Cluster Analysis (Sect. 9.6/Chap. 14 of Wilks) **Notes by Hong Li**

1) Introduction

Cluster analysis deals with separating data into groups whose identities are not known in advance. In general, even the 'correct' number of groups into which the data should be sorted is not known in advance. Most commonly implemented cluster analysis procedures are *hierarchical*.

Hierarchical clustering is a way to investigate grouping your data by creating a cluster tree. The tree is a multi-level hierarchy, where clusters at one level are joined as clusters at the next higher level.

Steps: assume a (N×K) data set **X** with N observations and K-dimensions.

Step 0. N observations are separated into N groups (one observation of dimension K in each group).

Step 1. Combine the two groups which are closest in their K-dimensional space into a new group. Now we have N-1 groups, one of which has two members.

……repeat Step1: in each step, two groups that are closest are merged to form a single larger group.

Step G. obtain N-G groups.

……

Step N-1: all N obs have been aggregated into a single group.

In practice, the procedure will not continue to the final, N-1 step, but rather be cutoff at some step G where the N data vectors cluster into N-G groups. $1<G<N-1$.

2) Distance Measures and Clustering Methods

X (N, K) data set

A. Distance between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$

You can choose any of the following measures to define the distance

- Euclidean distance: $d_{i,j} = \left[\sum_{k=1}^{K} (x_{i,k} - x_{j,k})^2 \right]^{1/2}$
- Correlation: $d_{i,j} = 1 - r^2$

  r : correlation coefficient
- Angle between pairs of vectors: $\theta = \cos^{-1}\left[\frac{x^T y}{\|x\| . \|y\|} \right]$
- Or any other measure of distance appropriate for the particular problem.

B. Cluster-to-cluster distance

Consider two groups G_1 and G_2 with n_1 and n_2 members, respectively. Even after the distance measure has been defined, there are several choices for the distance between groups:

- Single-linkage, or minimum-distance clustering (the distance between members of the two groups that are closest)

$$d_{G_1,G_2} = \min_{i \in G_1, j \in G_2} [d_{i,j}]$$

- Complete-linkage, or maximum-distance clustering (the distance between members of the two groups that are furthest away)

$$d_{G_1,G_2} = \max_{i \in G_1, j \in G_2} [d_{i,j}]$$

- Average-linkage clustering (the average distance between members of the two groups)

$$d_{G_1,G_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{i,j}$$

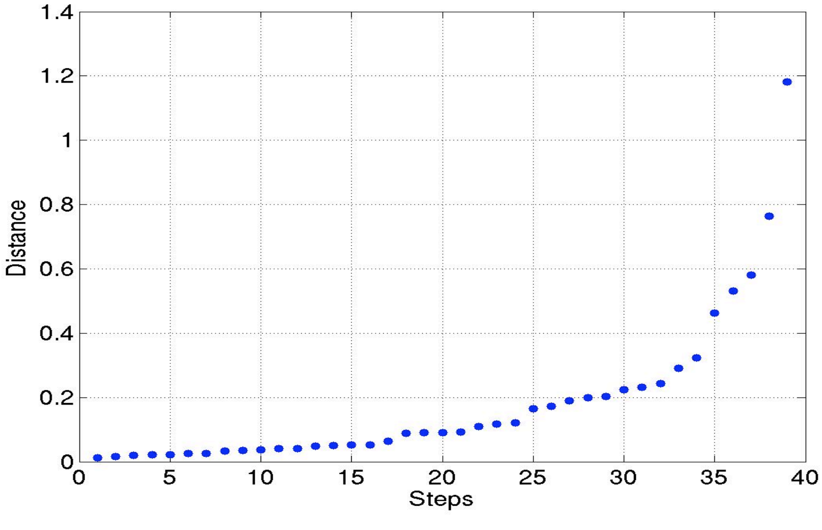
- Centroid clustering (the distance between the average of the members of the two groups)

$$d_{G_1,G_2} = \| \bar{x}_{G_1} - \bar{x}_{G_2} \|$$

- Or other appropriate choice

3) How many clusters should be chosen?

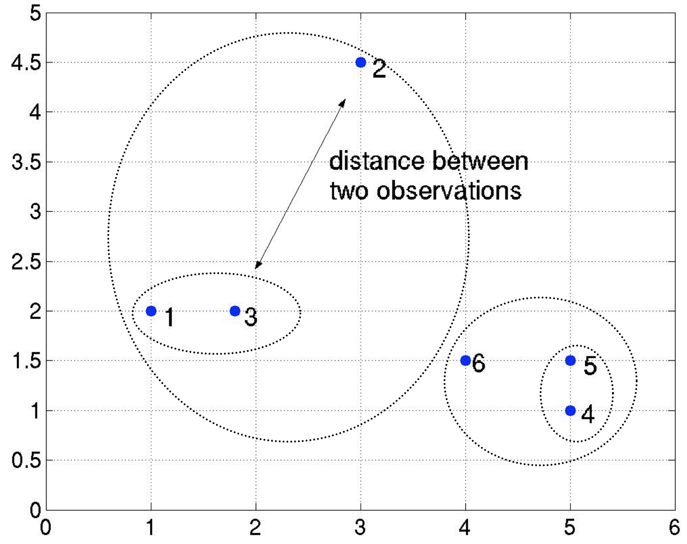
Generally the stopping point will require a subjective choice. To guide you, create a plot with the distance of the latest merged two clusters as a function of the step of the analysis. When the distance between the latest merged clusters jumps markedly, the process can be stopped just before these distances become large. In the case shown in the following figure (total 40 observations), the process could be cutoff after step 34 (6 clusters), or after 24 (16 clusters).



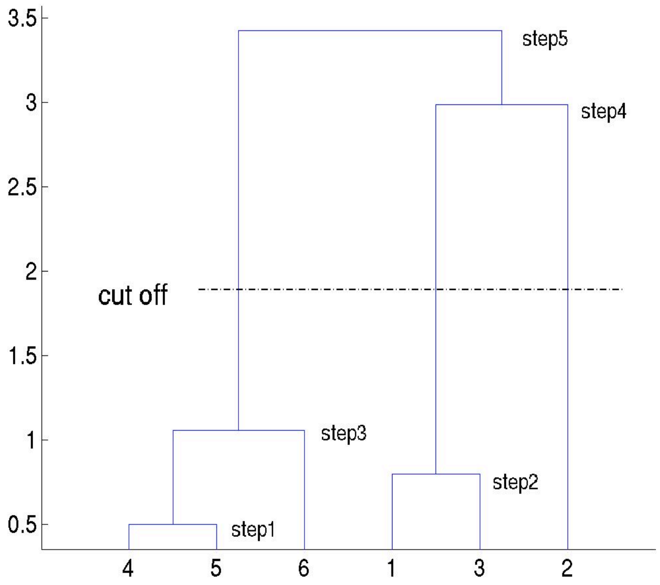
4) ***Example 1***:

Data set with 6 observations and 2-dimensions

$$X=\begin{pmatrix} 1 & 2 \\ 3 & 4.5 \\ 1.8 & 2 \\ 5 & 1 \\ 5 & 1.5 \\ 4 & 1.5 \end{pmatrix}$$



Here we use Euclidean distance as distance between two observations and Average-linkage clustering to calculate the cluster-to-cluster distance. (Note that the x and y axes have different scales).



We may cutoff the process after step 3 due to a distance jump between step3 and step4 (distance=2.98 in step4 compared with distance=1.06 in step3), so we separate the 6 observations into 3 groups (1, 3) (4, 5, 6) and (2)

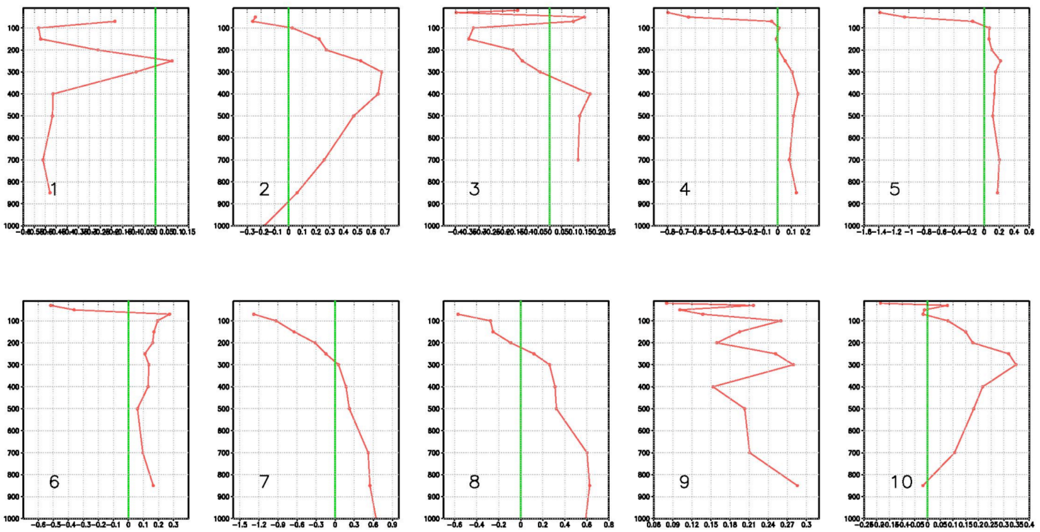
Note: the K dimensions may have different physical units. For example, the dimensions of the K-dimensional observation vector could be temperature, precipitation, humidity etc. It is advisable in this case to normalize the data before subjecting them to a clustering algorithm.

Example 2:

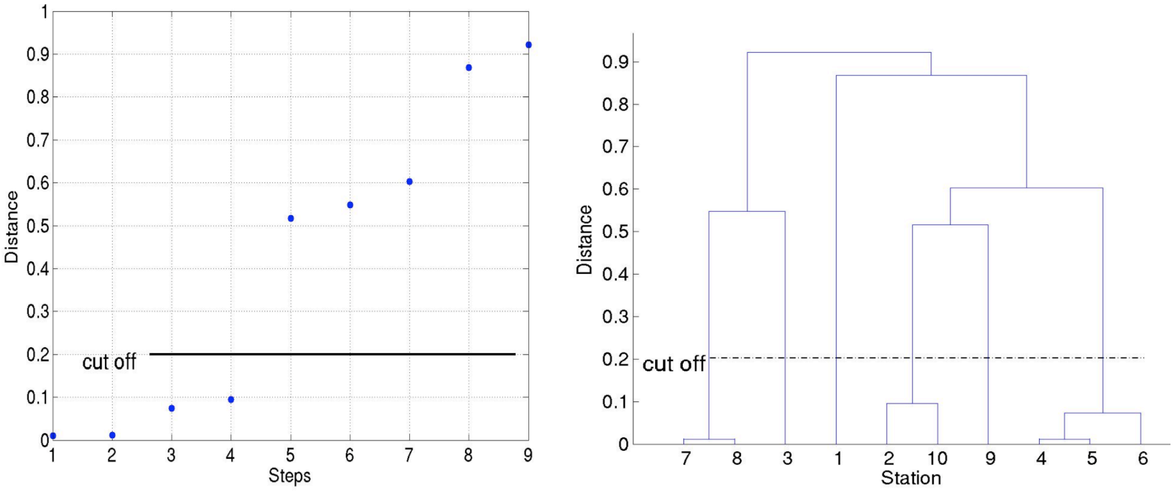
Temperature trend data set

$$X= \begin{pmatrix} x_{1,1} & x_{1,2} & .... & x_{1,15} \\ x_{2,1} & x_{2,2} & .... & x_{2,15} \\ .... & .... & & \\ x_{10,1} & x_{10,2} & .... & x_{10,15} \end{pmatrix}$$

We have temperature trends for 10 rawinsonde stations and 15-dimensions (15 level: 1000mb, 850mb, 700mb, 500mb, …10mb) for each station.



Use anomaly correlation distance and average-linkage clustering method.



If we cutoff after step 4, get 10-4=6 groups (7,8) (2,10) (4,5,6) (3) (1) (9)
If we cutoff after step 7 we get 3 groups: (7,8,3), (1), (2,10,9,4,5,6), a less satisfactory classification. In reality, with about a thousand rawinsonde trends, the first three groups (7,8), (2,10), (4,5,6) became densely populated compared with the outliers (3), (1) and (9).