, estas modalidades se quedan en el baricentro de los individuos en su representación plana. De esta manera, una modalidad puede considerarse como el individuo medio obtenido a partir del conjunto de los individuos que la poseen. En este sentido, la representamos sobre el gráfico de los individuos.
La información procedente de una variable cualitativa suplementaria puede también ser representada a través de un código de color : el conjunto de los individuos que poseen la misma modalidad se colorea con el mismo color. Esto permite visualizar la dispersión alrededor de los baricentros asociados a las modalidades.
En el ejemplo, podemos introducir la variable *condición de conservación* que toma las modalidades *ambiente* y *fresco* así como la variable *origen* de los zumos de frutas que toma las modalidades *Florida* y *Otro*. Parece haber una percepción sensorial diferente entre productos según su embalaje (aunque todos ellos hubieran sido degustados a la misma temperatura). La segunda bisectriz separa los productos comprados en la parte fresca de los otros.

	Condición de conservación	Origen
Pampryl amb.	Del tiempo	Otro
Tropicana amb.	Del tiempo	Florida
Fruvita fr.	Fresco	Florida
Joker amb.	Del tiempo	Otro
Tropicana fr.	Fresco	Florida
Pampryl fr.	Fresco	Otro

Tabla 1.10 – Datos de los jugos de naranja : variables cualitativas suplementarias.

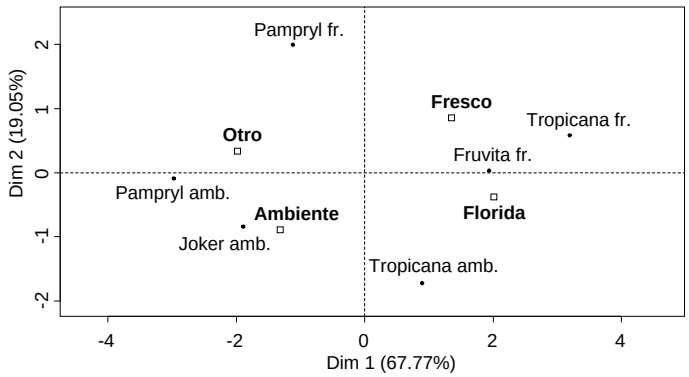

FIGURE 1.12 – Datos de los jugos de naranja : representación plana de la nube de individuos con dos variables cualitativas suplementarias.

Representación de individuos suplementarios

Del mismo modo que para las variables, podemos utilizar una fórmula de transición para calcular la coordenada de un individuo suplementario i' sobre el eje de rango s :

$$F_s(i') = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^{K} x_{i'k} G_s(k).$$

Precisemos que el centrado y la reducción (eventual), se hacen con relación a las medias y a las desviaciones-típicas calculadas sobre los individuos activos únicamente. Además, el cálculo de la coordenada de i' se hace únicamente a partir de las variables activas. No es necesario disponer de valores tomados por los individuos suplementarios para las variables suplementarias.

Observación
Una modalidad suplementaria puede ser considerada como un individuo suplementario que tomaría, para cada variable activa, la media calculada sobre el conjunto de los individuos que poseen esta modalidad.

1.6.3 Descripción automática de los ejes

Los ejes obtenidos por el análisis factorial pueden ser descritos de modo automático por el conjunto de las variables, ya sean cuantitativas o cualitativas, activas o suplementarias.
Para una variable cuantitativa, el principio es el mismo, sea la variable activa o suplementaria. Calculamos el coeficiente de correlación entre las coordenadas de los individuos sobre el eje s y cada una de las variables. Clasificamos las variables por su coeficiente de correlación, del más elevado al más débil y conservamos las variables que tienen los coeficientes más altos (en valor absoluto).

Observación
Recordemos que los ejes factoriales, como variables sintéticas, son combinaciones lineales de variables activas. Someter a un test la significación del coeficiente de correlación entre un componente y una variable es pues un procedimiento por construcción erróneo. Sin embargo, es útil clasificar y seleccionar las variables activas de ese modo para describir los ejes. En cambio, para las variables suplementarias, el test descrito para la significación del coeficiente de correlación entre dos variables corresponde al utilizado más a menudo.

Para una variable cualitativa, efectuamos un análisis de varianza a 1 factor donde procuramos explicar las coordenadas de los individuos (sobre el eje de rango s) por la mencionada variable cualitativa ; utilizamos la restricción $\sum_{i=1}^{I} \alpha_i = 0$. Luego, para cada modalidad, se construye un test t de Student que permite comparar la media de los individuos que poseen la modalidad a la media general (probamos $\alpha_i = 0$; para esto consideramos las varianzas de las coordenadas iguales para cada modalidad). Las modalidades positivas (resp. negativas) luego son clasificadas por probabilidad crítica creciente (resp. decreciente).
Estas ayudas para la interpretación son particularmente útiles para interpretar las dimensiones cuando el número de variables es importante.
Los datos utilizados contienen pocas variables ; no obstante, damos como ejemplo, las salidas del procedimiento de descripción automática del primer eje. Las variables que más caracterizan el eje 1 son el tipo de olor, el carácter azucarado, el carácter amargo y el carácter ácido (cf. tabla 1.11).

	Correlación	p.value
Tipo.olor	0.9854	0.0003
Carácter.azucarado	0.9549	0.0030
pH	0.8797	0.0208
Carácter.ácido	−0.9127	0.0111
Carácter.amargo	−0.9348	0.0062

Tabla 1.11 – Datos de los jugos de naranja : descripción de la primera dimensión por las variables cuantitativas.

El primer eje también se caracteriza por la variable cualitativa *Origen* ya que el test de correlación es significativamente diferente de 0 (probabilidad crítica del test igual a 0.00941) ; los zumos de naranja de Florida tienen coordenadas significativamente más elevadas que la media en el primer eje mientras que los zumos de naranja de otras procedencias tienen coordenadas inferiores a la media (cf. tabla 1.12).

```
$Dim.1$quali
           R2   p.value
Origen  0.8458    0.0094

$Dim.1$category
          Estimate p.value
Florida     2.0031  0.0094
Otro       -2.0031  0.0094
```

Tabla 1.12 – Datos de los jugos de naranja : descripción de la primera dimensión por las variables cuantitativas.

1.7 Puesta en práctica con FactoMineR

En esta sección, mostramos cómo efectuar un ACP con FactoMineR y cómo encontrar los resultados obtenidos sobre el juego de datos de los jugos de naranja. Primero cargamos FactoMineR y luego importamos los datos precisando que el nombre de los individuos está en la primera columna (`row.names=1`) :

```
library(FactoMineR)
naranja <- read.table("http://factominer.free.fr/libro/naranja.csv",
   header=TRUE, sep=";", dec=".", row.names=1)
summary(naranja)
```

El ACP se obtiene precisando que aquí las variables de 8 a 14 son cuantitativas suplementarias y las variables 15 y 16 son cualitativas suplementarias :

```
res.pca <- PCA(naranja,quanti.sup=8:14,quali.sup=15:16)
```

Esta instrucción aplica el ACP y proporciona el gráfico de las variables (con las variables activas y suplementarias, cf. figura 1.11) y el gráfico de los individuos (con los individuos y las modalidades de las variables cualitativas suplementarias, cf. figura 1.12). Para dibujar el gráfico solamente con los individuos (cf. figura 1.7), utilizamos la función **plot.PCA** :

```
plot(res.pca, invisible="quali")
```

Las tablas 1.4, 1.5, 1.6, 1.7 y 1.8 se obtienen por las líneas de código siguientes :

```
round(res.pca$var$coord[,1:2],2)
round(res.pca$eig,2)
round(res.pca$ind$dist,2)
round(res.pca$ind$contrib[,1:2],2)
round(res.pca$var$contrib[,1:2],2)
```

La función **dimdesc** proporciona la descripción automática de las dimensiones de las variables cuantitativas (cf. tabla 1.11) y cualitativas (cf. tabla 1.12). La función **lapply** permite únicamente redondear (gracias a la función **round**) todos los términos de una lista (aquí ; en el interior de una lista de lista !) :

```
lapply(dimdesc(res.pca),lapply,round,2)
```

1.8 Complementos

1.8.1 Test de significación de los ejes

Puede ser interesante comparar el porcentaje de inercia asociado a un eje o a un plano en el cuartil 0.95 de la distribución de estos porcentajes obtenida simulando tablas de datos de dimensiones equivalentes teniendo como base una ley multinormal. Estos cuartiles son reunidos en las tablas de la página 190 hasta la página 193 y un ejemplo es illustrado en § 1.9.4.

1.8.2 Resultados sobre las variables : loadings *vs.* correlación

El punto de vista que adoptamos es el de interesarse a las correlaciones entre las variables y los factores y es muy utilizado particularmente en Francia. Sin embargo, existen otros puntos de vista y los anglosajones, particularmente, prefieren interesarse a los «loadings». Los «loadings» se interpretan como los coeficientes de la combinación lineal de las variables iniciales que permiten la construcción de los factores. De un punto de vista numérico, los «loadings» son iguales a las coordenadas de las variables divididas por la raíz cuadrada del valor propio asociado al eje. Los «loadings» son las salidas por defecto de las funciones **princomp** y **prcomp** de R.
Este punto de vista algébrico no permite introducir variables suplementarias ya que estas variables no intervienen en la construcción de los ejes y como consecuencia no intervienen en la combinación lineal.

Para ir más lejos. EL ACP corresponde a un cambio de base que permite pasar de las variables iniciales a sus combinaciones lineales tales que la inercia de la nube de puntos proyectada sea máxima. Así, la matriz de los loadings corresponde a la matriz de paso de la antigua a la nueva base. Esta matriz corresponde a las coordenadas de los vectores propios que provienen de la diagonalización de la matriz de varianza-covarianza (ACP no normado) o de correlación. Podemos pues escribir (en el caso del ACP normado) :

$$F_s(i) = \sum_{k=1}^{K} L_s(k) \frac{(x_{ik} - \bar{x}_k)}{s_k}$$

con $L_s(k)$ el coeficiente de la combinación lineal (loading) de la variable k sobre el eje de rango s.

1.8.3 Representación simultánea : gráfico biplot

El biplot es un gráfico en el cual se representan dos conjuntos de objetos de naturaleza diferente. Cuando el número de individuos y el número de variables son débiles, puede ser interesante representar simultáneamente la nube de individuos y la nube de variables en un biplot. Sin embargo, esta representación superpuesta es ficticia ya que ambas nubes no evolucionan en el mismo espacio (una pertenece a $\mathbb{R}^K$ y la otra a $\mathbb{R}^I$). Nos fijamos entonces en interpretar sólo las direcciones de las variables en función de los individuos : un individuo está del lado de las variables para las cuales toma grandes valores. Sin embargo

las distancias entre individuos están deformadas a causa de una dilatación de cada eje por el inverso de la raíz cuadrada del valor propio que se le asocia : esta deformación es muy importante ya que las inercias de los ejes de representación son muy diferentes. Además, no es posible representar variables cuantitativas suplementarias. Para obtener una representación simultánea de las nubes, se puede utilizar la función **biplot**.

1.8.4 Datos ausentes

Es muy frecuente que haya datos ausentes en una tabla de datos. El modo más simple para manejar los datos ausentes es reemplazar cada dato ausente por la media de la variable para la cual este dato es ausente. Este modo de proceder da resultados correctos si el número de datos ausentes no es demasiado importante.
Más allá de esta técnica un poco grosera, existen otras metodologías más sofisticadas que sacan provecho de la estructura de la tabla y que se revelan generalmente más adecuadas. Indiquemos sucintamente dos ideas. Consideremos dos variables x y y estrechamente correlacionadas cuando tomamos en cuenta los individuos completos para ellas. En ausencia de valor de y para el individuo i, es natural estimar este dato ausente a partir del valor de x para el mismo individuo (por ejemplo con la ayuda de una regresión simple). Consideremos ahora dos individuos i y l de los cuales todos los valores presentes son muy próximos. En ausencia de valor de l para la variable k, es normal considerarlo por el valor de i para la misma variable k. Integrando estas ideas para aprehender el conjunto de los datos, podemos construir algoritmos de estimación de datos ausentes. Estos algoritmos son, en el momento en el que se están escribiendo estas líneas, el objeto de búsquedas activas y su implantación en paquete missMDA está realizándose ; su descripción va más allá del presente estudio.

1.8.5 Juego de datos de grandes dimensiones

Las tablas de datos en ciertas disciplinas, por ejemplo en genómica, contienen muchas más variables que individuos (es frecuente tener algunas decenas de filas y algunos millones de columnas). En este caso, los algoritmos diagonalizan la matriz de los productos escalares en lugar de la matriz de correlación, lo que disminuye los tiempos de cálculo.
Cuando el número de individuos y el número de variables son simultáneamente grandes, podemos recurrir a los algoritmos iterativos evocados en el párrafo precedente sobre los datos ausentes.

1.8.6 Rotación varimax

La práctica de rotación de los ejes inicialmente procedente del análisis en factores comunes y específicos (otro método de análisis de datos pero que está fundado sobre un modelo) y es utilizado en ACP por los Anglosajones.
Es posible efectuar una rotación de la representación de la nube de variables obtenida por ACP de manera que los ejes sean interpretables más fácilmente. Numerosos procedimientos son disponibles ; el más conocido ciertamente está fundado sobre el criterio varimax (y el procedimiento se llama, por abuso de lenguaje, procedimiento varimax). La rotación varimax es la rotación que maximiza la suma de los cuadrados de los loadings. Para efectuar el

procedimiento varimax en R, utilizamos la función **varimax**. Este procedimiento necesita definir previamente el número de ejes retenidos (para representar la nube de variables).
Este procedimiento tiene la ventaja de proporcionar ejes que están muy vinculados a ciertas variables y muy poco vinculados a otras, y tiene el inconveniente de que no proporcina soluciones encajadas : los dos primeros ejes de la solución en dos dimensiones no corresponden con los dos primeros ejes de la solución en tres dimensiones.
Este procedimiento privilegia el estudio de la nube de variables, particularmente dimensión por dimensión y sin la ayuda de gráficos.

1.9 Ejemplo : datos de los gastos del hogar

1.9.1 Descripción de los datos

El juego de datos procede de una encuesta «presupuesto por familia» llevada por el INSEE en 2006 (`http://www.insee.fr/fr/bases-de-donnees/`). Estas encuestas permiten conocer el peso de los grandes secciones de consumo en el presupuesto de la casa. La tabla 1.13 da el reparto del gasto anual medio (en Euros) por cuidado del hogar según la edad de la persona de referencia.
En esta tabla, una fila corresponde a un grupo de edad, una columna a una variable que corresponde a una sección de gasto ; x_{ik} corresponde al gasto medio en euros de un cuidado del hogar del grupo de edad i para el sección de gasto k. Disponemos de 30 variables cuantitativas (26 corresponden a diferentes rúbricas, tres corresponden a totales parciales y uno al gasto total).
Disponemos además del reparto del gasto anual para el conjunto de la población francesa así como del reparto de los gastos en función de la renta del hogar. Las rentas son reagrupadas por decilo : El decilo 1 corresponde al 10 % de los hogares que tiene las rentas más débiles, el decilo 2 concierne a las personas que tienen las rentas entre 10 y 20 % más débiles... y el decilo 10 corresponde al 10 % que tienen las rentas más elevadas.
Para cargar el package FactoMineR e importar el juego de datos, realizamos los comandos siguientes :

```
> library(FactoMineR)
> gastos <- read.table("http://factominer.free.fr/libra/gastos.csv",
    header=TRUE, sep=";", row.names=1)
```

1.9.2 Problemática

A partir de la tabla de los gastos brutos en euros, podemos construir una tabla de porcentajes para estudiar la parte del presupuesto para cada rúbrica. Trabajar en la tabla bruta o en la tabla de porcentajes no responde exactamente a los mismos objetivos : Si utilizamos los porcentajes, nos interesamos al reparto de los gastos por puesto, mientras que si utilizamos los datos brutos, podemos estudiar simultáneamente el reparto y el nivel de gasto, es decir, los gastos relativos y absolutos. En el marco de esta presentación, analizamos los datos brutos pero estudiar los porcentajes también daría resultados interesantes.
Nos interesamos aquí en la evolución del perfil de los gastos según la edad. Este objetivo bastante general puede declinarse según varias cuestiones. ¿Hay grupos de edad que tienen

	pan.cereales	carnes	pescados.mariscos	leche.quesos.huevos	aceites.grasas	frutas	legumbres	azúcar.productos	otros_productos	café.té.cacao	agua.bebidas	otros_gastos_alimentación	bebidas_alcohólicas	restauración	tabaco	ropa.artículos_calzado	vivienda.agua.gas.electricidad	mobiliario.equipamiento_doméstico	salud	transportes	comunicaciones	ocio.cultura	enseñanza	servicios.alojamiento	otros_bienes.servicios	fuera_campo_consumo_final	total_productos_alimenticios	total_alimentación	total_no_alimentación	gasto_total
Menos de 25 años	545	405	64	297	34	84	177	139	108	39	137	103	188	1185	518	1618	5855	1239	415	3474	1042	1715	354	70	1856	2585	2321	3506	18156	21662
De 25 a 34 años	741	734	153	522	56	156	297	230	187	78	204	151	233	1663	463	2690	5693	2000	805	4750	1214	2452	76	135	3656	7548	3742	5404	23935	29339
De 35 a 44 años	1005	1079	231	691	88	223	410	319	163	110	248	202	322	1867	502	2950	4681	2309	987	5551	1134	2872	272	270	3694	10474	5091	6958	25223	32181
De 45 a 54 años	944	1199	291	662	96	279	430	297	130	118	241	235	487	1761	436	2668	4526	2243	1054	5505	1257	2985	456	317	3716	11365	5410	7171	25163	32334
De 55 a 64 años	769	1167	325	613	102	305	443	253	126	109	209	184	481	1102	249	2080	3982	2187	1155	4818	961	2775	112	190	3632	10667	5087	6189	22141	28331
De 65 a 74 años	668	1061	286	525	111	315	410	219	97	90	161	141	450	608	85	1234	3630	1697	981	2845	650	2131	11	138	3207	6942	4535	5143	16609	21752
75 años y más	562	912	251	422	90	294	336	175	81	88	118	99	301	531	57	716	3493	1418	1015	1292	513	1311	6	73	2537	5486	3728	4259	12431	16690
Conjunto	795	1015	250	574	89	254	384	251	132	98	200	172	376	1320	328	2132	4428	1986	981	4285	989	2460	189	195	3384	8798	4590	5910	21357	27267
D1	581	687	132	382	67	169	272	171	91	62	148	73	225	532	299	1272	4072	875	486	1723	727	1019	118	61	1951	2175	3061	3593	12604	16197
D2	665	770	175	466	78	192	291	202	100	68	162	75	199	579	350	1256	4441	1051	537	2089	796	1279	134	69	2155	3056	3443	4023	14157	18180
D3	713	926	190	515	90	213	337	210	113	81	174	95	291	868	347	1517	4363	1188	959	3067	806	1469	138	115	2481	4261	3945	4813	16450	21263
D4	763	947	210	540	87	219	338	241	126	90	185	108	277	879	393	1638	4355	1618	879	3659	895	1803	136	118	2635	5020	4130	5008	18130	23138
D5	795	1040	223	556	89	237	367	243	136	91	197	143	366	1037	373	1867	4296	1677	865	3712	994	2076	125	139	3211	6169	4483	5520	19334	24853
D6	823	988	234	591	85	237	358	247	144	97	208	164	367	1201	329	2005	4270	1806	892	4651	966	2219	169	160	3222	7245	4543	5744	20687	26431
D7	887	1107	265	642	95	263	407	286	143	115	229	177	397	1444	349	2422	4366	2305	1028	4932	1132	2679	247	205	3740	9046	5014	6458	23405	29863
D8	913	1151	305	663	98	283	443	282	153	119	234	177	418	1736	315	2613	4501	2296	1086	5411	1147	3160	207	277	4028	10041	5296	7032	25041	32073
D9	892	1198	339	676	94	340	474	303	152	117	231	265	498	2155	294	2897	4466	2734	1306	6148	1199	3926	236	344	4684	14281	5579	7734	28234	35968
D10	913	1342	427	713	103	386	549	329	159	142	236	390	720	2768	232	3837	5147	4312	1776	7458	1230	4968	377	464	5737	26695	6409	9177	35538	44715

Tabla 1.13 – Datos gastos : reparto de gastos anuales medios de los franceses por grupos de edades (26 primeras columnas de la tabla 1.13).

perfiles de gastos muy próximos ? Podemos así obtener una tipología de los grupos de edad según los diferentes puestos de gastos ; la tipología siendo construida de modo que dos grupos de edad son tan próximos que tienen perfiles de gastos similares. ¿Hay puestos de gastos que evolucionan del mismo modo según la edad ? Es decir, queremos obtener un balance de las relaciones entre los puestos de gastos a través del estudio de los coeficientes de correlación entre las variables tomadas dos a dos. Además, ¿podemos construir indicadores sintéticos que permiten resumir la evolución de los perfiles de gastos ? Podemos interesarnos por los indicadores *a priori* como los totales parciales o el gasto total, pero el análisis permite construir indicadores *a posteriori* (los factores del ACP) que podremos intentar unir a otros indicadores (por ejemplo, la renta media).
Ambos resultados, el primero sobre los individuos y el segundo sobre las variables, están confrontados para describir la tipología de los individuos a partir de las variables y recíprocamente.
También podremos unir la tipología de los individuos con las variables cuantitativas que no participaron a la construcción de las distancias interindividuales (las variables que corresponden a los totales) así como los individuos suplementarios (los decilos de renta).

1.9.3 Elección del análisis

Elección de los elementos activos

Para obtener una tipología de las clases de edad fundada sobre sus gastos, definimos la distancia entre dos grupos de edad únicamente teniendo como base sus gastos en las diferentes rúbricas. Así, las otras rúbricas, que corresponden a las rúbricas de los totales, no son variables activas sino variables suplementarias que pueden ser útiles para la interpretación. Maticemos sin embargo el carácter ilustrativo de estas variables que indirectamente participan a la construcción de los ejes ya que son sumas de variables activas. Una variable como la de la renta sería una «verdadera» variable ilustrativa. En cuanto a los individuos, los grupos de edad son unos individuos activos y los decilos de renta se consideran como individuos suplementarios.

Elección del peso sobre los individus

En la inmensa mayoría del ACP, la elección del peso de los individuos se impone igual a $1/I$. Aquí, nos preguntamos : ¿qué peso conceder a cada individuo (más precisamente a cada individuo activo, *i.e.*, ¿A cada grupo de edad) ? ¿Concedemos el mismo peso a cada grupo de edad o entonces concedemos un peso igual a la proporción de jefes del hogar (en la población francesa) que pertenecen a cada grupo de edad ? Conceder el mismo peso a cada grupo de edad permite centrar la atención en la evolución del perfil de los gastos entre cada grupo de edad. En cambio, conceder un peso proporcional al total de jefes del hogar que pertenecen al grupo de edad vuelve a describir los gastos del conjunto de la población francesa. Recordemos que si los grupos de edad son equilibrados (sería el caso si se estudiasen en activo los grupos de rentas que son decilos), Ambas posibilidades conducen al mismo resultado. Si los grupos de edad contienen un efectivo muy desequilibrado, el riesgo es de tener una clase que contribuya demasido en el análisis. Escogemos aquí tomar pesos equilibrados. Anotemos que, desde el punto de vista del programa, basta con utilizar el

argumento `row.w` para precisar un vector de peso (sobre los individuos activos) ; cuando el argumento no es precisado, el peso $1/I$ es afectado por defecto.

Estandarización o no estandarización de variables

Las diferentes rúbricas son medidas en las mismas unidades (en euros), entonces es posible estandarizarlas o no (cuando las variables están en unidades diferentes, es indispensable estandarizarlas). Si se estandardizan las variables, su influencia en el cálculo de las distancias entre grupos de edad es equilibrado desde el punto de vista de su desviación-tipo respectiva : sin estandarización, la variable *Fuera de campos consumo final*, con una desviación-tipo de 2961.62, tendría una influencia más de 100 veces superior a la de la variable *café.té.cacao* (desviación-tipo de 26.77).Esta elección de estandardizar o no, tiene pues un impacto importante sobre los resultados del análisis. La tabla siguiente muestra que las desviaciones-tipos son globalmente proporcionales a la media (el coeficiente de variación es muy similar de una variable a otra). Así, concediendo más peso a las variables que tienen una desviación-tipo fuerte, concedemos más peso a las variables que corresponden a los puestos importantes de gasto. Este punto de vista puede ser adoptado en una perspectiva «económica». En una perspectiva más bien «sociológica», un puesto de gasto débil puede merecer la atención. Escogemos aquí este segundo punto de vista estandarizando.

```
> apply(gastos[1:7,],2,mean)
> apply(gastos[1:7,],2,sd)*sqrt(6/7)
> apply(gastos[1:7,],2,sd)*sqrt(6/7)/apply(gastos[1:7,],2,mean)
```

	pan.cereales	carnes	pescados.mariscos	leche.quesos.huevos	aceites.grasas	frutas	legumbres	azúcar.productos	otros_productos	café.té.cacao	agua.bebidas	otros_gastos_alimentación	bebidas_alcohólicas	restauración	tabaco	ropa.artículos_calzado	vivienda.agua.gas.electricidad	mobiliario.equipamiento_doméstico	salud	transportes	comunicaciones	ocio.cultura	enseñanza	servicios.alojamiento	otros_bienes.servicios	fuera_campo_consumo_final
media	748	937	229	533	82	237	358	233	127	90	188	159	352	1245	330	1994	4551	1870	916	4034	967	2320	184	170	3185	7867
desv.tip	164	262	84	129	25	81	88	59	34	25	47	47	113	502	183	776	871	393	226	1456	263	584	164	88	671	2962
cv	0,22	0,28	0,37	0,24	0,31	0,34	0,25	0,25	0,27	0,27	0,25	0,29	0,32	0,40	0,56	0,39	0,19	0,21	0,25	0,36	0,27	0,25	0,89	0,51	0,21	0,38

1.9.4 Puesta en práctica del análisis

Para efectuar el análisis, utilizamos la función **PCA** del package FactoMineR cuyos principales parámetros de entrada son : la tabla de datos, la elección de estandardizar o no las variables, los índices de los individuos suplementarios, los índices de las variables cuantitativas suplementarias, los índices de las variables cualitativas (necesariamente suplementarias). Por defecto, las variables son estandarizadas (`scale.unit=TRUE`), ningún individuo es suplementario (`ind.sup=NULL`) y ninguna variable es suplementaria (`quanti.sup=NULL` y `quali.sup=NULL`, es decir todas las variables son cuantitativas y activas).
En el ejemplo, precisamos que los individuos de 8 a 18 (los decilos de renta) son suplementarios y que las variables de 27 a 30 (los diferentes totales) son cuantitativas suplementarias :

```
> res.pca <- PCA(gastos,ind.sup=8:18,quanti.sup=27:30)
```

La función **PCA** proporciona el gráfico de los individuos y el gráfico de las variables así como las salidas numéricas siguientes contenidas en el objeto `res.pca` :

```
> res.pca
**Results for the Principal Components Analysis (PCA)**
The analysis was performed on 18 individuals, described by 30 variables
*The results are available in the following objects:

   name                 description
1  "$eig"               "eigenvalues"
2  "$var"               "results for the variables"
3  "$var$coord"         "coord. for the variables"
4  "$var$cor"           "correlations variables - dimensions"
5  "$var$cos2"          "cos2 for the variables"
6  "$var$contrib"       "contributions of the variables"
7  "$ind"               "results for the individuals"
8  "$ind$coord"         "coord. for the individuals"
9  "$ind$cos2"          "cos2 for the individuals"
10 "$ind$contrib"       "contributions of the individuals"
11 "$ind.sup"           "results for the supplementary individuals"
12 "$ind.sup$coord"     "coord. for the supplementary individuals"
13 "$ind.sup$cos2"      "cos2 for the supplementary individuals"
14 "$quanti.sup"        "results for the supplementary quantitative variables"
15 "$quanti.sup$coord"  "coord. for the supplementary quantitative variables"
16 "$quanti.sup$cor"    "correlations suppl. quantitative variables - dimensions"
17 "$call"              "summary statistics"
18 "$call$centre"       "mean of the variables"
19 "$call$ecart.type"   "standard error of the variables"
20 "$call$row.w"        "weights for the individuals"
21 "$call$col.w"        "weights for the variables"
```

Elección del número de dimensiones a estudiar

La inercia de los ejes factoriales indica por una parte, si las variables son estructuradas (presencia de correlaciones entre variables) y por otra parte, sugiere el número de componentes principales por interpretar.
El objeto `res.pca$eig` contiene el valor propio (*i.e.*, la inercia o la varianza explicada) asociado a cada uno de los ejes, el porcentaje de inercia que representa en el análisis así como la acumulación de estos porcentajes. Damos aquí los resultados redondeados de los dos primeros decimales con la ayuda de la función **round** :

```
> round(res.pca$eig,2)
       eigenvalue  percentage  cumulative percentage
                      variance            of variance
comp 1      15.52       59.69                 59.69
comp 2       8.67       33.34                 93.03
comp 3       1.22        4.67                 97.71
comp 4       0.38        1.48                 99.18
comp 5       0.14        0.55                 99.73
comp 6       0.07        0.27                100.00
```

Podemos visualizar estos valores propios con la ayuda de un diagrama en barras (cf. figura 1.13) utilizando el siguiente comando :

```
> barplot(res.pca$eig[,1], main="Valores propios",
    names.arg=paste("dim",1:nrow(res.pca$eig)))
```

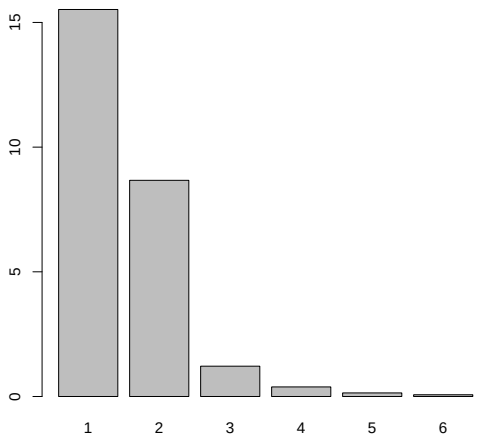

FIGURE 1.13 – Datos gastos : valor propio asociado a cada dimensión del ACP.

Los dos primeros ejes expresan 93.03% de la inercia total ; en otros términos, el 93.03% de la variabilidad total de la nube de los individuos (o de las variables) está representado por el primer plano. Este porcentaje es extremadamente importante y el primer plano representa la variabilidad contenida en el conjunto del juego activo de datos. Según la tabla de la página 193, el cuantilo 95 % obtenido para 7 individuos y 25 variables vale 56.4%. El porcentaje de inercia explicado por el primer plano de nuestro juego de datos es pues significativo. Los ejes 3 y 4 expresan sólo 4.7% y 1.5% de inercia y no aportarán mucha información. No obstante, es prudente representarlos para asegurarse de que no son interesantes de interpretar.

Plano 1-2

Estudio de la nube de los individuos activos. La representación de la nube de los individuos pone en evidencia eventuales particularidades del juego de datos : por ejemplo, la presencia de una partición sobre los individuos, los individuos extremos,etc. La función **PCA** proporciona por defecto un gráfico con los individuos activos y suplementarios. Podemos construir el gráfico de los individuos activos únicamente con la ayuda de la función **plot.PCA** (que puede ser llamada **plot** o **plot.PCA**). Precisamos entonces que construimos el gráfico de los individuos (`choix="ind"`) y que hacemos invisible a los individuos suplementarios (`invisible="ind.sup"`) :

```
> plot.PCA(res.pca, choix="ind", invisible="ind.sup")
```

Este gráfico de los individuos (cf. figura 1.14) presenta una disposición remarcable : el primer eje opone los grupos de edad extremos con los grupos de edad medios. La relación entre la edad y el primer eje no es lineal. El segundo eje ordena los grupos de edad del más alto al

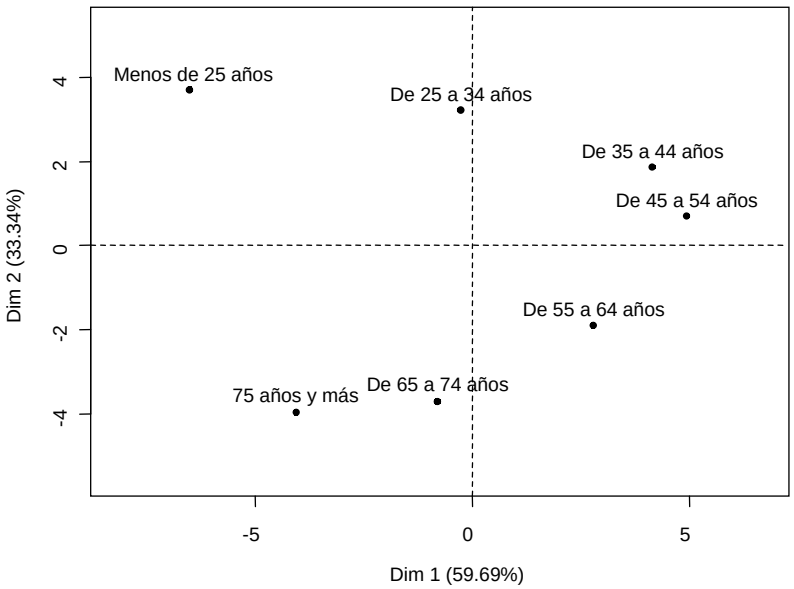

FIGURE 1.14 – Datos gastos : gráfico de los individuos.

más bajo. El objeto `res.pca$ind` contiene las coordenadas, los cosenos al cuadrado y las contribuciones para cada individuo. Damos aquí estos tres indicadores para los individuos activos y los tres primeros ejes.

```
> round(cbind(res.pca$ind$coord[,1:3],res.pca$ind$cos2[,1:3],
    res.pca$ind$contrib[,1:3]),2)
              Dim.1 Dim.2 Dim.3  Dim.1 Dim.2 Dim.3  Dim.1 Dim.2 Dim.3
Menos de 25 años -6.53  3.70  1.19   0.74  0.24  0.02  39.29 22.61 16.59
De 25 a 34 años  -0.30  3.22 -2.04   0.01  0.70  0.28   0.08 17.08 49.07
De 35 a 44 años   4.11  1.87 -0.29   0.78  0.16  0.00  15.53  5.75  0.97
De 45 a 54 años   4.90  0.71  1.66   0.87  0.02  0.10  22.10  0.84 32.25
De 55 a 64 años   2.75 -1.89 -0.22   0.62  0.29  0.00   6.97  5.88  0.59
De 65 a 74 años  -0.84 -3.68 -0.11   0.05  0.91  0.00   0.65 22.26  0.14
75 años y más    -4.09 -3.94 -0.18   0.50  0.47  0.00  15.38 25.58  0.39
```

Estudio de la nube de las variables. Las representaciones de la nube de las variables permiten visualizar rápidamente las correlaciones entre variables, la presencia de grupos de variables muy relacionadas entre ellas, etc.
La función **PCA** proporciona por defecto el primer plano (ejes 1 y 2) con las variables activas y suplementarias (las variables activas en negro y en lineas contínuas y las variables suplementarias en azul y en lineas discontínuas). Podemos construir el gráfico de las variables activas únicamente con la ayuda de la función **plot.PCA**. Entonces, precisamos que construimos el gráfico de las variables (`choix="var"`) y que hacemos invisibles las variables suplementarias (`invisible="quanti.sup"`) :

```
> plot.PCA(res.pca, choix="var", invisible="quanti.sup")
```

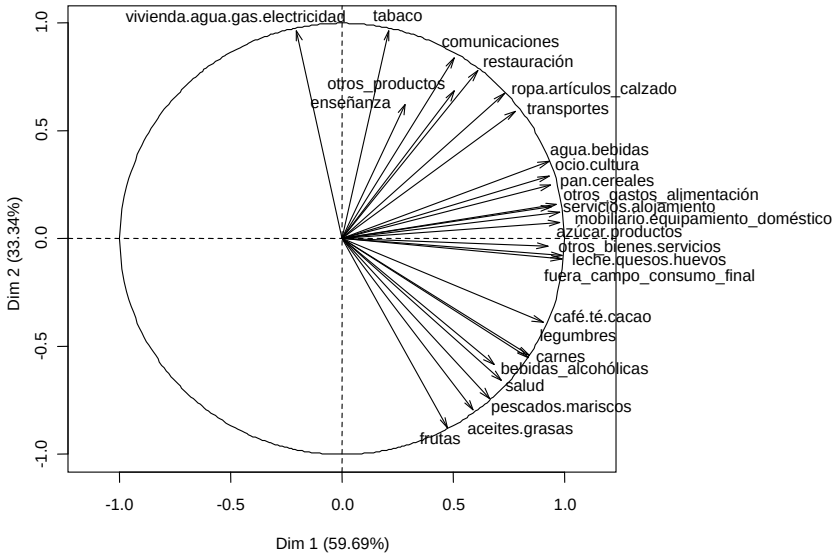

FIGURE 1.15 – Datos de gastos : gráfico de variables acivas.

La figura 1.15 pone en evidencia una correlación positiva entre el primer componente principal y todas las variables excepto una (*vivienda agua gas electricidad*). Así, este eje opone grupos de edad que consumen poco (grupos que tienen coordenadas negativas en el primer eje) con grupos de edad que consumen mucho (en cualquier tipo de gastos).
Todas las variables están bien representadas en el plano 1-2, excepto la variable *enseñanza*. Podemos leer esta calidad de representación en el gráfico gracias a la aproximación entre el extremo de la flecha y el círculo de radio 1. Encontramos las coordenadas de las variables activas en el objeto `res.pca$var` así como sus calidades de representación (cosenos al cuadrado) y sus contribuciones en la construcción de los ejes (para no cargar mucho el texto, damos indicadores para ocho variables) :

```
> round(cbind(res.pca$var$coord[,1:3],res.pca$var$cos2[,1:3],
    res.pca$var$contrib[,1:3]),2)
                    Dim.1 Dim.2 Dim.3  Dim.1 Dim.2 Dim.3  Dim.1 Dim.2 Dim.3
pan.cereales         0.94  0.25  0.05   0.88  0.06  0.00   5.67  0.70  0.20
carnes               0.83 -0.55  0.06   0.70  0.30  0.00   4.48  3.46  0.27
pescados.mariscos    0.66 -0.74  0.05   0.44  0.55  0.00   2.82  6.31  0.17
leche.quesos.huevos  0.99 -0.08 -0.09   0.97  0.01  0.01   6.28  0.07  0.62
aceites.grasos       0.59 -0.80  0.04   0.35  0.63  0.00   2.24  7.29  0.16
frutas               0.48 -0.88  0.04   0.23  0.77  0.00   1.45  8.87  0.12
legumbres            0.84 -0.54 -0.01   0.70  0.29  0.00   4.54  3.36  0.01
enseñanza            0.28  0.62  0.72   0.08  0.39  0.52   0.52  4.49 42.81
```

Las variables estando bien representadas, lo mismo ocurre con el ángulo entre dos variables, es decir, con la correlación entre dos variables. Ciertos gastos están fuertemente correlados entre ellos : *pan.cereales* y *leche.quesos.huevos* están correlados positivamente (*i.e.*, ciertos grupos de edad gastan poco en estos dos puestos de gastos y otros gastan mucho). La variable

vivienda.agua.gas.electricidad es ortogonal a estas dos variables y por consecuencia está muy poco correlada con ellas. Esto puede verificarse a partir de la matriz de correlación y basta de calcular una parte (aquí nos limitamos a las variables 1, 4 y 17) :

```
> round(cor(depenses[,c(1,4,17)]),2)
                                  pan        leche.quesos      vivienda.agua
                                cereales          huevos     gas.electricidad
pan.cereales                       1.00            0.95             0.11
leche.quesos.huevos                0.95            1.00            -0.06
vivienda.agua.gas.electricidad     0.11           -0.06             1.00
```

Las variables suplementarias son útiles aquí para simplificar la lectura del gráfico de las variables. En efecto, en este ejemplo, las variables suplementarias son variables totales que resumen varias variables. El gráfico de las variables suplementarias (cf. figura 1.16) es obtenido por :

```
> plot.PCA(res.pca, choix="var", invisible="var")
```

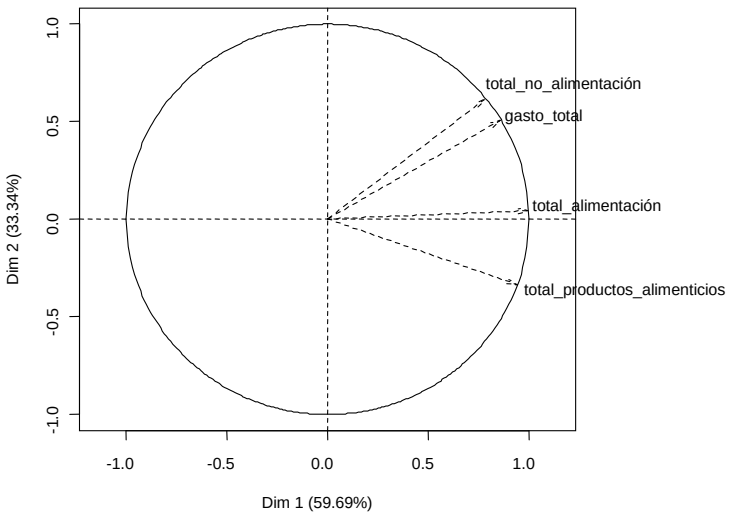

FIGURE 1.16 – Datos gastos : gráfico de las variables suplementarias.

Las coordenadas (y las calidades de representación) de variables suplementarias están disponibles en el objeto `res.pca$quanti.sup` :

```
> round(cbind(res.pca$quanti.sup$coord[,1:3],res.pca$quanti.sup$cos2[,1:3]),2)
                           Dim.1 Dim.2 Dim.3  Dim.1 Dim.2 Dim.3
total_productos_alimenticios  0.94 -0.34  0.04   0.88  0.11  0.00
total_alimentación            1.00  0.04  0.03   0.99  0.00  0.00
total_no_alimentación         0.78  0.61 -0.05   0.60  0.38  0.00
gasto_total                   0.86  0.51 -0.04   0.73  0.26  0.00
```

Descripción automática de las dimensiones. Recordemos que es posible obtener una descripción automática de los ejes factoriales (cf. § 1.6.3) por las variables activas o suplementarias con la ayuda de la función **dimdesc** :

```
> dimdesc(res.pca)
$Dim.1
$Dim.1$quanti
                                  correlation  p.value
total_alimentación                      0.996 1.85e-06
fuera_campo_consumo_final               0.988 3.21e-05
leche.quesos.huevos                     0.987 3.70e-05
azúcar.productos                        0.976 1.69e-04
mobiliario.equipamiento_doméstico       0.976 1.70e-04
otros_gastos_alimentación               0.964 4.48e-04
servicios.alojamiento                   0.941 1.54e-03
total_productos_alimenticios            0.941 1.60e-03
pan.cereales                            0.938 1.78e-03
agua.bebidas                            0.931 2.35e-03
ocio.cultura                            0.929 2.48e-03
otros_bienes.servicios                  0.924 2.92e-03
café.té.cacao                           0.903 5.36e-03
gasto_total                             0.856 1.41e-02
legumbres                               0.839 1.82e-02
carnes                                  0.834 1.97e-02
transportes                             0.778 3.93e-02
total_no_alimentación                   0.777 3.98e-02

$Dim.2
$Dim.2$quanti
                                  correlation  p.value
vivienda.agua.gas.electricidad          0.967 3.79e-04
tabaco                                  0.967 3.80e-04
comunicaciones                          0.840 1.81e-02
restauración                            0.780 3.86e-02
aceites.grasos                         -0.795 3.26e-02
frutas                                 -0.877 9.53e-03
```

Esta función es más útil cuantas más variables hayan y por lo tanto el gráfico de las variables esté cargado. Vemos aquí que el primer eje está muy vinculado a la variable *total_alimentación* (coeficiente de correlación de 0.996) y a la variable *fuera_campo_consumo_final*, etc. El segundo eje está muy vinculado a las variables *vivienda agua gas electricidad* y *tabaco*.

Análisis conjunto de la nube de individuos y de la nube de variables. La representación de la nube de individuos y de la nube de variables deben ser analizadas conjuntamente ; es decir las distancias entre individuos pueden ser explicadas por las variables y las relaciones entre variables ilustradas por los individuos.
Así como lo vimos gracias al gráfico de las variables, el primer eje opone los grupos de edad que gastan poco a los grupos de edad que gastan mucho.
Más en el detalle, este primer componente está sobre todo vinculado a los gastos alimentarios, lo que está bien resumido por la variable total alimentación que está muy vinculada al primer componente principal (correlación = 0.996). Recordemos aquí, que el primer componente

principal es la combinación lineal de las variables que las sintetiza mejor. En este ejemplo, la síntesis automática proporcionada por el ACP casi coincide con el total alimentación. Una gran parte de las diferencias (cf. el porcentaje de inercia de 59.69%) que existen entre los puestos de gastos de un grupo de edad al otro, puede ser resumida por la sola variable *total alimentación*.

El eje 2 opone los gastos de *fruta*, *aceites.grasos*, *pescado.marisco* y *salud* los gastos *vivienda.agua.gas.electricidad.conbustible*, *tabaco*, *communicación* y *restauración*. Este eje separa entre ellos sobre todo los grupos de edad que gastan globalmente poco (los grupos de edad extremos). Los grupos de edad medios se separan sobre otros ejes (contribuyen mucho en la construcción de los ejes 3 y 4, cf. tabla de contribuciones de la página 32). Entre los presupuestos más débiles, encontramos presupuestos más especializados : los que tienen menos de 25 años (coordenada positiva en el eje 2) gastan más (que la media) en communicación, vivienda, restauración, (variables correladas positivamente con el factor 2) y menos (que la media) en pescado, fruta, salud, aceites y grasa (variables correladas negativamente con el factor 2). Las personas mayores (coordenadas negativas) presentan un perfil de gasto opuesto.

Esto puede ser ilustrado comparando los*menos de 25 años* y los*75 años y más* a partir de los datos centrados reducidos (ver tabla página 39). Anotemos que estos dos grupos de edad tienen coordenadas débiles sobre el eje 1 y gastan globalmente poco ; sus gastos importantes para ciertos puestos, exacerbados por el eje 2, tienen que relativizarse con relación a su gasto global.

Individuos suplementarios. Podemos proyectar como individuos suplementarios, el individuo *Conjunto* así como los decilos (cf. figura 1.17 que corresponden a uno de los gráficos por defecto de la función **PCA**).

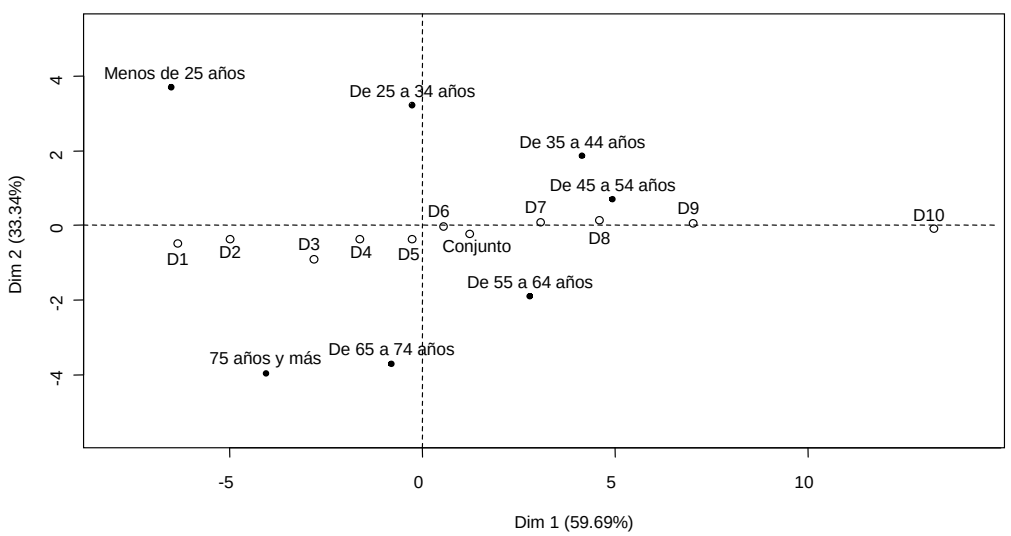

FIGURE 1.17 – Datos gastos : gráfico de individuos y proyección de individuos suplementarios.

Fundamentalmente, la introducción de elementos (individuos o variables) suplementarios pretende enriquecer el análisis. Su elección no es tan crucial como el de los elementos activos en el sentido siguiente. Si, *in fine*, el análisis revela que algunos de estos elementos no aportan nada, basta con no tomarlos en consideración en la interpretación que permanece pues intacta. Estos datos «gastos» ilustran dos ejemplos de peso para introducir individuos suplementarios. 1) Cuando se menciona un gasto en comunicación superior a la media, se trata de una media «artificial» en el sentido en el que no se sabe destinarlo a una población : es la media équiponderada de clases de efectivos diferentes. Introduciendo, en suplemento, el perfil de gasto de la población total, podremos interpretar la separación entre el perfil medio y la media «artificial» representada por el origen de los ejes. 2) Estudiamos la evolución del perfil de gasto en función de la edad. Pero existe una variable que juega, por construcción, un papel superior en los gastos : la renta. Introducir, en líneas suplementarias, los perfiles de gastos de los tramos de renta, permite responder a cuestiones del tipo : ¿tal evolución de perfil de gasto, puesta en evidencia a partir de los tramos de edad, corresponde a una evolución de renta ?

El individuo *Conjunto* está en el centro del gráfico, lo que era de esperar ya que corresponde al perfil Francia entera. El punto no está exactamente en el centro de gravedad de la nube porque la ponderación del ACP utilizado $(1/I)$ no corresponde al peso utilizado para construir el perfil Francia entera (peso igual al porcentaje de cabezas de familia que pertenece al grupo de edad). La diferencia es débil entre el perfil medio (baricentro) y el perfil Francia entera (individuo *Conjunto*).

A lo largo del primer eje, todos los decilos sin excepción están ordenados del más pequeño al más elevado. Así, el primer eje opone las rentas más débiles a las rentas más elevadas. Todos los decilos son repartidos de modo homogéneo, excepto el último decilo que es más extremo que lo esperado : hay un salto entre las rentas muy elevadas (decilo 9) y las rentas más elevadas (decilo 10) en el consumo. Esta distancia más grande entre los perfiles de gastos de los decilos 9 y 10 corresponde sin duda a una separación más grande de renta. Encontramos los resultados de individuos suplementarios en el objeto `res.pca$ind.sup` bajo la forma de una tabla que contiene las coordenadas y los cosenos al cuadrado.

```
> round(cbind(res.pca$ind.sup$coord[,1:4],res.pca$ind.sup$cos2[,1:4],2)
         Dim.1 Dim.2 Dim.3  Dim.1 Dim.2 Dim.3
Conjunto  1.21 -0.22 -0.04   0.96  0.03  0.00
D1       -6.38 -0.48  0.52   0.94  0.01  0.01
D2       -5.01 -0.35  0.27   0.89  0.00  0.00
D3       -2.83 -0.89  0.10   0.79  0.08  0.00
D4       -1.66 -0.35 -0.22   0.63  0.03  0.01
D5       -0.30 -0.36 -0.29   0.12  0.17  0.11
D6        0.53 -0.01 -0.23   0.52  0.00  0.10
D7        3.04  0.11 -0.17   0.99  0.00  0.00
D8        4.57  0.15 -0.18   0.99  0.00  0.00
D9        7.00  0.07 -0.19   0.96  0.00  0.00
D10      13.23 -0.08  0.04   0.90  0.00  0.00
```

Plano 2-3

Estudio de la nube de individuos y de variables. Podemos también interesarnos por la dimensión siguiente y construir el gráfico 2-3 de individuos (cf. figura 1.18) y el de variables

(cf. figura 1.19) precisando en la función **plot.PCA** los ejes de representación (`axes=2:3`) :

```
> plot(res.pca, choix="ind", axes=2:3)
> plot(res.pca, choix="var", axes=2:3)
```

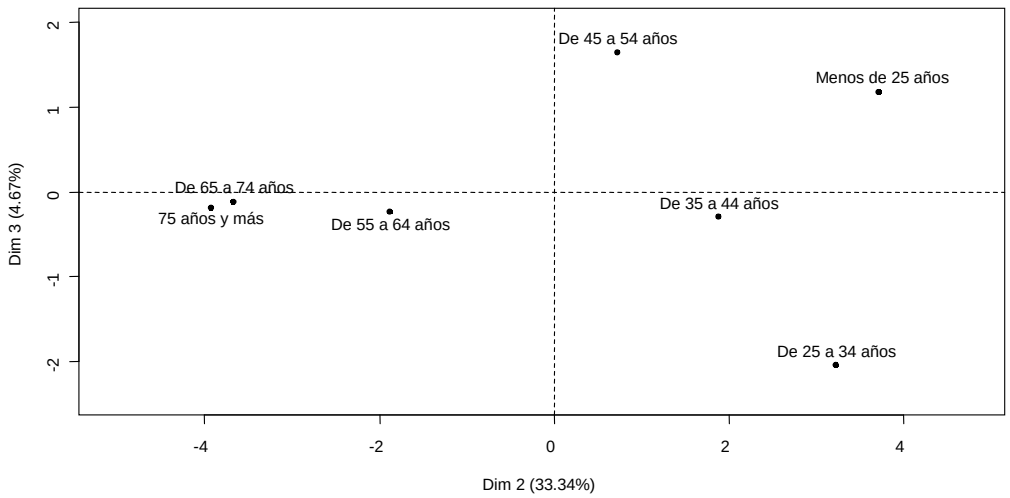

FIGURE 1.18 – Datos gastos : gráfico de individuos en el plano (2, 3).

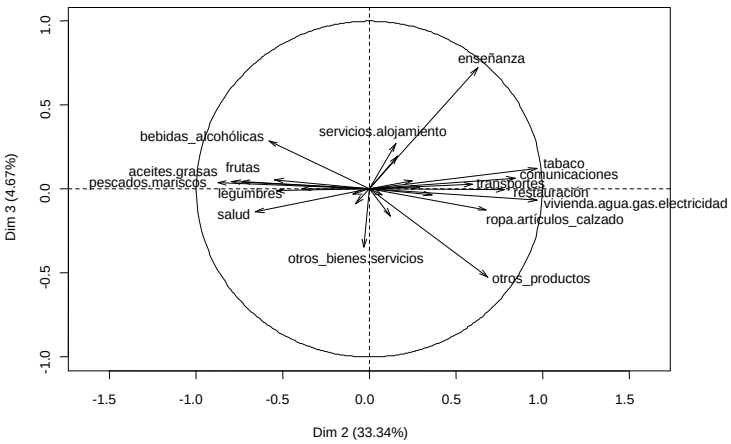

FIGURE 1.19 – Datos gastos : gráfico de variables en el plano (2, 3).

Análisis conjunto de la nube de individuos y de la nube de variables. El eje 3 está esencialmente vinculado a la variable *enseñanza* (correlación positiva) y a un menor grado

a la variable *otros_productos_alimenticios* (correlación negativa). Principalmente opone el tramo de edad 25-34 años a los tramos menos de 25 años y 45-54 años (suma de las contribuciones al eje 3 de estos tres puntos : 97.91%, cf. taba de la página 32). Los grupos de edad *menos de 25 años* y *45-54 años* gastan más que los otros grupos de edad en enseñanza. Podemos suponer que son o bien estudiantes o bien padres de estudiantes o de miños escolarizados. Recíprocamente, podemos suponer que el tramo de edad 25-34 años contiene pocos o ningun estudiante y/o padre. Podremos concluir que este tercer eje es interpretable ; no obstante, el débil valor propio recuerda que se trata de una dimensión de variabilidad de débil importancia comparada con la de los ejes precedentes (lo que concierne prácticamente a una sola variable).

Vuelta a los datos

La representación de individuos y la representación de variables, son representaciones aproximadas de la tabla de datos, por una parte, de la matriz de correlación (o de varianza-covarianza) por otra parte. Es prudente insistir en la interpretación de los resultados volviendo a los datos. Más abajo, recapitulamos las líneas de comandos que permiten obtener las medias y las desviaciones típicas por variable, los datos centrados- reducidos (en su conjunto o en una parte) y la matriz de correlación (completa o reducida a una selección de variables).

```
> res.pca$call$centre
> res.pca$call$ecart.type
pan.cereales  carnes  pescados.mariscos leche.quesos.huevos  aceites.grasos
      747.71  936.71             228.71              533.14           82.43
pan.cereales  carnes  pescados.mariscos leche.quesos.huevos  aceites.grasos
      163.51  262.39              84.34              128.84           25.40
```

El cálculo de datos centrados-reducidos es útil para comparar fácilmente los datos con la media en número de desviaciones típicas pero sobre todo para comparar los valores de una variable a otra. Para tener los datos centrados-reducidos, es necesario multiplicar por $\sqrt{(I-1)/I}$ ya que por defecto, la función **scale** considera que la desviación-típica es estimada a partir de una muestra.

```
> round(scale(gastos[1:7,c(5,6,15,17,19,21)])*sqrt(6/7),2)
                 aceites frutas tabaco    vivienda.agua salud comunicaciones
                  grasos                gas.electricidad
Menos de 25 años   -1.63  -1.62   0.88              1.28 -1.90           0.24
De 25 a 34 años    -0.89  -0.85   0.62              1.12 -0.42           0.80
De 35 a 44 años     0.19  -0.14   0.80              0.13  0.27           0.54
De 45 a 54 años     0.46   0.45   0.50             -0.03  0.52           0.94
De 55 a 64 años     0.66   0.72  -0.38             -0.56  0.91          -0.02
De 65 a 74 años     0.96   0.83  -1.15             -0.91  0.25          -1.03
75 años y más       0.26   0.61  -1.28             -1.04  0.37          -1.48
```

Ilustremos la tabla de datos centrados-reducidos por la variable *Comunicaciones*. Los individuos cuyo valor centrado-reducido es negativo (resp. positivo) tienen un presupuesto más débil (resp. más elevado) que la media del conjunto de los grupos de edad para esta variable : son los tramos de más edad (resp. los más jóvenes).

Los individuos *75 años y más*, *De 65 a 74 años* y *De 45 a 54 años* son los más particulares, sus valores centrados reducidos son los más grandes en valor absoluto. Podemos representar los valores de esta variable (cf. figura 1.20) con los comandos :

```
> par(las=2)
> plot(gastos[1:7,21],type="b",axes=F,ylab="Communicaciones (en Euros)",xlab="",bty="o")
> axis(2)
> axis(1,1:7,rownames(gastos)[1:7])
> par(las=0)
```

Visualizamos así una «fractura numérica» entre los que tienen más de 64 años y los otros. El gráfico de variables en el plano 1-2 podía hacer pensar que eran los jóvenes quienes gastaban mucho en comunicaciones. Pero de hecho, no es nada de esto, principalmente son los que tienen más edad quienes gastan menos que los otros.

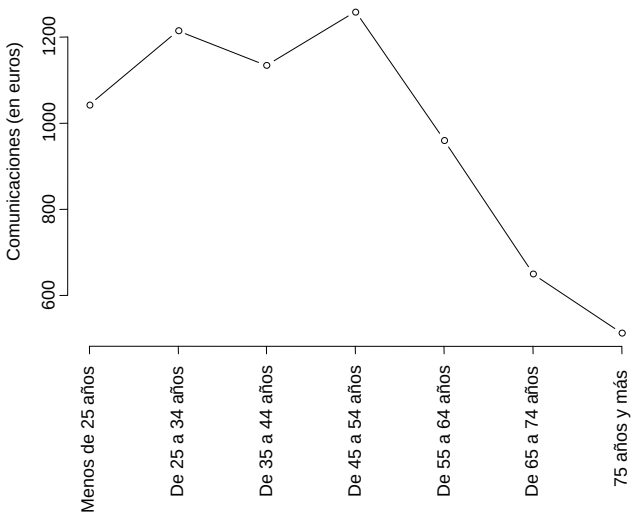

FIGURE 1.20 – Datos gastos : evolución de los gastos en comunicación.

Los datos centrados-reducidos permiten comparar igualmente los gastos du un mismo tramo de edad respecto a dos rúbricas de gastos. Los menores de 25 años gastan más en *vivienda.gas.electricidad* y en *tabaco* con relació a la media (valor centrado-reducido positivo para estas dos variables), pero son más notables por sus gastos en *vivienda* que en *tabaco* (el valor centrado-reducido para la variable *vivienda*, 1.28, es más extremo que para la variable *tabaco*, 0.88).

La matriz de correlación puede obtenerse con la función **cor**. Además, las relaciones entre variables pueden visualizarse dos a dos con la ayuda de la función **pairs** (cf. figura 1.21) :

```
> pairs(gastos[1:7,1:4])
```

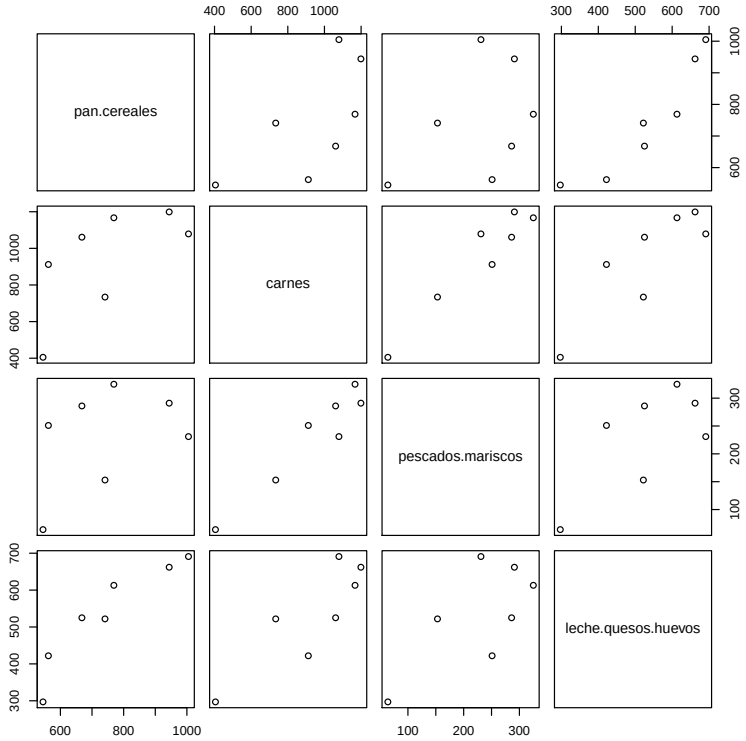

FIGURE 1.21 – Datos gastos : nube de puntos de variables *pan.cereales*, *carnes*, *pescados.mariscos* y *leche.quesos.huevos*.

Comentario sobre los porcentajes de inercia

Los individuos suplementarios y su posición, muy fácilmente interpretable en el plano 2-3, validan y justifican la interpretación de estos ejes. A partir de los valores propios y los porcentajes de inercia explicados por el eje 3, podríamos haber pensado que no es interesante interpretarla. El último criterio para saber si nos interesamos o no por un eje es finalmente la facultad que tenemos para interpretar el eje. Es importante anotar que los datos utilizados aquí son las medias sobre muchos individuos. Esto confiere una cierta «robustez» a los datos y puede explicar que porcentajes de inercia muy débiles tengan una interpretación clara.

1.10 Ejemplo : datos sobre temperaturas

1.10.1 Descripción de datos-problemática

Nos interesamos por el clima de los diferentes países de Europa. Para ello, cogimos las temperaturas medias mensuales (en grados centígrados) de las principales capitales europeas así como para ciertas grandes ciudades. Además de las temperaturas mensuales, damos, para

cada ciudad, la temperatura media anual así como la amplitud térmica (diferencia entre la media mensual máxima y la media mensual mínima de una ciudad). Damos también dos variables cuantitativas de localización (la longitud y la latitud) así como una variable cualitativa (la pertenencia a una región de Europa, variable cualitativa con cuatro modalidades : Europa del norte, del sur, del este y del oeste). Un extracto de los datos es proporcionado en la tabla 1.14.

	Ene	Feb	Marz	Abr	...	Nov	Dic	Med	Amp	Lat	Lon	Reg
Ámsterdam	2.9	2.5	5.7	8.2	...	7.0	4.4	9.9	14.6	52.2	4.5	Oeste
Atenas	9.1	9.7	11.7	15.4	...	14.6	11.0	17.8	18.3	37.6	23.5	Sur
Berlín	−0.2	0.1	4.4	8.2	...	4.2	1.2	9.1	18.5	52.3	13.2	Oeste
Bruselas	3.3	3.3	6.7	8.9	...	6.7	4.4	10.3	14.4	50.5	4.2	Oeste
Budapest	−1.1	0.8	5.5	11.6	...	5.1	0.7	10.9	23.1	47.3	19.0	Este
Copenhague	−0.4	−0.4	1.3	5.8	...	4.1	1.3	7.8	17.5	55.4	12.3	Norte
Dublín	4.8	5.0	5.9	7.8	...	6.7	5.4	9.3	10.2	53.2	6.1	Norte
Helsinki	−5.8	−6.2	−2.7	3.1	...	0.1	−2.3	4.8	23.4	60.1	25.0	Norte
Kiev	−5.9	−5.0	−0.3	7.4	...	1.2	−3.6	7.1	25.3	50.3	30.3	Este
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabla 1.14 – Datos temperaturas : extracto de algunas de las 23 capitales ; las temperaturas son medidas en grados Celsius.

1.10.2 Elección del análisis

Elección de los elementos activos

El estudio de las ciudades. Deseamos aprehender la variabilidad de las temperaturas mensuales de un país a otro de manera multidimensional,*i.e.*, teniendo en cuenta los 12 meses del año simultáneamente. Un país será representado por el clima de su capital ; los datos de las otras ciudades no se tendrán en cuenta, para evitar dar más peso a los países de los cuales tenemos información sobre varias ciudades. Así, las capitales serán consideradas como individuos activos mientras que las otras ciudades serán consideradas como individuos suplementarios (*i.e.*, que no intervienen en la construcción de los ejes). Desde un punto de vista multidimensional, cuanto más dos ciudades presentan el mismo conjunto de temperaturas mensuales, más próximas son. Una manera sintética de abordar estos datos es de poner en evidencia los principales factores de variabilidad de las capitales. Podríamos así responder a cuestiones del tipo : ¿Cuáles son las desigualdades más grandes entre países ? Estos factores podrán servir de base en la construcción de una tipología sobre los países.
El estudio de variables. Cada variable mide las temperaturas mensuales de las 23 capitales. La relación entre las variables es aprehendida a partir de las capitales (*i.e.*, de individuos activos) y no del conjunto de las ciudades. Estas relaciones constituyen un objetivo esencial en este tipo de estudios. Dos variables son correladas positivamente si globalmente, las ciudades con más calor según una variable son las de más calor según la otra variable (por ejemplo, ¿hace calor en agosto o hace calor en enero ?). Naturalmente, queremos obtener una visión del conjunto de estas relaciones, sin analizar cada pareja de variables.

Esta visión del conjunto puede hacerse a través de variables sintéticas. La cuestión es entonces : ¿Podemos resumir las temperaturas mensuales por un pequeño número de componentes ? Si la respuesta es sí, examinaremos las relaciones entre las variables iniciales y las variables sintéticas : este exámen indirecto es más cómodo que el exámen directo (con 12 variables iniciales y 2 variables sintéticas, examinaremos 24 relaciones en lugar de $(12 \times 11)/2 = 66$).
Nos interesamos por los perfiles de las temperaturas de las ciudades, por eso tomaremos como variables activas únicamente las variables que conciernen la temperatura (lo que elimina variables como la latitud, la longitud). Para las otras variables propuestas como suplementarias (temperatura media anual y amplitud anual), se considerar como indicadores sintéticos que serán interesantes de confrontar con los componentes principales pero que tampoco pertenecen al perfil propiamente dicho. Además, son variables que utilizan una información ya presente en otras variables.

Estandarizar o no estandarizar las variables

El centrado-reducción es indispensable sólo cuando las variables activas no están en las mismas unidades : así, las variables suplementarias deberán ser analizadas a través de sus coeficientes de correlación con los factores y en este sentido deberán ser automáticamente reducidas.
La homogeneidad de las unidades de medida no es necesariamente un criterio decisivo para no reducir : ¿1 grado representa la misma cosa en enero y en julio ? No reducir vuelve a conceder a cada variable un peso proporcional a su varianza. Anotemos que las desviaciones-típicas difieren bastante poco de un mes al otro (como máximo del simple al doble) para que sea razonable pensar que el hecho de reducirse o no influencia poco los resultados del análisis. Según otro punto de vista, no reducir vuelve, en el cálculo de la distancia entre dos ciudades, a conceder la misma influencia con una diferencia de 1 grado, cualquiera que sea el mes del año ; reduciendo, esta diferencia es tan ampliada que aparece durante un mes en el que las temperaturas varían poco de una ciudad a otra. Aquí, ha sido escogido reducir, o sea, conceder el mismo peso a cada mes.

1.10.3 Puesta en práctica con FactoMineR

He aquí las líneas de código que permiten obtener los gráficos y las salidas del análisis que comentamos :

```
> library(FactoMineR)
> temperaturas <- read.table("http://factominer.free.fr/libra/temperaturas.csv",
    header=TRUE, sep=";", dec=".", row.names=1)
> res <- PCA(temperaturas, ind.sup=24:35, quanti.sup=13:16, quali.sup=17)
> plot.PCA(res, choix="ind", habillage=17)
> dimdesc(res)
> res$eig
> res$ind
> res$ind.sup
> res$var
> res$quanti.sup
> res$quali.sup
> scale(temperaturas[1:23,1:16])*sqrt(22/23)
```

```
> cor(temperaturas[1:23,1:16])
```

Estas líneas de código permiten :
- importar el juego de datos (precisando que el nombre de las variables está presente, el separador de campos es « ;» el separador de decimal es «.» y el nombre de los individuos está presente en la primera columna) ;
- poner en marcha el ACP con individuos suplementarios de 24 a 35 (ciudades que no son capitales), las variables cuantitativas suplementarias de 13 a 16 y la variable 17 como cualitativa suplementaria ;
- construir el gráfico de individuos vistiendo los individuos en función de la variable *Región* ;
- describir las dimensiones a partir de las variables ;
- recuperar la tabla con varianzas explicadas por cada eje ;
- recuperar la tabla con los resultados para los individuos activos ;
- recuperar la tabla con los resultados para los individuos suplementarios ;
- recuperar la tabla con los resultados para las variables (cuantitativas) activas ;
- recuperar la tabla con los resultados para las variables cuantitativas suplementarias ;
- recuperar la tabla con los resultados para las variables cualitativas suplementarias ;
- calcular los datos centrados-reducidos para las variables cuantitativas sobre los individuos activos únicamente ;
- calcular la matriz de correlación.

El primer factor es preponderante : él solo expresa el 82.9% de la inercia total (cf. los gráficos o el objeto `res$eig`). El segundo factor es relativamente importante ya que expresa el 15.4% de la inercia total. Estos dos factores expresan $82.9 + 15.4 = 98.3\%$ de la inercia total lo que justifica limitarse a ellos. Es decir, a partir de 2 variables sintéticas, resumimos casi toda la información aportada por las doce variables iniciales. Estamos aquí en un caso de escuela donde el resumen aportado por el ACP es casi exhaustivo. Esto tiene como consecuencia que las variables y los individuos en el primer plano están muy bien proyectados y la proximidad de dos individuos en el plano nos indica una proximidad en el espacio completo, de la misma manera que el ángulo entre dos variables en el plano da una aproximación muy buena del ángulo en el espacio.

Primer eje

Todas las variables activas tienen una coordenada del mismo signo (cf. figura 1.22) : estamos en presencia de un efecto tamaño. Ciertas ciudades tienen temperaturas fuertes en cualquier mes del año, otras tienen temperaturas débiles en cualquier mes. En otros términos, los meses son de una manera general, correlados positivamente dos a dos. Podemos resumir este eje con el término : temperatura media anual. Este resumen está confirmado por el coeficiente de correlación de 0.998 entre este factor y la variable ilustrativa del mismo nombre (el gráfico parece mostrar una correlación más débil pero el objeto `res$quanti.sup$coord` indica la correlación : 0.998). Más detalladamente, observamos que el mes de Septiembre, de Octubre y de Abril están más relacionados que los otros meses a este primer eje : «representan» mejor las temperaturas anuales. Excepto la temperatura media anual ya citada, otra variable cuantitativa suplementaria está vinculada al primer factor : la latitud. La correlación entre la latitud y el primer factor es de -0.85 lo que significa que las ciudades que están más al Sur (latitud más pequeña) tienen una coordenada más elevada en el primer eje y por eso son las ciudades con más calor : esto, evidentemente, no es una sorpresa.

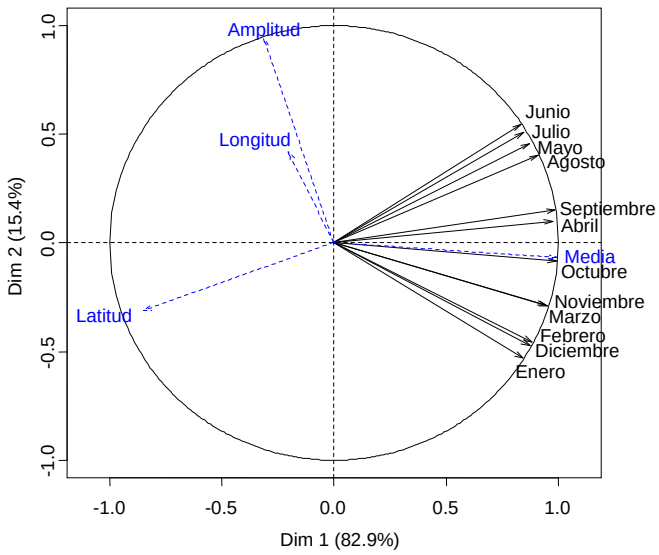

FIGURE 1.22 – Datos temperaturas : gráfico de variables.

Observación
El efecto tamaño da más información que el resumen *temperaturas anuales* ya que indica que las ciudades con más calor anualmente también lo son (más o menos) cada mes.

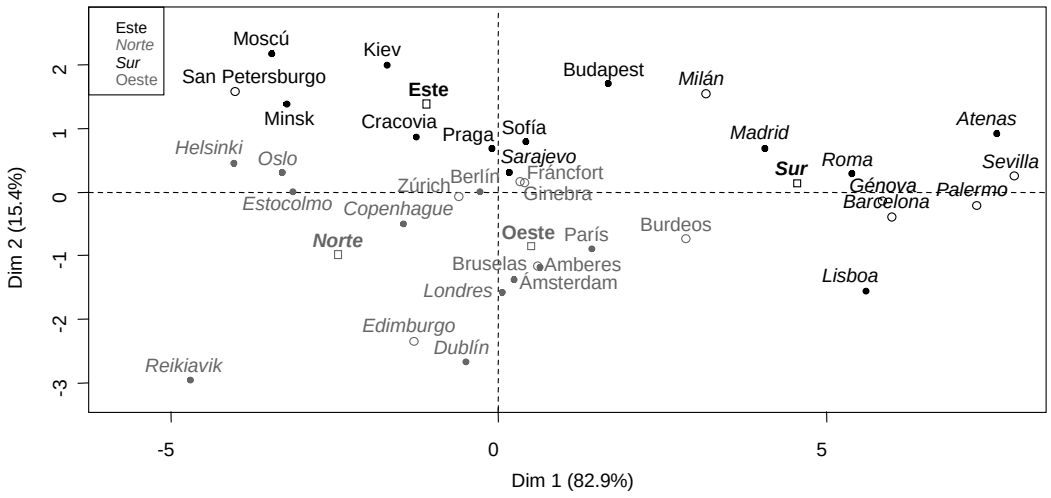

FIGURE 1.23 – Datos temperaturas : gráfico de individuos.

A causa de la dualidad, la coordenada de*Helsinki* (resp. *Atenas*) nos indica una ciudad donde hace frío (resp. calor) durante todo el año (cf. figura 1.23). Esto es claramente visible en los datos : cualquiera que sea el mes, *Helsinki* (resp. *Atenas*) es una ciudad con más frío (resp. con más calor) que la media. Esto se visualiza más fácilmente a partir de datos centrados-reducidos que obtenemos por :

```
> scale(temperaturas[1:23,1:12])*sqrt(22/23)
```

¡Atención ! Aquí precisamos que centramos y reducimos únicamente a partir de individuos activos, *i.e.*, a partir de 23 primeros individuos.

Segundo eje

Opone por una parte, los meses del período Mayo-Julio y por otra parte, los meses del período Noviembre-Marzo. Este factor puede ser resumido por la oposición «temporada buena»-«temporada mala». Atención : esta oposición no tiene nada que ver con una evolución de las medias ya que los datos son centrados antes del análisis. Esta oposición nos indica el hecho de que, a temperaturas anuales iguales, ciertas ciudades son relativamente más calrosas en la buena temporada del año y otras más bien frías. El adverbio relativamente es necesario por el centrado de datos : la oposición anteriormente citada entre ciudades, sólo puede leerse directamente cuando los datos son centrados-reducidos ; fuertes variaciones de las medias entre los meses pueden hacerla difícil de ilustrar sobre los datos brutos.

La *amplitud térmica anual* está vinculada a este eje, lo que puede unirse a los dos siguientes hechos : los valores más fuertes de esta variable son observados en las ciudades más continentales (arriba en el eje) y los valores más débiles son observados en las ciudades próximas del Atlántico (abajo en el eje). La *longitud* está vinculada a este eje pero la relación no es muy fuerte (correlación = 0.4196).

Así, a causa de la dualidad, ciudades como *Kiev*, *Moscú* o *Budapest* tienen valores centrados-reducidos más bien elevados durante la buena temporada y más bien débiles durante la estación de invierno ; por el contrario, ciudades como *Dublín* o *Reykjavik* tienen valores centrados-reducidos más bien débiles durante la temporada buena y suaves en invierno. De hecho, en estos datos, esta oposición puede leerse directamente en los datos brutos. Este eje individualiza las ciudades oceánicas de amplitud térmica débil con las ciudades continentales de fuerte amplitud térmica. Las temperaturas de ciudades oceánicas (por ejemplo *Dublín* o *Reykjavik*) son en conjunto medias o débiles (lo indicado en el primer eje) y son muy débiles durante la temporada buena. Por el contrario, ciudades continentales (como *Kiev*, *Moscú* o *Budapest*) son en conjunto medias o débiles y son muy débiles en invierno y elevadas en verano.

La descripción de los ejes (**dimdesc(res)**) obtenida a partir de variables cuantitativas confirma la descripción hecha de los ejes. La variable cualitativa *Región* permite igualmente caracterizar los ejes. Las modalidades *Norte* y *Sur* caracterizan el primer eje : la modalidad *Sur* (resp. *Norte*) tiene una coordenada significativamente positiva (resp. negativa) en el primer eje, lo que interpretamos como : las ciudades de Europa del Sur (resp. del Norte) tienen temperaturas más calientes (resp. más frías) todo el año. La modalidad *Este* (resp. *Norte*) tiene una coordenada significativamente positiva (resp. negativa) en el segundo eje, lo que interpretamos como : las ciudades de Europa del Este (resp. del Norte) tienen amplitudes térmicas importantes (resp. más débiles). El conjunto de los resultados con respecto a la variable cualitativa *Región* pueden encontrarse en el objeto `res$quali.sup`.

```
$Dim.1
$Dim.1$quanti
          correlation  p.value
Media           0.998  9.58e-26
Octubre         0.992  3.73e-20
Septiembre      0.986  1.06e-17
Abril           0.974  5.30e-15
Noviembre       0.952  2.66e-12
Marzo           0.945  1.15e-11
Agosto          0.909  1.90e-09
Febrero         0.884  2.18e-08
Diciembre       0.873  5.45e-08
Mayo            0.870  7.01e-08
Julio           0.844  4.13e-07
Enero           0.842  4.59e-07
Junio           0.833  7.96e-07
Latitud        -0.852  2.57e-07

$Dim.1$quali
             R2   p.value
Región   0.679 6.282e-05

$Dim.1$category
      Estimate   p.value
Sur      4.183 2.282e-05
Este    -1.478 4.090e-02
Norte   -2.823 4.983e-04

$Dim.2
$Dim.2$quanti
          correlation   p.value
Amplitud       0.9444 1.296e-11
Junio          0.5453 7.120e-03
Julio          0.5087 1.319e-02
Mayo           0.4578 2.804e-02
Longitud       0.4196 4.621e-02
Febrero       -0.4558 2.882e-02
Diciembre     -0.4729 2.268e-02
Enero         -0.5314 9.077e-03

$Dim.2$quali
              R2  p.value
Región    0.546  0.00153

$Dim.2$category
      Estimate   p.value
Este    1.4620 0.0004473
Norte  -0.9064 0.0166600
```

Conclusión

El balance de las relaciones entre las temperaturas pone en evidencia correlaciones positivas entre las diferentes temperaturas mensuales y más finamente, dos períodos : la temporada buena (de Mayo a Agosto) y la temporada mala (de Noviembre a Marzo). Esta separación en dos periodos no está hecha en función de la evolución de la temperatura media, con la cual

no corresponde. Dentro de cada período, las temperaturas están más vinculadas entre ellas que de un período a otro. El conjunto de temperaturas puede ser resumido por dos variables sintéticas : la temperatura media anual y la amplitud térmica. Con la ayuda de estas dos variables, podemos esbozar una tipología de las ciudades. Reagrupando las ciudades a la vez próximas en el primer plano factorial y respetando la situación geográfica, podemos proponer la tipología siguiente :

- Ciudades de Europa del Sur caracterizadas por temperaturas elevadas a lo largo del año ;
- Ciudades de Europa del oeste caracterizadas por temperaturas medias durante todo el año ;
- Ciudades de Europa del Norte caracterizadas por temperaturas frías sobre todo en verano.
- Ciudades de Europa del Este caracterizadas por temperaturas frías sobre todo en invierno.

La ciudad de *Sarajevo* es una ciudad de Europa del Sur cuyo perfil de temperatura se parece más a las ciudades de Europa del Oeste que a las ciudades de Europa del Sur. Podemos anotar que las ciudades que no participaron en la construcción de los ejes (los individuos suplementarios del análisis) tienen un perfil de temperaturas próximo al de la capital del mismo país.

Las variables *Noviembre* y *Marzo* son muy correladas : en efecto, las puntas de las flechas son próximas del círculo de correlación, etonces el ángulo entre los vectores *Noviembre* y *Marzo* en el espacio $\mathbb{R}^K$ (espacio de los individuos) es próximo del ángulo en el plano, es decir, próximo de 0. Como el coeficiente de correlación es igual al coseno del ángulo en el espacio de los individuos, entonces el coeficiente de correlación es próximo de 1. Esto significa que las ciudades dónde hace frío en Noviembre son también ciudades en las que hace frío en Marzo.

La correlación entre *Enero* y *Junio* es próxima de 0 ya que en el eje, el ángulo es próximo de $\pi/2$ y las variables están bien proyectadas.

Para ir más lejos. Dos elipses de confianza pueden ser trazadas alrededor de las modalidades de una variable cualitativa suplementaria (*i.e.* alrededor del baricentro de los individuos que poseen la modalidad). Estas elipses son adaptadas a representaciones planas y permiten visualizar si dos modalidades son significativamente diferentes o no (cf. figura 1.24).

Para una modalidad, consideramos el vector de sus coordenadas y la matriz de varianza-covarianza asociada y consideramos que sus coordenadas siguen una ley multinormal. Esta hipótesis es razonable ya que se trabaja sobre baricentros y por consiguiente, sobre medias. Conociendo la ley de la posición de una modalidad, podemos trazar su elipse de confianza.

En la práctica, es necesario construir una tabla (`data.frame`) con la variable cualitativa y las coordenadas de los individuos en cada uno de los ejes factoriales. El cálculo de las elipses de confianza es efectuado y, por fin, las elipses son trazadas :

```
> concat.data <- cbind.data.frame(temperaturas[1:23,17],res$ind$coord)
> ellipse.coord <- coord.ellipse(concat.data,bary=TRUE)
> plot.PCA(res, habillage=17, ellipse=ellipse.coord, cex=0.8)
```

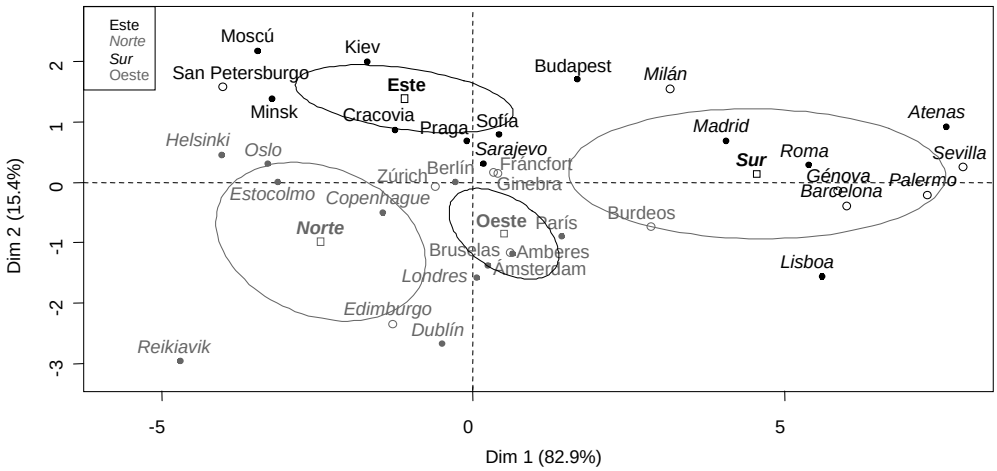

FIGURE 1.24 – Datos temperaturas : elipses de confianza alrededor de modalidades en el plano (1, 2).

1.11 Ejemplo : datos genómicos

1.11.1 Descripción de los datos y problemática

Cuarenta y tres pollos experimentaron uno de los seis regímenes siguientes : régimen normal (N), ayuno durante 16 horas (J16), ayuno durante 16 horas y realimentados 5 horas (J16R5), ayuno durante 16 horas y realimentados 16 horas (J16R16), ayuno durante 48 horas (J48), ayuno durante 48 horas y realimentados 24 horas (J48R24). Al final de este régimen, un análisis de los genes por chip ADN ha sido efectuado y la expresión de 7407 genes ha sido conservada para el conjunto de los pollos. Una selección de los genes ha sido efectuada por el biólogo ya que al principio, más de 20000 genes son medidos por los chips ADN. Después, los datos fueron pre-tratados de modo estándar para los chips ADN (estandarización, eliminación del efecto chip, etc.).

La tabla de datos que hay que analizar es una tabla rectangular con muchos menos individuos que variables : 43 líneas (pollos) y 7407 columnas (genes). Añadamos a esto la variable cualitativa *Régimen* que corresponde a una de las 6 situaciones de estrés o regímenes definidos anteriormente.

El objetivo del estudio es ver si los genes se expresan de modo distinto según la situación de estrés en la cual esté puesto el pollo. Más precisamente, puede ser interesante ver cuánto tiempo debe ser realimentado el pollo después de un ayuno antes de recobrar un estado normal, *i.e.*, un estado comparable al estado del pollo en un régimen normal. ¿Acaso algunos genes son subexpresados durante el ayuno y sobre expresados en el momento de la fase de realimentación ?

1.11.2 Elección del análisis

Elección de elementos activos

En este estudio, todos los pollos son considerados como individuos activos y todos los genes como variables activas. La variable *Régimen* es necesariamente ilustrativa ya que es cualitativa.

Estandarización o no estandarización de variables

Las variables son estandarizadas aquí para dar la misma influencia a cada gen.

1.11.3 Puesta en práctica

Un problema técnico puede presentarse para importar este tipo de juego de datos con muchas columnas ya que ciertas hojas de cálculo no soportan más de 128 columnas. Por es es frecuente que la tabla sea realizada con los genes en línea y los individuos en columna. La variable cualitativa *régimen* no debe tenerse en cuenta en esta tabla, si no, todas las variables serían consideradas como cualitativas en el momento de la importación (para una variable, si un dato es cualitativo, el conjunto de la variables se considera como cualitativo). Podemos importar la tabla (gen×pollo) y transponerla haciendo :

```
> pollos <- read.table("http://factominer.free.fr/libra/pollos.csv",header=TRUE,
    sep=";",dec=".",row.names=1)
> pollos <- as.data.frame(t(pollos))
```

Luego hay que concatenar la variable cualitativa *régimen* a esta tabla (después de haberla creado) :

```
> regimen <- as.factor(c(rep("N",6),rep("J16",5),rep("J16R5",8),rep("J16R16",9),
    rep("J48",6),rep("J48R24",9)))
> pollos <- cbind.data.frame(regimen,pollos)
> colnames(pollos)[1] <- "Régimen"
```

Luego podemos realizar el ACP y construir el gráfico de individuos coloreando los individuos en función de la variable régimen (aquí la primera variable de la tabla ; modificamos la talla de la fuente por el parámetro `cex` (`"cex=0.7"` en lugar de 1 por defecto) :

```
> res.pca <- PCA(pollos,quali.sup=1)
> plot(res.pca, habillage=1, cex=0.7)
```

El plano principal expresa 29.1% de la inercia total (cf. los gráficos o el objeto `res.pca$eig`). Anotar que aquí obtenemos como máximo 42 dimensiones, lo que corresponde al total de individuos −1 (y no al total de variables) : en efecto, los 43 individuos están como máximo en un subespacio de 42 dimensiones.
En el plano principal del ACP (cf. figura 1.25), la nube de observaciones (pollos) se divide en dos subgrupos. El primero, muy disperso, contiene todos los pollos que sufrieron un estrés muy fuerte, el segundo, concentrado y próximo del origen, contiene los pollos que no sufrieron estrés. Más detalladamente, el primer eje separa los pollos en tres grupos :

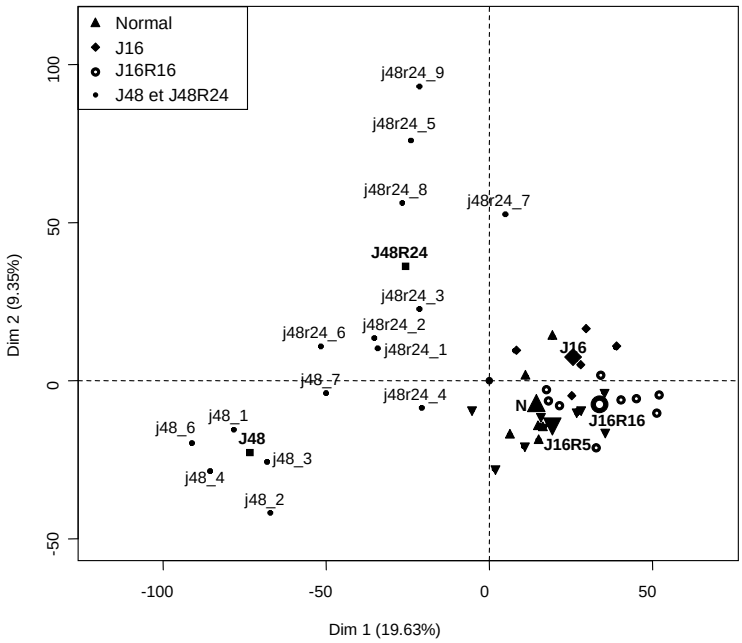

FIGURE 1.25 – Datos genómicos : gráfico de individuos en el primer plano.

los pollos que sufrieron un estrés muy fuerte pero no fueron realimentados (J48), los pollos que sufrieron un estrés muy fuerte y fueron realimentados (J48R24), y los otros pollos. Los pollos realimentados tienden a recuperarse del estrés muy fuerte y su estado de salud tiende a acercarse al de un pollo normal. Sin embargo, la realimentación durante 24 horas no es suficiente para que el estado del pollo vuelva a ser completamente normal. Esto significa que ciertos genes son específicos de un estado de estrés muy fuerte : ciertos genes son sobre expresados en estado de estrés cuando otros son subestimados (el gráfico de las variables muestra que ciertas variables son correladas negativamente cuando otras son correladas positivamente). El segundo eje es específico de los pollos J48R24.
El gráfico de las variables no es legible aquí debido a una gran cantidad de variables. Para representarlo y ver si existe una estructura sobre las variables, podemos representar un punto por variable (sin flecha y sin etiqueta) con el encargo :

```
> plot(res.pca, choix="var", invisible="var")
> points(res.pca$var$coord[,1:2], cex=0.5)
```

Esta nube presenta un ritmo regular que no necesita comentarios particulares (pero había que asegurarse de ello). Es entonces necesario caracterizar los ejes con la ayuda de la función **dimdesc** (damos aquí sólo las variables cuantitativas que más caracterizan las dimensiones y la totalidad de las modalidades que caracterizan las dimensiones) :

```
> dimdesc(res.pca,proba=1e-5)
 $Dim.1$quanti       $Dim.2$quanti       $Dim.3$quanti
```

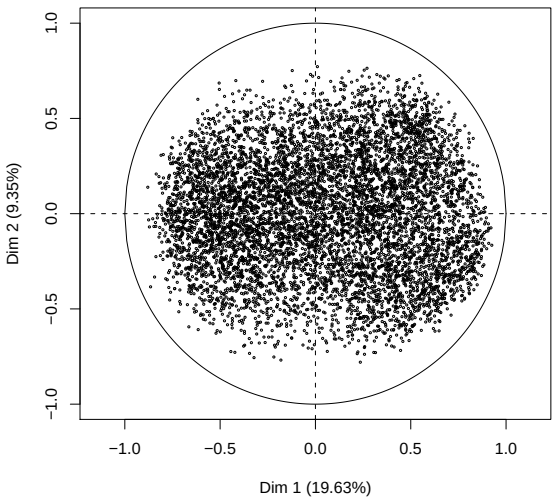

FIGURE 1.26 – Datos genómicos : gráfico de variables en el primer plano (un punto corresponde a una variable).

```
         Dim.1              Dim.2              Dim.3
                                 ⋮                  ⋮
HS2ST1    0.93     MPHOSPH9  0.77       AGL     -0.83
TTC151    0.92     BNC2      0.76       LRRC8A  -0.83
PRC1      0.91     XRCC6     0.75       ARFIP1  -0.84
KLHL101   0.91     FBXW4     0.75       PRDM161 -0.85
C6orf66   0.91     OFD11     0.75       PDE4B   -0.86
C16orf48  0.91     USP53     0.73
                   ⋮                    GLI2    -0.87
TTC8      0.91     ⋮         ⋮          PRKAA2  -0.87
KCNJ15    0.90     DNAH10   -0.75       PCSK51  -0.89
GRIA3     0.90     RHOT2    -0.76       DUSP14  -0.89
C7orf30   0.90     PDCD11   -0.77
 ⋮         ⋮                           HIPK2   -0.90
                   PHYHD1   -0.78

  $Dim.1$quali       $Dim.2$quali       $Dim.3$quali
         Dim.1              Dim.2              Dim.3
J16R16    2.98     J48R24    4.59     J16       3.58
J48R24   -2.24     J48      -2.25     J16R5     2.30
J48      -5.02                        N        -3.85
```

Los genes más correlados al primer eje son todos correlados positivamente : estos genes son subexpresados cuando los pollos sufrieron un ayuno de 48 horas. Para el eje 2, ciertos genes son sobre expresados (MPHOSPH9, BNC2, etc.) cuando los pollos se realimentan después de un ayuno de 48 horas mientras que otros genes son subexpresados (PHYHD1, PDCD11, etc.). Evidentemente, aquí el estadista debe llamar al biólogo para analizar por qué son estos genes los que son subexpresados o sobre expresados. Varias modalidades de la variable *Régimen* son características de los ejes 1 y 2 : encontramos aquí el resultado visualizado en el plano, pero con un test (mientras que visualmente no podemos decir si las diferencias son

significativas o no). Los pollos que sufrieron un estrés durante 48 horas (realimentados o no) tienen una coordenada significativamente más débil que otros sobre el eje 1, mientras que los pollos que sufrieron un estrés durante 16 horas y que fueron realimentados 16 horas tienen una coordenada significativamente positiva. El eje 2 separa los pollos que sufrieron un estrés durante 48 horas : este eje opone los pollos realimentados (con coordenada significativamente positiva) con los pollos no realimentados (con un coordenada significativamente negativa). También es posible visualizar el plano 3-4 del ACP :

```
> plot(res.pca, habillage=1, axes=3:4)
> plot(res.pca, choix="var", invisible="var", axes=3:4)
> points(res.pca$var$coord[,3:4], cex=0.5)
```

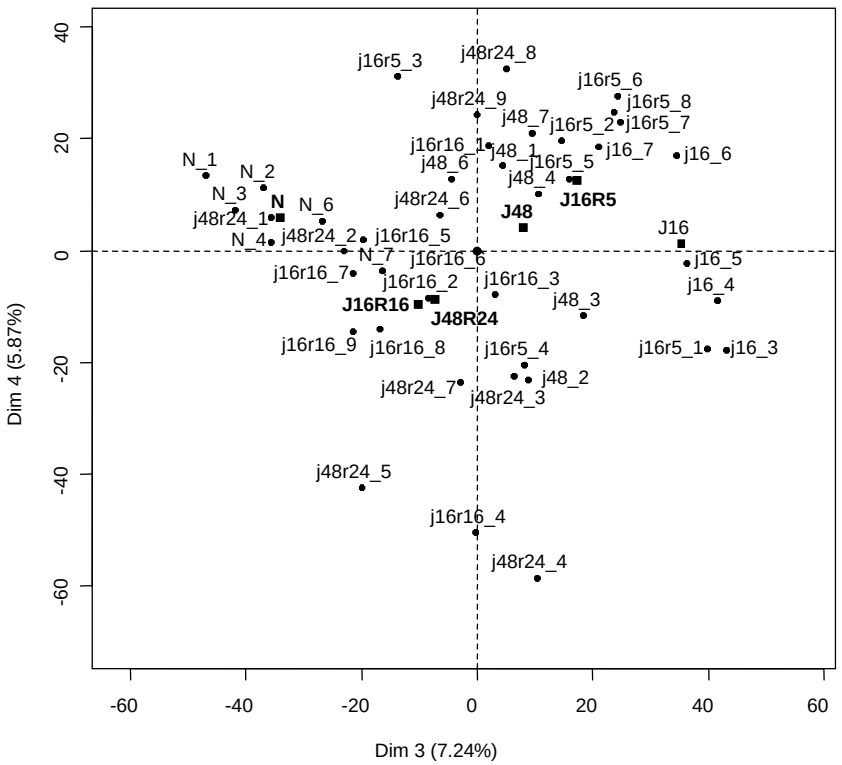

FIGURE 1.27 – Datos genómicos : gráfico de individuos en el plano 3-4.

Este plano 3-4 (cf. figura 1.27), y más particularmente el eje 3, separa los regímenes no diferenciados del primer plano. Los pollos que siguieron un régimen normal tienen coordenadas negativas en el eje 3 y los pollos que sufrieron un estrés durante 16 días tienen coordenadas positivas en el eje 3. Los pollos realimentados después de un estrés de 16 días están entre estos dos grupos, con un gradiente según el tiempo de realimentación : los pollos realimentados 5 horas están más próximos de los pollos no realimentados y los pollos realimentados

16 horas están más próximos de los pollos que no sufrieron estrés. Parece pues que ciertos genes sean expresados de otro modo según si hubo un estrés durante 16 horas o no, y ciertos genes toman poco a poco una expresión «normal». Sin embargo, incluso después de 16 horas de realimentación, los genes no funcionan todavía de modo normal.

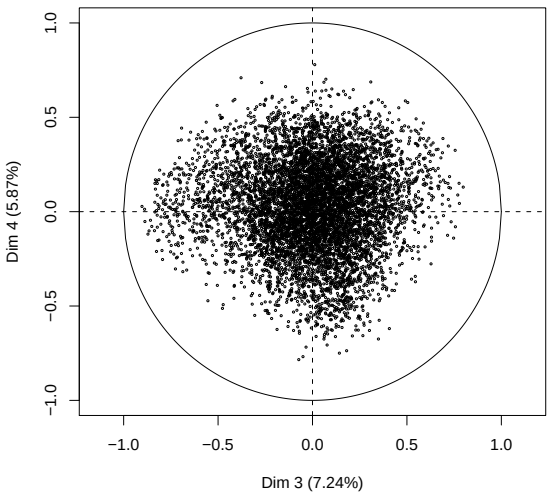

FIGURE 1.28 – Datos genómicos : gráfico de variables en el plano 3-4 (un punto corresponde a una variable).

Igual que para el primer plano, la nube de variables en el plano 3-4 presenta un ritmo regular que no necesita comentarios particulares. Es más fácil caracterizar los ejes de modo automático con la ayuda del procedimiento **dimdesc**. La variable HIPK2, DUSP14 o todavía PCSK51 caracterizan el eje 3 : son los genes más vinculados al eje (correlación negativa). Las modalidades que caracterizan el eje son los pollos que no sufrieron estrés (con una coordenada significativamente negativa), así como los pollos que tuvieron un estrés durante 16 horas y que no se realimentaron, y en un menor grado, los pollos que tuvieron un estrés durante 16 horas y que fueron realimentados 5 horas (con coordenada significativamente positiva).

Es posible construir elipses de confianza alrededor de los baricentros de la representación del conjunto de los pollos que siguieron el mismo régimen. Para eello, creamos una tabla con la variable *Régimen* y las coordenadas de los individuos de los ejes 1 y 2 ; luego calculamos las coordenadas de las elipses (con la función **coord.elipse** y el argumento `bary=TRUE` para precisar que son elipses construidas alrededor de baricentros) antes de construir el gráfico del ACP (con la función **plot.PCA** y el argumento `ellipse=bb` que precisa que las coordenadas de las elipses están en el objeto `bb`) :

```
> aa <- cbind.data.frame(pollos[,1], res.pca$ind$coord[,1:2])
> bb <- coord.ellipse(aa,bary=TRUE)
> plot.PCA(res.pca, habillage=1, ellipse=bb)
```

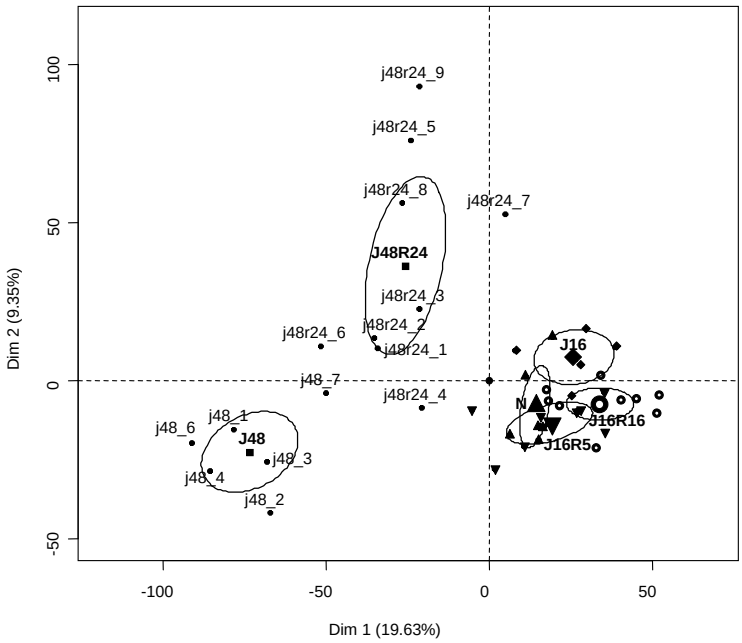

FIGURE 1.29 – Datos genómicos : elipses de confianza alrededor de las modalidades de la variable *Régimen* en el plano 1-2.

Estas elipses de confianza (cf. figura 1.29) confirman la impresión visual de que los regímenes de estrés importantes (J48 y J48R24) son muy diferentes de los otros. Del mismo modo, las elipses de confianza son disjuntas para los regímenes J16 y J16R16, para los regímenes J16R16 y N o para los regímenes J16 y J16R5 : esta diferenciación de los regímenes no era nada evidente sin las elipses de confianza.
Para tener las elipses de confianza en los ejes 3-4, creamos una tabla con la variable *Régimen* y las coordenadas de los individuos en los ejes 3 y 4, calculamos el trazado de las elipses y dibujamos el plano 3-4 del ACP con el añadido de las elipses :

```
> aa <- cbind.data.frame(pollos[,1], res.pca$ind$coord[,3:4])
> bb <- coord.ellipse(aa, bary=TRUE)
> plot.PCA(res.pca, habillage=1, ellipse=bb, axes=3:4)
```

En el plano 3-4, varias modalidades de la variable*Régimen* están bien diferenciadas (cf. figura 1.30) : el régimen N es diferente de todos los demás regímenes y particularmente del régimen J16R16 ; esto quiere decir que los pollos que sufrieron un estrés durante 16 horas y que fueron realimentados 16 horas no se recuperaron del estrés.

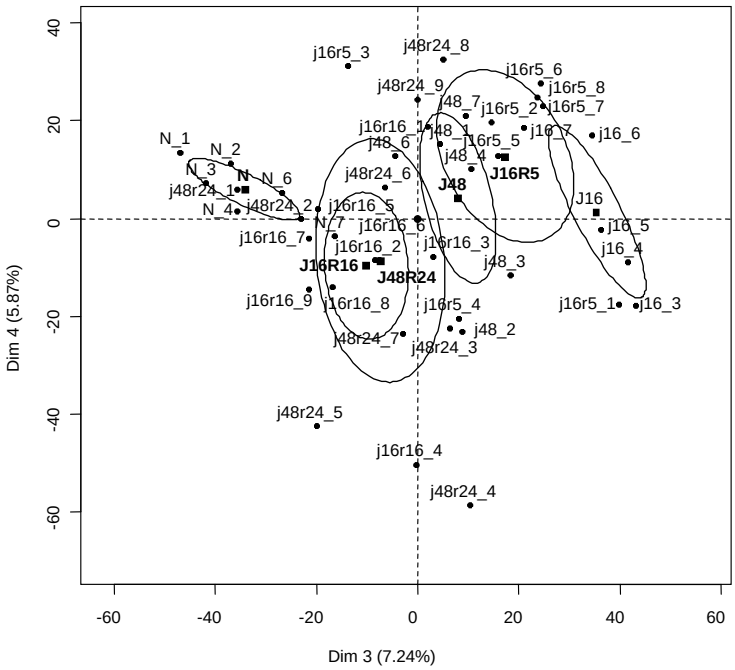

FIGURE 1.30 – Datos genómicos : elipses de confianza alrededor de las modalidades de la variable*Régimen* en el plano 3-4.

Chapitre 2

Análisis factorial de las correspondencias (AFC)

2.1 Datos y notaciones

Para ilustrar este capítulo, utilizamos una pequeña tabla de datos extraídos de los resultados de una encuesta antigua pero notable (Tabard, 1974) [1]. Se interrogó a 1724 mujeres con la ayuda de un largo cuestionario que contiene, entre otras cosas, gran cantidad de preguntas relativas a su actitud con respecto al trabajo femenino. Estos datos presentan un carácter histórico sobre el plano sociológico ; el fin de los años sesenta ve el resultado de varios combates de las élites feministas, particularmente el acceso al trabajo asalariado de las mujeres (que en Francia pueden trabajar sin el consentimiento de su marido sólo desde 1965) ; de ahí el interés de estudios de opinión de las mujeres en general en este momento sobre tal tema. De este conjunto, se extrajeran dos preguntas, cuya redaccion, así como las respuestas, respectivas, aparecen en la tabla 2.1.
De este conjunto se extrajeran dos preguntas, cuya redaccion asi como las respuestas, respectivas, aparecen en la tabla

Imagen ideal que usted se hace de una familia :	Actividad que conviene más a una madre de familia cuando los niños son pequeños			
	Quedarse en el hogar	Trabajo de medio tiempo	Trabajo de tiempo completo	Suma
Ambos cónyuges trabajan por igual	13	142	106	261
Trabajo del marido más absorbente	30	408	117	555
Sólo el marido trabaja	241	573	94	908
Suma	284	1123	317	1724

Tabla 2.1 – Tabla que cruza las respuestas a dos preguntas de opinión.

Esta tabla se llama «tabla cruzada» en la terminología usual de los organismos que realizan encuestas y «tabla de contingencia» en la jerga de los estadistas. La tabla reagrupa las

1. N. Tabard (1974). Necesidades y aspiraciones de las familias y de los jóvenes. París : CREDOC.

respuestas simultáneas a las dos preguntas : así, 241 es el número de personas que ha respondido a la vez *Sólo el marido trabaja* a la pregunta de la familia ideal y *Quedarse en el hogar* a la pregunta de la actividad de una madre de familia. Esta tabla es completada por la suma de los términos de una misma fila (908 personas respondieron *Sólo el marido trabaja* ; estos números constituyen el margen columna) o de una misma columna (284 personas respondieron *Quedarse en el hogar* ; estos números constituyen el margen fila). La redacción exacta de las preguntas es la siguiente :
Entre los tres modelos siguientes, ¿cuál es el que más se acerca de la imagen ideal que usted se hace de una familia ? :

1. Una familia donde ambos cónyuges tienen una profesión que los absorbe tanto al uno como al otro y donde los quehaceres domésticos y el cuidado de los niños son compartidos entre los dos.
2. Una familia donde la mujer tiene una profesión menos absorbente que la del hombre y donde ella se coupa de una parte más grande de los quehaceres domésticos y del cuidado de los niños.
3. Una familia donde sólo el hombre ejerce una profesión y donde la mujer se queda en el hogar.

Distinguiendo el periodo de cuando los niños son pequeños y el periodo de donde todos los niños van a la escuela, ¿cuál es, según usted, el tipo de actividad que conviene mejor a una madre de familia :

1. Quedarse en el hogar.
2. Trabajo de medio tiempo.
3. Trabajo de tiempo completo.

Más generalmente, una tabla de contingencia está construida del modo siguiente (cf. figura 2.1). Disponemos para n individuos de su valor para dos variables cualitativas anotadas $V1$ (presentando I modalidades o niveles) y $V2$ (presentando J modalidades). La tabla de contingencia tiene como término general x_{ij}, número de individuos que posee la modalidad i de $V1$ y j de $V2$.
Los márgenes de la tabla se anotan reemplazando por un punto, en x_{ij}, el índice sobre el cual se efectúa la suma. Así :

$$x_{i\bullet} = \sum_{j=1}^{J} x_{ij} \qquad x_{\bullet j} = \sum_{i=1}^{I} x_{ij} \qquad n = x_{\bullet\bullet} = \sum_{i,j} x_{ij}.$$

Para terminar, en el análisis factorial de correspondencias (AFC), consideramos la tabla de probabilidades[2] asociada a la tabla de contingencia, de término general $f_{ij} = x_{ij}/n$, probabilidad de poseer a la vez las modalidades i (de $V1$) y j (de $V2$). Los márgenes de esta tabla, llamadas también probabilidades marginales, se definen por

$$f_{i\bullet} = \sum_{j=1}^{J} f_{ij} \qquad f_{\bullet j} = \sum_{i=1}^{I} f_{ij} \qquad f_{\bullet\bullet} = \sum_{i,j} f_{ij} = 1.$$

2. En este ejemplo, el término «probabilidad» puede parecer abusivo, ya que designa una cantidad establecida a partir de una muestra. Pero además de que es cómodo, el término corresponde al hecho de que en el AFC los datos son considerados poblaciones, es decir, sin aspectos inferenciales.

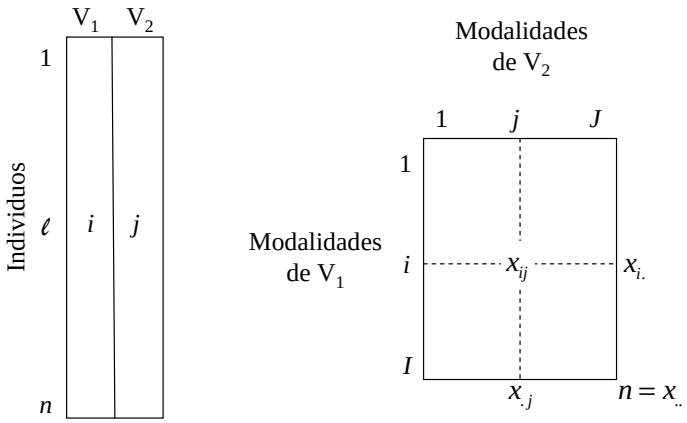

FIGURE 2.1 – Notaciones generales para una tabla de contingencia que cruza dos variables cualitativas ($V1$ y $V2$) definidas sobre n individuos ; el individuo l posee las modalidades i (para $V1$) y j (para $V2$) : es contabilizado en x_{ij}.

Observación
El término «análisis de correspondencias» procede del hecho de que se analiza una tabla que pone en correspondencia dos conjuntos : el representado por las filas y el representado por las columnas (desempeñan papeles simétricos).

2.2 Objetivos y modelo de independencia

2.2.1 Objetivos

La restitución usual de los resultados de una encuesta se resume por lo general en la enumeración de las respuestas a algunas preguntas (juiciosamente) elegidas. En el ejemplo, las respuestas a la pregunta 1 (sobre la familia ideal) muestran con claridad que las mujeres (en Francia y en 1970, precisión que no se repetirá siempre con el fin de no volver pesado el texto) son mayoritariamente hostiles hacia el trabajo femenino (52.7% escogió *Sólo el marido trabaja*) ; pero las respuestas a la pregunta 2 (sobre la actividad de una madre de familia) muestran también que las mujeres son mayoritariamente favorables al trabajo femenino (apenas 16.47% de ellas escogió *Quedarse en el hogar*). Podremos afirmar que la enumeración de las respuestas a una sola pregunta de opinión sólo puede llevarnos a resultados muy frágiles (algunos dicen sin interés). Es, pues, necesario tomar en consideración simultáneamente las respuestas a varias preguntas (dos en este capítulo ; más de dos en el capítulo siguiente, dedicado al análisis de correspondencias múltiples). En nuestro ejemplo, esperamos que el cruce de respuestas a ambas preguntas nos ayude a comprender las imágenes contradictorias aportadas por cada una de estas dos preguntas.
De modo empírico, el análisis de esta tabla consiste en comparar los números. Si se tratase de una simple serie numérica (*i.e.*, nueve números no estructurados en filas y en columnas) enfocaríamos la atención en los valores más grandes y en los más pequeños. Así, el valor

más grande de la tabla, 573, parece sugerir una «atracción», término que queda por definir precisamente entre las modalidades *Sólo el marido trabaja* y *Trabajar de medio tiempo*, lo que parece confirmar el hecho de que *Trabajar de medio tiempo* es la respuesta más frecuente entre las personas que han respondido *Sólo el marido trabaja* y recíprocamente. Pero la consulta de los márgenes muestra que estas dos respuestas son, por separado, ampliamente mayoritarias. De ahí la pregunta : ¿el valor elevado 573 no se debe al hecho de que estas respuestas son cada una por separado muy frecuentes, más bien que una «atracción» entre estas modalidades ? Resulta que los números de una tabla de contingencia se pueden comparar entre ellos sólo recordando los márgenes que les corresponden. El análisis de tal tabla no es simple : se requieren una formalización del objetivo del estudio y una metodología adaptada.

2.2.2 Modelo de independencia y test de χ^2

El mismo principio de construcción de una tabla de contingencia (cruce de dos variables cualitativas) implica que el estudio de tal tabla tiene como objeto la relación entre las respuestas a dos preguntas. Indiquemos de entrada que aquí, como en la inmensa mayoría de las tablas sometidas a un AFC, estamos seguros de la existencia de una relación. Teniendo en cuenta el significado de las preguntas, demostrar la ausencia de relación en la tabla 2.1, vía un test de χ^2, sería un scoop o, más probablemente, dejaría dudas sobre la calidad de los datos.

Estudiar la relación entre dos variables vuelve a situar los datos en relación con una situación de referencia que es la ausencia de relación. El modelo de independencia especifica esta situación de referencia. La relación usual de independencia entre dos acontecimientos (P[A y B] =P[A] P[B]) se extiende directamente a dos variables cualitativas. Dos variables cualitativas son independientes si verifican :

$$\forall i, j \quad f_{ij} = f_{i\bullet} f_{\bullet j}.$$

Así, la independencia estipula que la probabilidad conjunta (f_{ij}) depende sólo de probabilidades marginales ($f_{i\bullet}$ y $f_{\bullet j}$), lo que está de acuerdo con nuestro comentario sobre el número 573.

Estudiar una relación equivale a comparar los efectivos observados ($x_{ij} = nf_{ij}$) y los efectivos teóricos correspondientes al modelo de independencia ($nf_{i\bullet}f_{\bullet j}$). La tabla 2.2 reagrupa estas dos tablas para nuestro ejemplo.

Comentemos algunas diferencias entre la tabla de los efectivos observados y la tabla de los efectivos teóricos :

- 13 personas respondieron a la vez *Ambos cónyuges trabajan por igual* y*Quedarse en el hogar* : si las preguntas fueran independientes, observaríamos (por término medio) a 43 personas que hubiesen dado esta pareja de respuestas. El efectivo observado es sensiblemente inferior al efectivo teórico, lo que se esperaba dado el significado de estas respuestas. Decimos que estas modalidades se rechazan : cuando escogemos una modalidad, tendemos a no escoger la otra.
- 241 personas respondieron a la vez *Sólo el marido trabaja* y *Quedarse en el hogar*, valor sensiblemente superior al efectivo teórico 149.6 obtenido (por término medio) con la hipótesis de independencia (aquí todavía este resultado es esperado visto el significado

Efectivos observados

	Quedarse en el hogar	Trabajo de medio tiempo	Trabajo de tiempo completo	Suma
Dos cóny. trab. por igual	13	142	106	261
Trab. marido + absorbente	30	408	117	555
Sólo el marido trab.	241	573	94	908
Suma	284	1123	317	1724

Efectivos teóricos

	Quedarse en el hogar	Trabajo de medio tiempo	Trabajo de tiempo completo	Suma
Dos cóny. trab. por igual	43,0	170,0	48,0	261
Trab. marido + absorbente	91,4	361,5	102,1	555
Sólo el marido trab.	149,6	591,5	167,0	908
Suma	284	1123	317	1724

Tabla 2.2 – De los efectivos observados a los efectivos teóricos.

de estas respuestas). Decimos que estas modalidades se atraen : cuando escogemos una, tendemos a escoger la otra.

– 573 personas respondieron a la vez *Sólo el marido trabaja* y *Trabajar de medio tiempo*, efectivo inferior (muy ligeramente) al efectivo 591.5 teórico.

Este último resultado es muy interesante desde el punto de vista metodológico. El valor más elevado de la tabla, es 573, lo que para un observador superficial, sugiere una atracción entre estas dos respuestas. De hecho, no es nada de eso ya que, al contrario, estas modalidades se rechazan (muy ligeramente). El valor fuerte (573) puede atribuirse, entonces, al hecho de que ambas modalidades (consideradas por separado) son muy frecuentes (respectivamente 52.7 y 65.1% de las respuestas) y no al hecho deque se atraigan. Este resultado, que podía presentir, está aquí claramente cuantificado gracias a la formalización (relación entre dos variables ; desviación del modelo de independencia).

El criterio χ^2 permite someter a un test la significación de la distancia global entre la tabla observada y el modelo de independencia. Se escribe :

$$\begin{aligned}\chi^2 &= \sum_{i,j} \frac{(\text{efectivos observados} - \text{efectivos teóricos})^2}{\text{efectivos teóricos}}, \\ &= \sum_{i,j} \frac{(nf_{ij} - nf_{i\bullet}f_{\bullet j})^2}{nf_{i\bullet}f_{\bullet j}} = n\sum_{i,j} \frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}} = n\Phi^2,\end{aligned}$$

donde Φ^2 corresponde a una medida de relación independiente del efectivo y a una inercia total (ver más lejos). En el ejemplo, el χ^2 vale 233.43, valor altamente significativo (probabilidad crítica de ser sobrepasado : 2.4×10^{-49}), resultado esperado dado la significación de las preguntas. El detalle del cálculo (cf. tabla 2.3) pone en evidencia la contribución de las celdas aparte de la independencia (es la asociación entre *Los dos cónyuges trabajan por igual* y *Trabajar de tiempo completo* que expresa el valor que más se aleja de la hipótesis de independencia : 30.04 del total) pero también la de las filas y de las columnas (observamos la débil contribución, 4.78%, de *Trabajar de medio tiempo*).

	Quedarse en el hogar	Trabajo de medio tiempo	Trabajo de tiempo completo	Suma
Dos cóny. trab. por igual	20,93	4,62	70,12	95,66
Trab. marido + absorbente	41,27	5,98	2,19	49,44
Sólo el marido trab.	55,88	0,58	31,88	88,34
Suma	118,07	11,17	104,19	233,43

	Quedarse en el hogar	Trabajo de medio tiempo	Trabajo de tiempo completo	Suma
Dos cóny. trab. por igual	8,96	1,98	-30,04	40,98
Trab. marido + absorbente	17,68	-2,56	-0,94	21,18
Sólo el marido trab.	-23,94	0,25	13,66	37,84
Suma	50,58	4,78	44,63	100,00

Tabla 2.3 – Descomposición de χ^2, por celda, fila y columna (valores brutos y porcentajes). Cuando el efectivo observado es inferior al efectivo teórico, añadimos el signo − a cada valor.

2.2.3 Modelo de independencia y AFC

El análisis de una tabla de contingencia debe hacerse, pues en referencia a la situación de independencia. Es lo que hace el AFC al escribir el modelo de independencia en la forma siguiente :

$$\forall i, j \quad \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}.$$

La cantidad $f_{ij}/f_{i\bullet}$ es la probabilidad condicional de poseer la modalidad j (de la variable 2) cuando se posee la modalidad i (de la variable 1). Hay independencia cuando, para todas las celdas, la probabilidad condicional es igual a la probabilidad marginal. Este punto de vista sobre la independencia es próximo a la intuición : hay independencia si la probabilidad de poseer j (de $V2$) no depende de la modalidad poseída por $V1$.
De modo simétrico, el modelo de independencia puede escribirse así :

$$\forall i, j \quad \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}.$$

El AFC considera simultáneamente ambas escrituras, utilizando la terminología de perfil-fila $\{f_{ij}/f_{i\bullet}\ ;\ j = 1, J\}$, perfil-columna $\{f_{ij}/f_{\bullet j}\ ;\ i = 1, I\}$ y perfil medio (fila o columna) para la distribución de toda la población para una variable, sean $\{f_{i\bullet}\ ;\ i = 1, I\}$ y $\{f_{\bullet j}\ ;\ j = 1, J\}$. El modelo de independencia estipula, pues, que los perfiles-filas, por una parte, y los perfiles-columnas, por otra parte, son iguales al perfil medio correspondiente.

2.3 Las nubes y su ajuste

2.3.1 Nube de perfiles-filas

A partir de la tabla de los perfiles-filas construimos una nube de puntos (N_I), en el espacio $\mathbb{R}^J$, donde cada dimensión corresponde a una modalidad de la variable $V2$. Esta construcción es completamente análoga a la de la nube de los individuos en ACP. A cada fila i le

corresponde un punto cuya coordenada para la $j$ésima dimensión es $f_{ij}/f_{i\bullet}$; esta nube es completada por el punto medio (G_I), cuya $j$ésima coordenada vale$f_{\bullet j}$ (cf. figura 2.2).
Además de la transformación en perfiles, en relación con la nube de los individuos en ACP, la nube de las filas en AFC presenta las dos particularidades esenciales siguientes :

1. Cada punto i es afectado por el peso $f_{i\bullet}$; este peso es impuesto y es una parte integral del AFC ; a perfil igual, damos a una modalidad una influencia tan grande como frecuente ; con estos pesos, el perfil medio (G_I) es el centro de gravedad de N_I. Este punto G_I es tomado como origen de los ejes (como en ACP para los individuos).
2. La distancia de la que se provee el espacio $\mathbb{R}^J$ consiste en dar el peso $1/f_{\bullet j}$ a la dimensión j. El cuadrado de la distancia (dicha de χ^2) entre los puntos i y l se escribe :

$$d^2_{\chi^2}(i,l) = \sum_{j=1}^{J} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{lj}}{f_{l\bullet}} \right)^2 .$$

La principal justificación de esta distancia es indirecta y reside en la propiedad siguiente. Teniendo en cuenta el peso $f_{i\bullet}$, la inercia del punto i con respecto a G_I se escribe :

$$\begin{aligned} \text{Inercia}(i/G_I) = f_{i\bullet}\ d^2_{\chi^2}(i,G_I) &= f_{i\bullet} \sum_{j=1}^{J} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 , \\ &= \sum_{j=1}^{J} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}} . \end{aligned}$$

A parte del coeficiente n, reconocemos la contribución de la fila i al χ^2, de ahí el nombre de distancia de χ^2. De esto resulta que la inercia total de la nube de puntos N_I en relación con G_I es igual (excepto el coeficiente n) al criterio χ^2 (o si se prefiere, esta inercia es igual a Φ^2). Examinar la dispersión de N_I alrededor de G_I lleva a estudiar la diferencia entre los datos y el modelo de independencia. Esto es lo que hace el AFC, poniendo en evidencia las direcciones de inercia más grande de N_I.

Observación sobre la inercia total de N_I.
Esta inercia, igual a Φ^2, es una información importante, pues mide la intensidad de la relación entre ambas variables que se cruzan para obtener la tabla de contingencia. Hay aquí una gran diferencia con el ACP normado, en el cual la inercia total, igual al número de variables, sólo depende del formato de los datos y no de los datos mismos.

2.3.2 Nube de perfiles-columnas

En una tabla de contingencia, filas y columnas desempeñan papeles simétricos : podemos estudiar indistintamente $V1 \times V2$ o $V2 \times V1$. Es aquí donde hay una mayor diferencia con ACP, en el cual, filas (individuos) y columnas (variables) no se analizan del mismo modo ; así, por ejemplo, calculamos distancias entre individuos y correlaciones entre variables. El AFC construye también la nube de perfiles-columnas de modo perfectamente simétrico con respecto al utilizado para los perfiles-filas. O (cf. figura 2.3) :

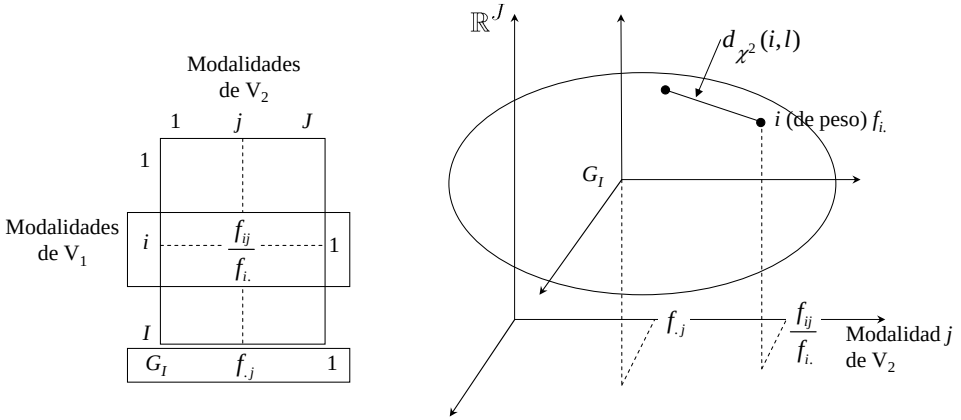

FIGURE 2.2 – Nube de perfiles-filas en AFC.

1. Consideramos los perfiles-columnas (así, según nos interesemos por las filas o por las columnas, no analizamos la misma tabla, $f_{ij}/f_{i\bullet}$ en un caso, $f_{ij}/f_{\bullet j}$ en el otro ; es aquí donde hay una mayor diferencia con el ACP, en el cual la misma transformación de los datos –el centrado y la reducción– se utiliza tanto para estudiar los individuos como las variables).
2. A cada columna le corresponde un punto de $\mathbb{R}^I$, cuya coordenada sobre la dimensión i es $f_{ij}/f_{\bullet j}$; estos puntos constituyen la nube N_J.
3. Cada punto j es afectado por un peso de $f_{\bullet j}$; con estos pesos, el centro de gravedad de la nube, anotado G_J, es igual al perfil medio. Situamos el origen de los ejes G_J.

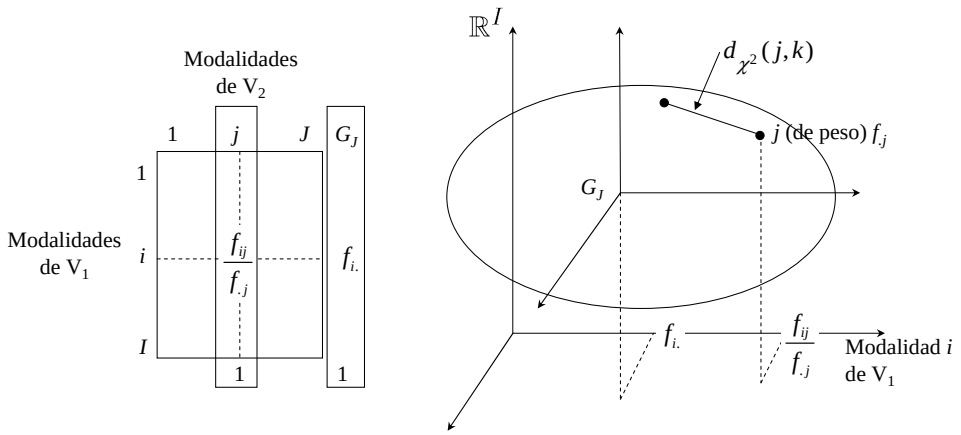

FIGURE 2.3 – Nube de perfiles-columnas en AFC.

En $\mathbb{R}^I$, la distancia afecta un peso de $1/f_{i\bullet}$ a la i^e dimensión. La distancia (al cuadrado) entre dos columnas j y k se escribe :

$$d^2_{\chi^2}(j,k) = \sum_{i=1}^{I} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ik}}{f_{\bullet k}} \right)^2 .$$

La inercia de la columna j con respecto al punto G_J se escribe :

$$\begin{aligned} \text{Inercia}(j/G_J) = f_{\bullet j}\ d^2_{\chi^2}(j,G_J) &= f_{\bullet j} \sum_{i=1}^{I} \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - f_{i\bullet} \right)^2 \\ &= \sum_{j=1}^{J} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}. \end{aligned}$$

Reconocemos la contribución (excepto el coeficiente n) de la columna j al χ^2. La inercia total de N_J es, pues, la misma que la de N_I ($= \frac{1}{n}\chi^2$) : estudiar la dispersión de N_J alrededor de G_J lleva a estudiar la relación entre ambas variables $V1$ y $V2$.

2.3.3 Ajuste de las nubes N_I y N_J

Procedemos como para el ajuste de la nube de individuos en ACP (cf. § 1.3.2, p. 6). Las etapas son descritas a continuación para la nube de los perfiles-filas.
El origen de los ejes está situado en el centro de gravedad G_I de la nube N_I, evolucionando en $\mathbb{R}^J$. Buscamos una serie de ejes ortogonales de inercia máxima. Sea u_s el vector unitario del eje de rango s y H_i^s la proyección del perfil i sobre este eje u_s que hace máxima la cantidad siguiente :

$$\sum_{i=1}^{I} f_{i\bullet} \left(OH_i^s\right)^2 \quad \text{máximo.}$$

La nube N_I es proyectada sobre los ejes u_s. Representamos estas proyecciones sobre planos que asocian dos ejes, en primer lugar el plano (u_1,u_2). Como en ACP, a causa de la ortogonalidad entre los ejes, este primer plano también hace máxima la inercia proyectada de N_I ; es decir, obtenemos el mismo plano buscando directamente (y no eje por eje) el plano de inercia máxima. Esta propiedad se llama «encaje de soluciones» : en el sentido de la inercia proyectada, el mejor eje es incluido en el mejor plano.
La inercia total mide la intensidad de la relación (en el sentido de Φ^2) entre las dos variables $V1$ y $V2$, en tanto que la inercia λ_s asociada al eje de rango s mide la parte de la relación expresada por este eje. La naturaleza de esta parte de la relación puede describirse mediante las coordenadas de los perfiles-filas : la distancia de un perfil al origen es una desviación al perfil medio y, una contribución a la relación entre $V1$ y $V2$. La proximidad entre dos perfiles-filas i y l expresa el mismo modo de desviarse del perfil medio : estas modalidades i y l (de $V1$) se asocian de modo privilegiado (*i.e.*, más que si hubiese independencia) a las mismas modalidades de $V2$. Paralelamente, son las mismas modalidades de $V2$ con las cuales i y l se asocian menos que en el modelo de independencia. El alejamiento del origen de dos perfiles-filas i y l expresa dos modos opuestos de desviarse del perfil medio : las

modalidades de $V2$ con las cuales i se asocia de modo privilegiado son aquellas con las que l se asocia menos que si hubiese independencia.
El AFC procede de modo simétrico para ajustar la nube N_J. Las principales etapas se resumen a continuación. En $\mathbb{R}^I$, el origen de los ejes está situado en G_J, centro de gravedad de N_J. Buscamos una continuación de los ejes ortogonales de inercia máxima. Sea v_s el vector unitario del eje de rango s y H_j^s la proyección del perfil j sobre este eje v_s que hace máxima la cantidad siguiente :

$$\sum_{j=1}^{J} f_{\bullet j} \left(OH_j^s\right)^2 \quad \text{máximo.}$$

La nube N_J es proyectada sobre los planos factoriales constituidos por parejas $(v_s,\ v_t)$, principalmente el primero de ellos $(v_1,\ v_2)$.

Observación sobre el número de ejes.
La nube N_I evoluciona en el espacio $\mathbb{R}^J$ a J dimensiones. Por lo tanto podemos pensar que, por regla general, J ejes son necesarios para representarla perfectamente. De hecho, otros dos elementos intervienen :
– La suma de las coordenadas de un perfil es igual a 1 ; la nube N_I pertenece, pues, a un subespacio de dimensión $J-1$.
– La nube N_I contiene I puntos ; siempre es posible representarlos todos con $I-1$ dimensiones.
Así, el número máximo de ejes necesarios para representar N_I perfectamente es pues $\min\{(I-1),(J-1)\}$. Obtenemos el mismo valor razonando a partir de N_J.

Observación sobre la puesta en práctica de los cálculos.
Podemos mostrar que la base del AFC es una diagonalización de matriz cuyos valores propios son las inercias proyectadas, de ahí la terminología «valor propio», que se encuentra en los listados en lugar de «inercia proyectada» : como son inercias, estos valores propios son positivos (veremos que son inferiores a 1) y los clasificamos por orden decreciente (el primer eje corresponde a la inercia proyectada máxima). Las coordenadas, filas y columnas se deducen de vectores propios asociados a estos valores propios. La dimensión de esta matriz es $\min\{I,J\}$: el tiempo de cálculo depende pues, principalmente, de la dimensión más pequeña de la tabla analizada (como en ACP).

2.3.4 Ejemplo : la actitud de las mujeres con respecto al trabajo femenino en Francia en 1970

El AFC aplicado a la tabla 2.1 conduce a dos gráficos reunidos en la figura 2.4. Vista la dimensión de la tabla (3×3), un plano es por construcción suficiente para representar perfectamente cada una de las nubes. Limitamos la interpretación al primer eje. Es indiferente comenzar el comentario por las filas o por las columnas. Apoyaremos la interpretación del AFC con las tablas de los perfiles-filas y de los perfiles-columnas (cf. tabla 2.4).

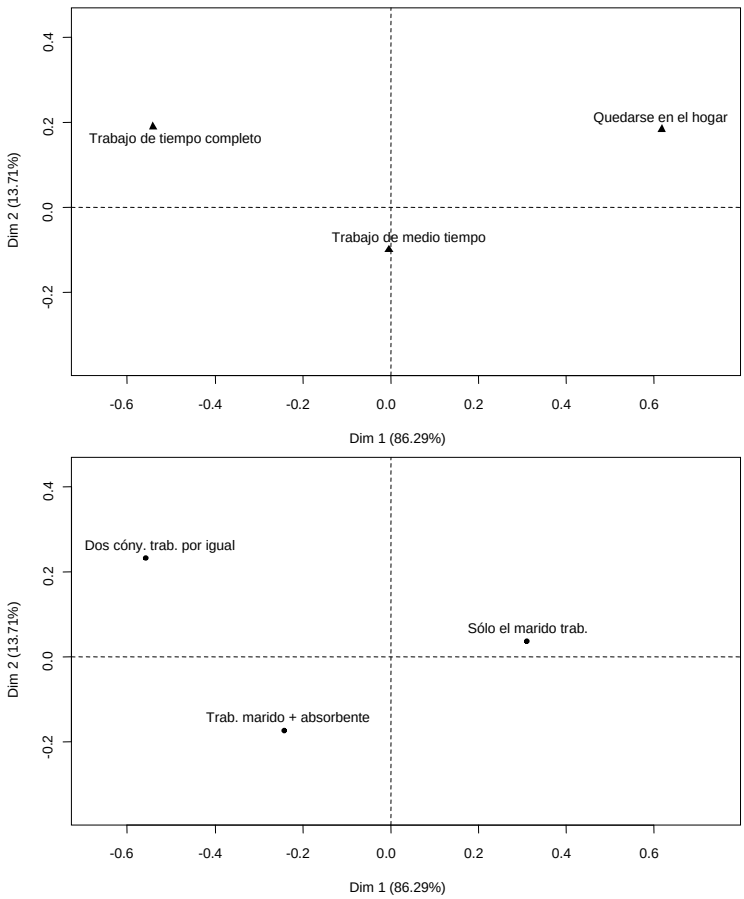

FIGURE 2.4 – Primer (y único) plano factorial procedente del AFC de la tabla 2.1. Arriba : representación de las columnas ; abajo : representación de las filas.

Representación de las columnas (actividad de la madre de familia)

El primer eje opone las modalidades *Quedarse en el hogar* y *Trabajar de tiempo completo*. Esta oposición sobre el gráfico implica necesariamente una oposición en cuanto al perfil ; así, las mujeres que han contestado *Quedarse en el hogar* (perfil-columna 1) responden :

– *Sólo el marido trabaja* más a menudo que el conjunto de la población (=perfil-columna medio) : 84.9% contra 52.7% ;
– *Ambos cónyuges trabajan por igual* menos a menudo que el conjunto de la población (4.6% contra 15.1%).

Recíprocamente, las mujeres que han respondido *Trabajar de tiempo completo* responden :

– *Sólo el marido trabaja* menos a menudo que el conjunto de la población (29.7% contra 52.7%) ;

Perfiles-filas

	Quedarse en el hogar	Trabajo de medio tiempo	Trabajo de tiempo completo	Suma
Dos cóny. trab. por igual	0,050	0,544	0,406	1,000
Trab. marido + absorbente	0,054	0,735	0,211	1,000
Sólo el marido trab.	0,265	0,631	0,104	1,000
Perfil medio	0,165	0,651	0,184	1,000

Perfiles-columnas

	Quedarse en el hogar	Trabajo de medio tiempo	Trabajo a tiempo completo	Perfil medio
Dos cóny. trab. por igual	0,046	0,126	0,334	0,151
Trab. marido + absorbente	0,106	0,363	0,369	0,322
Sólo el marido trab.	0,849	0,510	0,297	0,527
Suma	1,000	1,000	1,000	1,000

Tabla 2.4 – Perfiles-filas (A) y perfiles-columnas (B) de la tabla 2.1.

– *Ambos cónyuges trabajan por igual* más a menudo que el conjunto de la población (33.4% contra 15.1%).

Esta oposición entre perfiles es el aspecto más importante (ya que está bien valorada por el primer eje) de la desviación entre la tabla de contingencia y la independencia, o mejor, de la relación entre ambas variables.

Este aspecto concierne a las modalidades extremas (lo que podía esperarse razonablemente), esto es, la modalidad media desempeñando un papel neutro en esta oposición. Más generalmente, es decir, sobre el plano, la modalidad *Trabajar de medio tiempo* está muy próxima al centro de gravedad, lo que indica un perfil muy próximo del perfil medio (esto puede comprobarse directamente sobre la tabla y medirse por la contribución de esta modalidad al χ^2 : 4.78% ; cf. tabla 2.3). Esto se puede expresar así : el conjunto de las mujeres que han respondido *Trabajar de medio tiempo* no se distingue (desde el punto de vista de sus respuestas a la pregunta 1) del conjunto de la población. Esta formulación sugiere, a su turno, que la respuesta *Trabajar de medio tiempo* ha sido escogida en parte por el hecho de lo que N. Tabard llama su «carácter moderado» (en particular, por aquellas que han respondido *Sólo el marido trabaja*). Finalmente, esta respuesta no parece muy informativa : cuando una mujer contestó, esto no sugiere nada en cuanto a lo contestó a la pregunta 1 (técnicamente : la distribución condicional de *Trabajar de medio tiempo* es igual a la distribución marginal). La contradicción entre las imágenes dadas por ambas preguntas es así bien aclarada (hay que saber que las respuestas a la pregunta sobre la familia dan una imagen de los encuestados más desfavorable al trabajo femenino que las respuestas a la otra pregunta).

De modo sintético, podemos decir que el primer eje clasifica las modalidades de la segunda variable desde la más desfavorable al trabajo femenino hasta la más favorable. Como en ACP, es cómodo nombrar un eje por una (o algunas) palabra(s) que resume(n) el significado : aquí, es natural llamar este eje «actitud con respecto al trabajo femenino». La palabra «actitud» hay que considerarla en el sentido de los psicólogos, según el cual todo objeto (aquí el concepto de trabajo femenino) es dotado, para un individuo, de connotación (positiva o negativa) ; resulta que las opiniones a propósito de este objeto se organizan según tal valencia de un modo esencialmente unidimensional. La actitud (de un individuo con respecto a un

objeto) es su posicionamiento sobre dicha dimensión.

Representación de las filas (trabajo de los cónyuges)

El primer eje ordena las modalidades de la más desfavorable al trabajo femenino (*Sólo el marido trabaja*) a la más favorable (*Ambos cónyuges trabajan por igual*). Aquí todavía, sin ser un azar, podemos nombrar este eje «actitud con respecto al trabajo femenino». Podemos ilustrar dicha disposición refiriéndonos a los perfiles-filas : dejamos al lector el cuidado de hacerlo, al haber sido ilustrado el paso por las columnas. Observemos simplemente que la modalidad intermedia no está muy próxima al origen de los ejes (a diferencia de la nube de las columnas) sino claramente del lado favorable para el trabajo femenino.

2.3.5 Representación superpuesta de filas y columnas

Hasta aquí, consideramos por separado la nube de las filas N_I en $\mathbb{R}^J$ y la de las columnas N_J en $\mathbb{R}^I$. Cada una de estas nubes ha sido proyectada sobre sus direcciones de inercia máxima, proyecciones que se han comentado por separado ; tienen cada una su propia optimalidad (cada una hace máxima la inercia proyectada). Sin embargo en AFC, como en todo análisis factorial y entre otros, en ACP, el análisis de la nube de las filas por las partes y el de la nube de las columnas por otra parte están estrechamente vinculados por relaciones de dualidad. Dualidad, o carácter doble, proviene del hecho de que se analiza la misma tabla de datos, pero según dos puntos de vista (el de las filas y el de las columnas) ; la dualidad es clara y fecunda en AFC, ya que las filas y las columnas de una tabla de contingencia son intrínsecamente objetos de la misma naturaleza, esto es, modalidades de variables cualitativas.
La primera relación ya se ha presentado : ambas nubes, N_I y N_J, tienen la misma inercia total. En AFC, la interpretación clara y crucial de esta inercia total (Φ^2 = desviación de la independencia) muestra que se estudia la misma cosa, vía N_I por una parte o vía N_J por otra parte.
La segunda relación indica que, al proyectar sobre el eje de rango s (u_s para N_I en $\mathbb{R}^J$; v_s para N_J en $\mathbb{R}^I$), la inercia de N_I es igual a la de N_J y es anotada λ_s. Sea :

$$\sum_{i=1}^{I} f_{i\bullet} \left(OH_i^s\right)^2 = \sum_{j=1}^{J} f_{\bullet j} \left(OH_j^s\right)^2 = \lambda_s.$$

Así, no sólo las nubes N_I y N_J tienen la misma inercia total sino también la misma inercia en proyectar sobre los ejes factoriales del mismo rango. Esta propiedad caracteriza los ejes factoriales : ningún otro par de direcciones (uno en $\mathbb{R}^J$, el otro en $\mathbb{R}^I$) la posee.
La tercera relación, la clave de la interpretación, comunica las coordenadas de las filas a las coordenadas de las columnas sobre los ejes del mismo rango. Sea :

$$\begin{aligned} F_s(i) &= \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \frac{f_{ij}}{f_{i\bullet}} G_s(j), \\ G_s(j) &= \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{\bullet j}} F_s(i). \end{aligned}$$

Con $F_s(i)$ la coordenada del perfil-fila i sobre el eje de rango s (en $\mathbb{R}^J$) ; $G_s(j)$ la coordenada del perfil-columna j sobre el eje de rango s (en $\mathbb{R}^I$) ; λ_s la inercia de N_I (resp. de N_J) proyectada sobre el eje de rango s en $\mathbb{R}^J$ (resp. en $\mathbb{R}^I$). Esta propiedad es la base de la representación superpuesta, decimos también «simultánea», de las filas y columnas (cf. figura 2.5, superposición de los gráficos de la figura 2.4). Así, para el s^e eje de esta representación superpuesta, exceptuando el coeficiente $1/\sqrt{\lambda_s}$:

– una fila i está en el baricentro de las columnas, cada columna j con peso $f_{ij}/f_{i\bullet}$ *i.e.*, su término en el perfil de i (estos términos son positivos y su suma es igual a 1) ;
– una columna j está en el baricentro de las filas, cada línea i con peso $f_{ij}/f_{\bullet j}$ *i.e.*, su término en el perfil de j (estos términos también son positivos y su suma es igual a 1).

Esta propiedad denominada baricéntrica (algunas veces casi – baricéntrica, para recordar el coeficiente $1/\sqrt{\lambda_s}$; decimos también relaciones de transición porque permiten transitar de un espacio – $\mathbb{R}^I$ o $\mathbb{R}^J$ – al otro) permite interpretar la posición de una fila en relación con las posiciones del conjunto de las columnas, por una parte, y la posición de una columna con respecto a las posiciones del conjunto de las filas, por otra parte : una fila (resp. una columna) está del lado de las columnas (resp. filas) con las cuales se asocia más y en oposición a las columnas (resp. filas) con las que se asocia menos. Así, en el ejemplo :

– *Quedarse en el hogar* está del lado de *Sólo el marido trabaja*, modalidad con la que se asocia mucho, y en oposición a las dos otras modalidades, con las cuales se asocia poco ;
– *Ambos cónyuges trabajan por igual* está del lado de *Trabajar de medio tiempo* y en oposición a *Quedarse en el hogar*.

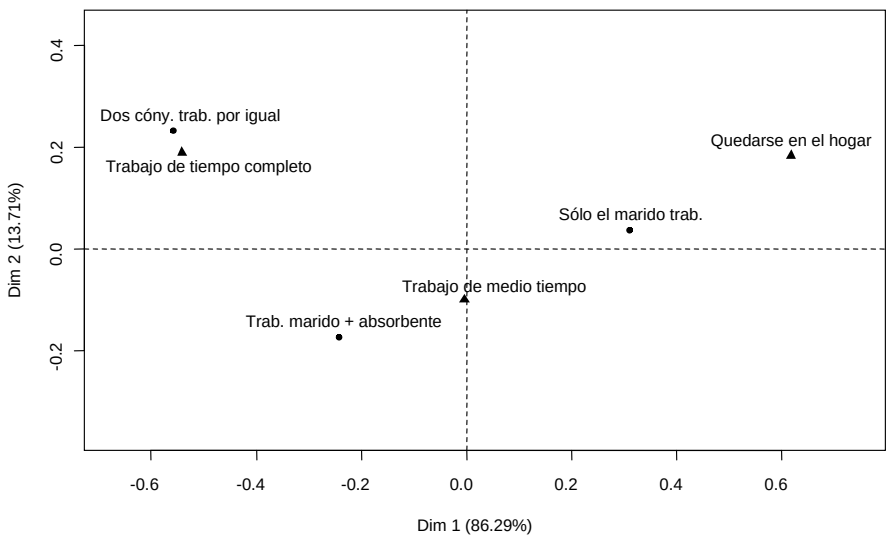

FIGURE 2.5 – Representación simultánea de filas y columnas (= superposición de gráficos de la figura 2.4).

Recordemos que el origen de los ejes es confundido con el perfil medio =(baricentro) de cada una de ambas nubes. Así, cuando un perfil-fila i tiene una coordenada positiva, se asocia

globalmente :
- más que en el modelo de independencia a las modalidades j, teniendo una coordenada positiva ;
- menos que en el modelo de independencia a las modalidades j, teniendo una coordenada negativa.

La palabra «globalmente» en la frase antes citada es importante. La coordenada de un perfil-fila está determinada por el conjunto de coordenadas de las columnas : podemos así comentar la posición de una fila con respecto a las de todas las columnas, pero formalmente no podemos decir nada en cuanto a la proximidad de una fila y de una columna particulares. En concreto, verificaremos en los datos las asociaciones sugeridas por proximidades particulares, entre una fila y una columna, que deseemos comentar.

Baricentro y casi–baricentro.
El coeficiente $1/\sqrt{\lambda_s}$ no debe ser olvidado en la interpretación. Indiquemos que, en AFC, los valores propios están comprendidos entre 0 y 1 (este aspecto se detallara más tarde). De ello resulta que en relación con los baricentros exactos, la representación del AFC está dilatada. Así, en el ejemplo, $1/\sqrt{\lambda_1} = 2.93$ y $1/\sqrt{\lambda_2} = 7.33$; también :
- la modalidad (columna) *Quedarse en el hogar*, que se asocia casi exclusivamente a la modalidad (fila) *Sólo el marido trabaja*, casi se confundiria con esta última en una representación baricéntrica exacta ; su posición sobre el plano es mucho más excéntrica ;
- la modalidad (fila) *Ambos cónyuges trabajan por igual* se asocia, en partes más o menos iguales (142 y 106), a las modalidades *Trabajar de medio tiempo* y *Trabajar de tiempo completo* y sería, en una representación baricéntrica exacta, situada más o menos a mitad del camino entre estas dos modalidades ; sobre el plano es mucho más excéntrica, y a lo largo del eje 1, aparece (ligeramente) más allá de *Trabajar de tiempo completo.*

Podemos preguntarnos si no sería preferible representar los baricentros exactos más bien que los casi–baricentros. Pero, en este caso, dos gráficos son necesarios y en cada uno de ellos las filas y las columnas no desempeñan papeles simétricos ; en particular, el conjunto de las filas y el de las columnas no tienen la misma inercia, la nube de los baricentros está (en relación con la representación usual) contraída alrededor del origen, lo que hace más difícil la lectura de las asociaciones entre modalidades (cf. figura 2.6).
El interés de una representación baricéntrica exacta es la visualización de la intensidad de la relación expresada por el plano (en el sentido de Φ^2). Una nube de baricentros (por ejemplo, la de las filas para fijar las ideas), muy reagrupada alrededor del origen (a lo largo del eje de rango s), pone en evidencia una débil relación (se trata de la parte expresada por el eje de rango s) entre ambas variables $V1$ y $V2$ (cada perfil-fila, próximo al origen, difiere poco del perfil medio). Pero, en este caso, las asociaciones entre filas y columnas son difíciles de ver, lo que permite justamente la dilatación por el coeficiente $1/\sqrt{\lambda_s}$, dilatación que es más fuerte cuanto la (parte de) relación expresada por el eje es más débil. Resulta así que la representación simultánea del AFC es concebida para visualizar la naturaleza de la relación entre las variables (*i.e.*, las asociaciones entre filas y columnas) y no dice nada en cuanto a su intensidad. Esta intensidad es medida por los valores propios (que son componentes de Φ^2) y desde este punto de vista, en la práctica usual del AFC, ambos aspectos de la relación, la naturaleza y la intensidad, están bien identificados por instrumentos separados (gráficos por una parte y valores propios por otra parte).

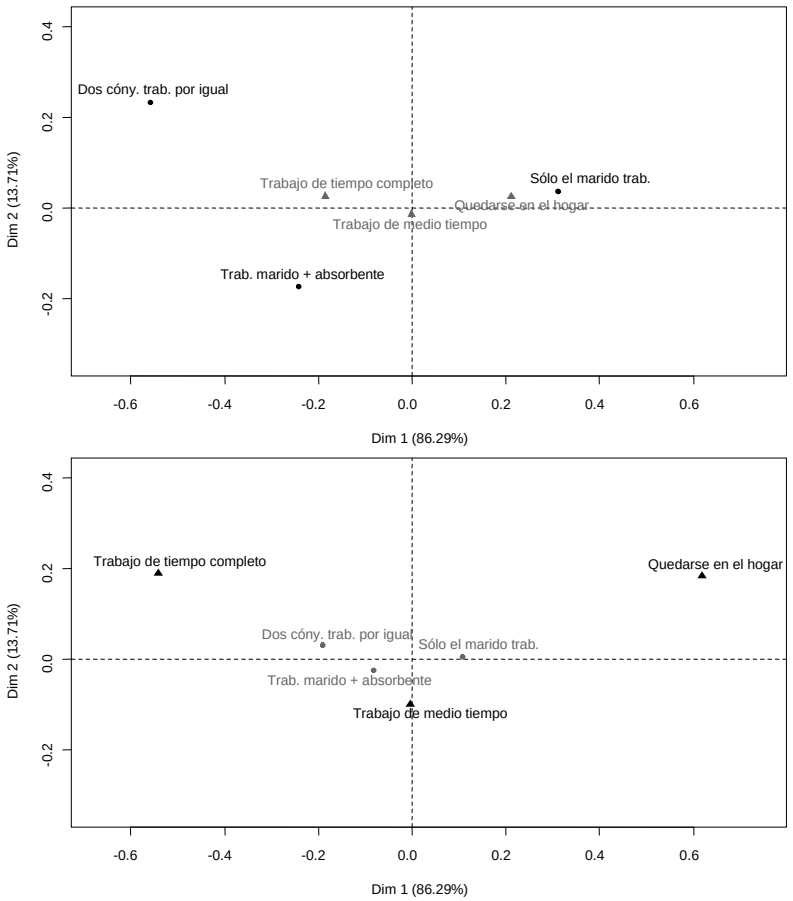

FIGURE 2.6 – Representación de los baricentros exactos. Figura de arriba para las filas; figura de abajo para las columnas del AFC de la tabla 2.1.

Otra ventaja decisiva de la representación casi–baricéntrica aparece en la interpretación sintética de la representación simultánea de este pequeño ejemplo. El primer eje opone, por un lado, las modalidades desfavorables al trabajo femenino, y por el otro, las modalidades favorables. Más precisamente, ordena las modalidades de ambas variables desde el más desfavorable al trabajo femenino (*Quedarse en el hogar*) hasta el más favorable (*Ambos cónyuges trabajan por igual*). En esta perspectiva, el AFC sugiere que *Quedarse en el hogar* es una respuesta mucho más desfavorable al trabajo femenino que *Sólo el marido trabaja*. Esta interclasificación da informaciones sobre el modo en el que los encuestados perciben las modalidades de respuestas. Conviene, pues, encontrar en los datos el origen de la diferencia hecha por el AFC entre estas dos modalidades. El alejamiento más grande, con respecto al origen, de *Quedarse en el hogar* corresponde a una desviación más grande del perfil medio, lo que se puede medir por la contribución a χ^2 (118.07 para *Quedarse en el hogar*; 88.34

para *Sólo el marido trabaja*). De un modo más directamente vinculado a los datos, podemos observar que las personas que han respondido *Quedarse en el hogar* casi todas (84.9%) han respondido *Sólo el marido trabaja* : acumulan así dos respuestas desfavorables al trabajo femenino. En cambio, las personas que han respondido *sólo el marido trabaja* acumulan en el 26.5% de los casos solamente dos respuestas desfavorables. En este sentido podemos decir que *quedarse en el hogar*, que predispone más a una segunda respuesta desfavorable al trabajo femenino, es ella misma más desfavorable al trabajo femenino que *sólo el marido trabaja*.
No está en nuestras atribuciones proponer una explicación psico–sociológica a estas características de la tabla. Retendremos sobre todo que el AFC, vía la representación simultánea, pone en evidencia de modo claro y simple características de la tabla analizada que no aparecen forzosamente por la sola inspección directa de los datos. Esto, ya visible sobre una tabla de dimensión muy pequeña, es tan flagrante y preciado como el aumento de la dimensión de la tabla.

2.4 Ayudas a la interpretación

Como para todo análisis factorial, la interpretación de un AFC se funda esencialmente en las inercias y las representaciones gráficas (es decir, las coordenadas de las filas y de las columnas sobre los ejes). No obstante, en el momento de la interpretación se siente la necesidad de disponer de indicadores para responder a algunas preguntas particulares. Enumeramos a continuación los principales indicadores, y damos algunas pautas para su utilización.

2.4.1 Inercias asociadas a los ejes (valores propios)

De la doble propiedad baricéntrica resulta una propiedad importante del AFC, que introducimos con la ayuda del razonamiento siguiente :

1. Consideramos la proyección de N_I sobre el eje de rango s.
2. Colocamos N_J en los exactos baricentros. La nube N_J está, pues, «en el interior» de N_I y esta última no puede ser una nube de baricentros de N_J.
3. La propiedad doble, «N_I en los baricentros de N_J y N_J en los baricentros de N_I», puede verificarse sólo con la excepción de coeficiente. Este coeficiente debe dilatar la nube de exactos baricentros y ser positivo. De ahí $\lambda_s \leq 1$.

El caso $\lambda_s = 1$ es particular. Una vez situada la nube N_I, el único modo para que la nube N_J en calidad de baricentros no esté dentro de la nube N_I es una asociación mutua exclusiva entre filas y columnas. En la figura 2.7 se muestra la estructura de los datos que corresponde a este caso : el conjunto I de filas (resp. J columnas) puede dividirse en dos subconjuntos $I1$ y $I2$ (resp. $J1$ y $J2$) ; $I1$ (resp. $I2$) se asocia exclusivamente a $J1$ (resp. $J2$). Esta estructura de datos expresa una relación fuerte entre ambas variables $V1$ y $V2$, lo que el AFC pone en evidencia por un eje que opone, por una parte, $I1$ y $J1$, y por otra parte, $I2$ y $J2$.
En práctica, los valores propios de un AFC no son casi nunca exactamente iguales a 1 ; pero un valor propio elevado es la señal de una estructura parecida a la de la figura 2.7, información capital en el análisis de una tabla de contingencia. La consulta de los valores

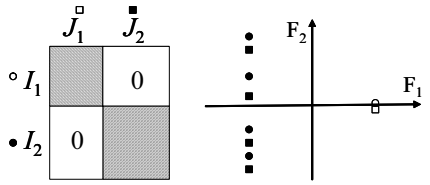

FIGURE 2.7 – Caso de un valor propio igual a 1. Estructura de los datos y plano factorial ($\lambda_1 = 1$).

propios es, pues, importante en AFC. En el ejemplo, los valores propios son débiles (cf. tabla 2.5). Incluso el primero, aunque asociado a una estructura clara, es débil : esto que se evidencia es sólo una tendencia, aunque sea altamente significativa (cf. el test de χ^2).
Aquí todavía no le corresponde al estadista proponer una interpretación psico–sociológica de esta «débil» intensidad de relación : ¿esto se debe simplemente al hecho de que las preguntas no son las mismas, o al «ruido» que está siempre presente en las respuestas a las preguntas de opinión ?

	valor propio	variación porcentual	variación porcentual acumulada
dim 1	0,117	86,29	86,29
dim 2	0,019	13,71	100,00

Tabla 2.5 – Valores propios (=inercias proyectadas) del AFC de la tabla 2.1.

Después de insistir en el hecho de que la inercia asociada a un eje es una parte de la relación entre ambas variables $V1$ y $V2$, es natural expresar esta parte en porcentajes (cf. tabla 2.5). En el ejemplo, resulta así que el primer eje representa 86.29%, es decir, la casi totalidad de la distancia entre la tabla de datos y la independencia. Hay aquí un argumento para tener en cuenta sólo este eje en la interpretación. De modo más general, los valores propios miden la importancia relativa de los ejes : su secuencia sugiere los ejes sobre los cuales enfocar la atención. En el mismo tipo de ideas, representamos esta secuencia por un diagrama de barras. En la figura 2.8 se muestra un caso histórico (doce marcas de cigarrillos en potencia ; Benzécri, 1973, tomo 2 p. 339) en la cual este diagrama sugiere 5 ejes sensiblemente más importantes que los otrosy se presenta el lento decrecimiento de los valores propios más allá del quinto lo que da a entender que los ejes correspondientes representan sólo al «ruido».
En el estudio de tal caso, es prudente examinar el eje 6 por lo menos superficialmente porque, *in fine*, una interpretación clara de este eje incitará a conservarlo en el comentario de los resultados. Este uso habitual (tomar en consideración los ejes interpretables incluso si corresponden a una inercia débil) no está desprovisto de sentido común (es difícil apartar del comentario una dimensión que se sabe interpretar bien), pero dio lugar a numerosos debates.
Al ser los ejes ortogonales, se pueden adicionar las inercias proyectadas sobre varios ejes. En el ejemplo, la parte de la relación expresada por el plano es 100 %, lo que no es una característica de los datos pero proviene de la dimensión de la tabla (3×3 ; cf. observación sobre el número de ejes, sección 2.3.3). De modo más general, para cuantificar la parte de

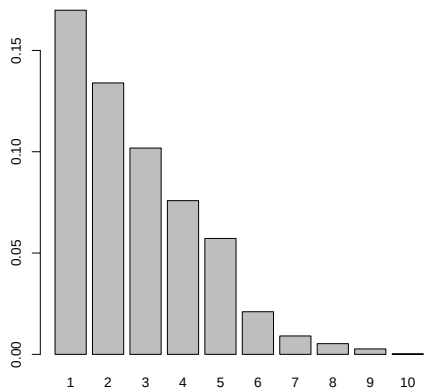

FIGURE 2.8 – Ejemplo de diagrama que ilustra la secuencia de los valores propios de un AFC.

inercia tomada en consideración en un comentario de los S primeros ejes, utilizamos la suma de los S primeros porcentajes de inercia.

Volviendo a la interpretación geométrica de los valores propios como inercia proyectada, el porcentaje de inercia asociado al eje se escribe :

$$\frac{\text{inercia proyectada de } N_I \text{ (o } N_J\text{) sobre el eje de rango } s}{\text{inercia total de } N_I \text{ (o de } N_J\text{)}} \times 100.$$

Este criterio aparece aquí como una medida de la calidad global de representación de la nube N_I (o N_J) por el eje de rango s. Más generalmente, podemos considerar la proyección sobre un plano. En el presente caso, este criterio responde a la pregunta siguiente : si sabemos que al proyectar la nube N_I (o N_J) sobre un plano (generalmente el primero, construido a partir de los ejes 1 y 2) lo deformamos (recordemos que la operación de proyección sólo puede reducir las distancias entre puntos), ¿es esta deformación importante ? Es decir, ¿las proximidades entre puntos (del mismo conjunto, las filas o las columnas) sobre un plano reflejan bien las proximidades en el espacio de salida ($\mathbb{R}^J$ o $\mathbb{R}^I$) ? Si la respuesta es sí, la interpretación es simple en el sentido de que las proximidades sobre el plano se encuentran muy fácilmente en los datos aunque los valores propios sean débiles. Si la respuesta es no, el interés *a priori* de la representación no se pone en duda ; simplemente, la débil calidad de representación indica que otros fenómenos, visibles sobre los planos siguientes, se añaden a lo que muestra el plano estudiado. En tal caso, encontrar en los datos los hechos puestos en evidencia por el plano será menos simple en el caso de valores propios débiles (es siempre fácil en el caso de valores propios próximos a 1).

Finalmente, la calidad de representación asociada a un plano es una característica que hay que tomar en consideración pero no constituye de ninguna manera un juicio de valor sobre el interés del plano. El pequeño ejemplo utilizado es una ilustración perfecta (aunque en un caso límite) : la calidad de representación de 100 % se debe a la débil dimensión de la tabla y no prejuzga para nada el interés del análisis.

Observación sobre el valor máximo de Φ^2.
La tabla de dimensión $I \times J$ genera como máximo $\inf(I-1, J-1)$ valores propios no nulos. Cada uno de estos valores propios es inferior o igual a 1. El valor máximo de Φ^2 es, pues, $\inf(I-1, J-1)$. Llevando el valor observado de Φ^2 a su máximo teórico, se llega al indicador estadístico, llamado V de Cramer, definido así :

$$V = \left(\frac{\Phi^2}{\inf\{(I-1);(J-1)\}} \right)^{1/2}.$$

El interés de este criterio es variar entre 0 (independencia) y 1 (relación máxima en el sentido siguiente : cada modalidad de la variable que tiene el número más grande de modalidades se asocia exclusivamente con una sola modalidad de la otra variable). A causa de su zona de variación, V de Cramer desempeña un papel análogo, hasta cierto punto, al del coeficiente de correlación. Así, frente a varias variables cualitativas (definidas sobre los mismos individuos), podemos editar una matriz de V (como se edita una matriz de correlación).

2.4.2 Contribución de un punto a la inercia de un eje

La inercia asociada a un eje puede descompuerse por puntos. La contribución del punto i a la inercia del eje de rango s se define generalmente por (retomando las notaciones de la sección 2.3.3) :

$$\begin{aligned} \mathrm{ctr}_s(i) &= \frac{\text{inercia de } i \text{ proyectada sobre el eje de rango } s}{\text{inercia de } N_I \text{ proyectada sobre el eje de rango } s}, \\ &= \frac{f_{i\bullet} \left(OH_i^s\right)^2}{\sum_{i=1}^{I} f_{i\bullet} \left(OH_i^s\right)^2} = \frac{f_{i\bullet} \left(OH_i^s\right)^2}{\lambda_s}. \end{aligned}$$

Esta contribución es a menudo multiplicada por 100 o 1000 para facilitar la edición de las tablas. Se denomina a veces «relativa», ya que se trae al conjunto de la nube ; la contribución «absoluta» es, entonces, la inercia proyectada por el punto $(f_{i\bullet} \left(OH_i^s\right)^2)$. Esta distinción de «relativa» y «absoluta» no la hacen con estos términos todos los autores. La mayoría de las veces, «contribución» (o incluso «contribución absoluta») significa lo que se llama en este libro «contribución relativa».
Las contribuciones son calculadas tanto para las filas como para las columnas. Pueden ser acumuladas sobre varios puntos (de la misma nube). Son útiles sobre todo cuando hay muchos puntos. Seleccionar los puntos más contributivos a menudo facilita un primer enfoque de la interpretación. El caso particular de un eje debido esencialmente a uno o dos puntos se detecta inmediatamente : la interpretación puede enfocarse entonces, en este punto, evitando generalizaciones arriesgadas. En esta misma idea, el número mínimo de puntos que acumulan un porcentaje fijado (por ejemplo, 50 %) de la inercia de un eje es un indicador de la «generalidad» de dicho eje.
A causa de su pequeña dimensión, el análisis de datos sobre las opiniones con respecto al trabajo femenino no necesita las contribuciones pero ésta bastan para ilustrar los cálculos : por ejemplo, el detalle del cálculo de las contribuciones de *Sólo el marido trabaja* y de *Ambos cónyuges trabajan por igual* sobre el primer eje muestra el papel respectivo de los pesos y de las distancias en la formación de dos contribuciones cercanas.

$$\text{ctr}_1(\textit{Sólo el marido trabaja}) = \frac{0.5267 \times 0.3096^2}{0.1168} = \frac{0.5267 \times 0.0958}{0.1168} = 0.432$$
$$\text{ctr}_1(\textit{Ambos cónyuges trabajan por igual}) = \frac{0.1514 \times 0.5586^2}{0.1168} = \frac{0.1514 \times 0.312}{0.1168} = 0.404$$

El punto *Ambos cónyuges trabajan por igual* está (más o menos) dos veces más alejado del origen que el otro, lo que sugiere una influencia más grande ; pero el peso de *Ambos cónyuges trabajan por igual* es (más o menos) tres veces más débil, lo que sugiere a su turno una influencia más débil ; en la inercia (criterio utilizado para definir los ejes), la distancia interviene en su cuadrado : finalmente, ambas contribuciones están equilibradas.

	Coordenadas		Contribuciones		Calidad de representación	
	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2
Dos cóny. trab. por igual	-0,56	0,23	40,43	44,43	0,85	0,15
Trab. marido + absorbente	-0,24	-0,17	16,37	51,44	0,67	0,33
Sólo el marido trab.	0,31	0,04	43,20	4,13	0,99	0,01
	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2
Quedarse en el hogar	0,62	0,18	53,91	29,61	0,92	0,08
Trabajo de medio tiempo	0,00	-0,10	0,01	34,85	0,00	1,00
Trabajo de tiempo completo	-0,54	0,19	46,08	35,53	0,89	0,11

Tabla 2.6 – Coordenadas, contribuciones relativas (=en %) y calidad de representación para cada modalidad y para cada eje.

Observación
En AFC, los puntos generalmente tienen el mismo peso y los cálculos de contribución desempeñan un papel más importante que en el ACP normado usual (en el cual los elementos tienen el mismo peso) : en efecto, en este último caso, la contribución es proporcional al cuadrado de la distancia al origen y se lee (más o menos) sobre las representaciones factoriales.

2.4.3 Calidad de representación de un punto por un eje o un plano

El porcentaje de inercia asociado a un eje se ha presentado, entre otras cosas, como un indicador de calidad de representación de una nube por un eje. Podemos aplicar este indicador para un solo punto y calcular así la calidad de representación de un punto i por el eje de rango s que se anota $\text{cali}_s(i)$ (cf. § 1.6.1) ; sea :

$$\text{cali}_s(i) = \frac{\text{inercia de } i \text{ proyectada sobre el eje de rango } s}{\text{inercia total de } i} = \frac{(OH_i^s)^2}{(Oi)^2} = \cos^2(\overrightarrow{Oi}, \overrightarrow{OH_i^s}).$$

Esta relación indica en qué medida la desviación de la modalidad i al perfil medio se expresa sobre el eje de rango s. Aquí todavía este indicador no es verdaderamente útil para los resultados del AFC aplicado a la tabla de las opiniones sobre el trabajo femenino ; esto se debe a la pequeña dimensión de la tabla que conduce a una representación perfecta de las nubes (y de cada punto) sobre el primer (y único) plano. Pero estos datos permiten ilustrar simplemente el significado de este indicador, lo que se detalla en dos puntos :

1. Las cuatro modalidades extremas están bien representadas por el primer eje (calidad de representación > 0.85) ; la desviación de cada una de ellas al perfil medio (*i.e.*, las modalidades a las cuales se asocia más, o menos, que si hubiese independencia) está bien descrita por este eje ; el otro eje aporta relativamente poco sobre estas modalidades.

2. La modalidad *Trabajar de medio tiempo* está muy mal representada por el primer eje ; pero esto no significa que se deba apartar tal modalidad de la interpretación (al contrario, la posición central de esta modalidad se ha comentado suficientemente) ; esto ilustra bien la primacía de las coordenadas en la interpretación ; simplemente, la desviación de esta modalidad al perfil medio puede leerse sólo a través de otros ejes.

En la práctica, utilizamos las calidades de representación principalmente en los casos siguientes :
– Nos interesamos por una modalidad en particular ; la calidad de representación permite seleccionar el plano sobre el cual esta modalidad se expresa mejor.
– Buscamos un pequeño número de modalidades para ilustrar el significado de un eje s con la ayuda de los datos brutos, lo que es muy útil particularmente para comunicar los resultados ; seleccionamos primero las modalidades que tienen las coordenadas más extremas (ya que el efecto representado por el eje s estudiado es muy fuerte aquí), modalidades que se ordenan luego en beneficio de las modalidades mejor representadas (puesto que el efecto del eje s es único aquí).

Observemos que estos comentarios, hechos en el contexto del AFC, se transponen fácilmente a otros métodos factoriales (reemplazando, por ejemplo para el ACP, la noción de perfil medio del AFC por la noción de «individuo medio»).

2.4.4 Distancia e inercia en el espacio inicial

Quizás antes de obtener los resultados del AFC, podemos preguntarnos qué modalidades son las más –o las menos– «responsables» de la desviación a la independencia. Dos puntos de vista se pueden adoptar :
– El de la inercia ya se ha utilizado a través de la descomposición de χ^2 por filas y por columnas ; así, la tabla 2.7 pone en evidencia el papel relativamente equilibrado desempeñado por cada una de las cuatro modalidades extremas.
– El de la distancia al perfil medio ; aquí no tomamos en consideración el efectivo de la modalidad ; estas distancias se reúnen en la tabla 2.7, que pone en evidencia una distancia al principio comparable para ambas modalidades : *Sólo el marido trabaja* y *El marido tiene un trabajo más absorbente* (el débil número de filas limita el interés de este indicador en la interpretación ; la modalidad *Sólo el marido trabaja*, mayoritaria (52.7%), no puede diferir mucho del perfil medio del que es parte integrante).

En la práctica, las distancias al origen permiten seleccionar la fila o la columna que se parece más –o menos– al perfil medio ; lo que es un modo cómodo de ilustrar la diversidad de los perfiles.

	Dos cony. trab. por igual	Trab. marido + absorbente	Sólo el marido trab.
Distancia	0,3665	0,0891	0,0973
Inercia	0,0555	0,0287	0,0512

	Quedarse en el hogar	Trabajo de medio tiempo	Trabajo de tiempo completo
Distancia	0,4158	0,0099	0,3287
Inercia	0,0685	0,0065	0,0604

Tabla 2.7 – Distancia (al cuadrado) del perfil medio e inercia (en los espacios de salida, $\mathbb{R}^I$ y $\mathbb{R}^J$).

2.5 Elementos suplementarios (=ilustrativos)

Como en todo análisis factorial (cf. ACP § 1.6.2), podemos introducir elementos (*i.e.*, de filas o columnas) suplementarios, terminología que se refiere a su estatus : no intervienen en la construcción de los ejes (lo que no impide proyectarlos como los otros –llamados elementos activos– sobre los ejes encontrados). Los llamamos también «ilustrativos» en referencia a su función más frecuente : enriquecer, ilustrar la interpretación de los ejes.
En AFC, los elementos suplementarios son generalmente tablas de contingencia. Su posición sobre el plano se calcula utilizando las propiedades baricéntricas. Observemos que en esta relación el coeficiente de dilatación, $1/\sqrt{\lambda_s}$, depende de la relación entre las variables activas $V1$ y $V2$ y no de los elementos suplementarios. De esto resulta que la representación de las modalidades de una tabla de contingencia introducida en columnas suplementarias (que cruza $V1$ y una tercera variable $V3$, por ejemplo) toma en consideración la intensidad de la relación entre $V1$ y $V2$. Así, la nube de las modalidades de $V3$ (cf. figura 2.9) estará más (resp. menos) concentrada alrededor del origen que en las modalidades de $V1$ si la relación (más exactamente la parte de la relación expresada por el eje considerado) entre $V1$ y $V3$ es menos (resp. más) intensa que la relación entre $V1$ y $V2$. Podríamos pensar en utilizar otro coeficiente de dilatación para los elementos suplementarios con el pin de visualizar «mejor» las asociaciones, por ejemplo entre columnas suplementarias y filas activas, pero esto no nos serviría mucho ya que no podríamos comparar las posiciones relativas de las columnas activas y de las columnas suplementarias.
Esto puede ilustrarse con un ejemplo. En su obra, N. Tabard publica otra tabla, que cruza $V1$, y una nueva pregunta (cf. tabla 2.8) que llamaremos $V3$. Esta nueva variable es de un formato muy clásico en los cuestionarios de opinión. Proponemos a los que van a responder una lista de opiniones : para cada una de ellas, la persona que responde expresa su acuerdo o su desacuerdo con la ayuda de una escala, en cuatro puntos que van de *Para nada de acuerdo* a *Completamente de acuerdo*. La redacción exacta de la pregunta es : *¿Qué piensa usted de la siguiente opinión escuchada algunas veces : las mujeres que no trabajan se sienten aisladas del mundo ?*

1. Completamente de acuerdo
2. Más bien de acuerdo
3. No muy de acuerdo
4. Para nada de acuerdo

Anotemos de entrada que la relación entre $V1$ y $V3$ es altamente significativa ($\chi^2 = 162.19$;

p-crítico 2.04×10^{-32}) pero poco intensa ($\Phi^2 = 0.094$; $V(V1, V3) = 0.217$), en particular menos intensa que la relación entre $V1$ y $V2$ ($\Phi^2(V1, V2) = 0.135$; $V(V1, V2) = 0.260$). Más allá del significado de las preguntas, esta relación más débil reenvía el «ruido» que acompaña las respuestas a las preguntas de tipo $V3$. Las modalidades que expresan un acuerdo pueden tener como origen una preocupación general de no oponerse; las que expresan un desacuerdo pueden originarse en una hostilidad con respecto al cuestionario en general. De ahí las respuestas contradictorias que tienen como efecto de ocultar la relación entre las preguntas.

	Imagen ideal de una familia :			
¿Las mujeres en el hogar se sienten aisladas del mundo?	Ambos cónyuges trabajan por igual	Trabajo del marido más absorbente	Sólo el marido trabaja	Suma
Completamente de acuerdo	107	192	140	439
Más bien de acuerdo	75	175	215	465
No muy de acuerdo	40	100	254	394
Para nada de acuerdo	39	88	299	426
Suma	261	555	908	1724

Tabla 2.8 – Tabla de opiniones dispuestas en columnas suplementarias en el AFC de la tabla 2.1 (Tabard, 1974).

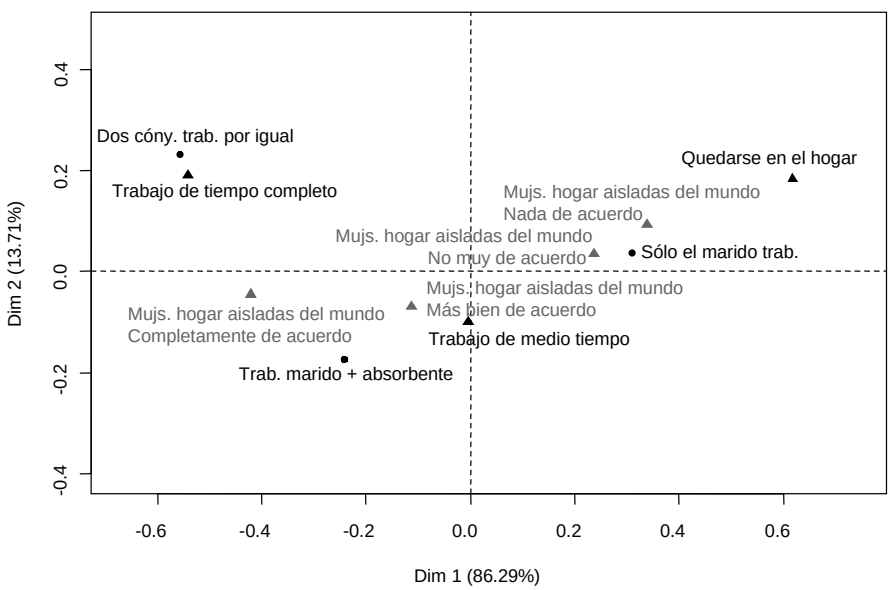

FIGURE 2.9 – Representación de la figura 2.5 completada por las modalidades de la variable suplementaria *Las mujeres que se quedan en el hogar se sienten aisladas del mundo.*

Limitamos el comentario sobre la proyección de las modalidades de la variable $V3$ a los siguientes puntos :

– Las modalidades que expresan el acuerdo con la opinión *Las mujeres que se quedan en el hogar se sienten aisladas del mundo* se encuentran del lado de las actitudes desfavorables con respecto al trabajo femenino e inversamente para las modalidades que expresan un desacuerdo. La interpretación del eje es enriquecida.
– La nube de las modalidades de $V3$ está más concentrada alrededor del origen que las otras dos variables ; encontramos el hecho de que la relación entre $V1$ y $V3$ es menos intensa que la relación entre $V1$ y $V2$.
– La modalidad *Completamente de acuerdo* está más alejada del origen de los ejes que la modalidad *Para nada de acuerdo* ; parece ser más característica de una actitud favorable al trabajo femenino que *Para nada de acuerdo* de una actitud desfavorable.

Observación sobre el campo de aplicación del análisis de correspondencias.
El análisis de las correspondencias se concibió para tratar tablas de contingencia y su justificación teórica completa se sitúa en este marco. Sin embargo, la puesta en práctica de un programa de AFC proporciona resultados útiles sobre muchas otras tablas desde el momento en el que contienen números positivos y que sus márgenes se interpretan. Citemos, entre otros, el caso de una matriz de incidencia asociada a un gráfico (en el que el término general x_{ij} vale 1 si una arista une los vértices i y j y 0 si no los une).
Para justificar la aplicación del AFC sobre tal tabla, y de ser susceptible de interpretar los resultados obtenidos, el usuario deberá preguntarse el significado de las principales propiedades del AFC. Así, en el caso de una matriz de incidencia : 1) la propiedad baricéntrica asegura la interpretabilidad de los planos factoriales, cada vértice aparece en el baricentro de aquellos con los que está unido por una arista ; 2) la inercia máxima asegura el interés en particular del primer plano, sabiendo que acerca al máximo los vértices unidos por varios caminos de longitud 2 y separa los otros.

2.6 Puesta en marcha con FactoMineR

En esta sección mostramos cómo efectuar un AFC con FactoMineR. Encontramos entonces los diferentes resultados del AFC de la tabla 2.1 que se han comentado en las secciones precedentes.

```
> library(FactoMineR)
> trabajo <- read.table("http://factominer.free.fr/libra/trabajo.csv",
    header=TRUE, row.names=1, sep=";")
> summary(trabajo)
```

El test de χ^2 y la tabla 2.2 se obtienen utilizando únicamente las tres primeras columnas del juego de datos :

```
> res.test.chi2 <- chisq.test(trabajo[,1:3])
> res.test.chi2
> round(res.test.chi2$expected,1)
```

La tabla 2.3 se obtiene por :

```
> round(res.test.chi2$residuals^2, 2)
> round(100 * res.test.chi2$residuals^2 / res.test.chi2$stat, 2)
```

La tabla 2.4 se obtiene, después de haber transformado la tabla de datos en matriz, por :

```
> dd <- rbind(trabajo,apply(trabajo[,1:3],2,sum))
> rownames(dd)[4] <- "Perfil medio"
> round(prop.table(as.matrix(dd),margin=1),3)

> dd <- cbind(trabajo,apply(trabajo[,1:3],1,sum))
> colnames(dd)[4] <- "Perfil medio"
> round(prop.table(as.matrix(dd),margin=2),3)
```

El AFC se realiza después ; el AFC proporciona por defecto el gráfico de la representación superpuesta (cf. figura 2.5).

```
> res.ca <- CA(trabajo[,1:3])
```

El gráfico de la representación de las filas y el de la representación de las columnas (cf. figura 2.4) se obtienen empleando utilizando la función **plot.CA**.

```
> plot(res.ca, invisible="col")
> plot(res.ca, invisible="row")
```

Los gráficos de la representación de exactos baricentros (cf. figura 2.6) se obtienen por :

```
> plot(res.ca,invisible="col")
> coord.col = sweep(res.ca$col$coord,2,sqrt(res.ca$eig[,1]),FUN="*")
> points(coord.col, pch=17, col="red")
> text(coord.col, rownames(coord.col), col="red")

> plot(res.ca,invisible="row")
> coord.row = sweep(res.ca$row$coord,2,sqrt(res.ca$eig[,1]),FUN="*")
> points(coord.row, pch=20, col="blue")
> text(coord.row, rownames(coord.row), col="blue")
```

El cuadro de los valores propios (cf. tabla 2.5) y el gráfico de los valores propios se obtienen por :

```
> res.ca$eig
> barplot(res.ca$eig[,1], main="Valores propios", names.arg=1:nrow(res.ca$eig))
```

La tabla de contribuciones y la de las calidades de representación de las filas y de las columnas (cf. tabla 2.6) se obtienen por :

```
> cbind(res.ca$row$coord,res.ca$row$contrib,res.ca$row$cos2)
> cbind(res.ca$col$coord,res.ca$col$contrib,res.ca$col$cos2)
```

Las inercias de las filas y de las columnas (cf. tabla 2.7) se logran directamente mientras que las distancias al cuadrado deben calcularse de nuevo utilizando el margen fila y el margen columna :

```
> res.ca$row$inertia
> res.ca$col$inertia
> res.ca$row$inertia/res.ca$call$marge.row
> res.ca$col$inertia/res.ca$call$marge.col
```

El gráfico de la figura 2.9 se obtiene realizando un nuevo AFC precisando que las columnas a partir de la cuarta son suplementarias :

```
> res.ca2 <- CA(trabajo,col.sup=4:ncol(trabajo))
```

2.7 AFC y tratamiento de datos textuales

Reagrupamos con la denominación de análisis de datos textuales las metodologías centradas sobre el análisis de un conjunto de textos desde el punto de vista de las palabras que los contienen. La base de estas metodologías es el análisis de correspondencias de la tabla (llamada tabla léxica) que cruzan textos y palabras, de término general x_{ij}, número de veces que la palabra j ha sido utilizada en el texto i. A primera vista, se trata de un simple campo de aplicación de métodos de análisis de datos. De hecho, los datos textuales presentan numerosas particularidades que necesitan metodologías específicas ; desde este punto de vista, se trata totalmente de una disciplina científica (que posee sus propios congresos : Jornadas de Análisis de Datos Textuales, JADT). Nuestra presentación se articula sobre este razonamiento : ámbito de aplicación y disciplina científica.

Retomemos la tabla de léxica mencionada anteriormente. Puede verse como una tabla de contingencia (y tener que ver con un AFC) adoptando el razonamiento siguiente. La unidad estadística elemental es la forma gráfica, secuencia de caracteres comprendida entre dos separadores (esencialmente los signos de puntuación y el espacio). Cada forma gráfica se caracteriza por dos variables cualitativas : la variable «texto» (las modalidades son los textos mismos) y la variable «diccionario» (las modalidades son las palabras). La tabla de léxica reparte las formas gráficas según estas dos variables y, con este título, es una tabla de contingencia.

El AFC está bien adaptado al estudio de este tipo de tabla (históricamente se ha imaginado para ello : la primera aplicación de AFC publicada, en la tesis de Brigitte Escofier, 1965 [3], es de este tipo) ; más precisamente, describe la desviación entre esta tabla y el modelo de independencia. El punto de vista del AFC sobre esta desviación se apoya en la noción de perfil : aquí hablamos del perfil léxico de un texto (conjuntos de frecuencias de las palabras en este texto) y del perfil de utilización de una palabra (conjuntos de frecuencias de esta palabra en los textos).

Hay independencia cuando todos los perfiles (léxicos por una parte, de utilización por otra) son idénticos entre ellos, y entonces, idénticos al perfil medio (número total de palabras de cada texto ; frecuencia global de utilización de cada palabra). La desviación de la independencia es más grande cuanto más los perfiles difieren del perfil medio y el AFC analiza precisamente esta desviación para proporcionar una síntesis en la forma de una visualización organizada en una serie de dimensiones ; una dimensión puede poner en evidencia, por ejemplo, un grupo de textos que tiene en común una frecuencia elevada (*i.e.*, más elevada en la del perfil medio) para ciertas palabras y una frecuencia débil (*i.e.*, más débil que en la del perfil medio) para otras palabras ; esta misma dimensión pone en evidencia también, por dualidad, un grupo de palabras que tienen en común una frecuencia elevada (*i.e.*, más elevada que en el perfil medio) para ciertos textos : son las palabras que caracterizan los textos puestos en evidencia por esta misma dimensión. Así, la visualización proporcionada por el AFC corresponde perfectamente a lo que podemos esperar de un análisis exploratorio de un conjunto de textos.

La especificidad de los datos textuales aparece en la construcción de la tabla, en otras palabras, en la elección de las filas y la de las columnas.

¿Qué textos ?

3. B. Escofier, 1965. El análisis de las correspondencias, tesis de tercer ciclo, Universidad de Rennes.

Hasta aquí, nosotros hemos llamado por comodidad «texto» a una fila de la tabla de léxica. La definición de estos textos no es siempre tan evidente, y es lo que ilustramos con estos dos ejemplos.

En la primera aplicación del AFC (citada anteriormente), el cuerpo inicial es la pieza de teatro Phèdre. Para analizar este cuerpo, hubo que subdividirlo. El criterio elegido era el personaje : una fila de la tabla (un texto) es el conjunto de las réplicas de un personaje dado. Así fue posible establecer una cartografía de los personajes en función del único vocabulario que utilizan ; la primera dimensión corresponde al estatus social : los personajes más importantes (el de Phèdre, pero hacer generalizaciones es tentador) no utilizan las mismas palabras que sus súbditos (¡comenzando por el tuteo y el tratamiento de usted !). Otros recortes eran posibles : por escena (para visualizar el desarrollo de la intriga) o, más finamente, por personajes que se cruzan en los actos, para seguir la evolución de los personajes a lo largo de la obra.

Una aplicación capital del análisis de los datos textuales consiste en analizar las preguntas abiertas en los cuestionarios. Un ejemplo famoso es el de una pareja siguiendo preguntas formuladas sucesivamente : ¿Qué es, para usted, la derecha ? ¿Qué es, para usted, la izquierda ? El interés de este tipo de preguntas es unánimemente reconocido : la espontaneidad de las respuestas es una prueba de la importancia concedida (por el que contesta) a los aspectos que evoca, información que es difícil de obtener de otro modo. En el ejemplo anterior, ¿se destacan más bien los aspectos económicos, sociales y políticos ? ¿Y esto indiferentemente para la izquierda y la derecha ?

Antes que todo, podemos pensar en considerar cada encuestado como una fila de la tabla. Pero esta tabla está generalmente muy vacía (numerosas casillas), y su análisis por AFC es a menudo arduo y decepcionante (muchos ejes ponen en evidencia pequeños grupos de individuos que tienen algunas palabras particulares en común) pero poco sintético. Una metodología recomendada consiste en reagrupar los encuestados según un criterio que cruza todas o una parte de las variables clásicas siguientes : género (hombre/mujer), nivel de estudios y edad («dividido» en clases). Otras reagrupaciones son por supuesto posibles y el usuario deberá hacer esta elección con cuidado porque condiciona fuertemente los resultados. Un texto es, entonces, la concatenación de las respuestas de una de las categorías procedentes de la reagrupación de los que responden.

¿Qué palabras ?

De nuevo, por comodidad, llamamos «palabra» a una columna de la tabla de léxica. En la práctica, la definición de lo que debe representar una columna no es simple, ya que hay numerosos puntos de vista, cada uno de los cuales presenta aspectos interesantes. El usuario deberá, pues, tomar las decisiones que le parezcan más convenientes a sus datos y a sus objetivos. Citamos a continuación algunos puntos claves.

Selección según la frecuencia global. Las palabras raras, interesan poco al usuario en un procedimiento de síntesis y a la vez pueden influir en el AFC. En efecto, una palabra utilizada en un solo texto que contendría sólo esta palabra engendra un eje asociado al valor propio (máximo) de 1 ; aunque este caso jamás se produce en la práctica, constituye una referencia útil que contiene valores propios elevados, del orden de 0.5, observados regularmente en este tipo de análisis. Eliminaremos entonces las palabras muy poco frecuentes (raras) ; la noción de rareza, al ser relativa, implica definir el umbral en cada caso, en función del conjunto de la frecuencia de las palabras.

Palabras herramientas. ¿Hay que conservar los artículos, las conjunciones, etc. ? A primera vista, el usuario intenta eliminar esas palabras que no parecen importantes. Indiquemos, a pesar de todo, que si esas palabras están repartidas como el perfil de repartición medio (cuando su caso depende sólo de la longitud del texto), entonces están cerca del centro de gravedad de la nube de las palabras e influyen poco en el análisis. En cambio, si su frecuencia de utilización depende del texto, entonces son la marca de algo y merecen ser conservados.
Lematización. ¿Hay que reagrupar, por ejemplo, las formas gráficas correspondientes al singular o al plural del mismo nombre ? ¿Las correspondientes al mismo verbo ? La lematización consiste en reagrupar las formas gráficas relativas a la misma entrada en el diccionario. Tiene sus partidarios y sus oponentes. Indiquemos la propiedad de equivalencia distribucional, propiedad general del AFC valorada precisamente en referencia a las tablas léxicas, ilustrándolas por un ejemplo «textual» : si *día* y *días* tienen el mismo perfil, entonces es lo mismo considerarlos juntos o por separado. Esta propiedad es un argumento en desventaja de la lematización : en caso de igualdad de perfil, no ganamos nada ; sino que perdemos un matiz. En la práctica, no obstante, hay que asegurarse de que el matiz citado merece la atención, lo que no es el caso para las palabras menos frecuentes (excepto si se pone el límite inferior de selección exageradamente elevado).
Stematización. Consiste en reagrupar las formas gráficas que poseen la misma raíz. Así, en comentarios de degustación de vinos, podemos querer reagrupar *verde* y *falta de madurez*. La stematización llama observaciones análogas a las emitidas a propósito de la lematización ; pero aquí la toma de riesgo (de confundir nociones distintas) es más importante.
Segmentos repetidos. Ciertas palabras a menudo aparecen de modo combinado y esta combinación (hablamos de segmentos repetidos) es a la vez más evocadora que las palabras consideradas por separado y puede eliminar las posibles ambigüedades. Así, en relación con las descripciones de vinos, *frutas rojas* es precisamente más evocador que *frutas* (pensar en *frutas confitadas* de los vinos untuosos) y que *rojo* (el color *rojo* no implica, en principio, una nota aromática *frutas rojas*). El ejemplo más conocido de segmento repetido es sin duda *seguridad social*, cuyo significado no se deduce fácilmente de *seguridad* y de *social*. Por eso es muy útil considerar los segmentos repetidos, asignándole una columna a cada uno de ellos.
Las consideraciones que preceden no agotan el tema del análisis de los datos textuales sino que dan los puntos de referencia claves para su puesta en práctica. Lo hemos comprendido : lo más importante del trabajo se sitúa más arriba del AFC, en la constitución de la tabla léxica a partir de un conjunto de textos.
El paquete tm (para text mining) está dedicado al análisis de datos textuales. La función **textual** de FactoMineR es una función lexical que permite construir una tabla de contingencia. Ilustremos esta función a partir del pequeño ejemplo siguiente que contiene dos variables cualitativasy una variable textual :

```
> vino
    Año de cosecha    Vino                Texto
1 Año de cosecha 1  Vino 1      Acidez,afrutado
2 Año de cosecha 2  Vino 1      Afrutado,ligero
3 Año de cosecha 1  Vino 1 Con toques de madera
4 Año de cosecha 2  Vino 1               Acidez
5 Año de cosecha 1  Vino 2            Azucarado
6 Año de cosecha 2  Vino 2   Azucarado,licoroso
7 Año de cosecha 1  Vino 2      Ligero,afrutado
```

```
8 Año de cosecha 2  Vino 2     Azucarado,ligero
```

La función **textual** permite construir la tabla de contingencia para cada modalidad de una o varias variables cualitativas o para cada combinación de modalidades de dos variables cualitativas. El argumento `sep.word` permite precisar los separadores de palabras y el argumento `maj.in.min` permite poner todas las palabras en minúsculas. La fila de encargo siguiente construye una tabla de contingencia con las palabras en columna, en tanto que las modalidades en fila de la segunda variable y las combinaciones de modalidades de la primera con la segunda variable. Ella representa también el número de veces que se usa una palabra (objeto `nb.words`) y en cuántas filas se ha empleado (salida útil para textos pero sin interés para preguntas abiertas, ya que la misma palabra no se repite muchas veces).

```
> textual(vina,num.text=3,contingence.by=list(2,1:2),
    sep.word=",",maj.in.min=TRUE)
$cont.table
                         acidez afrutado azucarado con toques de madera licoroso ligero
Vino 1                        2        2         0                     1        0      1
Vino 2                        0        1         3                     0        1      2
Año de cosecha 1.Vino 1       1        1         0                     1        0      0
Año de cosecha 1.Vino 2       0        1         1                     0        0      1
Año de cosecha 2.Vino 1       1        1         0                     0        0      1
Año de cosecha 2.Vino 2       0        0         2                     0        1      1

$nb.words
                      words nb.list
ligero                    3       3
azucarado                 3       3
afrutado                  3       3
acidez                    2       2
licoroso                  1       1
con toques de madera      1       1
```

2.8 Ejemplo : datos de Juegos Olímpicos

2.8.1 Descripción de datos

La tabla de datos cruza en filas las pruebas de atletismo y en columnas los diferentes países. Cada casilla contiene el número total de medallas (oro, plata y bronce) obtenidas en las Olimpiadas desde 1992 hasta el 2008 (Barcelona 1992, Atlanta 1996, Sídney 2000, Atenas 2004, Pekín 2008). Proporcionamos un extracto del juego de datos en la tabla 2.9. En las cinco Olimpiadas, 58 países obtuvieron por lo menos una medalla en una de las 24 pruebas realizadas : 10.000 m, 100 m, 110 m vallas, 1500 m, 200 m, 20 km, 3000 m carrera de obstáculos, 400 m, 400 m vallas, 4×100 m, 4×400 m, 5000 m, 50 km, 800 m, decatlón, disco, salto de altura, jabalina, salto de longitud, maratón, martillo, salto con pértiga, peso, salto de triple. La tabla contiene muchos 0, ya que el número total de medallas otorgadas es de 360 mientras que el número de casillas de la tabla es de 1392 :

```
> library(FactoMineR)
> JO <- read.table("http://factominer.free.fr/libra/JO.csv", header=TRUE, sep=";",row.names=1)
```

	usa	ken	rus	gbr	eti	cub	mar	ale	jam	pol
10000 m	0	4	0	0	8	0	2	0	0	0
100 m	5	0	0	1	0	0	0	0	1	0
110 m vallas	9	0	0	0	0	3	0	1	0	0
1500 m	0	5	0	0	0	0	3	0	0	0
200 m	8	0	0	1	0	0	0	0	1	0
20 km	0	0	3	0	0	0	0	0	0	1
3000 m Obstáculos	0	12	0	0	0	0	1	0	0	0
400 m	11	1	0	1	0	0	0	0	1	0
400 m vallas	7	0	0	1	0	0	0	0	2	0
4x100 m	4	0	0	1	0	2	0	0	1	0
4x400 m	5	0	1	2	0	1	0	0	2	0
5000 m	0	5	0	0	4	0	3	1	0	0
50 km	0	0	4	0	0	0	0	1	0	3
800 m	1	5	1	0	0	0	0	1	0	0
Decatlón	5	0	0	0	0	1	0	1	0	0
Disco	0	0	0	0	0	1	0	3	0	1
Salto de altura	3	0	3	2	0	2	0	0	0	1
Jabalina	0	0	2	3	0	0	0	0	0	0
Salto de longitud	7	0	0	0	0	2	0	0	1	0
Maratón	1	3	0	0	3	0	1	1	0	0
Martillo	1	0	0	0	0	0	0	0	0	1
Pértiga	4	0	3	0	0	0	0	1	0	0
Peso	8	0	0	0	0	0	0	0	0	1
Salto de triple	3	0	2	3	0	2	0	0	0	0

Tabla 2.9 – Datos de Juegos Olímpicos : número de medallas obtenidas por disciplina y por país durante cinco olimpiadas. Extracto : los diez países que ganaron más medallas

2.8.2 Problemática

Se trata de una tabla de contingencia. Los individuos son las 360 medallas. A cada medalla le son asociadas dos variables cualitativas : la prueba a la cual se refiere y el país al cual pertenece el que la obtuvo. La tabla cruza estas dos variables.

Desde un punto de vista un poco formal, la problemática asociada a tal tabla consiste en el estudio de la relación entre ambas variables : prueba y país. Pero esta manera de redactar es difícil de entender. Podemos hacerla más concreta así : existen asociaciones notables como «pruebas-países» en un sentido (*i.e.*, tal país obtiene medallas sólo en tal prueba), o en el otro (tal país no gana medallas en tal prueba mientras que consigue medallas en otras pruebas).

El recurso a la noción de perfil, la clave del AFC, es aquí más evidente, más preciso y más rico. Primero definimos el perfil atlético de un país por el conjunto de sus medallas, distribuidas por pruebas (concretamente, una columna de la tabla). La problemática se convierte entonces en la siguiente : ¿podemos considerar que todos los países tienen el mismo perfil atlético o, por el contrario, ciertos países alcanzan mejores resultados en ciertas pruebas ? Y, en este último caso, podemos sintetizar dichas «especializaciones» ? Por ejemplo, poniendo de manifiesto oposiciones, por un lado, entre países que tienen los mismos perfiles (*i.e.*, habiendo ganado las mismas pruebas) y por otro lado, los que tienen el perfil opuesto (*i.e.*, no habiendo ganado las mismas pruebas).

De modo dual, la distribución de las medallas de una prueba por países define el «perfil

geográfico» de la prueba (concretamente, una fila de la tabla). ¿Podemos considerar que todas las pruebas tienen el mismo perfil geográfico o, por el contrario, ciertas pruebas son la especialidad de ciertos países ? Podemos sintetizar estas especializaciones poniendo de manifiesto oposiciones, por un lado, entre pruebas que tienen el mismo perfil (*i.e.*, ganadas por los mismos países), y por otro lado, las pruebas que tienen un perfil opuesto (*i.e.*, ganadas por otros países) ?
Los dos puntos de vista anteriores se apoyan implícitamente sobre una noción de semejanza entre perfiles. En esta semejanza, el número total de medallas de un país no debe intervenir porque conduciría a separar los países que habrían obtenido muchas medallas de otros, por eso el AFC no es útil. Así mismo, la noción de perfil precedente hay que comprender la en el sentido del AFC, es decir, en el sentido de la probabilidad condicional o, más simplemente, en el sentido de los porcentajes (de medallas obtenidas para cada prueba por un país).

Observación sobre los márgenes.
En estos datos, por construcción, el margen columna debe ser constante e igual a : 3 (tipos de medallas) $\times$ 5 (olimpiadas) = 15 (no obstante, hay algunas excepciones debido a las anulaciones de medalla). Esto implica dos consecuencias. En primer lugar, las pruebas tienen el mismo peso en el análisis (y a perfil constante, la misma influencia). Luego, el perfil atlético «medio», que sirve de referencia (situado en el origen de los ejes) es un perfil constante. El AFC, que pone en evidencia las diferencias del perfil medio, hará desempeñar un papel importante a los países que tendrán un perfil atlético muy especializado (el caso más extremo es aquel en que todas las medallas de un país provienen de la misma prueba).
El margen fila contiene el número total de medallas de cada país. Estos números son muy variables (1 medalla para 18 países y 82 medallas para Estados Unidos). Los pesos de los países son muy diferentes unos de otros : a perfil constante, los país es con más medallas tienen una influencia más fuerte en el análisis. El perfil de referencia (situado en el origen de los ejes) contiene las proporciones de medallas obtenidas por los países (muy diferente de un perfil constante) : así, una prueba A puede ser más caracterizada por un país X que por un país Y, aunque X obtuvo menos medallas que Y en esta prueba (porque Y ganó, en total, muchas más medallas que X).
Los márgenes pueden calcularse una vez realizado el análisis de correspondencias (ver el final de esta sección para la obtención de los márgenes).

2.8.3 Elección del análisis

2.8.4 Puesta en práctica del análisis

Aquí consideramos todas las filas y todas las columnas como activas. Para efectuar este análisis, utilizamos la función **CA** del paquete FactoMineR, donde los principales parámetros de entrada son la tabla de datos, los índices de filas suplementarias y los índices de columnas suplementarias. Por defecto, ninguna fila ni ninguna columna son suplementarias (`row.sup=NULL` y `col.sup=NULL`), es decir, todos los elementos son activos.

```
> res.ca <- CA(JO)
```

La función **CA** proporciona el gráfico del AFC que representa las filas y las columnas, así como las salidas numéricas siguientes contenidas en el objeto `res.ca` :

```
> res.ca
**Results of the Correspondence Analysis (CA)**
The variable in rows have 24 categories, the variable in columns 58 categories
The chi square of independence between the two variables is equal to 2122.231
(p-value = 2.320981e-41).
*The results are available in the following objects:

   name              description
1  "$eig"            "eigenvalues"
2  "$col"            "results for the columns"
3  "$col$coord"      "coord. for the columns"
4  "$col$cos2"       "cos2 for the columns"
5  "$col$contrib"    "contributions of the columns"
6  "$row"            "results for the rows"
7  "$row$coord"      "coord. for the rows"
8  "$row$cos2"       "cos2 for the rows"
9  "$row$contrib"    "contributions of the rows"
10 "$row.sup$coord"  "coord. for the supplementary rows"
11 "$row.sup$cos2"   "cos2 for the supplementary rows"
12 "$call"           "summary called parameters"
13 "$call$marge.col" "weights of the columns"
14 "$call$marge.row" "weights of the rows"
```

Previamente al AFC, el test de χ^2 indica si la diferencia de la tabla de independencia puede atribuirse o no a «fluctuaciones aleatorias» (ya que este test tiene en cuenta el efectivo global, al contrario del AFC). El estadístico de χ^2 vale 2122 y se le asocia a una probabilidad crítica de 2.32×10^{-41}.
Pero aquí el efectivo total ($5 \times 5 \times 24 = 360$ medallas) es muy débil con respecto al número de casillas de la tabla ($24 \times 58 = 1392$). Estamos, pues, muy lejos de las condiciones de validez del test (incluso los más «laxistas» que suponen que el 80 % de los efectivos teóricos son superiores a 5 y que los otros son superiores a 1) y la probabilidad crítica puede ser considerase sólo a título indicativo. No obstante, es tan débil que la significación de la diferencia de esta tabla a la independencia está fuera de duda.

Elección del número de dimensiones por estudiar

Como en todos los análisis factoriales, el estudio de la inercia de los ejes permite, por una parte, ver si existe una estructura en los datos, y por otra parte, determinar el número de componentes principales por interpretar.
El objeto `res.ca$eig` contiene el valor propio (*i.e.*, la inercia o la varianza explicada) asociado a cada dimensión, el porcentaje de inercia que representa en el análisis, así como la acumulación de estos porcentajes. Damos aquí los resultados redondeados a los dos primeros decimales con la ayuda de la función **round** :

```
> round(res.ca$eig,2)
        eigenvalue  percentage  cumulative percentage
                      variance            of variance
dim 1         0.82        13.85                 13.85
```

```
dim 2        0.62        10.53                24.38
dim 3        0.54         9.23                33.62
dim 4        0.48         8.16                41.78
dim 5        0.40         6.72                48.50
dim 6        0.36         6.17                54.67
dim 7        0.33         5.55                60.23
dim 8        0.32         5.35                65.58
dim 9        0.27         4.56                70.14
dim 10       0.24         4.16                74.29
dim 11       0.23         3.91                78.20
dim 12       0.18         3.11                81.31
dim 13       0.16         2.78                84.09
dim 14       0.14         2.46                86.55
dim 15       0.13         2.22                88.77
dim 16       0.12         2.06                90.82
dim 17       0.10         1.76                92.58
dim 18       0.09         1.58                94.16
dim 19       0.08         1.44                95.60
dim 20       0.08         1.35                96.95
dim 21       0.07         1.27                98.21
dim 22       0.06         1.05                99.27
dim 23       0.04         0.73               100.00
```

Podemos visualizar estos valores propios con la ayuda de un diagrama en barras (cf. figura 2.10) :

```
> barplot(res.ca$eig[,1], main="Valores propios", names.arg=paste("dim",1:nrow(res.ca$eig)))
```

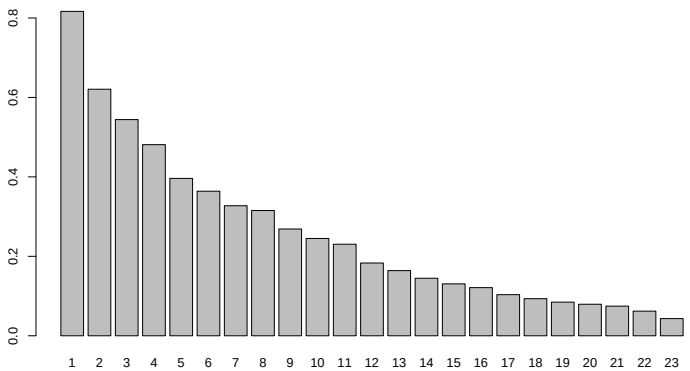

FIGURE 2.10 – Datos de Juegos Olímpicos : valores propios asociados a cada dimensión del AFC.

Los dos primeros ejes expresan 24.40% de la inercia total. Puede ser interesante interpretar los ejes siguientes, que expresan igualmente un porcentaje importante de inercia total.

Estudio de la representación superpuesta

La representación superpuesta del AFC (cf. figura 2.11) es una salida por defecto de la función **CA**.

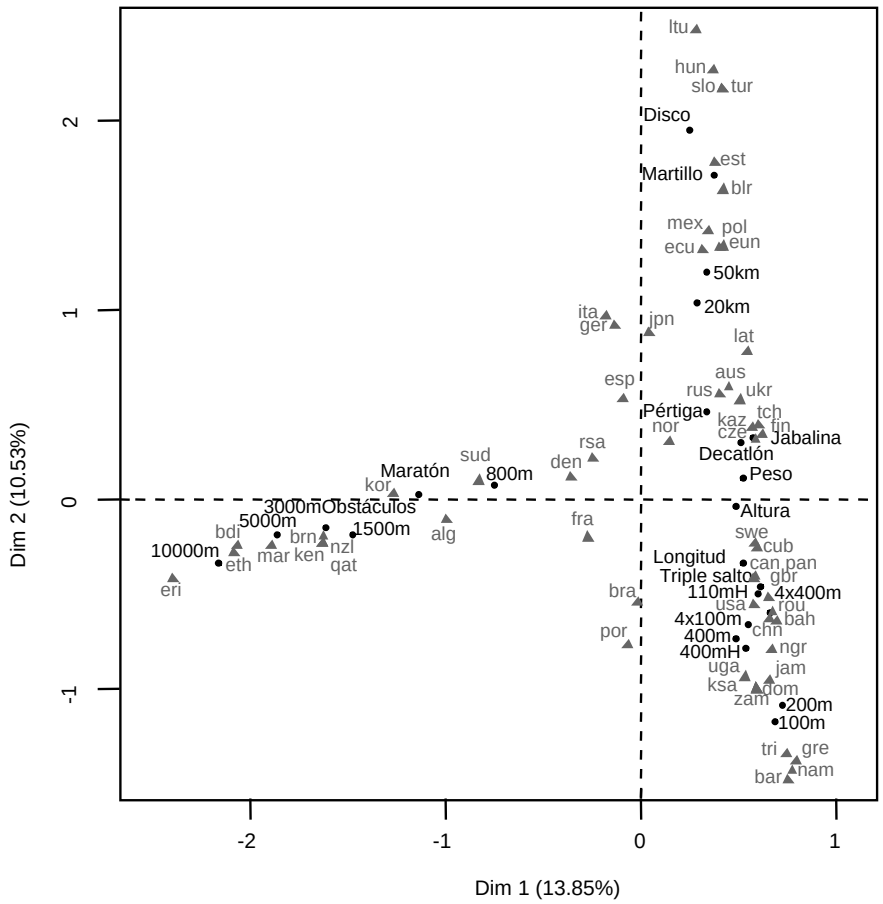

FIGURE 2.11 – Datos de Juegos Olímpicos : representación superpuesta.

Podemos encontrar el conjunto de las coordenadas de las filas (resp. columnas) en el objeto `res.ca$row` (resp. `res.ca$col`). Obtenemos entonces una tabla con las coordenadas, las contribuciones (lo que indica en qué medida un individuo contribuye a la construcción de un eje), los cosenos al cuadrado (lo que mide la calidad de la proyección de los individuos sobre un eje) y las inercias para cada elemento (lo que corresponde a la distancia al baricentro ponderado por el peso del elemento).

Podemos así construir el gráfico con los ejes 3 y 4. Utilizamos la función **plot.CA** (que puede ser llamarse **plot** o **plot.CA**). Precisamos entonces los ejes de representación (`axes = 3:4`) :

```
> plot(res.ca, axes = 3:4)
```

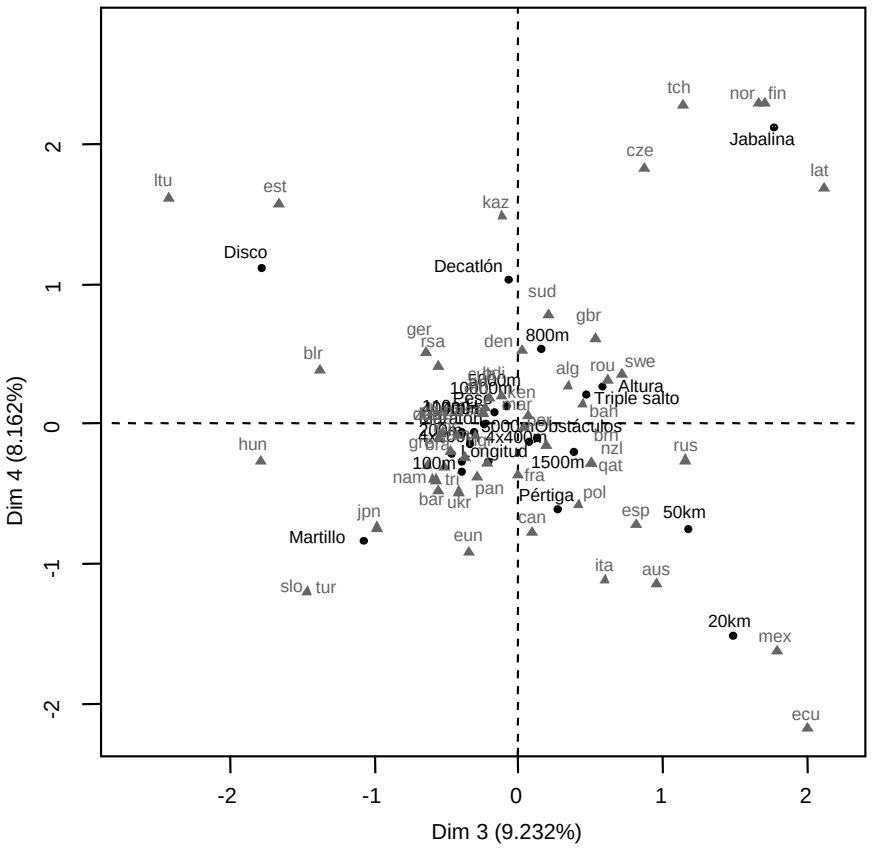

FIGURE 2.12 – Datos de Juegos Olímpicos : representación superpuesta sobre el plano (3, 4).

Comentarios sobre los datos

Ante todo, podemos interesarnos por las proyecciones de las diferentes pruebas sobre el primer plano factorial. Los resultados son bastante espectaculares, ya que las pruebas de carreras de fondo están bien separadas de las otras pruebas sobre el primer eje factorial. Además, hay un gradiente entre estas pruebas, empezando por los 10.000 m y yendo hasta los 800 m. Todas las pruebas son clasificadas de la distancia más larga a la más corta, sin ninguna excepción. Esto muestra que los resultados de los 10.000 m son más particulares que las otras pruebas de fondo. No obstante, podemos anotar que el maratón está más próximo al centro del gráfico que lo esperado. Esto se explica por el hecho de que no es una prueba de fondo como las otras.

Los países que tienen coordenadas negativas sobre el primer eje son aquellos que ganan numerosas medallas en las pruebas de fondo en comparación con los resultados obtenidos

por estos países en otras pruebas, pero también en comparación con el número de medallas ganadas por otras naciones en las pruebas de fondo. Encontramos numerosos países africanos especialistas en las pruebas de fondo (Eritrea, Etiopía, Burundi, Marruecos, Catar, Kenia) y también Nueva Zelanda (atención, Nueva Zelanda obtuvo sólo una medalla en los 1500 m, de ahí su coordenada extrema).
Es interesante ver aquí las contribuciones de los diferentes países. Recordemos que en AFC, contrariamente al ACP, los elementos más extremos no son necesariamente los que más contribuyeron a la construcción de los ejes ya que los pesos son diferentes de un elemento al otro. Las contribuciones de los trece países que más contribuyeron a la construcción del eje 1 se indican más abajo (los países son clasificados por contribución decreciente) :

```
> res.ca$col$contrib[rev(order(res.ca$col$contrib[,1])),1]
   ken    eth    mar    usa    gbr    eri    cub    bdi    alg    jam    tri    kor
31.387 22.072 12.160  9.149  2.139  1.947  1.683  1.452  1.352  1.313  1.119  1.089
```

Etiopía, Kenia y Marruecos contribuyeron un 65 % en la construcción de la primera dimensión. Son países que obtuvieron muchas medallas. Ellos tres ganaron 60 medallas en total, de las que 59 corresponden a las pruebas de fondo.
El segundo eje separa, en cuanto a él, las pruebas de velocidad de las pruebas de lanzamientos de disco y de martillo y de las pruebas de marcha (20 km y 50 km). Aquí existe un gradiente entre las pruebas de velocidad : la carrera de 100 m es más extrema que las de 200 m y las de 400 m. Las pruebas de relevo son también menos extremas que las pruebas individuales. Podemos anotar aquí que los 400 m es una prueba de velocidad, mientras que los 800 m es una prueba de fondo. De la misma manera, la marcha (20 km y 50 km) está separada de las pruebas de fondo y de velocidad. Aquí la prueba de los 50 km es más extrema que la de 20 km.
Los países que ganan medallas en velocidad son Barbados, Namibia, Trinidad y Tobago, Jamaica, República Dominicana, etc. Las contribuciones de los países en la construcción de la segunda dimensión son mucho más homogéneas que para el primer eje. Los Estados Unidos contribuyó mucho en la construcción de este eje, aunque su coordenada es relativamente próxima a 0. Esto se explica por la gran cantidad de medallas que obtuvo : 82 en total, de las que 49 corresponden a las pruebas de velocidad (a comparar con el porcentaje de las pruebas de velocidad : 7/24). Damos más adelante los quince países que más contribuyeron en la construcción del eje 2 :

```
> res.ca$col$contrib[rev(order(res.ca$col$contrib[,2])),2]
   usa    ltu    blr    hun    pol    eun    tri    est    ger    nam    jam    mex
11.324 10.942  7.175  6.911  6.314  5.582  4.790  4.234  3.766  3.643  3.629  3.608
```

Para las pruebas de lanzamiento de martillo y de disco, observamos que los países más eficientes son Lituania, Hungría, Eslovenia y Turquía.
Los ejes 3 y 4 separan de nuevo disco y martillo de las pruebas de marcha (20 km y 50 km). La jabalina es una prueba de lanzamiento verdaderamente diferente de las pruebas de martillo y de disco. Los países nórdicos (Noruega, República Checa, Finlandia, Letonia) son los más eficientes en el lanzamiento de la jabalina.
Es posible obtener los márgenes filas y los márgenes columnas (así como el número de medallas obtenidas por países multiplicando el margen columna por el número total de medallas, 360) :

```
> res.ca$call$marge.row
> res.ca$call$marge.col[rev(order(res.ca$call$marge.col))]
  usa   ken   rus   eth   gbr   cub   ger   mar   jam   pol   esp   ita
0.228 0.097 0.053 0.042 0.042 0.039 0.028 0.028 0.025 0.022 0.022 0.019
> res.ca$call$marge.col[rev(order(res.ca$call$marge.col))]*360
usa ken rus eth gbr cub ger mar jam pol esp ita
 82  35  19  15  15  14  10  10   9   8   8   7
```

Comentario sobre los datos

El aficionado al atletismo podría decepcionarse en una primera lectura de este ejemplo. De hecho, el AFC devuelve las grandes tendencias que se liberan de los datos : que las pruebas de fondo son dominadas por los atletas africanos, que las de velocidad son monopolizar por Estados Unidos, que velocidad, fondo y lanzamientos son pruebas bastante diferentes. Esto es lo que se pide a un método de análisis : encontrar las principales características.
Sin embargo, más detalladamente, ciertos resultados son interesantes y pueden despertar la curiosidad del aficionado al atletismo (incluso la del especialista). Listamos algunos resultados y dejamos al especialista ir más lejos en la interpretación.

- Los resultados del AFC muestran una separación bastante clara entre las pruebas de fondo (1500 m, 3000 m obstáculos, 5000 m, 10.000 m y maratón) y las pruebas de velocidad (100 m y 200 m). Ambas pruebas de fondo, 400 m y 800 m, no son reagrupadas y el conjunto de pruebas de carreras son separadas en dos, con un límite que se sitúa entre los 400m y los 800 m (la prueba de 800 m está próximo a las carreras de fondo mientras que la de 400 m está próxima a las pruebas de velocidad). La manera de gastar la energía es distinta para estas dos pruebas.
- Por otro lado, el maratón es una prueba de fondo que no se comporta como las otras : tiene una posición mucho menos extrema que lo esperado. Igualmente, las pruebas de marcha (20 km y 50 km) no son de fondo, como las carreras.
- Los atletas que corren las pruebas de velocidad a menudo tienden a «hacer dos pruebas» y a correr los 100 m y 200 m o 200 m y 400 m. El gráfico muestra que los 100 m y 200 m son dos pruebas muy próximas, más que las de 200 m y 400 m. Los 100 m y 200 m son efectivamente dos pruebas de potencia, mientras que la de 400 m es una prueba de fondo.
- Las dos pruebas de vallas (110 m y 400 m) son bastante diferentes : la carrera de 110 m vallas está relativamente alejada de los 100 m, mientras que la de 400 m vallas está muy próxima a los 400 m. Las pruebas de los 100 m y de los 110 m vallas utilizan son muy diferentes : prueba muy técnica para los 110 m vallas y prueba explosiva para los 100 m, lo que explica que ningún atleta participa en estas dos pruebas. En cambio, los 400 m vallas es una prueba mucho menos técnica que los 110 m vallas ; tiene características comunes con los 400 m, lo que explica que ciertos atletas puedan correr estas dos pruebas.
- En los lanzamientos, martillo y disco son unas pruebas muy próximas, mientras que bala y jabalina son muy diferentes. Martillo y disco son dos lanzamientos en rotación (con un efecto de palanca), mientras que la jabalina es lanzar en línea y la bala un lanzamiento con rotación o sin ella (y sin efecto de palanca, la bala debe estar pegada al cuello durante el lanzamiento).
- El decatlón, prueba completa por excelencia, es opuesto a las pruebas de fondo sobre el primer eje. Los atletas de fondo, pues, no son favorecidos en esta prueba. Efectivamente,

los atletas de decatlón tienen una masa muscular importante y características físicas de explosión que no les favorecen en las pruebas de fondo : estos atletas tienen dificultades para terminar la prueba de 1500 m.

Recordemos que todos estos comentarios se hacen a partir del número de medallas obtenidas por país y por disciplina, sin referencia a las características físicas de los atletas de las diferentes pruebas.

2.9 Ejemplo : diez vinos blancos del Valle del Loira

2.9.1 Descripción de los datos y problemática

En el marco de una investigación sobre la singularidad de vinos que proviene de la cepa chenin, en el Valle del Loira (investigación iniciada por C. Asselin, Interloire, Angers y realizada en el Agrocampus), estudiamos un conjunto de diez vinos blancos secos de Touraine, cinco Touraine DOC, procedentes de la cepa sauvignon, y cinco Vouvray DOC, procedentes de la cepa chenin (cf. tabla 2.10).

Estos vinos escogió por J.P. Gouvazé (Interloire, Tours) con el fin de ilustrar la diversidad, en el seno de cada cepa, de los vinos del Loira (no obstante, hay una restricción en esta diversidad ; viendo su profesión, podemos pensar que la persona encargada de la selección ha descartado los vinos que no le gustaban). Observamos que hay una confusión total (en el sentido de la planificación experimental) entre la denominación y la cepa. Más tarde, para simplificar, hablaremos sólo del factor cepa.

Numéro	Nombre	Cepa	Denominación	Observación
1	Michaud	sauvignon	Touraine	
2	Renaudie	sauvignon	Touraine	
3	Trotignon	sauvignon	Touraine	
4	Buisse	sauvignon	Touraine	
5	BuisseCristal	sauvignon	Touraine	
6	Aub. Silex	chenin	Vouvray	7g de azúcar residual
7	Aub. Marigny	chenin	Vouvray	Elaboración en barrica
8	Font Domaine	chenin	Vouvray	
9	Font Brûlés	chenin	Vouvray	
10	Font Coteaux	chenin	Vouvray	Elaboración en barrica

Tabla 2.10 – Datos de vinos : los diez vinos estudiados.

Estos vinos dieron lugar a numerosos análisis sensoriales, combinando diferentes tipos de jurado y diferentes protocolos. Los datos analizados aquí provienen de una degustación que reúne a doce profesionales y contiene un aspecto «textual». La pregunta formada era : para cada vino, dé una (o varias) palabra(s) que, según usted, caracteriza(n) sensorialmente este vino. Como es habitual, la degustación se efectuó «a ciegas» y los vinos se presentaron sin ninguna indicación. No obstante, en esta degustación, que se efectuó durante un salón de vinos del Loira, los catadores pensaron que se trataba de vinos del Loira aunque esto no se les había dicho ; pero la diversidad de los vinos del Loira, desde el punto de vista de las tierras, las cepas y las prácticas viti–vinícolas, autoriza a pensar que este nivel general de información sobre el conjunto de los vinos no tiene repercusión sobre las caracterizaciones relativas de los vinos.

Se trata, en cierto modo, de un cuestionario que contiene diez preguntas abiertas (una por vino). Estos datos son reunidos en una tabla cuyas filas son los vinos, cuyas columnas son las palabras y cuyo término general x_{ij} es el número de veces que la palabra j ha sido asociada al vino i (cf. tabla 2.11, que figura de un modo transpuesto por razones de presentación de la página).

	1S-Mic	2S-Ren	3S-Tro	4S-Bui	5S-Bui	6C-Aub	7C-Aub	8C-Fon	9C-Fon	10C-Fon	Suma
Afrutado	1	5	5	3	4	0	1	4	3	1	27
Azucarado, sutil, licoroso	0	1	1	0	0	11	1	2	1	1	18
Con toques de madera	1	0	0	0	2	0	7	0	1	5	16
Ligero, suave	1	0	2	2	1	2	0	0	4	0	12
Acidez	1	0	1	2	1	0	2	1	2	1	11
Cítrico	2	3	1	1	1	0	0	3	0	0	11
Amarillo dorado	2	0	0	1	0	1	2	1	2	2	11
Alegre	2	3	0	1	3	1	1	0	0	0	11
Aromas afrutados	2	1	2	1	0	1	0	1	1	0	9
Fino, discreto	0	2	1	4	0	0	0	1	1	0	9
Amargo	1	1	0	0	0	0	0	1	2	3	8
Floral	0	1	2	0	2	0	0	1	1	1	8
Graso, falta de frescura	0	0	0	0	0	2	2	1	2	1	8
Amarillo pálido, claro	1	2	2	0	1	2	0	0	0	0	8
Fresco en boca	1	2	2	2	0	0	0	0	0	0	7
Largo, muy largo	1	1	1	0	0	0	2	0	1	1	7
Floral, con fondo de flores blancas	2	1	1	0	1	0	0	0	0	1	6
Seco	0	0	0	3	1	0	0	1	1	0	6
Intenso, amplio	1	0	0	0	0	1	1	0	1	1	5
Miel	0	1	0	0	0	1	1	1	1	0	5
Complejo, corto	0	0	0	0	0	3	0	2	0	0	5
Abierto, expresivo	2	0	1	0	0	0	0	1	1	0	5
Con buena expresión aromática	1	1	1	1	0	0	0	0	0	0	4
Sabor extraño (cera, neumático)	0	0	0	0	0	0	3	0	0	1	4
Sabor poco maduro	2	0	2	0	0	0	0	0	0	0	4
Yodado	1	1	0	1	1	0	0	0	0	0	4
Poca acidez	1	0	0	1	2	0	0	0	0	0	4
Poco carácter, poca expresión	0	0	0	0	0	1	2	0	1	0	4
Sauvignon	1	1	1	0	0	0	0	0	0	1	4
Olor persistente	1	0	0	0	2	0	1	0	0	0	4
Suma	28	27	26	23	22	26	26	21	26	20	245

Tabla 2.11 – Datos de vinos : número de veces que cada palabra se ha utilizado para cada vino (30 palabras).

Esta tabla puede verse como una tabla de contingencia, considerando que se dispone de n descripciones sensoriales (una descripción es la asociación entre una palabra y un vino) y que estas descripciones se clasifican según dos variables cualitativas : el vino al cual se refieren y la palabra utilizada. El AFC va a analizar la diferencia entre esta tabla y el modelo de independencia, modelo según el cual cada vino tiene el mismo perfil de palabras y cada palabra se usa, en proporción, el mismo número de veces para cada vino.

Este tipo de tabla a menudo está constituida y analizada por AFC (históricamente, el primer AFC publicado trataba de una tabla análoga, Escofier, 1965), pero habitualmente con efectivos mucho más importantes. Estamos aquí en condiciones límites a causa de un número total de casos ($n = 245$) muy débil. Sin embargo, el análisis es posible por el hecho de que se trata de un vocabulario más bien estereotipado de los profesionales del vino, lo que conduce a un número total de palabras no demasiado elevado y, entonces, un número

«suficiente» de palabras que presentan un efectivo no muy débil. Además, antes del análisis, algunas palabras «vecinas» han sido reagrupadas (por ejemplo, *Azucarado*, *Sutil* y *Licoroso*, que reenvían la misma percepción, la del sabor azucarado). En este texto, con la intención de una simplificación, guardamos el término «palabra» para las filas de la tabla 2.11, incluso cuando representan grupos de palabras que figuran tal cual en los cuestionarios (*Falta de frescura*) o procedentes de una reagrupación *a posteriori* (*Azucarado*, *Sutil*).
En este tipo de análisis, eliminamos las palabras menos utilizadas. Teniendo en cuenta los débiles efectivos, el límite se ha fijado en 4, límite por debajo del cual las palabras no se toman en consideración. La determinación de este límite tiene siempre un carácter empírico : con estos datos, poner el límite en 5 no cambia fundamentalmente la representación de los vinos pero priva de palabras importantes (por ejemplo, «sauvignon»), y el límite en 3 conduce a gráficos muy cargados, que además tienen palabras cuyas coordenadas son frágiles.
El objetivo de este análisis es proporcionar una imagen sintética de la diversidad de estos vinos. Habitualmente, la diversidad sensorial de los vinos es estudiada con la ayuda de un protocolo mucho más pesado : se establece un cuestionario, que contiene una lista de descriptores (acidez, amargura, etc.) ; un jurado está preparado para la evaluación de los vinos con la ayuda de estos descriptores, y se hace la evaluación final. Uno de los objetivos de este estudio es también metodológico : ¿es posible obtener con un protocolo muy simplificado (los catadores no se preparan de manera conjunta ; utilizan su propio vocabulario) una imagen interesante de la diversidad de los vinos ?

2.9.2 Márgenes

El examen de los márgenes es importante tanto desde el punto de vista de su interpretación directa (¿cuáles son las palabras más utilizadas ? ¿Ciertos vinos son objeto de más comentarios que otros ?) como del de su influencia en AFC (como peso).
La palabra más utilizada es «afrutado», lo que está en consonancia con la observación habitual de comentarios de degustación (para convencerse, leer la etiqueta de cualquier botella : es difícil de escapar de «maravilloso afrutado»). Luego viene el conjunto *Azucarado, Sutil, Licoroso*. Recordemos que estos vinos son secos y una percepción de azucarado es, *de facto*, una característica notable. Por fin, la percepción de *Con toques de madera*, asociada a una elaboración en barrica, está bien identificada por los profesionales, lo que favorece la aparición de una citación elevada frecuente (por oposición a un olor fácilmente percibido pero no reconocido que engendra palabras diferentes según los catadores). Prolongar el comentario de estos efectivos marginales sobrepasa el marco de un libro de estadística. Sobre el plano técnico, desde el punto de vista del AFC, las palabras tendrán, a perfil igual, un peso tan importante que hace que hayan sido citadas frecuentemente, lo que es deseable.
En cambio, el número de palabras por vino parece homogéneo. Ningún vino parece atraer más comentarios que otros, lo que es sin duda una consecuencia (deseada) de la forma de la pregunta realizada («Para cada vino, dar una o algunas palabras...»). Por tener la conciencia tranquila, podemos realizar un test χ^2 de ajuste de los diez efectivos observados (última fila de la tabla 2.11) a una ley uniforme. La probabilidad crítica (0.97) confirma que no hay que prestar atención a las diferencias entre los efectivos de las palabras por vino. Desde el punto de vista del AFC, podemos considerar que los vinos tendrán, a perfil igual, más o menos la misma influencia en el análisis. Cuando el número de palabras por vino es diferente, el análisis

concede a un vino un peso importante, hecho que fue objeto de numerosos comentarios (su perfil se conoce mejor).

2.9.3 Inercias

La inercia total (Φ^2) vale 1.5 lo que lo conduce a un χ^2 de 368.79 ($n = 245$), asociado a una probabilidad crítica de 1.23×10^{-5}. La tabla está en condiciones de validez del test muy malas (en principio, por lo menos 80 % de los efectivos teóricos debe ser superior a 5 y ninguno debe ser nulo), pero la probabilidad crítica es tan débil que el interés del AFC sobre estos datos está fuera de duda. Observemos que, para el que conoce la diversidad entre los vinos del Valle del Loira (sobre todo teniendo en cuenta el hecho de que estos vinos han sido escogidos para ilustrar esta diversidad), la relación entre las palabras y los vinos es la esperada. La pregunta «realizada al χ^2» no es tanto la de la existencia de una relación como la de la aptitud de un conjunto tan limitado de datos para poner en evidencia esta relación. Podemos considerar aquí que la respuesta es positiva pero que los datos no tienen la «solidez estadística» de la tabla sobre las opiniones con respecto al trabajo femenino (recordemos : para esta última tabla, p-crítico $= 10^{-49}$). También aumentaremos la prudencia en la interpretación (lo que concretamente quiere decir : vueltas frecuentes a los datos brutos y puesta en relación con informaciones exteriores a los datos). Estas observaciones son muy importantes ya que el AFC, teniendo en cuenta sólo las probabilidades, no da ninguna garantía desde el punto de vista de la significación.
La intensidad de la relación, medida por el V de Cramer, es más bien elevada : 0.409 (el valor 1 correspondería a una asociación exclusiva entre cada vino y un grupo de palabras, máximo impensable para el que conoce la dificultad de una degustación a ciegas) ; es más elevada, por ejemplo, que la de los datos sobre el trabajo femenino (0.26).
La puesta en práctica del AFC se obtiene con los comandos siguientes :

```
> library(FactoMineR)
> vinos = read.table("http://factominer.free.fr/libra/vinos.csv",header=T,row.names=1,sep=";")
> colnames(vinos)=c("1S.Michaud","2S.Renaudie","3S.Trotignon","4S.Buisse","5S.BuisseCristal",
    "6C.AubSilex","7C.Aub.Marigny","8C.FontDomaine","9C.FontBrûlés","10C.FontCoteaux","Suma")
> res.ca=CA(vinos,col.sup=11,row.sup=nrow(vinos))
> barplot(res.ca$eig[,1],main="Valores propios", names.arg=1:nrow(res.ca$eig))
```

La secuencia de los valores propios (cf. figura 2.13 y tabla 2.12) muestra dos ejes de inercia mucho más importantes que los ejes siguientes, lo que añadido al porcentaje de inercia acumulado de 53.6% incita a concentrar la interpretación en el primer plano. Cada uno de estos dos ejes tiene una inercia bastante elevada (0.436 y 0.371) : las asociaciones entre vinos y palabras deberían aparecer claramente.

2.9.4 Representación sobre el primer plano

Varias interpretaciones del análisis son posibles. En vez de una interpretación por eje, preferimos, para comenzar, una interpretación por grupos fundada sobre los vinos. Tres grupos aparecen :

– Aubuissières Silex (6). Caracterizado por *Azucarado*, citado once veces para este vino ; es el único que contiene azúcar residual con una concentración claramente perceptible ;

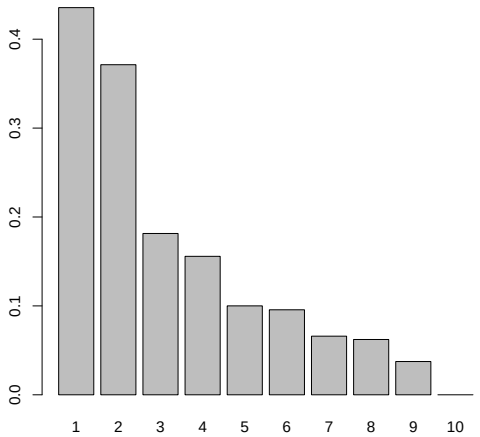

FIGURE 2.13 – Datos de vinos : diagrama de los valores propios del AFC de la tabla 2.11.

```
> round(res.ca$eig,3)
        eigenvalue  percentage  cumulative percentage
                      variance            of variance
dim 1        0.436      28.932                 28.932
dim 2        0.371      24.666                 53.598
dim 3        0.181      12.055                 65.653
dim 4        0.156      10.348                 76.001
dim 5        0.100       6.645                 82.646
dim 6        0.096       6.353                 88.999
dim 7        0.066       4.382                 93.380
dim 8        0.062       4.133                 97.513
dim 9        0.037       2.487                100.000
dim 10       0.000       0.000                100.000
```

Tabla 2.12 – Datos de vinos : valores propios.

esta característica, insólita (pero autorizada) en un vino seco, se destaca claramente en este sentido y es relativamente poco citada para otros vinos (siete veces pero nunca más de dos veces para el mismo vino) y constituye más del tercio de las palabras asociadas a este vino. El gráfico valoriza la falta de aromas de este vino ; pero como dicho término es citado sólo tres veces para este vino, le concedemos sólo un segundo puesto (además, esta característica es más bien una ausencia de característica, menos evocadora ; hablaremos de ello posteriormente).

– Aubuissières Marigny (7) + Fontainerie Coteaux (10). Estos dos vinos, caracterizados principalmente *Con toques de madera*, se citan respectivamente siete y cinco veces para cada uno, cuando esta palabra sólo se ha utilizado tres veces por los otros. Tal descripción, evidentemente, tiene que relacionarse con el hecho de que estos dos vinos son los únicos que se han elaborado en barrica. Según el plano, *Sabor extranjero* caracteriza mejor estos vinos, pero lo citamos sólo en segundo lugar a causa de su débil frecuencia de citación (4), aunque este término ha sido citado sólo para estos dos vinos. Observemos de paso que el efecto de la elaboración en barrica, bastante buscado por la profesión, no engendra

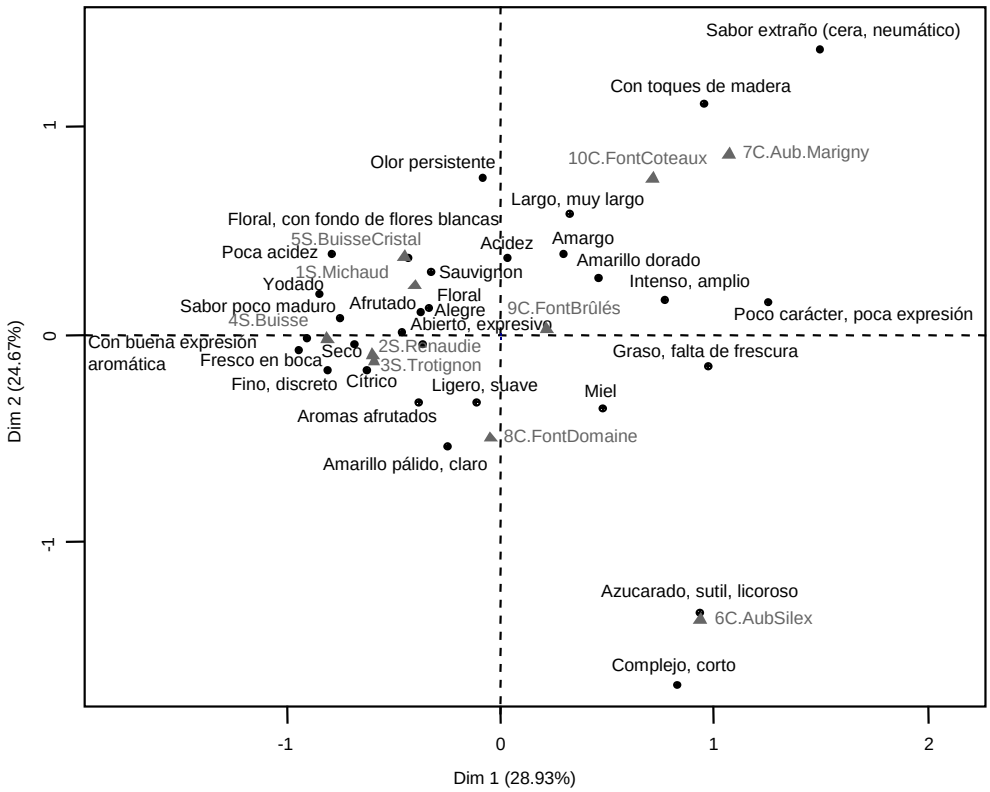

FIGURE 2.14 – Datos de vinos : primer plano factorial del AFC de la tabla 2.11.

solamente características positivas.

– Los cinco Touraine (sauvignon ; 1-5). Para estos vinos, las caracterizaciones son menos evidentes. Citemos *Buena expresión aromática*, *Fresco en boca*, *Cítrico*, *Fino, Discreto* que está de acuerdo con la clásica imagen de sauvignon, según la cual esta cepa nos dirige a vinos frescos muy aromáticos. Añadamos a esto dos características poco citadas : *Falta de frescura* (resp. *Poco carácter*), citada ocho veces (resp. cuatro veces) en total, y nunca para los vinos sauvignon.

Una vez establecidos estos tres grupos, podemos intentar calificar los ejes. El primero separa los vinos sauvignon de los vinos chenin sobre una base de frescura y de expresión aromática. El segundo opone los vinos chenin elaborados en barrica (con sabor a toques de madera) a los vinos que presentan azúcar residual (o sabor azucarado).

Una vez establecidas estas grandes líneas, la frase *Falta de aromas en nariz* utilizada para los vinos 6 y 8 aparece bien en su sitio, es decir, lejos de los vinos aromáticos, ya sean estos aromas de un sauvignon o inducidos por la elaboración en barrica.

Finalmente, este plano propone una imagen de los vinos blancos de Touraine según la cual los vinos sauvignon son homogéneos y los vinos chenin diversos. Lo que, *en definitiva*, podrá

interpretarse de varios modos, no contradictorios :
– Sólo hay una manera de hacer un sauvignon y numerosas maneras de hacer un chenin.
– Los viticultores «trabajan» más su chenin, cepa blanca noble de Touraine, intentando diversas técnicas.
Después de decir esto, salimos de nuestro papel de estadistas, pero quisimos evocar algunos modos de los que el usuario final puede apropiarse los resultados.

Ejes 3 y 4.
Con la intención de ser exhaustivos, podemos consultar rápidamente los ejes siguientes. En este enfoque, las contribuciones son útiles para resumir un eje a lo esencial.
Así, desde el punto de vista de las contribuciones, el eje 3 opone los vinos 1 y 4 y, para las palabras, *Seco* y *Fino* a *Poco maduro*. Encontramos estas asociaciones/oposiciones en los datos. Pero, además de que conciernen efectivos débiles, no nos sugieren ninguna interpretación. Por su parte, el eje 4 pone de relieve el vino 5, que se asocia a las palabras *Alegre* y *Poca acidez*. Aquí todavía esta asociación es (poco) visible en los datos, pero concierne a efectivos débiles y no evoca nada (al contrario, generalmente *Alegre* y *Poca acidez* se oponen).

Conclusiones.
Desde un punto de vista viti–vinícola, este análisis muestra un singularidad poco característica de la cepa chenin : esta cepa conduce, en la práctica, a vinos bastante diferentes que se separan de los vinos sauvignon, sobre todo porque estos últimos son homogéneos y bien caracterizados.
Desde un punto de vista sensorial, es posible obtener una imagen fiable (la fiabilidad es sugerida aquí por las relaciones claras entre las «descripciones» sensoriales y las informaciones «externas» disponibles, que conciernen a la cepa y la elaboración en barrica o no), con un protocolo muy ligero (una sola sesión) comparado con el protocolo habitual.
Desde un punto de vista estadístico, el AFC parece bien adaptado al análisis de matrices dispersas (presentando muchas casillas de efectivos débiles o nulos). Recordemos, no obstante, que eliminamos las palabras de efectivos muy débiles (≤ 3).

2.10 Ejemplo : causas de mortalidad de los franceses

2.10.1 Descripción de los datos y problemática

Disponemos para cada año, desde 1979 hasta 2006, de la tabla de contingencia que cruza, para la población francesa, las causas de defunción y la edad subdividida por grupos de edad. En cada tabla (correspondiente a un año), encontramos en la intersección de la fila i y de la columna j, el número de individuos que pertenece al grupo de edad j y que han muerto (el año estudiado) de la causa i. Para simplificar, principalmente consideramos las dos tablas correspondientes a los años 1979 y 2006, así como la suma de las dos). Consideramos la tabla que cruza los años y los grupos de edad siempre para la población francesa, pero esta vez sobre el conjunto del periodo que va del 1979 al 2006 cualquiera que sea la causa de defunción. El término general de esta última tabla está en la intersección de la fila i y de la columna j, el número de individuos que pertenece al grupo de edad j y que ha fallecido

en el año i (cualquiera que sea la causa). Estas tablas se yuxtaponen en columnas según la figura 2.15. Los datos provienen del Centro de Epidemiología sobre las Causas Médicas de Defunción (Cepidc), que ofrece en su página web un fácil acceso a algunos de sus datos (`http://www.cepidc.vesinet.inserm.fr/`).

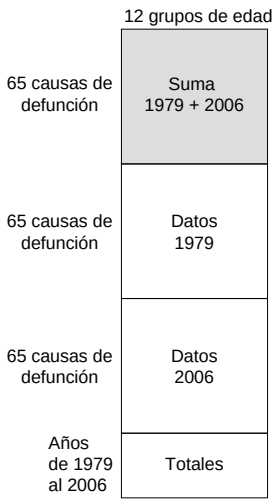

FIGURE 2.15 – Datos de defunción : estructura de la tabla de datos.

El centro de la problemática consiste en el estudio de la relación entre la edad y la causa de la defunción. Inicialmente, la variable edad es cuantitativa : la transformación de esta variable, a través de un recorte en intervalos de su ámbito de variación en una variable cualitativa, permite poner en evidencia, de modo simple y natural, los aspectos no lineales de esta relación. Tal previsión de una relación no lineal resulta de un conocimiento *a priori* del fenómeno estudiado ; se traduce en particular en la definición de los grupos de edad, que se supone que deben reagrupar individuos relativamente homogéneos en relación con las causas de defunción. Así, definimos grupos de diez años sobre lo esencial en el ámbito de variación. Como casi siempre ocurre en un recorte por grupos, las excepciones se encuentran en las extremidades, pero aquí revisten significados muy diferentes : la reagrupación de los individuos de más de 95 años responde a la preocupación de no construir grupos de efectivo demasiado débiles ; al contrario, los más jóvenes son objeto de un recorte más fino porque hay buenas razones para pensar que, por una parte, los recién nacidos (0-1 año), y por otra parte, los niños pequeños (1-5 años), son asociados a causas de defunción que les son específicas.

Introduciendo en activo la tabla que reagrupa las defunciones que se produjeron en 1979 y en 2006, nos libramos de una particularidad eventual de un año y damos por este motivo más importancia a los resultados. De la misma manera, era posible analizar simultáneamente cada año del periodo considerado y no solo los dos años extremos : la elección hecha aquí es de orden pedagógico y pretende no ofrecer datos demasiado voluminosos (conservando una variabilidad *a priori* máxima de las tablas anuales, lo que es esperado con la hipótesis, razonable, de una evolución regular).

Los factores del AFC de la tabla activa proporcionan un marco para el análisis de la relación entre la edad y la causa de defunción, y esto para el periodo estudiado. La introducción de las tablas anuales como filas suplementarias permite analizar, en este marco, la evolución de dicha relación desde el punto de vista de las causas de defunción. Precisemos este punto de vista : a una fila de la tabla activa, *i.e.*, una causa de defunción, le corresponde la distribución de los individuos «que pertenecen» a esta causa según los grupos de edad, lo que llamamos «perfil de edad». El objeto del AFC puede expresarse como la puesta en evidencia de las principales dimensiones de variabilidad de estos perfiles. Esperamos, por ejemplo, una dimensión que opone perfiles «jóvenes» (las causas de defunciones características de los jóvenes) y perfiles «mayores» (las causas de defunciones características de las personas mayores).

Las filas suplementarias también son perfiles de edad ; cada perfil corresponde a una causa de defunción para un año dado. Así, para cada causa de defunción, disponemos de varios perfiles de edad (concretamente, disponemos de varios puntos sobre el gráfico) y será posible analizar la evolución de estos perfiles con observaciones del tipo : tal causa de defunción, muy característica de los jóvenes en 1979, lo es menos en 2006.

2.10.2 Márgenes

Los márgenes indican los grupos de edad más afectados y las causas de defunción más frecuentes. También dan el peso de cada modalidad en el AFC. Los dos márgenes son muy variables (cf. figura 2.16 y figura 2.17). Los resultados numéricos y las figuras se pueden obtener con los comandos siguientes :

```
> library(FactoMineR)
> defuncion <- read.table("http://factominer.free.fr/libra/defuncion.csv",
    header=TRUE, sep=";",row.names=1)
> colnames(defuncion) = c("0-1","1-4","5-14","15-24","25-34","35-44","45-54","55-64","65-74",
    "75-84","85-94","95 y más")
> res.ca=CA(defuncion,row.sup=66:nrow(defuncion), graph=FALSE)
> round(res.ca$call$marge.col,3)
> round(res.ca$call$marge.row[order(res.ca$call$marge.row)],3)
> par(las=1)
> barplot(res.ca$call$marge.col,horiz=TRUE)
> barplot(res.ca$call$marge.row[order(res.ca$call$marge.row)],horiz=TRUE)
> par(las=0)
```

La causa de defunción más frecuente esta relacionada con las enfermedades cerebrovasculares. El grupo de edad para el cual el número de defunciones es el más importante es el grupo 75-84 años. En los grupos de edad superiores (85-94 años y 95 años y más) hay menos defunciones porque el número de personas en estos grupos de edad es muy inferior. Podemos observar que el número de defunciones en el grupo de edad 0-1 año es relativamente importante con respecto a los grupos de edad siguientes. Esto es bastante notable, ya que este grupo de edad concierne a sólo un año mientras que los siguientes conciernen a 4 años y luego a 10 años. El porcentaje de niños de edad de 0-1 año que fallecen es mucho superior al porcentaje de niños de 1 a 4 años o de 5 a 14 años que fallecen.

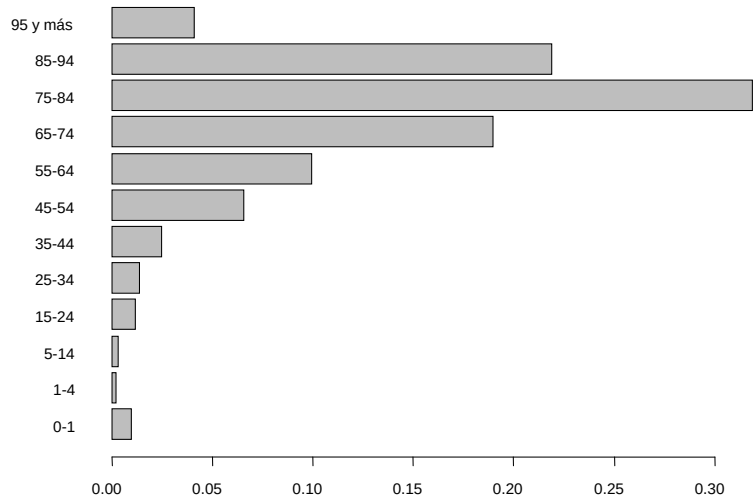

FIGURE 2.16 – Datos de defunciones : margen de los grupos de edad.

2.10.3 Inercias

Aplicado sobre los datos activos, el test de independencia de χ^2 muestra que la relación entre ambas variables es significativo. El χ^2 observado vale 1080254 y la probabilidad crítica asociada muy próxima de 0 (el programa da 0). La significación del test fue la prevista viendo a la vez lo que cada uno puede comprobar según sus conocimientos (aunque sólo fuese por la existencia del término «enfermedad infantil») y del número muy importante de observaciones. Aunque las hipótesis del test no son verificadas (muchas casillas tienen efectivos teóricos inferiores a 5), la probabilidad crítica es tan débil que la significación queda fuera de duda. La inercia total es igual a $\Phi^2 = 1.0213$; la intensidad de la relación, medida por el V de Cramer, es más bien elevada : 0.305 (el valor 1 correspondería a una asociación exclusiva entre cada grupo de edad y un grupo de causas de defunción).

```
> res.ca=CA(defuncion,row.sup=66:nrow(defuncion))
> barplot(res.ca$eig[,1],main="Valores propios", names.arg=1:nrow(res.ca$eig))
       eigenvalue percentage of  cumulative percentage
                          variance             of variance
dim 1      0.5505       53.9002                53.9002
dim 2      0.2570       25.1628                79.0630
dim 3      0.1385       13.5653                92.6283
dim 4      0.0338        3.3141                95.9424
dim 5      0.0199        1.9439                97.8863
dim 6      0.0143        1.4022                99.2885
dim 7      0.0037        0.3665                99.6550
dim 8      0.0017        0.1624                99.8174
dim 9      0.0013        0.1256                99.9430
dim 10     0.0004        0.0439                99.9868
dim 11     0.0001        0.0132               100.0000
```

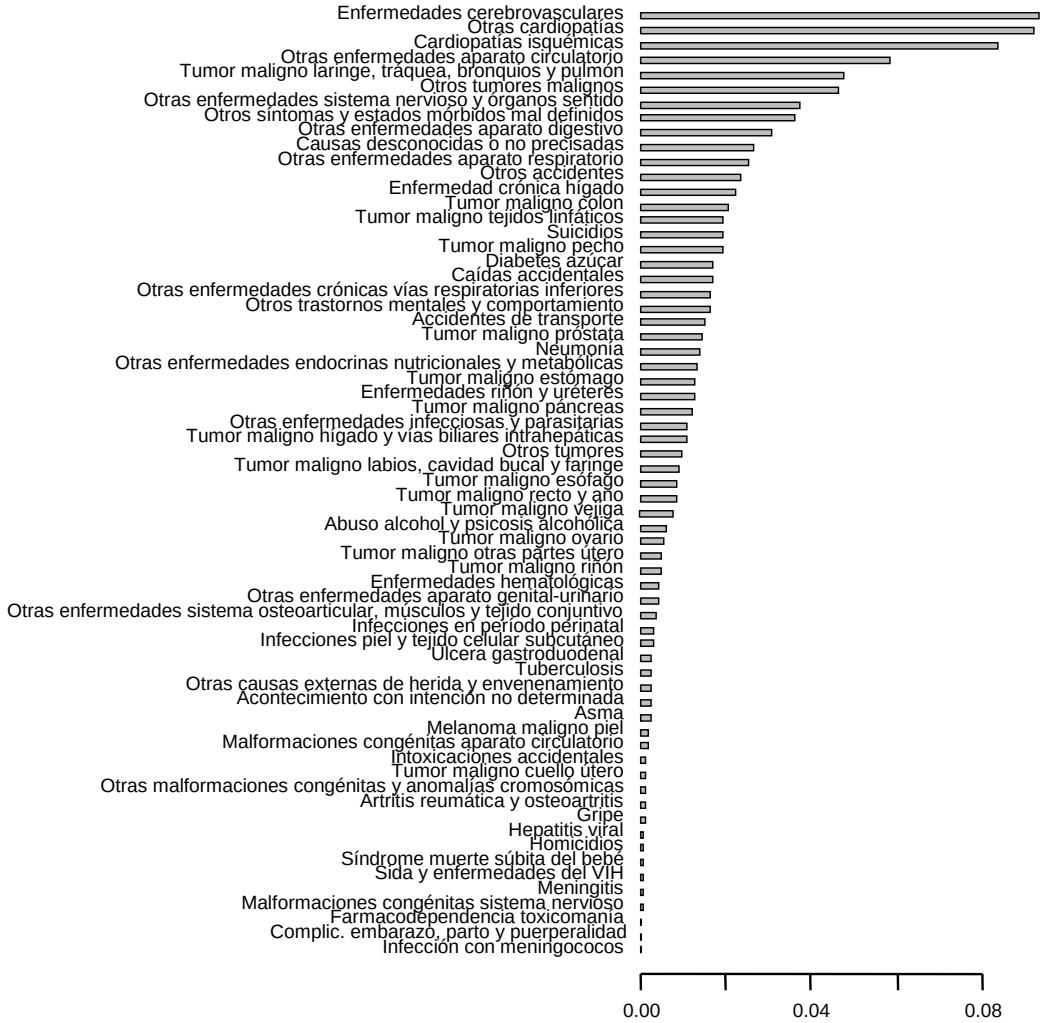

FIGURE 2.17 – Datos de defunciones : margen de las causas de defunción.

La secuencia de valores propios (cf. figura 2.18) distingue tres ejes de inercia. Estos tres ejes representan 92.6% de la inercia total y resume bien el conjunto de la variabilidad (contenida en un espacio de $12 - 1 = 11$ dimensiones). Podremos, pues, concentrarnos en la interpretación de estos tres primeros ejes.

Previamente al AFC, es decir, en los espacios completos, es interesante descomponer esta inercia por fila y por columna. Los objetos `res.ca$row$inertia` y `res.ca$col$inertia` contienen la inercia total descompuesta por fila y por columna. Es interesante expresar estas inercias en porcentaje. Para las columnas, obtenemos :

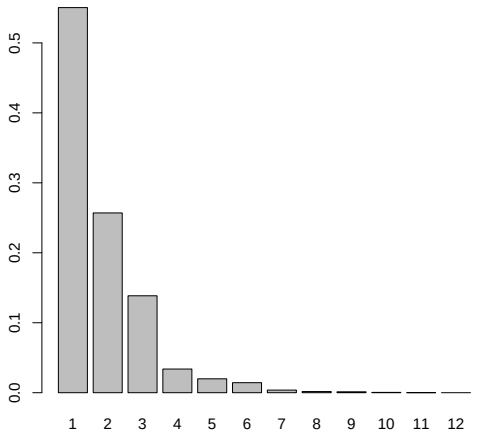

FIGURE 2.18 – Datos de defunciones : diagrama de valores propios.

```
> res.ca$col$inertia/sum(res.ca$col$inertia)
    0-1       1-4      5-14     15-24     25-34      35-44
 0.5262    0.0216    0.0167    0.1222    0.0618     0.0399
  45-54     55-64     65-74     75-84     85-94   95 y más
 0.0456    0.0397    0.0208    0.0239    0.0534     0.0282
```

La inercia del grupo de edad 0-1 año es muy importante, puesto que 52.6% de la inercia total se debe a este grupo de edad. La «mitad» de la relación entre edad y causa de defunción reside en la particularidad del grupo de edad, que tendrá una influencia importante en los resultados del AFC. Después del primer grupo, los dos otros grupos de edad que contribuyen más a la relación son 15-24 años y 25-34 años. Estos grupos de edad tienen un perfil de defunción muy particular y desempeñarán también un papel importante en el AFC.
Para las causas de defunción (hay 65 causas), damos a continuación únicamente las cinco inercias más fuertes (en el espacio completo), clasificadas por orden decreciente :

```
> res.ca$row$inertia[rev(order(res.ca$row$inertia))]/sum(res.ca$row$inertia)
              Infecciones en periodo perinatal  0.3241
                      Accidentes de transporte  0.1370
               Síndrome muerte súbita del bebé  0.0794
Malformaciones congénitas aparato circulatorio  0.0654
                                    Suicidios  0.0500
```

La infección en el periodo perinatal tiene una inercia fuerte comparada con otras causas de defunción (32.41%), mientras que su peso es relativamente débil (su margen vale 0.00336). Esta causa de defunción presenta un perfil de edad muy particular (como su nombre lo indica).
En el marco de una inspección minuciosa de los datos, podemos poner de manifiesto el detalle del cálculo de estas inercias en la forma de una tabla que recapitula para cada fila y para cada columna, el peso (igual al margen expresado en porcentaje), la distancia al origen y la inercia (bruta y en porcentaje). Sea, entonces, para las filas :

```
> bb<-round(cbind.data.frame(res.ca$call$marge.col,
    sqrt(res.ca$col$inertia/res.ca$call$marge.col),
    res.ca$col$inertia,res.ca$col$inertia/sum(res.ca$col$inertia)),4)
> colnames(bb)<-c("Peso","Distancia","Inercia","% del inercia")
           Peso Distancia Inercia % del inercia
0-1      0.0099   7.3829  0.5374        0.5262
1-4      0.0021   3.2375  0.0221        0.0216
5-14     0.0032   2.3039  0.0170        0.0167
15-24    0.0118   3.2583  0.1248        0.1222
25-34    0.0140   2.1275  0.0632        0.0618
35-44    0.0251   1.2736  0.0408        0.0399
45-54    0.0657   0.8413  0.0465        0.0456
55-64    0.0994   0.6390  0.0406        0.0397
65-74    0.1900   0.3342  0.0212        0.0208
75-84    0.3189   0.2765  0.0244        0.0239
85-94    0.2189   0.4993  0.0546        0.0534
95 y más 0.0410   0.8375  0.0288        0.0282
```

Así las cosas, que la fuerte contribución del grupo de edad 15-24 años proviene principalmente de la distancia al origen, de un perfil de causas de defunción muy particular.

2.10.4 Primer eje factorial

El primer eje separa los bebés de 0 a 1 años de otros grupos de edad (cf. figura 2.19). En la figura 2.20 se evidencian las causas de defunciones específicas de este grupo de edad, como las enfermedades infantiles muy particulares que afectan exclusivamente o casi exclusivamente a este grupo de edad (muerte súbita del bebé, infección en el periodo perinatal). El AFC revela un fenómeno específico de una modalidad.

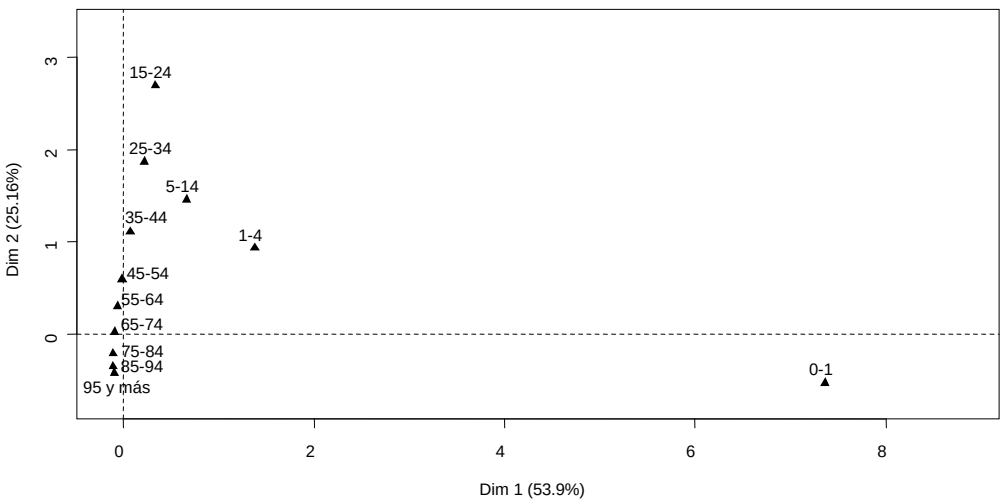

FIGURE 2.19 – Datos de defunciones : representación de los grupos de edad sobre el primer plano.

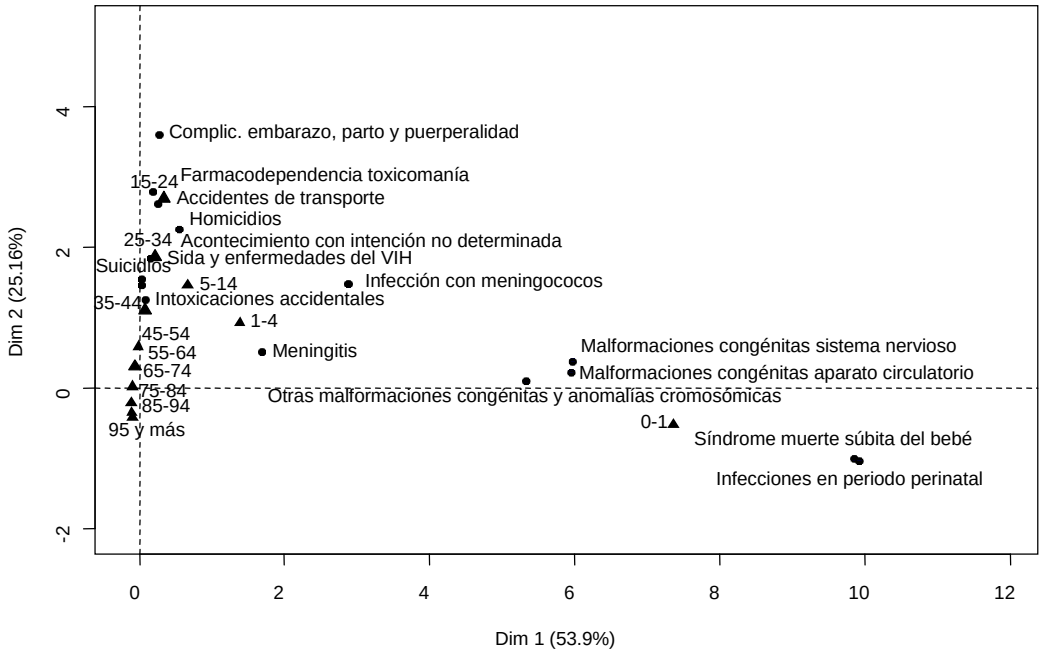

FIGURE 2.20 – Datos de defunciones : representación sobre el primer plano de los grupos de edad y de las causas de las defunciones más específicas.

En AFC, como los elementos (filas o columnas) no tienen el mismo peso, es necesario consultar las contribuciones antes de proponer una interpretación. Los objetos `res$col$contrib` y `res$row$contrib` contienen las contribuciones de las filas y de las columnas para los diferentes ejes. Las contribuciones son expresadas en porcentaje (y algunas veces llamadas contribuciones relativas). Presentamos las contribuciones de las columnas en su orden «natural». Sea :

```
> round(res.ca$col$contrib[,1],3)
     0-1      1-4      5-14     15-24     25-34     35-44
  97.071    0.730    0.256     0.240     0.122     0.024
   45-54     55-64    65-74     75-84     85-94  95 y más
   0.004     0.068    0.306     0.660     0.451     0.069
```

Las contribuciones confirman que el grupo de edad 0-1 año contribuyó él solo (casi) al primer eje (lo que sugiere la figura 2.19) ; a esta edad, las causas de defunción son muy particulares. Este resultado concuerda con la parte de inercia del grupo de edad en el espacio completo (0.5262) comentado anteriormente.
Al ser numerosas las causas de defunción, presentamos las contribuciones clasificadas por orden decreciente, limitándonos a las cinco más grandes (estas cinco causas de defunción contribuyen en un 95.56% a la construcción del primer eje). Sea :

```
> res.ca$row$contrib[rev(order(res.ca$row$contrib[,1])),1]
```

```
                        Infecciones en periodo perinatal  59.101
                         Síndrome muerte súbita del bebé  14.440
          Malformaciones congénitas aparato circulatorio  11.512
Otras malformaciones congénitas y anomalías cromosómicas   7.428
               Malformaciones congénitas sistema nervioso   3.079
```

El eje demuestra las causas de defunciones específicas (casi por definición, como lo muestran los términos «perinatal», «bebé» del grupo de edad 0-1 año. Estas contribuciones completan el gráfico e indican el papel clave que cumplen de las infecciones.

2.10.5 Plano 2-3

El primer eje pone de relieve el rasgo más destacado de la distancia a la independencia : las causas de defunciones específicas del bebé. En este momento, a los facultativos se les presentan dos opciones :

1. La especificidad del grupo de edad 0-1 año está bien establecida, eliminamos este grupo de edad de los datos y rehacemos el análisis. Haciendo esto, modificamos el campo del estudio : nos interesamos por la población de individuos de más de un año. El práctico facultativo está a menudo tentado por esta táctica, cartesiana, que descompone un campo en elementos simples antes de estudiarlo.
2. Continuar la investigación de este AFC, la ortogonalidad de los ejes asegurando que la especificidad de 0-1 año se expresó sobre el eje 1, no «contaminará» los ejes siguientes. Es la táctica que seguimos (y que recomendamos de manera general).

Nos interesamos de ahora en adelante por el plano 2-3 (cf. figura 2.21). La representación de los grupos de edad pone en evidencia un efecto Guttman. El segundo eje (eje de abscisas) opone los grupos de edad más jóvenes a los grupos de edad de las personas de mayor edad, mientras que el tercer eje opone las edades extremas a los grupos de edad media.
A lo largo del eje 2, los grupos de edad «adultos» ($\geq$ 15 años) están situados según su orden «natural». Esto significa que hay una evolución regular del perfil de mortalidad de acuerdo con la edad. De modo más detallado, esta representación sugiere dos observaciones.

1. La disimetría entre ambas curvas de la parábola proviene, en primer lugar, de las diferencias de efectivos ; los grupos más jóvenes son los más raros (no hay por qué lamentarse : ¡se trata del número de defunciones !) y el origen de los ejes que se halla en el centro de gravedad de las nubes (tanto de las filas como de las columnas) se encuentra *de facto* del lado de las modalidades más numerosas (recordemos de paso que la edad media de defunción es de 70.98 años, que está de acuerdo con el gráfico). Otro punto de vista sobre esta diferencia de efectivos entre los grupos jóvenes y los grupos de mayor edad es que estos últimos están «mecánicamente» más próximos al perfil medio, puesto que influyen más sobre él.
2. Sin embargo, el gráfico sugiere claramente que los grupos de personas mayores son más próximos entre ellos que los grupos de jóvenes adultos.

   Podemos verificar este hecho en el espacio completo, calculando las distancias entre los grupos de edad en dicho espacio. Encontramos más abajo el comando que permite obtener esta matriz, así como la matriz misma.

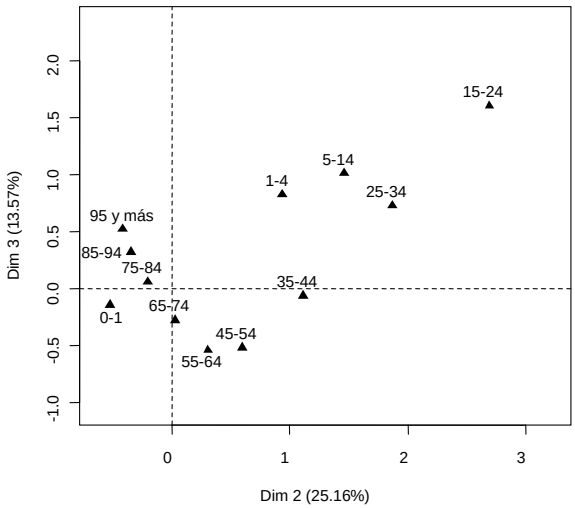

FIGURE 2.21 – Datos de defunciones : representación de los grupos de edad sobre el plano 2-3.

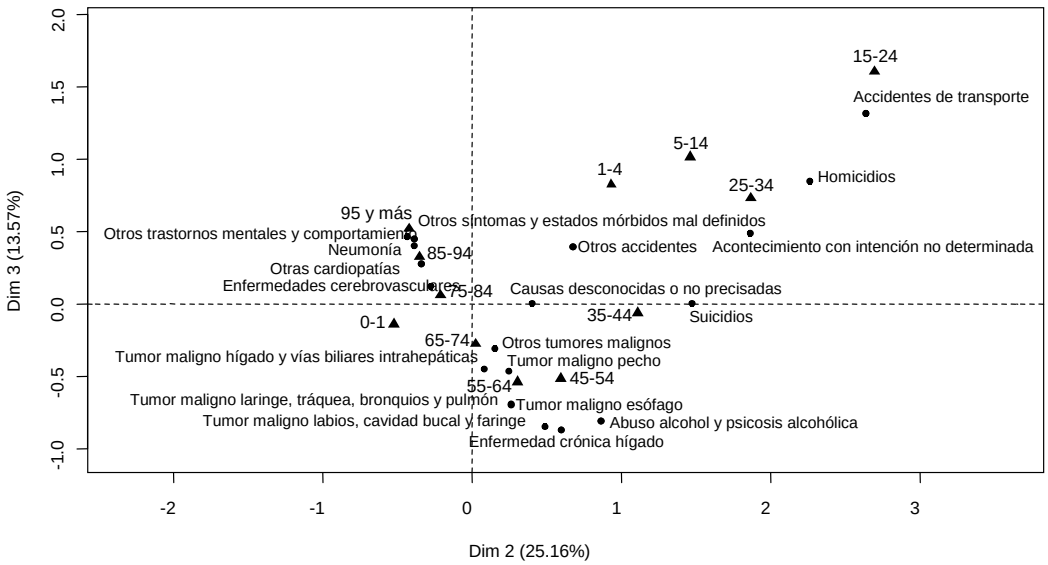

FIGURE 2.22 – Datos de defunciones : plano 2-3 con la representación de los grupos de edad y las causas de defunción que tienen una contribución superior a 1.5% sobre uno de los dos ejes.

```
> res.ca=CA(defuncion,row.sup=c(66:nrow(defuncion)),ncp=Inf)
> round(dist(res.ca$col$coord),3)
```

```
           0-1   1-4  5-14 15-24 25-34 35-44 45-54 55-64 65-74 75-84 85-94
1-4      6.818
5-14     7.221 2.069
15-24    7.965 3.656 2.008
25-34    7.611 3.263 1.874 1.840
35-44    7.495 3.241 2.118 2.694 1.250
45-54    7.480 3.322 2.352 3.166 1.944 0.874
55-64    7.483 3.354 2.428 3.329 2.171 1.175 0.412
65-74    7.480 3.346 2.428 3.374 2.249 1.343 0.767 0.445
75-84    7.480 3.342 2.445 3.410 2.312 1.496 1.073 0.827 0.422
85-94    7.486 3.351 2.485 3.449 2.373 1.619 1.282 1.094 0.754 0.380
95 y más 7.505 3.390 2.562 3.508 2.463 1.766 1.491 1.355 1.098 0.807 0.474
```

Esta matriz muestra primero la gran distancia entre el grupo de edad 0-1 año y todos los demás grupos, de acuerdo con el primer eje. Más allá indica que la distancia entre grupos de edad consecutivos disminuye regularmente entre 1 año y 54 años, y después se estabiliza alrededor de un valor débil. Esto está de acuerdo con nuestra observación sobre el plano 2-3, que concierne a los grupos de edad a partir de 15 años (para 1-4 años y 5-15 años, otros ejes, incluyendo el primero, son necesarios para mostrar esta particularidad).

Las contribuciones en la construcción de los ejes, al igual que las calidades de representación son las siguientes para los grupos de edad :

```
> round(cbind(res.ca$col$contrib[,2:5],res.ca$col$cos2[,2:5]),3)
           Contribuciónes        Calidades de representación (cos2)
          Dim 2  Dim 3  Dim 4  Dim 5   Dim 2 Dim 3 Dim 4 Dim 5
0-1       1.060  0.146  0.015  0.599   0.005 0.000 0.000 0.000
1-4       0.711  1.031  2.089 58.057   0.083 0.065 0.032 0.523
5-14      2.659  2.375  4.075 15.458   0.401 0.193 0.081 0.180
15-24    33.216 21.793 13.518  0.920   0.684 0.242 0.037 0.001
25-34    18.946  5.357  4.207  6.381   0.771 0.118 0.023 0.020
35-44    12.049  0.074 19.113  1.596   0.759 0.003 0.159 0.008
45-54     9.017 12.762 11.460  2.453   0.498 0.380 0.083 0.010
55-64     3.585 20.883  0.002  2.923   0.227 0.713 0.000 0.014
65-74     0.038 10.562 11.896  0.471   0.005 0.690 0.190 0.004
75-84     5.439  0.719  9.790  5.097   0.573 0.041 0.136 0.042
85-94    10.447 16.309  6.272  0.298   0.492 0.414 0.039 0.001
95 y más  2.832  7.988 17.564  5.747   0.253 0.385 0.207 0.040
```

Para las causas de defunción, las contribuciones son clasificadas por orden decreciente y las cinco contribuciones más fuertes son presentadas para los ejes 2 y 3 :

```
> cbind(res.ca$row$contrib[,2],res.ca$row$cos2[,2],res.ca$call$marge.row)
    [rev(order(res.ca$row$contrib[,2])),]
                                          contrib   cos2 eff. en %
Accidentes de transporte                   41.048  0.754     0.015
Suicidios                                  16.250  0.818     0.019
Otras cardiopatías                          4.272  0.546     0.092
Otros accidentes                            4.130  0.592     0.024
Acontecimiento con intención no determinada 3.390  0.886     0.003

> cbind(res.ca$row$contrib[,3],res.ca$row$cos2[,3],res.ca$call$marge.row)
    [rev(order(res.ca$row$contrib[,3])),]
                                                 contrib  cos2 eff. en %
```

```
Accidentes de transporte                              19.199 0.190     0.015
Tumor maligno laringe, tráquea, bronquios y pulmón 16.503 0.818     0.048
Enfermedad crónica hígado                             12.206 0.625     0.022
Otros síntomas y estados mórbidos mal definidos        5.312 0.351     0.036
Otras cardiopatías                                     5.071 0.349     0.092
```

A lo largo del segundo eje, los grupos de edad entre 15 y 44 años tienen una contribución acumulada de 64.211% y la interpretación puede fundarse en ellas. Las contribuciones de estos tres grupos concuerdan con las coordenadas (los tres efectivos marginales son similares) y el grupo de edad 15-24 años es un extremo sobre el cual podemos enfocar la atención para ilustrar el eje.
Los accidentes de transporte contribuyen de modo determinante a este eje (41.05%) y tienen la coordenada más elevada. Esta causa de defunción es característica de los jóvenes adultos (coordenada elevada) ; esto, unido al hecho de que su frecuencia es relativamente elevada (cf. figura 2.17), hace que los jóvenes adultos constituyan una dimensión esencial (la segunda) de la diferencia a la independencia (contribución elevada). Esto puede ilustrarse directamente a partir de los datos (cf. tabla 2.13) : el porcentaje de los jóvenes con defunciones debidas a accidentes de transporte es muy superior al de los jóvenes con defunciones en general.
Lo paralelo puede hacerse con los «Homicidios», cuya coordenada elevada indica una causa característica de los jóvenes adultos. Pero la débil frecuencia de esta causa (cf. figura 2.17) engendra una contribución débil (1.86%) : no es esta causa la que hace característicos a los jóvenes adultos. En la tabla 2.13 se ilustran estos resultados de modo directo a partir de los datos ; con respecto a «Accidentes de transporte», el porcentaje más débil del grupo de edad 15-24 para los homicidios (14.56 en lugar de 28.80) está de acuerdo con la posición menos excéntrica de los «Homicidios».
La causa «Suicidios» es sensiblemente menos característica de los jóvenes adultos (posición más central vinculada al porcentaje entre los jóvenes más débil que para las dos causas precedentes) ; pero su frecuencia relativamente grande (1.93%) hace que esta causa contribuya de modo importante al particularismo de los jóvenes adultos.

	15-24	25-34	35-44	Otros	Totales
Accidentes de transporte	4653	2451	1841	7211	16156
Homicidios	144	199	180	466	989
Suicidios	1431	2693	3280	13003	20407
Otros	6203	9415	21299	983288	1020205
	15-24	25-34	35-44	Otros	Totales
Accidentes de transporte	0.288	0.152	0.114	0.446	1.000
Homicidios	0.146	0.201	0.182	0.471	1.000
Suicidios	0.070	0.132	0.161	0.637	1.000
Otros	0.006	0.009	0.021	0.964	1.000

Tabla 2.13 – Datos de defunciones : extracción de algunos datos que conciernen a causas que caracterizan a los jóvenes adultos ; datos totales y frecuencias.

2.10.6 Proyección de elementos suplementarios

Para analizar los datos de 1979 y de 2006, existen varias posibilidades. Así, podemos realizar por separado el AFC de cada una de las dos tablas, o el AFC de su yuxtaposición. Escogemos

aquí introducir las tablas anuales como filas suplementarias en el AFC de su suma. El interés es : 1) de no multiplicar los análisis ; 2) de analizar simultáneamente las dos tablas en un marco «medio» ya interpretado.
Cada fila suplementaria es asociada a una pareja (causa, año) que llamaremos «causa–anual». En la figura 2.23 se muestra la evolución de algunas causas de defunción. Una causa de defunción, correspondiente a los datos acumulados de 1979 y 2006, se une a los puntos suplementarios de la misma causa de defunción en 1979 y en 2006. Mencionemos una propiedad del AFC cuando se representan varios perfiles y sus sumas : el punto medio (*i.e.*, correspondiente a la suma) está en el baricentro de los puntos que constituyen la suma, es decir, los dos puntos 1979 y 2006. Así, por ejemplo, el punto farmacodependencia toxicomanía 2006 está más próximo al punto medio que el punto farmacodependencia toxicomanía 1979 : hay más defunciones atribuidas a «farmacodependencia y toxicomanía» en 2006 (189) en relación con 1979 (33). *Por el contrario*, las defunciones debidas a la gripe retrocedieron fuertemente (117 en 2006 contra 1062 en 1979).

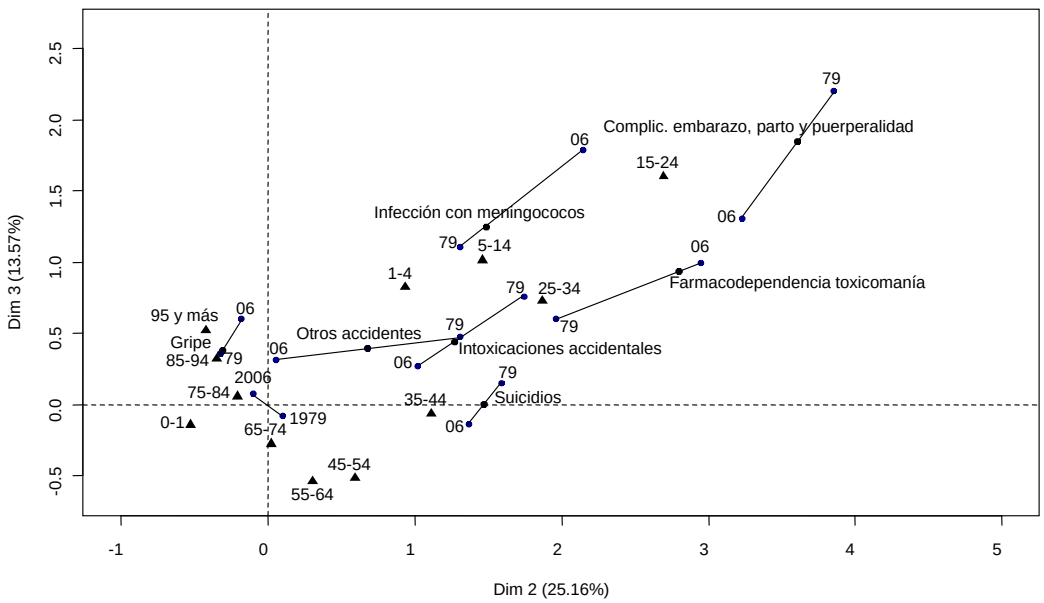

FIGURE 2.23 – Datos de defunciones : proyección de algunos elementos suplementarios.

Consideremos dos causas–anuales relativas a la misma causa. Más allá de su posición con respecto a su punto medio, es sobre todo interesante su distancia sobre el plano, ya que indica una evolución de los perfiles de edad correspondiente. Las causas que aparecen sobre la figura 2.23 han sido seleccionadas precisamente por su gran evolución de perfil de edad entre 1975 y 2006. Comentemos dos ejemplos.
Farmacodependencia toxicomanía. El gráfico sugiere una evolución del perfil de edad hacia los jóvenes. Esto puede verificarse directamente sobre los datos, pero reagrupando –para simplificar– las edades en dos grupos : $\leq$ 44 años y $>$ 44 años (el límite de 44 años es

sugerido por los datos totales). El aumento de esta causa entre los jóvenes es sensible, en frecuencia absoluta (pasamos de 13 a 167) o relativa (el porcentaje de esta causa entre los jóvenes pasa de 39 a 88 %, cf. tabla 2.14).

	Efectivos			Porcentaje		
	15-44	Otros	Totales	15-44	Otros	Totales
79_Farmacodependencia toxicomanía	13	20	33	0.394	0.606	1
06_Farmacodependencia toxicomanía	167	22	189	0.884	0.116	1
Farmacodependencia toxicomanía	180	42	222	0.811	0.189	1

Tabla 2.14 – Datos de defunciones : extracción de algunos datos concernientes a la farmacodependencia toxicomanía.

Como los efectivos son débiles, es prudente verificar la relación entre la edad y el año a partir del test de χ^2 realizado sobre la tabla «Totales», en la figura 2.15. Conduce a un valor de 43.913 (probabilidad crítica : 3.4×10^{-11}) altamente significativo.
Suicidios. El gráfico sugiere una evolución opuesta a la precedente, es decir, una disminución relativa de esta causa entre los jóvenes. Tal evolución parece mucho menos importante que la de la causa precedente, pero como la causa «Suicidio» es muy frecuente, merece atención. La tabla 2.15 cruza la edad (reagrupada en dos grupos, repartidos esta vez alrededor de 34 años, límite sugerido por los datos totales) y el año. La tabla muestra que, en el periodo 1979 – 2006, el porcentaje de jóvenes en las defunciones por suicidios evoluciona del 24.6 al 16.0%. Esta evolución es menos espectacular que la de la toxicomanía (los Φ^2 calculados a partir de las tablas valen 0.198 para el primero y 0.011 para el segundo) pero, a causa de los efectivos más importantes, es todavía más significativa (probabilidad crítica inferior a 2.2×10^{-16}).

	Efectivos			Porcentaje		
	15-34	Otros	Totales	15-34	Otros	Totales
79_Suicidios	2461	7531	9992	0.246	0.754	1.000
06_Suicidios	1663	8752	10415	0.160	0.840	1.000
Suicidios	4124	16283	20407	0.202	0.798	1.000

Tabla 2.15 – Datos de defunciones : extracción de algunos datos relativos a los suicidios.

Además de las «causas-anuales», el perfil de edad media (*i.e.*, cualquiera que sea la causa de defunción) de cada año puede introducirse en suplementario. Para los años 1979 y 2006, estos perfiles son los márgenes filas de las tablas 1979 y 2006. Estas tablas permiten estudiar la evolución, entre los dos años, de la distribución de las defunciones según los grupos de edad. La figura 2.23 muestra que, entre 1979 y 2006, el perfil de edad media se desplaza hacia los grupos de edades elevadas : esto tiene que ver con : 1) el envejecimiento de la población (no olvidemos que nuestros datos son efectivos y no índices) ; 2) el aumento de la esperanza de vida.
Ya indicamos que, de hecho, estos datos están disponibles para cada año comprendido entre 1979 y 2006 ; sólo estos dos años extremos fueron introducidos en el análisis, con el fin de que los resultados no fuesen demasiado complejos. En cambio es posible, sin complicar demasiado el análisis, introducir como filas suplementarias los márgenes filas de cada tabla anual. El gráfico de la figura 2.24 se obtiene rehaciendo el análisis con los mismos elementos

activos pero introduciendo sólo los perfiles anuales de edad en forma suplementaria.
La secuencia de los años presenta una trayectoria asombrosamente regular que muestra una evolución hacia perfiles de edad más avanzada ; lo es hasta tal punto que son las irregularidades en esta las trayectoria que merecen atención. Mencionemos, por ejemplo, el cambio de dirección en 1999 de la trayectoria de los años : aunque en todo rigor la posición sobre el plano de una fila (*i.e.*, un año) deba interpretarse en función del conjunto de columnas (los grupos de edad), la figura 2.24 muestra una trayectoria que, hasta 1999, se aleja de los grupos de edad 45-54 y 55-64 años, y no se aleja más. El examen de la evolución de las defunciones del grupo de edad 45-64 años (cf. figura 2.25) muestra en efecto un decrecimiento hasta 1999-2000, y un ascenso a partir de esta fecha. Propongamos una pista de interpretación : este ascenso tiene que unirse sin duda con la llegada del grupo de edad de la generación (numerosa) de la posguerra.

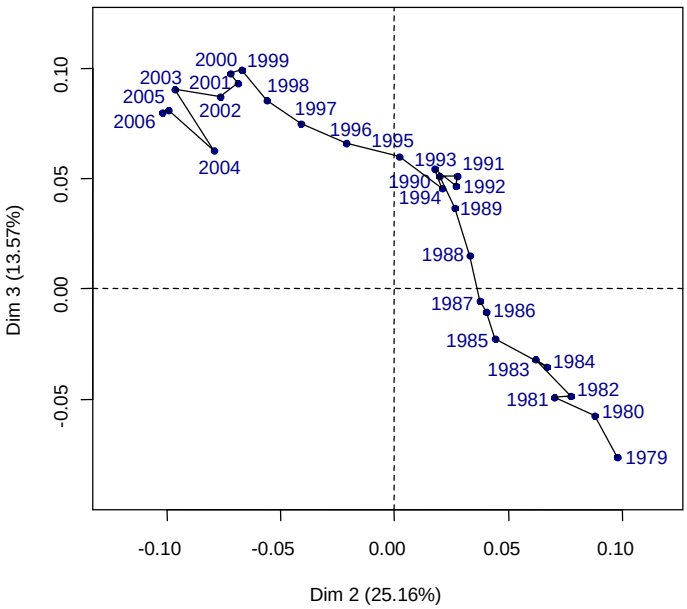

FIGURE 2.24 – Datos de defunciones : evolución del número total de muertes por año y por grupo de edad.

Técnicamente, dos posibilidades son factibles para construir el gráfico 2.24 :

1. Suprimimos los elementos suplementarios que no corresponden a los totales de las defunciones por año entre 1976 y 2006 :

```
> res.ca$row.sup$coord <- res.ca$row.sup$coord[130:157,]
> plot.CA(res.ca,invisible=c("row","col"),axes=2:3)
> points(res.ca$row.sup$coord[,2:3],type="l")
```

2. Rehacer un AFC con el número total de defunciones por año entre 1976 y 2006 como filas suplementarias. Construimos luego un gráfico sin los elementos activos gracias al

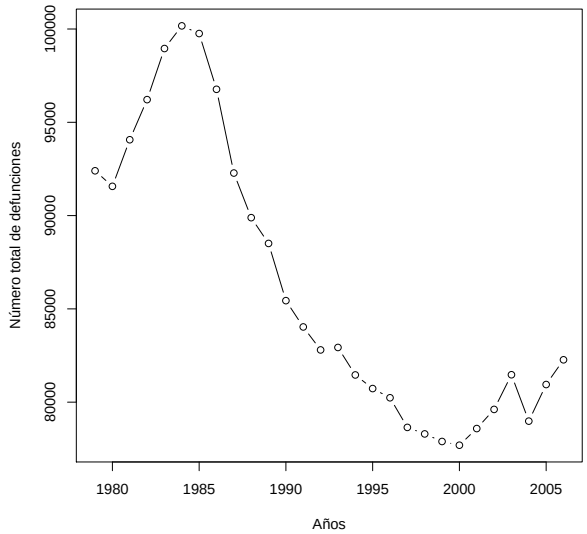

FIGURE 2.25 – Datos de defunciones : evolución de las defunciones del grupo de edad 45-64 años.

argumento `invisible=c('row','col')`. Así, hacemos visibles solamente los elementos suplementarios y unimos los puntos :

```
> tab.evol <- defuncion[-(66:194),]
> res.evol <- CA(tab.evol,row.sup=66:nrow(tab.evol),graph=FALSE)
> plot.CA(res.evol,invisible=c("row","col"),axes=2:3)
> points(res.evol$row.sup$coord[,2:3],type="l")
```

2.10.7 Conclusión

Este ejemplo ilustra bien la naturaleza de las síntesis que el AFC puede ofrecer a partir de una tabla compleja. Los ejes también pueden poner en evidencia un caso particular si este presenta una especificidad (el grupo de edad 0-1 año) con respecto a otros fenómenos más globales.
La elección de los elementos activos y suplementarios es crucial y traduce un objetivo preciso. Varias elecciones son posibles. En una fase de aprendizaje de los métodos o de la apropiación de los datos, el usuario podrá confrontar varios puntos de vista ; en el momento de la comunicación de los resultados, generalmente deberá escoger un solo resultado si no quiere desanimarse. Es imperativo especificar bien el objetivo del análisis escogido. En el ejemplo de las tablas anuales de defunción, confrontemos la metodología escogida (análisis de la tabla suma e introducción de las tablas anuales en suplementarias) con una segunda metodología (análisis de una yuxtaposición –en columna– de las tablas anuales e introducción de la tabla suma en suplementario).
Como ya se indicó, el AFC de la tabla suma estudia la relación entre las variables causas y edad para el periodo considerado. En este marco se examina la evolución anual de esta

relación a través de la de los perfiles de edad de las causas de defunción. En dicho análisis, las evoluciones que no se inscriben en la relación global (*i.e.*, sobre todo el periodo) no pueden aparecer.
La segunda metodología, el AFC de las tablas anuales yuxtapuestas en columnas (cf. figura 2.15), aprehende a la vez la relación global y su evolución mediante la de los perfiles de las causas. Esta presentación de objetivo es más completa que la más formal pero intrínseca en el AFC, fundada sobre la relación entre la edad y la variable que cruza las causas y el tiempo. En este análisis, las dimensiones específicas de la evolución anual (*i.e.* no vinculadas a la relación global) pueden aparecer.
Señalemos de paso que las tablas anuales pueden también estar yuxtapuestas en fila, mostrando la evolución de la relación edades-causas a través de la de los perfiles de defunción de los diferentes grupos de edad. Esto sugiere realizar un tercer AFC (las tablas anuales yuxtapuestas en fila siendo activas), pero también la introducción como columnas suplementarias de los grupos de edad anuales en el AFC de la tabla suma. Este primer análisis se encuentra aquí enriquecido pero conservando su sencillez (debida a la de los elementos activos) y es recomendada, por lo menos en una primera etapa.
Recordemos finalmente que el AFC (como otros métodos del análisis de datos multidimensionales) proporciona una visualización de los datos. Esta visualización es de gran valor y sugiere interpretaciones más allá de los datos pero no las «demuestra». El ejemplo presentado ilustra bien este hecho, poniendo en evidencia evoluciones globales anuales ; pero el AFC no dice nada en cuanto a la parte de la evolución de la pirámide de edades y la parte de la evolución de los índices de mortalidad por causa y por grupo de edad. El AFC permitió responder a la pregunta inicial (qué relación hay entre la edad y la causa de defunción) pero *en definitiva* sugiere nuevas cuestiones. El usuario puede, entonces, tener la impresión de insatisfacción. Pero ¿no es esta la señal de toda investigación ?

Chapitre 3

Análisis de correspondencias múltiple (ACM)

3.1 Datos y notaciones

El Análisis (Factorial) de Correspondencias Múltiples (ACM o AFCM) no es un nuevo método matemático sino una aplicación particular del AFC de tablas que cruzan individuos y sus respuestas con varias variables cualitativas. Se le considera como un método global por sus propiedades específicas y los resultados interesantes que da. El ACM se aplica a tablas que cruzan individuos en fila y variables cualitativas en columnas. La aplicación más frecuente del ACM concierne el tratamiento de encuestas : dentro de este contexto, una pregunta corresponde a una variable y una respuesta posible a esta pregunta corresponde a una modalidad de la variable. A la pregunta «¿A qué categoría socio-profesional pertenece ?» se asocian un conjunto de 8 respuestas posibles (modalidades) que son : agricultor, estudiante, obrero, cuadro medio, cuadro superior, empleado, otro activo, no activo. Para cada una de estas variables, el individuo elige una y una sola modalidad.
Presentaremos en la parte dedicada a los ejemplos aplicaciones del ACM con datos que no proceden de las encuestas.
Consideramos x_{ij} la modalidad tomada por el individuo i para la variable j, i varía de 1 a I y j de 1 a J. Consideramos que la variable cualitativa j tiene K_j modalidades. Ilustramos este capítulo de ACM tratando datos que provienen de una encuesta realizada a 300 consumidores de té. Las diferentes preguntas que se realizaron trataban de qué manera consumían té, la imagen que tenían del producto y su descriptivo socioeconómico. En el análisis realizado después sólo, las variables del comportamiento del consumo son introducidas como activas y las variables de imagen y embalaje como variables suplementarias.
Diecinueve preguntas conciernen el modo en el que consumen el té.
– «¿Qué variedad de té consume la mayoría de las veces (té negro, té verde, té perfumado) ?»
– «¿Cómo consume el té la mayoría de las veces (puro, con limón, con leche, otro) ?»
– «¿Bajo qué forma consume el té (en bolsita, a granel, en bolsita y a granel) ?»
– «¿Le echa azúcar a su té (sí, no) ?»

– «¿Dónde compra el té (en el supermercado, en las tiendas especializadas, los dos) ?»
– «¿Qué tipo de té compra (gama baja, marca de distribuidor (MDD), marca conocida, gama alta, variable, no sabe) ?»
– «¿Con qué frecuencia bebe té (más de 2 veces al día, 1 vez al día, 3 a 6 veces a la semana, 1 a 2 veces a la semana) ?»
– Seis cuestiones conciernen el lugar de degustación del producto : «¿Consume té en casa ?», «¿Consume té en su lugar de trabajo ?», «¿Consume té en un salón de té o una cafetería ?», «¿Consume té en casa de sus amigos ? », «¿Consume té en el restaurante ?», «¿Consume té en un bar ?». Para estas seis preguntas, los consumidores debían responder por sí o no.
– Seis preguntas conciernen al momento de degustación del producto : «¿Consume té en el desayuno ?», «¿Consume té en la merienda ?», «¿Consume té por la tarde ?», «¿Consume té después del almuerzo ?», «¿Consume té después de la cena ?», «¿Consume té a cada momento del día ?». Para estas seis preguntas, los consumidores debían responder por sí o no.

Para la imagen que tienen del producto, doce preguntas han sido realizadas : «¿Asocia el té con la evasión o el exotismo ?», «¿Asocia el té con la espiritualidad ?», «¿El té es bueno para la salud ?», «¿El té es diurético ?», «¿Asocia el té con la convivencia ?», «¿El té impide la absorción de hierro ?», «¿El té es femenino ?», «¿El té es refinado ?», «¿El té adelgaza ?», «¿El té es excitante ?», «¿El té es relajante ?», «¿El té es no tiene ningún efecto sobre la salud ?». Para estas doce preguntas, los consumidores debían responder por sí o no.
Cuatro variables sobre el descriptivo socioeconómico también han sido realizadas : el sexo, la categoría socio-profesional (agricultor, estudiante, obrero, cuadro medio, cuadro superior, empleado, otro activo, no activo), la edad y la práctica regular de un deporte (sí, no).

3.2 Objetivos

Los datos pueden ser estudiados a partir de individuos, variables y modalidades ; esto lleva a hacerse varios tipos de preguntas relativas a estos objetos de naturaleza diferente.

3.2.1 Estudio de individuos

El estudio de individuos consiste en comprender las semejanzas entre individuos desde el punto de vista del conjunto de las variables. En otros términos, a establecer una tipología de los individuos : ¿Cuáles son los individuos más próximos (resp. los más alejados) ? ¿Existen unos grupos de individuos homogéneos desde el punto de vista de sus semejanzas ? En el ejemplo, dos consumidores de té tienen semejanzas ya que respondieron del mismo modo a las preguntas que se les han realizado.
Comparamos los individuos según la presencia-ausencia de las modalidades que escogieron. Sobre esta sola base, la distancia entre dos individuos dependería exclusivamente de sus características y no de las características de otros individuos. Sin embargo es importante tener en consideración las características de otros individuos en el cálculo de esta distancia.
Tomemos cuatro ejemplos para comprender cómo calcular la distancia entre dos individuos :
– si dos individuos toman las mismas modalidades, queremos que la distancia que les separa sea nula ;

– si dos individuos tienen en común un gran número de modalidades, queremos que sean próximos ;
– si dos individuos tienen en común todas las modalidades salvo una que es tomada por uno de los individuos y raramente por el conjunto de otros, nos gustaría a alejarlos con el fin de tener en consideración la especificidad de uno de los dos ;
– si dos individuos tienen en común una modalidad rara, tenemos ganas de acercarlos cualesquiera que sean sus diferencias con el fin de tener en consideración su especificidad común.

Estos diferentes ejemplos permiten comprobar que es necesario comparar los individuos modalidad por modalidad y teniendo en cuenta la rareza o el carácter general de la modalidad.

3.2.2 Estudio de variables y de modalidades

Como para el ACP, procuramos establecer un balance de las relaciones entre variables. Estas relaciones se estudian dos a dos (ver el capítulo del AFC) o globalmente. En este último caso, buscamos variables sintéticas que resumen la información contenida en varias variables. La información llevada por una variable puede ser estudiada a nivel de las modalidades. En ACM, nos centramos esencialmente en el estudio de las modalidades ; la modalidad representa a la vez una variable y un grupo de individuos (el conjunto de los individuos que toman esta modalidad).

Para estudiar las proximidades entre modalidades, la primera etapa es definir una distancia entre estas modalidades. Sean dos modalidades k y k' asimiladas cada una a un grupo de individuos. Un modo de comparar estas dos modalidades es contar los individuos que toman a la vez ambas modalidades : diremos que dos modalidades están más alejadas (distancia grande) cuanto menos individuos tienen en común. Es decir, que el número de individuos que toman o la modalidad k, o la modalidad k' (la una o la otra) es grande ; denotamos este número $I_{k \neq k'}$.

Sin embargo, es importante tener en cuenta el tamaño de cada grupo de individuos en el cálculo de esta distancia. Tomemos un ejemplo con tres modalidades k, k' y k'' constituidas respectivamente por 10, 100 y 100 individuos. Si las modalidades k y k' no tienen ningún individuo en común, $I_{k \neq k'} = 110$. Si las modalidades k' y k'' tienen 45 individuos en común, $I_{k' \neq k''} = 110$. Sin embargo, k y k' tienen 0 % de individuos en común mientras que k' y k'' tienen 45 % de individuos en común. Deseamos que las modalidades k y k' estén más alejados que las modalidades k' y k''. Por esta razón es importante tener en cuenta el efectivo de cada modalidad.

3.3 Definición de una distancia entre individuos y de una distancia entre modalidades

Como vimos en los objetivos, principalmente nos centramos en los individuos y en las modalidades durante el estudio de una tabla de individuos × variables cualitativas. Es lógico construir la tabla disyuntiva completa (TDC) que cruza en filas los individuos y en columnas las modalidades de todas las variables, a partir de la tabla de datos individuos × variables.

El elemento x_{ik} de esta tabla vale 1 si el individuo i posee la modalidad k y si no, vale 0. Esta tabla es de dimensión $I \times K$ (con $K = \sum_{j=1}^{J} K_j$) y está constituida sólo por 0 y por 1.

3.3.1 Distancia entre individuos

Retomando las notaciones de la TDC y los objetivos definidos anteriormente, la distancia entre individuos se calcula sumando las diferencias entre modalidades, es decir $(x_{ik}-x_{i'k})^2$, y ponderando por una función inversamente proporcional a I_k (con I_k el número de individuos que toman la modalidad k). Esta distancia (al cuadrado) se escribe :

$$d^2_{i,i'} = C \sum_{k=1}^{K} \frac{(x_{ik} - x_{i'k})^2}{I_k},$$

con una constante C.

3.3.2 Distancia entre modalidades

La distancia entre dos modalidades k y k' se calcula contando los individuos que toman sea la modalidad k sea la modalidad k' (es decir $I_{k \neq k'}$), y ponderando por una función inversamente proporcional a I_k e $I_{k'}$. Esta distancia puede escribirse :

$$d^2_{k,k'} = C' \frac{I_{k \neq k'}}{I_k I_{k'}},$$

con una constante C'. Ahora bien, según la codificación ($x_{ik} = 0$ o 1), el número de individuos que toma una y una sola de ambas modalidades es igual a $I_{k \neq k'} = \sum_{i=1}^{I} (x_{ik} - x_{ik'})^2$. Podemos pues escribir :

$$d^2_{k,k'} = C' \frac{1}{I_k I_{k'}} \sum_{i=1}^{I} (x_{ik} - x_{ik'})^2.$$

Desarrollando esta ecuación, tenemos :

$$\begin{aligned} d^2_{k,k'} &= C' \frac{1}{I_k I_{k'}} \sum_{i=1}^{I} (x^2_{ik} + x^2_{ik'} - 2 x_{ik} x_{ik'}), \\ &= C' \frac{\sum_{i=1}^{I} x^2_{ik} + \sum_{i=1}^{I} x^2_{ik'} - 2 \sum_{i=1}^{I} x_{ik} x_{ik'}}{I_k I_{k'}}. \end{aligned}$$

Utilizando las propiedades de la codificación ($x_{ik} = 0$ o 1 y entonces $x^2_{ik} = x_{ik}$ y como consecuencia $\sum_i x^2_{ik} = \sum_i x_{ik} = I_k$), podemos escribir :

$$d^2_{k,k'} = C' \left(\frac{1}{I_{k'}} + \frac{1}{I_k} - 2 \frac{\sum_{i=1}^{I} x_{ik} x_{ik'}}{I_k I_{k'}} \right).$$

Ahora bien

$$\frac{1}{I_k} = \frac{I_k}{I_k^2} = \frac{\sum_{i=1}^{I} x_{ik}^2}{I_k^2}.$$

La distancia (al cuadrado) entre dos modalidades se puede escribir :

$$\begin{aligned} d^2_{k,k'} &= C' \left(\frac{\sum_{i=1}^{I} x_{ik'}^2}{I_{k'}^2} + \frac{\sum_{i=1}^{I} x_{ik}^2}{I_k^2} - 2\frac{\sum_{i=1}^{I} x_{ik} x_{ik'}}{I_k I_{k'}} \right), \\ &= C' \left(\sum_{i=1}^{I} \left(\frac{x_{ik'}}{I_{k'}} \right)^2 + \sum_{i=1}^{I} \left(\frac{x_{ik}}{I_k} \right)^2 - 2\sum_{i=1}^{I} \left(\frac{x_{ik}}{I_k} \times \frac{x_{ik'}}{I_{k'}} \right) \right), \\ &= C' \sum_{i=1}^{I} \left(\frac{x_{ik}}{I_k} - \frac{x_{ik'}}{I_{k'}} \right)^2. \end{aligned}$$

3.4 AFC sobre la tabla disyuntiva completa

3.4.1 Relación entre ACM y AFC

Si se considera en las expresiones de más abajo que la constante $C = I/J$, la distancia (al cuadrado) entre dos individuos i y i' se escribe :

$$\begin{aligned} d^2_{i,i'} &= \frac{I}{J} \sum_{k=1}^{K} \frac{1}{I_k} \left(x_{ik} - x_{i'k} \right)^2, \\ &= \sum_{k=1}^{K} \frac{IJ}{I_k} \left(\frac{x_{ik}}{J} - \frac{x_{i'k}}{J} \right)^2, \\ &= \sum_{k=1}^{K} \frac{1}{I_k/(IJ)} \left(\frac{x_{ik}/(IJ)}{1/I} - \frac{x_{i'k}/(IJ)}{1/I} \right)^2. \end{aligned}$$

Con las notaciones de la tabla de contingencia introducidas en AFC aplicadas a la tabla disyuntiva completa, tenemos :

$$\begin{aligned} f_{ik} &= x_{ik}/(IJ), \\ f_{\bullet k} &= \sum_{i=1}^{I} x_{ik}/(IJ) = I_k/(IJ), \\ f_{i\bullet} &= \sum_{k=1}^{K} x_{ik}/(IJ) = 1/I. \end{aligned}$$

Reconocemos entonces la distancia de χ^2 entre los perfiles filas i y i' calculados sobre la tabla disyuntiva completa :

$$d^2_{\chi^2}(\text{perfil fila } i, \text{perfil fila } i') = \sum_{k=1}^{K} \frac{1}{f_{\bullet k}} \left(\frac{f_{ik}}{f_{i\bullet}} - \frac{f_{i'k}}{f_{i'\bullet}} \right)^2 .$$

Además, si consideramos que la constante $C' = I$, la distancia (al cuadrado) entre dos modalidades k y k' se escribe :

$$\begin{aligned} d^2_{k,k'} &= I \sum_{i=1}^{I} \left(\frac{x_{ik}}{I_k} - \frac{x_{ik'}}{I_{k'}} \right)^2 , \\ &= \sum_{i=1}^{I} \frac{1}{1/I} \left(\frac{x_{ik}/(IJ)}{I_k/(IJ)} - \frac{x_{ik'}/(IJ)}{I_{k'}/(IJ)} \right)^2 . \end{aligned}$$

Reconocemos aquí la distancia de χ^2 entre los perfiles columnas k y k' calculados sobre la tabla disyuntiva completa :

$$d^2_{\chi^2}(\text{perfil columna } k, \text{perfil columna } k') = \sum_{i=1}^{I} \frac{1}{f_{i\bullet}} \left(\frac{f_{ik}}{f_{\bullet k}} - \frac{f_{ik'}}{f_{\bullet k'}} \right)^2 .$$

La elección pertinente de las constantes C y C' nos lleva a la distancia de χ^2 sobre perfiles filas y perfiles columnas, lo que nos conduce al Análisis Factorial de Correspondencias. Desde el punto de vista de los cálculos (*i.e.*, del programa), el ACM se apoya pues en un Análisis Factorial de Correspondencias aplicada a la tabla disyuntiva completa.

3.4.2 Nube de individuos

Una vez la nube de individuos construida como en AFC (transformación en perfiles, distancia de χ^2, peso = margen), lo representamos según el procedimiento del Análisis Factorial ya visto en ACP y AFC : maximizar la inercia de la nube de individuos proyectados sobre una serie de ejes ortogonales (ver la puesta en práctica en § 3.6).
El grafo de los individuos para los dos primeros ejes factoriales (17.99 % de inercia explicada) es corresponde la figura 3.1 para el ejemplo del té. Como en la inmensa mayoría de los tratamientos de datos de encuesta, la nube de individuos contiene muchos puntos y únicamente queremos observar una forma particular incluso de grupos de individuos particulares. En el ejemplo, no hay un grupo de individuos particulares : la nube de puntos tiene más bien una forma homogénea.
Para ilustrar la noción de distancia entre individuos, podemos interesarnos por los cuatro individuos siguientes : 200, 262 (a la extremidad negativa del primer eje factorial) y 265, 273 (a la extremidad positiva del primer eje factorial). Los individuos 200 y 262 (resp. 265 y 273) están próximos porque tienen muchas modalidades comunes. Las parejas de individuos 200-262 y 265-273 están alejadas una de la otra (opuestas sobre el primer eje) porque tienen muy pocas modalidades en común (cf. figura 3.2).

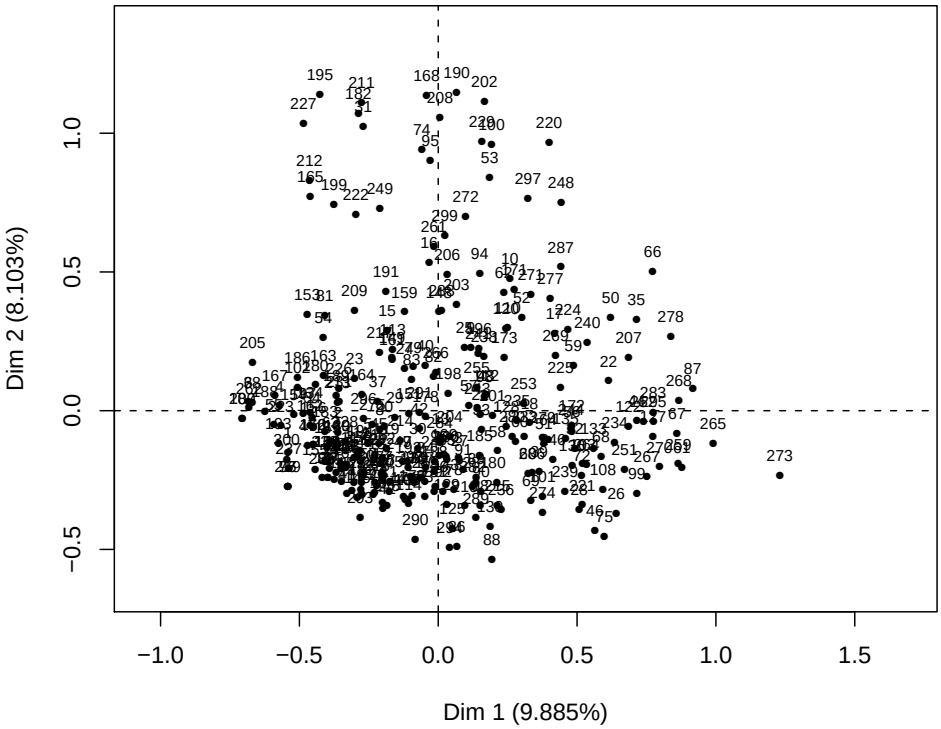

FIGURE 3.1 – Datos té : representación plana de la nube de individuos.

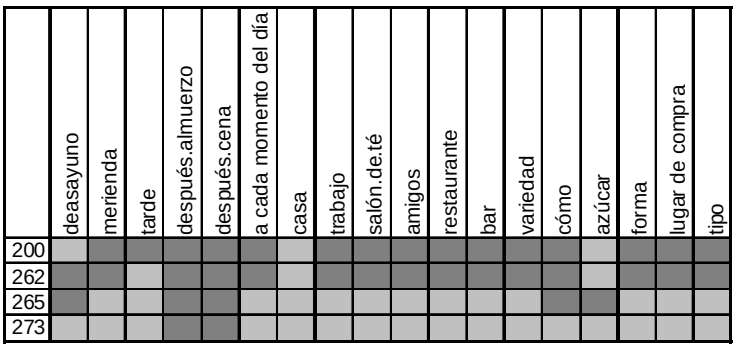

FIGURE 3.2 – Datos té : comparación de individuos 200, 262, 265 y 273 (gris claro = presencia de la modalidad).

Podemos, como en todo análisis factorial, interpretar las dimensiones del ACM a partir de los individuos. Los individuos 265 y 273 son unos apasionados del té que beben té regularmente y en cada ocasión. Los individuos 200 y 262 beben té sólo en casa en el desayuno o por

la tarde. Este procedimiento exploratorio no es cómodo debido al número importante de individuos y se generaliza por el estudio de las modalidades a través de los individuos a los que representan.

3.4.3 Nube de variables

Las variables pueden ser representadas calculando las razones de correlación entre las coordenadas de individuos sobre un eje y cada una de las variables cualitativas. Si la razón de correlación entre la variable j y el eje s es próxima de 1, los individuos que poseen la misma modalidad (para esta variable cualitativa) tienen coordenadas próximas en el eje s. El gráfico de las variables corresponde a la figura 3.3 para el ejemplo del té.

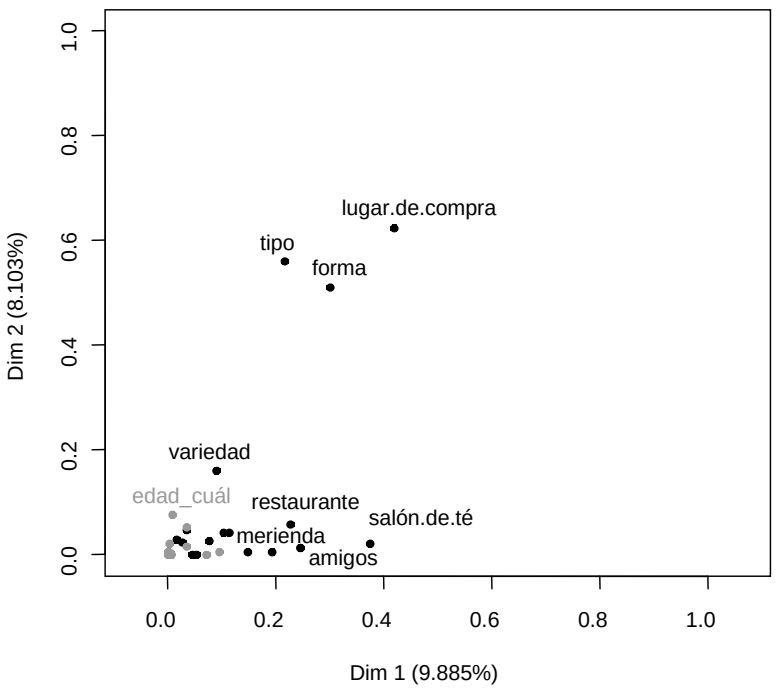

FIGURE 3.3 – Datos té : representación plana de la nube de variables.

Las variables *tipo*, *forma* y *lugar de compra* están muy vinculadas a cada uno de los dos primeros ejes ; pero no sabemos cómo (esto aparece en la representación de las modalidades). También este gráfico es sobre todo valioso en el marco del primer desbrozo frente a un gran número de variables.

3.4.4 Nube de modalidades

Del mismo modo que para las variables cualitativas suplementarias en ACP, podemos representar las modalidades en el baricentro de los individuos que las tomaron. Esta representa-

ción es óptima ya que corresponde, exceptuando un factor multiplicador, a la representación obtenida maximizando la inercia de la nube de las modalidades sobre una serie de ejes ortogonales (cf. § 3.4.5).

El gráfico de las modalidades se encuentra en la figura 3.4 para el ejemplo del té. El primer eje opone las modalidades *salón de té*, *GMS + tienda especializada*, *bolsita + a granel*, *bar*, *restaurante*, *trabajo* a las modalidades *No.amigos*, *No.restaurante*, *No.trabajo*, *No.casa*. Este primer eje opone pues los bebedores de té regulares con los bebedores de té ocasionales. En cuanto al segundo eje, distingue las modalidades *tienda especializada*, *a granel*, *tipo gama alta* y en menor medida *verde* y *después de cenar* del conjunto de otras modalidades.

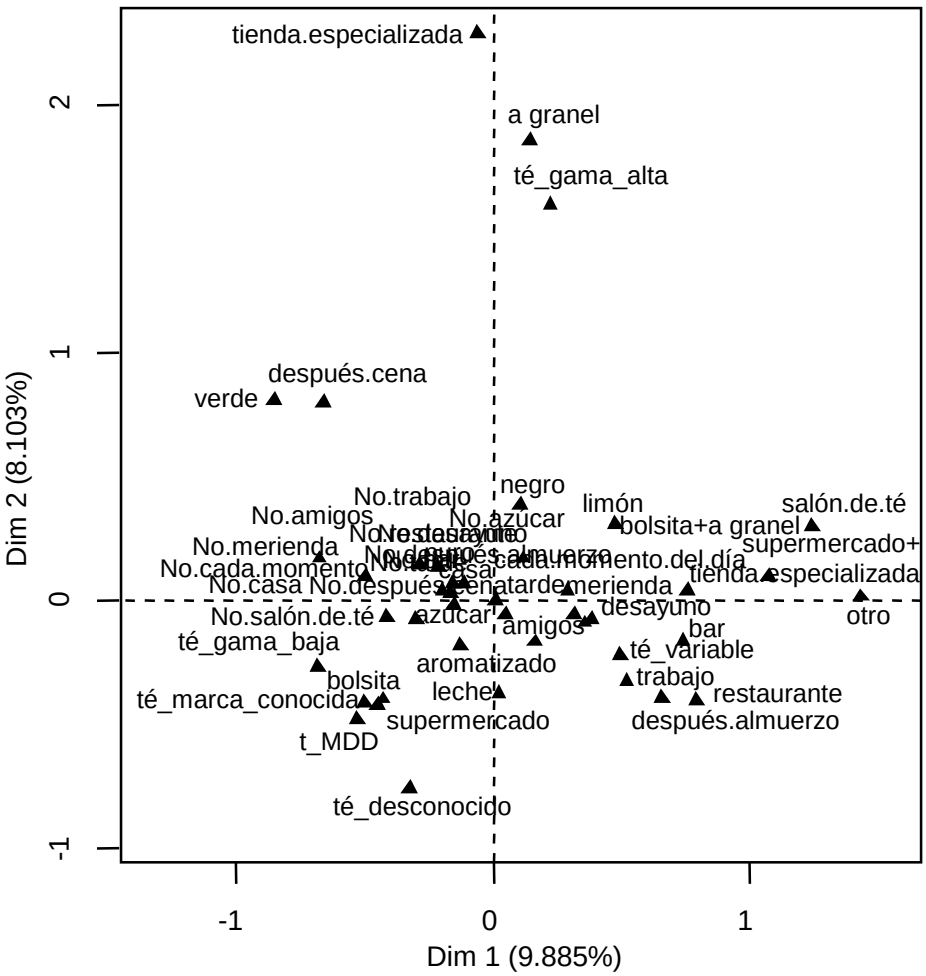

FIGURE 3.4 – Datos té : representación plana de la nube de modalidades.

Observación
El baricentro de todas las modalidades de una misma variable está en el centro de gravedad de la nube de individuos. Se confunde pues con el origen de los ejes.

Calculemos la inercia de una modalidad k comenzando por la distancia (al cuadrado) de k en el centro de gravedad de la nube de modalidades cuyas coordenadas son totalmente iguales a $1/I$ (*i.e.*, vector medio del conjunto de las modalidades) :

$$\begin{aligned} d^2_{k,G_K} &= I \sum_{i=1}^{I} \left(\frac{x_{ik}}{I_k} - \frac{1}{I} \right)^2, \\ &= I \left(\sum_{i=1}^{I} \frac{x_{ik}^2}{I_k^2} - \frac{2}{I} \frac{x_{ik}}{I_k} + \frac{I}{I^2} \right), \\ &= I \left(\frac{1}{I_k} - \frac{2}{I} + \frac{1}{I} \right), \\ &= \frac{I}{I_k} - 1. \end{aligned}$$

Esta distancia es más grande cuanto menos individuos posean la modalidad k. Recordemos que en AFC, el peso de un perfil-columna corresponde a su margen (aquí, $I_k/(IJ)$). Podemos entonces escribir la inercia de la modalidad k :

$$\text{Inercia}(k) = d^2_{k,G_K} \times \frac{I_k}{IJ} = \frac{I_k}{IJ} \left(\frac{I}{I_k} - 1 \right) = \frac{I - I_k}{IJ} = \frac{1}{J} \left(1 - \frac{I_k}{I} \right).$$

Esta fórmula muestra que la inercia de una modalidad es más importante cuanto más esta modalidad es rara : por ejemplo si 1 % de los individuos toman la modalidad k y 50 % de los individuos toma la modalidad k', la inercia asociada a k será dos veces más importante que la asociada a k'. Es entonces frecuente que las principales dimensiones del ACM estén engendradas por algunas modalidades raras presentes en el análisis. Esto es casi sistemático si estas modalidades raras son compartidas por los mismos individuos, lo que es bastante corriente cuando estas modalidades son datos ausentes (por ejemplo, el mismo individuo no respondió a varias preguntas en una encuesta). Los ejes, al estar determinados sólo a partir de algunos individuos, puede ser preferible «eliminar» estas modalidades raras para interesarse por el fenómeno general. Para ello, es posible reagrupar ciertas modalidades, lo que es lógico, concretamente en el caso de modalidades ordenadas (por ejemplo, podemos reagrupar los *60-75 años* con los *más de 75 años*). También es posible repartir de modo aleatorio los individuos asociados a las modalidades raras en otras modalidades (respetando las proporciones asociadas con cada modalidad), método llamado ventilación (cf. § 3.7.1).
La inercia del conjunto de K_j modalidades de una variable j, denominada inercia de la variable j, vale :

$$\text{Inercia}(j) = \sum_{k=1}^{K_j} \frac{1}{J} \left(1 - \frac{I_k}{I} \right).$$

Como $\sum_{k=1}^{K_j} I_k = I$, tenemos :

$$\text{Inercia}(j) = \frac{K_j - 1}{J}.$$

Así, la inercia de una variable depende sólo del número de modalidades que la constituyen : es más grande cuanto más este número es grande. En el ejemplo, la variable *tipo* (que tiene 6 modalidades) tiene 5 veces más inercia que la variable *azucarado* (que tiene 2 modalidades).

Observación
Podemos recomendar construir cuestionarios con un número equilibrado de respuestas por pregunta (con el fin de tener un número equivalente de modalidades por variable) pero esta recomendación no es una exigencia. En efecto, en la práctica, si una variable tiene muchas modalidades, estas modalidades se reparten sobre muchas dimensiones (número de dimensiones igual al número de modalidades menos 1). De este hecho, esta variable no influirá sistemáticamente en la construcción de los ejes factoriales.

Por fin podemos calcular la inercia asociada al conjunto de las modalidades, que corresponde a la inercia de la nube de modalidades (N_K) :

$$\text{Inercia}(N_K) = \sum_{j=1}^{J} \frac{K_j - 1}{J} = \frac{K}{J} - 1.$$

Esta inercia depende sólo de la estructura del cuestionario, más precisamente, del número medio de modalidades por variables. Por ejemplo, si todas las variables tienen el mismo número de modalidades ($\forall j,\ K_j = c$), la inercia de la nube será igual a $c - 1$.

3.4.5 Relaciones de transición

Como para el ACP o el AFC, las relaciones de transición vinculan la nube de los individuos N_I a la nube de las modalidades N_K. En las fórmulas siguientes, obtenidas aplicando en la TDC las relaciones del AFC, $F_s(i)$ (resp. $G_s(k)$) designa la coordenada del individuo i (resp. de la modalidad k) sobre el eje de rango s.

$$\begin{aligned} F_s(i) &= \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \sum_{k=1}^{K_j} \frac{x_{ik}}{J} G_s(k), \\ G_s(k) &= \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{x_{ik}}{I_k} F_s(i). \end{aligned}$$

Sobre el eje de rango s, exceptuando el coeficiente $\frac{1}{\sqrt{\lambda_s}}$, la primera relación expresa que el individuo i está en el centro de gravedad de las modalidades que posee (ya que $x_{ik} = 0$ para las modalidades que no posee).
Sobre el eje de rango s, exceptuando el coeficiente $\frac{1}{\sqrt{\lambda_s}}$, la segunda relación expresa que la modalidad k está en el centro de gravedad de los individuos que la poseen. Como las modalidades corresponden a grupos de individuos, es lógico representarlas sobre el gráfico de los individuos. Las relaciones de transición muestran que dos representaciones son posibles :

dibujar las modalidades en el centro de gravedad de los individuos o dibujar los individuos en el centro de gravedad de las modalidades. Estos dos gráficos son interesantes pero, como en el AFC, no es posible tener estas dos propiedades simultáneamente. Construimos entonces un gráfico compromiso del modo siguiente : construimos el gráfico de los individuos, y situamos las modalidades multiplicando su coordenada sobre el eje de rango s por el coeficiente $\sqrt{\lambda_s}$ (cf. figura 3.5). Así dilatamos la nube de modalidades por un coeficiente diferente en cada eje. Este gráfico evita tener las modalidades concentradas en el centro del gráfico. Anotemos sin embargo que la mayoría de las veces nos interesamos rápidamente por la forma de la nube de individuos (la mayoría de las veces, los individuos son anónimos) antes de interpretar detalladamente la nube de modalidades.

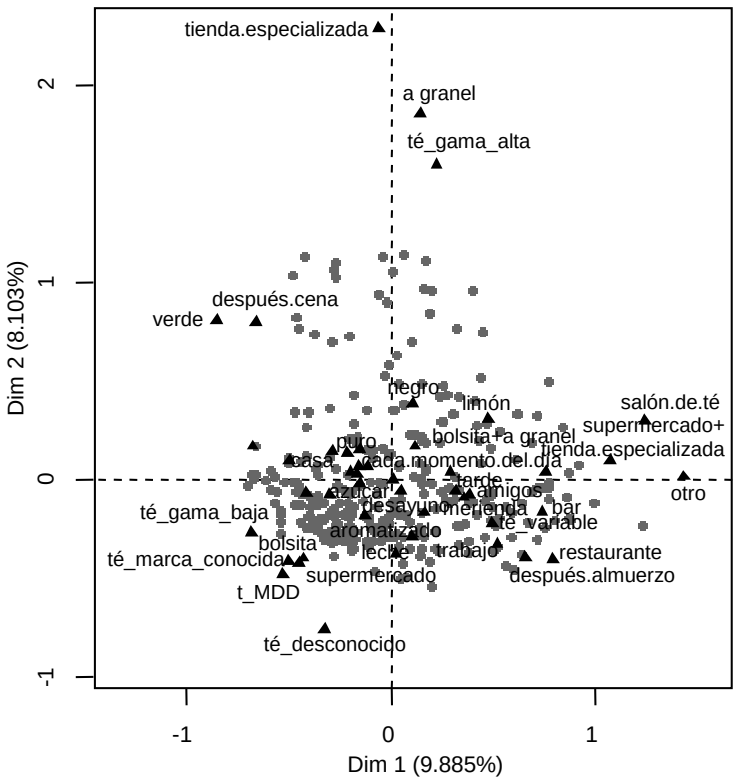

FIGURE 3.5 – Datos té : representación plana de la nube de individuos (puntos grises) y de modalidades.

La segunda relación de transición está en acuerdo con el objetivo fijado en § 3.2.2 : dos modalidades son próximas si las poseen los mismos individuos. También sugiere un modo de interpretar la proximidad entre dos modalidades en el caso de que estas modalidades pertenezcan a la misma variable. En efecto, en este caso, ambas modalidades no pueden ser tomadas por los mismos individuos (elección exclusiva), lo que las aleja una de la otra

por construcción. Sin embargo, como una modalidad representa un grupo de individuos, dos grupos de individuos pueden estar próximos si tienen los mismos perfiles.
En el ejemplo, las modalidades *marca de distribuidor (MDD)* y *marca conocida* asociadas a la pregunta «¿Qué tipo de té compra ? (gama baja, marca de distribuidor, marca conocida, gama alta, variable, no sabe) ?» son representadas una al lado de la otra (cf. figura 3.4). En efecto, estas dos modalidades agrupan consumidores con perfiles similares : tienden a comprar más en supermercado y menos en tienda especializada, a consumir el té exclusivamente en bolsita, a azucararlo (cf. tabla 3.1). La influencia del conjunto de estas variables acerca estas dos modalidades, y es el aspecto multidimensional el que sobresale sobre el aspecto exclusivo de las respuestas aportadas a la misma pregunta.

	marca conocida	MDD	Global
lugar de compra=supermercado	86.32 %	95.24 %	64.00 %
forma=bolsita	73.68 %	76.19 %	56.67 %
azúcar=azucarado	52.63 %	61.90 %	48.33 %
forma=bolsita+a granel	21.05 %	19.05 %	31.33 %
lugar de compra=tienda especializada	2.11 %	0.00 %	10.00 %
lugar de compra=supermercado+tien.espec.	11.58 %	4.76 %	26.00 %

Tabla 3.1 – Datos té : comparación del perfil de consumidores que compran marcas conocidas (resp. MDD) con el perfil medio. 86.32 % (resp. 95.24 %) de consumidores que compran marcas conocidas (resp. MDD) compran en GMS contra 64 % para el conjunto de consumidores.

3.5 Ayuda a la interpretación

3.5.1 Indicadores numéricos

Porcentaje de inercia asociado a un eje

El porcentaje de inercia asociado a un eje se calcula como en todo análisis factorial (cf. § 1.6.1). En ACM, los porcentajes de inercia asociados a los primeros ejes son generalmente mucho más débiles que en ACP. La razón es que en ACP, sólo las relaciones lineales son estudiadas : en última instancia un solo eje puede representar todas las variables si estas últimas están estrechamente correlacionadas entre ellas. En ACM, estudiamos las relaciones mucho más generales y por lo menos $\min(K_j, K_l) - 1$ dimensiones son necesarias para representar la relación entre dos variables que tienen respectivamente K_j y K_l modalidades. Por este hecho, a menudo debemos interpretar un número más grande de dimensiones en ACM que en ACP. En el ejemplo, el 17.99 % de los datos son representados por los dos primeros ejes (9.88 % + 8.10 % = 17.99 %). Podemos ver (cf. tabla 3.2 o figura 3.6) que el decrecimiento de los valores propios es regular. Interpretamos aquí sólo los dos primeros ejes factoriales aunque sea interesante interpretar los ejes siguientes.

Contribución y calidad de representación de un individuo o de una modalidad

El cálculo y la interpretación de las contribuciones y de las calidades de representación de un individuo o de una modalidad son los mismos que para el AFC. Sin embargo, a causa

	Valor propio	Porcentaje de inercia	Porcentaje de inercia acumulada
dim 1	0.15	9.88	9.88
dim 2	0.12	8.10	17.99
dim 3	0.09	6.00	23.99
dim 4	0.08	5.20	29.19
dim 5	0.07	4.92	34.11
dim 6	0.07	4.76	38.87
dim 7	0.07	4.52	43.39
dim 8	0.07	4.36	47.74
dim 9	0.06	4.12	51.87
dim 10	0.06	3.90	55.77

Tabla 3.2 – Datos té : descomposición de la variabilidad para los 10 primeros ejes.

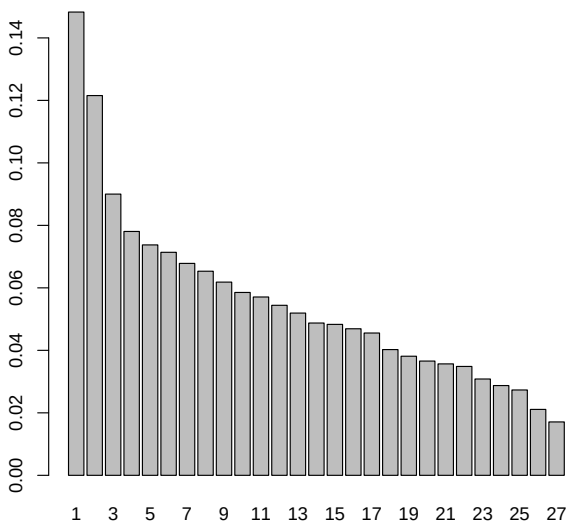

FIGURE 3.6 – Datos té : diagrama de valores propios.

de la dimensionalidad del juego de datos, la calidad de representación sobre un plano es a menudo muy débil comparada con las calidades de representación obtenidas en AFC (o ACP). Para la contribución, la dimensionalidad del juego de datos no se tiene en cuenta ya que la contribución es calculada eje por eje. Anotemos que se puede calcular la contribución de una variable cualitativa a la construcción de un eje sumando las contribuciones de sus modalidades. La contribución (al eje de rango s) de una variable cualitativa dividida por $J\lambda_s$ es igual a la razón de correlación entre el componente principal y la variable cualitativa. En ACP, llamamos componente principal al vector de las coordenadas de los individuos sobre el eje de rango s ; este concepto se transpone directamente en ACM.

3.5.2 Elementos suplementarios

Así como para el ACP, los elementos suplementarios pueden ser individuos, variables cualitativas y/o cuantitativas.
Para un individuo suplementario i' y una modalidad suplementaria k', las fórmulas de transición se escriben :

$$F_s(i') = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \sum_{k=1}^{K_j} \frac{x_{i'k}}{J} G_s(k),$$

$$G_s(k') = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{x_{ik'}}{I_{k'}} F_s(i).$$

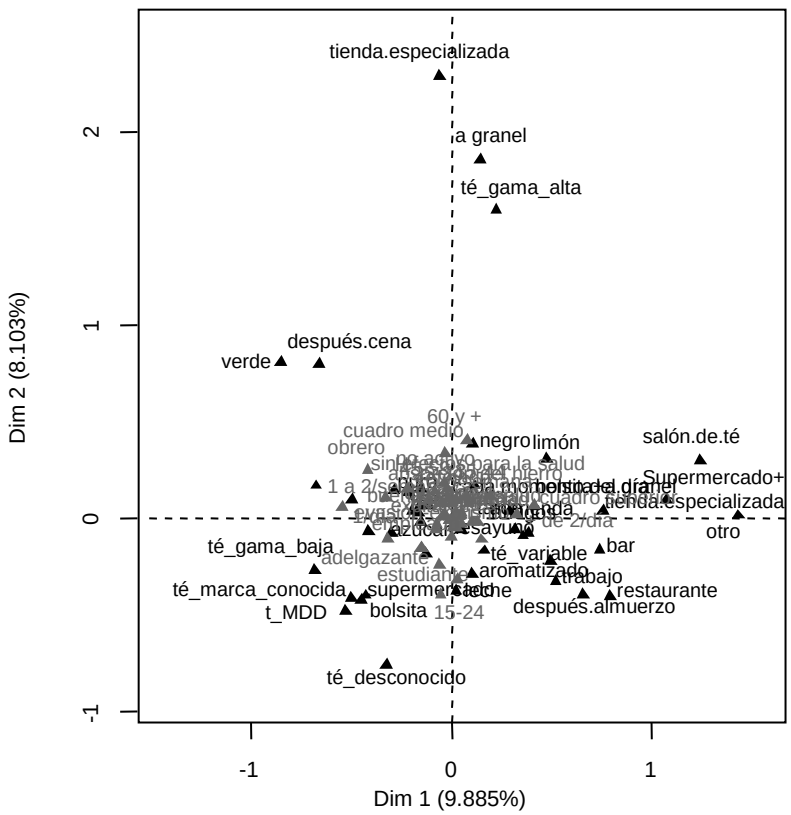

FIGURE 3.7 – Datos té : representación de las modalidades activas y suplementarias.

Estas fórmulas de transición son idénticas a las de los elementos (individuos y modalidades) activos. En el ejemplo (cf. figura 3.7), podemos proyectar las modalidades de las variables

que conciernen a la imagen del té. Estas modalidades están en el centro del gráfico, lo que muestra que será bastante difícil relacionar las variables de comportamiento por una parte y las variables de imagen y del descriptivo socioeconómico por otra parte.
Las variables cuantitativas suplementarias se representan de la misma manera que en ACP (cf. § 1.6.2) : sobre un círculo de correlación con la ayuda de los coeficientes de correlación entre la variable y los factores. En el ejemplo, el círculo de correlación (cf. figura 3.8) permite representar la variable cuantitativa *edad*. Esta variable no está bien representada ; sin embargo, la correlación con el segundo factor (0.204) es significativa debido al número importante de individuos. Los jóvenes tienden más bien a no comprar su té en tienda especializada. ¡Podemos decir también que los adultos compran preferentemente un té de gama alta, a granel, en tiendas especializadas !

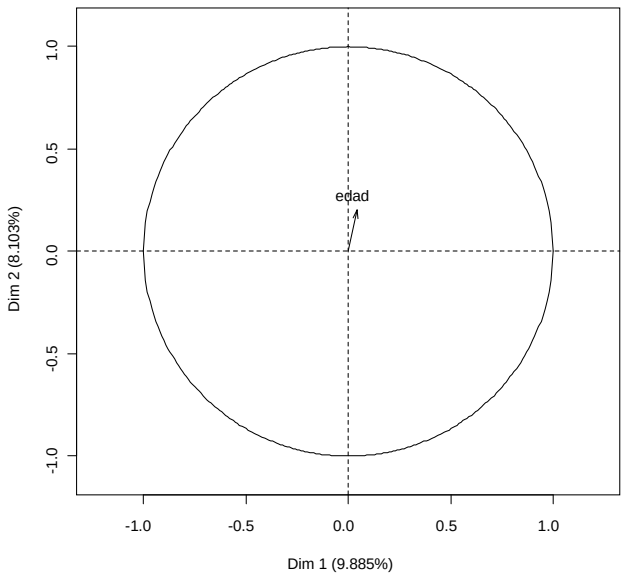

FIGURE 3.8 – Datos té : representación de la variable suplementaria *edad*.

Observación
La variable *edad* ha sido codificada en clase (*15-24 años*, *25-34 años*, *35-44 años*, *45-59 años*, *60 años y más*) y representada como una variable cualitativa suplementaria. Esta codificación puede ser interesante para poner en evidencia relaciones no lineales. Si observamos en detalle las modalidades suplementarias (cf. figura 3.9), podemos ver que las modalidades de la variable *edad* se reparten en su orden natural a lo largo del segundo eje (cf. figura 3.9). Esto está en acuerdo con la correlación positiva entra la variable *edad* y el segundo factor.

3.5.3 Descripción automática de los ejes

De la misma manera que en ACP (cf. § 1.6.3), los ejes proporcionados por el ACM pueden ser descritos de modo automático por el conjunto de las variables, sean cuantitativas o cuali-

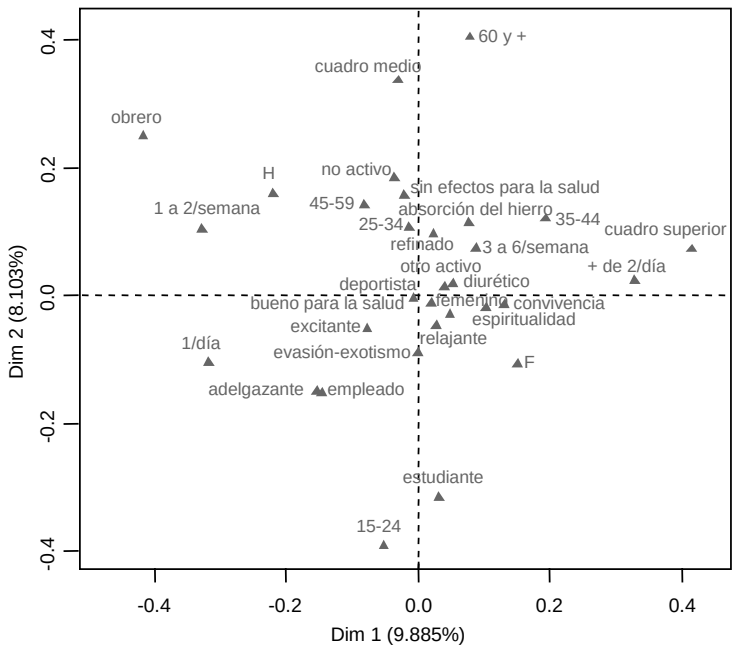

FIGURE 3.9 – Datos té : representación de las modalidades suplementarias.

tativas (en este último caso, utilizamos también las modalidades), activas o suplementarias. En el ejemplo (cf. tabla 3.3), el primer eje es caracterizado por las variables *lugar de compra*, *salón de té*, etc. Observamos que ciertas variables suplementarias están bien vinculadas a este eje (*sexo* y *convivencia*). Como la inmensa mayoría de las variables tienen dos modalidades, la caracterización por las modalidades (cf. tabla 3.4) es similar a la de las variables pero explicita el sentido del eje : por ejemplo, la coordenada de *salón de té* es positiva mientras que la coordenada de *No salón de té* es negativa ; así, los individuos que tienen una coordenada positiva tienden más bien a ir a los salones de té.

3.6 Puesta en práctica con FactoMineR

Mostramos en esta sección cómo efectuar un ACM con FactoMineR y cómo encontrar los resultados obtenidos sobre el juego de datos té.

```
> library(FactoMineR)
> te <- read.table("http://factominer.free.fr/libra/te.csv",header=TRUE,sep=";")
> summary(te)
```

EL ACM es obtenido precisando que aquí la variable 22 es cuantitativa suplementaria y las variables 19 a 21 y 23 a 36 son cualitativas suplementarias :

```
> res.mca<-MCA(te, quanti.sup=22, quali.sup=c(19:21,23:36))
```

```
$`Dim 1`$quali
                           R2  p.value
lugar.de.compra        0.4180 1.26e-35
salón.de.té            0.3720 6.08e-32
forma                  0.2990 1.27e-23
amigos                 0.2430 8.62e-20
restaurante            0.2260 2.32e-18
merienda               0.1920 1.65e-15
tipo                   0.2160 4.05e-14
bar                    0.1470 5.85e-12
trabajo                0.1120 3.00e-09
cómo                   0.1030 4.80e-07
variedad               0.0895 8.97e-07
después.almuerzo       0.0746 1.57e-06
frecuencia             0.0944 1.85e-06
convivencia            0.0713 2.71e-06
tarde                  0.0531 5.59e-05
a.cada.momento.del.día 0.0448 2.22e-04
sexo                   0.0334 1.49e-03
después.cena           0.0329 1.61e-03
deasayuno              0.0254 5.67e-03
azúcar                 0.0153 3.23e-02
```

Tabla 3.3 – Datos té : descripción de la primera dimensión por las variables cualitativas.

```
$`Dim 1`$category
                                  Estimate  p.value
salón de té                         0.2970 6.08e-32
supermercado+tienda.especializada.  0.3390 1.76e-25
amigos                              0.2000 8.62e-20
restaurante                         0.2080 2.32e-18
merienda                            0.1700 1.65e-15
bolsita+a granel                    0.2350 2.72e-12
bar                                 0.1810 5.85e-12
trabajo                             0.1420 3.00e-09
té_variable                         0.2760 1.20e-07
después.almuerzo                    0.1490 1.57e-06
convivencia                         0.1300 2.71e-06
+ de 2/día                          0.1490 1.46e-05
tarde                               0.0935 5.59e-05
otro                                0.3820 6.34e-05
aromatizado                         0.1220 1.18e-04
a cada momento del día              0.0858 2.22e-04
té_gama_alta                        0.1710 7.32e-04
negro                               0.1240 8.90e-04
F                                   0.0716 1.49e-03
No.después.cena                     0.1370 1.61e-03
desayuno                            0.0614 5.67e-03
cuadro superior                     0.1680 6.09e-03
No.azúcar                           0.0476 3.23e-02
```

Tabla 3.4 – Datos té : descripción de la primera dimensión por las modalidades sobreexpresadas.

Este código ejecuta el ACM y ilustra el gráfico de las variables (con las variables activas y suplementarias, cf. figura 3.3), el gráfico de los individuos (con los individuos, las modalidades de las variables activas y suplementarias, cf. figura 3.5) así como el gráfico de las variables cuantitativas suplementarias (cf. figura 3.8). Para dibujar el gráfico con ciertos elementos solamente, utilizamos la función **plot.MCA**. Los códigos siguientes permiten encontrar el gráfico de los individuos (cf. figura 3.1), el de las modalidades activas (cf. figura 3.4), el de la representación superpuesta (cf. figura 3.5), de las modalidades activas y suplementarias (cf. figura 3.7), de las modalidades suplementarias (cf. figura 3.9) :

```
> plot(res.mca,invisible=c("var","quali.sup"),cex=0.7)
> plot(res.mca,invisible=c("ind","quali.sup"))
> plot(res.mca,invisible="quali.sup")
> plot(res.mca,invisible="ind")
> plot(res.mca,invisible=c("ind","var"))
```

La tabla de los valores propios (cf. figura 3.6) :

```
> round(res.mca$eig,2)
> lapply(dimdesc(res.mca),lapply,round,4)
```

El código **dimdesc** proporciona la descripción automática de las dimensiones por las variables cualitativas (cf. tabla 3.3) o las modalidades (cf. tabla 3.4). La función **lapply** permite únicamente poder redondear dentro de una lista (aquí dentro de una lista de listas !) :

```
> lapply(dimdesc(res.mca),lapply,signif,3)
```

Para ir más lejos. Las elipses de confianza pueden ser trazadas alrededor de las modalidades de una variable cualitativa (*i.e.*, alrededor del baricentro de los individuos que poseen la modalidad) según el mismo principio descrito en ACP (cf. p. 48). Estas elipses son adaptadas a representaciones planas y permiten visualizar si dos modalidades son significativamente diferentes o no. Es posible construir elipses de confianza para el conjunto de las modalidades de varias variables cualitativas gracias a la función **plotellipses** (cf. figura 3.10) :

```
> plotellipses(res.mca,keepvar=c("restaurante","lugar.de.compra","relajante",
    "categoria.profesional"))
```

Es también posible construir elipses de confianza para las modalidades de una sola variable cualitativa. Para ello, reutilizamos las instrucciones de la función **plot.PCA** : realizamos un ACP no normado sobre los componentes del ACM (lo que proporciona los mismos resultados que el ACM) y luego reconstruimos un gráfico de ACP con elipses de confianza (cf. figura 3.11) :

```
> res.mca <- MCA(te, quanti.sup=22, quali.sup=c(19:21,23:36), graph=FALSE)
> new.data <- cbind.data.frame(te[,11],res.mca$ind$coord)
> res.pca <- PCA(new.data,quali.sup=1,scale=FALSE,graph=FALSE)
> res.pca$eig[1:5,]=res.mca$eig[1:5,]
> concat.data <- cbind.data.frame(te[,11],res.mca$ind$coord)
> ellipse.coord <- coord.ellipse(concat.data,bary=TRUE)
> plot.PCA(res.pca, habillage=1, ellipse=ellipse.coord, cex=0.8,label="none")
```

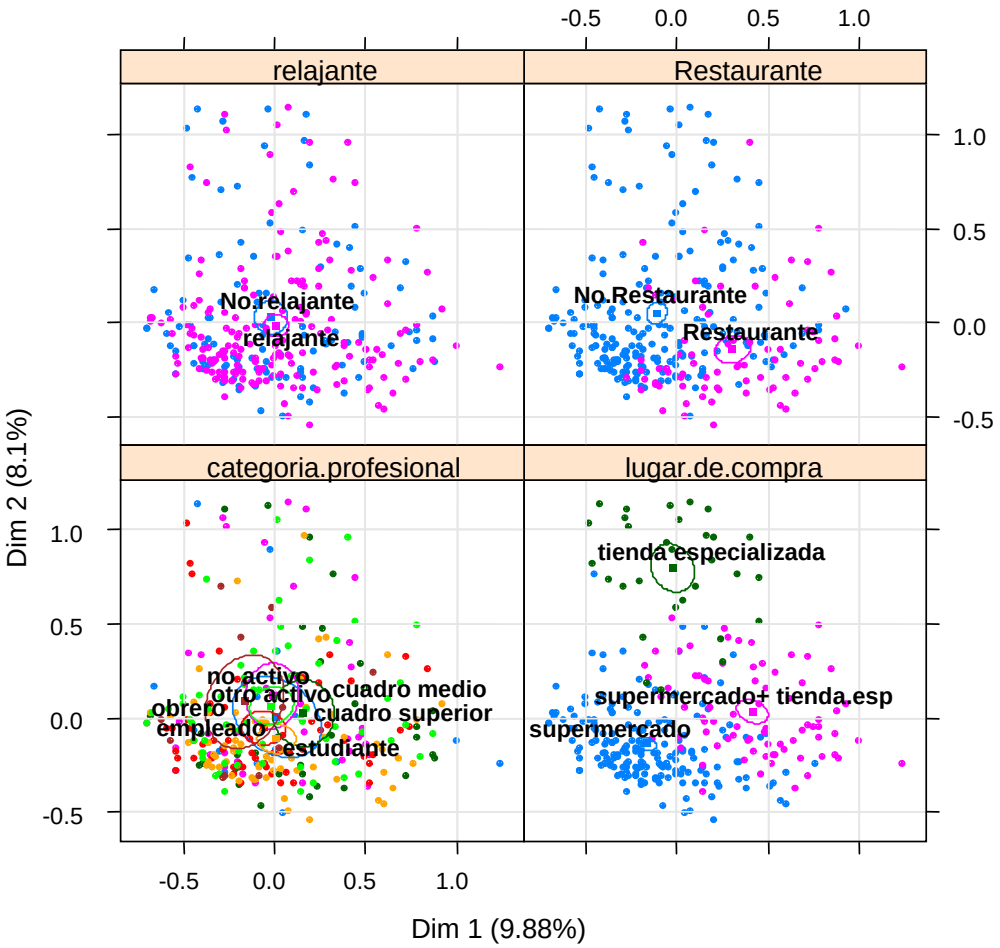

FIGURE 3.10 – Datos té : representación de las elipses de confianza para algunas variables.

3.7 Complementos

3.7.1 Análisis de una encuesta

Construcción del cuestionario - elección de la codificación

Cuando se redacta un cuestionario, es frecuente querer utilizar preguntas llamadas a elección múltiple. Por construcción, estas preguntas pueden dar un número diferente de respuestas por persona interrogada. En el ejemplo, la pregunta inicialmente realizada respecto a la imagen del té era : «¿he aquí una lista de palabras, cuáles son aquellas que usted asocia a la imagen del té ?». La persona interrogada puede entonces escoger entre la lista siguiente : evasión.exotismo, espiritualidad, bueno.para.la.salud, diurético, convivencia, ab-

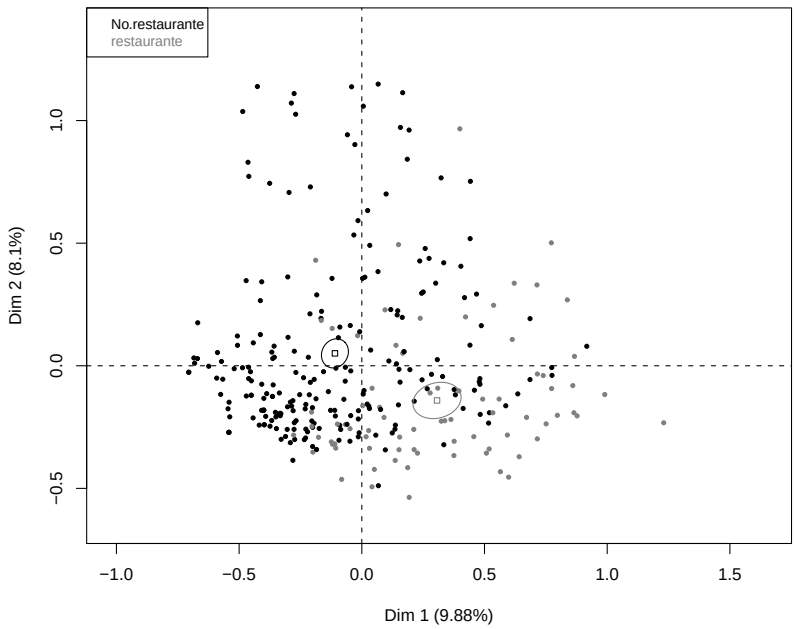

FIGURE 3.11 – Datos té : representación de las elipses de confianza alrededor de las modalidades de la variable *Restaurante*.

sorción.hierro, femenino, refinado, adelgazante, excitante, relajante, sin.efecto.salud. Para utilizar esta información es necesario considerar cada palabra como una pregunta binaria («¿asocia la evasión.exotismo a la imagen del té ? Sí/no»). Esta pregunta a elección múltiple se transforma entonces en 12 preguntas binarias. Desde el punto de vista de la tabla de datos, tendremos pues una columna (una variable) por palabra.

Es también posible explotar la información procedente de una pregunta llamamda abierta, *i.e.* para la cual ninguna respuesta es propuesta. En el ejemplo, los consumidores debían responder a la pregunta siguiente : «¿cuáles son las razones por las cuales usted bebe té ?». En este ejemplo, la explotación de esta información se ha hecho del modo siguiente. Ponemos en una lista el conjunto de las palabras utilizadas y seleccionamos las que tienen una frecuencia bastante elevada. A partir de esta lista creamos tantas preguntas binarias como palabras. Si la palabra es citada por un consumidor afectamos la modalidad *sí* y si no *no*. Esta pregunta abierta entonces es tratada como tantas preguntas binarias que hay palabras escogidas. Esta práctica puede rápidamente llevar a considerar un gran número de variables binarias lo que conlleva por construcción representar a los individuos en espacios cada vez más grandes. Además, la modalidad «sí» de estas variables generalmente tiene una frecuencia débil y su introducción en activo es raramente satisfactoria. En este caso, puede ser interesante reagrupar las palabras con arreglo a su significado (lematización cf. § 2.7). Sin embargo, es preferible no tener demasiadas preguntas abiertas. Cuando queremos considerar una variable cuantitativa como activa, es posible recortar esta variable en clases con el fin de hacerla cualitativa. Varios recortes son posibles : por clases de extensiones planas, por clases de

efectivos planos, por clases con cortes naturales (estos cortes pueden ser hechos visibles con la ayuda de un histograma u obtenidos de modo automático por un método de clasificación, cf. § 4.10).
Cuando las preguntas realizadas son condicionadas por la respuesta a una pregunta precedente j (hablamos entonces de preguntas imbricadas), un modo de analizar los datos es considerar a cada una de las subpoblaciones inducidas por cada modalidad de j. En el ejemplo, la pregunta «¿bebe usted té ?» dividió la población en dos y nos interesamos sólo por los bebedores de té. Si hubiéramos estudiado el conjunto de la población, los primeros ejes del ACM justo habrían puesto en oposición a los bebedores de los no bebedores, en la medida en que los no bebedores responden sistemáticamente no a los diferentes lugares propuestos de consumo, a los diferentes momentos del día, etc. Es pues preferible limitar este tipo de preguntas.
Anotemos por fin que el número de modalidades puede ser diferente de una variable a otra : en efecto, las variables que tienen más modalidades tienen una inercia más importante pero esta inercia se reparte sobre un número más importante de ejes. Así, las primeras dimensiones serán construidas tanto con las variables que tienen pocas modalidades como con las variables que tienen muchas.

La toma en consideración de las modalidades raras

Cuando ciertas variables admiten modalidades con efectivos débiles, varias soluciones son factibles para evitar que estas modalidades no influyan demasiado en el análisis.
- Reagrupación natural de ciertas modalidades. Esta solución es preconizada en el caso de modalidades ordenadas : reagrupación por ejemplo de las modalidades *70-85 años* y *85 años y más* ;
- Ventilación. El principio de la ventilación es afectar de modo aleatorio a los individuos asociados con las modalidades raras en otras modalidades. Para ello, las proporciones de otras modalidades son calculadas y sirven en el momento de la afectación de los individuos que tienen modalidades raras.
- Supresión de los individuos que toman modalidades raras. Esta solución hay que evitarla. Es factible sólo si el conjunto de las modalidades raras está realizada por un pequeño número de individuos (situación que se presenta algunas veces por la ausencia de respuestas).

3.7.2 Descripción de una variable cualitativa y de una subpoblación

El análisis multidimensional es a menudo completado por análisis univariados que permiten caracterizar algunas variables específicas. Nos interesamos entonces por la descripción de una variable cualitativa particular así como por los grupos de individuos definidos por las modalidades de esta variable. Para ello, podemos utilizar variables cuantitativas, variables cualitativas o todavía modalidades de variables cualitativas. En calidad de ejemplo, vamos a describir más particularmente la variable *tipo* (gama baja, gama alta, MDD, etc. ; uno de los intereses de esta variable es que tiene más de dos modalidades). Detallamos más abajo los resultados de la función **catdes** aplicada sobre la variable *tipo* :

```
> catdes(te, num.var = 18)
```

Descripción de una variable cualitativa por una variable cualitativa

Para evaluar la relación entra la variable cualitativa de interés *tipo* y otra variable cualitativa, podemos construir un test de χ^2. Cuanto más pequeña es la probabilidad crítica asociada al test de χ^2, más la hipótesis de independencia está en duda y más la variable cualitativa caracteriza la variable *tipo*. Las variables cualitativas pueden entonces ser clasificadas por probabilidad crítica creciente. En el ejemplo (cf. tabla 3.5), la variable *lugar de compra* es la más vinculada a la variable *tipo*.

```
$test.chi2
                     p.value df
lugar.de.compra 1.1096e-18 10
forma           8.4420e-11 10
salón.de.té     1.6729e-03  5
amigos          4.2716e-02  5
adelgazante     4.3292e-02  5
variedad        4.9635e-02 10
```

Tabla 3.5 – Datos té : descripción de la variable *tipo* por las variables cualitativas.

Descripción de una subpoblación (de una modalidad) por una variable cuantitativa

Para cada modalidad de la variable cualitativa *tipo* y para cada variable cuantitativa (anotada X), calculamos el valor-test definido por :

$$\text{valeur-test} = \frac{\bar{x}_q - \bar{x}}{\sqrt{\frac{s^2}{I_q}\left(\frac{I-I_q}{I-1}\right)}},$$

con $\bar{x}_q$ la media de la variable X para los individuos de la modalidad q, $\bar{x}$ la media de X sobre el conjunto de los individuos, I_q el número de individuos que ha tomado la modalidad q. Este valor permite someter a un test la hipótesis nula siguiente : *los valores de X para los individuos que toman la modalidad q son tirados al azar entre el conjunto de los valores posibles de X*. Consideramos entonces la variable aleatoria $\bar{X}_q$, media de los individuos de la modalidad q. Su esperanza matemática y su varianza son :

$$\mathbb{E}(\bar{X}_q) = \bar{x} \quad \text{y} \quad \mathbb{V}(\bar{X}_q) = \frac{s^2}{I_q} \times \frac{I-I_q}{I-1}.$$

El valor-test puede pues ser visto como una desviación «normalizada» entre la media de los individuos que poseen la modalidad q y la media general. Podemos además asociar una probabilidad al valor-test. Si en la población, la distribución de X es normal, entonces bajo la hipótesis nula la ley de $\bar{X}_q$ es la siguiente :

$$\bar{X}_q = \mathcal{N}\left(\bar{x}, \frac{s}{\sqrt{I_q}}\sqrt{\frac{I-I_q}{I-1}}\right).$$

Si la distribución de X no es normal, podemos a pesar de todo utilizar la distribución normal como ley aproximada de $\bar{X}_q$. Consideramos el valor-test una estadística del test de H_0 («la media de X para la modalidad q es igual a la media general» es decir «la variable X no caracteriza la modalidad q») y así calculamos una probabilidad crítica.

Observación
Caso particular cuando las clases proceden de una clasificación : este test puede aplicarse en todo rigor sólo a las variables suplementarias (*i.e.*, que no sirvieron para construir las clases) pero las calculamos también para las variables activas a título indicativo.

El interés de la probabilidad crítica es que proporciona una indicación en cuanto a la «significación» de una desviación. Es pues posible clasificar las variables cuantitativas por valores-tests decrecientes limitándose a las probabilidades críticas inferiores a 5 %.
En el ejemplo (cf. tabla 3.6), la modalidad *té_gama_alta* es la única que se caracteriza por una variable cuantitativa. Es caracterizada por individuos de más edad que la media ya que el valor-test es positivo. La edad media de los compradores de esta clase es de 43.4 años mientras que en la población total, la edad media es de 37.1 años (media calculada con los individuos que toman la modalidad *té_gama_alta*). Las desviaciones-tipos de la clase (16.95) y de la población (16.8) son también proporcionadas.

```
> catdes(te,num.var=18)
$quanti$té_desconocido
NULL

$quanti$té_gama_alta
      v.test  Mean in category  Overall mean  sd in category  Overall sd  p.value
edad    3.02             43.40         37.05           16.95       16.84  0.00256

$quanti$té_gama_baja
NULL

$quanti$té_marca_conocida
NULL

$quanti$té_MDD
NULL

$quanti$té_variable
NULL
```

Tabla 3.6 – Datos té : descripción de la variable *tipo* por la variable cuantitativa edad.

Descripción de una subpoblación (de una modalidad) por las modalidades de una variable cualitativa

La descripción de una variable cualitativa puede ser afinada gracias al estudio de las relaciones entre modalidades. Caracterizamos entonces cada modalidad de la variable de interés (*tipo*) por las modalidades de las variables cualitativas.
Ilustramos los cálculos a partir de la variable *lugar de compra* y de la tabla cruzada entre las variables *tipo* y *lugar de compra*.

	supermercado	supermercado+ tienda.especializada	tienda.especializada	Suma
té_desconocido	10	1	1	12
té_gama_alta	12	20	21	53
té_gama_baja	6	1	0	7
té_marca_conocida	82	11	2	95
t_MDD	20	1	0	21
t_variable	62	44	6	112
Suma	192	78	30	300

Tabla 3.7 – Datos tá : tabla que cruza las variables *tipo* y *lugar de compra*.

Interesémonos por la modalidad *té_gama_alta* y consideremos la variable *lugar de compra* que posee las modalidades *supermercado*, *supermercado+tienda especializada* y *tienda especializada* y más particularmente la modalidad *tienda especializada*. La pregunta que nos realizamos es : «¿acaso la modalidad *tienda especializada* caracteriza la modalidad *té_gama_alta*?». La idea consiste en comparar la proporción de individuos que compran en *tienda especializada* entre los que compran té de gama alta I_{qt}/I_q al porcentaje global de individuos que compran en tienda especializada I_t/I.

	tienda especializada	otro	Suma
té_gama_alta	$I_{qt} = 21$	32	$I_q = 53$
otro	9	238	247
Suma	$I_t = 30$	270	$I = 300$

Bajo la hipótesis nula de independencia, estas dos proporciones son iguales :

$$\frac{I_{qt}}{I_q} = \frac{I_t}{I}.$$

Se trata una extracción aleatoria sin reposición de I_q individuos (los que toman la modalidad de interés *té_gama_alta*) entre I (la población total) ; nos interesamos por la variable aleatoria X igual al número I_{qt} de apariciones de individuos que tienen el carácter estudiado (compra en tienda especializada) sabiendo que su efectivo en la población es I_t. Bajo la hipótesis nula, la variable aleatoria X sigue una ley hipergeométrica $\mathcal{H}(I, I_t, I_q)$. Podemos entonces calcular la probabilidad de tener un valor todavía más extremo que el observado. Todas las modalidades de las variables cualitativas que caracterizan a cada modalidad de la variable de interés pueden ser clasificadas por probabilidad crítica creciente. La primera fila de la tabla 3.8 indica que 70 % (21/30 cf. tabla 3.7 o el extracto) individuos que compran su té en tienda especializada provienen de la clase *té_gama_alta* ; 39.6 % (21/53 cf. tabla 3.7) individuos de la clase *té_gama_alta* compran su té en tienda especializada ; 10 % (30/300 cf. tabla 3.7) de personas compran su té en tienda especializada. La probabilidad critica del test (1.58×10^{-11}) así como el valor-test (6.64) asociado es proporcionado. El valor-test corresponde aquí al cuantil de la ley normal asociada con la probabilidad crítica ; el signo indica una sobre o una subrepresentación (Lebart *et al.*, 2006).
Las modalidades del conjunto de las variables cualitativas son clasificadas de la que más caracteriza a la que menos caracteriza cuando la modalidad es sobrerepresentada en la clase (*i.e.*, la modalidad de interés) con relación a otras clases (el valor-test es entonces positivo) y

de la que menos caracteriza a la que más caracteriza cuando la modalidad es subrepresentada en la clase (el valor-test entonces es negativo). Lo que caracteriza más a los individuos que compran té de gama alta es que no compran en supermercado (el valor-test de supermercado es negativo y el más elevado en valor absoluto).

```
$category$té_gama_alta
                                                  Cla/Mod Mod/Cla Global  p.value v.test
lugar.de.compra=tienda especializada                70.00    39.6   10.0 3.16e-11   6.64
forma=a granel                                      55.60    37.7   12.0 5.59e-08   5.43
variedad=negro                                      28.40    39.6   24.7 1.15e-02   2.53
edad_cual=60 y +                                    31.60    22.6   12.7 3.76e-02   2.08
sin.efectos.para.la.salud=sin efectos para la salud 27.30    34.0   22.0 3.81e-02   2.07
sin.efectos.para.la.salud=No.sin efectos para la salud 15.00 66.0   78.0 3.81e-02  -2.07
variedad=aromatizado                                12.40    45.3   64.3 2.86e-03  -2.98
edad_cual=15-24                                      7.61    13.2   30.7 2.48e-03  -3.03
forma=bolsita                                        8.24    26.4   56.7 1.90e-06  -4.76
lugar.de.compra=supermercado                         6.25    22.6   64.0 2.62e-11  -6.67
```

Tabla 3.8 – Descripción de las modalidades de la variable *tipo* por las modalidades de las variables cualitativas.

3.7.3 Tabla de Burt

La tabla de Burt es una tabla cuadrada de dimensiones $K \times K$ donde cada fila y cada columna corresponde a K modalidades del conjunto de las variables. En la casilla (k, k') encontramos el número de individuos que toman simultáneamente las modalidades k y k'. Esta tabla es una extensión de la tabla de contingencia en el caso de más de dos variables cualitativas : yuxtapone (en fila y en columna) el conjunto de las tablas de contingencia de las variables tomadas dos a dos. Un análisis de las correspondencias efectuado sobre esta tabla permite obtener una representación de las modalidades. Como esta tabla es simétrica, la representación de la nube de los perfiles-filas es estrictamente idéntica a la de la nube de los perfiles-columnas (guardamos pues sólo la una o la otra de ambas representaciones). Esta representación es muy próxima de la representación de las modalidades proporcionada por el ACM y mostramos que los factores del mismo rango obtenidos por estos dos métodos son colineales. Sin embargo, las inercias asociadas a cada eje difieren de un coeficiente λ_s : si λ_s designa la inercia del eje s para el ACM, la inercia del eje s del AFC de la tabla de Burt será de λ_s^2. Podemos observar que los porcentajes de inercia asociados a los primeros ejes del AFC de la tabla de Burt son superiores a los porcentajes de inercia asociados a los primeros ejes del ACM. En el ejemplo, los porcentajes de inercia asociados a los dos primeros ejes del ACM valen 9.88 % y 8.10 % contra 20.73 % y 14.11 % para el del AFC.

La tabla de Burt presenta pues un interés para el almacenamiento de datos. En efecto, más que conservar la tabla completa de individuos $\times$ variables, basta con construir la tabla de Burt que contiene la misma información a nivel de las asociaciones entre modalidades tomadas dos a dos en la perspectiva de un análisis factorial. Cuando el número de individuos es muy importante, es frecuente que las respuestas individuales no se tengan en cuenta y que las únicas asociaciones entre modalidades sean estudiadas.

3.8 Encuesta sobre la percepción de los OGM

3.8.1 Descripción de los datos y problemática

Los franceses se plantean numerosas cuestiones sobre los organismos genéticamente modificados (OGM) como lo demuestra la existencia de la dirección internet interministerial sobre los OGM (`www.ogm.gouv.fr`). En 2008, estas cuestiones eran tanto más importantes que el ministro de Agricultura, Miguel Barnier, anunciaba el 5 de febrero en el Senado la reanudación «a partir del 2008» de pruebas de culturas OGM en el campo al aire libre, reanudación que va en contra de los compromisos tomados por la decisión política de Medio ambiente. Este mismo año, se realizaba el proceso de los «segadores voluntarios» (movimiento en el que los adherentes se comprometieron a destruir las parcelas de pruebas transgénicas y de culturas de OGM en el campo al aire libre) que habían destruido el año precedente una parcela de prueba de maíz OGM de Monsanto, empresa especializada en las biotecnologías vegetales.

Es en este contexto de tensión, que dos estudiantes del Agrocampus realizaron una encuesta sobre una muestra de 135 personas que pretendía tener una visión del conjunto de sus diferentes posición políticas que concernían los OGM. Se han realizado un conjunto de 21 preguntas cerradas que repartimos en dos grupos.

El primer grupo está compuesto de dieciséis preguntas en relación directa con el informe de los OGM que tienen las personas interrogadas :

- «¿se siente concernido (a) por la polémica sobre los OGM (mucho, más o menos, un poco, en absoluto) ?»
- «¿Cuál es su posición en cuanto a la cultura de los OGM en Francia (muy favorable, favorable, más bien desfavorable, nada favorable) ?»
- «¿Cuál es su posición en cuanto a la incorporación de materia prima OGM en los productos alimenticios destinados a la alimentación humana (muy favorable, favorable, más bien desfavorable, nada favorable) ?»
- «¿Cuál es su posición en cuanto a la incorporación de materia prima OGM en los productos alimenticios destinados a la alimentación animal (muy favorable, favorable, más bien desfavorable, nada favorable) ?»
- «¿ya participó en una manifestación contra los OGM (sí, no) ?»
- «¿Considera que los medios de comunicación comunican bastante sobre el tema (sí, no) ?»
- «¿Usted tiene la iniciativa de informarse sobre el tema (sí, no) ?»
- «¿Piensa que los OGM pueden permitir la reducción del uso de fungicidas (sí, no) ?»
- «¿Piensa que los OGM pueden permitir la reducción de los problemas de hambre en el mundo (sí, no) ?»
- «¿Piensa que los OGM pueden permitir el mejoramiento de las condiciones de vida de los agricultores (sí, no) ?»
- «¿Piensa que los OGM pueden permitir futuros progresos científicos (sí, no) ?»
- «¿Piensa que los OGM representan un peligro eventual para nuestra salud (sí, no) ?»
- «¿Piensa que los OGM representan una amenaza para el medio ambiente) (sí, no) ?»
- «¿Piensa que los OGM representan un riesgo económico para los agricultores (sí, no) ?»
- «¿Piensa que los OGM representan un procedimiento científico inútil (sí, no) ?»
- «¿Piensa que nuestros abuelos tenían una alimentación más sana (sí, no) ?»

Un segundo grupo está compuesto de cinco variables de descriptivo socioeconómico en un sentido amplio :

- Sexo (masculino, femenino)
- Categoría socio-profesional (agricultor, estudiante, obrero, cuadro, funcionario público, liberal, técnico, comerciante, otro activo, no activo, jubilado)
- Edad (-25 años, 25-40 años, 40-60 años, +60 años)
- «¿Realiza usted estudios o una profesión en relación con la agricultura, la industria agroalimentaria o la industria farmacéutica (sí, no) ?»
- «¿Con qué partido político se identifica usted más (extrema izquierda, verdes, PS, centro, UMP, FN) ?»

A través de este cuestionario, procuramos obtener una tipología de las personas interrogadas en función de su relación a los OGM por una parte ; por otra parte procuramos verificar si esta tipología tiene alguna relación con las variables de descriptivo socioeconómico. La pregunta «¿Realiza usted estudios o una profesión en contacto con la agricultura, la industria agroalimentaria o la industria farmacéutica ?» tiene su importancia en la interpretación de los resultados ya que es normal suponer que las personas que contestaron afirmativamente a esta pregunta son susceptibles de tener un conocimiento científico sobre los OGM superior a otras personas.

La primera toma de contacto con los datos consiste en realizar la tabla de frecuencias sobre el conjunto de las preguntas con el fin de ver cómo se reparten las respuestas a cada una de las preguntas. Para ello, utilizamos la línea de código siguiente que permite obtener los efectivos asociados a cada una de las modalidades de las 16 primeras variables :

```
> library(FactoMineR)
> ogm <- read.table("http://factominer.free.fr/libra/ogm.csv",header=TRUE,sep=";")
> summary(ogm[,1:16])
    Concernido          Posición.cultura             Posición.Al.H             Posición.Al.A
En absoluto:15  Favorable            :45  Favorable            :37  Favorable            :44
Más o menos:53  Más bien desfavorable:54  Más bien desfavorable:47  Más bien desfavorable:39
Mucho      :36  Muy favorable        : 3  Muy favorable        : 1  Muy favorable        : 8
Un poco    :31  Nada favorable       :33  Nada favorable       :50  Nada favorable       :44

 Manif  Media.activa   Info.activa  Productos.fitosanitarios  Hambre  Mejoramiento.Agr
No:122         No:78         No:82                     No:56   No:67             No:93
Sí: 13         Sí:57         Sí:53                     Sí:79   Sí:68             Sí:42

Futur.progreso  Peligro   Amenaza  Riesgo.eco  Proced.inútil  Abuelos
         No:54    No:39     No:48       No:67          No:123    No:49
         Sí:81    Sí:96     Sí:87       Sí:68          Sí: 12    Sí:86
```

El resumen del juego de datos activo incita a reagrupar algunas modalidades entre ellas a causa de su débil efectivo (cf. § 3.7.1). A la pregunta «¿Cuál es su posición en cuanto a la incorporación de materia prima OGM en los productos alimenticios destinados a la alimentación humana ?», por ejemplo, una sola persona declaró estar muy favorable. Estamos pues en presencia de una modalidad de débil efectivo y es entonces aconsejable reagruparla con otra. En este caso particular, la reagrupación se hace relativamente fácilmente en la medida en que la variable concernida está constituida por modalidades ordenadas : no traicionamos completamente el pensamiento de una persona reemplazando *Muy favorable* por *Favorable*.

Para ello, utilizamos la línea siguiente de código que permite reagrupar las modalidades *Muy favorable* y *Favorable* en una única bajo la modalidad *Favorable* :

```
> levels(ogm$Posición.Al.H)[3] <- levels(ogm$Posición.Al.H)[1]
```

Ocurre lo mismo para la pregunta «¿Cuál es su posición en cuanto a la cultura de OGM en Francia ?», reagrupamos las modalidades *Muy favorable* y *Favorable* en una única bajo la modalidad *Favorable* :

```
> levels(ogm$Posición.Cultura) <- c("Favorable",
    "Más bien desfavorable","Favorable","Nada favorable")
```

El resumen de juego de datos una vez las modalidades recodificadas es el siguiente :

```
> summary(ogm[,1:16])
    Concernido          Posición.cultura               Posición.Al.H
En absoluto:15   Favorable            :78      Favorable            :87
Más o menos:53   Nada favorable       :54      Más bien desfavorable:47
Mucho      :36   Más bien desfavorable: 3      Muy favorable        : 1
Un poco    :31

           Posición.Al.A   Manif  Media.activa  Info.activa  Productos.fitosanitarios
Favorable             :44  No:122         No:78        No:82                      No:56
Más bien desfavorable:39  Sí: 13         Sí:57        Sí:53                      Sí:79
Muy favorable         : 8
Nada favorable        :44

Hambre  Mejoramiento.Agr  Futur.progreso  Peligro  Amenaza  Riesgo.eco  Proced.inútil  Abuelos
 No:67             No:93           No:54    No:39    No:48       No:67         No:123    No:49
 Sí:68             Sí:42           Sí:81    Sí:96    Sí:87       Sí:68         Sí: 12    Sí:86
```

De modo general, para una pregunta dada, cuando las modalidades son cualesquiera (cuando no existe relación de orden entre ellas por ejemplo), podemos reemplazar la modalidad raramente utilizada por otra escogida aleatoriamente entre las restantes que han sido utilizadas más frecuentemente.
La línea de código siguiente proporciona la tabla de frecuencias de las variables de descriptivo socioeconómico :

```
> summary(ogm[,17:21], maxsum=Inf)
Sexo         Edad                    CSP     Relación          Parti.Político
F:71   [26; 40]:24   Comerciante        : 3   No:79    Centro           :32
H:64   [41; 60]:24   Cuadro             :17   Sí:56    Extrema izquierda: 9
       < 25    :73   Estudiante         :69             PS               :47
       > 60    :14   Funcionario público: 9             UMP              :40
                     Inactivo           : 4             Verdes           : 7
                     Jubilado           :14
                     Liberal            : 3
                     Obrero             : 1
                     Otro               : 9
                     Técnico            : 6
```

En la siguiente parte, vamos a ver que no es necesario hacer reagrupaciones de modalidades para estas últimas.

3.8.2 Elección del análisis y puesta en práctica

Según los objetivos anunciados en la parte precedente, es lógico describir los individuos en función de sus respuestas a las 16 primeras preguntas, las relativas a sus posiciones políticas con relación a los OGM. Las 16 primeras preguntas se considerarán como variables activas, las 5 preguntas siguientes se considerarán como variables ilustrativas. Las variables ilustrativas no participan en la construcción de los ejes factoriales por definición, igualmente para las modalidades que le son asociadas y no es pues necesario proceder a reagrupaciones para las modalidades de débil efectivo en este caso.
La línea de código siguiente permite de realizar tal análisis :

```
> res <- MCA(ogm, ncp=5, quali.sup=17:21, graph = FALSE)
> res
**Results of the Multiple Correspondence Analysis (MCA)**
The analysis was performed on 135 individuals, described by 21 variables
*The results are available in the following objects:

   name                 description
1  "$eig"               "eigenvalues"
2  "$var"               "results for the variables"
3  "$var$coord"         "coord. of the categories"
4  "$var$cos2"          "cos2 for the categories"
5  "$var$contrib"       "contributions of the categories"
6  "$var$v.test"        "v-test for the categories"
7  "$ind"               "results for the individuals"
8  "$ind$coord"         "coord. for the individuals"
9  "$ind$cos2"          "cos2 for the individuals"
10 "$ind$contrib"       "contributions of the individuals"
11 "$quali.sup"         "results for the supplementary qualitative variables"
12 "$quali.sup$coord"   "coord. for the supplementary categories"
13 "$quali.sup$cos2"    "cos2 for the supplementary categories"
14 "$quali.sup$v.test"  "v-test for the supplementary categories"
15 "$call"              "intermediate results"
16 "$call$marge.col"    "weights of columns"
17 "$call$marge.li"     "weights of rows"
```

Anotemos que es también posible reagrupar las modalidades de modo automático a través de la ventilación evocada en § 3.7.1. Esta repartición es entonces aleatoria o tiene en cuenta la relación de orden entre modalidades en el seno de una variable cuando la variable es ordenada (`ordered` en R). Para ejecutar reagrupaciones de modo automático, podemos aplicar la línea de código siguiente :

```
> res <- MCA(ogm, ncp=5, quali.sup=17:21, graph=FALSE, level.ventil=0.05)
```

donde `level.ventil` designa el límite inferior por debajo del cual una modaidad es ventilada. En el ejemplo, si una modalidad es tomada por menos de 5 % de individuos, estos últimos son repartidos en el seno de las modalidades restantes.

3.8.3 Análisis del primer plano

Con el fin de visualizar la nube de los individuos, ejecutamos la línea de código siguiente :

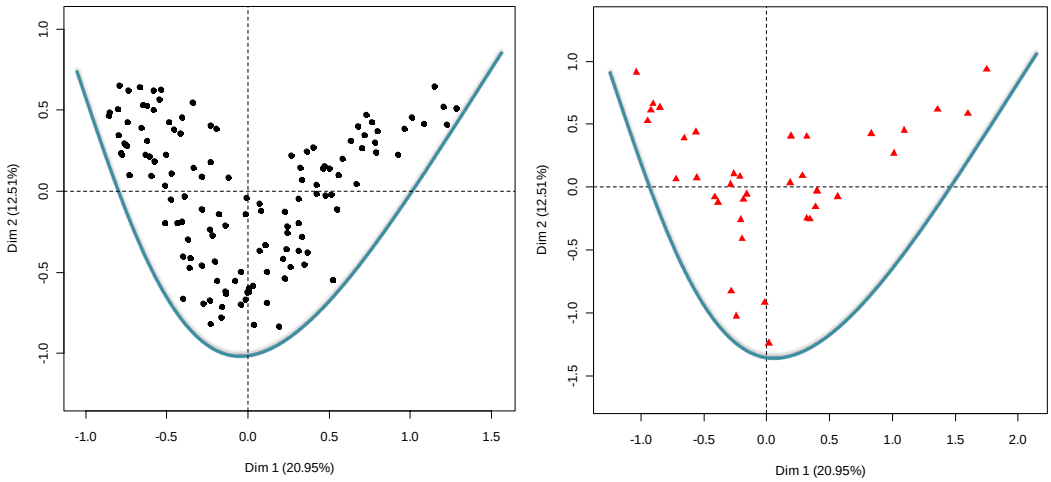

FIGURE 3.12 – Datos OGM : representación de los individuos (gráfico de la izquierda) y modalidades activas (gráfico de la derecha) sobre el primer plano.

```
> plot.MCA(res,invisible=c("var","quali.sup"),label=FALSE)
```

La forma de la nube de los individuos sobre el primer plano (cf. gráfico de la izquierda figura 3.12) recuerda la forma de una parábola : es lo que se llama el efecto Guttman . Este efecto traduce una redundancia entre las variables activas, es decir, una nube de individuos extremadamente estructurada según el primer eje factorial. En el ejemplo, esto traduce por una parte dos tipos extremos de posicionamiento con relación a los OGM que se reparten en los dos lados del primer eje factorial, y por otra parte un posicionamiento más moderado situado a lo largo del segundo eje factorial. No podemos decir nada más a simple vista de la nube de los individuos que hay que interpretar conjuntamente con la nube de las modalidades.
Con el fin de visualizar la nube de las modalidades activas, ejecutamos la línea de código siguiente :

```
> plot.MCA(res,invisible=c("ind","quali.sup"),label=FALSE)
```

Igualmente que para la nube de individuos, la forma de la nube de las modalidades sobre el primer plano (cf. gráfico de la derecha figura 3.12 o figura 3.13) recuerda la forma de una parábola, lo que corresponde al efecto Guttman.
Con el fin de interpretar los ejes factoriales, es indispensable representarlos asociados a su etiqueta, lo que se hace con la ayuda de la línea de código siguiente :

```
> plot.MCA(res,invisible=c("ind","quali.sup"))
```

Vemos pues (cf. figura 3.13), del lado positivo del primer eje factorial, las personas (a través de las modalidades que escogieron) que se sienten concernidas por la pregunta de los OGM y que son más bien desfavorables a su utilización ; del lado negativo del mismo eje factorial,

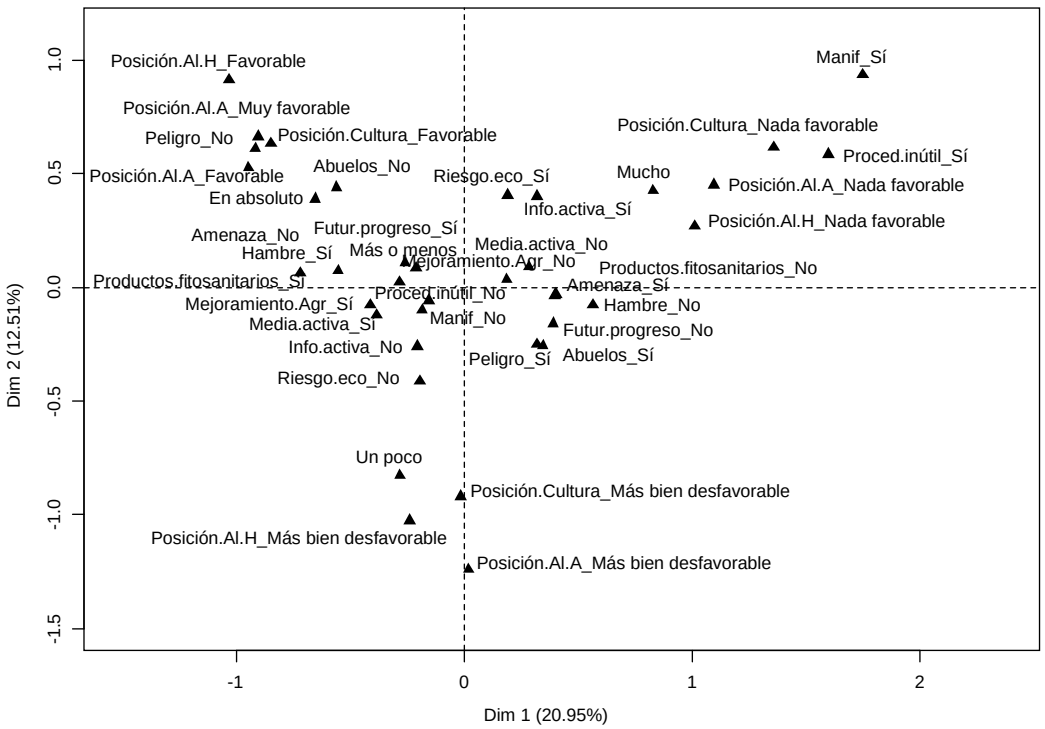

FIGURE 3.13 – Datos OGM : representación de las modalidades activas de su identificador sobre el primer plano.

las personas que no se sienten concernidas por la pregunta de los OGM y que son más bien favorables para su utilización.
También vemos, a lo largo del segundo eje factorial, a personas con la opinión menos tajante que se sienten un poco concernidas por la pregunta de los OGM y que son más bien desfavorables a su utilización.

3.8.4 Proyección de variables suplementarias

Ahora es interesante ver si esta estructura fuerte que percibimos a nivel de los individuos por lo que refiere a su posición relacionada con los OGM puede ser puesta en contacto con lo que son, es decir, ¿podemos explicar la relación a los OGM por el descriptivo socioeconómico ? Para ello, visualizamos la nube de las modalidades ilustrativas sobre el primer plano a partir de la línea de código siguiente :

```
> plot.MCA(res, col.quali.sup="brown", invisible=c("quanti.sup","ind","var"))
```

Esta representación de las variables suplementarias (cf. figura 3.14) es particularmente notable ya que aporta dos cosas. Por una parte, revela una estructura fuerte para ambas

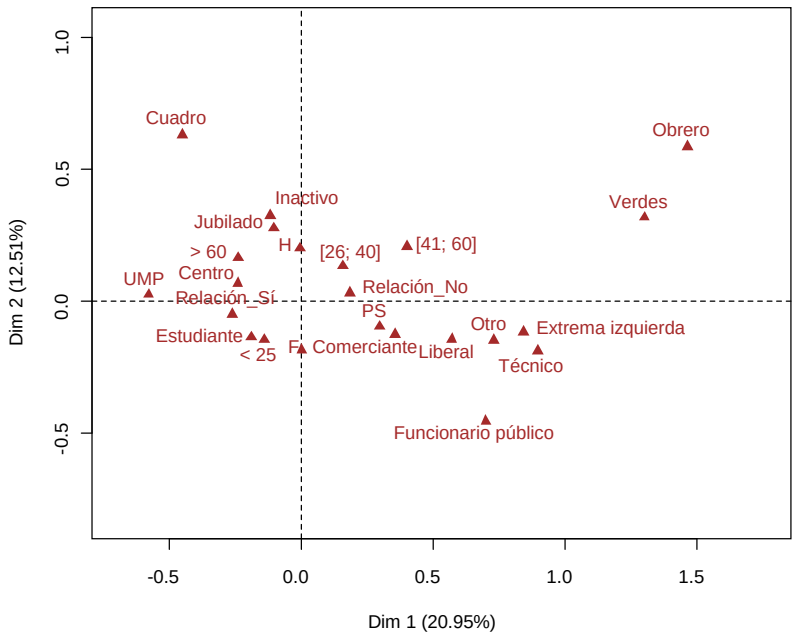

FIGURE 3.14 – Datos OGM : representación de modalidades ilustrativas y de su redacción sobre el primer plano.

variables *CSP* y *identificación a un partido político* y por otra parte, no pone en evidencia estructura particular con las variables de edad, de sexo, y de profesión en relación con la agricultura, la industria agroalimentaria o la industria farmacéutica.

Las modalidades *Ejecutivo*, *Inactivo* y *Jubilado* se oponen a las modalidades *Técnico* y *Obrero*, con la modalidad *Funcionario* situada en el medio ; igualmente, la modalidad *UMP* se opone a las modalidades *Verdes* y *Extrema izquierda*, con la modalidad *PS* situada en el medio.

3.8.5 Conclusión

La puesta en relación de estas tres nubes de puntos permite identificar tres posicionamientos diferentes con relación a los OGM. Estos posicionamientos tienen que ponerse en relación directa a la vez con el CSP del encuestado y el partido político con el que se identifica más ; estas dos últimas variables parecen estar particularmente vinculadas. En cambio, estos tres posicionamientos no parecen poder ser explicados por el sexo, ni por la edad, ni por el hecho de que la profesión ejercida esté en relación con la agricultura, la industria agroalimentaria o la industria farmacéutica, lo que se supone debe aportar un conocimiento científico suplementario sobre los OGM.

	Juez 18	Juez 31	Juez 40	Juez 93
Angel	1	1	6	1
Aromatics Elixir	2	2	5	2
Chanel N.˚5	1	3	5	2
Cinéma	1	4	3	3
Coco Mademoiselle	1	5	2	3
J'adore (agua de perfume)	3	5	1	1
J'adore (agua de colonia)	1	5	1	3
L'instant	2	5	2	1
Lolita Lempicka	3	4	3	3
Pleasures	1	5	1	1
Pure Poison	3	2	2	3
Shalimar	2	5	4	4

Tabla 3.9 – Datos perfume : categorización de los jueces 18, 31, 40 y 93.

3.9 Ejemplo : categorización

3.9.1 Descripción de los datos y problemática

La categorización es un proceso cognitivo por el cual diferentes objetos son reagrupados según sus similitudes por un conjunto de sujetos. Algunas veces es calificada de enfoque holístico en el sentido que los objetos que hay que categorizar son tomados en consideración en su carácter global. La categorización se utilizar para recoger datos, particularmente en análisis sensorial donde se procura comprender un conjunto de productos según sus propiedades sensoriales. En este contexto particular, esta prueba consiste en pedir a los consumidores / sujetos / jueces hacer grupos de productos en función de sus semejanzas sensoriales. Este párrafo presenta una aplicación un poco inhabitual del ACM con datos tan singulares donde cada variable puede ser considerada como una partición sobre un conjunto de objetos, lo que detallaremos más tarde.

Los datos de los que disponemos proceden de una colección de datos sensoriales organizada en el Agrocampus. 98 consumidores realizaron una prueba de categorización sobre 12 perfumes de lujo : Angel, Aromatics Elixir, Chanel 5, Cinéma, Coco Mademoiselle, L'instant, Lolita Lempicka, Pleasures, J'adore (agua de perfume), J'adore (agua de olor), Pure Poison, Shalimar (evidentemente las etiquetas de los perfumes fueron escondidas). Se les pidió, además de reagrupar los perfumes según sus semejanzas sensoriales, de caracterizar con palabras cada grupo así constituido.

Primeramente, los datos pueden ser reagrupados en una tabla de 12 filas y 98 columnas, en la cual, cada fila i corresponde a un perfume, cada columna j corresponde a un consumidor, una casilla (i, j) corresponde al grupo en el cual el producto i ha sido colocado por el consumidor j (cf. tabla 3.9). Cada consumidor j puede ser asociado así a una variable cualitativa j a K_j modalidades, donde K_j designa el número de grupos utilizados por el consumidor j en el curso de su categorización. Por ejemplo, en el tabla 3.9, podemos ver que el juez 31 ($j = 31$) categorizó los perfumes según 5 grupos ($K_{31} = 5$) y puso los perfumes J'adore (agua de perfume) y J'adore (agua de colonia) en el mismo grupo.

En segundo lugar, de modo equivalente, el índice del grupo en el cual el producto i ha sido colocado por el consumidor j puede ser reemplazado por las palabras que caracterizan el mismo grupo : del mismo modo, cada consumidor j puede ser asimilado a una variable

cualitativa a K_j modalidades (cf. tabla 3.10). Obtenemos así una tabla idéntica a la precedente pero con una codificación más explícita. Es esta segunda tabla la que se analiza en los párrafos siguientes.

	Juez 18	Juez 31	Juez 40	Juez 93
Angel	dinámico vida	fuerte	Gr6	floral fuerte
Aromatics Elixir	abuela	picante	Gr5	químico
Chanel N.˚5	dinámico vida	jabón	Gr5	químico
Cinéma	dinámico vida	Gr4	Gr3	floral débil
Coco Mademoiselle	dinámico vida	suave	Gr2	floral débil
J'adore (agua de perfume)	suave agradable bebé	suave	gel-ducha	floral fuerte
J'adore (agua de colonia)	dinámico vida	suave	gel-ducha	floral débil
L'instant	abuela	suave	Gr2	floral fuerte
Lolita Lempicka	suave agradable bebé	Gr4	Gr3	floral débil
Pleasures	dinámico vida	suave	gel-ducha	floral fuerte
Pure Poison	suave agradable bebé	picante	Gr2	floral débil
Shalimar	abuela	suave	limón	fuerte

Tabla 3.10 – Datos perfumes : ejemplo de datos de categorización con verbalización.

Uno de los objetivos principales de este estudio es el de proporcionar una imagen sintética del conjunto de los 12 perfumes de lujo teniendo como base las categorizaciones producidas por los 98 consumidores. Una vez obtenida esta imagen, para comprender las razones por las cuales dos perfumes se oponen, las dimensiones sensoriales de la imagen deben ser unidas a los términos utilizados en el momento de la caracterización de los grupos. Y para ir más lejos, veremos en este contexto sensorial particular cómo es posible explotar las propiedades baricéntricas del ACM con el fin de sacar el máximo provecho de estos datos.

3.9.2 Elección del análisis

En este estudio, los 12 perfumes son considerados como individuos estadísticos (activos), los 98 consumidores como variables cualitativas (activas) ; la tabla de datos es de tipo individuos×variables cualitativas y depende pues del análisis de las correspondencias múltiples. En el análisis, recordemos que estos datos son tomados en consideración a través de la tabla disyuntiva completa que contiene aquí $I = 12$ filas y $K = \sum K_j$ columnas : el consumidor j es representado por el conjunto de sus K_j variables indicadoras, cada variable corresponde a un grupo y toma el valor 1 si el perfume pertenece al grupo k y 0 si no es así (cf. § 3.4). La distancia entre dos perfumes es tal que :

1. dos perfumes i y l son confundidos si han sido puestos juntos por todos los consumidores ;
2. dos perfumes i y l son más próximos cuanto más han sido colocados en el mismo grupo por un gran número de consumidores ;
3. dos productos son más alejados cuanto más han sido colocados en dos grupos diferentes por un gran número de consumidores.

Para realizar el ACM, ejecutamos la línea de código siguiente que almacena los resultados del ACM en el objeto `res.parfums` :

```
> library(FactoMineR)
> perfume <- read.table("http://factominer.free.fr/libra/perfume.csv",
    header=TRUE, sep=";", row.names=1)
> res.perfume <- MCA(perfume)
```

Por defecto, la función considera todas los variables como activas y sólo necesita el nombre del juego de datos como parámetro de entrada.

3.9.3 Representación de los individuos sobre el primer plano

Con el fin de visualizar la nube de los individuos, ejecutamos la instrucción de código siguiente :

```
> plot.MCA(res.perfume, invisible="var", col.ind="black")
```

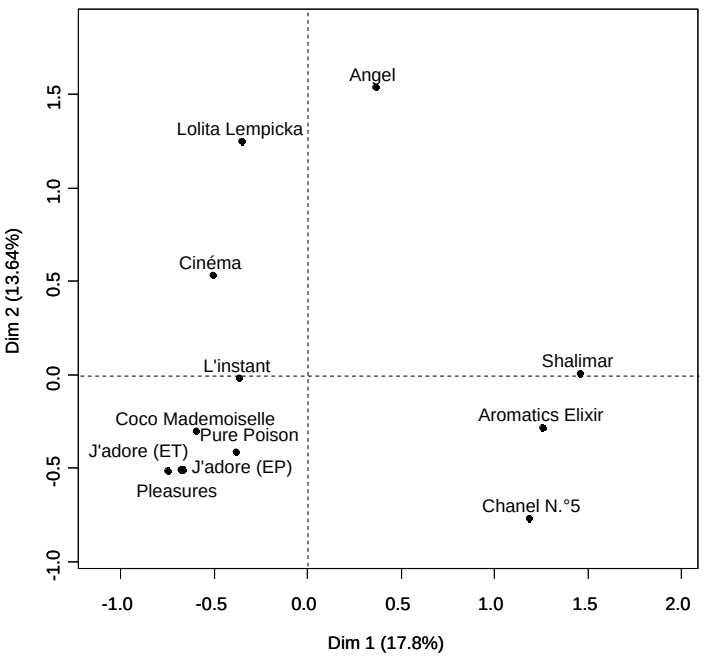

FIGURE 3.15 – Datos perfumes : representación de los perfumes sobre el primer plano.

El primer eje factorial opone el perfume Shalimar, Aromatics Elixir y Chanel 5 con otros (cf. figura 3.15). El segundo eje factorial opone a Angel, Lolita Lempicka y en una menor medida Cinéma con otros perfumes. Estas posiciones alejadas de algunos perfumes tienen que relacionarse con el número de veces que estos perfumes pertenecen a un grupo de un solo elemento : es el caso por ejemplo de Shalimar, Chanel 5 y Angel, que respectivamente han sido aislados por 24, 17 y 13 consumidores. Las proximidades tienen que relacionarse con la frecuencia de pertenencia a la misma clase : es el caso de Aromatics Elixir asociado

42 veces con Shalimar y 51 veces con Chanel 5 ; y de Lolita Lempicka asociado 36 veces con Angel. Anotemos igualmente la proximidad entre los dos J'adore, puestos juntos 56 veces.

3.9.4 Representación de las modalidades

La representación de los perfumes es completada superponiéndole la representación de las modalidades (que a partir de ahora denominaremos palabras) : por construcción, un perfume está en el baricentro de las palabras a los cuales ha sido asociado. Con el fin de hacer visible la nube de las modalidades y de interpretar las oposiciones entre perfumes, ejecutamos la línea de código siguiente :

```
> plot.MCA(res.perfume, invisible="ind", col.var="black")
```

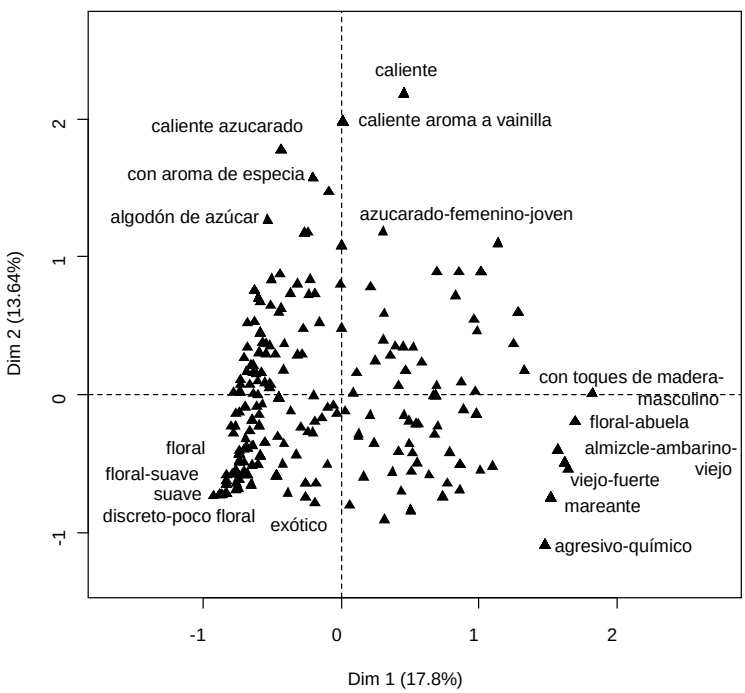

FIGURE 3.16 – Datos perfumes : representación de las palabras sobre el primer plano.

La nube «en bruto» de las modalidades es inexplotable directamente tal y como es proporcionada por la función **plot.MCA** debido al gran número de palabras. La figura 3.16 es una representación simplificada.
El primer eje opone los perfumes asociados con las palabras *fuerte*, *viejo*, con los perfumes más bien *florales*, *suaves*. El segundo eje opone los perfumes asociados con las palabras *calor*, *azucarado*, *con toques de especias* a los otros (cf. figura 3.16).

3.9.5 Representación de las variables

Las variables pueden ser representadas calculando las razones de correlación entre las coordenadas de los individuos sobre un eje y cada una de las variables cualitativas (cf. § 3.4.3). En el ejemplo, cada consumidor es representado por un punto y dos consumidores son más próximos cuanto más categorizaciones similares hayan realizado.
La figura 3.17 pone en evidencia diferentes tipos de categorizaciones. Sobre el primer eje, los consumidores 93 y 40, que tienen una coordenada elevada se oponen a los consumidores 18 y 31. La coordenada de un consumidor sobre un eje siendo igual a la razón de correlación entre su variable de particionamiento y el eje, nos indica que los consumidores 40 y 93 individualizaron claramente los perfumes Shalimar, Aromatics Elixir y Chanel 5, contrariamente a los consumidores 18 y 31 (cf. tabla 3.9). Según el segundo eje, los consumidores 31 y 40, que tienen una fuerte coordenada, se oponen a los consumidores 18 y 93. En efecto, los consumidores 31 y 40 individualizaron bien Angel y en menor grado Lolita Lempicka y Cinéma, lo que no es el caso de los consumidores 18 y 93 (cf. tabla 3.9).

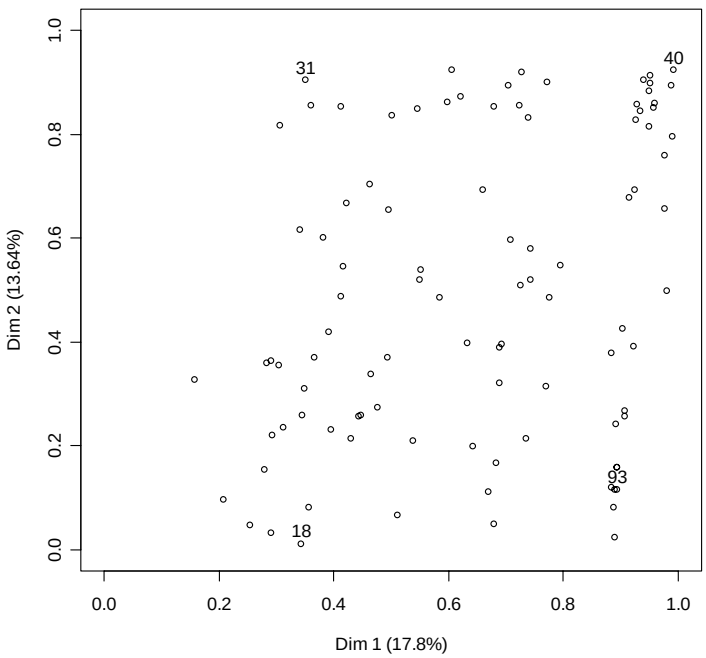

FIGURE 3.17 – Datos perfumes : representación de los consumidores sobre el primer plano.

Chapitre 4

Clasificación

4.1 Datos y problemática

Los métodos del análisis multidimensional de los datos (ADM) principalmente proporcionan representaciones sintéticas de objetos (estos objetos son esencialmente individuos, variables o modalidades de variables cualitativas) correspondientes a las filas y a las columnas de una tabla numérica. En ADM, el modo emblemático de la representación de un conjunto de objetos es una nube de puntos (cada punto es un objeto) evolucionando en un espacio euclidiano (pudiendo reducirse a un plano) ; el término euclidiano significa aquí que las distancias entre puntos (y los ángulos para las variables cuantitativas) se interpretan en términos de similitud para los individuos o las modalidades y en términos de correlación para las variables cuantitativas. Los métodos factoriales, entre los que los tres fundamentales (ACP, AFC y ACM) son descritos en los capítulos precedentes, proporcionan representaciones euclidianas.
Otro modo de representación de un conjunto de objetos, que pone en evidencia los parentescos entre ellos (similitudes o correlaciones) es el árbol jerárquico (cf. figura 4.1) ; decimos también, más simplemente, una jerarquía y, más precisamente una jerarquía indexada para recordar que la altura a la cual los objetos se reagrupan se interpreta (utilizamos también el término dendrograma). La utilización de este árbol es intuitiva : dos objetos son más parecidos cuanto más, para ir de uno a otro, no es necesario subir alto en el árbol. Así, en la figura 4.1 :
– los objetos A y B se parecen más que los objetos D y E ;
– el objeto C se parece más al conjunto de dos objetos D, E que al conjunto A, B.
Observemos que no se modifica la estructura de un árbol efectuando simetrías como lo muestran ambas representaciones de la figura 4.1 (un árbol jerárquico funciona desde este punto de vista como un móvil de Calder) : las proximidades laterales (por ejemplo entre B y C figura 4.1 a la izquierda) entre los objetos no se interpretan. Hay aquí un grado de libertad en la representación de un árbol que podemos utilizar si disponemos de un orden sobre los individuos procedente de una variable que desempeña un papel particularmente importante ; permutamos si es necesario las ramas de cada nudo para respetar más posible este orden lo.
El ejemplo más bello de árbol jerárquico es sin duda el de los seres vivos, en el que el

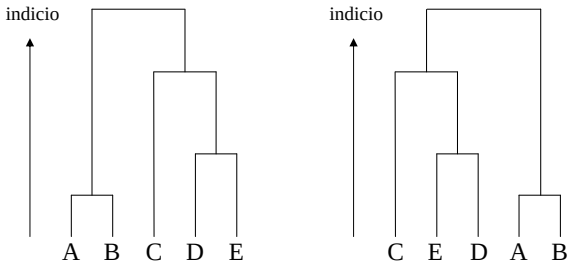

FIGURE 4.1 – Ejemplo del árbol jerárquico (sintetizando las similitudes entre cinco objetos : A, B, C, D, E).

primer nudo separa el reino animal y el reino vegetal. Es utilizado por todos los naturalistas. También son muy utilizados los ejemplos que describen las administraciones, lo que permite a cada uno conocer su (justo) puesto. Otro ejemplo : el árbol genealógico que describe la descendencia de un antepasado (forzosamente) ilustre. Finalmente la figura 4.2 es un buen ejemplo de visión sintética proporcinada por un árbol jerárquico.

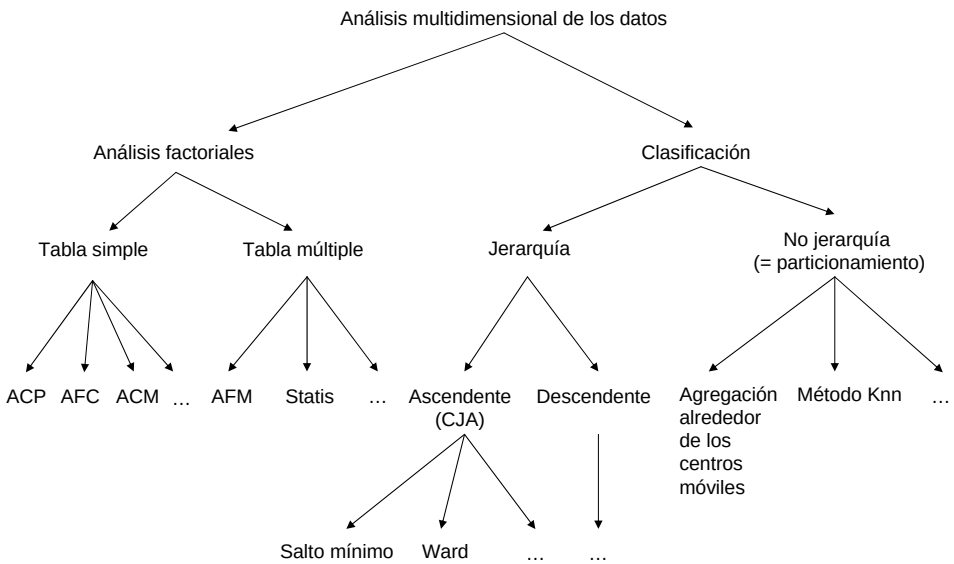

FIGURE 4.2 – Árbol jerárquico que ilustra las similitudes entre los principales métodos del análisis de los datos.

En estos ejemplos, los árboles han sido construidos por expertos según reglas establecidas en referencia a un modelo más o menos estricto. Para la representación de los seres vivos, por ejemplo, nos referimos a la evolución : idealmente, los diferentes nudos del árbol representan cada uno una etapa de la evolución, las más importantes correspondiendo a los nudos situados

en lo alto del árbol (que por ejemplo para el reino animal, separa, en primer lugar, los organismos unicelulares de los pluricelulares). La idea de evolución se encuentra en el orden lateral de los individuos : las ramas de un nudo están colocadas poniendo a la izquierda a los animales menos evolucionados.
A cada nudo se le asocia (por lo menos) un criterio y todos los individuos de una rama que derivan de este nudo presentan el mismo valor para este criterio. Tal conjunto de individuos se denomina monotético (por ejemplo : los organismos unicelulares, los vertebrados, los mamíferos, etc.). En este ejemplo en particular (pero también, aunque en menor grado, en otros precitados), el árbol jerárquico es el resultado de numerosas búsquedas, que permitieron entre otras cosas dar un valor a los criterios más importantes que definían los nudos de los niveles más elevados.
En este capítulo nos situamos en otra perspectiva, la misma que adoptamos en el análisis factorial, es decir, la exploración sin *a priori* de una tabla de datos (construida con *a priori*, de la entre los que emanan por ejemplo la elección de los individuos y la de las variables para definir la tabla que hay que analizar). Se trata de construir un árbol jerárquico (y no un plano factorial) para visualizar las similitudes entre objetos, que es un modo de estudiar la variabilidad contenida en la tabla. Esta problemática es la misma que en el análisis factorial : sólo el modo de representación difiere entre ambos enfoques.
Sin ideas *a priori*, procuraremos construir un árbol jerárquico en el que cada rama reúna individuos que constituyen un grupo politético (tal grupo es definido por un conjunto de propiedades tal que 1) cada elemento del grupo posee un gran número de estas propiedades 2) cada propiedad es poseída por un gran número de individuos del grupo).
Los algoritmos que construyen tales árboles son reagrupados bajo el término «clasificación jerárquica». Estos son numerosos : los más utilizados proceden de manera ascendente (reagrupando primero los objetos más semejantes y a continuación los grupos así constituídos) y son reagrupados bajo el término «Clasificación Jerárquica Ascendente» (CJA). Este capítulo describe e ilustra principalmente uno de los algoritmos más utilizados : el algoritmo de Ward (llamado también en Francia, «método de los momentos de orden 2»).
La tercera representación sintética de las similitudes entre objetos es la partición, conjunto de clases (de objetos) de manera que cada objeto pertenece solamente a una clase. Categóricamente, una partición es una variable cualitativa (cuyo valor, para cada objeto, es el nombre - o el número - de la clase a la cual pertenece). Así, en el momento de una encuesta de opinión distinguiremos por ejemplo a los hombres y a las mujeres, a los consumidores de tal producto de los que no lo consumen, etc. Pero estas clases, monotéticas, son interesantes sólo si la partición que constituyen está vinculada a un gran número de variables ; concretamente, en una encuesta de opinión, la partición hombres / mujeres es más interesante cuanto más el número de preguntas de opinión para las cuales las respuestas de los hombres difieren de las de las mujeres es importante.
Aquí todavía nos situamos en una perspectiva exploratoria : a partir de una tabla de datos rectangular que reúne las propiedades de un conjunto de objetos, queremos construir una partición de estos objetos tales que 1) dentro de cada grupo los individuos se parecen y 2) de un grupo al otro los individuos se diferencian. Varios algoritmos, reagrupados bajo el término de particionamiento, están disponibles ; nos limitamos en este capítulo al más utilizado de ellos : la agregación alrededor de los centros móviles.
Por no perder generalidad, hablamos hasta aquí de objetos, que pueden ser individuos es-

tadísticos, modalidades de variables cualitativas o variables cuantitativas. En efecto, un punto fuerte de los métodos de clasificación es que sus principios generales se aplican a objetos de naturalezas variadas. Pero esta generalidad perjudica al carácter concreto del planteamiento. A continuación también restringimos el planteamiento a los objetos constituidos por individuos estadísticos, descritos por un conjunto de variables cuantitativas o cualitativas : este caso es de lejos el más frecuente en la práctica.

Clasificar y asignar. Clasificar un conjunto de objetos consiste en establecer (o construir) clases o una jerarquía. Asignar un objeto consiste en poner este objeto en una de las clases de una partición definida *a priori*. La operación es denominada clasificación. En estadística, el término «discriminación» corresponde al problema de la búsqueda de reglas de clasificación (de individuos en una de las clases de una partición definida *a priori*) a partir de un conjunto de variables disponibles. Acordándose de que una partición puede ser vista como una variable cualitativa, el problema de la discriminación consiste en «predecir» una variable cualitativa (a partir de variables cuantitativas y/o cualitativas) del mismo modo que decimos que los métodos de regresión pretenden «predecir» una variable cuantitativa. El ejemplo emblemático de la discriminación es el diagnóstico médico : disponemos, para un enfermo, de sus valores para un conjunto de variables ; ¿cómo deducir la enfermedad que padece ? Cada enfermedad es una modalidad de la variable cualitativa que se puede llamar «nombre de la enfermedad» : el conjunto de las modalidades (una modalidad = una enfermedad) no es puesto en duda.

Clasificación supervisada o no supervisada. Recientemente, los investigadores introdujeron la terminología clasificación «no supervisada» para designar lo que es llamado desde hace tiempo (e igualmente en este libro) clasificación, el término «no supervisada» queriendo evocar el carácter exploratorio de los métodos. Esto por oposición a la clasificación «supervisada», que designa lo que es llamado desde hace tiempo (igualmente que en este libro) discriminación, el término «supervisada» queriendo evocar la focalización sobre una variable (la variable cualitativa que hay que predecir).

4.2 Formalización de la noción de similitud

4.2.1 Similitud entre individuos

En clasificación, jerárquica o no, es necesario especificar desde un principio lo que se entiende por similitud entre dos individuos. Esta necesidad existe también en el análisis factorial pero es menos visible porque esta especificación se incluye en el método. Al contrario, en clasificación, la elección es abierta, lo que es una ventaja frente a datos que presentan particularidades.

Distancias y distancias euclidianas

En el caso de una tabla que cruza I individuos y K variables cuantitativas en término general x_{ik} (valor del individuo i para la variable k) el ACP normado sitúa primero los I individuos en el espacio (vectorial) $\mathbb{R}^K$ y utiliza, para medir la similitud entre dos individuos i y l,

la distancia (euclidiana) usual en $\mathbb{R}^K$, es decir, anotando $\bar{x}_k$ (resp. s_k) la media (resp. la desviación-tipo) de la variable k :

$$\begin{aligned} d^2(i,l) &= \sum_{k=1}^{K}(x_{ik}-x_{lk})^2, \\ d^2(i,l) &= \sum_{k=1}^{K}(\frac{x_{ik}-\bar{x}_k}{s_k}-\frac{x_{lk}-\bar{x}_k}{s_k})^2, \\ d^2(i,l) &= \sum_{k=1}^{K}\frac{1}{s_k}(x_{ik}-x_{lk})^2. \end{aligned}$$

En la primera fórmula, la más general, los datos x_{ik} han sido previamente centrados y reducidos. En las dos otras fórmulas, ponemos de manifiesto explícitamente el centrado y la reducción, lo que será comentado más tarde.
Cuando se define la distancia d entre individuos de un espacio (aquí $\mathbb{R}^K$), decimos que proveemos este espacio de la distancia d (decimos también «métrica» y hablamos entonces de espacio métrico) porque en este espacio todos los cálculos relacionados a la noción de distancia deben ser hechos con esta distancia d. La función d de $I \times I$ en $\mathbb{R}^+$ definida en el ACP posee todas las propiedades matemáticas deseables, comenzando por la de una distancia (en el sentido matemático del término) sea :

$$\left\{ \begin{array}{lcl} d(i,l) & = & 0 \Longleftrightarrow i=l, \\ d(i,l) & = & d(l,i), \\ d(i,l) & \leq & d(i,j)+d(j,i) \text{ (desigualdad triangular).} \end{array} \right.$$

Se trata además de una distancia euclidiana, es decir, que permite definir una noción de ángulo y de proyección ortogonal (la definición axiomática del concepto de distancia euclidiana sobrepasa el marco de esta obra). Esta última noción estando en el centro del análisis factorial, todo método factorial necesariamente utiliza una distancia euclidiana (es también el caso de la distancia de χ^2 en AFC, por ejemplo). Pero, si no necesitamos la noción de proyección, que es el caso en clasificación, no tenemos la obligación de recurrir a una distancia euclidiana. Es aquí una propiedad de los métodos de clasificación interesante si queremos medir la similitud entre dos individuos por una distancia no euclidiana.

Ejemplo de distancia no euclidiana. En la distancia euclidiana usual, las distancias para cada variable intervienen al cuadrado. Esto aumenta la influencia de las grandes distancias de ahí la idea de hacer intervenir estas distancias por su valor absoluto. Esto conduce a la distancia siguiente, entre los individuos i y l :

$$d(i,l)=\sum_{k=1}^{K}|x_{ik}-x_{lk}|.$$

Esta distancia se llama Manhattan o aún city-block, en referencia a las ciudades americanas cuyas calles son o paralelas u ortogonales : para ir de un punto a otro, el camino recorrido

corresponde a la distancia que indicamos más arriba. Es de interpretación directa (*i.e.*, a partir de los datos) fácil.
La figura 4.3 ilustra sobre un caso elemental la diferencia entre ambas distancias. En este ejemplo, los individuos a y b difieren por una variable pero mucho ; los individuos a y c (o b y c) difieren por muchas variables pero poco. Para la distancia euclidiana usual, la distancia entre a y b es la más grande $d(a,b) = 2 > \sqrt{3} = d(a,c) = d(b,c)$; para la distancia city-block es a la inversa.

	V1	V2	V3
a	1	1	3
b	1	1	1
c	2	2	2

A

	a	b	c
a	0		
b	2	0	
c	$\sqrt{3}$	$\sqrt{3}$	0

B

	a	b	c
a	0		
b	2	0	
c	3	3	0

C

FIGURE 4.3 – Distancia euclidiana usual (B) y distancia city-block (C) ilustrada para tres individuos a, b, c descritos por tres variables $V1$, $V2$, $V3$ (A).

La distancia city-block no es una distancia euclidiana. ¿Qué debemos escoger ? Salvo una necesidad muy fuerte implicada por los datos (que jamás hemos encontrado) recomendamos utilizar una distancia euclidiana ya que permite poner en marcha conjuntamente una clasificación y un análisis factorial.

Otras distancias euclidianas. Existe una infinidad de distancias. Las más clásicas y fáciles de interpretar, consisten en partir de la distancia usual y en otorgar un peso a cada dimensión. Por ejemplo, en ACP normado, podemos también considerar que los datos son solamente centrados y que la distancia utilizada asigna a cada variable un peso inverso a su desviación-tipo (cf. la tercera fórmula de la definición de $d^2(i,l)$ más arriba). Estas fórmulas ilustran el hecho de que, en presencia de una distancia euclidiana cualquiera, podemos trabajar con la distancia usual por una transformación de los datos.

Similitudes y disimilitudes

Entre las primeras tablas que han sido objeto de una clasificación automática, figuran las tablas llamadas de presencia-ausencia en fitosociología. En una zona que estudiamos, definimos un conjunto de lugares que deben «representar» la diversidad de los medios que encontramos sobre la zona ; para cada lugar, hacemos una lista de las plantas presentes. Estos datos son reunidos en una tabla que cruza las I especies y los J lugares, cuyo término general x_{ij} vale 1 si la especie i está presente en el lugar j y 0 si no.
Uno de los objetivos generales de este tipo de estudio es la puesta en evidencia de asociaciones vegetales, es decir, de conjuntos de especies presentes en los mismos medios. De aquí la idea de clasificar especies ; dos especies se parecen más cuanto más las observamos en los mismos lugares (también podemos clasificar los lugares ; dos lugares son próximas si tienen numerosas especies en común). Queda por especificar esta noción de similitud.

Muy rápidamente, los fitosociólogos observaron que en la apreciación de la asociación entre dos especies, su presencia simultánea en el mismo lugar tiene más valor (= de significado ecológico) que su ausencia simultánea. De ahí la idea de construir una medida de similitud *ad hoc* tomando en consideración este aspecto. Numerosas medidas han sido propuestas. Cuando estas medidas no verifican la desigualdad triangular, las llamamos «disimilitudes» o «indicios de disimilitud» (o «indicio de similitud» cuando el valor es más grande cuanto más los individuos se parecen). La más antigua ha sido propuesta por Pablo Jaccard (en 1901). Anotando para una pareja de especies i y l : n_{++} el número de lugares donde ambas especies i y l son presentes y n_{+-} el número de lugares donde una única de ambas especies i y l está presente, el índice (de similitud) de Jaccard se escribe :

$$\frac{n_{++}}{n_{++} + n_{+-}}.$$

El índice no hace intervenir los lugares en los cuales las dos especies son ausentes.
Este tipo de enfoque se aplica más generalmente a las tablas de presencia-ausencia que cruzan individuos (a clasificar) y caracteres tales que la presencia de un carácter tiene, para el usuario, más «valor» que su ausencia. Sino, los caracteres pueden ser vistos como variables cualitativas con dos modalidades y el marco del ACM, en particular la distancia que se le asocia, convenie.
Hay otros casos donde la naturaleza de los objetos estudiados es tal que la medida de similitud que más se les adapta no es una distancia sino una disimilitud. Se proporciona un ejemplo por la semejanza entre genomas. Sin entrar en detalles, para un estadista, se trata de medir la similitud entre secuencias de letras que pertenecen al alfabeto {a, c, g, t }. Podemos pensar en contar en cada secuencia, la frecuencia de todas las sucesiones de n letras (con eventualmente varios valores de n) y utilizar entonces la distancia de χ^2. Pero resumir una secuencia a tal conjunto de frecuencias no es satisfactorio. Podemos tener la intuición que la semejanza entre dos genomas A y B será más cercano a través de la longitud de las largas sucesiones de letras comunes de A y B. A partir de estas longitudes construimos un indicador que satisface al genetista pero que no posee las propiedades de una distancia (incluso sin conocer precisamente cómo estas longitudes son tomadas en consideración en el indicador, lo que es un poco técnico, podemos tener la intuición de que la desigualdad triangular no será verificada). Los métodos de clasificación son verdaderamente valiosos en tales casos para respetar la medida de similitud adaptada a los objetos que hay que clasificar.

4.2.2 Similitud entre grupos de individuos

Para construir un árbol jerárquico, es necesario definir una distancia o una disimilitud entre grupos de individuos. Existen varias posibilidades : citamos sólo las más importantes. Sean dos grupos de individuos A y B. El salto mínimo entre A y B (= relación simple = single linkage) es el más pequeño de las distancias entre un elemento de A y un elemento de B. El diámetro entre A y B (= relación completa = complete linkage) es el más grande de las distancias entre un elemento de A y un elemento de B. La figura 4.4 ilustra estas definiciones. El mayor interés de las definiciones precedentes es que son aplicables a todas las distancias o disimilitudes. En el caso de las distancias euclidianas, existen otras posibilidades. Consideramos G_A y G_B los centros de gravedad de los conjuntos de individuos A y B. Una primera idea

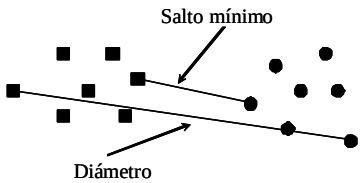

FIGURE 4.4 – Salto mínimo y diámetro entre dos grupos de individuos (identificados por símbolos diferentes).

consiste en medir la disimilitud entre A y B por la distancia entre sus centros de gravedad. Otro punto de vista, más completo, es el de la inercia : consiste en tomar en consideración los pesos de los grupos (en este capítulo, consideramos que los individuos tienen el mismo peso, caso más frecuente, y el peso de un grupo es proporcional a su efectivo ; mencionamos aquí que el punto de vista de la inercia permite tomar en consideración simplemente pesos diferentes de un individuo al otro).
Apliquemos al conjunto de los elementos de A y de B ($A \cup B$ de centro de gravedad G) el teorema de Huygens. Inercia total (de $A \cup B$ con relación a G) = Inercia inter (de $\{G_A, G_B\}$ con relación a G) + Inercia intra (inercia de A con relación a G_A más inercia de B con relación a G_B). Esta descomposición sugiere tomar la inercia inter como medida de disimilitud entre A y B. Daremos algunas propiedades de esta estrategia en la sección dedicada al método de Ward, fundada sobre este criterio.

4.3 Construcción de una jerarquía indiciada

4.3.1 Algoritmo clásico de construcción ascendente

El punto de partida es una matriz de disimilitudes D (estas disimilitudes pueden ser distancias euclidianas) entre individuos donde el término general $d(i,l)$ es la disimilitud entre los individuos i y l. Esta matriz es simétrica y contiene ceros sobre la primera diagonal : solamente es necesario una mitad, por convenio, la parte triangular baja.
Agregamos los individuos i y l más «similares» o «próximos» (en caso de *ex-æquo*, escogemos arbitrariamente uno de ellos) : constituimos así un nuevo elemento, (i,l) : este grupo de individuos no será puesto en duda más tarde. El valor $d(i,l)$ es el índice de la agregación entre i y l. Este valor es utilizado para definir la altura a la cual las ramas del árbol que corresponden a i y a l se reúnen.
Ponemos al día la matriz D suprimiendo las filas y las columnas que corresponden a los individuos i y l y creando una nueva fila y una nueva columna para el grupo (i,l) que completamos con disimilitudes entre este grupo y cada uno de los individuos restantes. Obtenemos la matriz D(1) en la cual buscamos la pareja de los elementos más próximos. Estos elementos son agregados y etcétera.
En calidad de ejemplo, aplicamos este algoritmo a un pequeño conjunto de datos que contiene seis individuos repartidos sobre un plano. Por razones de facilidad de cálculo, utilizaremos la distancia inicial city-block y el recálculo de las distancias según el diámetro. Las etapas

de la construcción del árbol son mostradas en la figura 4.5.

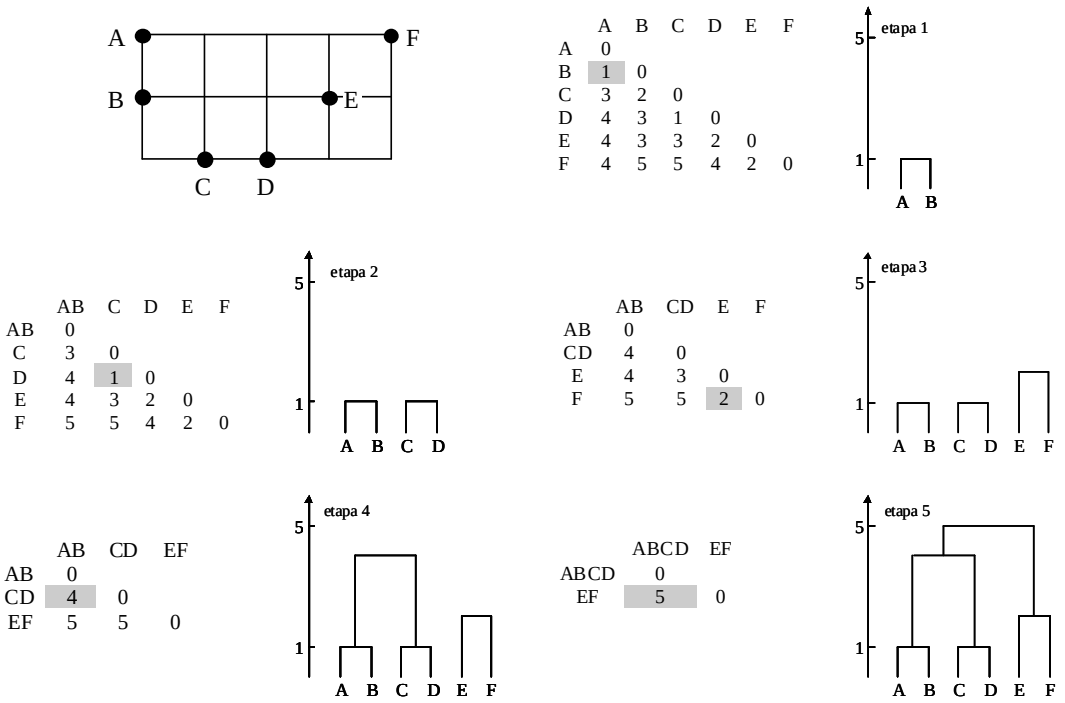

FIGURE 4.5 – Etapas de la construcción de un árbol jerárquico a partir de seis individuos repartidos sobre un plano.

4.3.2 Jerarquía y partición (figura 4.6)

Los puntos donde se reúnen las ramas que corresponden a los elementos que reagrupamos son llamados «nudos». Decimos también algunas veces «bifurcación» para expresar una descripción descendente del árbol. Los individuos que hay que clasificar son llamados algunas veces «nudos terminales». Con I individuos, hay $(I-1)$ nudos a menudo numerados de $I+1$ a $2\times I$ (cf. figura 4.6) por orden de aparición en la construcción del árbol (los I primeros números son reservados a los nudos terminales ; no obstante, en ciertos programas, la numeración de los nudos no toma en consideración los nudos terminales). Trazando una línea horizontal con un índice dado, definimos una partición (decimos que cortamos el árbol). Sobre la figura 4.6, el nivel de corte A define una partición en dos clases $\{1,2,3,4\}$ y $\{5,6,7,8\}$; El nivel de corte B define una partición más fina en cuatro clases $\{1,2\}$, $\{3,4\}$, $\{5,6\}$ y $\{7,8\}$. Por construcción, estas particiones son encajadas : cada clase del nivel B es incluída en la misma clase del nivel A.
Resulta así que un árbol jerárquico puede ser visto como una continuación de particiones encajadas, yendo de la más fina (en la cual cada individuo constituye una clase) a la más

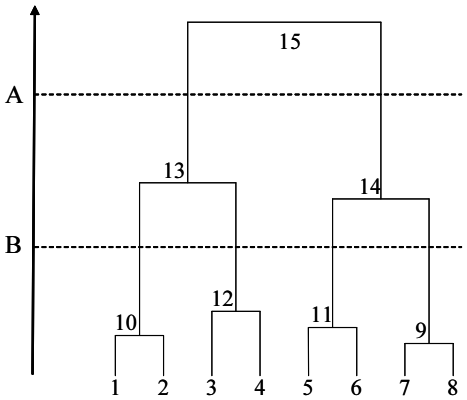

FIGURE 4.6 – Jerarquía y partición.

grosera (en la cual sólo hay una clase).

4.4 Método de Ward

Su principio ha sido esbozado más arriba. Este método se aplica a individuos situados en un espacio euclidiano. Es el caso más frecuente, una tabla en la cual un conjunto de individuos es descrito por un conjunto de variables. Cuando los datos son cuantitativos (resp. cualitativos), estudiamos la nube N_I evolucionando en $\mathbb{R}^K$ definido en § 1.3.1 (resp. § 3.4.2). Este método, ascendente, consiste, a cada paso, en reagrupar dos elementos (individuos aislados o clases ya formadas) maximizando la calidad de la partición obtenida.

4.4.1 Calidad de una partición

Una buena partición es tal que :
– dentro de una clase los individuos son homogéneos (variabilidad intra-clase débil) ;
– de una clase a otra los individuos son diferentes (variabilidad inter-clases elevada).
Si los individuos están en un espacio euclidiano, el teorema de Huygens proporciona un marco de análisis bien adaptado al estudio de una partición. Este teorema descompone la inercia total (de la nube de los individuos) en dos partes :
– la inercia intra-clase, fundada sobre la diferencia entre cada punto y el centro de gravedad de la clase a la cual pertenece ;
– la inercia inter-clases, fundada sobre la diferencia entre cada centro de gravedad de una clase y el centro de gravedad general.
De modo muy general se escribe :

$$\text{Inercia total} = \text{Inercia inter-clases} + \text{Inercia intra-clase}.$$

Si consideramos los individuos descritos por una sola variable cuantitativa (anotada y), encontramos la ecuación del análisis de varianza a un factor. Con I individuos (de mismo

peso 1) repartidos en Q clases anotamos : y_{iq} el valor (para y) del i individuo de la clase q ; $\bar{y}_q$ la media de y para los individuos de la clase q ; I_q el número de individuos de la clase q ; $\bar{y}$ la media general de y. El teorema de Huygens se escribe :

$$\sum_{q=1}^{Q}\sum_{i=1}^{I_q}(y_{iq}-\bar{y})^2 = \sum_{q=1}^{Q} I_q(\bar{y}_q-\bar{y})^2 + \sum_{q=1}^{Q}\sum_{i=1}^{I_q}(y_{iq}-\bar{y}_q)^2.$$

En general, disponemos de K variables cuantitativas y la nube de los individuos evoluciona en $\mathbb{R}^K$ (cf. la nube N_I en ACP § 1.3.1 ; veremos en § 4.7.1 cómo volver a este caso cuando las variables son cualitativas). Las dimensiones de $\mathbb{R}^K$ siendo ortogonales, el teorema de Huygens se obtiene sumando las inercias a lo largo de cada dimensión. Sea, anotando y_{iqk} el valor para la variable k del individuo i de la clase q :

$$\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\sum_{k=1}^{K}(y_{iqk}-\bar{y}_k)^2 = \sum_{q=1}^{Q}\sum_{k=1}^{K} I_q(\bar{y}_{qk}-\bar{y}_k)^2 + \sum_{q=1}^{Q}\sum_{i=1}^{I_q}\sum_{k=1}^{K}(y_{iqk}-\bar{y}_{qk})^2.$$

$$\text{Inercia total} = \text{Inercia inter-clases} + \text{Inercia intra-clase.}$$

Si se adopta esta descomposición como marco de análisis (*i.e.*, si se mide la variabilidad por la inercia) entonces, en la búsqueda de una buena partición, es indiferente minimizar la variabilidad intra-clase o maximizar la variabilidad inter-clases (ya que la variabilidad total es fijada por los datos). Esto es cómodo para el usuario, que tendría dificultades frente a una aplicación particular, de privilegiar uno de ambos criterios. De ello resulta que la calidad de una partición puede ser medida por :

$$\frac{\text{Inercia inter-clases}}{\text{Inercia total}}.$$

Este cociente indica la parte de variabilidad total expresada por la partición. A menudo es multiplicado por 100 para poder ser enunciado en término de porcentaje. En el caso unidimensional, se confunde con la (cuadrado de la) razón de correlación. Con los datos de la figura 4.5, utilizando esta vez la distancia euclidiana usual y considerando la partición en tres clases $\{A, B\}$, $\{C, D\}$ y $\{E, F\}$, este cociente vale 0.8846. Esta partición expresa pues 88.46 % de la variabilidad de los individuos. Es decir, si en lugar de considerar el conjunto de los seis individuos consideramos sólo las tres clases, representamos 88.46 % de la variabilidad de los datos. Este porcentaje debe ser apreciado comparado con el número de individuos con el número de clases. En efecto, aumentando el número de clases, podemos encontrar una partición que presenta un porcentaje (de inercia expresada) tan elevado como queremos. La partición en la cual cada individuo constituye una clase presenta un porcentaje de 100 % pero no tiene ningún interés práctico. En el pequeño ejemplo, podremos considerar que la partición de seis individuos en tres clases, que en cierto modo divide por 2 la complejidad de los datos pero que expresa 88.46 % es satisfactoria.

4.4.2 Agregación por la inercia

Situémonos en la etapa n del algoritmo ascendente. Los individuos son repartidos en $Q = (I - n + 1)$ clases obtenidas por las etapas precedentes. La cuestión es escoger las dos

clases (entre las Q) que se van a agregar. Agregando dos clases, pasamos de una partición en Q clases a una partición en $Q-1$ clases ; la inercia intra-clase sólo puede aumentar (resultado inmediato aplicando el teorema de Huygens sobre estas dos clases, lo que muestra también que el aumento es nulo si y sólo si ambas clases tienen el mismo centro de gravedad). La idea de la agregación por la inercia consiste en escoger las dos clases que hay que agregar para minimizar el crecimiento de inercia intra-clase. A causa del teorema de Huygens, esta agregación de dos clases conlleva una disminución de la inercia inter-clases, disminución que es minimizada.
Consideremos las clases p (de centro de gravedad g_p y de efectivo I_p) y q (de centro de gravedad g_q y de efectivo I_q). El aumento $\Delta(p,q)$ de inercia intra-clase engendrada por la reagrupación de las clases p y q puede escribirse :

$$\Delta(p,q) = \frac{I_p I_q}{I_p + I_q} d^2(g_p, g_q).$$

Escoger las clases p y q tales que $\Delta(p,q)$ sea mínimo se reduce a escoger :
– clases cuyos centros de gravedad son próximos ($d^2(g_p, g_q)$ pequeño) ;
– clases de efectivos débiles ($\frac{I_p I_q}{I_p+I_q}$ pequeño).
La primera propiedad es intuitiva. La segunda lo es menos pero presenta una consecuencia interesante : la agregación por la inercia tiende a producir árboles armoniosos en el sentido que las particiones son compuestas por clases de efectivos no demasiado diferentes. Aplicando este algoritmo sobre los datos de la figura 4.5, obtenemos el árbol de la figura 4.7 ; los índices de niveles y el detalle de su cálculo son recapitulados en la tabla 4.1.
El aspecto general del árbol es idéntico al obtenido en la figura 4.5 (con otra distancia y otro criterio de agregación) : cuando una estructura es fuerte, (casi) es puesta en evidencia cualquiera que sea el método empleado. La diferencia superior entre las dos jerarquías reside en la variabilidad de los niveles : la agregación por la inercia aumenta las diferencias entre los niveles más altos por una parte y los niveles más bajos por otra parte, y ello se debe al coeficiente $\frac{I_p I_q}{I_p+I_q}$ que (casi) crece «mecánicamente» entre los primeros niveles (que agregan elementos de efectivos débiles) y los últimos (que agregan - en general - clases de efectivos fuertes).
En ciertos programas, existe una opción de representación de un árbol jerárquico que utiliza, como índices de nivel de un nudo, la raíz cuadrada de la ganancia de inercia intra. Y de ello obtenemos un aspecto de los árboles más apretujado. En este libro, utilizamos el índice original, *i.e.* la ganancia de inercia intra.

4.4.3 Dos propiedades del índice de agregación

1. En la representación del árbol, la cantidad $\Delta(p,q)$ es utilizada como índice. Este índice va creciendo (anotando Δ_n el índice asociado en la etapa n tenemos : $\Delta_n \geq \Delta_{n-1}$), hecho que se puede intuir : agregamos primero clases próximas y de efectivos débiles ; luego acabamos por agregar clases alejadas y de efectivos importantes. Esta primera propiedad es importante : garantiza que el árbol no presenta «inversiones» (hay «inversión» por ejemplo cuando el elemento $\{c\}$ se agrega el grupo a, b a un nivel inferior al de la agregación entre a y b, cf. figura 4.8).

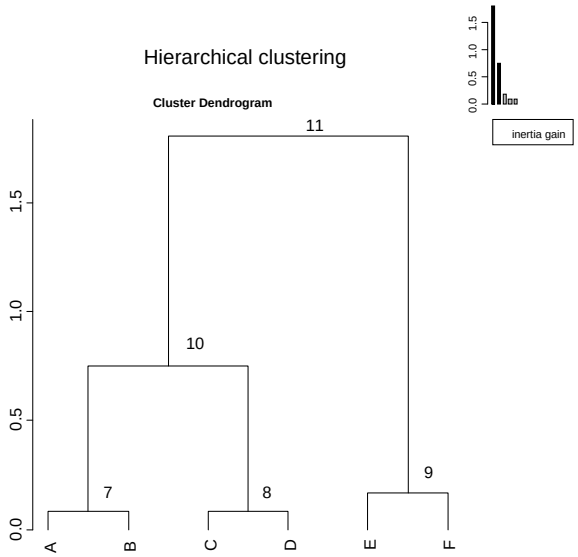

FIGURE 4.7 – Árbol procedente del algoritmo de Ward aplicado sobre los datos de la figura 4.5 y utilizando la métrica euclidiana usual. Arriba a la derecha : diagrama de los índices de nivel (del nudo de nivel más elevado al nudo de nivel más bajo). Los números de los nudos han sido añadidos sobre el gráfico.

2. La suma de todos los índices de agregación (de la misma jerarquía) es igual a la inercia total del conjunto de los individuos (con relación a su centro de gravedad). Sea :

$$\sum_{n=1}^{I-1} \Delta_n = \text{inercia total.}$$

Esta propiedad se obtiene fácilmente considerando la evolución de la partición de los individuos a lo largo de las etapas sucesivas en la construcción del árbol. En la etapa

N.° del nudo	p	q	$\frac{I_p I_q}{I_p + I_q}$	$d^2(g_p, g_q)$	Índices	en %	% acumulado	Inercia intra	Varianza intra
7	2	1	0,5	0,167	0,083	2,88	100	0,083	0,250
8	4	3	0,5	0,167	0,083	2,88	97,12	0,083	0,250
9	6	5	0,5	0,333	0,167	5,77	94,23	0,167	0,500
10	8	7	1	0,750	0,750	25,96	88,46	0,917	1,375
11	9	10	1,33	1,354	1,806	62,50	62,50	2,889	2,889
					2,889	100			

Tabla 4.1 – Indices asociados con la figura 4.7. Los individuos son considerados como nudos numerados en el orden de aparición del archivo (aquí, el orden alfabético).

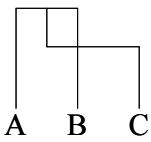

FIGURE 4.8 – Ejemplo de árbol que presenta una inversión.

0, cada individuo constituye una clase y la inercia intra-clase de la partición es nula. A medida que avanzamos en el algoritmo, el número de clases disminuye y la inercia intra-clase aumenta (de Δ_n en la etapa n) ; al final del algoritmo, todos los individuos están en la misma clase y la inercia intra-clase es igual a la inercia total. Concluimos que una jerarquía indexada (obtenida por este método) propone una descomposición de la inercia total (*i.e.*, la variabilidad de los datos) y, desde este punto de vista, se inscribe en una problemática global análoga a la del análisis factorial ; la diferencia es que la descomposición es realizada por clases en un caso y por ejes en la otra.

4.4.4 Análisis de una jerarquía, elección de una partición

Aunque construida de modo ascendente, una jerarquía es generalmente analizada de modo descendente. Recordemos el objetivo : dar una visualización de la variabilidad de los datos o, de otro punto de vista, del conjunto de las similitudes entre los individuos. En esta perspectiva, el último nudo de la jerarquía responde a la pregunta : ¿si hubiese que resumir la variabilidad en una partición de los individuos en dos clases, cual sería esta partición ? Observemos de paso que el término «nudo» evoca más la similitud de dos clases (óptica de la construcción ascendente) que una subdivisión en dos clases ; de ahí el término de «bifurcación» utilizado algunas veces en una investigación descendente.
Con la agregación por la inercia, el nivel de un nudo, visto de modo descendente, determina la cantidad que se gana (en inercia inter-clases o en disminución de inercia intra-clase) separando ambas clases que reagrupa. En el ejemplo (cf. figura 4.7 y tabla 4.1), la separación en dos grupos expresa 62.50 % de la variabilidad. Si consideramos la partición en tres clases, la separación inducida por el nudo 10 (considerar $\{a, b\}$ y $\{c, d\}$ más bien que $\{a, b, c, d\}$) representa 25.96 % de la variabilidad y conduce pues a un porcentaje de 62.50 %+25.96 % = 88.46 % para la partición en tres clases.
Percibimos aquí que una jerarquía es muy útil para razonar la elección de una partición ; incluso su interés es mayor en las aplicaciones donde los individuos son anónimos, como el caso de las encuestas por ejemplo. Concretamente tendremos en cuenta :
– el aspecto general del árbol ; en el ejemplo de la figura 4.7, sugiere una partición en tres clases ;
– los niveles de los nudos, para cuantificar el punto precedente ; estos niveles pueden ser representados por un diagrama en barras que hace visible su decrecimiento (gráfico representado arriba a la derecha de la figura 4.7) ; cada irregularidad en este decrecimiento sugiere un nivel de corte ;
– el número de clases que no debe ser demasiado elevado en cuyo caso el interés sintético del procedimiento disminuye ;

– la interpretabilidad de las clases : aunque corresponde a una ganancia de inercia inter-clases apreciable, no retendremos una subdivisión que no sabemos interpretar ; del mismo modo, retendremos una subdivisión claramente interpretable aunque corresponde a una ganancia de inercia poco elevada. Afortunadamente, en la práctica, tales dilemas son poco frecuentes.

El análisis visual del árbol jerárquico y del diagrama de los índices de nivel sugiere un nivel de corte en Q clases cuando la ganancia de inercia inter entre $Q-1$ y Q clases es mucho más importante que el nivel de corte entre Q y $Q+1$ clases. Adoptando un proceso descendente (*i.e.*, partiendo de la partición más grosera), esto vuelve a minimizar el criterio siguiente :

$$\min_{q_{min} \leq q \leq q_{max}} \frac{\Delta(q)}{\Delta(q+1)}$$

con $\Delta(q)$ la ganancia de inercia inter-clases cuando se pasa de $q-1$ a q clases, q_{min} (resp. q_{max}) el número mínimo (resp. máximo) de clases deseadas por el usuario. La función **HCPC** (Hierarchical Clustering Principal Components) pone en marcha este cálculo después de haber construido la jerarquía y propone un nivel de corte «óptimo». Es el estudio de un árbol, este nivel de corte generalmente corresponde a la intuición visual. En el estudio automático de un gran número de árboles que esto es más valioso.

4.5 Investigación directa de una partición : agregación alrededor de los centros móviles

4.5.1 Datos y problemática

Los datos son los mismos que para el análisis factorial : una tabla individuos × variables y una distancia euclidiana. Consideramos las variables cuantitativas, sin pérdida de generalidad ya que, como para el CJA, la sección 4.7.1 mostrará cómo volver a este proceso cuando las variables son cualitativas. Los algoritmos de particionamiento, que a partir de una tabla individuos × variables producen una partición de los individuos, se sitúan frente a la clasificación jerárquica principalmente según las dos preguntas siguientes :

– En práctica, una jerarquía indexada es utilizada muy a menudo como herramienta de investigación de una partición. ¿Hay algunas ventajas en buscar directamente una partición ?
– Cuando el número de individuos es grande, el tiempo de cálculo necesitado por la construcción de una jerarquía indexada puede ser prohibitivo. ¿Podemos esperar tiempos de cálculo más cortos por parte de los algoritmos de búsqueda directa de una partición ?

Existen varios algoritmos de particionamiento : limitaremos el planteamiento a uno de ellos. La agregación alrededor de los centros móviles (método llamado también «k-means»), es suficiente en la práctica.

4.5.2 Principio

El número Q de clases es fijado *a priori*. Podríamos pensar en calcular todas las particiones posibles y retener la que optimiza un criterio dado. De hecho, consideraciones combinato-

rias muestran que el tiempo de cálculo asociado con este proceso es prohibitivo cuando el número de individuos es un poco grande. Utilizamos pues el algoritmo iterativo descrito a continuación. Sea P_n la partición de los individuos en la etapa n del algoritmo y ρ_n el cociente [(inercia inter) / (inercia total)] de esta partición P_n

0. Consideramos una partición inicial P_0 ; calculamos ρ_0.

   En la etapa n del algoritmo :

1. Calculamos el centro de gravedad $g_n(q)$ de cada clase q de P_n ;
2. Volvemos a afectar cada individuo a la clase q de la que está más próxima (en término de distancia a los centros de gravedad $g_n(q)$) ; obtenemos una nueva partición P_{n+1} y calculamos su cociente ρ_{n+1} ;
3. Mientras $\rho_{n+1} - \rho_n >$ umbral (*i.e.*, la partición P_{n+1} es mejor que P_n) regresamos en 1. Si no, P_{n+1} es la partición buscada.

La convergencia de este algoritmo está asegurada por el hecho de que, a cada etapa, ρ_n disminuye. En la práctica, esta convergencia es rápida (generalmente menos de 5 iteraciones incluso para un conjunto importante de datos). La figura 4.9 ilustra este algoritmo sobre un conjunto de datos escogidos en un plano.

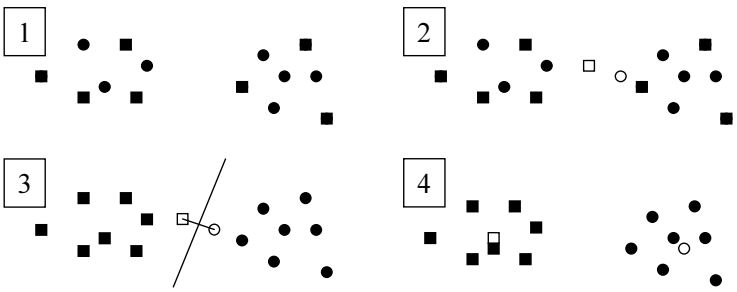

FIGURE 4.9 – Ilustración del algoritmo de la agregación alrededor de los centros móviles en un caso simple (los datos presentan un número de clases bien claro y correspondiente al número de clases impuesto al algoritmo).

Buscamos una partición en dos clases de los 14 individuos (clase de los círculos y clase de los cuadrados).

1. La asignación de los individuos a ambas clases es hecha al azar.
2. Calculamos los centros de gravedad de cada una de las clases (círculo y cuadrado vacíos).
3. Asignamos cada individuo a la clase de la cual está más próximo (representamos la mediatriz del segmento uniendo los centros de gravedad).
4. Calculamos los centros de gravedad de las nuevas clases.

Si se aplica de nuevo la etapa 3, no hay ningún cambio : el algoritmo ha convergido.

4.5.3 Metodología

El algoritmo descrito más arriba converge pero no necesariamente hacia un óptimo global. En práctica, ejecutamos muchas veces el algoritmo partiendo de particiones iniciales P_0

diferentes. Conservamos la mejor solución. Podemos también cruzar las particiones obtenidas en consecuencia de una serie de ejecuciones del algoritmo. Llamamos «formas fuertes» a los conjuntos de individuos que pertenecen a la misma clase cualquiera que sea la partición. Estas formas fuertes constituyen grupos de individuos cuya estabilidad frente a la partición inicial es interesante : ponen en evidencia zonas (del espacio) de densidad fuerte. Pero esta metodología conduce también a algunas clases de efectivo débil, a menudo reducido a un solo individuo, conteniendo individuos situados entre las zonas de densidad fuerte y cuya gestión es necesariamente empírica (las dos opciones principales son la asignación a la forma fuerte - de efectivo suficiente - la más próxima o la creación de una clase «residual» que reagrupa los individuos aislados).

4.6 Particionamiento y clasificación jerárquica

Frente a los métodos jerárquicos, los métodos de particionamiento presentan dos ventajas mayores :
- optimizan un criterio ; en CJA, optimizamos un criterio a cada paso pero no nos referimos a un criterio global que concierne al árbol mismo ;
- pueden tratar a números de individuos mucho más importantes.

Pero estos métodos necesitan fijar *a priori* el número de clases. De aquí la idea de combinar ambos procesos con el fin de obtener una metodología que presenta las ventajas de cada uno de ellos.

4.6.1 Consolidación de una partición

Al final de un CJA, el análisis de la jerarquía conlleva generalmente que el usuario se interese por una partición. Esta partición puede ser introducida como partición inicial de un algoritmo de particionamiento. Conservamos *en definitiva* la partición procedente de este algoritmo. En la práctica, la partición inicial jamás es modificada ; la partición es mejorada «al margen» (decimos «consolidada»), el aumento del cociente [(inercia inter) / (inercia total)] (aunque generalmente débil) asegura clases (un poco) más homogéneas y separadas. El inconveniente, menor, de esta metodología es que la jerarquía producida por el CJA (exactamente) no está en acuerdo con la partición escogida.

4.6.2 Algoritmo mixto

En presencia de un número demasiado grande de individuos para emprender directamente un CJA, podemos aplicar la metodología siguiente en dos etapas.
Etapa 1. Realizamos un particionamiento con un número de clases muy elevado (digamos 100 para fijar las ideas). La partición obtenida no es utilizable directamente en una perspectiva de interpretación : las clases son muy numerosas y muchas son muy próximas entre ellas. En cambio, cada una de ellas es muy homogénea (débil inercia intra-clase) y contiene individuos de los que estamos seguros que no hay que separarlos.
Etapa 2. Ponemos en ejecución un CJA tomando como elementos que hay que clasificar los grupos de individuos de la etapa 1 (cada elemento teniendo como peso el número, o más

generalmente, la suma de los pesos, de los individuos a los que representa). Obtenemos así una jerarquía que es, groseramente, lo alto de la jerarquía que se obtendría clasificando a los individuos mismos.
Una variante de la etapa 1 consiste en realizar varios particionamientos y en conservar las formas fuertes para la etapa 2.

4.7 Clasificación y análisis factorial

Hemos evocado muchas veces este punto : clasificación automática y análisis factorial se inscriben en la misma perspectiva (el análisis exploratorio de una tabla rectangular) y difieren según el modo de representación (nube euclidiana, jerarquía indexada o partición). De ahí la idea de combinar ambos enfoques para obtener una metodología rica, una calidad esencial, para nosotros, en estadística exploratoria ya que el hecho de disponer de varios puntos de vista refuerza la solidez de las conclusiones y permite escoger el más adaptado a un usuario dado (una partición es un instrumento grosero pero comunicable a un público sin cultura estadística). En este caso, utilizamos para cada método la misma distancia (euclidiana) entre individuos. Primeramente porque la elección de una distancia debe ser hecha previamente a los análisis ya que traduce la idea que se tiene de la similitud entre individuos. Y en segundo lugar, porque si queremos estudiar la influencia de la elección de una distancia, vale más hacerlo utilizando el mismo método de análisis, para evitar comparaciones poco sólidas.

4.7.1 Análisis factorial previo a una CJA

Sea una tabla X (de dimensión $I \times K$) en la que queremos clasificar sus filas (conjunto I). Realizamos el análisis factorial de X (ACP, AFC o ACM según la naturaleza de la tabla) y conservamos todos los factores (= coordenadas de las filas sobre los ejes factoriales ; en ACP estos factores se llaman componentes principales) de varianza no nula (en número S ; el factor de rango s es anotado F_s). Yuxtaponemos estos factores para constituir la tabla F (de dimensión $I \times S$). Las tablas X y F son equivalentes, es decir, definen la misma distancia entre los individuos. Además, la distancia utilizada a partir de las coordenadas incluidas en F es la distancia euclidiana usual, incluso si la distancia entre las filas de X no es la distancia euclidiana usual (por ejemplo la de χ^2 cuando la tabla X proviene del AFC). En efecto, los vectores u_s (asociados con F_s) que sirven para representar a los individuos (recordemos que las coordenadas están en los F_s) constituyen una base ortonormal (es por ello que los planos factoriales procedentes de un AFC son legibles, *i.e.*, se leen con la distancia euclidiana usual aunque inicialmente el espacio de las filas es dotado de la distancia de χ^2).
Como consecuencia, el programador encuentra facilidades y puede contentarse con escribir un solo programa de clasificación, tomando de entrada una tabla individuos $\times$ variables cuantitativas y la distancia usual entre individuos. La toma en cuenta de datos más variados se hace utilizando el análisis factorial adaptado antes (el AFC para una tabla de contingencia o una tabla de Burt o el ACM para una tabla individuos $\times$ variables cualitativas), como pretratamiento. Este encadenamiento de ambos procesos proporciona una posibilidad metodológica nueva : conservar para la CJA sólo una parte de S factores de inercia no nula. Para ello, se pueden realizar los dos razonamientos siguientes.

– Eliminar de la CJA, las únicas dimensiones de las que estamos (prácticamente) seguros que representan sólo «ruido», es decir, las últimas ; así conservaremos los factores que totalizarán un porcentaje muy elevado de la inercia, digamos 80 % o 90 % para fijar las ideas ; la jerarquía así obtenida debería ser más estable y más clara.
– Conservar para la CJA sólo los ejes que supimos interpretar, sea, en la práctica, un número pequeño (entre 2 y 5) ; la jerarquía así obtenida desempeña esencialmente el papel de ayuda a la interpretación del análisis factorial.

4.7.2 Análisis simultáneo de un plano factorial y de una jerarquía

Consiste simplemente en representar sobre el plano factorial los nudos más altos de la jerarquía como centro de gravedad de los individuos que reagrupan. Si se escogió una partición, nos limitamos al centro de gravedad de las clases de esta partición. En tal representación, la complementariedad entre ambos enfoques, aparece principalmente bajo dos aspectos :
– Disponemos a la vez de una visión continua (las «tendencias» materializadas por los ejes factoriales) y discontínua (las clases de la clasificación) del mismo conjunto de datos, todo ello en un marco único ;
– El plano factorial no proporciona ninguna información sobre la posición de los puntos en otras dimensiones ; las clases, establecidas a partir del conjunto de las dimensiones, aportan sobre el plano poca información «exterior al plano» : dos puntos próximos sobre el plano pudiendo estar en la misma clase (no demasiado alejados uno del otro a lo largo de otras dimensiones) o en dos clases diferentes (porque están alejados uno del otro a lo largo de otras dimensiones).

4.8 Ejemplo : datos sobre temperaturas

4.8.1 Descripción de los datos y problemática

Volvamos a utilizar el juego de datos sobre las temperaturas de las capitales europeas presentado en el capítulo del ACP (ver página 41). El objetivo es ahora reagrupar las capitales en grupos homogéneos de modo que las capitales del mismo grupo presentan temperaturas semejantes cada mes del año. Una vez estos grupos construidos, es importante caracterizar los grupos a partir de las variables o a partir de los individuos particulares. Con el fin de determinar en cuántos grupos debemos reagrupar las capitales, construimos en primer lugar una clasificación ascendente jerárquica.

4.8.2 Elección del análisis

La clasificación necesita escoger un indicio de agregación (escogemos aquí el indicio de agregación de Ward) así como una distancia entre individuos. La distancia euclidiana es adaptada pero es también necesario definir si hay que reducir o no las variables. Encontramos aquí el comentario efectuado en ACP (ver página 43) y escogemos trabajar con datos centrados-reducidos. Además, las distancias entre las capitales son definidas a partir de las doce variables de temperaturas mensuales únicamente, es decir, a partir de las variables escogidas como activas en el ACP.

Observación
Los individuos suplementarios (en el ejemplo, las ciudades que no son capitales) no son utilizados para construir las distancias entre individuos y no participan en el análisis.

Los dos primeros ejes del ACP realizado sobre las ciudades expresan más de 98 % de la información. Podemos conservar todas las dimensiones ya que esto no modifica la clasificación y permite descomponer la inercia total del ACP.

4.8.3 Puesta en marcha

Después de la importación de los datos, realizamos el ACP precisando que conservamos todos los ejes gracias al argumento `ncp=Inf` (`Inf` para infinito y el número de ejes conservados será igual al mínimo entre $I-1$ y K). Realizamos entonces la clasificación ascendente jerárquica a partir del objeto `res.pca` que contiene los resultados del ACP.

```
> library(FactoMineR)
> temperaturas <- read.table("http://factominer.free.fr/libra/temperaturas.csv",
    header=TRUE, sep=";", dec=".", row.names=1)
> res.pca <- PCA(temperaturas[1:23,], ncp=Inf, graph=FALSE,quanti.sup=13:16,quali.sup=17)
> res.hcpc <- HCPC(res.pca)
```

Observación
Anotemos que si se desea efectuar una clasificación jerárquica ascendente sobre un juego de datos brutos, es posible hacer un ACP no normado (con argumento `scale.unit=FALSE`) y conservar todos los ejes utilizando el argumento `ncp=Inf` (`Inf` para infinito). Es lo que se hace por defecto por la función **HCPC** cuando el objeto de entrada es una tabla de datos.

La forma del dendrograma (cf. figura 4.10) sugiere una partición de las capitales en tres grupos. El nivel óptimo de corte calculado por la función **HCPC** sugiere también tres grupos. Encontramos por ejemplo en el primer grupo las capitales más frías (las que tienen las coordenadas más débiles sobre el primer eje del análisis factorial). Así como se indica en la sección 4.1 y se representa en la figura 4.1, es posible permutar las ramas de cada nudo del árbol para ordenar los individuos en lo posible según el primer eje factorial. Es lo que se hace con el argumento `order=TRUE` (utilizado por defecto). Si queremos clasificar los individuos en función de otro criterio, hay que ordenar los individuos en la tabla en función de este criterio antes de hacer el ACP, y luego hacer la clasificación con el argumento `order=FALSE` en **HCPC**.
El objeto `call$t` contiene los resultados de la clasificación jerárquica ascendente. Particularmente :
– las salidas de la función de clasificación **agnes** (del paquete cluster) en `call$t$tree`
– el número de grupos «óptimo» calculado (`$call$t$nb.clust`) : este número es determinado entre el mínimo y el máximo de grupos definidos por el usuario y de modo que el cociente `$call$t$quot` sea lo más pequeño posible ;
– la inercia intra de la partición en n grupos (`$call$t$within`) ; para $n = 1$ grupo (la partición más grosera en un solo grupo) la inercia intra vale 12, para 2 grupos 5.237, etc.
– la ganancia de inercia inter cuándo se pasa de n grupos a $n + 1$ (`$call$t$inter`) ; para 2 grupos (*i.e.* Pasando de 1 a 2 grupos) la ganancia de inercia inter vale 6.763, para 3 grupos (*i.e.*, pasando de 2 a 3 grupos) la ganancia de inercia inter vale 2.356, etc.

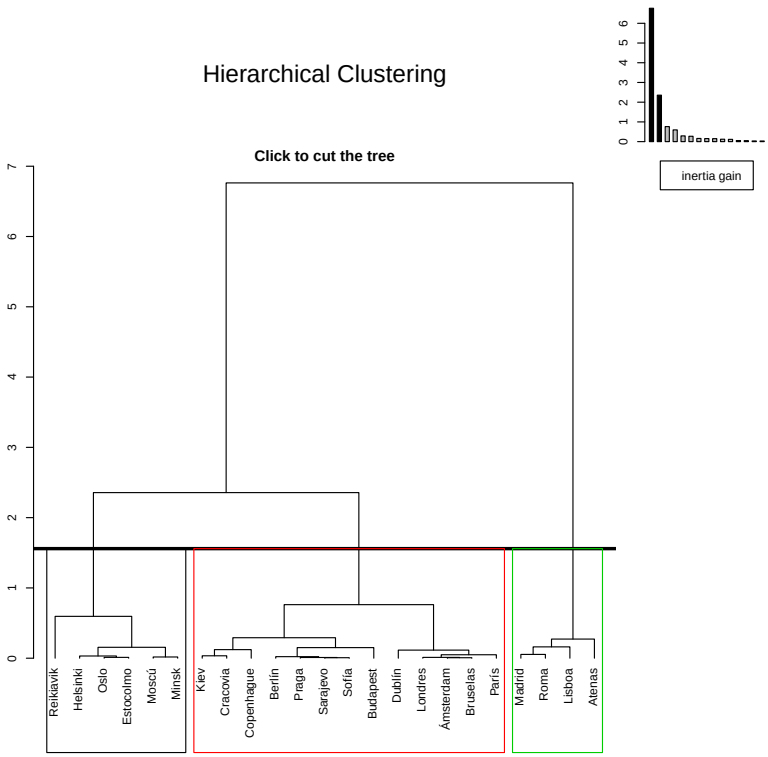

FIGURE 4.10 – Datos temperaturas : árbol jerárquico.

– el cociente de las dos inercias intra sucesivas ($0.550 = 2.881/5.237$).

```
$call$t$nb.clust
[1] 3

$call$t$within
 [1] 12.000  5.237  2.881  2.119  1.524  1.232  0.960  0.799  0.643  0.493
[11]  0.371  0.255  0.202  0.153  0.118  0.087  0.065  0.048  0.036  0.024
[21]  0.014  0.007  0.000

$call$t$inert.gain
 [1] 6.763 2.356 0.762 0.596 0.291 0.272 0.161 0.155 0.151 0.122 0.115 0.054
[13] 0.049 0.034 0.031 0.022 0.017 0.012 0.012 0.010 0.007 0.007

$call$t$quot
[1] 0.550 0.736 0.719 0.809 0.779 0.832 0.806 0.766
```

Para dibujar el árbol completo en tres dimensiones sobre el primer plano factorial (cf. figura 4.11), utilizamos el argumento `t.levels="all"` :

```
> res.hcpc <- HCPC(res.pca,t.levels="all")
```

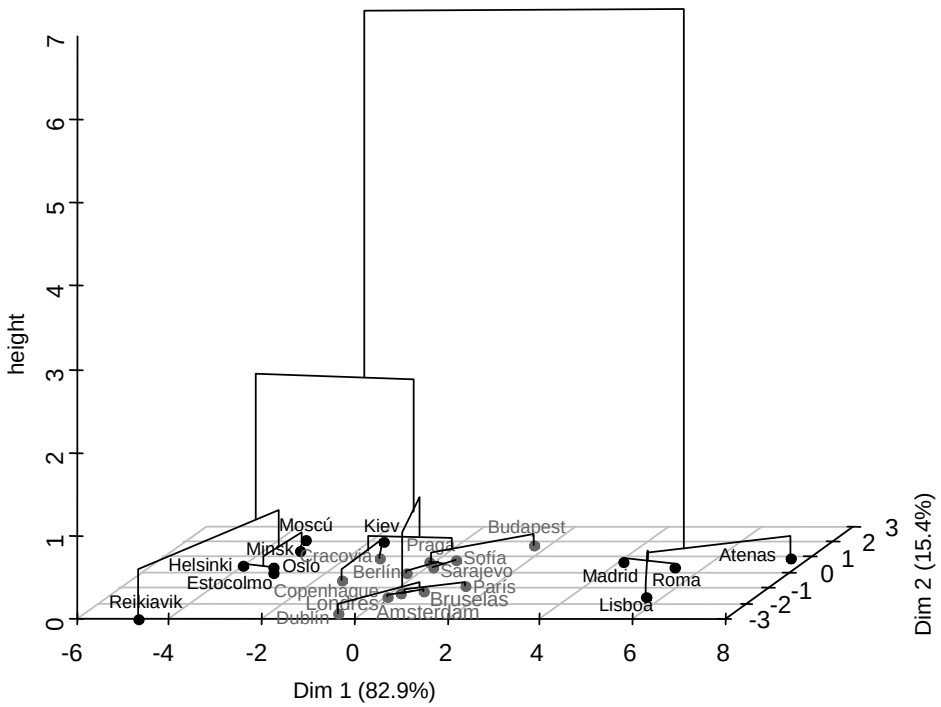

FIGURE 4.11 – Dendrograma en tres dimensiones sobre el primer plano factorial.

Descripción de los grupos Los grupos son descritos después y los resultados están en el objeto `desc.var`. Todas las variables del juego de datos iniciales son utilizadas, sean cuantitativas o cualitativas, activas o suplementarias. Para ello, la función devuelve los mismos resultados que la función **catdes** (cf. § 3.7.2). Estos resultados son reagrupados en la tabla 4.2. Ninguna variable caracteriza las ciudades del grupo 2. Las capitales de la clase 3 son características porque la temperatura media anual (15.7 grados) es más importante que para el conjunto de las capitales (9.37 grados). Este grupo es caracterizado por la modalidad *sur* de la variable cualitativa *Región* : hay más ciudades del sur en este grupo que en otros. En efecto, 80 % de las ciudades del sur pertenecen al grupo 3, 100 % de las ciudades del grupo 3 son ciudades del sur. Estos porcentajes son importantes ya que 21.7% de las ciudades están en el sur.

Los grupos pueden también ser descritos por los componentes principales. Para ello, una descripción idéntica a la realizada por las variables cuantitativas es efectuada a partir de las coordenadas de los individuos sobre los ejes factoriales. La tabla 4.3 muestra así como las capitales del grupo 1 (resp. 3) tienen una coordenada significativamente más débil (resp. fuerte) que otras sobre la primera dimensión. Las coordenadas sobre la tercera dimensión son más débiles para las capitales del grupo 2. Recordemos que la inercia explicada por el eje 3 es sólo 1 %, pues no iremos más lejos en el comentario de este resultado.

```
> res.hcpc$desc.var
$test.chi2
       p.value df
Región  0.0012  6

$category
$category$`1`
NULL
$category$`2`
NULL
$category$`3`
           Cla/Mod Mod/Cla Global p.value v.test
Región=Sur      80     100 21.739   0.001  3.256

$quanti
$quanti$`1`
$`1`
           v.test Mean in category Overall mean sd in category Overall sd  p.value
Latitud      2.78            56.13        49.88           5.85       6.98  0.00550
Amplitud     2.14            21.99        18.80           4.84       4.61  0.03200
Julio       -1.99            16.79        18.93           2.45       3.33  0.04600
Junio       -2.06            14.73        16.77           2.52       3.07  0.04000
Agosto      -2.48            15.49        18.30           2.26       3.53  0.01300
Mayo        -2.55            10.84        13.27           2.43       2.96  0.01100
Septiembre  -3.14            10.99        14.71           1.67       3.68  0.00170
Enero       -3.26            -5.14         0.17           2.63       5.07  0.00110
Diciembre   -3.27            -2.91         1.84           1.83       4.52  0.00110
Noviembre   -3.36             0.60         5.08           0.94       4.14  0.00078
Media       -3.37             5.50         9.37           0.77       3.56  0.00074
Abril       -3.39             4.67         8.38           1.55       3.40  0.00071
Febrero     -3.44            -4.60         0.96           2.34       5.01  0.00058
Octubre     -3.45             5.76        10.07           0.92       3.87  0.00055
Marzo       -3.68            -1.14         4.06           1.10       4.39  0.00024
$quanti$`2`
NULL
$quanti$`3`
           v.test Mean in category Overall mean sd in category Overall sd  p.value
Media        3.85            15.75         9.37           1.39       3.56  0.00012
Septiembre   3.81            21.23        14.71           1.54       3.68  0.00014
Octubre      3.72            16.75        10.07           1.91       3.87  0.00020
Agosto       3.71            24.38        18.30           1.88       3.53  0.00021
Noviembre    3.69            12.17         5.08           2.26       4.14  0.00022
Julio        3.60            24.50        18.93           2.09       3.33  0.00031
Abril        3.53            13.95         8.38           1.18       3.40  0.00041
Marzo        3.45            11.10         4.06           1.27       4.39  0.00056
Febrero      3.43             8.95         0.96           1.74       5.01  0.00059
Junio        3.39            21.60        16.77           1.86       3.07  0.00070
Diciembre    3.39             8.95         1.84           2.34       4.52  0.00071
Enero        3.29             7.92         0.17           2.08       5.07  0.00099
Mayo         3.18            17.65        13.27           1.55       2.96  0.00150
Latitud     -3.23            39.42        49.88           1.52       6.98  0.00130
```

Tabla 4.2 – Datos temperaturas : salida de la función **catdes** (cf. § 3.7.2) aplicada a la partición en tres grupos.

```
> res.hcpc$desc.axe
$quanti
$quanti$‘1‘
      v.test Mean in category Overall mean sd in category Overall sd    p.value
Dim.1  -3.32            -3.37     1.69e-16          0.849       3.15  0.0009087
$quanti$‘2‘
      v.test Mean in category Overall mean sd in category Overall sd    p.value
Dim.3  -2.41           -0.175    -4.05e-16          0.218      0.355  0.0157738
$quanti$‘3‘
      v.test Mean in category Overall mean sd in category Overall sd    p.value
Dim.1   3.86             5.66     1.69e-16           1.26       3.15 0.00011196
```

Tabla 4.3 – Descripción de los grupos (cf. § 3.7.2), procedentes de la clasificación, por los componentes principales.

Puede ser interesante ilustrar el grupo por individuos particulares de este grupo. Para ello, dos tipos de individuos particulares son propuestos : los modelos, es decir, los individuos más próximos del centro del grupo y los individuos específicos, es decir, los individuos más alejados de los centros de otros grupos. Para ello, el objeto `desc.ind$para` contiene los individuos ordenados por grupo y la distancia entre cada individuo y el centro de su grupo. El objeto `desc.ind$spec` contiene los individuos ordenados por grupo y la distancia entre cada individuo y el centro del grupo más próximo (cf. tabla 4.4). Así, Oslo es la ciudad que representa lo mejor posible las ciudades del grupo 1 mientras que Berlín y Roma son los modelos de los grupos 2 y 3. La ciudad de Reikiavik es específica del grupo 1, del que es la ciudad más alejada de los centros de los grupos 2 y 3 y que podemos considerar como la más particular del grupo 1. París y Atenas son específicas de los grupos 2 y 3.

4.9 Ejemplo : datos té

4.9.1 Descripción de los datos - problemática

Volvemos a examinar los datos sobre el consumo de té presentados en el capítulo de ACM página 119. El objetivo es ahora de proponer una clasificación de los 300 consumidores de té en algunos grupos correspondientes a perfiles distintos de consumo. Para el ACM, solamente las diecinueve preguntas que conciernen a la manera de cómo consumen el té han sido utilizadas como variables activas ; aquí todavía, únicamente estas variables servirán para la construcción de las grupos.

4.9.2 Construcción de la CJA

Al ser cualitativas las variables, el ACM efectuado antes de la clasificación permite utilizar las coordenadas factoriales como variables cuantitativas. Los últimos ejes del ACM son generalmente considerados como parásitos que es preferible suprimir para construir una clasificación más estable. Los primeros ejes son así conservados (aquí, escogemos 20 ejes que resumen 87 % de la inercia total). Después del ACM, construimos la clasificación ascendente jerárquica :

```
> res.hcpc$desc.ind
$para
cluster: 1
     Oslo  Helsinki Estocolmo     Minsk     Moscú
    0.339     0.884     0.922     0.965     1.770
------------------------------------------------
cluster: 2
   Berlín  Sarajevo  Bruselas     Praga Amsterdam
    0.576     0.716     1.040     1.060     1.120
------------------------------------------------
cluster: 3
     Roma    Lisboa    Madrid    Atenas
    0.36     1.74     1.84     2.17

$spec
cluster: 1
Reikiavik     Moscú  Helsinki     Minsk      Oslo
     5.47      4.34      4.28      3.74      3.48
------------------------------------------------
cluster: 2
    París  Budapest  Bruselas    Dublín Amsterdam
     4.38      4.37      4.35      4.28      4.08
------------------------------------------------
cluster: 3
  Atenas   Lisboa     Roma   Madrid
    7.67     5.66     5.35     4.22
```

Tabla 4.4 – Modelos e individuos específicos.

```
> library(FactoMineR)
> te <- read.table("http://factominer.free.fr/libra/te.csv",header=TRUE,sep=";")
> res.mca<-MCA(te, ncp=20, quanti.sup=22, quali.sup=c(19:21,23:36), graph=FALSE)
> res.hcpc <- HCPC(res.mca)
```

El aspecto del árbol jerárquico, igualmente que el diagrama de las inercias asociadas con los nudos, sugiere una partición en tres grupos (cf. figura 4.12).
Podemos colorear a continuación los individuos sobre el primer plano factorial en función del grupo al cual pertenecen (cf. figura 4.13).

```
> plot(res.hcpc,choice="map",ind.names=FALSE)
```

La inercia inter de la partición en dos grupos, 0.085 (primera parte de los resultados que figuran más abajo), es inferior al primer valor propio del ACM $\lambda_1 = 0.148$ (la segunda parte de los resultados de más abajo). Esto siempre se cumple y tiene la siguiente explicación : el eje factorial aporta matices comparado con la partición en dos grupos. Igualmente, el plano factorial expresa más inercia ($0.148 + 0.122 = 0.270$ que la partición en tres grupos ($0.085 + 0.069 = 0.154$). Es una ventaja cuando queremos resumir fácilmente la información, por ejemplo para una restitución de los resultados. El ACM será útil para una interpretación más fina de los resultados.

```
> round(res.hcpc$call$t$inert.gain,3)
 [1] 0.085 0.069 0.057 0.056 0.056 0.055 0.050
> round(res.mca$eig[,1],3)
 [1] 0.148 0.122 0.090 0.078 0.074 0.071 0.068
```

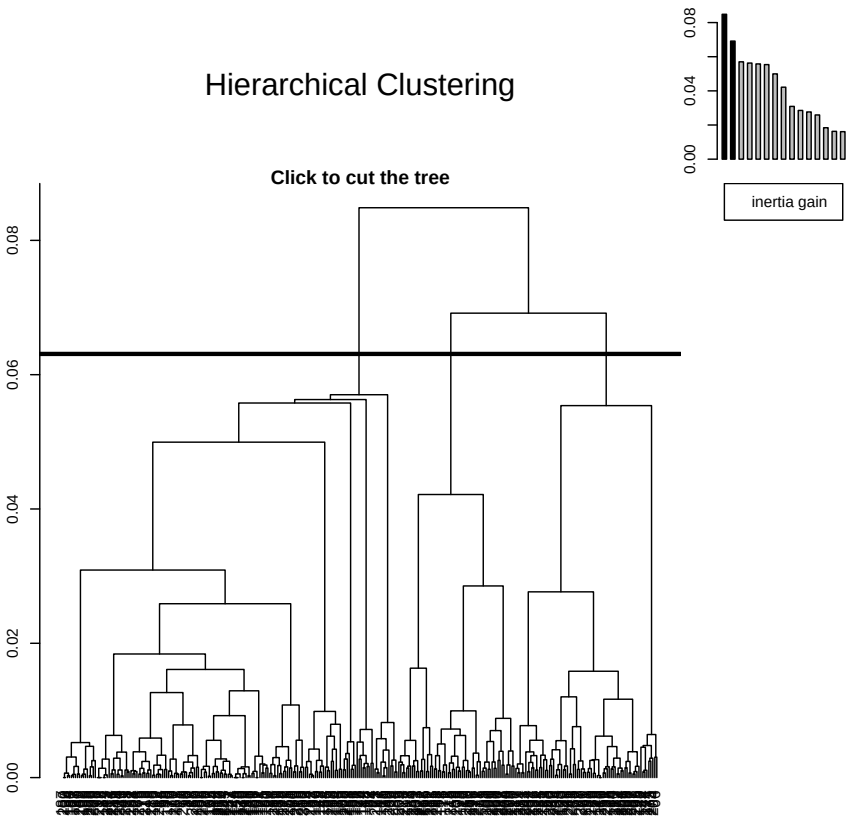

FIGURE 4.12 – Datos té : árbol jerárquico.

4.9.3 Descripción de los grupos

Para describir las características de los individuos de cada uno de los grupos, *i.e.*, su perfil de consumo de té, utilizamos la descripción de los grupos por las variables (objeto `res.hcpc$desc.var`, tabla 4.5) y por los ejes (objeto `res.hcpc$desc.axe`, tabla 4.7). La descripción de los grupos por individuos es menos interesante aquí porque los individuos no son conocidos y pueden servir de referencia. Las descripciones por las modalidades (tabla 4.6) son simplificadas únicamente conservando las modalidades sobreexpresadas asociadas a una probabilidad crítica inferior a 2 %.

Las variables *lugar de compra* y *forma* son las que mejor caracterizan la partición en tres grupos (probabilidades más débiles iguales a 8.47×10^{-79} y 3.14×10^{-47}, cf. tabla 4.5).

Si observamos más detalladamente, cada uno de los grupos es caracterizado por una modalidad de la variable *lugar de compra* y una modalidad de la variable *forma* : el primer grupo es caracterizado por los individuos que compran en *supermercado* su té en forma de *bolsita* : 85.9% de los individuos que compra en supermercado son en el grupo 1 y 93.8%

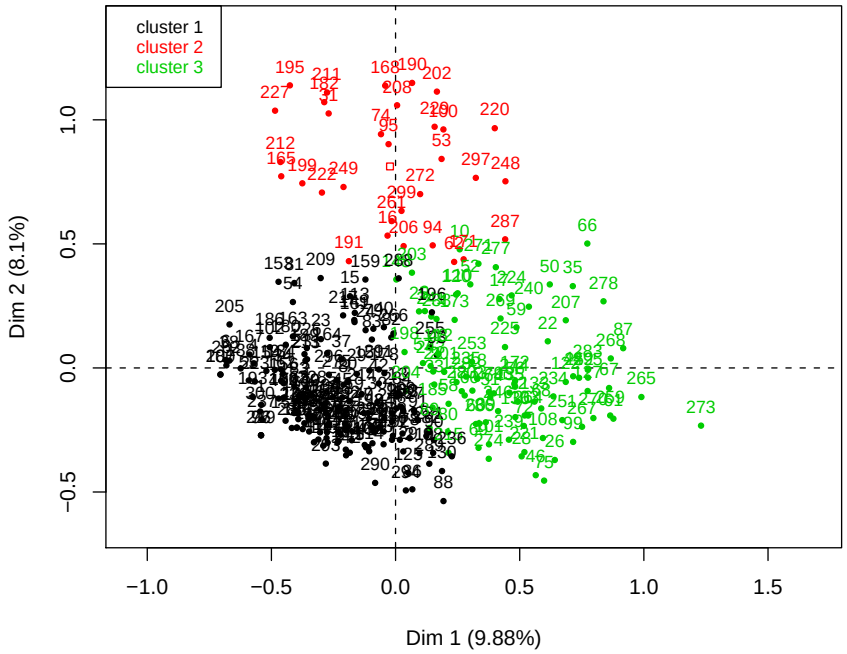

FIGURE 4.13 – Datos té : representación de la partición sobre el plano factorial.

de los individuos del grupo 1 compran en supermercado. Igualmente, el grupo 2 es caracterizado por los que compran en *tienda especializada* el té *a granel* mientras que el grupo 3 es caracterizado por los que compran en ambos tipos de tiendas (supermercado y tienda especializada) bajo ambas formas (bolsita y a granel). Otras variables y otras modalidades permiten caracterizar cada uno de los grupos pero de modo menos claro (probabilidad crítica más elevada).
La descripción de los grupos por los ejes factoriales (cf. tabla 4.7) muestra que los individuos del grupo 1 tienen coordenadas muy débiles sobre los ejes 1 y 2 (con relación a los individuos de otros grupos). Los individuos del grupo 2 tienen coordenadas fuertes sobre el eje 2 y los individuos del grupo 3 tienen coordenadas fuertes sobre el eje 1. Retenemos aquí las parejas grupo-eje que tienen un valor-test superior a 3 ya que los ejes sirvieron para construir los grupos.

4.10 Ejemplo : recorte en grupos de las variables cuantitativas

4.10.1 Recorte en grupos de una variable

En ciertos análisis, es recomendable transformar una variable cuantitativa en una variable cualitativa. Para ello, es necesario recortar la variable en grupos. La variable *edad* de los datos

```
> res.hcpc$desc.var$test.chi2
                          p.value df
lugar.de.compra          8.47e-79  4
forma                    3.14e-47  4
tipo                     1.86e-28 10
salón.de.té              9.62e-19  2
bar                      8.54e-10  2
amigos                   6.14e-08  2
restaurante              3.54e-07  2
cómo                     3.62e-06  6
variedad                 1.78e-03  4
sexo                     1.79e-03  2
frecuencia               1.97e-03  6
trabajo                  3.05e-03  2
merienda                 3.68e-03  2
después.almuerzo         1.05e-02  2
después.cena             2.23e-02  2
a.cada.momento.del.día 3.60e-02  2
azúcar                   3.69e-02  2
refinado                 4.08e-02  2
```

Tabla 4.5 – Datos té : descripción de la partición en tres grupos por las variables.

té (ver capítulo de ACM) ha sido declarada como cuantitativa en el cuestionario. Con el fin de poder poner en evidencia relaciones no lineales con esta variable, debe ser recodificada como cualitativa. Consideremos esta variable *edad* y transformémosla en variable cualitativa. Una primera estrategia es utilizar grupos «naturales» definidos *a priori* (por ejemplo, menos de 18 años, 18-30 años, etc.). La segunda estrategia es construir grupos equiprobables. Escogemos entonces un número de grupos *a priori*, generalmente entre 4 y 7, para tener suficientes grupos pero no demasiados :

```
> te <- read.table("http://factominer.free.fr/libra/te.csv",header=TRUE,sep=";")
> n.grupos <- 4
> grupos <- quantile(te[,22], seq(0,1,1/n.grupos))
> Xqual <- cut(te[,22],grupos, include.lowest=TRUE)
> summary(Xqual)
[15,23] (23,32] (32,48] (48,90]
     86      66      74      74
```

Una tercera estrategia es la de elegir el número de grupos y de sus límites a partir de los datos, *i.e.*, del histograma (cf. figura 4.14) que representa la distribución de la variable con la finalidad de definir los niveles de corte :

```
> hist(te$edad,col="grey",main="Histograma de la variable edad",
    freq=FALSE, xlab="edad", nclass=15)
```

La elección del recorte no es inmediata y es posible utilizar la clasificación para escoger un número de grupos antes de definirlos por un método K-means, por ejemplo.
Las líneas de código siguientes construyen la clasificación y consolidan los resultados por el método de K-means (en práctica, el método K-means converge muy rápidamente cuando se pone en práctica sobre una sola variable) :

```
> res.hcpc$desc.var$category
$category$‘1‘
                                              Cla/Mod Mod/Cla Global  p.value v.test
lugar.de.compra=supermercado                    85.90   93.80  64.00 4.11e-40  13.30
forma=bolsita                                   84.10   81.20  56.70 2.78e-25  10.40
salón.de.té=No.salón de té                      70.70   97.20  80.70 2.09e-18   8.75
tipo=té_marca_conocida                          83.20   44.90  31.70 2.78e-09   5.94
bar=No.bar                                      67.10   90.30  79.00 2.13e-08   5.60
amigos=No.amigos                                76.90   45.50  34.70 3.42e-06   4.64
restaurante=No.restaurante                      64.70   81.20  73.70 6.66e-04   3.40
tipo=té_MDD                                     90.50   10.80   7.00 2.40e-03   3.04
merienda=No.merienda                            67.90   50.60  43.70 5.69e-03   2.77
cómo=puro                                       64.10   71.00  65.00 1.32e-02   2.48
trabajo=No.trabajo                              63.40   76.70  71.00 1.41e-02   2.46
azúcar=azúcar                                   66.20   54.50  48.30 1.42e-02   2.45
a.cada.momento.del.día=No.a cada momento del día 64.00  71.60  65.70 1.45e-02   2.45
frecuencia=1 a 2/semana                         75.00   18.80  14.70 2.39e-02   2.26
frecuencia=1/día                                68.40   36.90  31.70 2.61e-02   2.22
tipo=té_desconocido                             91.70    6.25   4.00 2.84e-02   2.19
edad_cual=15-24                                 68.50   35.80  30.70 2.90e-02   2.18
después.almuerzo=No.después.almuerzo            61.30   89.20  85.30 3.76e-02   2.08
tipo=té_gama_baja                              100.00    3.98   2.33 4.55e-02   2.00

$category$‘2‘
                                              Cla/Mod Mod/Cla Global  p.value v.test
lugar.de.compra=tienda especializada            90.00   84.40   10.0 7.39e-30  11.40
forma=a granel                                  66.70   75.00   12.0 1.05e-19   9.08
tipo=té_gama_alta                               49.10   81.20   17.7 4.67e-17   8.39
variedad=verde                                  27.30   28.10   11.0 7.30e-03   2.68
refinado=refinado                               13.50   90.60   71.7 1.34e-02   2.47
sexo=H                                          16.40   62.50   40.7 1.43e-02   2.45
restaurante=No.restaurante                      13.10   90.60   73.7 2.59e-02   2.23
después.cena=después.cena                       28.60   18.80    7.0 3.10e-02   2.16
evasión.exotismo=No.evasión-exotismo            14.60   71.90   52.7 3.23e-02   2.14

$category$‘3‘
                                              Cla/Mod Mod/Cla Global  p.value v.test
lugar.de.compra=supermercado+tienda.especializada. 85.90 72.80  26.0 1.12e-33  12.10
forma=bolsita+a granel                          67.00   68.50   31.3 2.56e-19   8.99
salón.de.té=salón de té                         77.60   48.90   19.3 2.35e-16   8.20
bar=bar                                         63.50   43.50   21.0 1.95e-09   6.00
amigos=amigos                                   41.80   89.10   65.3 2.50e-09   5.96
tipo=té_variable                                51.80   63.00   37.3 2.63e-09   5.95
restaurante=restaurante                         54.40   46.70   26.3 3.92e-07   5.07
cómo=otro                                      100.00    9.78    3.0 3.62e-05   4.13
frecuencia=+ de 2/día                           41.70   57.60   42.3 6.13e-04   3.43
merienda=merienda                               38.50   70.70   56.3 1.22e-03   3.23
trabajo=trabajo                                 44.80   42.40   29.0 1.32e-03   3.21
sexo=F                                          37.10   71.70   59.3 4.90e-03   2.81
después.almuerzo=después.almuerzo               50.00   23.90   14.7 5.84e-03   2.76
cómo=limón                                      51.50   18.50   11.0 1.32e-02   2.48
azúcar=No.azúcar                                36.10   60.90   51.7 4.54e-02   2.00
```

Tabla 4.6 – Datos té : descripción de la partición en tres grupos por las modalidades (salida de la función **catdes** cf. § 3.7.2).

```
> res.hcpc$desc.axe
$quanti
$quanti$‘1‘
      v.test Mean in category Overall mean sd in category Overall sd  p.value
Dim.2  -7.80          -0.1320     4.93e-17          0.181      0.349 6.36e-15
Dim.1 -12.40          -0.2320    -2.00e-17          0.214      0.385 2.31e-35

$quanti$‘2‘
       v.test Mean in category Overall mean sd in category Overall sd  p.value
Dim.2   13.90           0.8120     4.93e-17          0.234      0.349 4.91e-44
Dim.4    4.35           0.2030    -3.35e-17          0.370      0.279 1.36e-05

$quanti$‘3‘
      v.test Mean in category Overall mean sd in category Overall sd  p.value
Dim.1  13.50           0.4520    -2.00e-17          0.252      0.385 1.89e-41
Dim.4  -4.73          -0.1150    -3.35e-17          0.292      0.279 2.30e-06
```

Tabla 4.7 – Datos té : descripción de la partición en tres grupos por los ejes factoriales.

```
> vari <- te[,22]
> res.hcpc <- HCPC(vari, iter.max=10)
```

Por defecto, la función **HCPC** construye un árbol jerárquico, la opción indicada aquí `iter.max=10` conlleva la ejecución de la agregación alrededor de los centros móviles. El árbol jerárquico (cf. figura 4.15) sugiere recortar la variable en cuatro grupos. Este árbol es construido en función de los valores de la variable *edad* sobre el eje de abscisas.
Después, podemos construir una nueva variable cualitativa `aa.cuali` de la manera siguiente :

```
> max.cla = unlist(by(res.hcpc$data.clust[,1],res.hcpc$data.clust[,2],max))
> breaks=c(min(vari),max.cla)
> aa.cuali = cut(vari, breaks, include.lowest=TRUE)
> summary(aa.cuali)
[15,28] (28,42] (42,57] (57,90]
    130      68      64      38
```

Este recorte parece de mejor calidad que el recorte en grupos equiprobables construido anteriormente ya que la clasificación jerárquica permitió detectar los «huecos» en la distribución (cf. el histograma de la figura 4.14).

4.10.2 Recorte automático de varias variables

Si queremos recortar en grupos numerosas variables cuantitativas, es fastidioso determinar el número de grupos que hay que escoger variable por variable, a partir del árbol jerárquico. Podemos entonces utilizar la función **HCPC** y tomar el número de grupos óptimo determinado por la función. Las líneas de código siguientes permiten recortar en grupos todas las variables cuantitativas del juego de datos `datos` :

```
> datos.cuali <- datos
> for (i in 1:ncol(datos.cuali)){
+   vari = datos.cuali[,i]
+   res.hcpc=HCPC(vari, nb.clust=-1, graph=FALSE)
+   maxi = unlist(by(res.hcpc$data.clust[,1], res.hcpc$data.clust[,2],max))
```

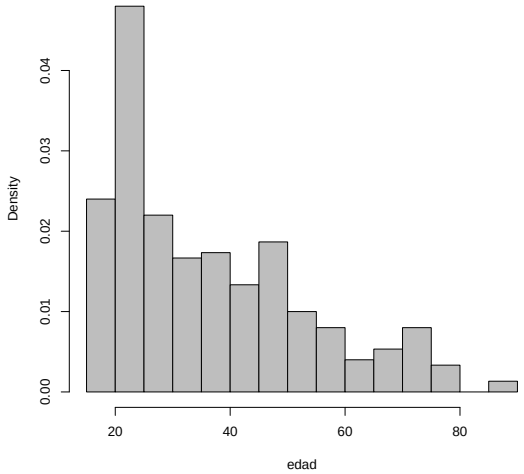

FIGURE 4.14 – Datos té : histograma de la variable edad.

```
+   breaks=c(min(vari),maxi)
+   aa.cuali = cut(vari, breaks, include.lowest=TRUE)
+   datos.cuali[,i] = aa.cuali
+ }
```

La tabla `datos.cuali` así creada contiene únicamente variables cualitativas que corresponden al recorte en grupos de cada una de las variables cuantitativas de la tabla inicial `datos`.

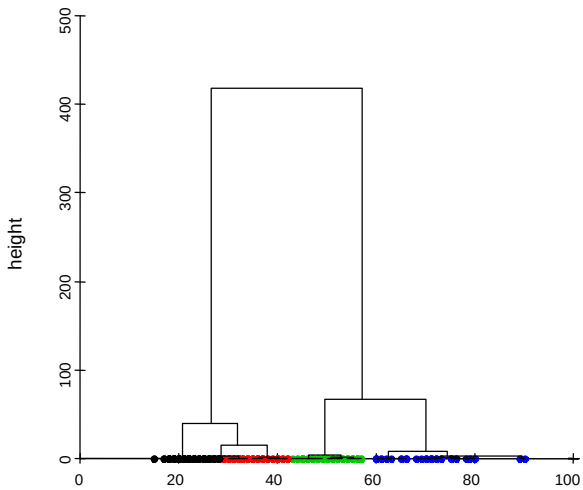

FIGURE 4.15 – Datos té : dendrograma de la variable edad.

Annexe A

A.1 Porcentaje de inercia explicado por un eje y por un plano

Nos interesamos aquí en testar el porcentaje de inercia explicado por un eje y luego el porcentaje de inercia explicado por el primer plano. Para ello, simulamos 10 000 juegos de datos para un número I de individuos y un número K de variables independientes que siguen una ley normal. Efectuamos a continuación un ACP normado (variables estandarizadas) por juego de datos y calculamos el porcentaje de inercia explicado por un eje y el porcentaje de inercia explicado por un plano. En las tablas A.1 y A.2 (resp. A.3 y A.4 damos el cuantil 0.95 de los 10 000 porcentajes de inercia del primer eje (resp. del primer plano) obtenido para una dimensión de tabla dada (I y K).
Así, comparar un porcentaje de inercia de un eje o de un plano con el valor asociado en la tabla corresponde a realizar un test de la hipótesis H_0 : el porcentaje de inercia explicado por el primer eje (resp. por el primer plano) no es significativamente superior al obtenido con variables (normales) independientes.

	Número de variables											
nbind	4	5	6	7	8	9	10	11	12	13	14	15	16
5	72.6	67.6	63.3	60.4	57.9	55.5	53.9	52.6	51.3	50.1	49.1	48.4	47.5
6	67.6	61.8	57.6	54.7	52.4	50.4	48.7	46.9	45.8	44.6	43.6	42.9	42.0
7	64.0	58.3	54.0	50.9	48.3	46.1	44.5	42.9	41.8	40.4	39.8	38.8	38.1
8	60.7	54.9	50.7	47.7	45.2	43.1	41.3	40.1	38.7	37.4	36.5	35.9	35.0
9	58.6	52.3	48.7	45.0	42.7	40.8	39.1	37.7	36.3	35.2	34.3	33.5	32.5
10	56.8	50.5	46.4	43.5	40.7	38.6	36.9	35.7	34.4	33.4	32.1	31.5	30.7
11	55.0	48.8	44.6	41.6	39.0	37.2	35.4	33.9	32.8	31.7	30.8	29.7	29.1
12	53.3	47.5	43.2	40.1	37.7	35.6	34.1	32.5	31.5	30.3	29.4	28.6	27.9
13	52.0	46.2	41.8	39.0	36.4	34.5	32.9	31.3	30.2	29.1	28.2	27.4	26.7
14	51.0	45.2	40.9	37.8	35.5	33.3	31.7	30.3	29.0	28.1	27.2	26.4	25.6
15	50.1	44.1	40.0	36.8	34.4	32.4	30.8	29.4	28.3	27.3	26.5	25.5	24.7
16	49.3	43.2	39.2	36.0	33.7	31.6	29.9	28.7	27.4	26.5	25.5	24.7	24.0
17	48.4	42.3	38.3	35.2	32.9	31.0	29.2	27.9	26.7	25.7	24.9	24.0	23.3
18	47.6	41.8	37.6	34.5	32.2	30.2	28.7	27.1	26.0	25.1	24.2	23.4	22.7
19	46.9	41.1	36.8	33.9	31.5	29.7	28.0	26.6	25.6	24.5	23.5	22.8	22.1
20	46.1	40.5	36.3	33.5	30.9	29.0	27.4	26.1	25.0	24.0	23.0	22.3	21.6
25	44.0	38.1	33.9	31.0	28.6	26.9	25.2	23.8	22.8	21.9	21.0	20.3	19.6
30	41.9	36.4	32.4	29.4	27.1	25.1	23.6	22.4	21.3	20.3	19.5	18.8	18.1
35	40.7	35.0	31.0	28.1	25.9	23.9	22.5	21.2	20.1	19.2	18.4	17.7	17.0
40	39.7	34.0	30.1	27.1	24.7	23.0	21.6	20.3	19.3	18.3	17.5	16.8	16.2
45	38.8	33.0	29.1	26.3	24.0	22.3	20.8	19.6	18.5	17.6	16.8	16.1	15.5
50	38.0	32.4	28.5	25.6	23.4	21.6	20.1	18.9	17.9	17.0	16.2	15.6	15.0
100	34.1	28.5	24.8	21.9	19.9	18.2	16.9	15.7	14.7	14.0	13.2	12.6	12.0

Tabla A.1 – Cuantil a 95 % del porcentaje de inercia explicado por el primer eje de 10 000 ACP efectuados sobre tablas constituidas por variables independientes (el número de individuos varía de 5 a 100 y el número de variables de 4 a 16) : por ejemplo, para una tabla con $I = 30$ individuos y $K = 10$ variables, 95 % de los porcentajes de inercia explicado por el primer eje son inferiores a 23.6%.

	Número de variables											
nbind	17	18	19	20	25	30	35	40	50	75	100	150	200
5	46.9	46.2	45.5	45.0	42.9	41.3	39.8	39.0	37.3	35.0	33.6	32.0	31.0
6	41.1	40.7	40.1	39.5	37.4	35.6	34.5	33.5	31.8	29.5	28.2	26.6	25.7
7	37.2	36.7	36.0	35.6	33.5	31.8	30.4	29.6	28.1	25.8	24.5	23.0	22.1
8	34.4	33.7	33.1	32.6	30.4	28.8	27.6	26.7	25.2	23.1	21.8	20.4	19.5
9	32.1	31.3	30.8	30.2	28.0	26.5	25.4	24.4	23.0	21.0	19.7	18.3	17.5
10	30.0	29.5	28.8	28.4	26.2	24.6	23.6	22.7	21.4	19.3	18.1	16.7	15.9
11	28.5	27.8	27.3	26.8	24.7	23.3	22.1	21.3	19.9	17.9	16.8	15.4	14.6
12	27.1	26.5	25.9	25.5	23.5	22.0	20.9	20.0	18.7	16.7	15.6	14.3	13.6
13	26.0	25.3	24.9	24.2	22.3	20.9	19.8	19.0	17.7	15.7	14.7	13.4	12.7
14	25.0	24.4	23.9	23.4	21.3	20.0	18.9	18.1	16.8	14.9	13.9	12.6	11.9
15	24.1	23.5	23.0	22.5	20.7	19.2	18.1	17.3	16.1	14.2	13.2	12.0	11.2
16	23.5	22.9	22.3	21.7	19.9	18.5	17.4	16.6	15.4	13.6	12.5	11.3	10.7
17	22.7	22.2	21.6	21.1	19.2	17.8	16.8	16.0	14.8	13.0	12.0	10.8	10.1
18	22.1	21.5	21.0	20.4	18.6	17.2	16.3	15.4	14.2	12.5	11.5	10.3	9.7
19	21.4	20.9	20.4	19.9	18.0	16.7	15.8	14.9	13.8	12.1	11.1	9.9	9.3
20	21.0	20.4	20.0	19.4	17.6	16.3	15.3	14.5	13.3	11.6	10.6	9.5	8.9
25	19.0	18.4	17.9	17.4	15.7	14.5	13.5	12.8	11.7	10.0	9.1	8.1	7.5
30	17.5	17.0	16.6	16.1	14.4	13.2	12.3	11.5	10.5	8.9	8.1	7.1	6.5
35	16.5	16.0	15.5	15.1	13.4	12.2	11.3	10.6	9.6	8.1	7.3	6.4	5.8
40	15.6	15.2	14.7	14.2	12.6	11.5	10.6	10.0	8.9	7.5	6.7	5.8	5.3
45	14.9	14.4	14.0	13.6	12.0	10.9	10.0	9.4	8.4	7.0	6.3	5.4	4.9
50	14.4	13.9	13.5	13.1	11.5	10.4	9.6	9.0	8.0	6.6	5.9	5.0	4.6
100	11.6	11.1	10.7	10.3	8.9	7.9	7.2	6.6	5.8	4.7	4.0	3.3	2.9

Tabla A.2 – Cuantil a 95 % del porcentaje de inercia explicado por el primer eje de 10 000 ACP efectuados sobre tablas constituidas por variables independientes (el número de individuos varía de 5 a 100 y el número de variables de 17 a 200) : por ejemplo, para una tabla con $I = 50$ individuos y $K = 30$ variables, 95 % de los porcentajes de inercia explicado por el primer eje son inferiores a 10.4%.

	Número de variables											
nbind	4	5	6	7	8	9	10	11	12	13	14	15	16
5	96.5	93.1	90.2	87.6	85.5	83.4	81.9	80.7	79.4	78.1	77.4	76.6	75.5
6	93.3	88.6	84.8	81.5	79.1	76.9	75.1	73.2	72.2	70.8	69.8	68.7	68.0
7	90.5	84.9	80.9	77.4	74.4	72.0	70.1	68.3	67.0	65.3	64.3	63.2	62.2
8	88.1	82.3	77.2	73.8	70.7	68.2	66.1	64.0	62.8	61.2	60.0	59.0	58.0
9	86.1	79.5	74.8	70.7	67.4	65.1	62.9	61.1	59.4	57.9	56.5	55.4	54.3
10	84.5	77.5	72.3	68.2	65.0	62.4	60.1	58.3	56.5	55.1	53.7	52.5	51.5
11	82.8	75.7	70.3	66.3	62.9	60.1	58.0	56.0	54.4	52.7	51.3	50.1	49.2
12	81.5	74.0	68.6	64.4	61.2	58.3	55.8	54.0	52.4	50.9	49.3	48.2	47.2
13	80.0	72.5	67.2	62.9	59.4	56.7	54.4	52.2	50.5	48.9	47.7	46.6	45.4
14	79.0	71.5	65.7	61.5	58.1	55.1	52.8	50.8	49.0	47.5	46.2	45.0	44.0
15	78.1	70.3	64.6	60.3	57.0	53.9	51.5	49.4	47.8	46.1	44.9	43.6	42.5
16	77.3	69.4	63.5	59.2	55.6	52.9	50.3	48.3	46.6	45.2	43.6	42.4	41.4
17	76.5	68.4	62.6	58.2	54.7	51.8	49.3	47.1	45.5	44.0	42.6	41.4	40.3
18	75.5	67.6	61.8	57.1	53.7	50.8	48.4	46.3	44.6	43.0	41.6	40.4	39.3
19	75.1	67.0	60.9	56.5	52.8	49.9	47.4	45.5	43.7	42.1	40.7	39.6	38.4
20	74.1	66.1	60.1	55.6	52.1	49.1	46.6	44.7	42.9	41.3	39.8	38.7	37.5
25	72.0	63.3	57.1	52.5	48.9	46.0	43.4	41.4	39.6	38.1	36.7	35.5	34.5
30	69.8	61.1	55.1	50.3	46.7	43.6	41.1	39.1	37.3	35.7	34.4	33.2	32.1
35	68.5	59.6	53.3	48.6	44.9	41.9	39.5	37.4	35.6	34.0	32.7	31.6	30.4
40	67.5	58.3	52.0	47.3	43.4	40.5	38.0	36.0	34.1	32.7	31.3	30.1	29.1
45	66.4	57.1	50.8	46.1	42.4	39.3	36.9	34.8	33.1	31.5	30.2	29.0	27.9
50	65.6	56.3	49.9	45.2	41.4	38.4	35.9	33.9	32.1	30.5	29.2	28.1	27.0
100	60.9	51.4	44.9	40.0	36.3	33.3	31.0	28.9	27.2	25.8	24.5	23.3	22.3

Tabla A.3 – Cuantil a 95 % del porcentaje de inercia explicado por el primer plano de 10 000 ACP efectuados sobre tablas constituidas por variables independientes (el número de individuos varía de 5 a 100 y el número de variables de 4 a 16) : por ejemplo, para una tabla con $I = 30$ individuos y $K = 10$ variables, 95 % de los porcentajes de inercia explicado por el primer plano son inferiores a 41.1%.

	Número de variables											
nbind	17	18	19	20	25	30	35	40	50	75	100	150	200
5	74.9	74.2	73.5	72.8	70.7	68.8	67.4	66.4	64.7	62.0	60.5	58.5	57.4
6	67.0	66.3	65.6	64.9	62.3	60.4	58.9	57.6	55.8	52.9	51.0	49.0	47.8
7	61.3	60.7	59.7	59.1	56.4	54.3	52.6	51.4	49.5	46.4	44.6	42.4	41.2
8	57.0	56.2	55.4	54.5	51.8	49.7	47.8	46.7	44.6	41.6	39.8	37.6	36.4
9	53.6	52.5	51.8	51.2	48.1	45.9	44.4	42.9	41.0	38.0	36.1	34.0	32.7
10	50.6	49.8	49.0	48.3	45.2	42.9	41.4	40.1	38.0	35.0	33.2	31.0	29.8
11	48.1	47.2	46.5	45.8	42.8	40.6	39.0	37.7	35.6	32.6	30.8	28.7	27.5
12	46.2	45.2	44.4	43.8	40.7	38.5	36.9	35.5	33.5	30.5	28.8	26.7	25.5
13	44.4	43.4	42.8	41.9	39.0	36.8	35.1	33.9	31.8	28.8	27.1	25.0	23.9
14	42.9	42.0	41.3	40.4	37.4	35.2	33.6	32.3	30.4	27.4	25.7	23.6	22.4
15	41.6	40.7	39.8	39.1	36.2	34.0	32.4	31.1	29.0	26.0	24.3	22.4	21.2
16	40.4	39.5	38.7	37.9	35.0	32.8	31.1	29.8	27.9	24.9	23.2	21.2	20.1
17	39.4	38.5	37.6	36.9	33.8	31.7	30.1	28.8	26.8	23.9	22.2	20.3	19.2
18	38.3	37.4	36.7	35.8	32.9	30.7	29.1	27.8	25.9	22.9	21.3	19.4	18.3
19	37.4	36.5	35.8	34.9	32.0	29.9	28.3	27.0	25.1	22.2	20.5	18.6	17.5
20	36.7	35.8	34.9	34.2	31.3	29.1	27.5	26.2	24.3	21.4	19.8	18.0	16.9
25	33.5	32.5	31.8	31.1	28.1	26.0	24.5	23.3	21.4	18.6	17.0	15.2	14.2
30	31.2	30.3	29.5	28.8	26.0	23.9	22.3	21.1	19.3	16.6	15.1	13.4	12.5
35	29.5	28.6	27.9	27.1	24.3	22.2	20.7	19.6	17.8	15.2	13.7	12.1	11.1
40	28.1	27.3	26.5	25.8	23.0	21.0	19.5	18.4	16.6	14.1	12.7	11.1	10.2
45	27.0	26.1	25.4	24.7	21.9	20.0	18.5	17.4	15.7	13.2	11.8	10.3	9.4
50	26.1	25.3	24.6	23.8	21.1	19.1	17.7	16.6	14.9	12.5	11.1	9.6	8.7
100	21.5	20.7	19.9	19.3	16.7	14.9	13.6	12.5	11.0	8.9	7.7	6.4	5.7

Tabla A.4 – Cuantil a 95 % del porcentaje de inercia explicado por el primer plano de 10 000 ACP efectuados sobre tablas constituidas por variables independientes (el número de individuos varía de 5 a 100 y el número de variables de 17 a 200) : por ejemplo, para una tabla con $I = 50$ individuos y $K = 30$ variables, 95 % de los porcentajes de inercia explicado por el primer plano son inferiores a 19.1%.

A.2 El lenguaje de programación R

A.2.1 Presentación general

El lenguaje de programación R es gratuito y puede descargarse en la dirección siguiente : `http://cran.r-project.org/`. El objetivo no está aquí en detallar el conjunto de las funcionalidades del programa sino más bien de presentar brevemente cómo realizar el conjunto de los análisis efectuados en este libro. Para una presentación más detallada de R, podremos referirnos al libro de (Sánchez *et al.*, 2008).
Describimos primeramente un ejemplo detalladamente antes de listar algunas funciones útiles para importar los datos, hacer gráficos, etc. En § A.2.2 presentamos el paquete Rcmdr que permite realizar estos análisis a partir de un menú deslizante y en § A.2.3 presentamos más detalladamente el paquete FactoMineR dedicado al análisis de los datos y utilizado a lo largo de esta obra. Para comenzar, partamos del ejemplo de ACP sobre las temperaturas (cf. § 1.10) y comentemos las líneas de códigos siguientes :

```
1 > library(FactoMineR)
2 > temperaturas <- read.table("http://factominer.free.fr/libra/temperaturas.csv",
      header=TRUE, sep=";", dec=".", row.names=1)
3 > res <- PCA(temperaturas, ind.sup=24:35, quanti.sup=13:16, quali.sup=17)
4 > plot.PCA(res, choix="ind", habillage=17, cex=0.7,title="Mi ACP")
5 > graph.var(res, draw=c("var","Media"),label=c("Mayo","Media"))
6 > write.infile(res, file="c:/essai.csv", sep = ";")
```

1. Carga del paquete FactoMineR.
2. Importación del juego de datos : la tabla de datos está en el archivo `http://factominer.free.fr/libra/temperaturas.csv` ; `header=TRUE` la primera fila del archivo contiene el nombre de las variables ; `sep=" ;"` el separador de campos es el carácter " ;" (formato clásico de importación con los ficheros de tipo csv), `dec="."` el separador de decimal es "." ; `row.names=1` la primera columna contiene el nombre de los individuos.
3. Ejecución del ACP vía la función **PCA** : los individuos de 24 hasta 35 (`24:35`) son suplementarios, las variables de 13 a 16 son cuantitativas suplementarias y la variable 17 cualitativa suplementaria. Por defecto, la función centra y reduce las variables (el argumento `scale.unit=TRUE` es utilizado por defecto y no es necesario precisarlo.
4. La función **plot.PCA** es valiosa para mejorar los gráficos por defecto : aquí, coloreamos los individuos en función de las modalidades de la variable 17 (variable cualitativa suplementaria), disminuimos el tamaño de los caracteres (`cex=0.7` más bien que 1 por defecto) y damos un título a cada gráfico.
5. Construcción de un gráfico de las variables : la función **graph.var** permite escoger las variables que se quieren dibujar sobre el gráfico de las variables. Aquí, todas las variables activas son dibujadas así como *Media* ; sólo las etiquetas de las variables *Mayo* y *Media* están presentes.
6. Exportación de los resultados : la función **write.infile** permite escribir el conjunto de los resultados contenidos en el objeto `res` (aquí en el arvhivo `c:/essai.csv`).

Exportación de los gráficos. Los gráficos pueden exportarse bajo diferentes formatos (pdf, emf, eps, jpg, etc.). Para ello, hacer clic en el gráfico y hacer `Archivo` y luego `Guardar como`. Otra posibilidad es hacer clic con el botón derecho del ratón en el gráfico y de `Copiar como metafile`. El gráfico puede entonces ser pegado directamente en un editor (Word o Powerpoint por ejemplo). Es así posible disociar el gráfico y retocarlo para mejorar la legibilidad.
Selección de individuos y/o de variables en un análisis. Es muy fácil realizar un análisis con una parte del juego de datos. Las líneas siguientes permiten ejecutar un ACP sobre una parte de la tabla de datos (entre los `[ , ]` los individuos se precisan antes de la coma y las variables después) :

```
1 > res<-PCA(temperaturas[,1:12])
2 > res<-PCA(temperaturas[c(1:10,15:20),1:12])
3 > res<-PCA(temperaturas[-c(4:6,8,10),1:12])
```

1. Sobre el conjunto de los individuos pero únicamente con las variables de 1 a 12.
2. Sobre los individuos de 1 a 10 y de 15 a 20 pero únicamente con las variables de 1 a 12.
3. Sobre todos los individuos salvo los individuos de 4, 5, 6, 8 y 10 y con las variables de 1 a 12.

Las funciones de importación y de exportación

Función	Descripción
read.table	importa una tabla de datos de un archivo y crea un data-frame (tabla que puede contener variables cuantitativas y/o cualitativas y que contiene informaciones tales como el nombre de las filas y el nombre de las columnas)
read.csv	importa una tabla de datos de un archivo que tiene una extensión csv y crea un data-frame
write.table	escribe una tabla en un archivo
write.infile	función del paquete FactoMineR que escribe todos los elementos de una lista en un archivo csv
save	salva objetos R en un archivo `.Rdata`
load	recupera los objetos guardados con la función `save`
history	recupera las últimas líneas de códigos ejecutados
save.history	salva el historial de las últimas líneas de código ejecutados

Las funciones de gestión de datos

Función	Descripción
cbind.data.frame	yuxtapone los data-frames en columnas (pega las columnas unas al lado de otras)

Función	Descripción
rbind.data.frame	yuxtapone los data-frames en filas ; los nombres de columnas de data-frames deben ser idénticos (pega las filas una debajo de la otra, las columnas son clasificadas en el mismo orden para todas las tablas con el fin de poner las variables en correspondencia antes de la concatenación)
sort	clasifica un vector por orden creciente (decreciendo si `decreasing = TRUE`)
order	clasifica una tabla en función de una o varias columnas (o filas) : `x[order(x[,3], -x[,6]), ]` clasifica la tabla `x` en función (creciente) de la tercera columna de `x` luego, en caso de igualdad en la tercera columna de `x`, en función (decreciente) de la sexta columna de `x`
dimnames	da los nombres de las dimensiones de un objeto (lista, matriz, data-frame, etc.)
rownames	da los nombres de las filas de una matriz o de un data-frame
colnames	da los nombres de las columnas de una matriz o de un data.frame
dim	da las dimensiones de un objeto
nrow	da el número de filas de una tabla
ncol	da el número de columnas de una tabla
factor	define un vector como un factor, *i.e.*, una variable cualitativa (si `ordered=TRUE` los niveles de los factores son considerados como ordenados)
levels	da las modalidades de una variable cualitativa (niveles de un factor)
nlevels	da el número de modalidad de una variable cualitativa
which	da las posiciones de los valores verdaderos de un vector o de una tabla lógica : el parámetro `arr.ind=TRUE` permite devolver los números de filas y de columnas de la tabla : **which**`(c(1,4,3,2,5,3) == 3)` devuelve 3 6 ; **which**`(matrix(1:12,nrow=4) ==3,arr.ind=TRUE)` devuelve (fila 3, columna 1)
is.na	comprueba si el dato está ausente

Las funciones estadísticas de base

Las funciones estadísticas siguientes permiten describir una variable cuantitativa `x`. Para el conjunto de estas funciones, el parámetro `na.rm=TRUE` permite eliminar los datos ausentes antes del cálculo. Si `na.rm=FALSE` y hay datos ausentes, entonces la función devuelve un mensaje de error.

Función	Descripción
mean`(x, na.rm=TRUE)`	media de `x` calculada sobre los datos presentes

Función	Descripción
sd(x)	desviación-tipo de x
var(x)	varianza de x, si x es un vector, o una matriz de varianza-covarianza, si x es una matriz (varianza sin sesgo)
cor(x)	matriz de correlación de x
quantile(x, probs)	cuantiles de x de orden probs
sum(x)	suma de los elementos de x
min(x)	mínimo de x
max(x)	máximo de x
scale(x, center=TRUE, scale=TRUE)	centra (center=TRUE) y reduce (scale=TRUE) x
colMeans(x)	calcula la media de cada columna de la tabla x
rowMeans(x)	calcula la media de cada fila de la tabla x
apply(x,MARGIN, FUN)	aplica la función FUN sobre las filas o sobre las columnas de la tabla x : **apply**(x, 2, mean) calcula las medias de cada columna de x ; **apply**(x, 1, sum) calcula las sumas de cada fila dex

Las funciones del análisis factorial

Función	Descripción
PCA	análisis en componentes principales con posibilidad de tener individuos suplementarios, variables cuantitativas y cualitativas suplementarias
CA	análisis factorial de correspondencias con posibilidad de tener filas y columnas suplementarias
MCA	análisis de correspondencias múltiples con posibilidad de tener individuos suplementarios, variables cuantitativas y cualitativas suplementarias
dimdesc	describe los ejes factoriales
catdes	describe una variable cualitativa en función de las variables cuantitativas y/o cualitativas
condes	describe una variable cuantitativa en función de las variables cuantitativas y/o cualitativas
HCPC	clasificación ascendente jerárquica sobre componentes principales
graph.var	dibuja el gráfico de las variables a partir de ciertas variables únicamente

Las funciones gráficas

Función	Descripción
x11()	crea una nueva ventana gráfica vacía
pdf, **postscript**, **jpeg**, **png**, **bmp**	guarda un gráfico en el formato pdf, postscript, jpeg, png, bmp ; todas las funciones se utilizan de la misma manera : **pdf("mongraphe.pdf")** ; *orden gráfico* ; **dev.off()**

Las funciones print y plot

Las funciones **print** y **plot** son funciones genéricas, es decir, funciones que dan resultados específicos según la clase de objeto al que son aplicadas.

Función	Descripción
print	escribe los resultados (el conjunto de los resultados o un extracto)
plot	construye un gráfico

Por ejemplo **print.PCA**, **print.CA**, **print.MCA**, pueden ser llamadas por la instrucción genérica **print**. Según la clase del objeto (resultados procedentes de un ACP, un AFC, un ACM), las salidas o los gráficos serán específicos. Para tener una ayuda sobre la función que escribe un objeto `PCA`, por ejemplo : `help ("print.PCA")`.

A.2.2 Paquete Rcmdr

El interfaz gráfico R Commander está disponible en el paquete Rcmdr. Este interfaz permite utilizar R con la ayuda de un menú deslizante de modo ameno. El interés de este paquete es también pedagógico ya que proporciona las líneas de código correspondientes a los análisis efectuados : nos familiarizamos así con la programación viendo las funciones empleadas. El interfaz Rcmdr no contiene ni todas las funciones disponibles bajo R, ni todas las opciones de las diferentes funciones pero las funciones más corrientes son programadas y las opciones más clásicas disponibles.
Como todo paquete, debe ser instalado una sola vez y luego cargado a cada utilización por :

```
> library(Rcmdr)
```

El interfaz (cf. figura A.1) se abre automáticamente. Este interfaz posee un menú deslizante, una ventana de guión y una ventana de salida. Cuando el menú deslizante es utilizado, el análisis es lanzado y las líneas de código que sirvieron para generar el análisis son escritas en la ventana de guión.
Para importar los datos con Rcmdr, lo mas simple es tener un archivo Excel :
`Datos` → `Importar datos` → `Desde conjunto de datos Excel`
Con un archivo en el formato txt o csv :
`Datos` → `Importar datos` → `Desde archivo de texto portapapeles o URL`
A continuación hay que precisar el separador de columnas (separador de campos) y el separador de decimales (un "." o una ",").
Para verificar que el juego de datos ha sido bien importado :
`Estadísticos` → `Resúmenes` → `Conjunto de datos activo`

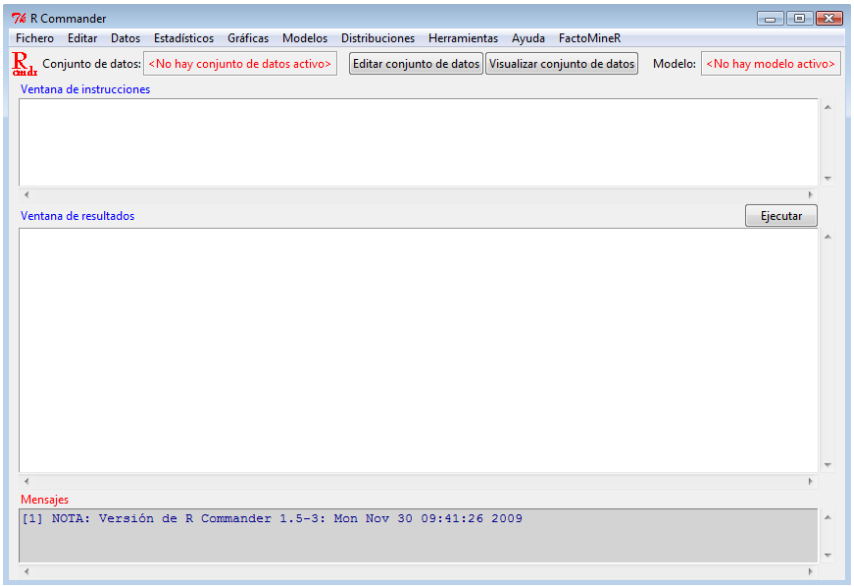

FIGURE A.1 – Ventana principal de Rcmdr.

Para importar un juego de datos en el formato `csv` que contiene la identificador de los individuos, no es posible precisar en el menú deslizante de Rcmdr que la primera columna contiene el identificador. Podemos entonces importar el juego de datos considerando la identificación como una variable. Modificamos entonces la línea de código escrita en la ventana de guión añadiendo el argumento `row.names=1` y haciendo clic sobre `Ejecutar`.
Para cambiar de juego de datos activo, hacer clic en el encuadrado `Datos`. Si se modifica el juego de datos activo (por ejemplo, convirtiendo una variable), es necesario validar esta modificación del juego de datos por :
`Datos` → `Conjunto de datos activo` → `Actualizar conjunto de datos activo`
La ventana de salida toma las líneas de código en rojo y los resultados en azul. Los gráficos son dibujados en R. Al final de una sesión Rcmdr, es posible guardar la ventana de guión, es decir, todas las instrucciones así como el archivo de salida, es decir, todos los resultados. Podemos cerrar a la vez R y Rmcdr haciendo `Fichero` → `Salir` → `De Commander y R`.

Observaciones

- Escribir en la ventana de guión de Rcmdr o en la ventana de R es totalmente equivalente. Si una instrucción es lanzada desde Rcmdr, también es reconocida en R y viceversa. Los objetos creados por Rcmdr pueden pues ser utilizados en R.
- Es posible que las ventanas de Rcmdr se abran mal escondiéndose detrás de ventanas ya abiertas. En este caso, bajo Windows, hacer clic con el botón derecho del ratón en el icono de `R` o en el atajo que permite lanzar R, y luego hacer clic sobre (Propiedades), y modificar `Blanco` añadiendo `"-sdi"` después del camino de acceso al archivo `Rgui.exe`, lo que da por ejemplo :
  `"C:\Program Files\R\R-2.9.0\bin\Rgui.exe" --sdi`

A.2.3 Paquete FactoMineR

El paquete FactoMineR

El paquete FactoMineR (Husson *et al.*, 2009) está dedicado al análisis de datos «a la francesa». Los métodos más clásicos de análisis de datos son programados aquí : Análisis en Componentes Principales (función **PCA**), Análisis Factorial de las Correspondencias (función **CA**), Análisis Factorial de Correspondencias Múltiples (función **MCA**) y construcción ascendente de una jerarquía (función **HCPC**). Hay otros métodos más avanzados que están también disponibles y permiten tomar en consideración una estructura sobre las variables o sobre los individuos :
Análisis Factorial Múltiple (función **MFA**), Análisis Factorial Múltiple Jerárquico (función **HMFA**) o Análisis Factorial Múltiple Dual (función **DMFA**). La función **catdes** permite describir una variable cualitativa en función de las variables cuantitativas y/o cualitativas. La función **condes** permite describir una variable cuantitativa en función de las variables cuantitativas y/o cualitativas.
En cada método, es posible añadir elementos suplementarios : individuos suplementarios, variables cuantitativos y/o cualitativos suplementarias. Para cada uno de estos análisis, las numerosas ayudas a la interpretación son proporcionadas : calidad de representación, contribución para los individuos y las variables. Las representaciones gráficas están en el centro de cada uno de los análisis y las numerosas opciones gráficas están disponibles : colorear los individuos en función de una variable cualitativa, representar sólo las variables mejor proyectadas sobre los planos factoriales, etc.
Como todo paquete, debe ser instalado una sola vez y después ser cargado a cada utilización por :

```
> library(FactoMineR)
```

Una página web está dedicada al paquete FactoMineR : `http://factominer.free.fr`. Los métodos son descritos y los ejemplos son detallados.

Observación
Varios paquetes de análisis de datos están disponibles en R. Mencionemos en particular el paquete ade4. Una página web está dedicada a este paquete y proporciona numerosos ejemplos detallados y comentados : `http://pbil.univ-lyon1.fr/ADE-4`. Existe otro paquete sobre R dedicado exclusivamente a la clasificación, jerárquica o no, llamado cluster. Realiza los algoritmos descritos en el libro de Kaufman y Rousseuw (1990). [1].

El menú desplegable

Un interfaz gráfico está también disponible y puede ser instalado en el interfaz del paquete Rcmdr (cf. § A.2.2). Para cargar el interfaz de FactoMineR, hay dos posibilidades :
– Instalar definitivamente el menú desplegable de FactoMineR en Rcmdr. Para ello, solamente hay que escribir o copiar-pegar la fila de código siguiente en una ventana R :
`> source("http://factominer.free.fr/install-facto.r")`

1. Kaufman L. & Rousseuw P.J. (1990). Finding groups in data. An introduction to cluster analysis. Wiley, New-York, 342 p.

Para las utilizaciones posteriores del menú desplegable de FactoMineR, basta con lanzar Rcmdr con el comando **library**(`Rcmdr`), y el menú desplegable entonces está presente por defecto.

– Instalar para la sesión utilizada en ese momento el menú desplegable de FactoMineR en Rcmdr. Para ello, hay que instalar una sola vez el paquete RcmdrPlugin. FactoMineR. Luego, cada vez que se quiere utilizar el menú desplegable de FactoMineR, hay que lanzar Rcmdr, luego hacer clic sobre `Herramientas` → `Cargar Plug-in Rcmdr`. Hay que escoger el Plug-in de FactoMineR en la lista, Rcmdr después debe reiniciarse para tener en cuenta este nuevo plug-in. Esto es más complicado, por eso aconsejamos optar más bien por la primera posibilidad. Una utilización del menú desplegable es propuesta para el ACP más abajo.

1. Importar los datos

   El menú desplegable de Rcmdr propone varios formatos para importar los datos. Cuando el archivo está en un formato de texto (txt, csv), no se puede precisar que la primera columna contiene el identificador de los individuos (lo que es frecuentemente el caso en el análisis de los datos). Preferiremos realizar la importación a través del menú de FactoMineR.

   `FactoMineR` → `Import data from txt file`

   Hacer clic sobre `Rownames in the first column` (si el nombre de los individuos está presente en la primera columna) y precisar el separador de columnas (separador de campos) y el separador de decimal.

2. El ACP con FactoMineR

   Haga clic en la pestaña `FactoMineR`. Elegir `Principal Components Analysis` para abrir la ventana principal del ACP (cf. figura A.2).

   Entonces es posible seleccionar variables cualitativas suplementarias (`Select supplementary factors`), variables cuantitativas suplementarias (`Select supplementary variable`) e individuos suplementarios (`Select supplementary individuals`). Por defecto, los resultados sobre las 5 primeras dimensiones son proporcionados en el objeto `res`, las variables son centradas-reducidas y los gráficos son proporcionados para el primer plano (ejes 1 y 2). Es preferible pulsar sobre `Apply` más bien que `Submit`, lo que permite lanzar el análisis guardando la ventana abierta y de modificar ciertas opciones sin tener que rehacer todo el parametraje.

   La ventana de las opciones gráficas (cf. figura A.3) está separada en dos partes. La parte izquierda concierne el gráfico de los individuos mientras que la parte derecha concierne el gráfico de las variables. Es posible representar sólo las variables cualitativas suplementarias (sin los individuos, en `Hide some elements :` seleccionar `ind`) ; también es posible omitir las etiquetas de los individuos (`Label for the active individuals`). Los individuos pueden ser coloreados en función de una variable cualitativa (`Coloring for individuals` : escoger la variable cualitativa).

   La ventana de las diferentes opciones de salida permite visualizar los diferentes resultados (valores propios, individuos, variables, descripción automática de los ejes). Todos los resultados también pueden ser exportados a un archivo `csv` (archivo legible por Excel).

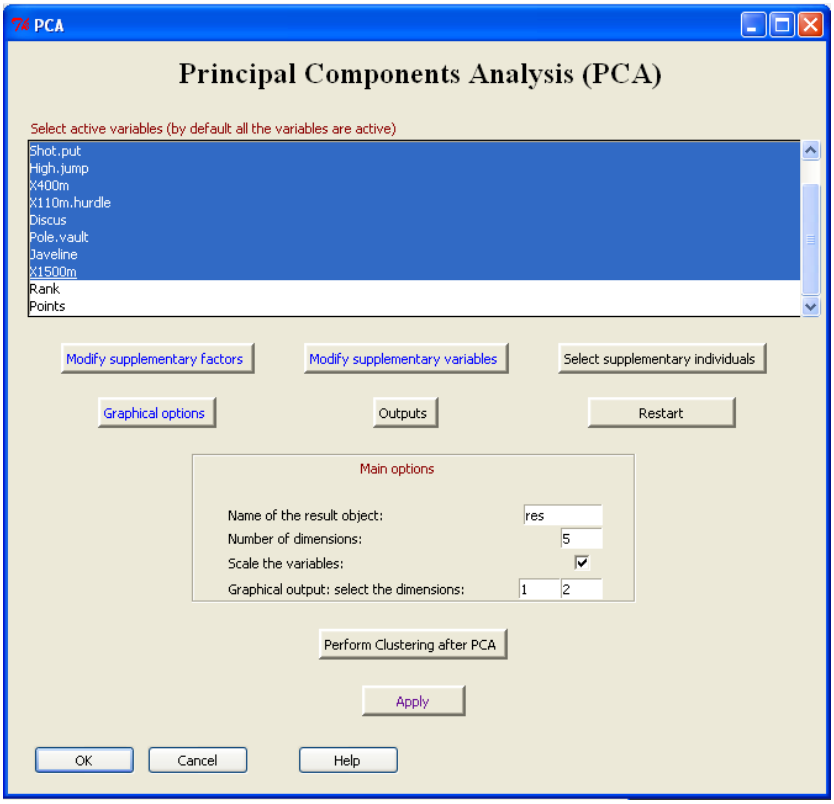

FIGURE A.2 – Ventana principal del ACP en el menu de FactoMineR.

El paquete dynGraph para gráficos interactivos

Existe un interfaz java que, en el momento en el que estas líneas son escritas, está en versión béta. Esta versión permite construir gráficos interactivos directamente a partir de las salidas de FactoMineR. Este interfaz java está disponible a través del paquete dynGraph.
A continuación hay que ir a la función **dynGraph**. Si los resultados de un análisis factorial son contenidos en un objeto `res`, basta con escribir :

```
> library(dynGraph)
> dynGraph(res)
```

El gráfico de los individuos se abre por defecto y es posible desplazar las etiquetas de los individuos para evitar que se sobrepongan, de colorear los individuos en función de una variable cualitativa, de representar los puntos con uno talla proporcional a una variable cuantitativa, etc. También podemos seleccionar individuos en una lista o directamente en la pantalla con el ratón y ponerlos en modo fantasma. El gráfico puede entonces ser guardado en diferentes formatos (`emf`, `JPEG`, `pdf`, etc.). El gráfico puede también ser guardado tal cual

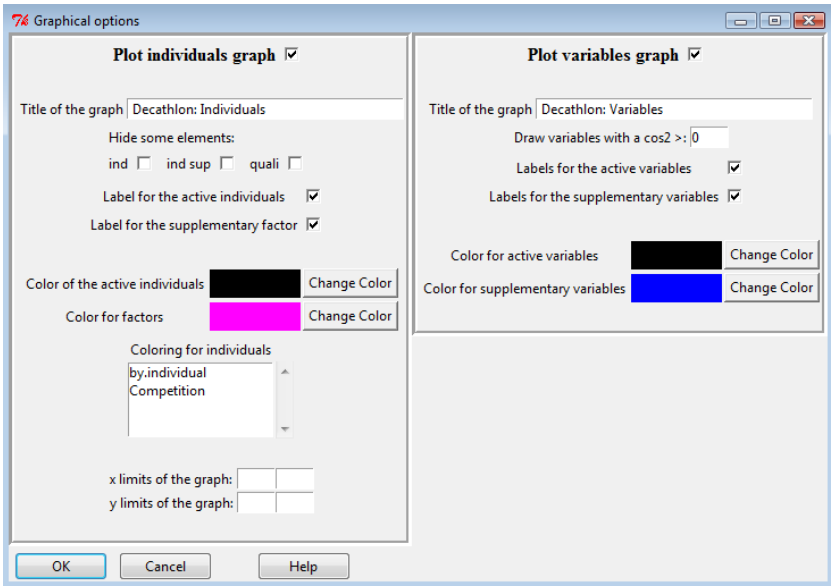

FIGURE A.3 – Ventana de las opciones gráficas del ACP.

y reabrirse posteriormente : esto es útil cuando los gráficos son lentos para pulir. La copia de seguridad es entonces un archivo `ser`.

Bibliografía sobre el paquete de R

He aquí una bibliografía de los principales paquetes que permiten realizar análisis factoriales o clasificaciones con R. Para una lista más detallada de los paquetes, remítase a la siguiente página web para los métodos de análisis factorial :

`http://cran.r-project.org/web/views/Multivariate.html`

y la página web siguiente para los métodos de clasificación :

`http://cran.r-project.org/web/views/Cluster.html`

- *El paquete ade4* propone funciones de análisis de datos para analizar datos ecológicos y medioambientales. El número de funciones disponibles es muy grande y muchas funciones pueden ser utilizadas en otros contextos fuera del contexto ecológico (funciones **dudi.pca**, **dudi.acm**, **dudi.fca**, **dudi.mix**, **dudi.pco**, etc.).
  Dray S. and Dufour A. B. (2007). The ade4 package : implementing the duality diagram for ecologists. *Journal of Statistical Software*. **22**, 1-20.
  Existe una página web dedicada a este paquete : `http://pbil.univ-lyon1.fr/ADE-4/`
- *El paquete ca*, propuesto por Greenacre y Nenadic, está dedicado al análisis simple de correspondencias (function **ca**) o múltiple (function **mjca**). Las numerosas extensiones para las variables cualitativas están disponibles en este paquete.
- *El paquete cluster* permite realizar clasificaciones estándares y más concretamente, la clasificación jerárquica ascendente con la función **agnes**.
- *El paquete dynGraph* es un programa de visualización que ha sido desarrollado inicialmente para el paquete FactoMineR. El principal objetivo de dynGraph permite al usuario explorar sus salidas gráficas multidimensionales de modo interactivo utilizando indicadores numéricos.
  Existe una página web dedicada a este paquete : `http://dyngraph.free.fr`
- *En este libro se ha utilizado el paquete FactoMineR* el cual permite realizar fácilmente análisis de datos multivariados (funciones **PCA**, **CA**, **MCA**, **HCPC**) proporcionando numerosos gráficos (funciones **plot**, **plotellipses**) y ayudas en la interpretación de los resultados (funciones **dimdesc**, **catdes**).
  Husson F., Josse J., Lê S. & Mazet J. (2009). *FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R.* R package version 1.12.
  Lê S., Josse J. & Husson F. (2008). FactoMineR : An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25, 1-18.

Existe una página web dedicada a este paquete : `http://factominer.free.fr`

- *El paquete homals* atañe al método *homogeneity analysis*. Es un método alternativo al análisis de las correspondencias múltiples para las variables cualitativas. Este método es a menudo utilizado en la psicometría.
  De Leeuw, J. & Mair P. (2009). Gifi methods for optimal scaling inR : The package homals. *Journal of Statistical Software*, **31**(4), 1–20.
- *El paquete hopach* construye (fonction **hopach**) árboles jerárquicos.
- *El paquete MASS* permite realizar análisis estándares. Las funciones **corresp** y **mca** permiten hacer el análisis de las correspondencias.
  Venables W.N. & Ripley B.D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.
- *El paquete missMDA* permite completar una tabla de datos incompleta con métodos de análisis de datos multivariados, e.g. según un modelo de ACP o un modelo de ACM. Combinándolo al paquete FactoMineR, permite administrar los datos ausentes en ACP y ACM.
- *El programa R* contiene algunas funciones de análisis de datos : **princomp** o **prcomp**, **hclust**, **kmeans**, **biplot**. Estas funciones son muy básicas y no hay ninguna ayuda disponible para la interpretación de los datos.
  R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- *El paquete Rcmdr* propone un interfaz gráfico (GUI) para R. Existen numerosos métodos de estadística clásica y varias extensiones disponibles para métodos específicos como por ejemplo RcmdrPlugin.FactoMineR.

Bibliografía

Esta bibliografía está dividida en varias secciones, cada una de ellas proporciona las referencias específicas asociadas a un método : análisis de componentes principales, análisis factorial de las correspondencias simples y múltiples y métodos de clasificación.

Referencias sobre el paquete R

– Sánchez A. L., Márquez M. M., Palacín F. F. & Navas A. S. (2008). *Estadística Básica con R y R-Commander*, UCA, Madrid
– Cornillon P.-A., Guyader A., Husson F., Jégou N., Josse J., Kloareg M., Matzner-Løber E. & Rouvière L. (2012). *R for Statistics*, CRC/PRESS Chapman & Hall, London.

Referencias sobre el conjunto de los métodos factoriales

– Abascal E. & Grande I. (1989). *Métodos multivariantes para la investigación comercial.* Ariel Economía, Barcelona.
– Aluja Banet T. & Morineau A. (1999). *Aprender de los datos : El Análisis de Componentes Principales ; una aproximación desde el Data Mining.* EUB, Barcelona
– Escofier B. & Pagès J. (1992). *Análisis Factoriales Simples y Múltiples. Objetivos, métodos e interpretación.* Servicio Editorial de la UPV/EHU, Bilbao.
– Escofier B. & Pagès J. (2008). *Analyses Factorielles Simples et Multiples : Objectifs, Méthodes et Interprétation.* Dunod, 4th edn, Paris.
– Gifi A. (1981). *Non-linear multivariate analysis.* D.S.W.O.-Press, Leiden.
– Govaert G. (2009). *Data Analysis.* Wiley.
– Hair J. F. (1999). *Analisis Multivariante De Datos 5E.* Prentice Hall, Madrid
– Lê S., Josse J., Husson F. (2008). FactoMineR : an R package for multivariate analysis. *Journal of Statistical Software.* 25 (1), 1-18.
– Lebart L., Morineau A. & Warwick K. (1984). *Multivariate descriptive statistical analysis.* Wiley, New-York.
– Lebart L., Piron M. & Morineau A. (2006). *Statistique exploratoire multidimensionnelle : visualisation et inférence en fouilles de données.* Dunod, 4^{e} édition, Paris.
– Lebart L., Morineau A. & Fénelon, J.P. (1984). *Tratamiento estadístico de datos.* Marcombo-Boixareu, Barcelona.

– Le Roux B. & Rouanet H. (2004). *Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis.* Dordrecht : Kluwer.
– Peña D. (2002). *Análisis de datos multivariantes.* McGraw-Hill, Madrid.

Referencias sobre el análisis de componentes principales

– Gower J. C. & Hand D. J. (1996). *Biplots.* Chapman & Hall/CRC, London.
– Jolliffe I. (2002). *Principal Component Analysis.* Springer. 2nd edn.

Referencias sobre el análisis factorial de las correspondencias y el análisis de correspondencias múltiple

– Benzécri J.P. (1973). *L'analyse des données. Tome 2 Correspondances.* Dunod, Paris.
– Benzécri J.P. (1992). *Correspondence Analysis Handbook.* (Transl : T.K. Gopalan) Marcel Dekker, New York.
– Greenacre M. (1984). *Theory and applications of correspondence analysis.* Acadamic Press.
– Greenacre M. (2007). *Correspondence Analysis in Practice.* Chapman & Hall/CRC.
– Greenacre M. & Blasius J. (2006). *Multiple Correspondence Analysis and related methods.* Chapman & Hall/CRC.
– Le Roux B. & Rouannet H. (2010). *Multiple Correspondence Analysis.* Sage, Series : Quantitative Applications in the Social Sciences, CA : Thousand Oaks Paris.
– Lebart L., Salem A. & Berry L. (2008). *Exploring Textual Data.* Kluwer Academic Publisher, Dordrecht, Boston.
– Murtagh F. (2005). *Correspondence Analysis and Data Coding with R and Java.* Chapman & Hall/CRC.

Referencias sobre los métodos de clasificación

– Hartigan J. (1975). *Clustering algorithms.* Wiley, New-York.
– Kaufman L. & Rousseeuw P. (1990). *Finding groups in data. An introduction to cluster analysis.* Wiley and sons, Inc. New-York.
– Lerman I. C. (1981). *Classification Automatique et Ordinale des Données.* Dunod, Paris.
– Mirkin B. (2005). *Clustering For Data Mining : A Data Recovery Approach.* Chapman & Hall/CRC.
– Murtagh F. (1985). *Multidimensional Clustering Algorithms.* Vienna : Physica-Verlag, COMPSTAT Lectures.

Índice

W