Universidad Carlos III de Madrid

Análisis Multivariante, Grado en Estadística y Empresa

Práctica III: Clasificación Jerárquica y No Jerárquica

Jorge Salas y Marc Pastor

Índice

1 Clasificación

En este apartado vamos a usar técnicas de clasificación jerárquicas y no jerárquicas, con el objetivo de extraer aun más información de los datos.

Para realizar esta práctica nos hemos basado únicamente en variables numéricas, ya que hemos probado a usar datos mixtos, y los resultados no eran muy fiables, dando lugar a matrices ultramétricas **U** con poca correlación cofenética respecto a la matriz de distancias originales **D** (del orden de 0.1, etc), lo cuál implicaba una total dispersión de las distancias originales.

En concreto hemos ceñido nuestro análisis a las 5 primeras componentes principales, que obtuvimos en la práctica anterior. En el siguiente gráfico podemos ver que 5 componentes explican alrededor del 88% de la variabilidad de los datos, lo cuál es muchísimo.

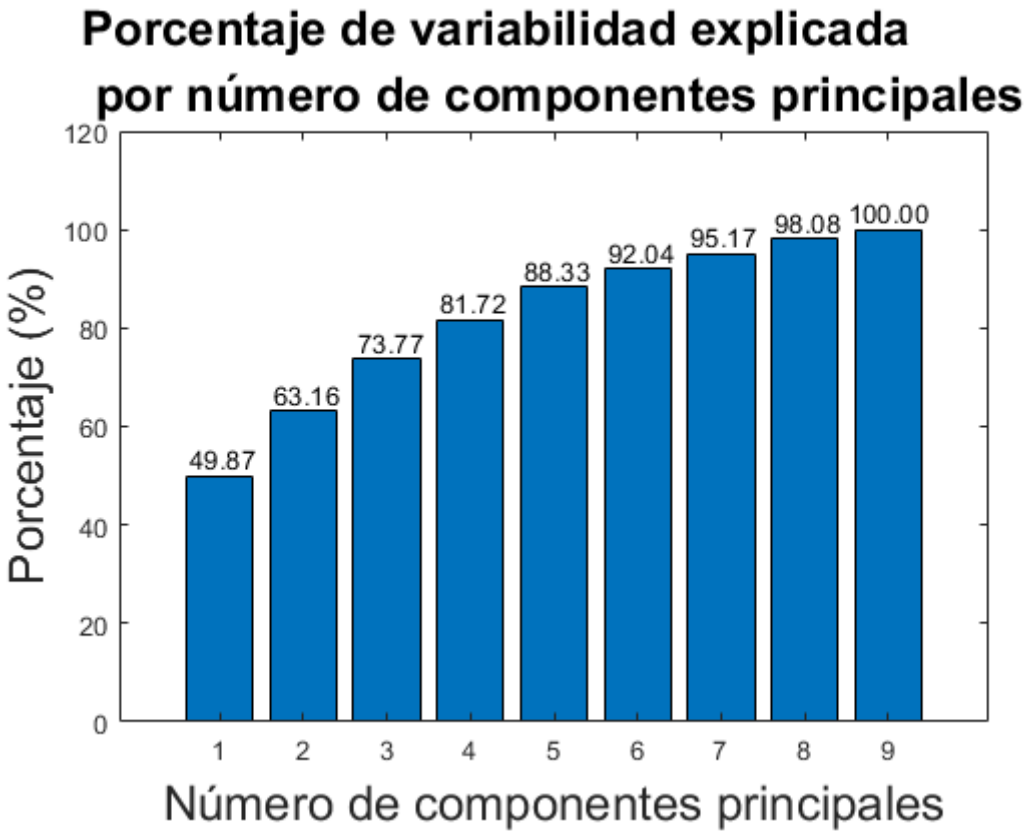

Figure 1: Porcentaje de variabilidad explicada por componentes principales

1.1 Clasificación Jerárquica:

Hemos realizado clasificación jerárquica sobre las 5 componentes principales utilizando 3 medidas de distancia distintas: **el método del mínimo, del máximo y el UPGMA (Unweighted Pair Group Method using arithmetic Averages)**, utilizando una matriz de distancias euclídeas, ya que podemos usarla porque los ejes definidos por los componentes principales son ortogonales.

```
% utilizaremos las 5 primeras componentes principales Y2 (solo variables
% numéricas)
Y2_5 = Y2(:, 1:5); % sacamos las 5 componentes principales
D_5 = pdist(Y2_5, 'euclidean'); % matriz de distancias de mahalanobis de
↪  las 3 componentes principales

Z_UPGMA = linkage(D_5, 'average');
Z_min = linkage(D_5, 'single');
Z_max = linkage(D_5, 'complete');
```

```
c_UPGMA = cophenet(Z_UPGMA, D_5)
dendrogram(Z_UPGMA, 0, 'colorthreshold', 3.5);
```

```
c_UPGMA =

    0.6631
```

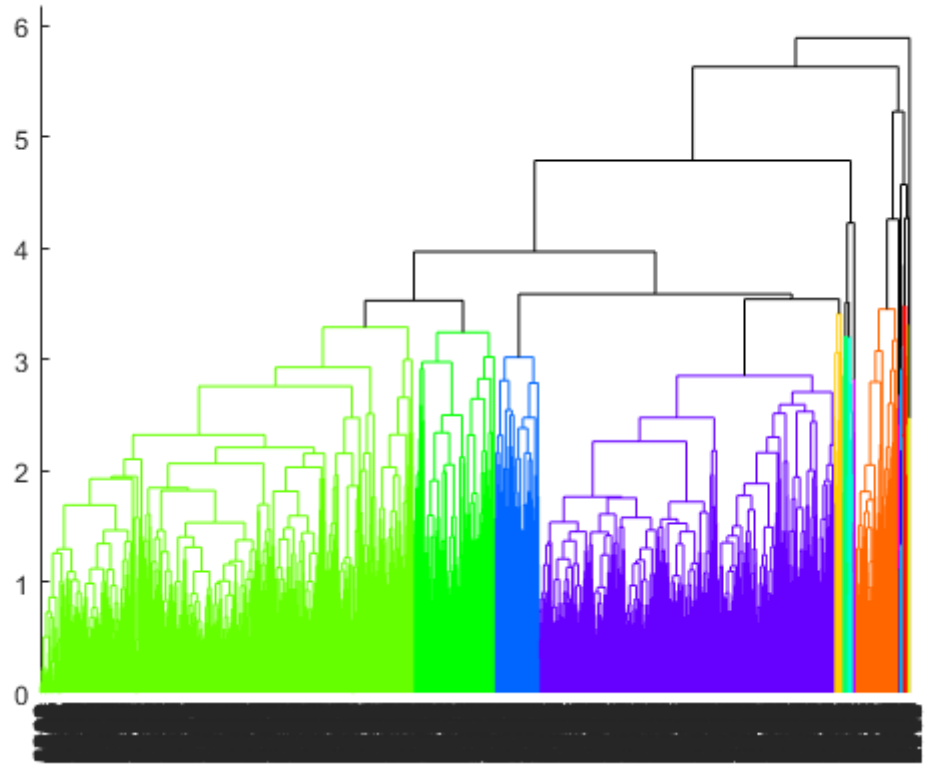

Figure 2: Dendrograma UPGMA

```
c_min = cophenet(Z_min, D_5)
dendrogram(Z_min, 0, 'colorthreshold', 0.25)
```

```
c_min =

    0.4189
```

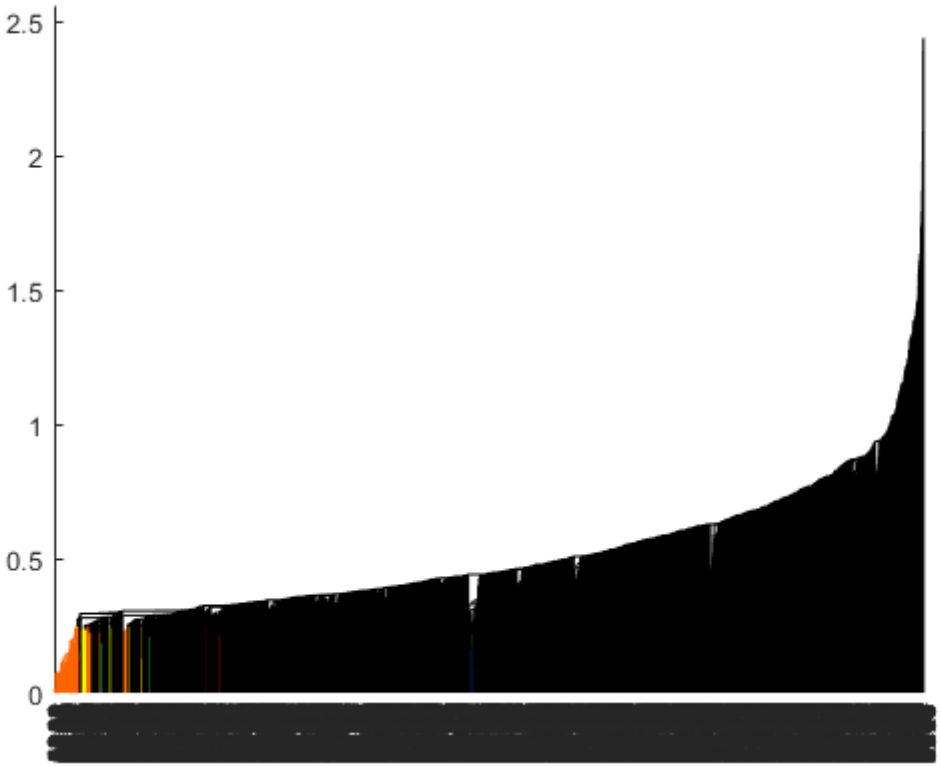

Figure 3: Dendrograma Distancia Máxima

```
c_min = cophenet(Z_max, D_5)
dendrogram(Z_max, 3.5, 'colorthreshold', 0.25)
```

```
c_min =

    0.4910
```

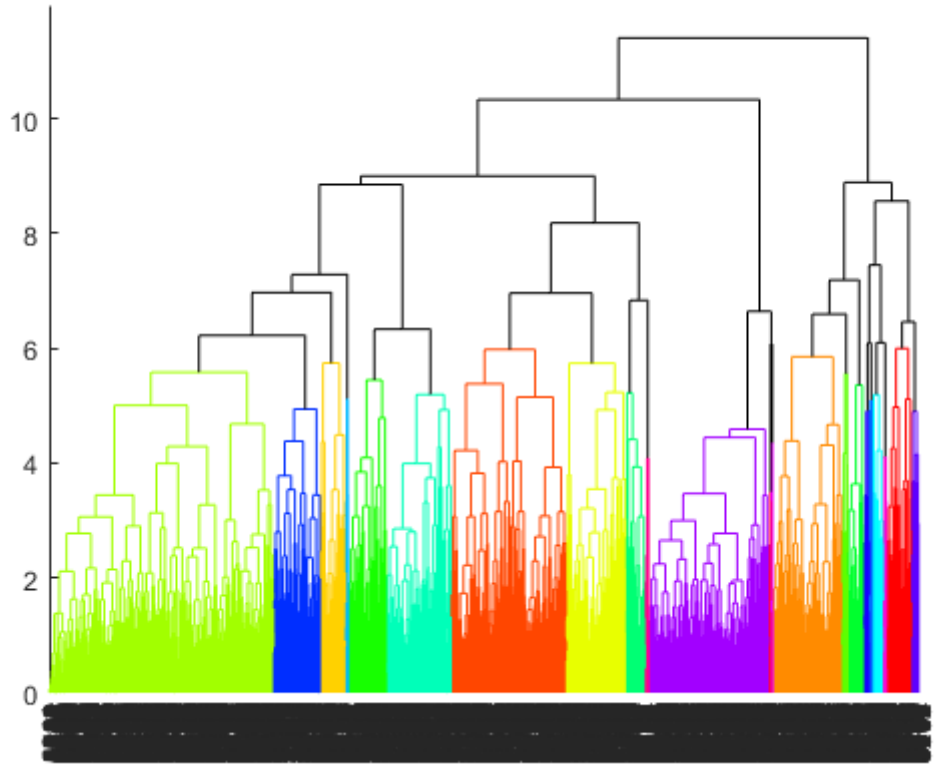

Figure 4: Dendrograma Distancia Mínima

Podemos ver claramente que el método más fiable es el UPGMA, ya que la correlación cofenética de la matriz ultramétrica **S** y la matriz de distancias original **D**, es de 0.61, siendo alrededor de un 50% mayor que en los otros dos métodos. Además, la representación es mucho más clara y nos permite distinguir unos 6 grupos distintos de apartamentos (aunque no logramos ver los atributos que forman cada grupo, ya que el eje horizontal del gráfico es totalmente negro, debido a la cantidad de valores superpuestos que hay).

1.2 Clasificación no Jerárquica:

Hemos usado el algoritmo de k-medias, para clasificar las observaciones en distintos grupos, en base a sus 5 componentes principales. Hemos establecido 2 como el número máximo de grupos, ya que el algoritmo no converge para más de 2 grupos (hemos probado y nos salta un warning de que el algoritmo diverge, y resulta en una clasificación en 2 grupos):

```
[C,s,IDX]=kmedias2(Y2_5,2);
C
s
```

```
C =

   -2.0467   -0.0184   -0.0191   -0.0590    0.0136
    1.4595    0.0131    0.0136    0.0421   -0.0097


s =

    0.5034
```

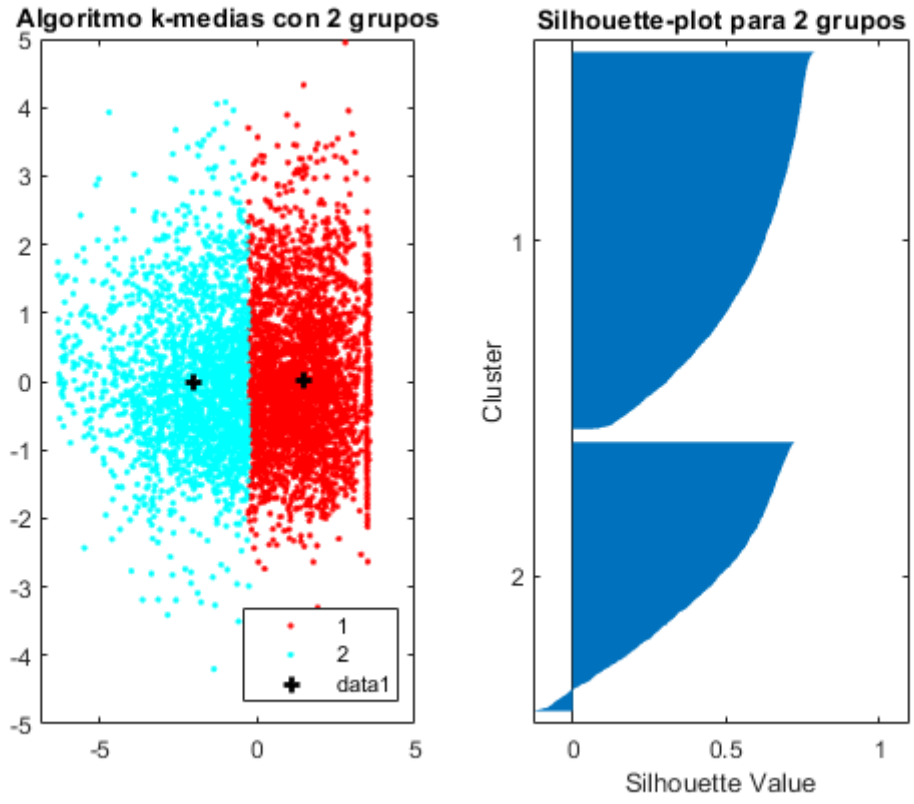

Figure 5: Gráfico de dispersión y de silueta

Vemos que la calidad de la clasificación no es muy buena, ya que el valor de la silueta promedio $\bar{s} = 0.5034$, lo cuál está más cerca de 1 que de 0, aunque la distancia a ambos extremos es muy parecida. Observando el gráfico podemos ver que hay una gran proporción de observaciones de cada grupo con siluetas por debajo a 0.5, con lo que no deberíamos

fiarnos mucho de las posibles interpretaciones.

Por otro lado, los centroides, son respecto a las componentes principales calculadas en base a las componentes principales de variables transformadas, con lo que no tienen una interpretación muy clara. Para poder interpretar los resultados, hemos calculado las medianas de cada variable cuantitativa usada en el análisis PCA (sin transformar) para ambos grupos obteniendo:

```
medianas=splitapply(@median,X_continuas(:,setdiff(1:size(X_continuas_transformada,
↪  2), [2,3])),IDX)
```

```
medianas =

    0.9900   78.0000    4.5000    4.6400    4.5800    4.7400    4.7400    4.8300    4.490
    0.9900   84.0000    4.8600    4.9200    4.9000    4.9400    4.9500    4.9400    4.800
```

La primera variable es ratio de aceptación del anfitrión y es igual en ambos grupos, la segunda es el precio por noche, el cuál es mayor en el primer grupo, y las restantes son puntuaciones del 1 al 5 de distintos aspectos del apartamento, siendo el primer grupo superior en todas.

Una posible interpretación, es que hay 2 grupos: pisos más caros y de mayor calidad, y pisos más baratos y de menor calidad, aunque las diferencias tampoco son muy significativas y habría que contrastar con algún otro método más fiable.