HDDA Tutorial: Clustering : Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 4

Concepts

1. Why might a researcher use cluster analysis?

Cluster analysis is a multivariate technique that attempts to split a large dataset into a smaller set of groups. The elements within the groups are similar (homogeneous) in their characteristics, while the characteristics between each of the groups are diverse (heterogeneous). So a researcher might use cluster analysis for data reduction (i.e. identify major groups within a population) or to examine hypotheses about the behavior of different groups in the population, identifying market segmentation and determining target markets. Cluster analysis is also useful for product positioning and new product development, and selecting test markets.

There is always a trade-off between homogeneity within clusters (achieved by having lots of clusters) and parsimony (having as few clusters as possible and less homogeneity within the cluster). Remember, clustering is like having mini samples within the overall sample. We need to make sure the mini-samples have enough elements in them to meaningfully analyse!

2. Distinguish between hierarchical and non-hierarchical cluster techniques. In what situations would you use each of these techniques?

Hierarchical: are sequential either building up from separate clusters or breaking down from the one cluster. The advantages of these methods is that all possible solutions (with respect to the number of clusters) is provided in a single analysis. This can be explored using the dendrogram. A disadvantage of hierarchical clustering is that it is a greedy algorithm that merges the closest clusters at each step. In this way it only explores a very narrow set of clustering solutions.

Non-hierarchical: For non-hierarchical clustering, the number of clusters is chosen a priori. An advantage is that it can explore cluster allocations that will never be visited by the path of hierarchical solutions.

Although both algorithms can be used in analysis, hierarchical are suited to small data sets (since the dendrogram can be more easily interpreted) while non-hierarchical methods are well suited to large data sets.

Application

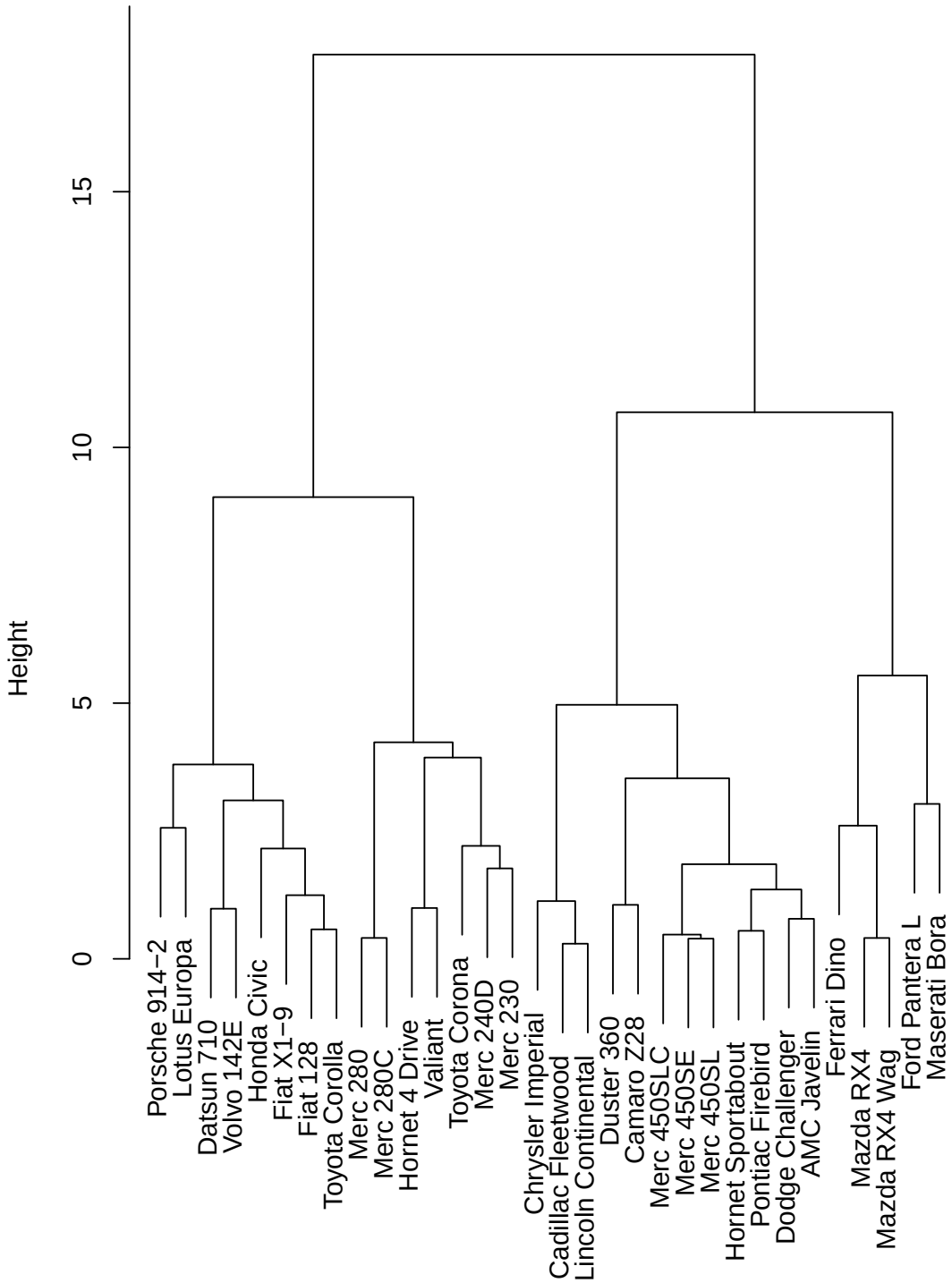
This section uses the `mtcars` dataset which is available with R.

1. Cluster the data using Ward's D2 method and Euclidean distance.

```
#First load required packages
library(dplyr) #dplyr is loaded for pipe only
mtcars%>%
  scale%>% #scale data
  dist(method = 'euclidean')->d #compute distance matrix
hclust(d,method='ward.D2')->hcl
```

2. Produce a dendrogram for the above analysis

```
#Simple plot the output of the previous code
plot(hcl)
```


Cluster Dendrogram
Height
0
5
10
15
Porsche 914−2
Lotus Europa
Datsun 710
Volvo 142E
Honda Civic
Fiat X1−9
Fiat 128
Toyota Corolla
Merc 280
Merc 280C
Hornet 4 Drive
Valiant
Toyota Corona
Merc 240D
Merc 230
Chrysler Imperial
Cadillac Fleetwood
Lincoln Continental
Duster 360
Camaro Z28
Merc 450SLC
Merc 450SE
Merc 450SL
Hornet Sportabout
Pontiac Firebird
Dodge Challenger
AMC Javelin
Ferrari Dino
Mazda RX4
Mazda RX4 Wag
Ford Pantera L
Maserati Bora
d
hclust (*, "ward.D2")

3. Discuss whether the following choices for the number of clusters are suitable or not

   - One-cluster
   - Two-cluster
   - Three-cluster
   - Four-cluster

One cluster is a poor choice since it defeats the purpose of doing a cluster analysis. The dendrogram shows that the two cluster solution prevails over a range of tolerance of about 10.5 to 17 while the 4 cluster solution prevails over a range of tolerance of about 5.5 to 9. The three cluster solution only prevails over a short range of tolerance. In this sense the two-cluster and four-cluster solutions are more stable than the three-cluster solution and should be preferred.

4. For the 2-cluster solution, store the cluster membership in a new variable.

```
memb_two_ward<-cutree(hcl,k = 2)
```

5. Repeat part 4 using:
   (a) Average Linkage
   (b) Centroid Method
   (c) Complete Linkage Method

```
hclust(d,method='average')%>%
  cutree(k = 2)->memb_two_al
hclust(d,method='centroid')%>%
  cutree(k = 2)->memb_two_cm
hclust(d,method='complete')%>%
  cutree(k = 2)->memb_two_cl
```

6. Find the adjusted Rand Index between your answer to question 4 and
   (a) Your answer to 5 (a)
   (b) Your answer to 5 (b)
   (c) Your answer to 5 (c)