

ANALISIS DE CONGLOMERADOS

El objetivo es estudiar técnicas para determinar grupos o subconjuntos de individuos en una muestra de datos.

Para lo anterior se puede usar un criterio que determine cuando algunos de los individuos en la muestra son "similares" ó bien cuando no lo son. Desde un punto de vista práctico, la situación a la cual se debería llegar, es cuando los subconjuntos ó conglomerados son lo más homogéneos que se pueda (dos individuos en un grupo se parecen), mientras que las diferencias entre dos grupos diferentes, son lo más grandes que se pueda.

Hay dos pasos fundamentales para hacer análisis de conglomerados

- 1 Selección de una "medida de proximidad", una función de dos argumentos que permita determinar cuando estos son "parecidos o cercanos".
- 2 Selección de un algoritmo de agrupamiento

Este algoritmo sienta las bases ó los pasos a seguir, para que usando la medida de proximidad, se asignen individuos a los grupos.

Para una matriz de datos X con n renglones (individuos) y p columnas (variables), la proximidad entre los individuos se puede describir usando una matriz D de dimensiones $n \times n$

$$D = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix}$$

la entrada i,j contiene el valor de la medida de similaridad ó proximidad (podría ser medida de disimilaridad⁽¹⁾, por ejemplo distancias) entre los individuos i y j $i,j=1,2,\dots,n$.

Distancia y similaridad son conceptos duales, ya que

(1) Una distancia mide disimilaridad, ya que entre mayor sea la distancia entre dos individuos, estos serían menos similares.

Si d_{ij} es una distancia, entonces $d'_{ij} = \max_{i,j \in A} \{d_{ij}\} - d_{ij}$
 es una medida de proximidad ($A = \{1, 2, \dots, n\}$).

La elección de una medida de proximidad, depende de la naturaleza de los datos. Por ejemplo, para datos que provienen de variables binarias, conviene usar medidas de similaridad, en general lo anterior funciona cuando las escalas de medición de las variables son Nominales. Cuando la escala de medición de las variables es continua, en general D es una matriz de distancias.

SIMILARIDAD ENTRE INDIVIDUOS CON COMPONENTES BINARIAS

Si x_i^t y x_j^t son observaciones $x_i^t = (x_{i1}, \dots, x_{ip})$
 $x_j^t = (x_{j1}, \dots, x_{jp})$, donde $x_{ik}, x_{jk} \in \{0, 1\}$, $\forall k=1, \dots, p$.

Entonces pueden suceder cuatro casos:

$$x_{ik} = x_{jk} = 1,$$

$$x_{ik} = 0 ; x_{jk} = 1,$$

$$x_{ik} = 1 ; x_{jk} = 0,$$

$$x_{ik} = x_{jk} = 0.$$

Se definen entonces a_1, a_2, a_3 y a_4 como

$$a_1 = \sum_{k=1}^p \mathbb{1}_{(x_{ik}=x_{jk}=1)},$$

$$a_2 = \sum_{k=1}^p \mathbb{1}_{(x_{ik}=0; x_{jk}=1)},$$

$$a_3 = \sum_{k=1}^p \mathbb{1}_{(x_{ik}=1; x_{jk}=0)},$$

$$a_4 = \sum_{k=1}^p \mathbb{1}_{(x_{ik}=x_{jk}=0)}$$

Cada a_i es función de (x_{ij}, x_{kj}) ; $i=1,2,3,4$.

Se puede definir una familia paramétrica de medidas de proximidad como

$$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)},$$

donde los parámetros δ y λ son pesos. La siguiente tabla muestra a varios elementos de esta familia paramétrica, dependiendo de los valores de δ y λ .

Nombre	δ	λ	Definición
Jaccard	0	1	$\frac{a_1}{a_1 + a_2 + a_3}$
Simple matching	1	1	$\frac{a_1 + a_4}{P}$
Kulczynski	-	-	$\frac{a_1}{a_2 + a_3}$

Nombre	δ	λ	Definición
Tanimoto	1	2	$\frac{a_1 + a_4}{a_1 + 2(a_2 + a_3) + a_4}$
Russel and Rao	—	—	$\frac{a_1}{P}$
Dice	0	$\frac{1}{2}$	$\frac{2a_1}{2a_1 + (a_2 + a_3)}$

Estas medidas tienen diferentes formas de ponderar discrepancias, así como coincidencias positivas (presencia de caracteres comunes) ó coincidencias negativas (ausencia de caracteres comunes).

En el capítulo 3 del libro de Everitt et.al. (2011)⁽¹⁾ "Cluster Analysis", Wiley, se discute con mayor profundidad sobre el tipo de datos binarios y las circunstancias o contextos en los cuales, algunas de estas medidas resultan adecuadas.

(1) Everitt, B.S., Landau, S., Leese, M., y Stahl, D. (2011) "Cluster Analysis", Wiley

Para el caso de datos continuos, se tienen las distancias provenientes de las normas en L^r ;
 $r \geq 1$

$$d_{ij} = \|x_i - x_j\|_r = \left\{ \sum_{k=1}^p |x_{ik} - x_{jk}|^r \right\}^{1/r}, \dots \quad (i)$$

estas distancias miden disimilitud entre individuos.

Al calcular distancias entre individuos con (i), se está asumiendo en forma implícita que las variables (los componentes de x_i) fueron medidas en la misma escala. Si este no fuera el caso, entonces se puede aplicar una "estandarización". Se considera una matriz \mathbf{M} que sea positivo definida y de dimensiones $p \times p$ para definir

$$d_{ij}^2 = \|x_i - x_j\|_{\mathbf{M}}^2 = (x_i - x_j)^T \mathbf{M} (x_i - x_j),$$

por ejemplo la norma en L^2 se obtiene si

$\mathbf{M} = \mathbb{I}_p$, pero en el caso de requerir una estandarización, se puede usar

$$\mathbf{M} = \text{diag}\left(\frac{1}{\widehat{\text{VAR}}(x_1)}, \frac{1}{\widehat{\text{VAR}}(x_2)}, \dots, \frac{1}{\widehat{\text{VAR}}(x_p)}\right)$$

de esta forma $d_{ij}^2 = \sum_{k=1}^p \left\{ \frac{(x_{ik} - x_{jk})^2}{\widehat{\text{VAR}}(x_k)} \right\}$. Este ejemplo es importante, porque se evita que las distancias dependan del tipo de unidades de medición.

ALGORITMOS PARA CONGLOMERADOS

Esencialmente hay dos tipos de algoritmos para formar grupos: algoritmos jerárquicos y algoritmos de particiones. A su vez, los algoritmos jerárquicos se pueden dividir en procedimientos aglomerativos y procedimientos separativos. El algoritmo jerárquico aglomerativo comienza por la partición más fina posible (aquella en la que cada grupo ó subconjunto tiene una sola observación o individuo) y con esta propone nuevos grupos con estructura más compleja. El algoritmo jerárquico separativo, comienza con la partición menos fina posible (aquella en la que sólo hay un grupo, el cual contiene a todas las observaciones o individuos) y procede a

separar este conglomerado en grupos más pequeños. Estos algoritmos trabajan intercambiando elementos entre grupos hasta que logran optimizar una función score. La diferencia principal entre los algoritmos jerárquicos y los algoritmos de particiones, consiste en que para los algoritmos jerárquicos una vez que se encuentra la estructura de conglomerados "óptima", esta ya no se puede modificar. Para los algoritmos de particiones si es posible modificar la estructura de conglomerados.

ALGORITMOS JERÁRQUICOS (AGLOMERATIVOS)

Algoritmo:

- 1 Construir la partición más fina
- 2 Calcular la matriz de distancias D
- 3 Repetir:
- 4 Encontrar aquellos dos grupos con distancia más pequeña entre ellos
- 5 Formar un grupo con los dos grupos seleccionados.

6) Calcular las distancias entre los nuevos grupos y construir una matriz de distancias (reducida) D

7) Detener el proceso cuando el único grupo que resulte es el de todas las observaciones X .

Si dos individuos o grupos P y Q forman un nuevo grupo, para calcular la distancia entre este nuevo grupo y el grupo R se utilize la siguiente distancia ponderada (Lance y Williams)

$$d(R, \{P, Q\}) = \delta_1 d(R, P) + \delta_2 d(R, Q) + \delta_3 d(P, Q) \\ + \delta_4 |d(R, P) - d(R, Q)|$$

Los pesos $\delta_1, \delta_2, \delta_3$ y δ_4 definen diferentes distancias ponderadas (diferentes algoritmos aglomerativos), para mencionar algunos ejemplos

sean $n_P = \sum_{i=1}^n \mathbb{1}_{[x_i \in P]}$, $n_Q = \sum_{i=1}^n \mathbb{1}_{[x_i \in Q]}$ y

$n_R = \sum_{i=1}^n \mathbb{1}_{[x_i \in R]}$ el número de elementos en P, Q y R,

Table 4.1 Standard agglomerative hierarchical clustering methods.

Method	Alternative name ^a	Usually used with:	Distance between clusters defined as:	Remarks
Single linkage Sneath (1957)	Nearest neighbour	Similarity or distance	Minimum distance between pair of objects, one in one cluster, one in the other	Tends to produce unbalanced and straggly clusters ('chaining'), especially in large data sets. Does not take account of cluster structure.
Complete linkage Sorensen (1948)	Furthest neighbour	Similarity or distance	Maximum distance between pair of objects, one in one cluster, one in the other	Tends to find compact clusters with equal diameters (maximum distance between objects). Does not take account of cluster structure.
(Group) Average linkage Sokal and Michener (1958)	UPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	Tends to join clusters with small variances. Intermediate between single and complete linkage. Takes account of cluster structure. Relatively robust.
Centroid linkage Sokal and Michener (1958)	UPGMC	Distance (requires raw data)	Squared Euclidean distance between mean vectors (centroids)	Assumes points can be represented in Euclidean space (for geometrical interpretation). The more numerous of the two groups clustered dominates the merged cluster. Subject to reversals.
Weighted average linkage McQuitty (1966)	WPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	As for UPGMA, but points in small clusters weighted more highly than points in large clusters (useful if cluster sizes are likely to be uneven).
Median linkage Gower (1967)	WPGMC	Distance (requires raw data)	Squared Euclidean distance between weighted centroids	Assumes points can be represented in Euclidean space for geometrical interpretation. New group is intermediate in position between merged groups. Subject to reversals.
Ward's method Ward (1963)	Minimum sum of squares	Distance (requires raw data)	Increase in sum of squares within clusters, after fusion, summed over all variables	Assumes points can be represented in Euclidean space for geometrical interpretation. Tends to find same-size, spherical clusters. Sensitive to outliers.

^aU = unweighted; W = weighted; PG = pair group; A = average; C = centroid.

entonces tenemos

TABLA 2

Nombre de la distancia	δ_1	δ_2	δ_3	δ_4
Single Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Average Linkage (unweighted)	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Average Linkage (weighted)	$\frac{n_p}{n_p+n_q}$	$\frac{n_q}{n_p+n_q}$	0	0
Centroid	$\frac{n_p}{n_p+n_q}$	$\frac{n_q}{n_p+n_q}$	$-\frac{n_p n_q}{(n_p+n_q)^2}$	0
Median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{n_R+n_P}{n_R+n_P+n_Q}$	$\frac{n_R+n_Q}{n_R+n_P+n_Q}$	$-\frac{n_R}{n_R+n_P+n_Q}$	0

Para el caso de Single Linkage y Complete Linkage, se suelen usar los siguientes algoritmos aglomerativos modificados

ALGORITMO (JERARQUICO) AGLOMERATIVO MODIFICADO

- 1 construir la partición más fina.
- 2 calcular la matriz de distancias D
- 3 Repetir:
- ④ Encontrar el mínimo (Single Linkage) / máximo (Complete Linkage) valor de d (entre dos individuos n y m) en D

- (5) Si m y n no están en el mismo grupo, entonces se forma un nuevo grupo con estos y se elimina el valor de d encontrado en 4
- (6) Construir la matriz de distancias con los nuevos grupos

7 Detener el proceso cuando:

- (a) Todos los grupos fueron agrupados en Z
- (b) El valor de d satisface un criterio preestablecido

Representación Gráfica de la sucesión de agrupamientos
DENDOGRAMA

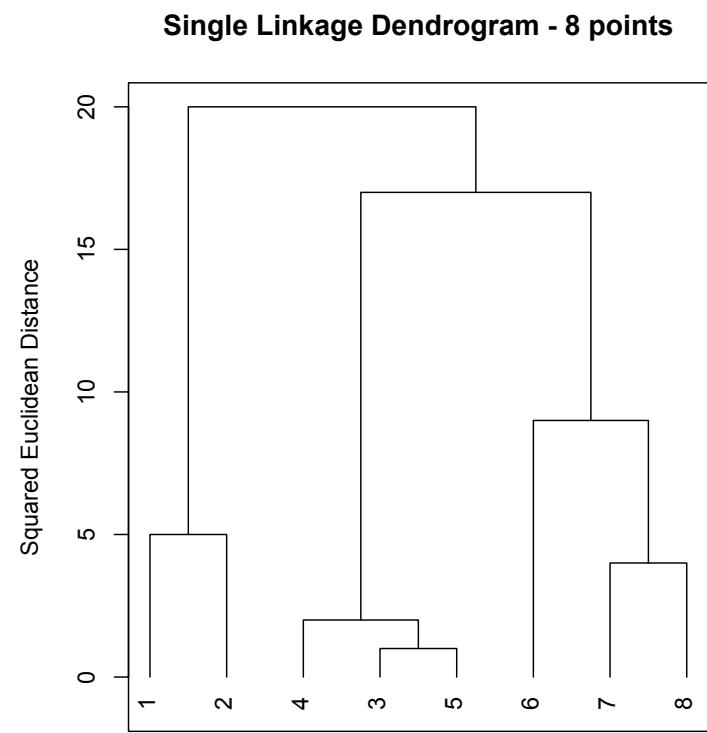
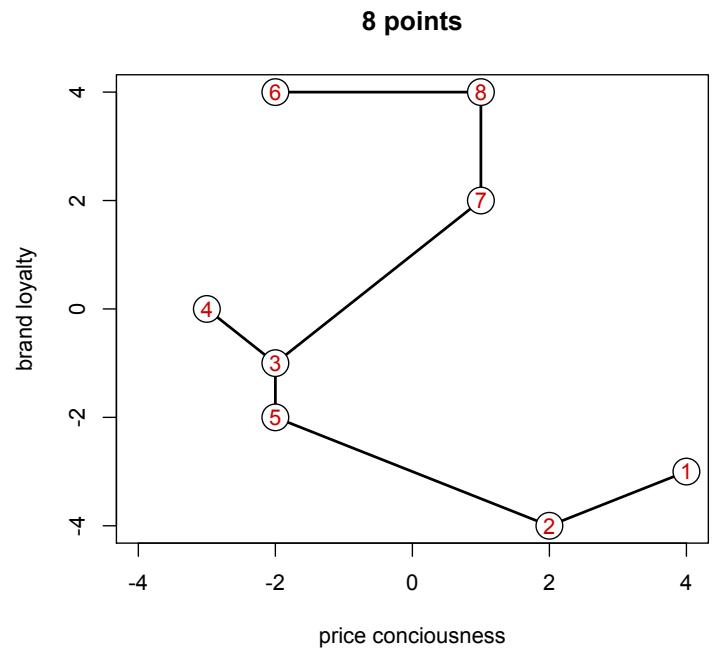
Una gráfica muy usada para representar la sucesión de agrupamientos es el dendograma. Esta gráfica despliega las observaciones, la sucesión de grupos producida por el algoritmo de agrupamiento y las distancias entre los grupos. En el eje horizontal

se suelen representar a los individuos con índices y en el eje vertical se representan las distancias entre los grupos (en algunos paquetes en el eje horizontal se representan distancias y en el eje vertical a los individuos). El dendograma es similar a un árbol que "ramifica" conforme las distancias se hacen pequeñas, al trazar una linea horizontal y cortar el árbol a una determinada distancia, los ramaes describen una estructura de agrupamientos.

Ejemplo: Se tiene un muestreo de 8 puntos
 $x_1, \dots, x_8 \quad x_i = (x_1, x_2)$

X_1 = Conciencia del Precio X_2 = Lealtad a la marca
VÉASE FIGURA A

Supóngase que se lleva a cabo un algoritmo aglomerativo usando single linkage. La matriz de distancias (nórmas L^2) está dada por



$$D = \begin{pmatrix} 0 & 10 & 53 & 73 & 50 & 98 & 41 & 65 \\ 0 & 25 & 41 & 20 & 80 & 37 & 65 \\ 0 & 2 & 1 & 25 & 18 & 34 \\ 0 & 5 & 17 & 20 & 32 \\ 0 & 36 & 25 & 45 \\ 0 & 13 & 9 \\ 0 & 4 \\ 0 \end{pmatrix}$$

Al aplicar el algoritmo aglomerativo se obtiene el dendograma en la figura A. Si por ejemplo el árbol se corta a nivel 10, tenemos una estructura de 3 grupos: $\{1,2\}$, $\{3,4,5\}$ y $\{6,7,8\}$

El algoritmo single Linkage, define la distancia entre dos grupos como el mínimo de las distancias individuales (las distancias entre individuos de los grupos), de la Tabla 2 (página 10) tenemos

$$d(R, \{P, Q\}) = \frac{1}{2} d(R, P) + \frac{1}{2} d(R, Q) - \frac{1}{2} |d(R, P) - d(R, Q)|$$

caso 1: $d(R, P) \geq d(R, Q)$

$$\begin{aligned} d(R, \{P, Q\}) &= \frac{1}{2} d(R, P) + \frac{1}{2} d(R, Q) - \frac{1}{2} (d(R, P)) + \frac{1}{2} (d(R, Q)) \\ &= d(R, Q) = \min(d(R, P), d(R, Q)) \end{aligned}$$

Caso 2 $d(R, P) \leq d(R, Q)$

$$\begin{aligned} d(R, \{P, Q\}) &= \frac{1}{2} d(R, P) + \frac{1}{2} d(R, Q) - \frac{1}{2} (d(R, Q) - d(R, P)) \\ &= d(R, P) = \min(d(R, P), d(R, Q)) \end{aligned}$$

$$\therefore d(R, \{P, Q\}) = \min(d(R, P), d(R, Q)) \dots \text{ (a)}$$

Por esta razón a este algoritmo también se le llama "algoritmo del vecino más cercano".

Como consecuencia de esta característica, el algoritmo single Linkage tiende a construir grupos numerosos. Puede suceder que los grupos no resulten "muy distantes" en el sentido de que existan puntos o individuos que estén "cercaos" pese a que estos pertenezcan a diferentes grupos, lo anterior puede ocasionar que el algoritmo reúna a los correspondientes grupos en uno solo.

El algoritmo complete Linkage trata de corregir el problema anterior al considerar el máximo de las distancias individuales

Se puede argumentar, igual que como se hizo para la ecuación (a) en el single linkage, que para el caso del algoritmo complete linkage

$$d(R, \{P, Q\}) = \max \{d(R, P), d(R, Q)\},$$

razón por la cual también se le conoce como el "algoritmo del vecino más lejano"

Se puede argumentar, igual que como se hizo para la ecuación (a) en el single linkage, que para el caso del algoritmo complete linkage

$$d(R, \{P, Q\}) = \max \{d(R, P), d(R, Q)\},$$

razón por la cual también se le conoce como el "algoritmo del vecino más lejano". En principio complete linkage producirá grupos de forma que todos los individuos en un grupo "son parecidos"

El algoritmo average linkage (weighted or unweighted) propone un punto intermedio entre el single linkage y el complete linkage, ya que para este algoritmo se calcula la distancia promedio

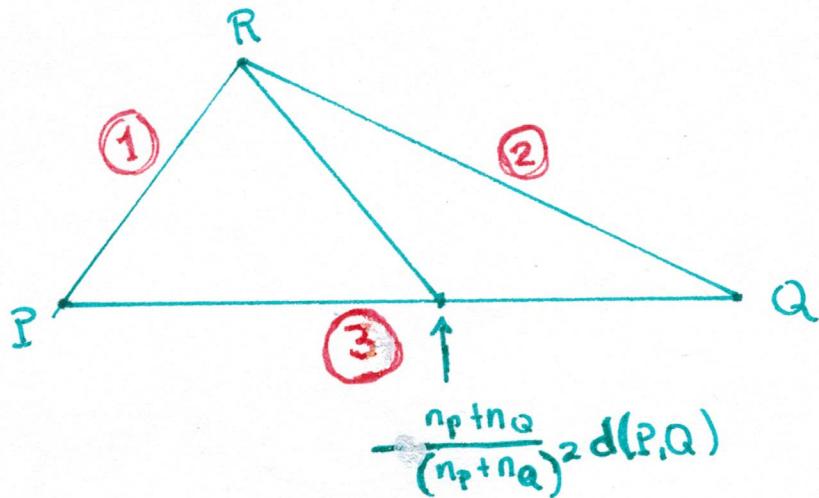
$$d(R, \{P, Q\}) = \frac{n_P}{n_P + n_Q} d(R, P) + \frac{n_Q}{n_P + n_Q} d(R, Q)$$

El algoritmo del centroide es muy similar al de average linkage pero tiene una corrección que corresponde a una proporción de la distancia entre P y Q

$$d(R, \{P, Q\}) = \frac{n_P}{n_P + n_Q} d(R, P) + \frac{n_Q}{n_P + n_Q} d(R, Q)$$

$$- \frac{n_P n_Q}{(n_P + n_Q)^2} d(P, Q)$$

(1) (2) (3)



El algoritmo de agrupamientos de Ward, cuyos pesos se enuncian en la tabla 2, tiene una forma más sofisticada⁽¹⁾, ya que decide unir dos grupos sólo si al calcular una "medida de heterogeneidad" asignada al nuevo grupo (la unión de los dos grupos), esta medida no es "muy grande". En otras palabras el algoritmo une a dos grupos solo si el grupo resultante es tan homogéneo como "sea posible".

La heterogeneidad de un grupo R se mide usando la "inercia"

$$I_R = \frac{1}{n_R} \sum_{i=1}^{n_R} d^2(x_i, \bar{x}_R),$$

(1) yo diría inteligente

$$\text{donde } \bar{x}_R = \frac{1}{n} \sum_{i=1}^{n_R} x_{ri} \quad R = \{x_{r1}, \dots, x_{rn_R}\}$$

I_R es una medida de dispersión del grupo al rededor de su centro de gravedad (\bar{x}_R). Si usamos d = distancia euclídea, entonces I_R representa la suma de las varianzas de las P componentes x_{ri}, \dots, x_{rp} de x_{ri} ; $i=1, 2, \dots, n_R$.

Cuando dos grupos P y Q son agrupados en $\{P, Q\}$, la inercia del nuevo grupo $\{P, Q\}$ se incrementa. Se puede probar que el correspondiente incremento está dado por

$$\Delta(P, Q) = \frac{n_P n_Q}{n_P + n_Q} d^2(P, Q).$$

El algoritmo de Ward une los grupos P y Q , sólo si $\Delta(P, Q)$ es mínimo.

Ejemplo Datos de comida en Francia

Se tienen registrados gastos (promedio) de comida para diferentes tipos de familias en Francia

Trabajadores Ménages = MA
Empleados = EM

Gerentes = CA

Además los datos se han registrado dependiendo del numero de hijos (2,3,4 ó 5).

	bread	veg.	fruits	meat	poultry	milk	wine
MA2	332	428	354	1437	526	247	427
EM2	293	559	388	1527	567	239	258
CA2	372	767	562	1948	927	235	433
MA3	406	563	341	1507	544	324	407
EM3	386	608	396	1501	558	319	363
CA3	438	843	689	2345	1148	243	341
MA4	534	660	367	1620	638	414	407
EM4	460	699	484	1856	762	400	416
CA4	385	789	621	2366	1149	304	282
MA5	655	776	423	1848	759	495	486
EM5	584	995	548	2056	893	518	319
CA5	515	1097	887	2630	1167	561	284

Haciendo un análisis de componentes principales normalizadas obtenemos los siguientes porcentajes de varianza muestral explicada

Valor Propio	Proporcion de Var	Proporcion de Var acumulada
4.33	0.6190	61.9
1.83	0.2620	88.1
0.631	0.09	97.1
0.128	0.0180	98.9
0.058	0.0080	99.7
0.019	0.0030	99.9
0.001	0.0001	100

Procedemos a trabajar con los primeros dos componentes los cuales explican el 88.1% de la varianza muestral

	$r_{X_i Z_1}$	$r_{X_i Z_2}$	$r_{X_i Z_1}^2 + r_{X_i Z_2}^2$
$X_1 = \underline{\text{Pan}}$	-0.499	0.842	0.957
$X_2 = \underline{\text{Vegetales}}$	-0.970	0.133	0.958
$X_3 = \underline{\text{Frutas}}$	-0.929	-0.278	0.941
$X_4 = \underline{\text{cerne}}$	-0.962	-0.191	0.962
$X_5 = \underline{\text{Aves}}$	-0.911	-0.266	0.901
$X_6 = \underline{\text{Leche}}$	-0.584	0.707	0.841
$X_7 = \underline{\text{Vinos}}$	0.428	0.648	0.604

$$Z_1 = -0.24 X_1 - 0.466 X_2 - 0.446 X_3 - 0.462 X_4 - 0.438 X_5 \\ - 0.281 X_6 + 0.206 X_7$$

$$Z_2 = 0.622 X_1 + 0.09 X_2 - 0.205 X_3 - 0.141 X_4 - 0.197 X_5 \\ + 0.523 X_6 + 0.479 X_7$$

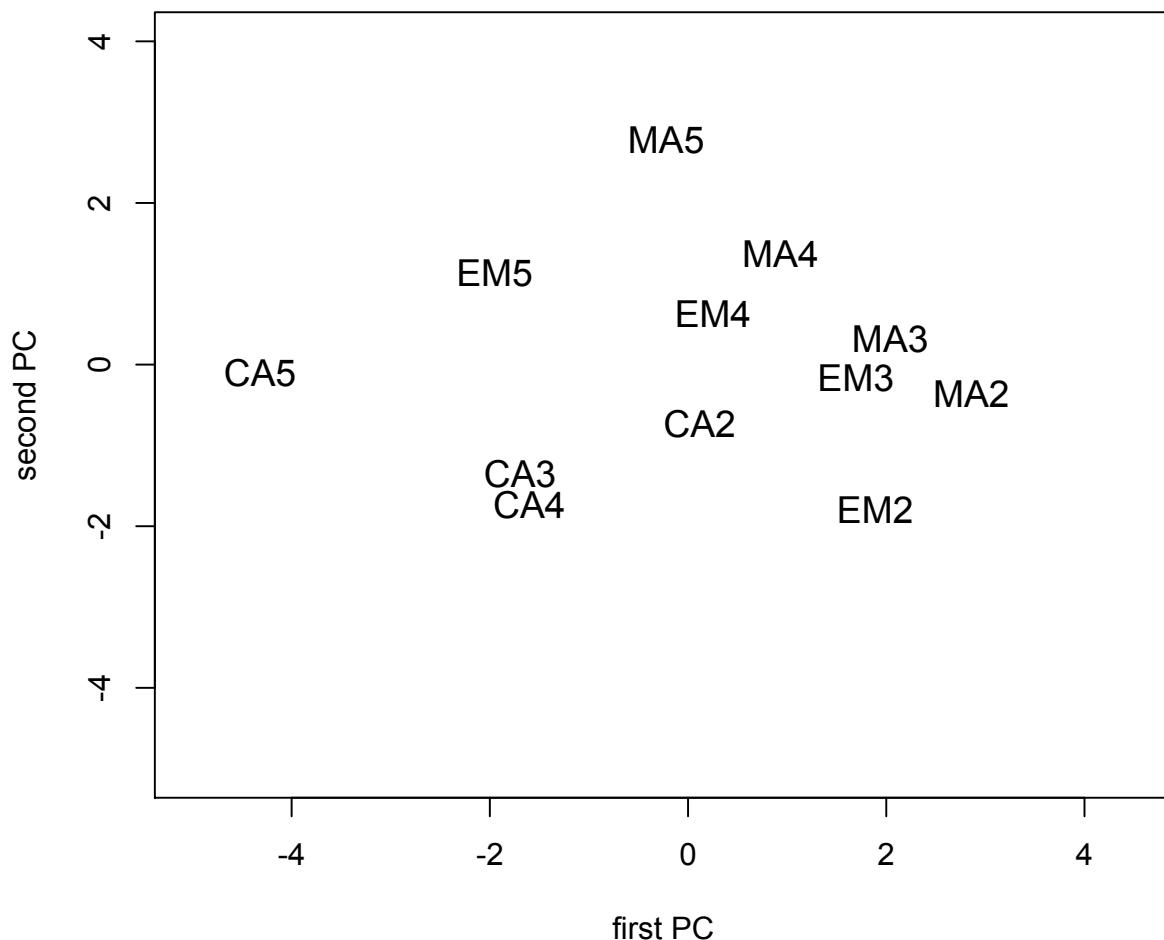
La primera componente depende fuertemente de las cantidades gastadas en vegetales, frutas, cerne y aves (mientras más grandes estos gastos, más pequeña es Z_1).

La segunda componente depende de los contenidos gestados en pan, vino y leche.

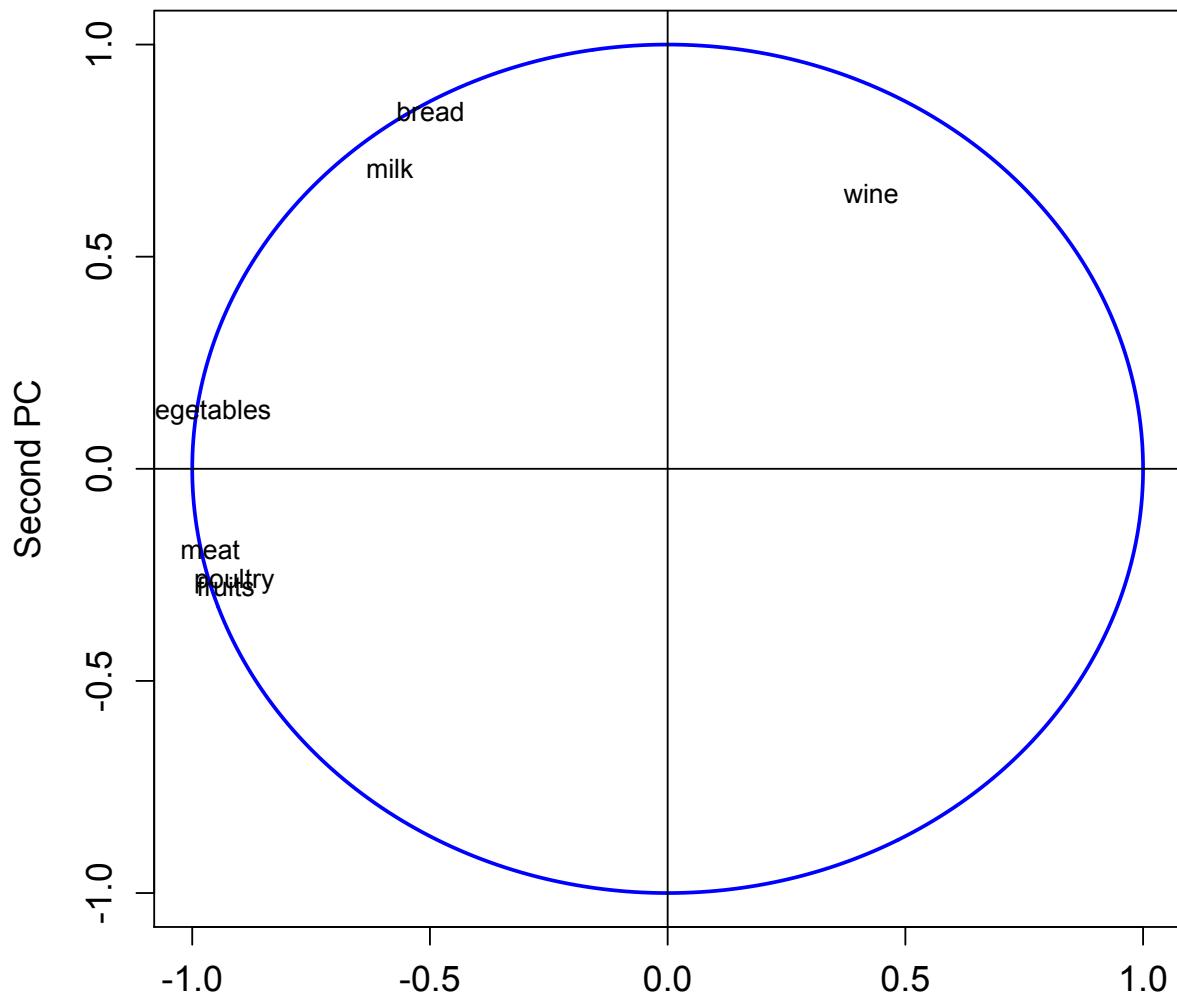
Si se usa la figura correspondiente a las dos primeras componentes principales (scatterplot) así como la información de las correlaciones, se puede inferir que las familias con mayor número de hijos quedan a la izquierda en la figura (los gastos en vegetales, fruta, carne y aves son mayores). Al mismo tiempo, las familias de los gerentes también ejercen mayor gasto en estos rubros y las familias de los trabajadores manuales ejercen menos gasto en esos mismos rubros, pero gastan más en pan, leche y vino (en fama similar para las familias de empleados).

Para llevar a cabo un análisis de agrupamientos, se consideran los datos estandarizados ó equivalentemente se usa $d_{ij}^2 = (x_{ij} - \bar{x}_{ij})^2 / \sigma^2_{ij}$ con $\sigma_{ij} = \text{diag}(\frac{1}{\widehat{\text{VAR}}(x_1)}, \dots, \frac{1}{\widehat{\text{VAR}}(x_p)})$

French Food data



French Food data



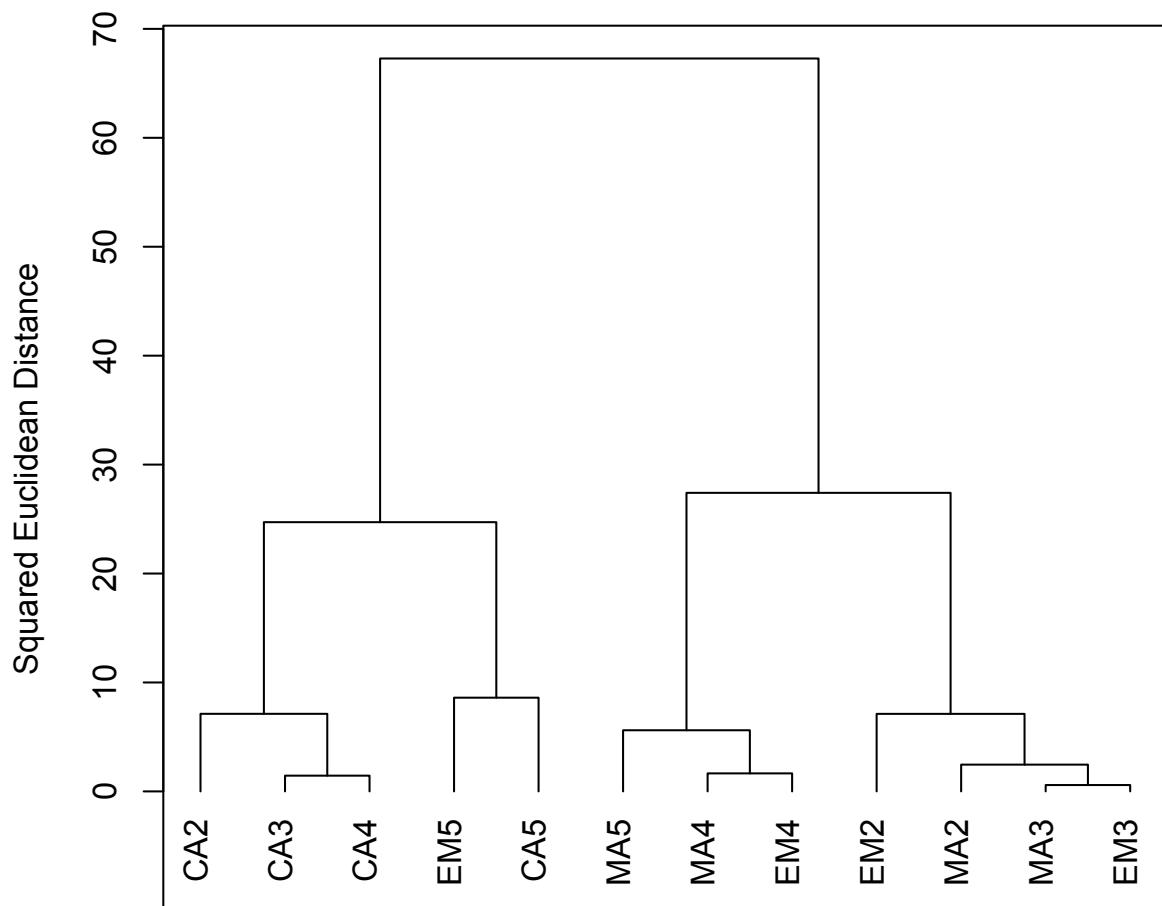
Al calcular el algoritmo jerárquico aglomerativo de Ward se obtiene el dendograma de la figura D1. Si se corta a una distancia de 30 hay dos grupos $\{CA_2, CA_3, CA_4, EM_5, CA_5\}$ y $\{MA_5, MA_4, EM_4, EM_2, MA_2, MA_3, EM_3\}$, el primer grupo sería el de familias cuyo gasto en comida primordialmente se da en frutas, vegetales, carne y aves. En el segundo grupo de familias se gasta más en pan, leche y vino. Si se corta a una distancia de 20, se obtienen 4 grupos $\{CA_2, CA_3, CA_4\}$, $\{EM_2, MA_2, EM_3, MA_3\}$, $\{EM_4, MA_4, MA_5\}$ y $\{EM_5, CA_5\}$ los factores que determinan cada grupo son el nivel social (social-profesional) y el tamaño de las familias.

Ejemplo Billetes del Banco Suizo

Se seleccionan al azar 20 billetes de la muestra $x_{G_1}, \dots, x_{G_{200}}$ de billetes del banco suizo con 10

Figura D1

Ward Dendrogram for French Food



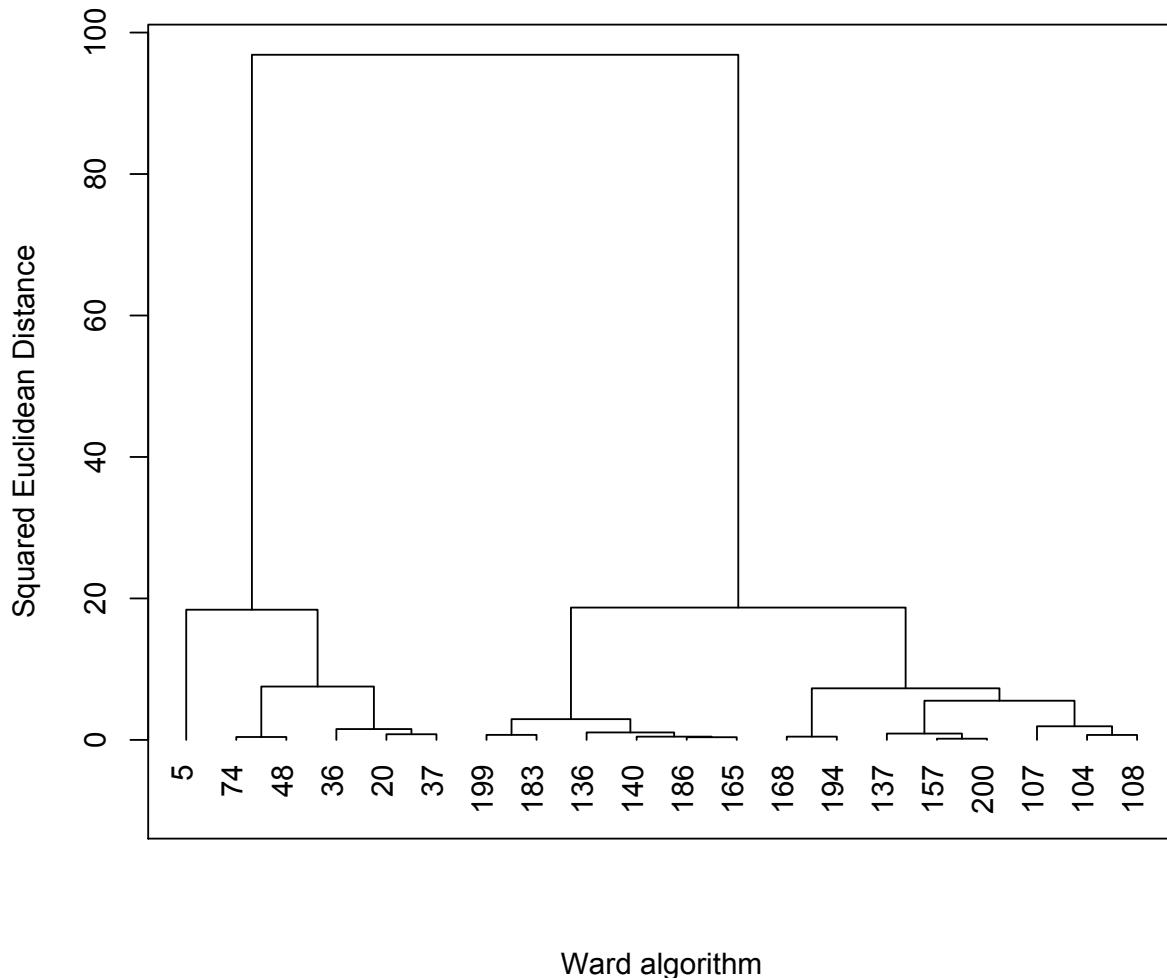
cual ya habíamos trabajado. Al aplicar el algoritmo jerárquico aglomerativo de Ward obtenemos el dendograma en la figura D2.

Si se corta en una distancia de 40 se obtienen dos grupos, los billetes en uno de los grupos tienen números entre 1 y 100 y los billetes del otro grupo, números entre 101 y 200.

Recordemos que se tenía información referente a que los billetes numerados 1, ..., 100 son verdaderos mientras que los numerados 101, ..., 200 son falsos.

Figura D2

Dendrogram for 20 Swiss bank notes



Ejemplo: Datos de Arrestos en los E.U.A. (1973) Estadísticas (por cada 100,000 habitantes) de asaltos, asesinatos y violaciones. La cuarta columna es el porcentaje de la población que vive en áreas urbanas.

La función "USarrest.R" estandariza los datos, calcula una matriz de distancias D y con esta obtiene un agrupamiento el cual depende del método jerárquico aglomerativo que se le indique a la función de R "hclust()", posteriormente se le puede aplicar la función "fuz_dend()"⁽¹⁾, al objeto que contiene el agrupamiento, con el fin de producir un dendograma. La figura en el archivo "Figura D1.pdf" contiene dendogramas producidos con los métodos jerárquicos aglomerativos: "average Linkage" (1), "complete linkage" (2) y "single Linkage" (3).

En algunos libros de Análisis Multivariado se

(1) Paquete "factoextra" de R

propone juzgar la "calidad" de un agrupamiento calculando la correlación entre las distancias D y las "distancias cofenéticas" en el agrupamiento. Para dos individuos x_{G_i} y x_{G_j} su distancia cofenética se define como la altura (en el dendograma) del nodo en donde x_{G_i} y x_{G_j} fueron "clasificados en el mismo grupo". Si T denota la matriz de distancias cofenéticas para un agrupamiento, se define el "coeficiente de correlación cofenético" ⁽¹⁾ c como

$$c = \frac{\sum_{i < j} (d_{ij} - \bar{d})(T_{ij} - \bar{T})}{\sqrt{\sum_{i < j} (d_{ij} - \bar{d})^2} \sqrt{\sum_{i < j} (T_{ij} - \bar{T})^2}}$$

donde \bar{d} es el promedio $\bar{d} = \frac{\sum_{i < j} d_{ij}}{n(n+1)/2}$

y \bar{T} es el promedio $\bar{T} = \frac{\sum_{i < j} T_{ij}}{n(n+1)/2}$

Si c tiene valores "grandes" (cercaos a 1)

(1) Sokal y Rohlf (1962) The comparison of dendograms by objective methods. Taxon, 11, 33-40

entonces se considera que el agrupamiento propuesto por el dendograma es una "buena descripción de los datos". En la práctica un valor arriba de 0,75 se considera "grande". La función "cophenD.R" calcula los coeficientes de correlación cophenética para los distintos métodos de agrupamiento jerárquico aglomerativo que tiene la función "hclust()". Al parecer el agrupamiento "más adecuado" para los datos de arrestos es el que se obtiene usando average linkage.

Se puede obtener un agrupamiento al cortar el dendograma a una altura determinada, o bien cuando se identifica la altura en que se forman un número de grupos K pre-determinado. Si se usa una altura ó un valor de K predeterminados, eso depende de la aplicación con la que se está trabajando.

No obstante se puede encontrar un valor "óptimo" de K (cuando no se tiene idea de este), usando una medida de "qué tan adecuado" resulta considerar a K como el número de grupos.

Esta medida se conoce como el "ancho promedio de la silueta"⁽²⁾ del agrupamiento

(2) Alan Julian Izenman "Modern Multivariate Statistical Techniques, Regression Classification and Manifold Learning". Springer Verlag páginas 426 - 429.