Paralelismo y sistemas distribuidos

Segundo cuatrimestre 2023-2024

Software para el análisis de datos: Spark(II)

Pau Fernández
Arnau Biosca
Pau Mateo

Universitat Politècnica de Catalunya

Facultat d'Informàtica de Barcelona

Correspondence Analysis and Cluster Analysis

1. Application of Correspondence Analysis.

a) Remove observations with missing values using function na.omit().

script

b) Apply correspondence analysis by using CA() function in FactoMineR.

script

c) How many dimensions would you need to extract? Why?

Tendríamos que extraer como mínimo 3 las tres primeras dimensiones para conseguir un porcentage decente de variabilidad explicada. Con sólo dos dimensiones tan solo llegaríamos a un 56%, que es demasiado poco, y con tres dimensiones llegamos a explicar un 69.757% de la variabilidad.

d) Plot the row and column profiles on the extracted dimensions.

script

e) Interpret the plot(s) considering

(i) similarities between countries,

Como ya era de esperar, podemos ver más separacion en las ciudades al plot de la dimensiones 1 y 2 que con las dimensiones 1 y 3 (ya que la dimensión dos explica más variabilidad de los datos). Por tanto, nos fijaremos primero en el plot de la dimensiones 1 y 2 y luego miraremos el segundo plot para terminar de comparar distancias entre ciudades que al primer plot estan cerca. És más fácil ver-lo en el plot 3D.

Al primer plot vemos que Portugal y Italia estan claramente distanciadas del resto de ciudades, y comprovando también que están cerca la una de la otra al segundo plot, sabemos que estod dos pasíes se encuentran cerca considerando las 3 dimensiones extraídas. Al otro extremo, vemos un caso similar con Norway y Denmark, que estan también distanciadas del resto de países y cerca entre ellos, tanto en el primer plot como en el segundo.

Luego hay otro grupo separado: Holland, England y Ireland, que en el primer plot parecen estar cerca entre todos. Pero al segundo plot vemos que realmente England y Ireland estan separadas de Holland.

El resto de países están bastant cerca los unos de los otros, más o menos centrados al origen, excepto quizás Francia, que se separa un poco del núcleo central.

(ii) similarities between food categories.

Siguiendo el mismo procedimiento obtenemos los siguiente:
Yougurt esta claramente separado de totos los otros alimentos.
Tin.soup esta también muy separado de los otros alimentos.
Olive Oil y Garlic están muy distanciados del resto de alimentos, y podemos considerar que estan relativamente cerca.
Frozen.fish, Frozen.bread y Frozen.veggies forman un pequño grupo separado del resto, pudiendo también añadir Sweetener, que se encuentra a igual distancia entre el núcelo central y este grupo.
Instant.coffee, Tinned.fruit y Jam se separan del núcleo, pero tampoco forman un grupo explícitamente separado de el resto de alimentos, pues tienen algunos cerca, como Biscuits o Powder.soup, que ya consideramos que estàn al núcleo.

El resto de alimentos se encuentran centrados al orígen, formando el núcleo mencionado, con algunos un poco más separados, como Real.coffee, Margarine y Potatoes.

Podemos ver una clara separacion entre alimentos que podríamos considerar "sanos", o más saludables, y alimentos que tiran más hacia comida no tan saludable. Considerando las dimensiones 1 y 2, se ve que los alimentos menos saludables se sitúan hacia el primer quartil. La dimension 3 no contribuye de forma tan clara en este sentido, al menos a simple vista.

(iii) the relationship between food categories and countries.

Algunos de los grupos de países que hemos identificado estan cerca de grupos de alimentos:
Portugal y Italy estan cerca de Garlic y Olive.oil. Más específicamente, Italy está muy cerca de Olive.oil y Portugal a medio camino entre Garlic y Olive.oil (considerando aún y así que está cerca de los dos). Eso indica que mucha gente de Italy y Portugal consume aceite de oliva y ajo.
Norway y Denmark estan cerca del grupo de Frozen.fish, Crosp.bread y Frozen.veggies. England y Holland estan cerca de Jam, Tinned.fruit y instant.coffe, y son los países que más cerca estan de Tin.soup. Holland también esta cerca de Tinned.fruit y Instant.coffee pero no tanto de Jam.

En conclusión, los resultados que hemos visto encajan bastante con lo que esperaríemos: los países nórdicos són los que comen más pescado, los países ingleses los que más comida en conserva consumen, Italia y Portugal los que más aceite de oliva y ajo utilizan... Y el resto de países son más neutros, lo que quiere decir que utilizan de forma más equilibrada todos los alimentos.

2. Application of Cluster Analysis.

a) Compute a distance matrix.

script

b) Apply hierarchical cluster analysis to group countries according to their food consumption and plot dendrogram. (Try different methods but in the report just show the one that you prefer. Explain why you have chosen that method.)

Tenemos diferentes alternativas que nos están llevando al mismo resultado. Aun así, hay métodos que según las características de nuestro dataset se adaptan mejor. Tenemos que tener en cuenta que nuestro dataset tiene pocos datos, y por lo tanto es probable que tengamos bastante ruido en nuestro análisis.

Hemos escogido el Ward's Method porque es poco sensible ante el ruido, y este nos ha dado unas agrupaciones que cuadran con lo que hemos visto en la parte de Correspodence Analysis.

Estas agrupaciones son las que hemos puesto en el siguiente apartado.

c) How many clusters do you think there are? (Considering the dendrogram plot)

Basándonos en el dendrograma usando "Ward's Method" creemos que hay 3 clusters: [(Italy, Portugal, Austria, Belgium); (England, Ireland); (Luxemburg, France, Switzerland, Holland, Germany, Denmark, Norway)]

d) Draw an elbow plot showing within sum of squares per dimension. Interpret it.

En el elbow plot lo que buscamos es el punto en que la total within sum of squares sea mínima en relación a la cantidad de clusters que tenemos que formar. Este punto lo encontramos buscando el sitio donde la pendiente de la recta pasa a ser más plana. En nuestro caso, este cambio de pendiente se encuentra en el k = 3.

A partir de que creamos más de 3 clusters, la distancia entre los puntos de cada cluster va a ser más pequeña, porque los clusters podrán contener menos puntos cada uno, pero la ganancia en disminución de suma de distancias va a ser inferior a medida que aumentemos el número de clusters.

Hay otro factor que tenemos que tener en cuenta, que no solo es la distancia de los puntos dentro de cada cluster, sino también la distancia llamada 'between distance', que lo que hace es medir las distancias entre puntos que están en clusters diferentes. Para hacer esto, tenemos el Pseudo F Index,

que es una función que nos da el mejor 'balance' entre estas dos sumas de distancias. En este caso el gráfico a partir de k = 3, siempre baja. Para k = 2 este balance es más óptimo, pero no lo consideramos porque entonces la suma de distancias within sería demasiado grande. Por lo tanto, considerando estos dos plots, creemos que la k más óptima es 3.

e) Apply k-means clustering according to the number of groups you have chosen.

script

f) Compute the centroids (means) of the clusters and show them in a table.

script

g) Interpret the most important characteristics of each cluster according to the summary table given in section (f).

Con los últimos resultados, después de haber aplicado "k-means" vemos que los 3 clusters que nos han salido son: {(Italy, Portugal, Austria), (Luxemburg, France, Switzerland, Holland, Germany, Belgium, Denmark, Norway), (England, Ireland)}.

Estos cuadran bastante con los grupos que habíamos interpretado en el Correspondance Analysis, de hecho, para el primer clúster vemos como el aceite y el ajo son los alimentos más consumidos (con un 71,3% y un 73,3% respectivamente) tal y como habíamos analizado en el Correspondance Analysis. Por otro lado, los alimentos como el café y el té (con un 90,75% y un 80,37%) son predominantes en los países del segundo clúster, seguidos por la mantequilla y la margarina.

Por último, en el tercer clúster vemos como el té es también predominante (con un 99%) pero, curiosamente, el consumo de café se queda muy atrás (seguramente debido a que en Irlanda y Inglaterra tienen muy interiorizado el tomar té durante el "tea time" por su cultura. Por lo que también nos sale un consumo del 85,5 % en galletas relacionado con este momento del día de la cultura inglesa). Además, tienen un alto consumo en mantequilla y mermelada (con un 96 % y 99 %), seguramente debido a que no tienen facilidad para acceder al aceite de oliva como aquí (así que usan mantequilla en su lugar) y la mermelada con tostadas es un desayuno muy típico de allí.