

## Hierarchical Agglomerative Clustering

“Clustering” is an analysis procedure that takes a set of data (or observations) and groups them into similar clusters in some sense. Many types of clustering algorithms have been established in the machine learning field. Here, we focus on one specific algorithm called “hierarchical agglomerative clustering”.

At a high level, this procedure iteratively combines clusters together by first treating each data point as its own cluster and then merging the most similar clusters together. This procedure is “hierarchical” because it merges clusters iteratively. It is “agglomerative” because it takes data points individually and then builds larger and larger clusters in a bottom-up fashion. The result of the procedure gives us information about the clusters formed at each iteration as well as the final cluster that each data point belongs to. Considering only the final cluster results, we can represent these results as a vector of cluster identifiers of size 1 by M. That is, for each community specified in the input data, a cluster ID is returned to indicate which cluster a particular community belongs to.

### Choosing K

The desired number of clusters, K, must be specified so that the procedure terminates when K clusters have been formed. If K is not specified, the procedure terminates when all the data combines into a single cluster (i.e.,  $K=1$ ). In general, determining the optimal value for K is an open research problem, although when extensive data is available, certain evaluation methods (such as cross validation) can be used to identify reasonable values for K.

### Defining Similarity

Linkage is the method used to define similarity between two clusters. “Average linkage” is defined using the average of the distances between all pairs of data in the two clusters under consideration. The pair of clusters with the smallest average distance is merged. There are other linkage methods available for hierarchical agglomerative clustering. The examples illustrated in this document focuses on average linkage only.

### Input Data with Examples

The input data needed in a clustering procedure is referred to as data set D, where D is a matrix of size M by N, with M being the number of communities (or alternatively, organizations or individuals) and N being the number of characteristics represented as variables. Table 1 shows an example with  $M=5$  and  $N=8$ , and Table 2 shows an example of  $M=4$  and  $N=2$ .

	char1	char2	char3	char4	char5	char6	char7	char8
comm1	.9283	.5915	.2723	.8348	.3948	.0655	.5152	.0367
comm2	.6835	.6233	.5369	.1346	.3481	.5227	.6032	.1764
comm3	.3331	.7339	.7059	.7326	.5677	.2407	.5901	.5361
comm4	.4156	.7436	.9666	.4758	.5470	.3979	.5258	.2244
comm5	.3672	.5973	.7281	.5518	.0360	.0900	.8015	.6278

Table 1: Example of data for 5 communities (comm) with 8 characteristics (char).

	char1	char2
comm1	.8433	.3044
comm2	.7265	.2102
comm3	.9437	.0780
comm4	.8316	.3902

Table 2: Example of data for 4 communities (comm) with 2 characteristics (char).

When  $N=2$ , we can view the data visually by representing the characteristics as x- and y-axes and plotting the communities as data points. To illustrate, we generated two sets of random data. The first data set has  $M=500$  and  $N=2$ , and the second has  $M=5000$  and  $N=2$ . Here, we assume the characteristics are continuous variables, so their values range from 0.0 to 1.0. Figure 1 shows these plots.

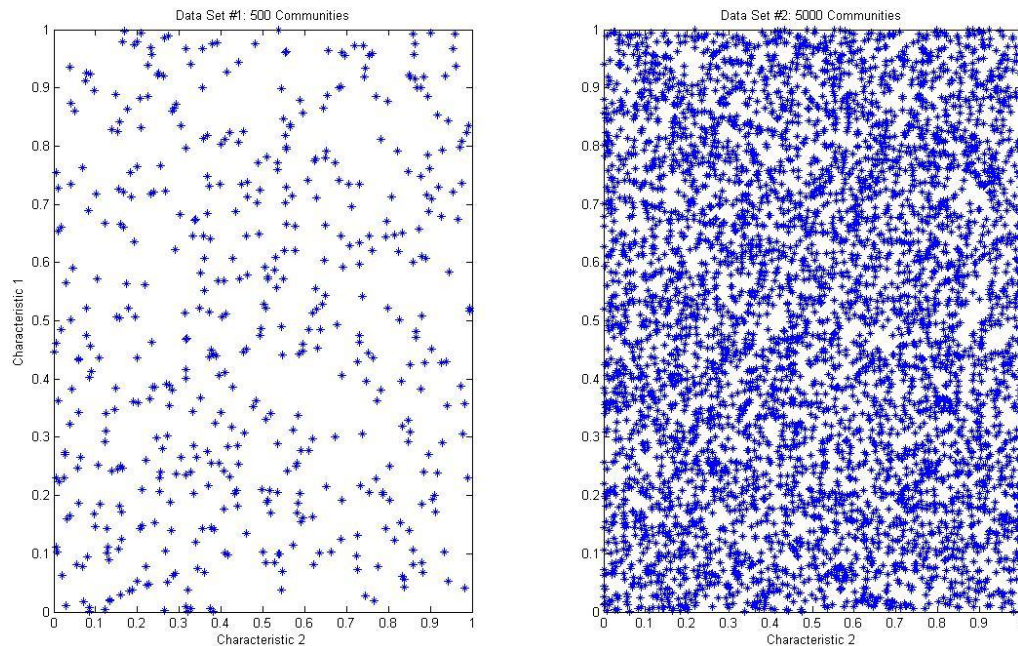
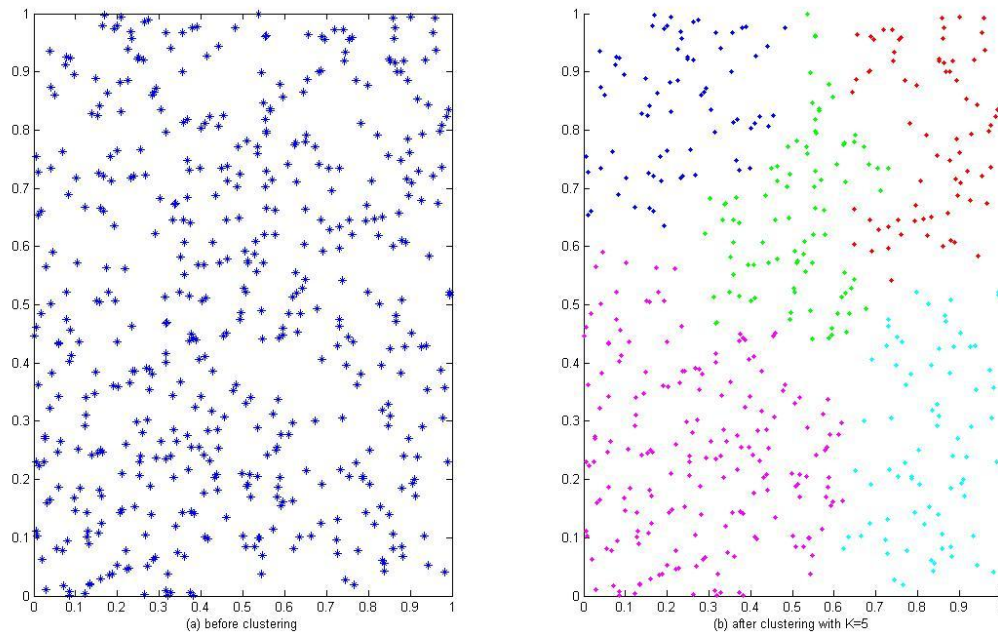


Figure 1: Examples of randomly generated data sets. Left:  $M=500$ . Right:  $M=5000$ .

In general,  $N$  can be arbitrarily larger than 2. However, when  $N>2$ , we cannot visualize the data easily, so the illustrations in this document consider the case of  $N=2$  only.

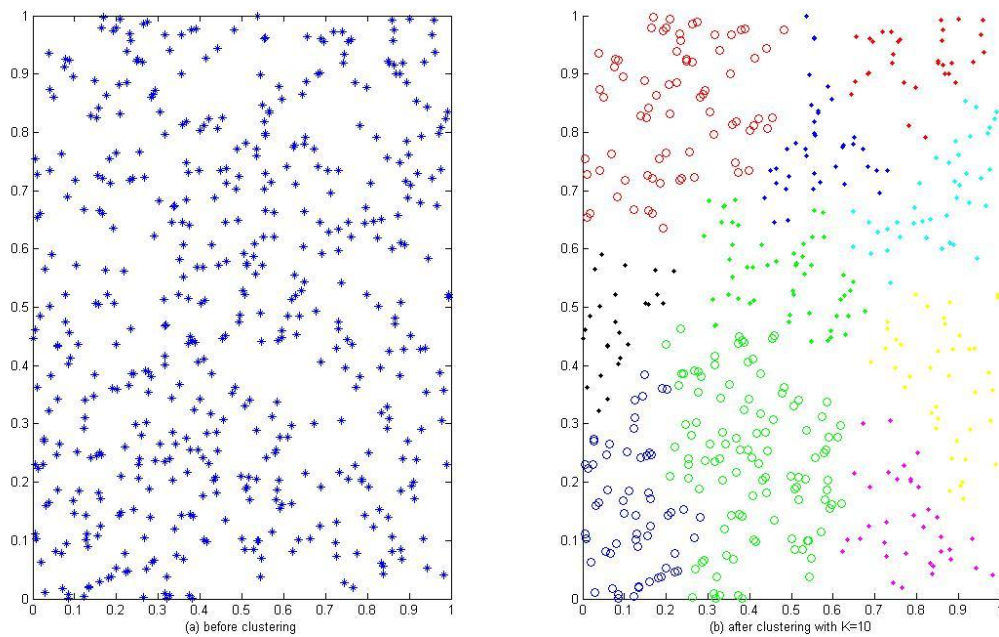
## Results with Examples

Using the cluster IDs, we illustrate the cluster analysis results for  $N=2$ . Figure 3 shows the first data set being clustered into  $K=5$  clusters using the method described above. Each cluster is represented by a different colour.

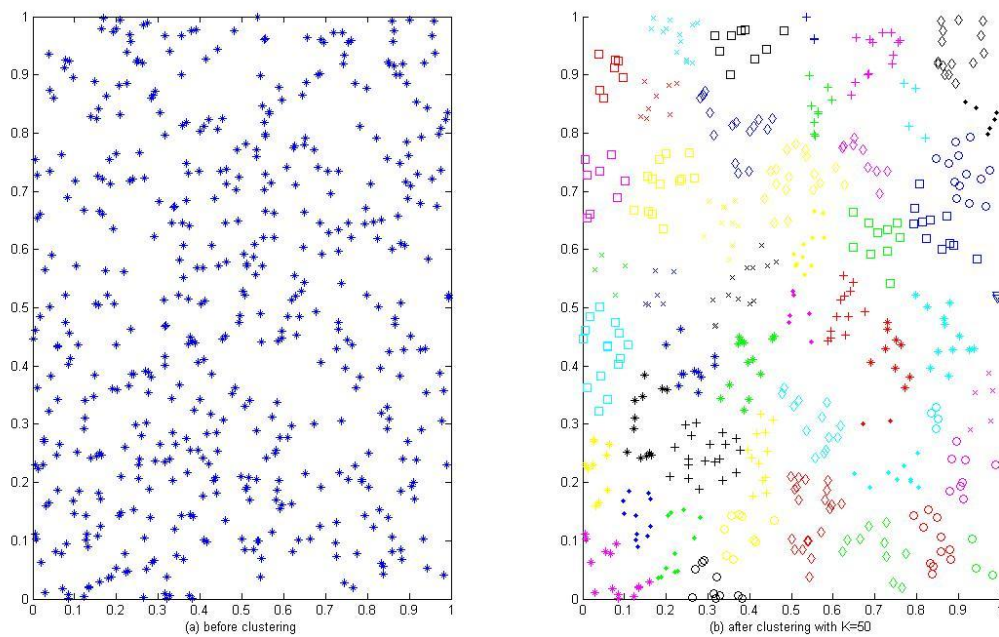


**Figure 3: Data with  $M=500$ ,  $N=2$ . Left: Original data. Right: Clustered data.**

For demonstrative purposes, we apply the same clustering procedure to the first data set for  $K=10$  and  $K=50$ . The corresponding results are shown in Figure 4 and Figure 5. Each cluster is represented by a different colour and/or shape.

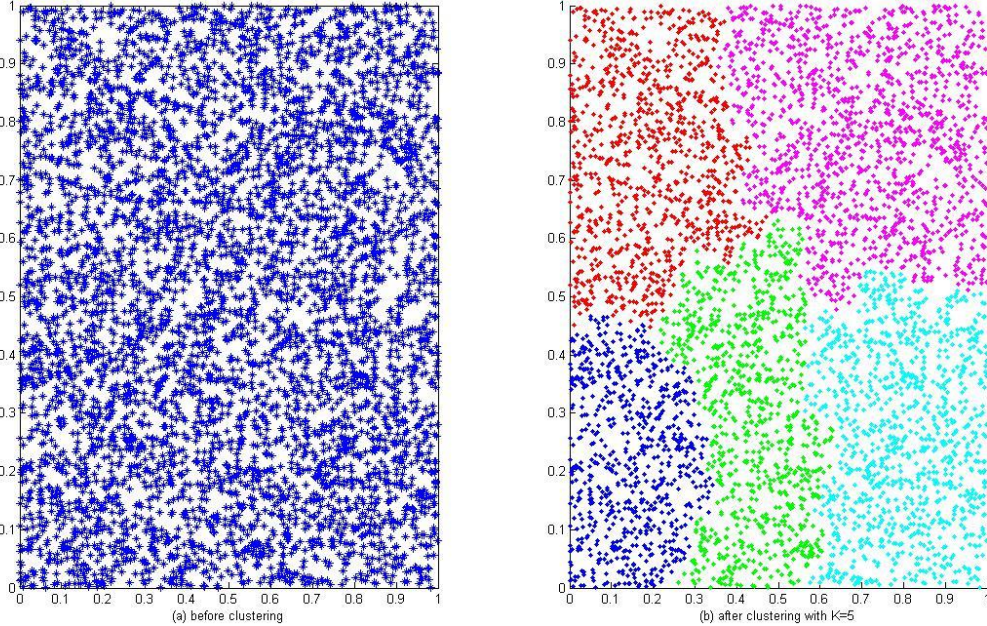


**Figure 4: Data with  $M=500$ ,  $N=2$ . Left: Original data. Right: Clustered data.**

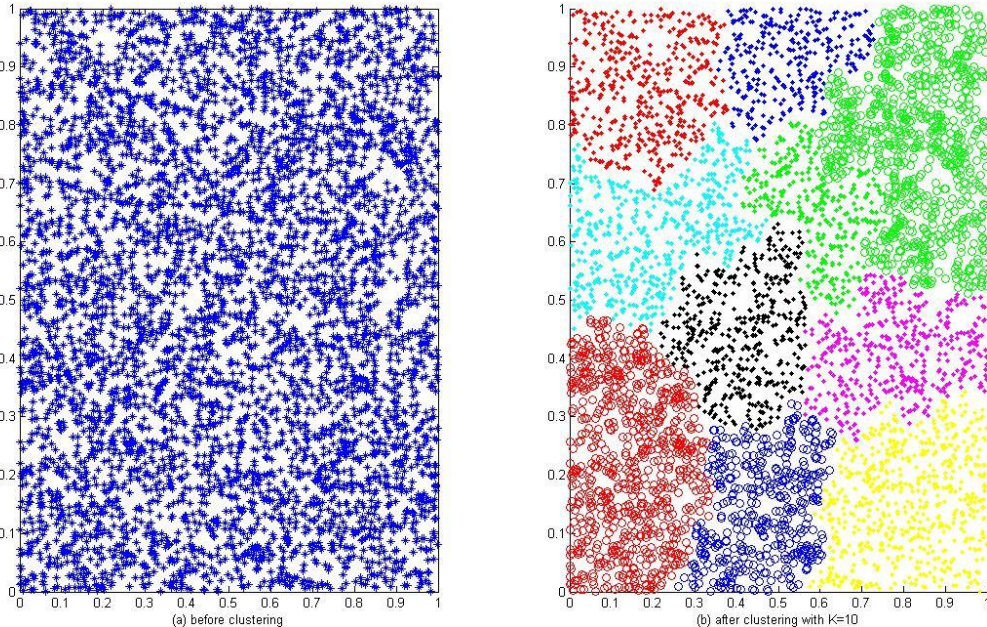


**Figure 5: Data with  $M=500$ ,  $N=2$ . Left: Original data. Right: Clustered data.**

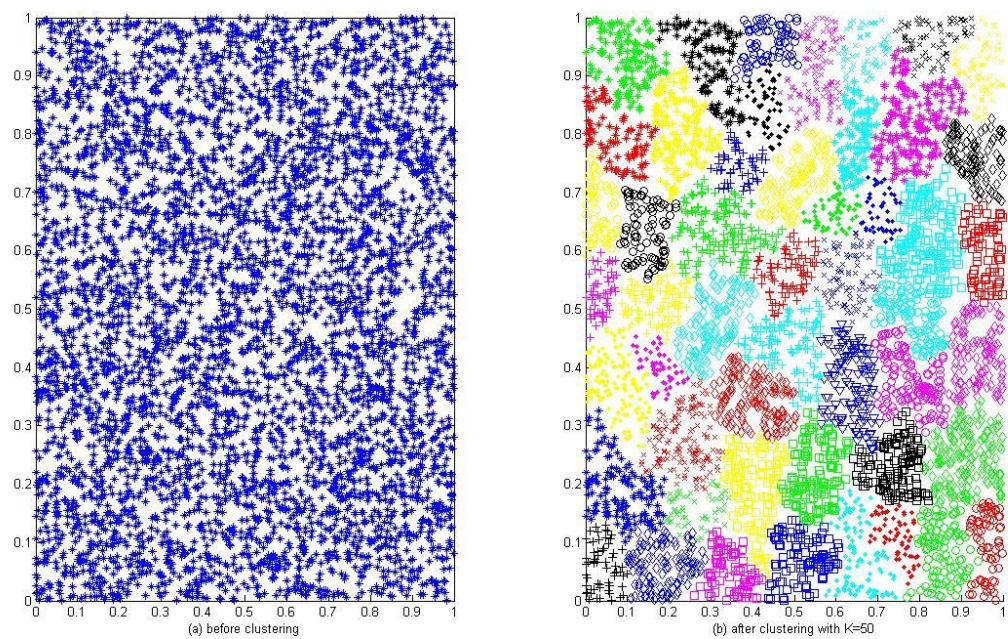
Using the second data set with  $M=5000$  and  $N=2$ , we also apply the same clustering procedure to obtain the corresponding plots. The results for  $K=5$  are shown in Figure 6, those for  $K=10$  are shown in Figure 7, and those for  $K=50$  are shown in Figure 8.



**Figure 6: Data with  $M=5000$ ,  $N=2$ . Left: Original data. Right: Clustered data.**



**Figure 7: Data with  $M=5000$ ,  $N=2$ . Left: Original data. Right: Clustered data.**



**Figure 8: Data with  $M=5000$ ,  $N=2$ . Left: Original data. Right: Clustered data.**