

Interactive Visualization of Hierarchical Clusters Using MDS and MST

Sung-Soo Kim

Department of Applied Statistics, Korea National Open University

Sunhee Kwon and Dianne Cook

Department of Statistics, Iowa State University

Abstract

In this paper, we discuss interactively visualizing hierarchical clustering using multidimensional scaling (MDS) and the minimal spanning tree (MST). We can examine the sequential process leading to agglomerative or divisive hierarchical clustering, compare the different agglomerative methods, and detect influential observations better than is possible with dendrograms.

Keywords: Hierarchical cluster analysis, multidimensional scaling (MDS), minimal spanning tree (MST), interactive visualization, statistical graphics, data mining, grand tour, dynamic graphics.

1 Introduction

Clustering is a valuable method for understanding the complex nature of multivariate relationships, and it is widely used in taxonomy and pattern recognition. It is enjoying a resurgence in popularity in the context of data mining. Cluster analysis is a procedure for grouping individuals or objects into similarity groups, without prior knowledge of groups. It is an exploratory tool, and also a promising data reduction technique for large data sets. In general, the methods are divided into two categories: hierarchical and non-hierarchical. We focus on hierarchical methods.

In hierarchical cluster analysis the algorithms begin (1) with all objects in a single large cluster and proceed to sequentially divide them into smaller

clusters, or equivalently, (2) with all the objects as individual clusters and proceed to sequentially fuse or agglomerate them into larger clusters, based on the interpoint distances. More emphasis has been placed on the rules governing the splitting or fusing process than on the adequacy with which each cluster accurately represents the objects in the measurement space (Zupan, 1982). Graphical representation of the clusters, consequently, provides visual grouping information, and plays a complementary role to the numerical algorithms in cluster analysis.

In hierarchical cluster analysis, the clustering process can be summarized diagrammatically in tree form, i.e. a dendrogram. Using a dendrogram, we can see the sequence of successive fusions or divisions that occur in the clustering process. For example, in Figure 6 following downwards from the top to the bottom of the dendrogram we can get the divisive process of clustering, while following upwards from the bottom to the top we can get the agglomerative process. Different agglomerative methods can produce radically different dendrograms, and single observation can dramatically affect the dendrogram. Essentially the dendrogram is good for displaying sequential steps in the hierarchical algorithms, but it lacks the context of the problem, i.e., relative positions of the points and their interpoint distances. For these reasons, the dendrogram is less helpful for comparing methods and detecting influential observations. Yet, these are important parts of a cluster analysis. Because cluster analysis is inherently exploratory, it is important to examine the results produced by different methods, and assess the impact of excluding certain observations. To extract this type of information different graphical representations are usually drawn: multidimensional scaling (MDS) views, profile plots, stars, Chernoff faces, Andrews curves and scatterplots. Adding interaction and motion to these graphical displays greatly enhances the exploratory process.

Buja, Cook and Swayne (1996) discuss an interactive system where the dendrogram is dynamically linked to a grand tour scatterplot display. The grand tour (Asimov, 1985) exposes clustering through motion of points as they are spun through high-dimensional euclidean space. Points that are “close” in the data space will have similar motion patterns. The dendrogram is also overlaid on the data as it moves in the grand tour. This helps in understanding the agglomerative process in terms of interpoint distance and can assist in detecting influential observations. Now, a more common approach to graphically representing the individuals is to use MDS to find a low-dimensional representation of the data that closely preserves the cluster

structure. We discuss this approach further in this paper, and also discuss overlaying a minimal spanning tree (MST) rather than a dendrogram. The MST provides strong visual cues for unraveling cluster structure in high-dimensional data.

This paper discusses using MST with MDS for interactive visualization of hierarchical clustering. We demonstrate comparing agglomerative methods (single, complete, average, centroid, median) and detecting influential observations. The work is implemented in S-Plus, and some of it is implemented in a prototype JAVA program.

2 MDS and MST

Multidimensional scaling (MDS) is a method that provides a low-dimensional visual representation of points that preserves their relative positions in high-dimensions, using a dissimilarity matrix or rank orderings. Because MDS is a way of representing interpoint distances it can naturally be used for visually identifying clusters. It is commonly used for this purpose. An interesting feature of MDS is that it depends only on the dissimilarity matrix, and as a consequence it is useful even in situations where a raw data matrix is non-existent. See Littman, Swayne and Buja (1992) for an example of an interactive system for generating MDS representations, and Bienfait and Gasteiger (1997) for an approach to visually describing the error in MDS representations. Note also that, MDS is a very computationally intensive procedure and is only feasibly computed in real-time for small data sets, and for larger data sets the MDS representation needs to be computed off-line. The MDS representation is independent of the hierarchical cluster analysis, so alone it is of little help to understand the agglomerative process. One needs to connect the idea of the dendrogram to the MDS representation.

An approach to solving this problem, along the lines of Buja, Cook and Swayne, is to overlay a representation of the dendrogram on the MDS plot. Here, an interesting point to note is that, the dendrogram is essentially the MST when the agglomerative method is single linkage. It is a very neat representation of the dendrogram in this case. The MST is a tree which completely connects the all objects without any closed loop and minimizes the sum of the edge lengths of the tree. Given n points, a tree spanning these points is any set of edges joining the pairs of points such that there are no closed loops, each point is visited by at least one edge, and the tree is connected (Gower

and Ross, 1969). When a set of n points and $n(n - 1)/2$ possible edges are given, a minimal spanning tree (MST) is a spanning tree for which the sum of edge lengths is smallest. MST is closely related to single linkage cluster analysis. Since single linkage joins clusters by the shortest length between them, single linkage cluster can be constructed from the MST. So a dendrogram for single linkage cluster can be drawn directly from the MST (see Gower and Ross, 1969). The MST has been used for identifying influential observations by Jolliffe *et al.* (1995), and for highlighting the inaccuracies present in the low-dimensional MDS representations of high-dimensional data by Gorden (1981), Krzanowski (1988), Bienfait and Gasteiger, (1997). The MST provides valuable visual cues to cluster structure when used in conjunction with a scatterplot.

The structure overlaid on the plot can be adapted from the strict MST to be a useful representation of the dendrogram even with other agglomerative methods. For example, within the groups use MST to visualize the nature of the interpoint connectedness here, and between groups connect the elements using a representation that matches the agglomerative method. We will simply connect the closest elements between groups by a line. This provides sufficient information for us to compare methods.

3 Adding Interaction

We provide MDS representations overlaid by the MST in an interactive setting, allowing the user to change the number of final clusters to examine the agglomerative or divisive sequence, compare different agglomerative methods, and the influence of particular objects on the final solution. We introduce an example to demonstrate the methods: villages data introduced by Morgan and Shaw (1982), and used again by Seber (1984) and Jolliffe *et al.* (1995).

The villages data comes as a similarity matrix, measuring the frequency that the same word is used for a range of 60 items amongst 25 villages, is given on our web page. The similarities (s_{ij}) were converted to dissimilarities (d_{ij}) by $d_{ij} = \sqrt{2(1 - s_{ij})}$. Figure 1 displays the 2-dimensional MDS representation, with the MST overlaid. A 2-cluster single linkage solution is displayed: points in the same cluster are connected by solid edges of the MST, and the dashed line represents the edge separating the two clusters. (On the computer screen color is also used.) Here we see that two clusters are separated between the points 13 and 24, which corresponds to the longest

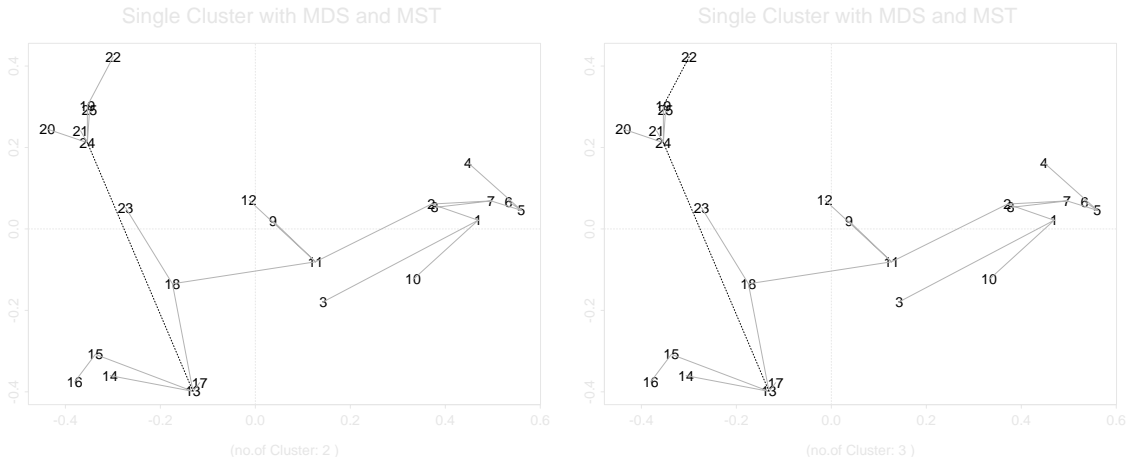


Figure 1: Single linkage (left) 2-cluster, (right) 3-cluster

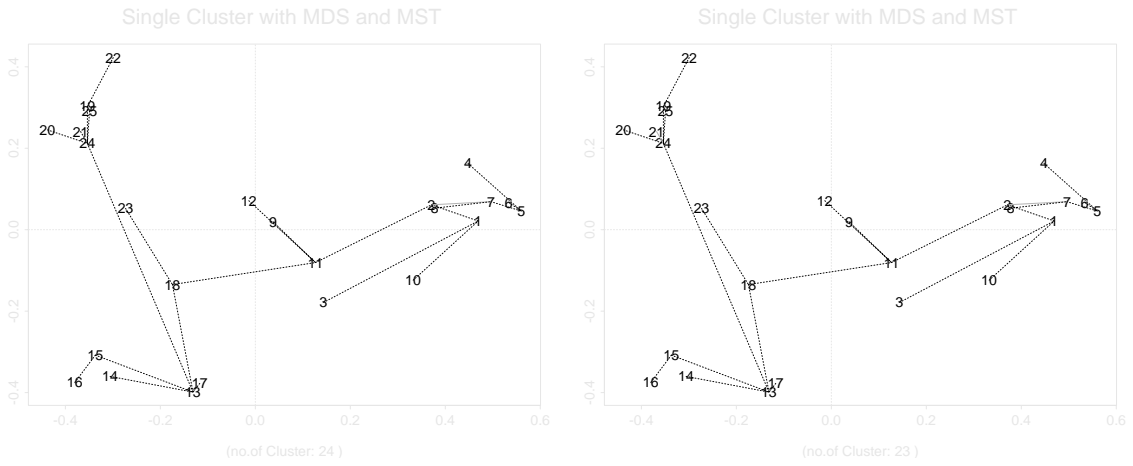


Figure 2: Single linkage (left) 24-cluster, (right) 23-cluster

edge of the MST. Figure 1 also displays the 3-cluster single linkage solution. The only difference from the 2-cluster solution is that point 22 is separated into its own cluster. We can sequentially interactively choose the number of clusters and watch the process of divisive clustering.

Conversely, if we start from 24 clusters and sequentially reduce the number of clusters we can see the steps of agglomeration. Figure 2 shows 24-cluster single linkage solution, where points 2 and 7 are connected, which means the distance of these two points is the shortest among all pairwise distances. Points 21 and 24 are connected in the second step as shown in the 23-cluster single linkage solution.

This is more useful than similarly working through the dendrogram because we can see the relative positions of points using the MDS representation. Also using MST superimposed on MDS we can assess the distortion that exists in a two-dimensional representation of dissimilarity matrix. For example, in Figure 1 the points 3 and 11 are close together in diagram, but MST shows that point 3 is closer to point 1 than to point 11.

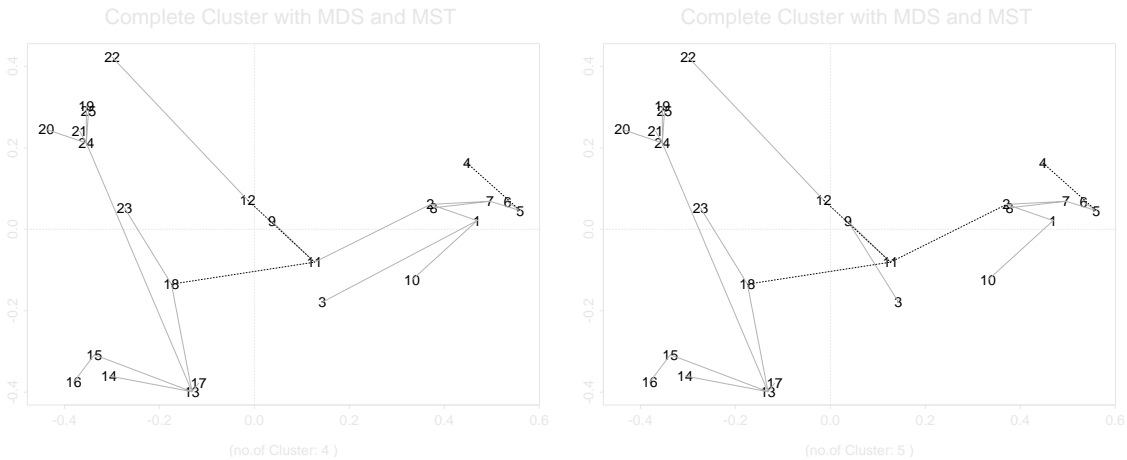


Figure 3: Complete linkage (left) 4-cluster, (right) 5-cluster

Single linkage cluster analysis is directly related to MST since the distance between groups are defined by the smallest distance between each element of groups, and so MST can be used to visualize single linkage clustering process easily. However, other hierarchical clustering methods, complete, average, centroid and median cluster methods, are not directly related to MST, and so MST cannot be used directly to visualize these clustering process interac-

tively. We modify the representation as follows:

- Within groups, the points are connected by the MST.
- Between groups, a line is drawn which connects the closest elements of the two groups.

Note that, MST is re-computed within each cluster, after sequential splits. Using this modified MST, we can visualize the process of clustering interactively for any hierarchical cluster analysis. Figure 3 shows the 4-cluster complete linkage solution where the points are clustered as $(1,2,3,5,6,7,8,9,10,11)$ $(13,14,15,16,17,18,19,20,21,23,24,25)$, $(12,22)$ and (4) . Here dashed lines between groups denote the smallest distance between two points connecting groups. Figure 3 shows 5-cluster complete linkage solution where the points are clustered as $(1,2,5,6,7,8,10)$, $(3,9,11)$, $(13,14,15,16,17,18,19,20,21,23,24,25)$, $(12,22)$ and (4) , which means that cluster $(1,2,3,5,6,7,8,9,10,11)$ is divided by two clusters of $(1,2,5,6,7,8,10)$ and $(3,9,11)$ at the next step. Using this sequence we can follow the hierarchical clustering process agglomeratively or divisively and get the difference between steps visually.

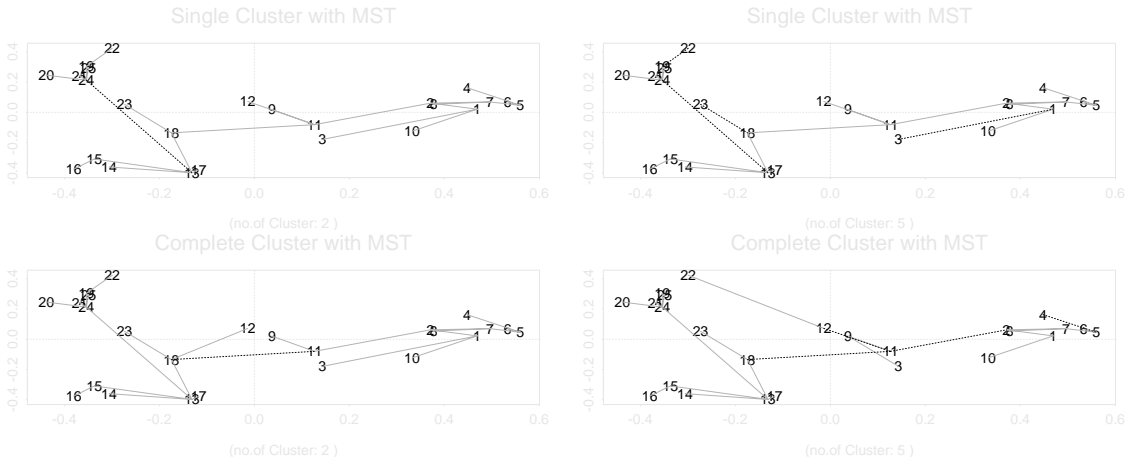


Figure 4: Single and complete linkage, (left) 2-cluster, (right) 5-cluster

For the purpose of comparing agglomeration methods we allow two plots to be made simultaneously. Figure 4 shows the 2-cluster solution of single linkage and complete linkage. Here we can see that using single linkage the two groups are divided by points $(13,24)$, and for complete linkage the two

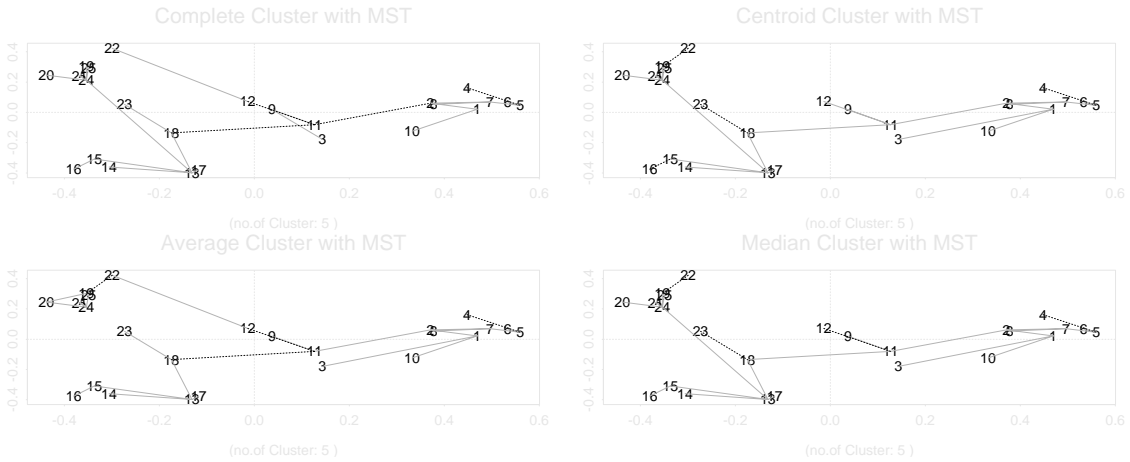


Figure 5: 5-cluster (left) Complete and Average, (right) Centroid and Median Linkage

groups are divided by points (11,18). It is generally well known that the single linkage method concentrates on seeking clusters which are isolated from each other, producing elongated clusters. On the other hand the complete linkage method concentrates on their internal cohesion, producing spherical clusters. Figure 4 shows this phenomenon. We can clearly see the difference of two methods as we increase the number of clusters to 5. Generally, it is well known that the complete, average and centroid methods lead to spherical clusters exhibiting high internal affinity, and the median methods weights towards the objects most recently added to the cluster. In our program we implemented five hierarchical cluster methods - single, complete, average, centroid and median linkage methods, so we can compare these clustering methods interactively. Figure 5 shows the 5-cluster comparisons of (complete,average) and (centroid, median) respectively. We can see an interesting fact that several points with only one degree comprise the separate clusters with centroid and median methods. (Degree is the number of edges incident with an observation.) Through interactively running the program, we can also see that in the centroid method points (4,22,23,16) are separated from other points sequentially, while in median method points (4,23,12,22) are separated sequentially, and these two methods are very similar in comprising clusters. From Figures 4 and 5, we can see the differences between several clustering analyses visually.

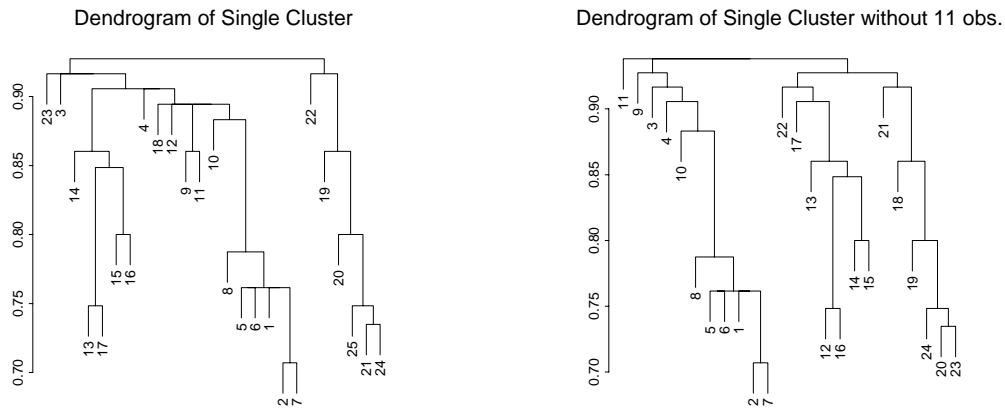


Figure 6: Dendrogram of single linkage, (left) all 25 villages, (right) without village 11

Detecting influential observations

Cluster analysis is very sensitive to one or two observations, in the sense that their removal may cause large changes in the results of the cluster analysis. Influential observations are defined as those which cause a large change in cluster solution when they are omitted. It is important here to recognize this definition. Points that are influential can be more insidious here than in other types of applications. Outliers to the general point cloud are not necessarily influential, but rather will be peeled off as individual clusters. More influential points can be found in the confluence of clusters, points that fall in “average” positions between cluster centers, or form daisy chains between clusters. Single linkage clustering is especially prone to influence from this type of point. Jolliffe *et al.*(1995) gives an example showing that removal of a single observation has a substantial effect on the results, using the similarity matrix for the village data. Table 1 shows the 5-cluster single linkage solutions with full data and without observation 11, and dendrograms with full data and without observation 11 are given in Figure 6 respectively. From Table 1, it is clear that the removal of single observation has a substantial impact on the results of single linkage cluster analysis for a 5-cluster solution. Despite large discrepancies near the top of these dendrograms there is a great deal of similarity between two solutions. This is very difficult to see from the dendrograms, and so it is not easy to determine how influential

an observation is based on the dendrograms. Jolliffe *et al.*(1995) considered using the MST to find potentially influential observations in a single linkage cluster. Points with a large degree within suitable radius may have a great effect on cluster analysis.

All villages	Without village 11
{1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18}	{1,2,3,4,5,6,7,8,10}
{3}	{9}
{19,20,21,24,25}	{19,20,21,22,24,25}
{22}	{12}
{23}	{13,14,15,16,17,18,23}

Table 1: Single linkage cluster analysis on 25 villages - 5-cluster solution with and without village 11

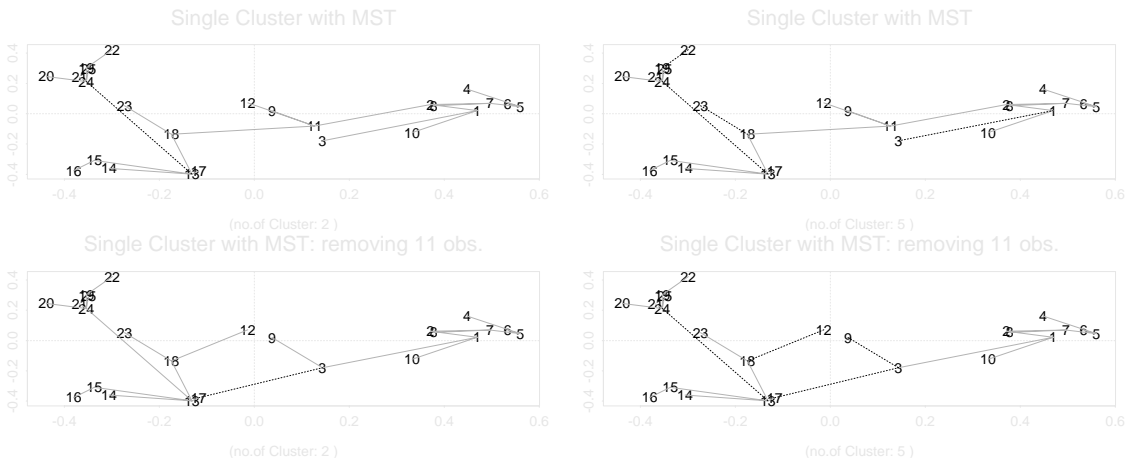


Figure 7: Single linkage with and without observation 11, (left) 2-cluster, (right) 5-cluster

The approach that we described in Section 2 also helps to visualize influential observations interactively. Figure 6 shows the single linkage solution with and without observation 11 for 2-cluster and 5-cluster solutions respectively. For all the data, the two clusters are divided between points (13,21) and for the data without observation 11, two clusters are divided between points (3,13). So the solutions are quite different when observation is excluded, and it is easy to understand what happens: observation 11 is

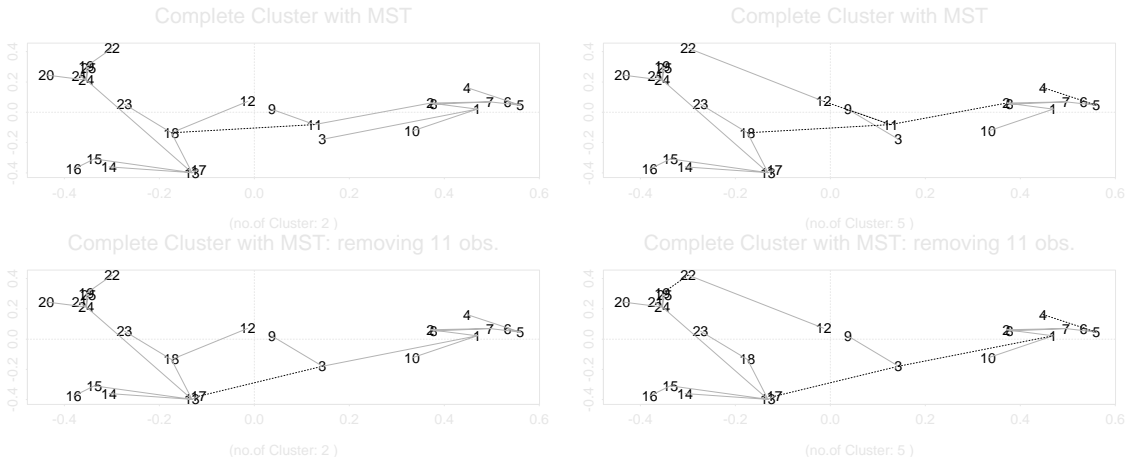


Figure 8: Complete linkage with and without observation 11, (left) 2-cluster, (right) 5-cluster

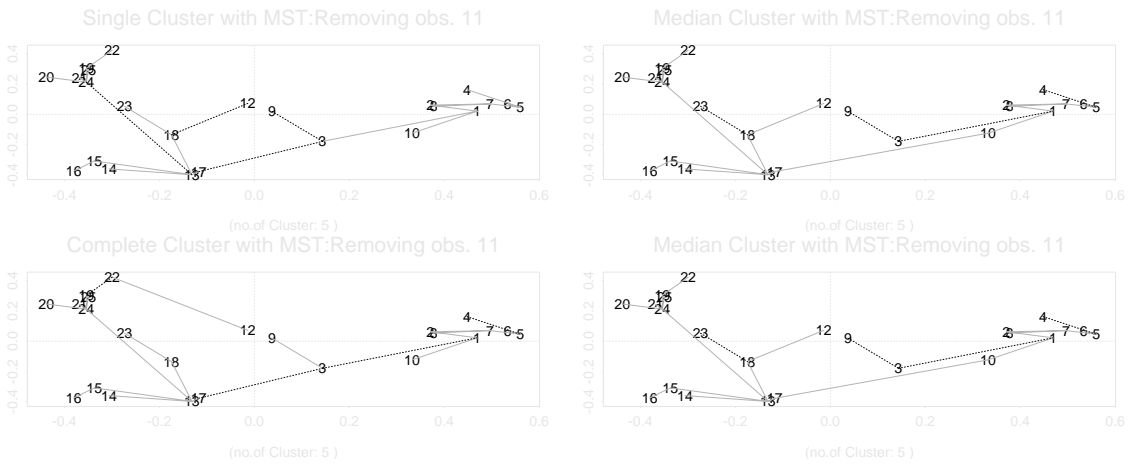


Figure 9: 5-cluster without observation 11 (left) Single and Complete, (right) Centroid and Median linkage

intermediate between 3 and 13 and acts like a connecting link in a chain. This effect is also present in the 5-cluster solution shown in Figure 7.

All of what has been described above can be done interactively. The interactive setting helps uncover and understand the nature of influential points, and also helps illuminate how persistent the impact is through numerous stages of the agglomerative clustering process. Information on running the software interactively is available on the web page.

Here we have one question that observation 11 is also influential in other cluster methods. Figure 8 shows the complete linkage solutions with and without observation 11 for 2-cluster and 5-cluster solutions respectively. We can see that there is little difference with and without observation 11, which means that observation 11 is not influential in complete linkage cluster. This is interesting fact that whether observations are influential or not depends on the cluster methods. We can also see the effect of observations in other cluster methods.

If we can compare cluster methods without observations in sequential order interactively, it is helpful to see the role of observations in different hierarchical cluster methods. The implemented graphic displays are similar to Figure 5 without a certain observation. For example, Figure 9 shows the 5-cluster solutions of (single, complete) and (centroid, median) respectively. Using this procedure, we can see the effect of observations visually and compare cluster methods without some observations interactively. We have considered only one observation to see the effect of removing observations, but it can be extended to see the effect of removing several multiple observations sequentially or simultaneously.

4 Adding Motion

It is important to go beyond 2-dimensional MDS representations to fully understand the inherent cluster structure. To demonstrate this we introduce a second example: flea beetles, first discussed by Lubischew (1962). In this data there are 3 different species of beetles and 6 measured variables, and there are 3 neatly separated clusters. We know the species identity so it is a good data set to use to test cluster algorithms, especially since the presence of a few influential points confounds all the hierarchical methods. Figure 10 shows the MDS plot, overlaid by the MST. Looking at this plot it is quite easy to recognize the three clusters, although two points (6,10) are not easily

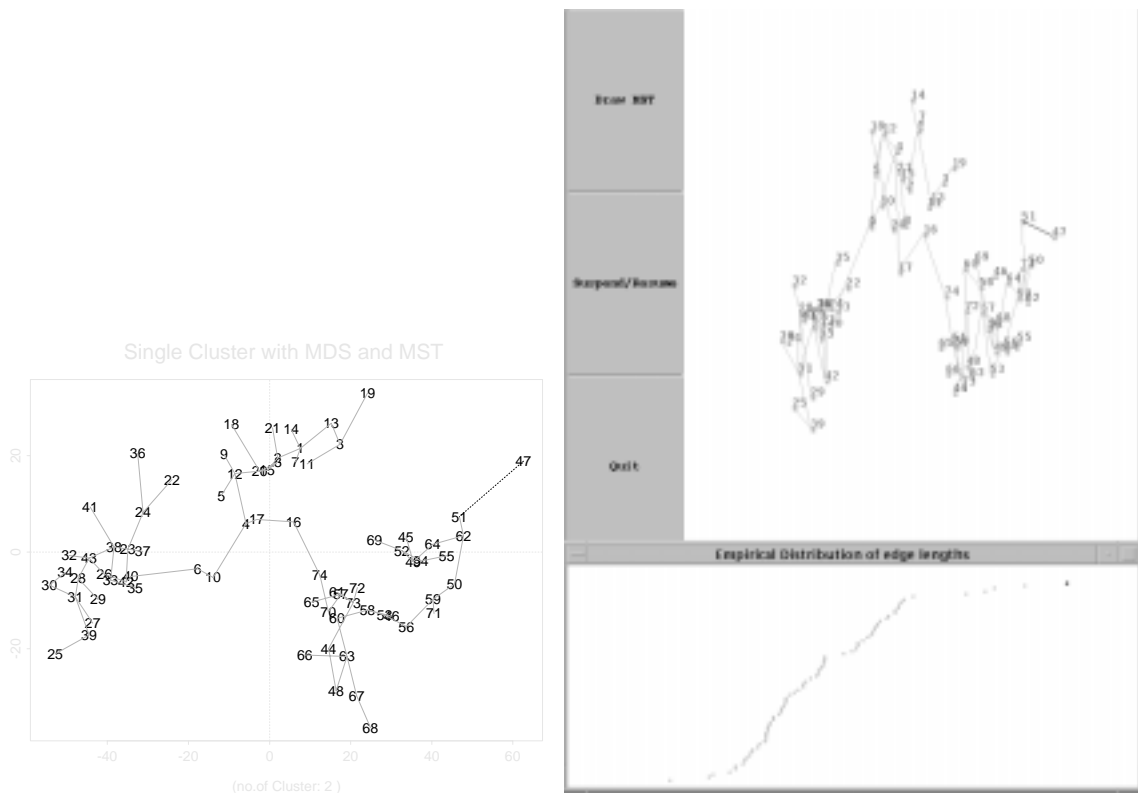


Figure 10: 2-cluster solutions in the flea beetle data, (left) MDS representation, (right) one projection from the grand tour

placed in a cluster. But MDS can be misleading because the construction only preserves “relative distance”, so points that are “far” in the original space could appear “closer” in the MDS space. Also even with MST overlaid in the left graph of Figure 10, we can’t identify which cluster the two points, 6 and 10, belong to, between the upper right one and the left one.

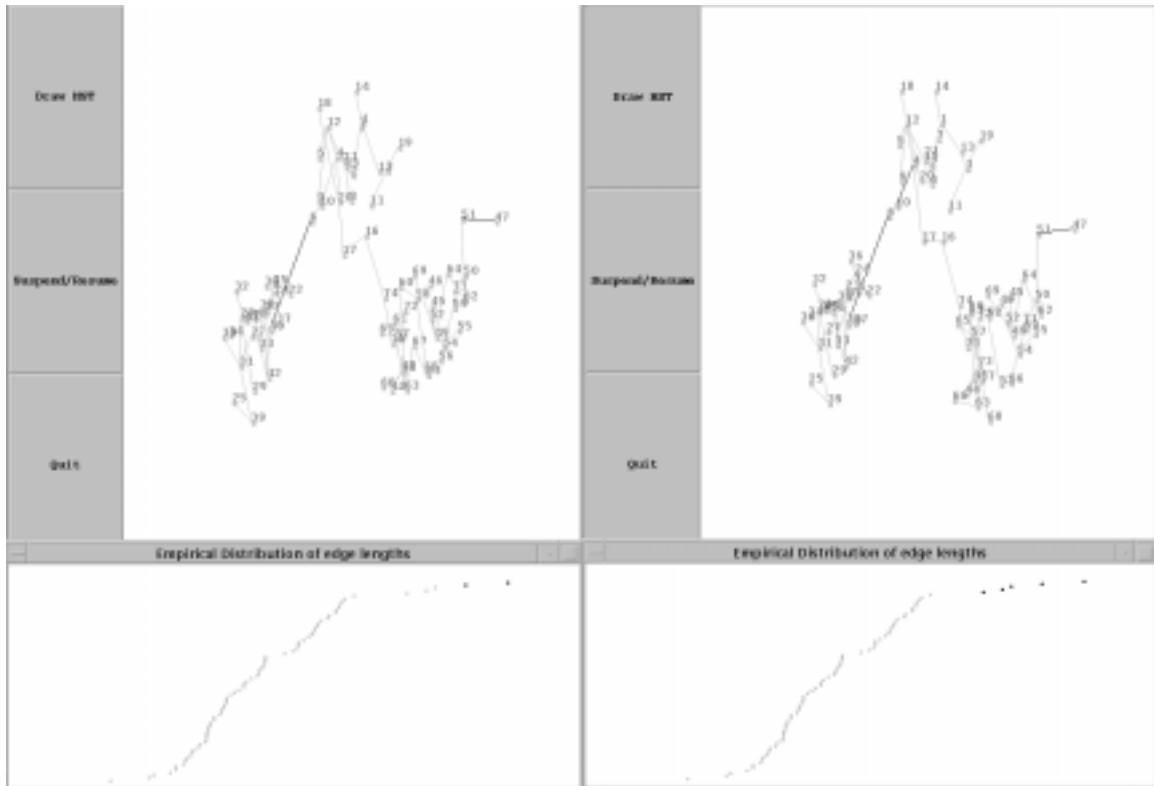


Figure 11: 3- and 6-cluster solutions in the flea beetle data, shown in projections from the grand tour.

This confusion can be clarified using a grand tour, where the points are placed in motion in the original data space. Similar motion paths are easily detected visually and cluster patterns can be better understood. Figure 10 shows one grand tour view where the cluster identity of points 6 and 10 is clear. We have a prototype JAVA program which runs the grand tour, calculates and overlay the MST in the main grand tour window. Also it displays a graph of the ordered MST edge lengths in the separate window. With its linked brushing between this window and the main grand tour window the

user can disconnect the edges and look at the resulting partitions either by scrolling the edge length graph or brushing the individual points representing the MST edges. Since the partitions made by scrolling the edges in the order of their lengths are same as single linkage clustering, we can examine the cases where single linkage clustering fails as follows.

Here the largest edge length is brushed and we can see that it corresponds to point 47 and point 51. This would be the 2-cluster single linkage solution. Further brushing reveals the 3-cluster and 6-cluster solutions (Figure 11). We see that the second longest edge is between point 6 and point 40, so at this stage we do get a “real” split of two clusters, rather than a peeling of outliers. Interestingly, the next real split doesn’t come until the 10th longest edge is trimmed, i.e. 11 clusters are made.

Our JAVA software can accept MDS result as input and perform grand tour for this. This way we can extend 2-dimensional MDS representation to higher dimension.

Also with motion added the clustering result can be less affected by the existence of a few influential points. To detect the effect of influential points on usual hierarchical methods (without motion), we take careful notice of points in intermediate locations between clusters, for example, 6, 10, 16, 74. (Note that, points 22, 41, 47 are primarily nuisance outliers that get peeled off by the hierarchical methods into individual clusters.) To assess these, we remove the points sequentially to assess their impact on the final solutions. Interestingly, removing observation 6 is sufficient to cleanly split two clusters, and removing observation 74 is sufficient to cleanly split the remaining two clusters (Figure 12). This suggests that removing these two points would enable the hierarchical methods to perfectly cluster this data.

But by performing grand tour guided by projection index which is one of exploratory clustering tools with motion, it is possible to perfectly cluster this data, without removing any observations.

5 Conclusions

Hierarchical cluster analysis can be summarized in a dendrogram, which gives the agglomerative and divisive process. However, it does not provide exploratory representations of data, and it becomes unwieldy for even moderate sample sizes. So, we need alternative methods to efficiently compare

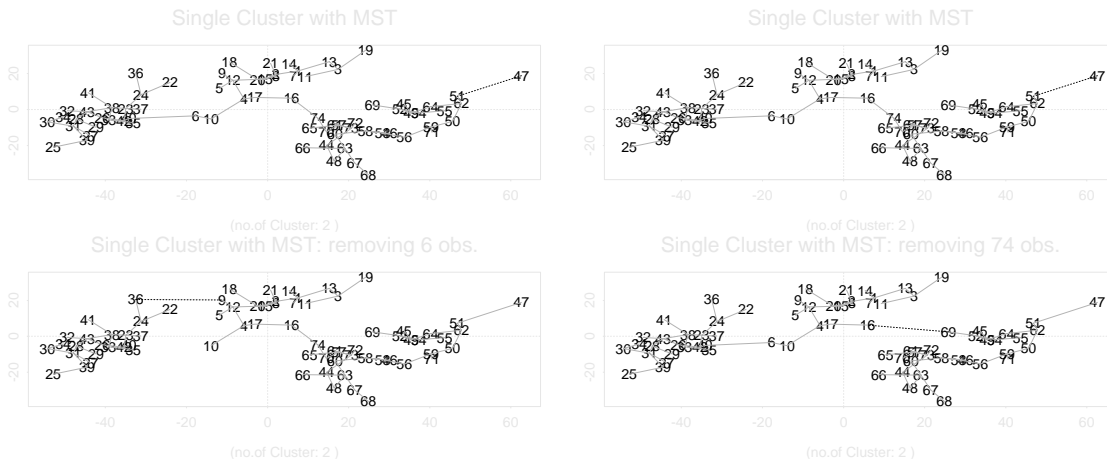


Figure 12: 2-cluster solutions flea beetle data with observation (left) 6, (right) 74, removed.

clustering methods and to see the effect of influential observations in cluster analysis.

In this paper we presented an approach for interactive visualization of hierarchical clusters using MDS and MST, from which we can obtain several benefits related to cluster analysis: (1) we can see the sequential processing of agglomerative or divisive hierarchical clustering interactively, (2) we can compare several cluster methods visually and interactively, (3) we can see the effect of influential observations in cluster analysis interactively, (4) we can examine relationships existing between MDS and cluster analysis visually and (5) we can assess the distortion that exists in a two-dimensional representation of high dimensional data.

We considered only hierarchical cluster analysis, however, it could be applied to non-hierarchical methods such as k -means clustering, where it would be helpful to see the effects as k changes.

The implemented S-plus and JAVA source programs and associated documentation can be obtained from Web site:

www.public.iastate.edu/~dicook/research/papers/Sungsoo.Kim/paper.html.

References

- [1] Afifi, A. A. and Clark, V.(1990) *Computer-Aided Multivariate Analysis*,

2-ed., Van Nostrand Reinhold Co. New York.

- [2] Asimov, D.(1985) *The Grand Tour: A Tool for Viewing Multidimensional Data*, 6(1), 128-143.
- [3] Bienfait, B. and Gasteiger, J. (1997) Checking the Projection Display of Multivariate Data with Colored Graphs, *Journal of Molecular Graphics and Modelling*, 15, 203-215 and 254-258.
- [4] Buja, A., Cook, D. and Swayne, D. F. (1996) Interactive High-Dimensional Data Analysis, *Journal of Computational and Graphical Statistics*, 5(1), 78-99.
- [5] Cheng, R. and Milligan, G. W.(1996) Measuring the influence of individual data points in a cluster analysis, *Journal of Classification*, 13, 315-335.
- [6] Gordon, A.D.(1981) *Classification*, Chapman and Hall, London.
- [7] Friedman, J. H. and Tukey, J. W. (1974) A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions on Computing C*, 23, 881-889.
- [8] Gower, J. C. and Ross, G. J. S.(1969) Minimum spanning trees and single linkage cluster analysis, *Applied Statistics*, 18, 54-64.
- [9] Hartigan, J. A.(1975) *Clustering Algorithms*, John Wiley & Sons, INC.
- [10] Johnson, R. A. and Wichern, D. W.(1992) *Applied Multivariate Statistical Analysis*, 3-ed., Prentice-Hall, Inc.
- [11] Jolliffe, I. T., Jones, B. and Morgan, B. J. T.(1995) Identifying influential observations in hierarchical cluster analysis, *Journal of Applied Statistics*, Vol 22, No. 1, pp.61-80.
- [12] Kruskal, J. B. and Wish, M. (1978) *Multidimensional Scaling*, Sage Pub.
- [13] Krzanowski, W. J.(1988) *Principles of Multivariate Analysis*, Oxford Science Pub.
- [14] Kwon, S. and Cook, D. (1998) Using a Grand Tour and Minimal Spanning Tree to Detect Structure in High-Dimensions, *Computing Science and Statistics*, 30, forthcoming.

- [15] Littman, M. L., Swayne, D. F., and Buja, A. (1992) Visualizing the Embedding of Objects in Euclidean Space, *Computing Science and Statistics*, 24, 208-217.
- [16] Lubischew, A. A. (1962) On the Use of Discriminant Functions in Taxonomy *Biometrics*, 18, 455-477.
- [17] Morgan, B. J. T. and Shaw, D.(1982) Illustrating data on English dialects, *Lore and Language*, 3, 14-29.
- [18] Ross, G. J. S.(1969) Algorithm AS 13, Minimum Spanning Tree, *Applied Statistics*, 18, 103-104.
- [19] Ross, G. J. S.(1969) Algorithm AS 15, Single Linkage Cluster Analysis, *Applied Statistics*, 18, 106-110.
- [20] Seber,G.A.F. (1984) *Multivariate Observations*, Wiley, New York.
- [21] Zahn, C. T.(1971) Graph-Theoretical Methods for detecting and describing gestalt clusters, *IEEE-Transactions on Computers*, Vol. C-20, No.1, 68-86.
- [22] Zupan, J.(1982) *Clustering of Large Data Sets*, Research Studies Press.