COS 424: Interacting with Data

Lecturer: David Blei
Scribe: Vaneet Aggarwal

Lecture # 9
March 6, 2007

1 Review of Clustering and K-means

In the previous lecture, we saw that clustering automatically segments data into groups of similar points. This is useful to organize data automatically, to understand the hidden structures in some data and to represent high-dimensional data in a low-dimensional space. In contrast to classification where we have descriptive statistics of data, these problems are solved widely even though there is no *label* information available as is there in classification. In classification, we have data as well as a label attached to each data point, which is not there in clustering.

We discussed *k-means* algorithm in the previous lecture. In the k-means algorithm, we first choose initial cluster means, and then repeat the procedure of assigning each data point to its closest mean and recomputing means according to this new assignment till the assignments do not change. We also saw an example of k-means where we decided to divide the data into 4 clusters where we scattered the initial cluster means all over the plane randomly. The *k-means* algorithm finds the local minimum of the objective function which is the sum of the squared distance of each data point to its assigned mean. Mean locations in the example is shown by boxes in the slides. If you see the objective function, it goes down with the iterations.

2 How do we choose number of clusters k?

Choosing k is a nagging problem in cluster analysis, and there is no agreed upon solution. Sometimes, people just declare it arbitrarily. Sometimes the problem determines k. For image we may have memory constraint that decide the limit on k. We may also have a constraint on the amount of distortion that we can accept and have no memory constraint which also puts a limit on the value of k we can choose. In another examples of clustering consumers, constraint can also be number of salespeople available. We try to choose a *natural* value for the number of clusters, but in general this notion is not well-defined.

Now we discuss what happens when the number of clusters increase. Let us consider the example in the slides where there are 4 clusters(Figure 1). There are many options for the fifth group.

1. One option is that the fifth group is small or empty. In this case, the objective function remains almost the same.

2. Fifth cluster center is in the center of the figure. In this case, fifth cluster draws points from all the other clusters. In this case, the objective function decreases as the points would not have otherwise shifted to the new mean and we would still be in the first case. But, there are many points that are far from the cluster means.

3. One of the cluster subdivides into two. In this, we decrease the objective function since all the points come closer to the means. Also, since all points will be closer to the means, this is a better option than the previous 2 since more data points are affected in this case than in the second case.

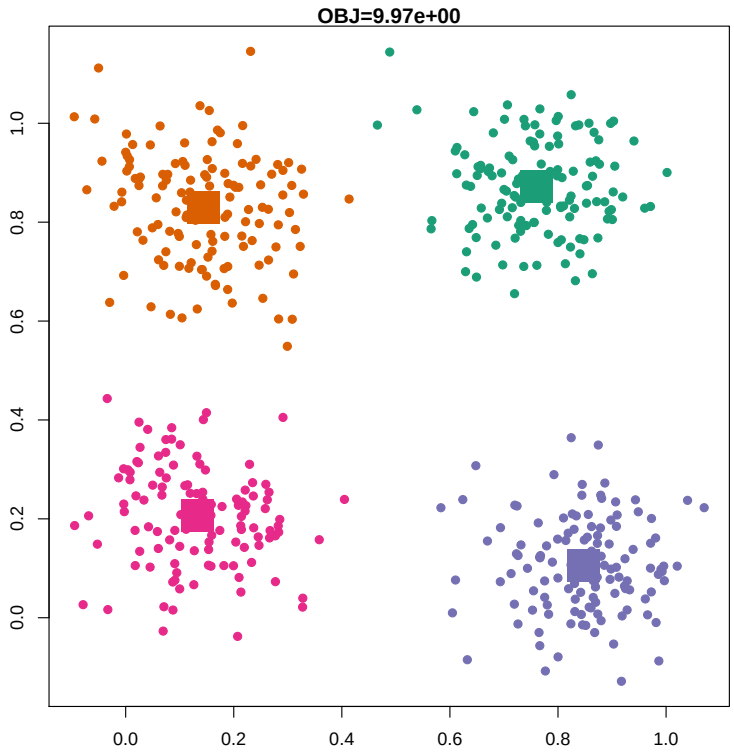

Figure 1: Division of data into four clusters

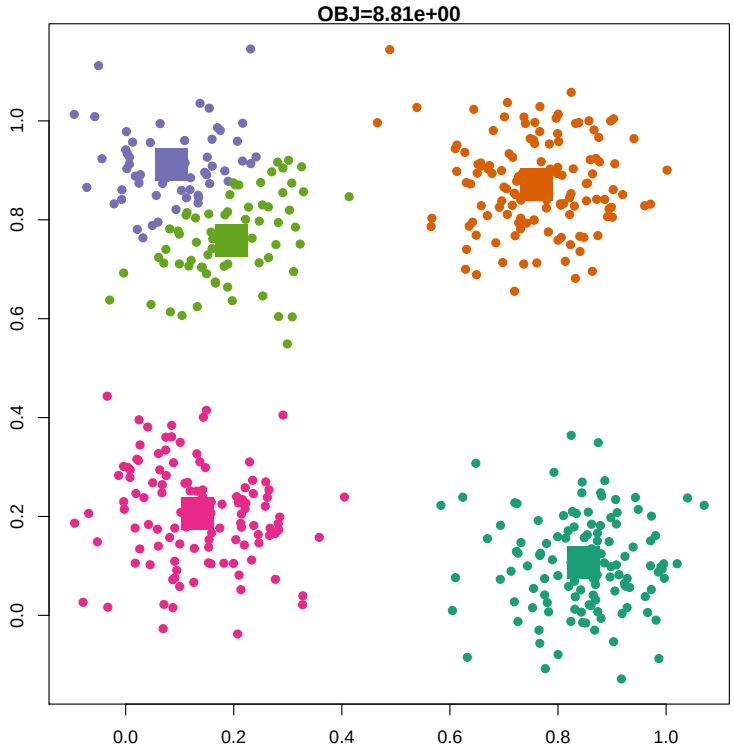

Figure 2: Division of data into five clusters

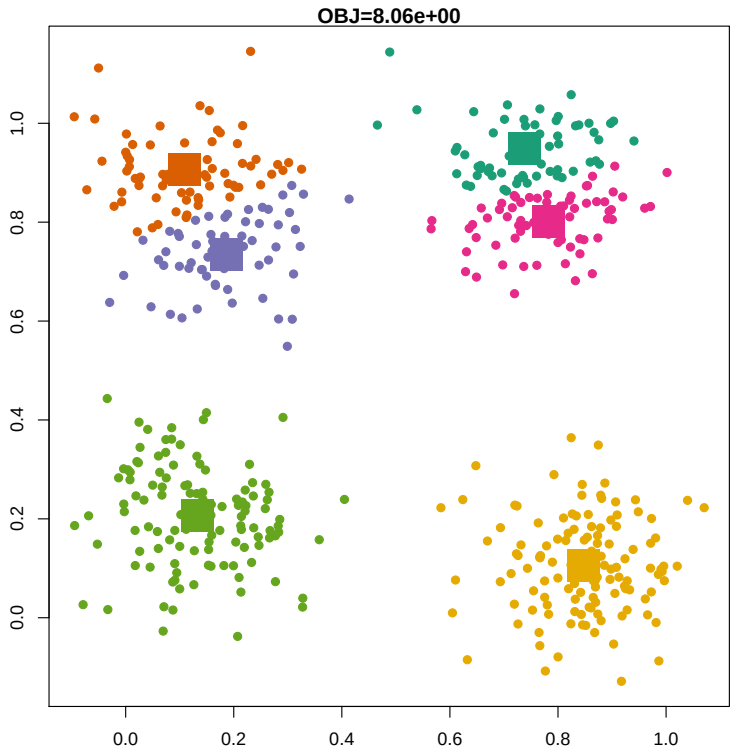

Figure 3: Division of data into six clusters

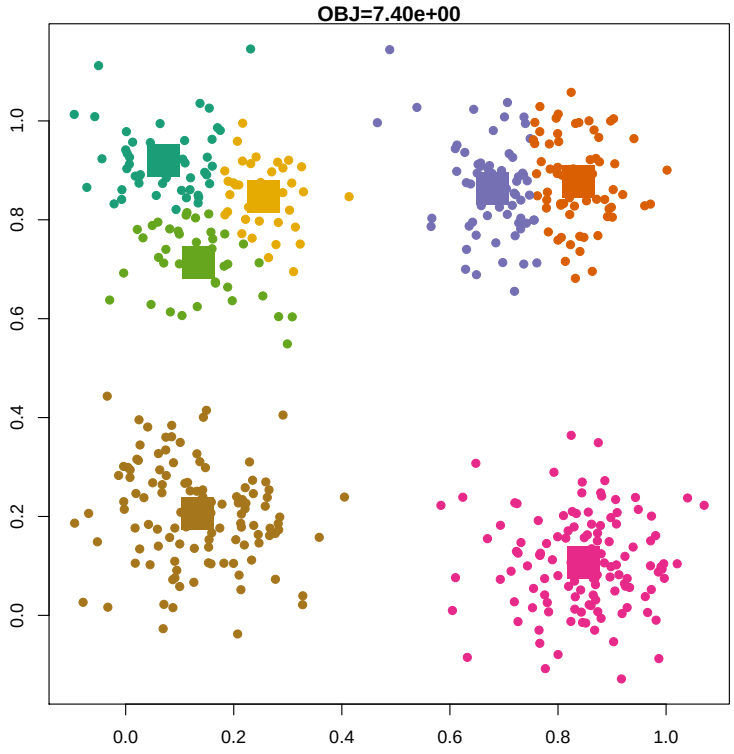

Figure 4: Division of data into seven clusters

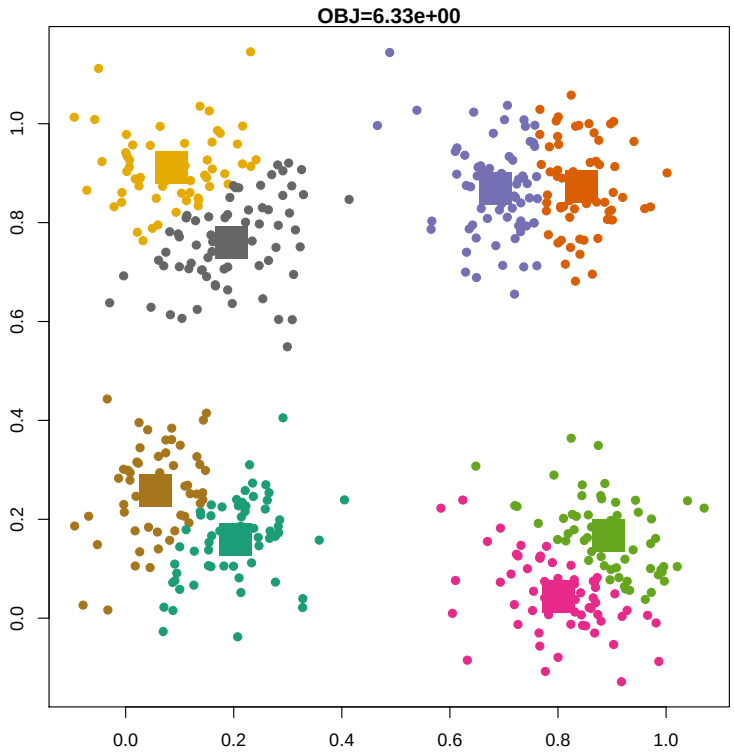

Figure 5: Division of data into eight clusters

We can see the effect of increase of the number of clusters from 4 to 5,6,7 and 8 respectively in figures 2, 3, 4 and 5 respectively. We find that one of the clusters get subdivided into two in all these cases.

When we plot the objective function against the number of clusters (Figure 6) , we find a kink between k=3 and k=5. This is because the decrease in the objective function when k increases from 3 to 4 is much higher than the decrease in the objective function when k increases from 4 to 5. This suggests that 4 is the right number of clusters. Tibshirani in 2001 presented a method of finding this kink.

3 Some applications of k-means

3.1 Archeology

This example is taken from "Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant" (Savage and Falconer) paper. The objective is to cluster the locations of archeological sites in Israel and to make inferences about political history based on the clusters. Number of clusters were chosen carefully with a complicated computational technique. The twenty-four clusters can be seen in figure 7. With these some speculations can be made, and they can be tested in actual going to the site. So, in a sense we can make some hypothesis using the clustering algorithm and must test them.

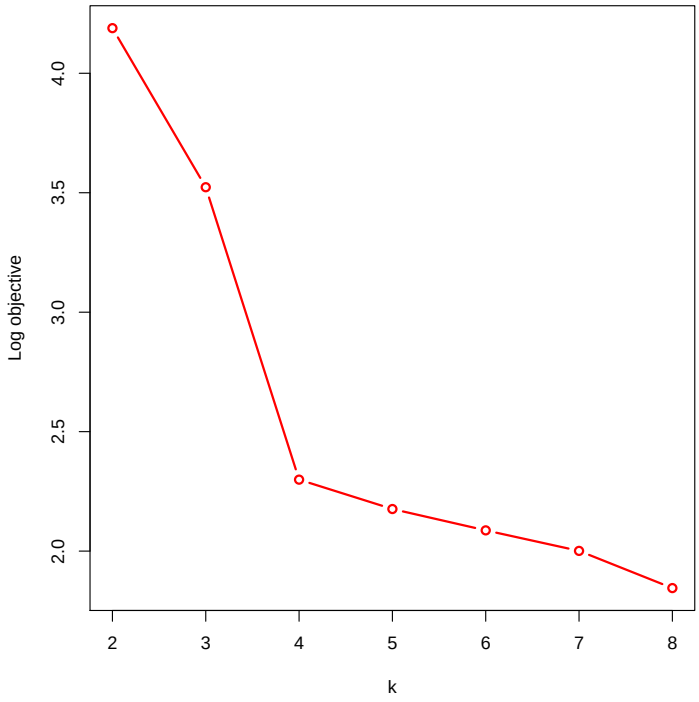

Figure 6: Plot of Log Objective function Vs. number of clusters

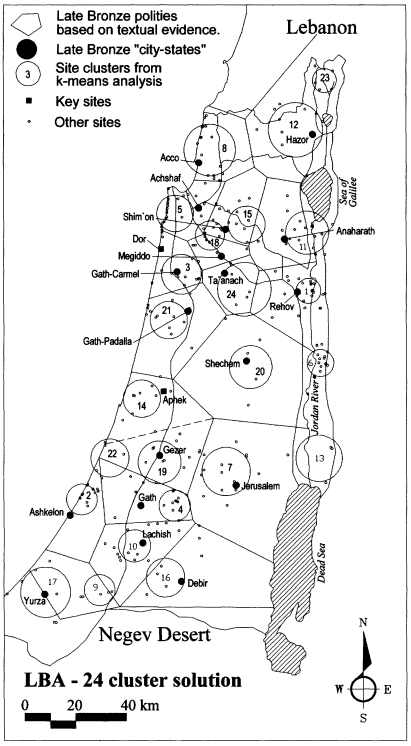

Figure 7: Clustering location of archeological sites in Israel

3.2 Computational Biology

This example is taken from "Coping with cold: An integrative, multitissue analysis of the transciptome of a poikilothermic vertebrate" (Gracey et al., 2004) paper. In this paper, carp to different levels of cold and genes were clustered based on their response in different tissues. The paper assumes 23 clusters without mentioning how it is chosen. The clustering

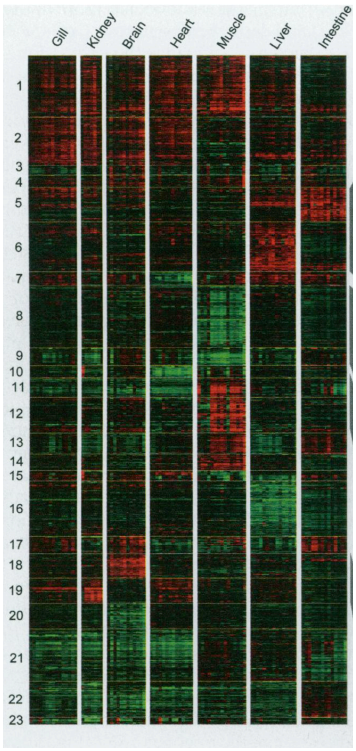

Figure 8: Clustering location of archeological sites in Israel

is shown in Figure 8. The green color represents that the gene is over-expressed while the red-color means that the gene is under-expressed. As we can see from the figure, there are some patterns in different tissues. We also see that clustering is a useful summarization tool as we are able to represent so much information in one plot. We can get some hypothesis from the clustering, which we can test later.

3.3 Education

This example is taken from "Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches" (Murdock and Miller, 2003) paper. Survey results of 206 students are clustered and these clusters are used to identify groups to buttress an analysis of what affects motivation. The number of clusters were selected to get some nice hypothesis. This hypothesis can be then verified.

3.4 Sociology

his example is taken from "Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases" (Halpert et

al., 2004) paper. Survey results of 13,998 students were clustered to understand patterns of drug abuse and sexual activity. Number of clusters were chosen for interpretability and "stability," which means that they could interpret multiple k-means runs on different data in the same way. The paper draws the conclusion that patterns exist, which is obvious since the clusters were chosen to get nice results. Also, k-means will find patterns everywhere!

4 Hierarchical clustering

Hierarchical clustering is a widely used data analysis tool. The main idea behind hierarchical clustering is to build a binary tree of the data that successively merges similar groups of points. Visualizing this tree provides a useful summary of the data. Recall that k-means or k-medoids requires the number of clusters k, an initial assignment of data to clusters and a distance measure between data $d(x_n, x_m)$. Hierarchical clustering only requires a measure of similarity between *groups* of data points. In this section, we will mainly talk about Agglomerative clustering.

4.1 Agglomerative clustering

The Agglomerative clustering algorithm can be given as:

1. Place each data point into its own singleton group
2. Repeat: iteratively merge the two closest groups
3. Until: all the data are merged into a single cluster

We can also see an example in which the similarity measure is the average distance of points in the two groups. This example can be seen in the slides.

Let us discuss some facts about the Agglomerative clustering algorithm. Each level of the resulting tree is a segmentation of the data. The algorithm results in a *sequence* of groupings. It is up to the user to choose a "natural" clustering from this sequence. Agglomerative clustering is *monotonic* in the sense that the similarity between merged clusters decreases monotonically with the level of the merge.

We can also construct a *dendrogram* which is a useful summarization tool, part of why hierarchical clustering is popular. The method to plot a dendrogram is to plot each merge at the (negative) similarity between the two merged groups. This provides an interpretable visualization of the algorithm and data. Tibshirani et al. in 2001 said that groups that merge at high values relative to the merger values of their subgroups are candidates for natural clusters. We can see the dendrogram of example data in figure 9.

4.2 Group Similarity

Given a distance measure between points, the user has many choices for how to define intergroup similarity. Three most popular choices are:

- *Single-linkage*: the similarity of the closest pair

$$d_{SL}(G, H) = \min_{i \in G, j \in H} d_{i,j}$$

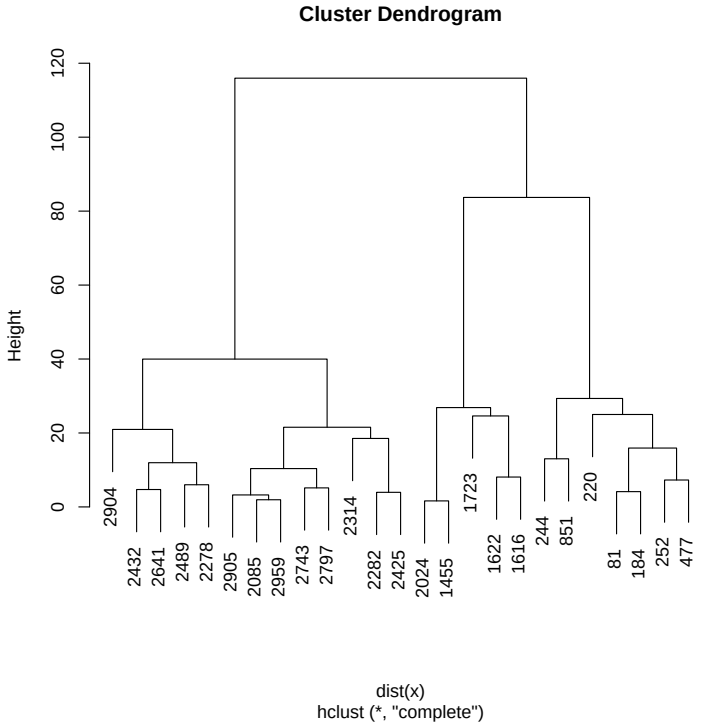

Figure 9: dendrogram of example data

- *Complete linkage*: the similarity of the furthest pair

$$d_{CL}(G, H) = \max_{i \in G, j \in H} d_{i,j}$$

- *Group average*: the average similarity between groups

$$d_{GA} = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d_{i,j}$$

4.2.1 Properties of intergroup similarity

- Single linkage can produce "chaining," where a sequence of close observations in different groups cause early merges of those groups. For example, in figure 10, suppose that the earlier grouping groups the two circled parts. The next grouping will group the two grouped parts and the indivisual point will be left alone.

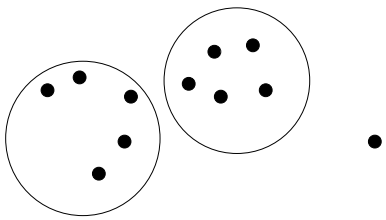

Figure 10: problem with single linkage

- Complete linkage has the opposite problem. It might not merge close groups because of outlier members that are far apart. For example, in figure 11, although groups 1 and 3 should have been clustered, but with complete linkage, groups 1 and 2 are actually clustered.

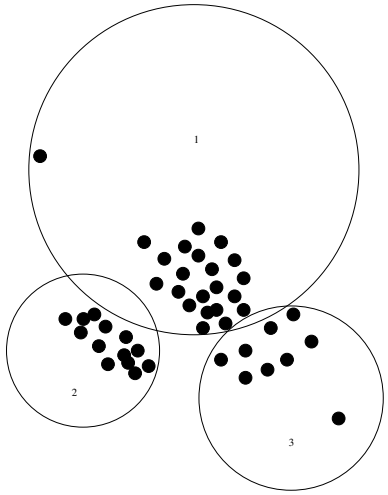

Figure 11: problem with complete linkage

- Group average represents a natural compromise, but depends on the scale of the similarities. Applying a monotone transformation to the similarities can change the results.

4.2.2 Caveats of intergroup similarity

- Hierarchical clustering should be treated with caution.
- Different decisions about group similarities can lead to vastly different dendrograms.
- The algorithm *imposes* a hierarchical structure on the data, even data for which such structure is not appropriate.

4.3 Examples of Hierarchical clustering

4.3.1 Gene Expression Data Sets

This example is taken from "Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets" (Sorlie et al., 2003) paper. In this paper, hierarchical clustering of gene expression data led to new theories which can be tested in the lab later. In general, clustering is a cautious way that leads to new hypothesis which can be tested later.

4.3.2 Roger de Piles

This example is taken from "The Balance of Roger de Piles" (Studdert-Kennedy and Davenport, 1974) paper. Roger de Piles rated 57 paintings along different dimensions. The authors of the above paper clustered them using different methods, including hierarchical

clustering. Being art critics, they also discussed the different clusters. They perform analysis cautiously, and mention that "The value of this analysis will depend on any interesting speculation it may provoke".

4.3.3 Australian Universities

This example is taken from "Similarity Grouping of Australian Universities" (Stanley and Reynlds, 1994) paper. In this paper, hierarchical clustering is used on Austrailian universities with the features such as # of staff in different departments, entry scores, funding and evaluations.

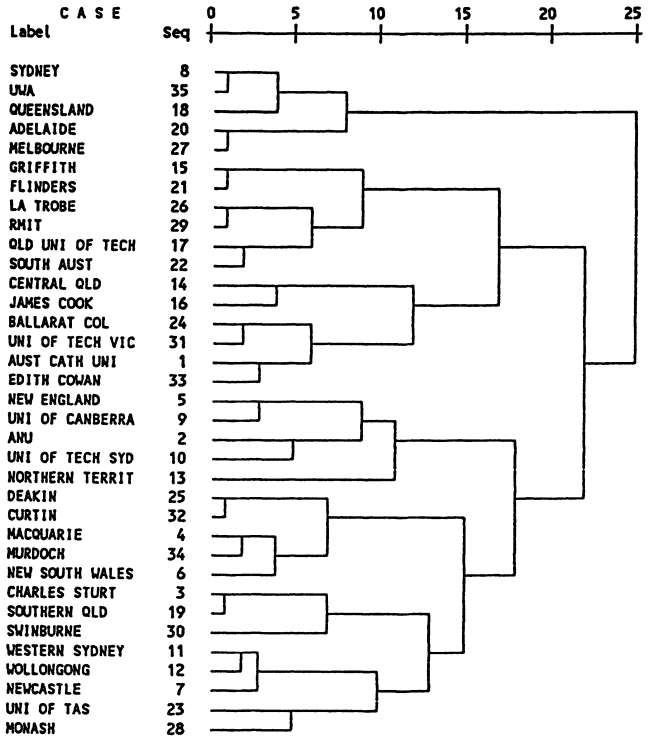

Figure 12: Dendrogram 1 for Australian Universities

The two dendograms can be seen in Figures 12 and 13 respectively. These two dendograms are different. Also, on seeing Agglomeration coefficient(Figure 14), the authors noticed that there's no kink and concluded that there is no cluster structure in Austrailian universities. The good thing about the paper is that it is a cautious interpretation of clustering, and the analysis of clustering is based on multiple subsets of the features. But, their conclusions are not good as the conclusion of "we can't cluster Australian universities" ignores all the algorithmic choices that were made. Another problem in the paper is that Euclidean distance is considered in the paper and there is no normalization. This would mean that some dimensions will dominate over others.

4.3.4 International Equity Markets

This example refers to the "Comovement of International Equity Markets: A Taxonomic Approach" (Panton et al., 1976) paper. In this paper, the data used is the weekly rates of

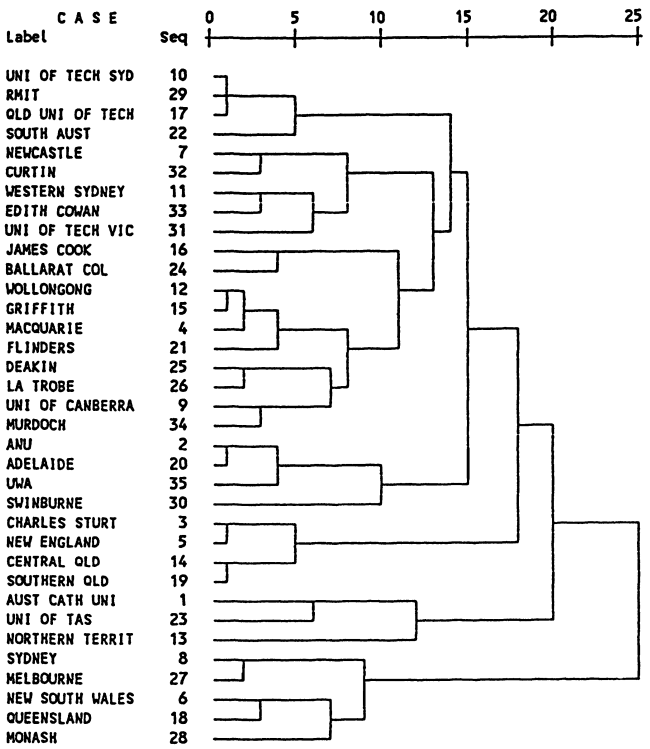

Figure 13: Dendrogram 2 for Australian Universities

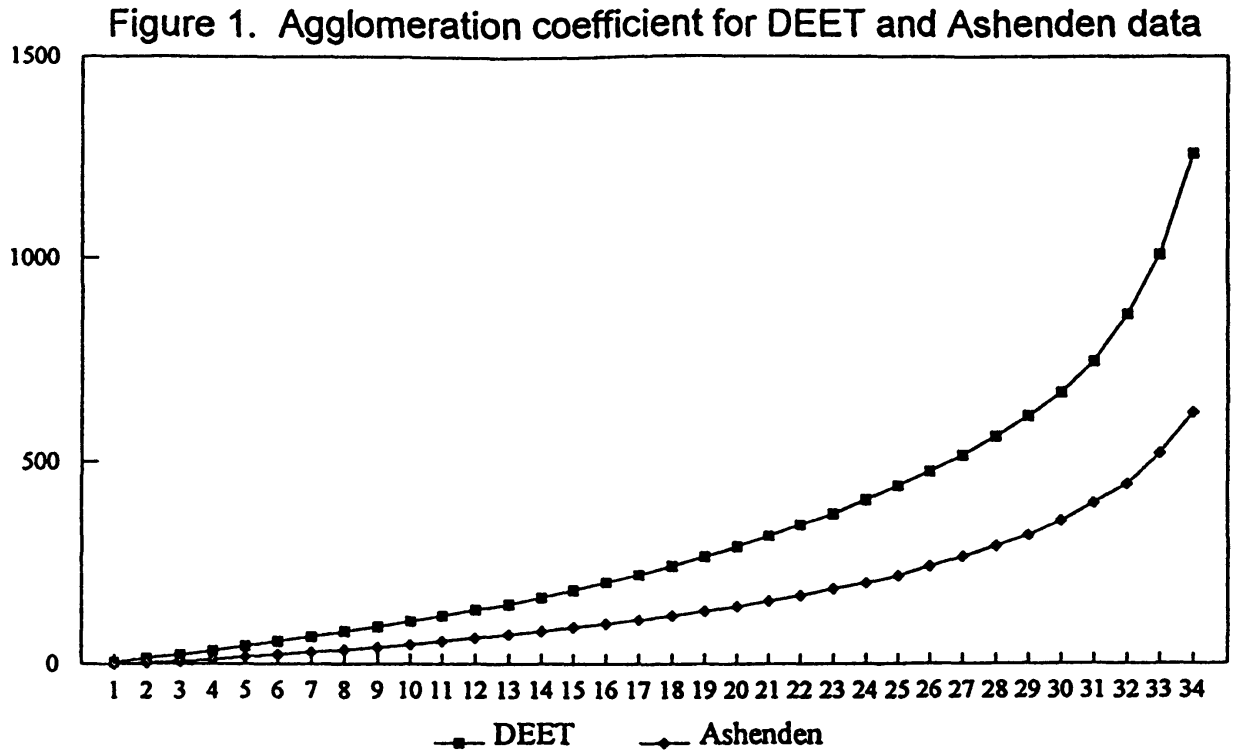

Figure 14: austrailian university agglomeration coefficient

return for stocks in twelve countries. The authors ran agglometerative clustering year by year and interpreted the structure and examined the stability over different time periods. These dendograms over the period of time can be seen in Figure 15.

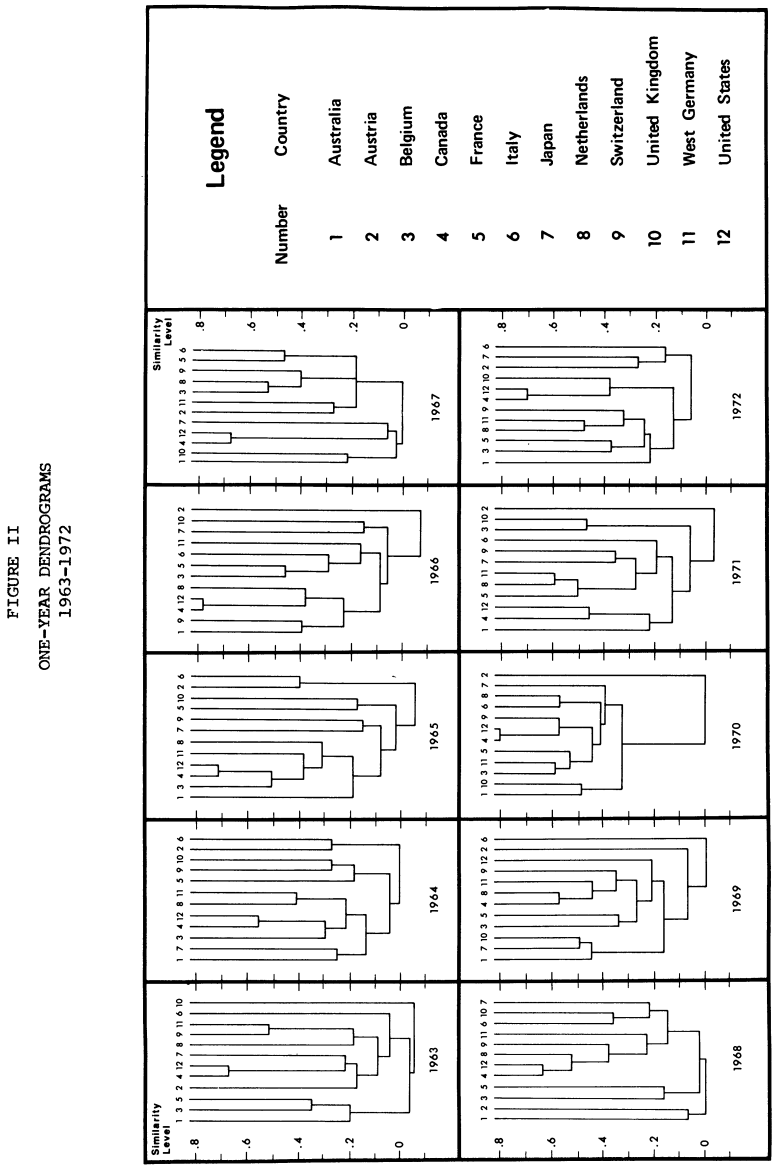

Figure 15: Dendrograms over time