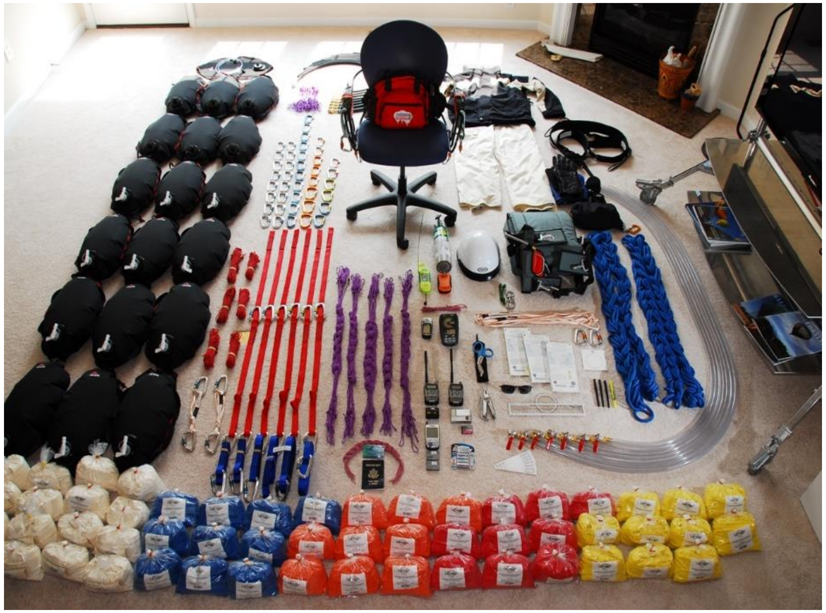
Clustering Approaches & Techniques

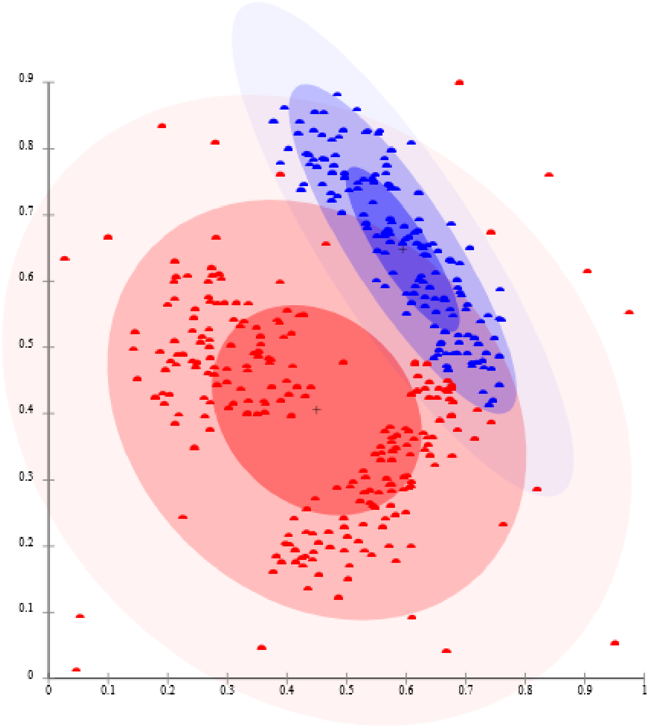
Presented by: Derek Kane

Overview of Topics

- Introduction to Clustering Techniques
  - K-Means
  - Hierarchical Clustering
  - Gaussian Mixed Model
- Visualization of Distance Matrix
- Practical Example



Introduction to Cluster Analysis



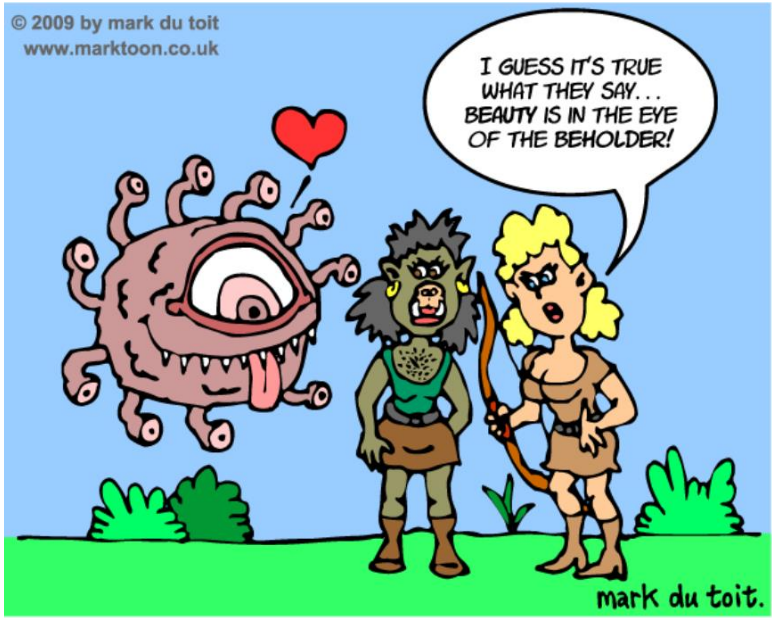
- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).
- It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.
- Cluster analysis itself is not one specific algorithm, but the general task to be solved.

Introduction to Cluster Analysis

There are many real world applications of clustering:

- Grouping or Hierarchies of Products
- Recommendation Engines
- Biological Classification
- Typologies
- Crime Analysis
- Medical Imaging
- Market Research
- Social Network Analysis
- Markov Chain Monte Carlo Methods

Introduction to Cluster Analysis


© 2009 by mark du toit
www.marktoon.co.uk
I GUESS IT'S TRUE WHAT THEY SAY... BEAUTY IS IN THE EYE OF THE BEHOLDER!
mark du toit.

- According to Vladimir Estivill-Castro, the notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms.

- There is no objectively "correct" clustering algorithm, but as it was noted, "clustering is in the eye of the beholder."

There are many different types of clustering models including:

- Connectivity models
- Centroid models
- Distribution models
- Density models
- Group models
- Graph-based models

Introduction to Cluster Analysis

- A "clustering" is essentially a set of such clusters, usually containing all objects in the data set.
- Additionally, it may specify the relationship of the clusters to each other, for example a hierarchy of clusters embedded in each other.

Clusterings can be roughly distinguished as:

- hard clustering: each object belongs to a cluster or not.
- soft clustering (also: fuzzy clustering): each object belongs to each cluster to a certain degree (e.g. a likelihood of belonging to the cluster).

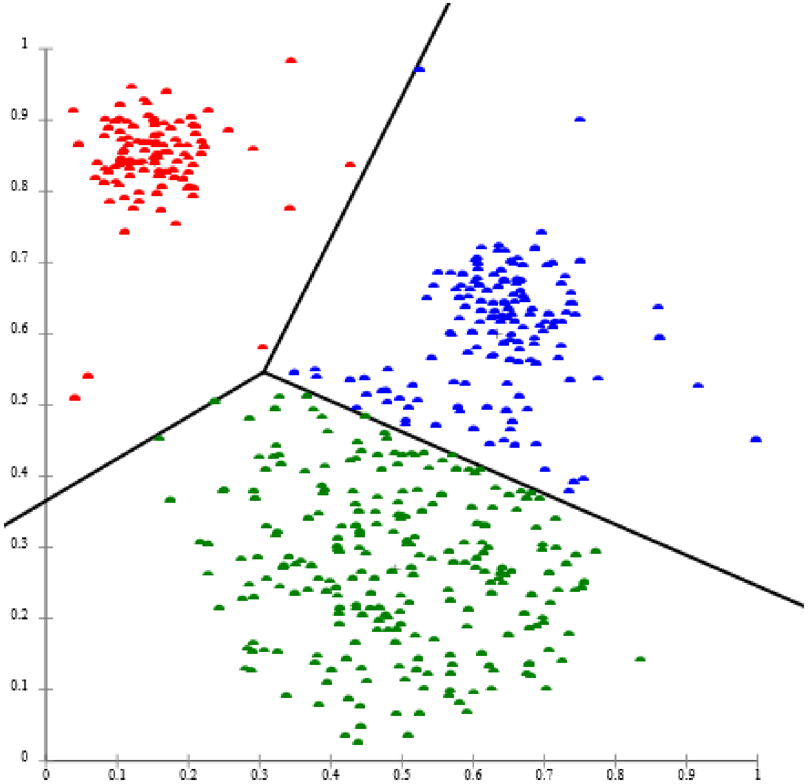
K-means

- K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem.

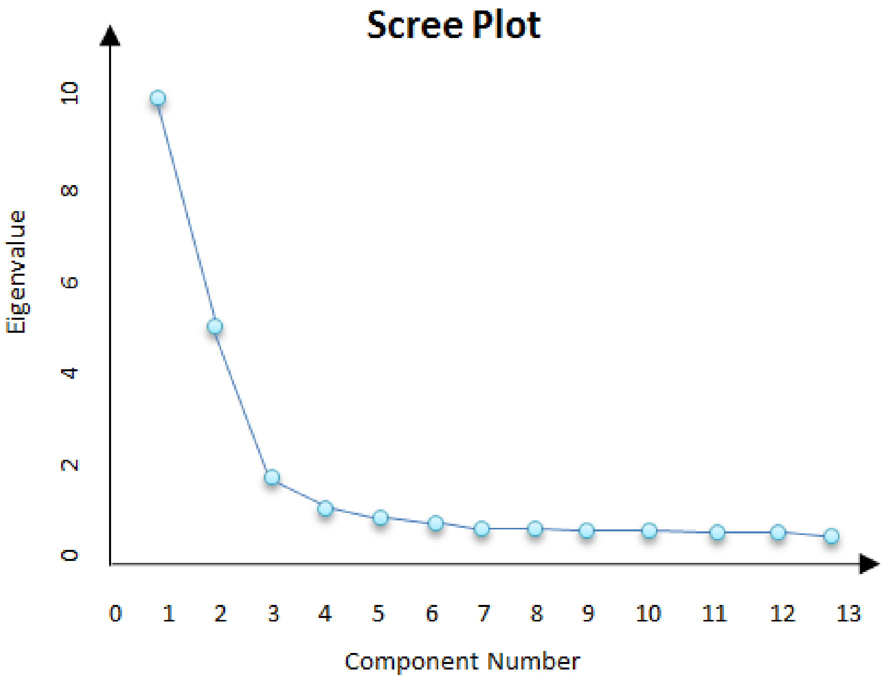
The algorithm is composed of the following steps:

  - Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
  - Assign each object to the group that has the closest centroid.
  - When all objects have been assigned, recalculate the positions of the K centroids.
  - Repeat Steps 2 and 3 until the centroids no longer move.

- The k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum.
- The algorithm is also significantly sensitive to the initial randomly selected cluster centers.



Scree Plot



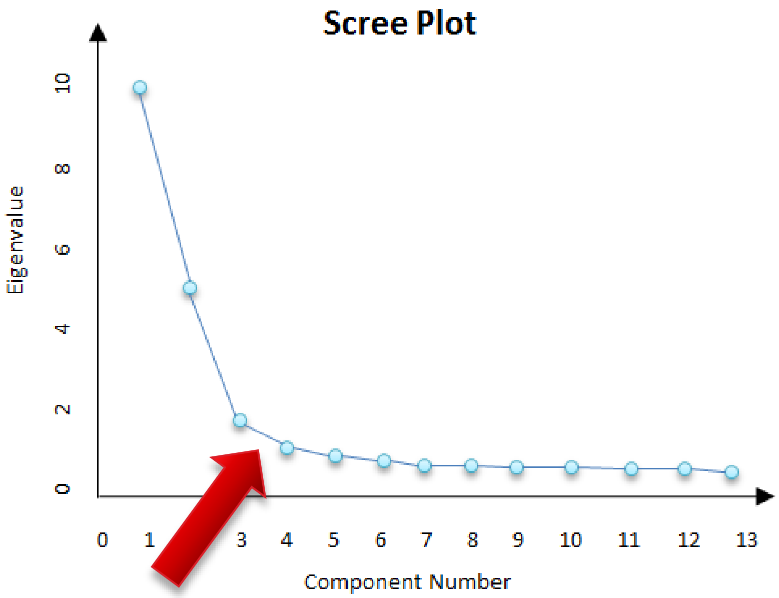
- A scree plot is a graphical display of the variance of each (cluster) component in the dataset which is used to determine how many components should be retained in order to explain a high percentage of the variation in the data.

- The variance of each component is calculated using the following formula: λi∑i=1nλi where λi is the ith eigenvalue and ∑i=1nλi is the sum of all of the eigenvalues.

- The plot shows the variance for the first component and then for the subsequent components, it shows the additional variance that each component is adding.

Scree Plot

- A scree plot is sometimes referred to as an “elbow” plot.

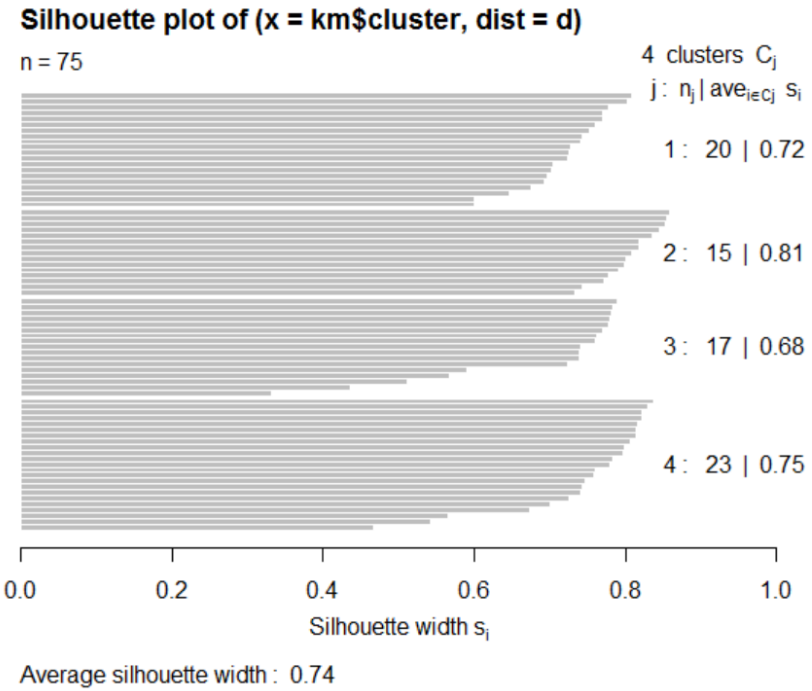


# of Clusters should be 3 or 4

- In order to identify the optimal number of clusters for further analysis, we need to look for the “bend” in the graph at the elbow.

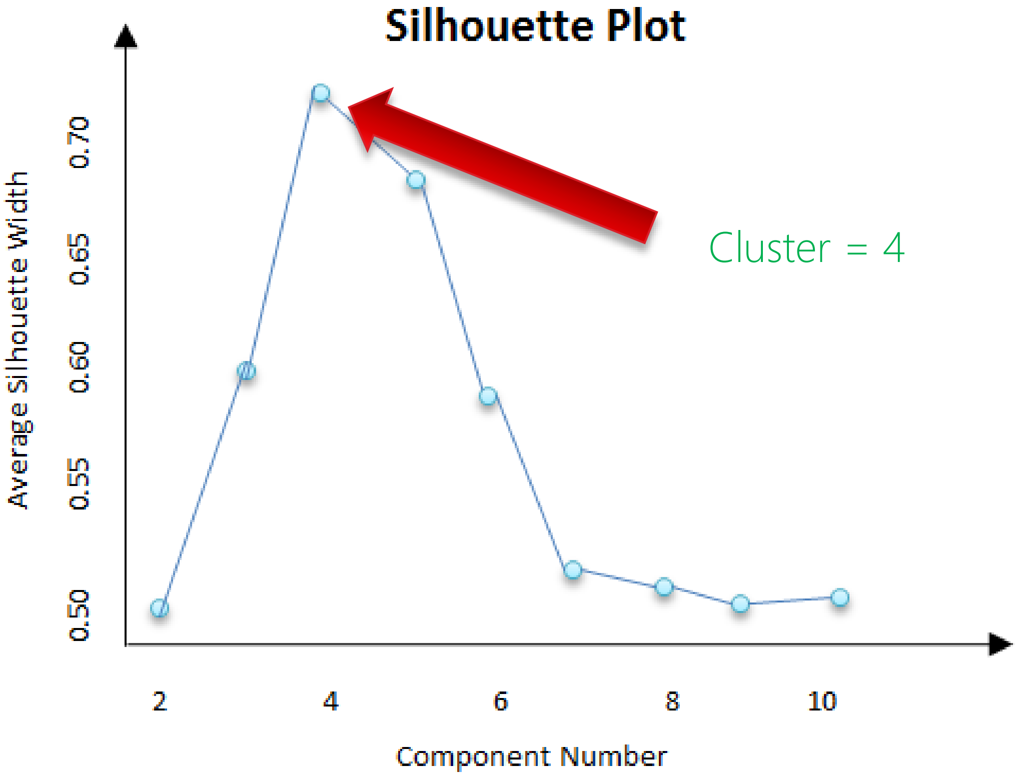
Silhouette Plot

- Silhouette refers to a method of interpretation and validation of clusters of data.
- The silhouette technique provides a succinct graphical representation of how well each object lies within its cluster and was first described by Peter J. Rousseeuw in 1986.



Silhouette Plot

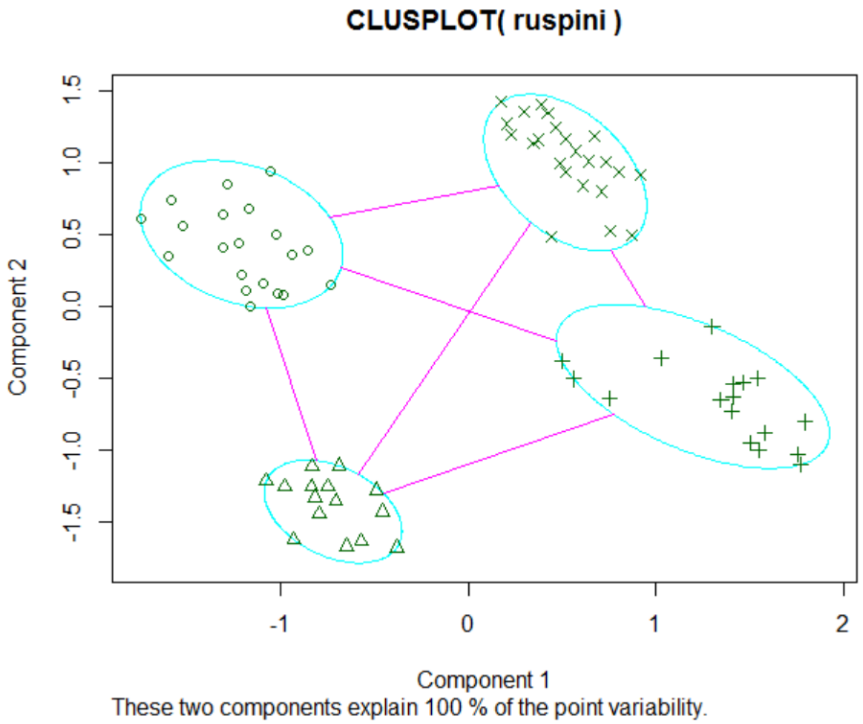
- The goal of a silhouette plot is to identify the point of the highest Average Silhouette Width (ASW). This is the optimal number of clusters for the analysis.

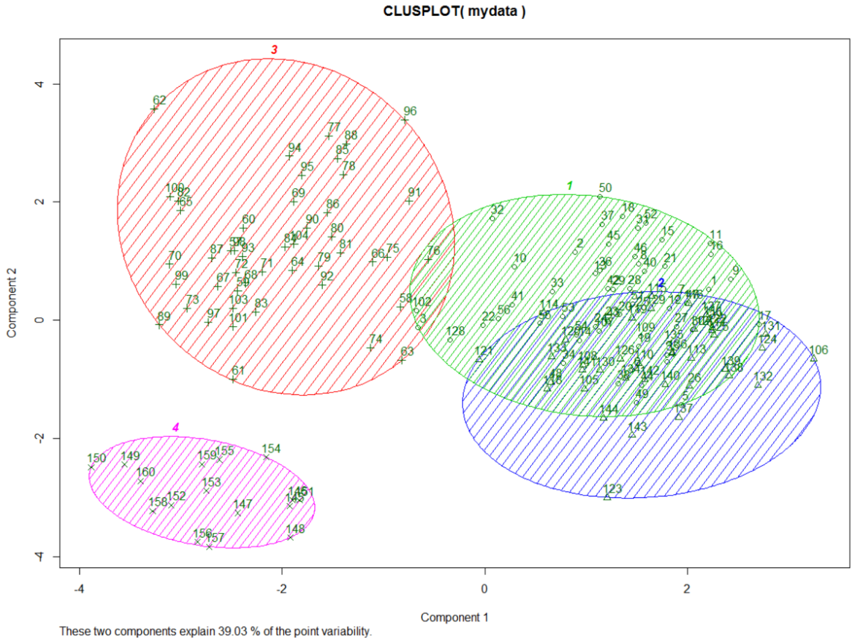


Cluster = 4

K-means

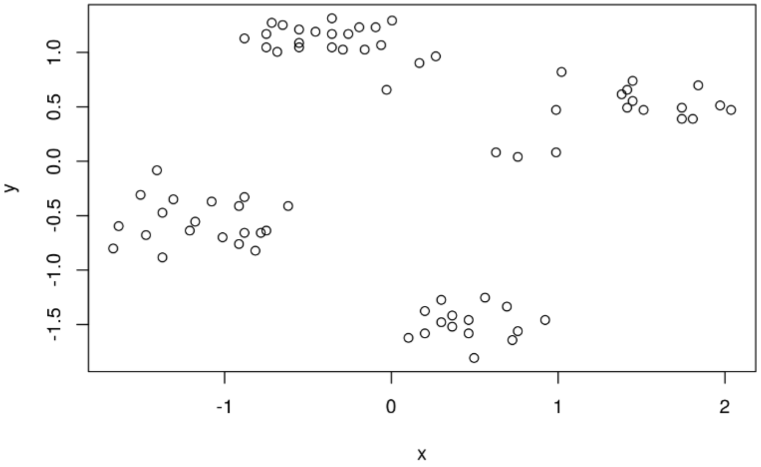
- The Cluster plot shows the spread of the data within each of the generated clusters.
- A Principal Component Analysis (greater than 2 dimensions) can be utilized to view the clustering.



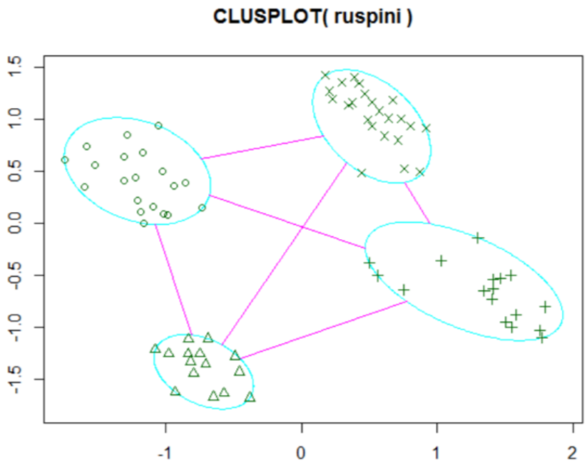


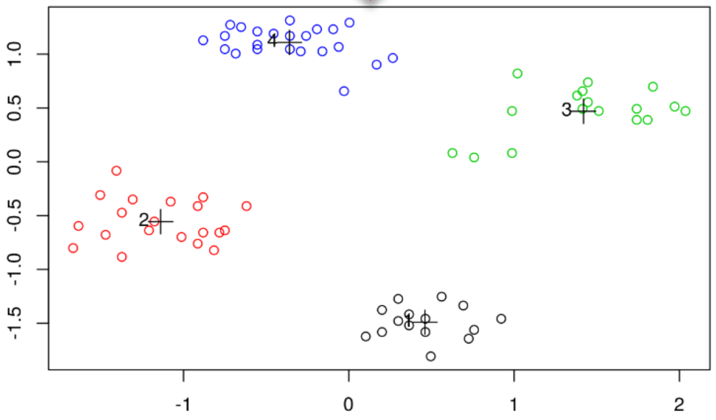
K-Means

- Here is an example of K-Means clustering based on the Ruspini data.



Cluster = 4





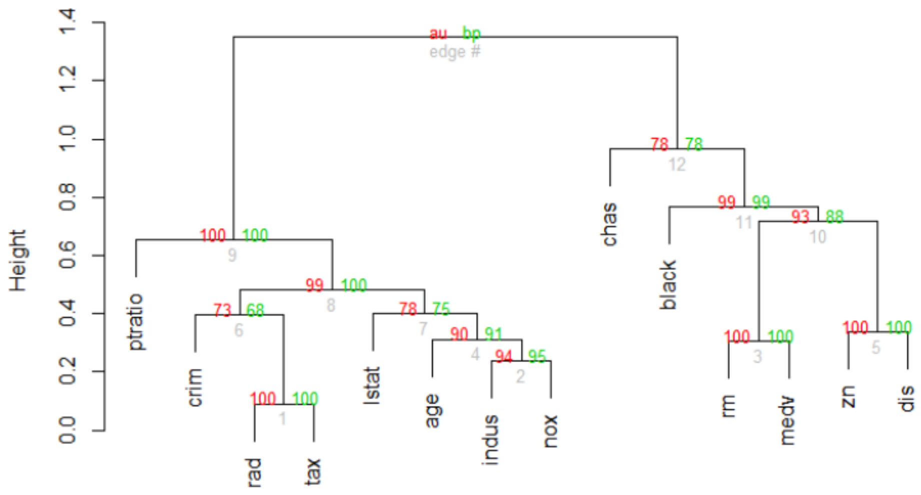
Hierarchical Clustering

- Hierarchical clustering (or hierarchic clustering) outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering.

- Hierarchical clustering does not require us to prespecify the number of clusters and most hierarchical algorithms that have been used in IR are deterministic.

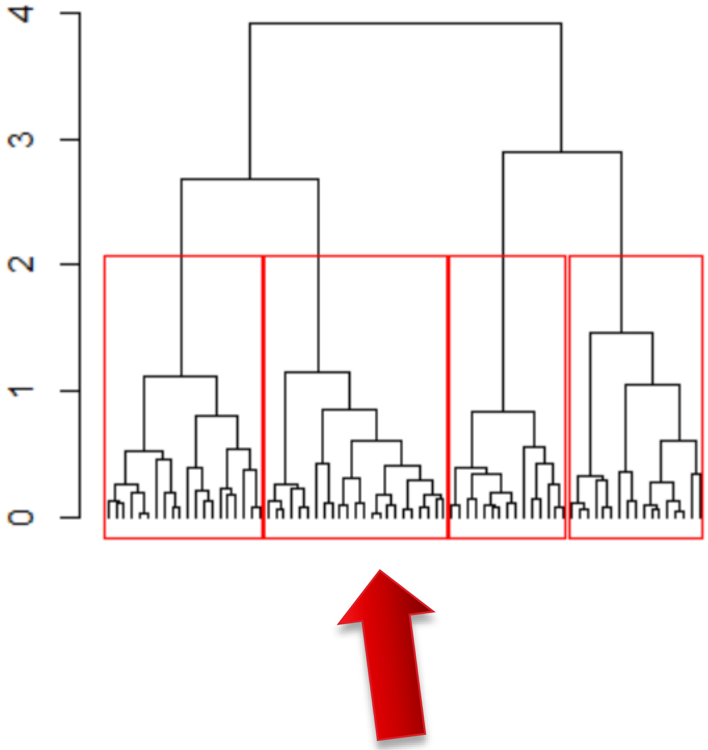
- These advantages of hierarchical clustering come at the cost of lower efficiency.

Cluster dendrogram with AU/BP values (%)



Distance: correlation
Cluster method: average

Hierarchical Clustering



Cluster = 4

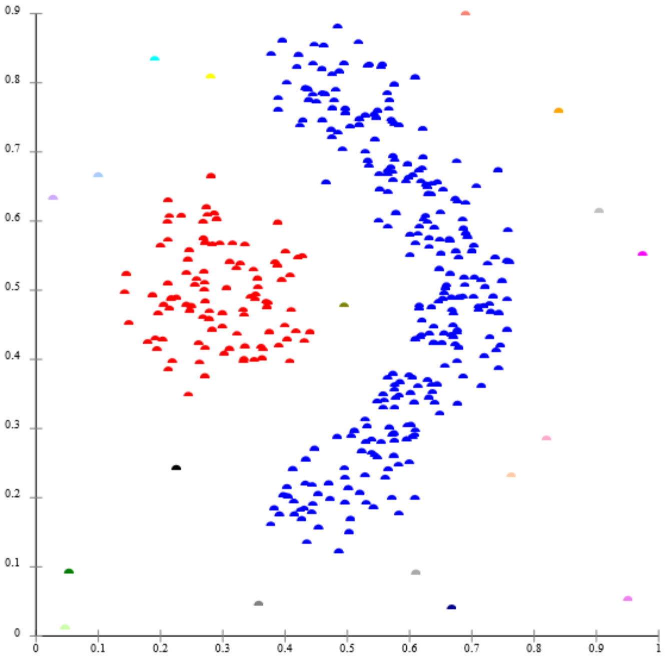
Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process of Johnson's (1967) hierarchical clustering is this:

- Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
- Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
- Compute distances (similarities) between the new cluster and each of the old clusters.
- Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Hierarchical Clustering

Step 3 can be done in different ways, which is what distinguishes single-linkage from complete-linkage and average-linkage clustering.

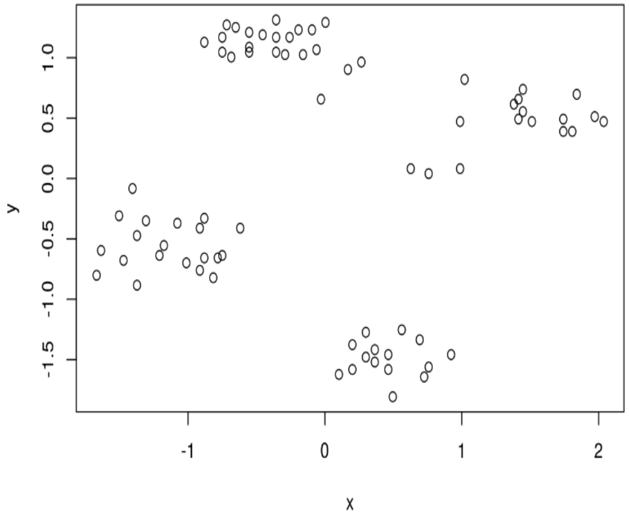
- In single-linkage clustering, we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.
- In complete-linkage clustering, we consider the distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster.
- In average-linkage clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.



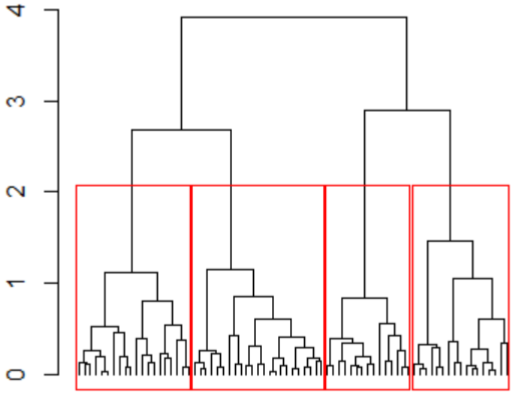
Single-linkage on density-based clusters.

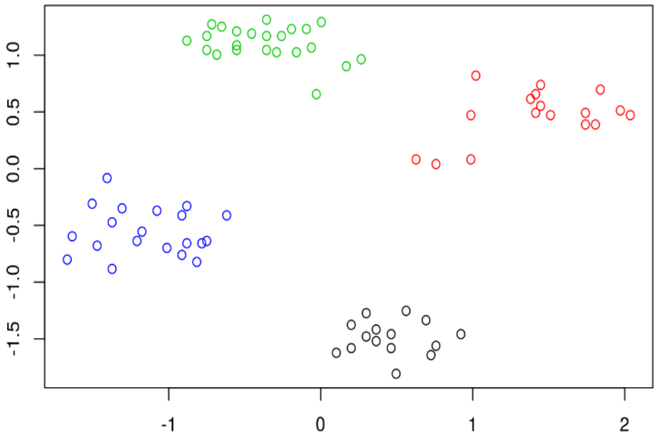
Hierarchical Clustering

- Here is an example of Hierarchical clustering based on the Ruspini data.



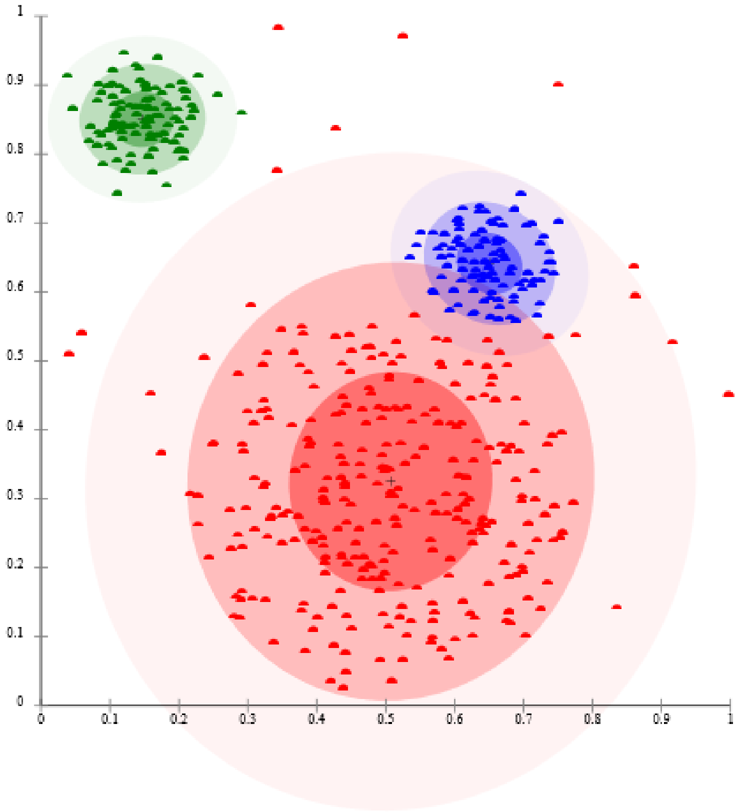
Cluster = 4





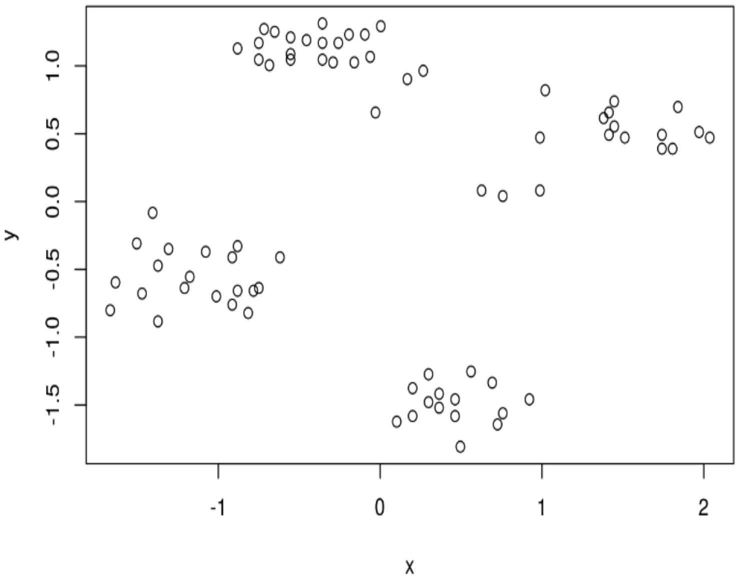
Gaussian Mixed Models

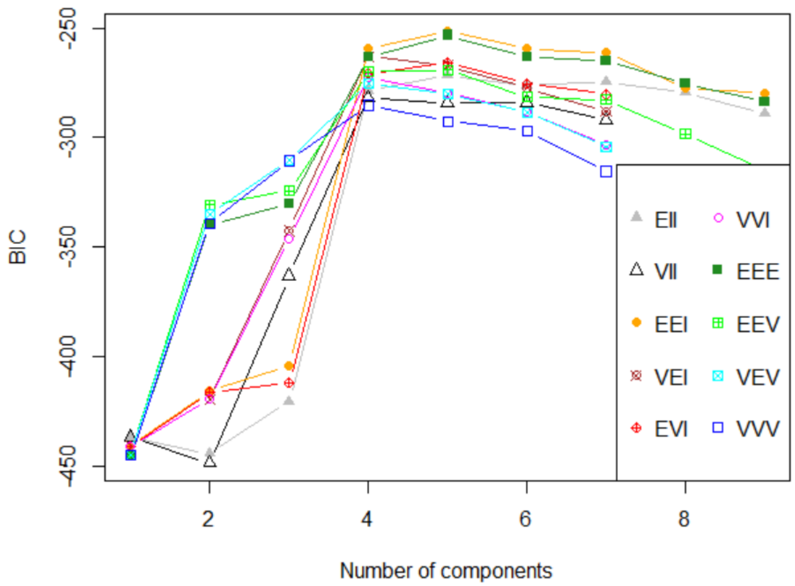
- Another technique we can use for clustering involves the application of the EM algorithm for Gaussian Mixed Models.
- The approach uses the Mclust package in R that utilizes the BIC statistic as the measurement criteria.
- The goal for identifying the number of clusters is to maximize the BIC.
- This approach is considered to be a fuzzy clustering method.



Gaussian Mixed Models

- Here are some diagnostics which are useful to establish the correct fit to the data.

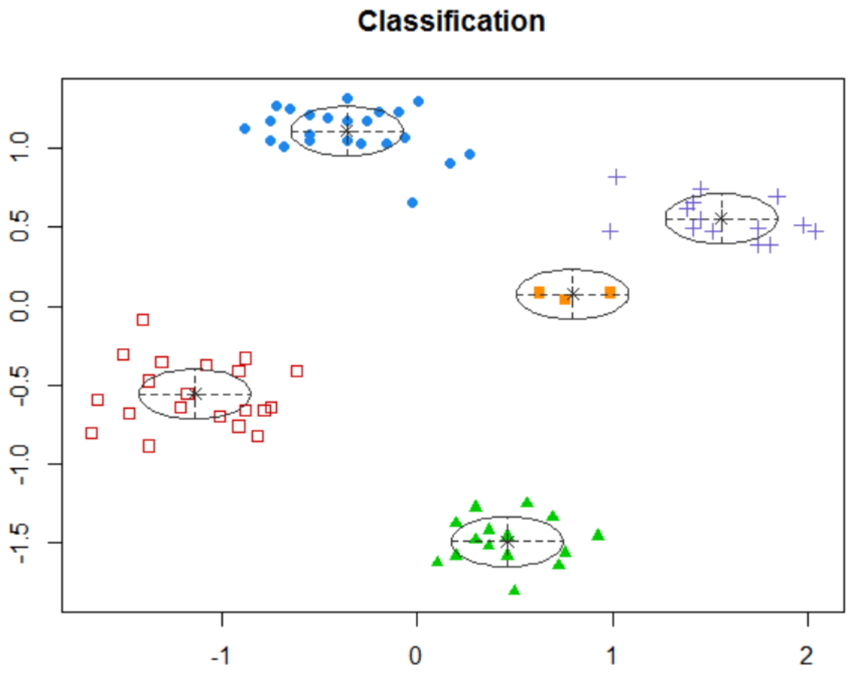


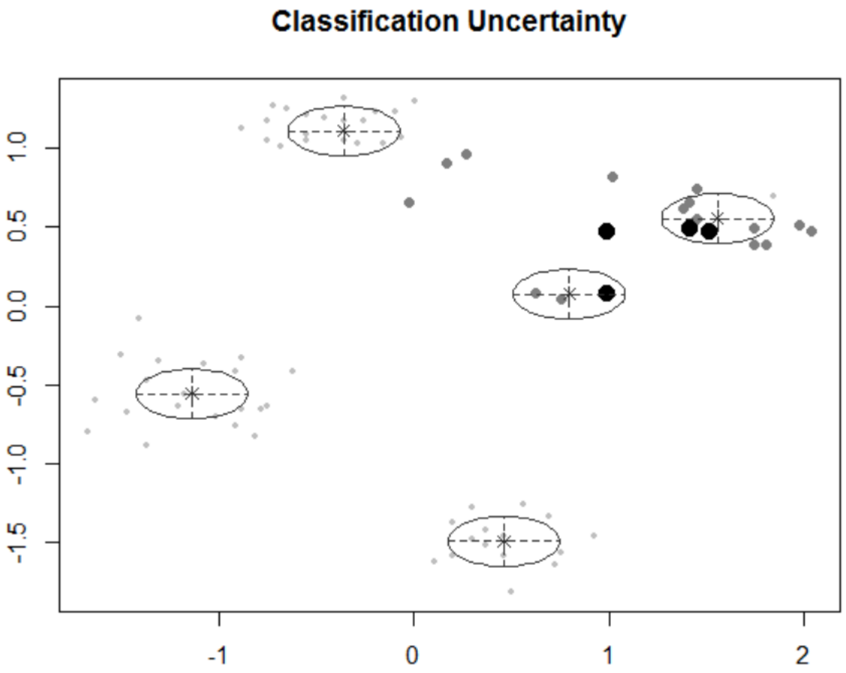


The maximum BIC indicates that a Cluster = 5 should be used

Gaussian Mixed Models

- The Mclust classification chart depicts the 5 centroid distribution and also the relative degree of the classification uncertainty.

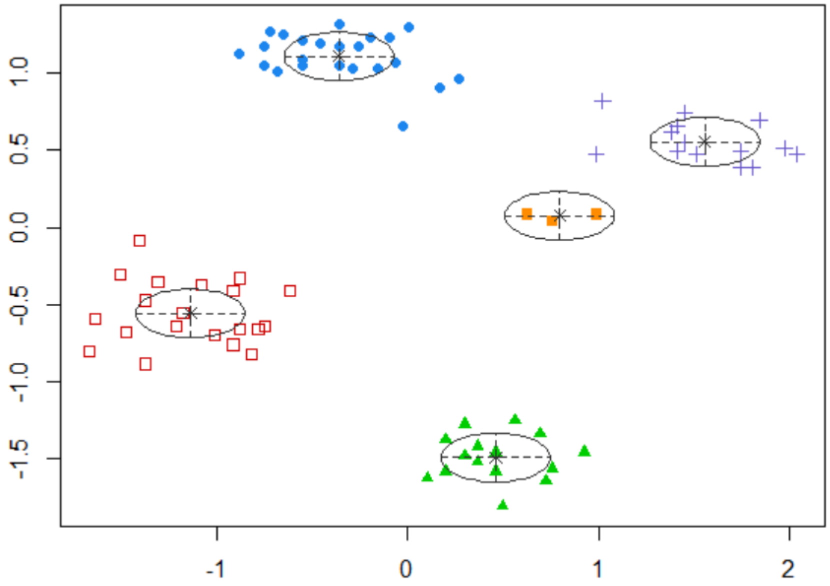

Classification


Classification Uncertainty

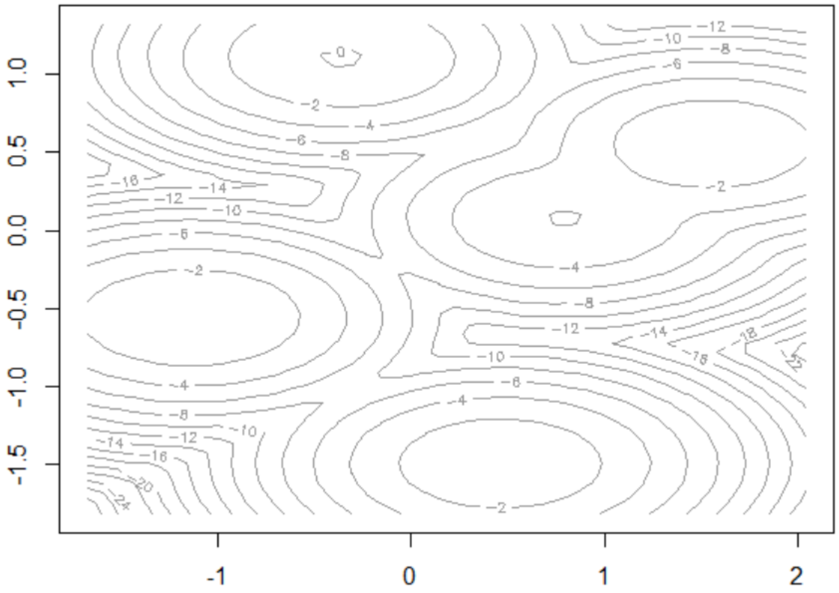
Gaussian Mixed Models

- The contour plot shows the density of the data points relative to each of the centroids.

Classification



log Density Contour Plot



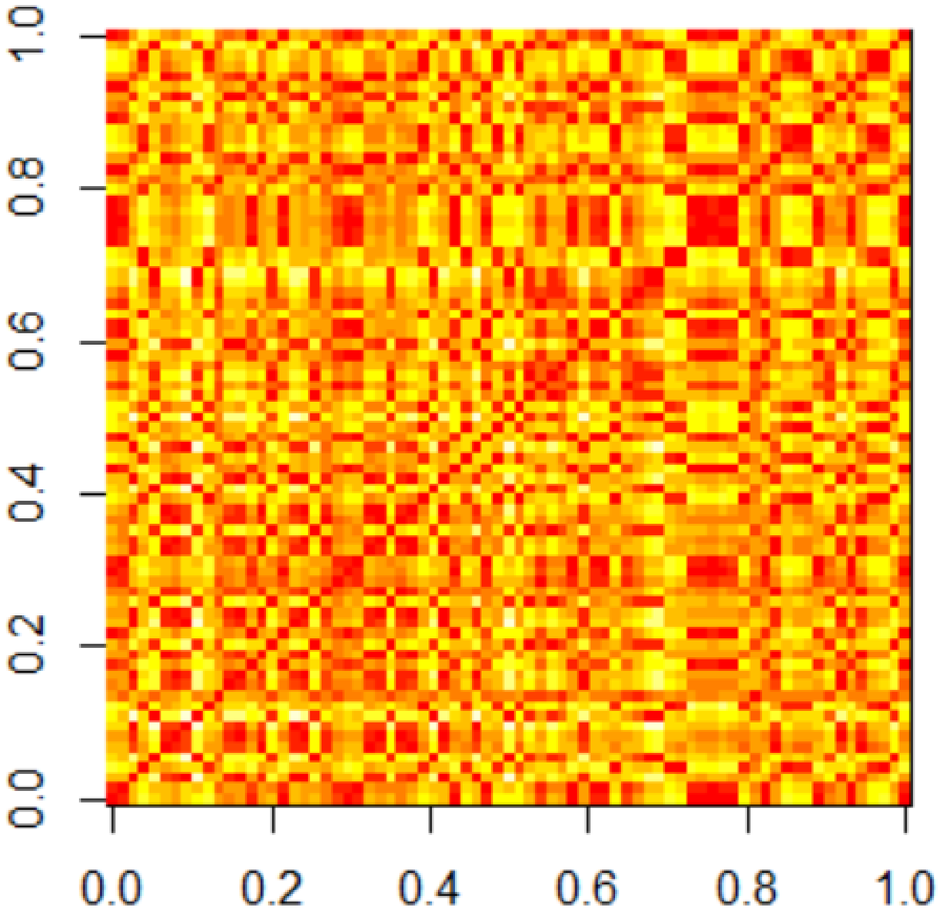
Distance & Similarity Matrices

- A similarity matrix is a matrix of scores that represent the similarity between a number of data points.

- Each element of the similarity matrix contains a measure of similarity between two of the data points.

- A distance matrix is a matrix (two-dimensional array) containing the distances, taken pairwise, of a set of points.

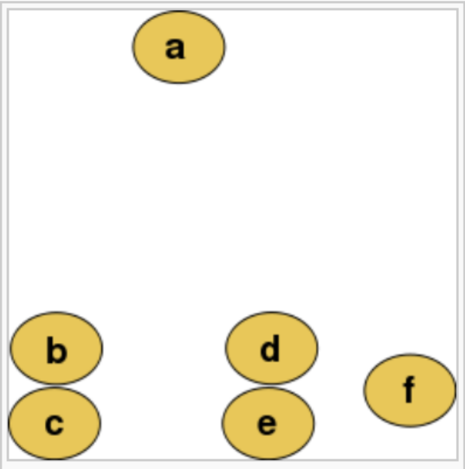
- This matrix will have a size of N×N where N is the number of points, nodes or vertices (often in a graph).



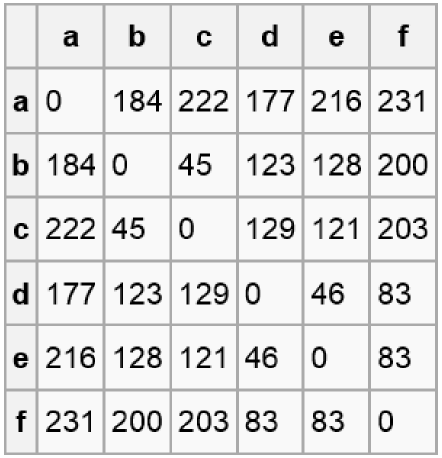
Heat map of Similarity Matrix

Distance Matrix Example

Ex. Points on a graph

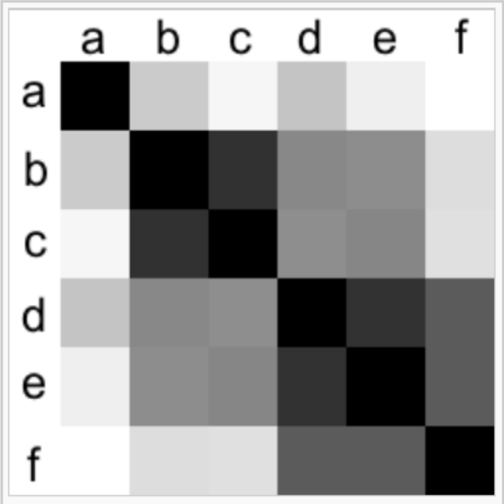


Distance Matrix



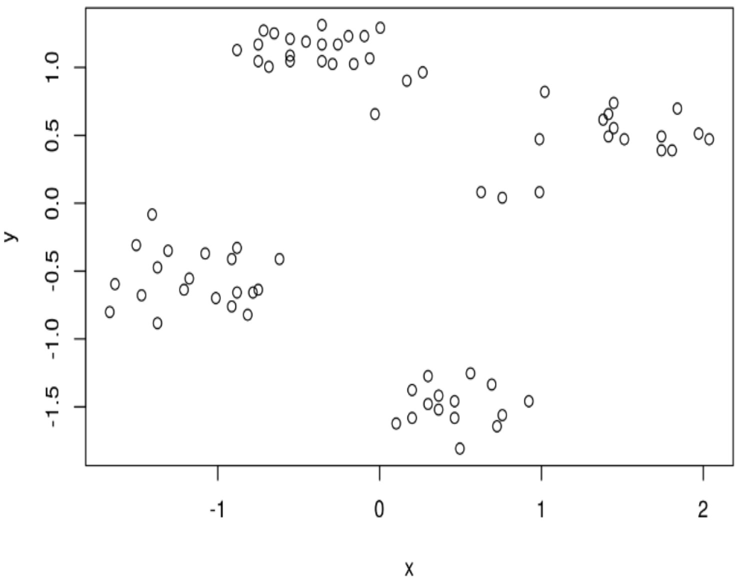
	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

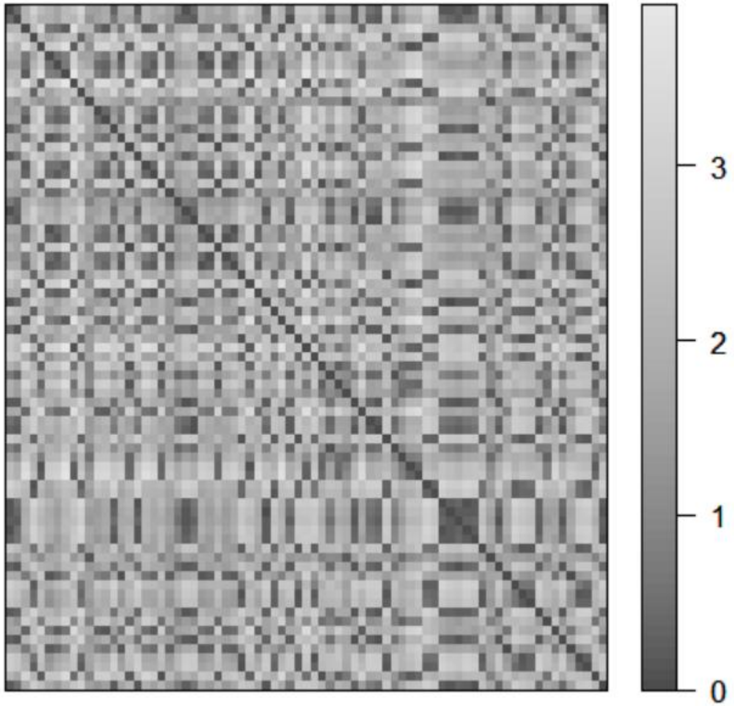
Heat map



Distance Matrix Example

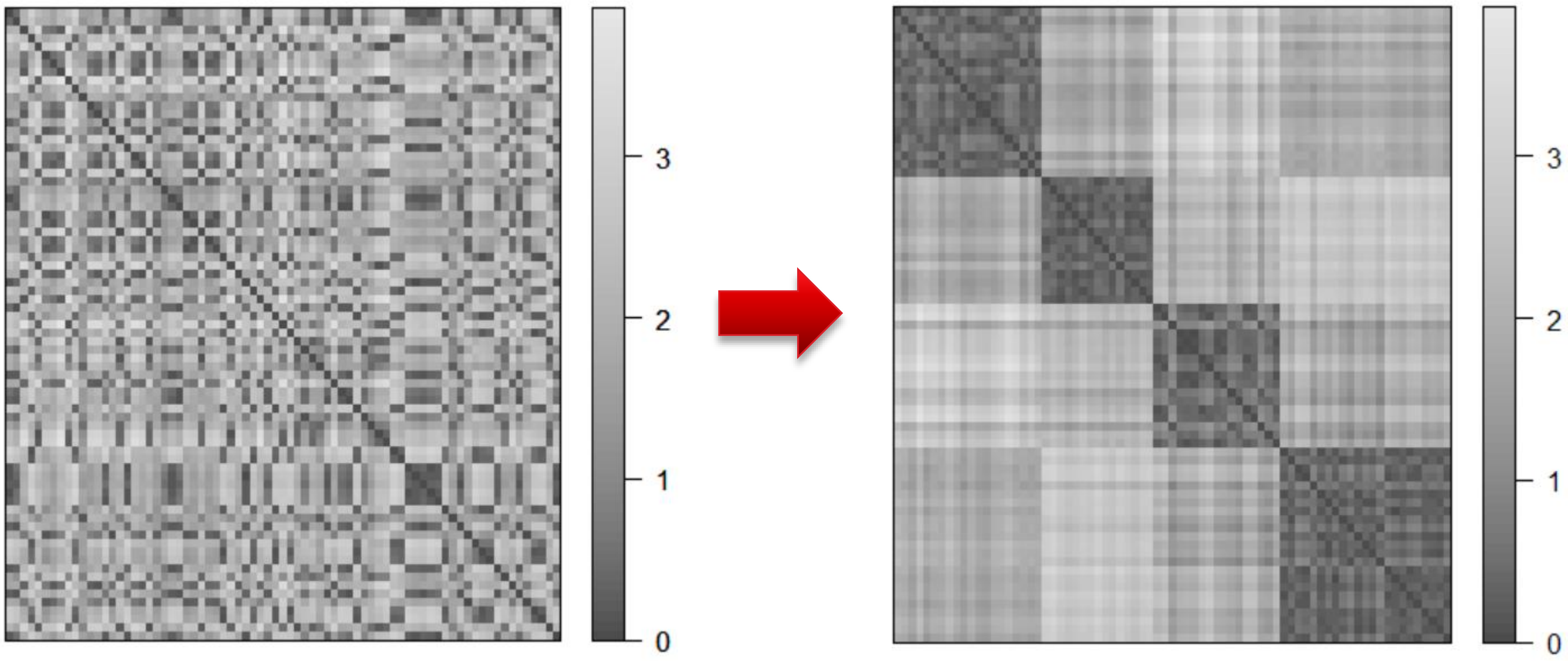
- Here is an application of the Distance Matrix to the Ruspini dataset.





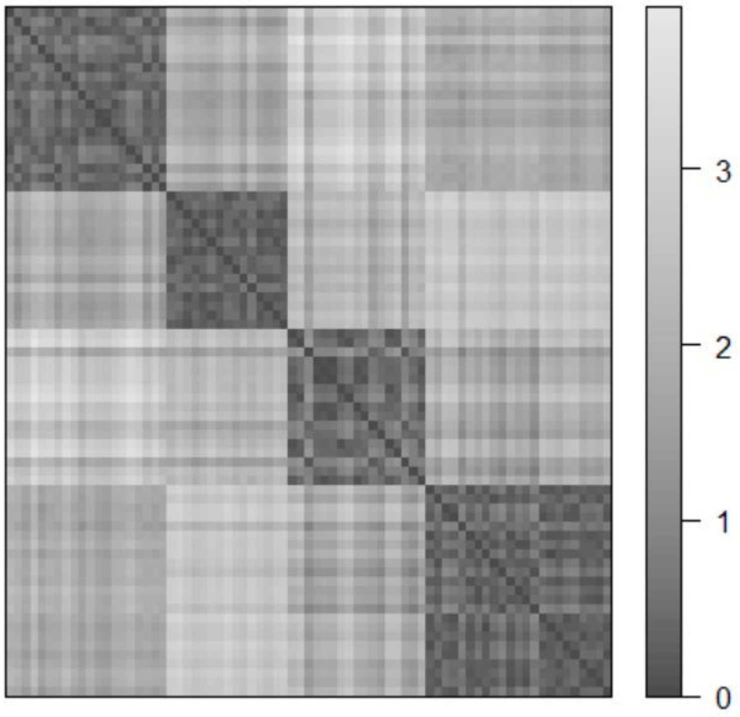
Distance Matrix Example

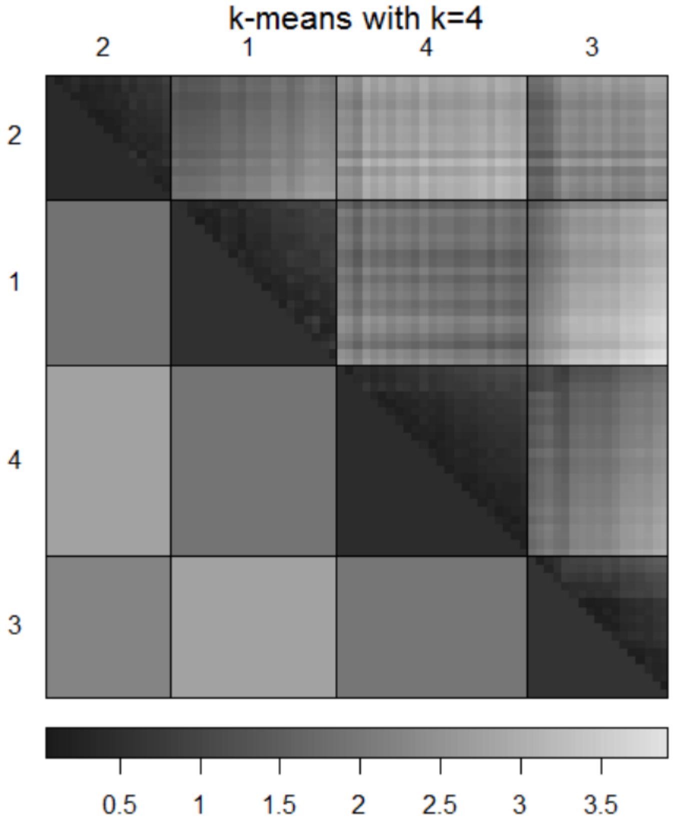
- Now we will reorder the distance matrix by the kmeans (k=4) algorithm.



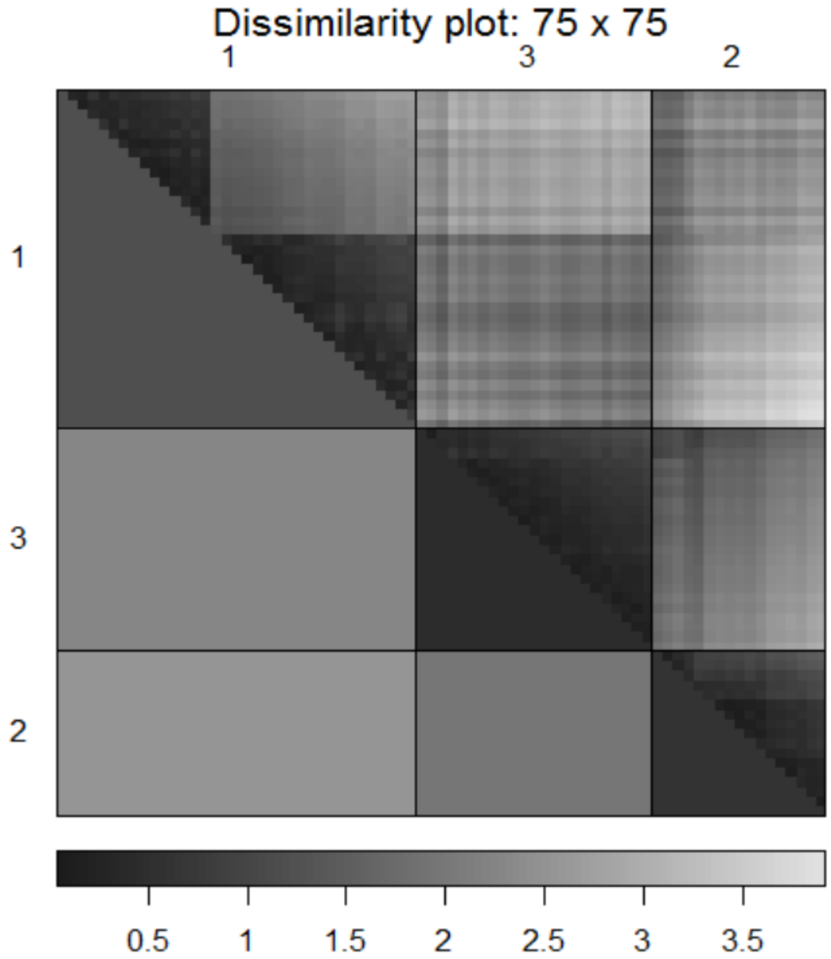
Distance Matrix Example

- Now we bring in the dissplot to compare the heatmap against expected values. We are looking for symmetry between the observed colors and the expected colors.




k-means with k=4

Distance Matrix Example



- This is an example of an incorrect kmeans specification of k = 3 instead of k = 4.
- The heatmap shows that within the 1st cluster there appears to be datapoints which should be separated into an additional distinct cluster.
- The similarity and distance matrices can be powerful visual aids when evaluating the fit of clusters.

Practical Example: Multiple Clustering Techniques

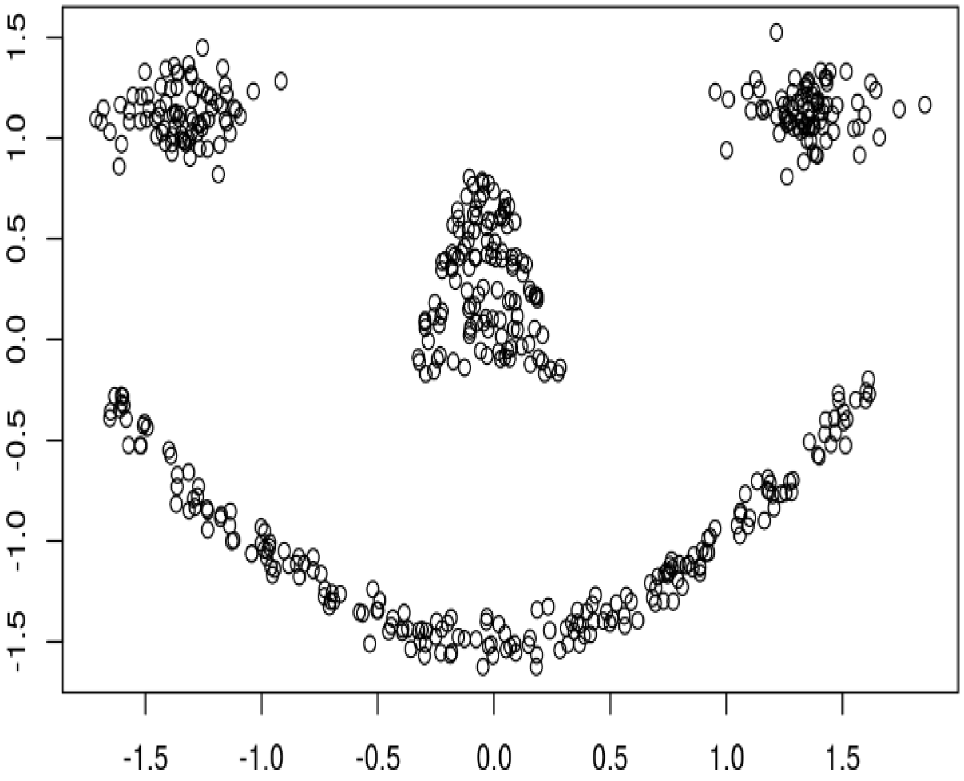
Understanding the Data

- The goal of a clustering exercise should be to identify the appropriate pockets or groups of data that should be grouped together.

- Because the structure of the underlying data is usually unknown, there could be many different clusters created by different clustering techniques.

- Therefore, lets use a dataset where it is obvious what the correct number of clusters should be.

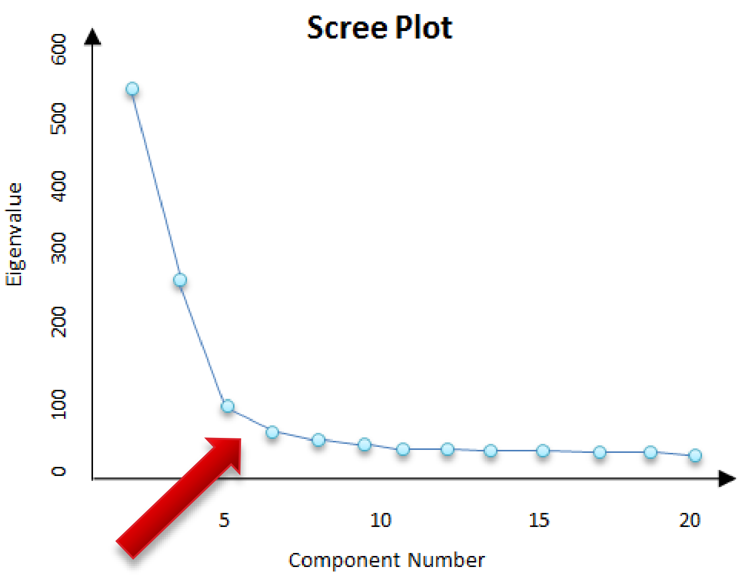
Understanding the Data



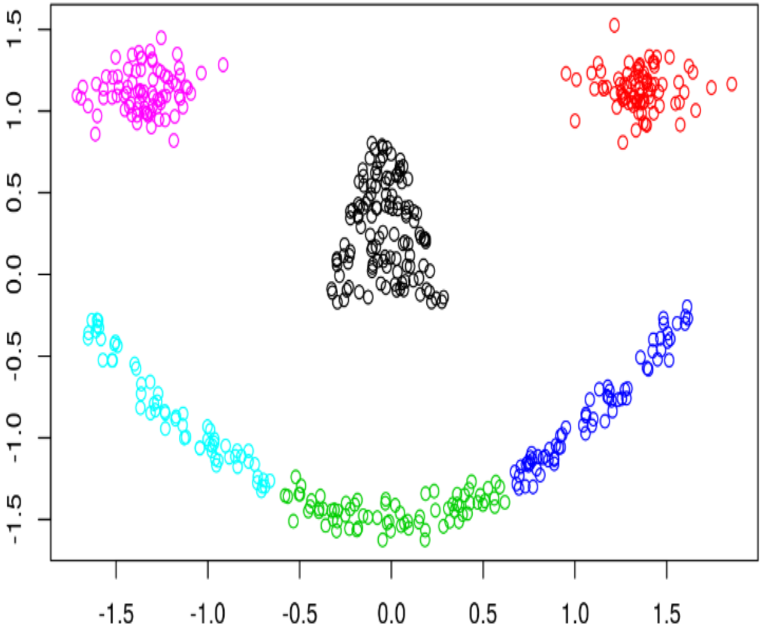
- This dataset clearly shows that there should be 4 clusters: 2 eyes, 1 nose, and 1 mouth.

K-Means

- Lets first run a k-means clustering algorithm:



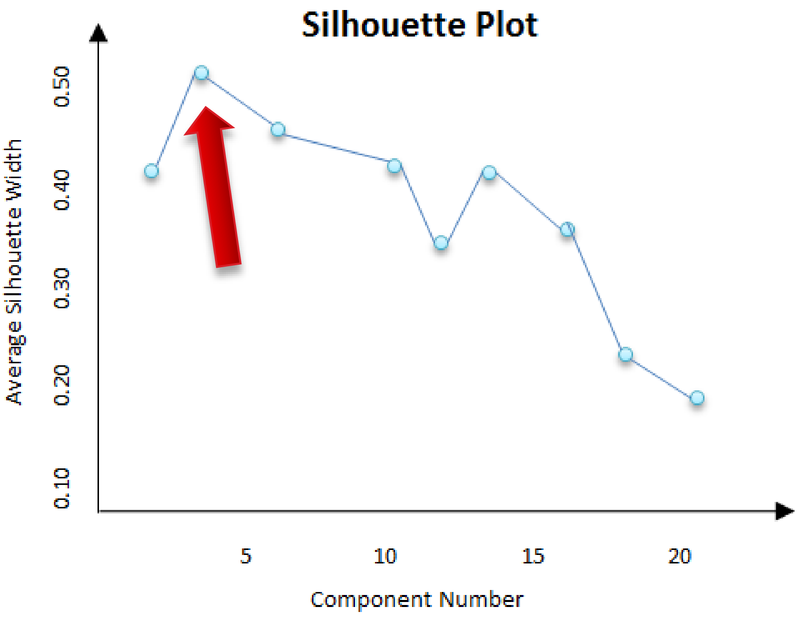
K = 6 Clusters

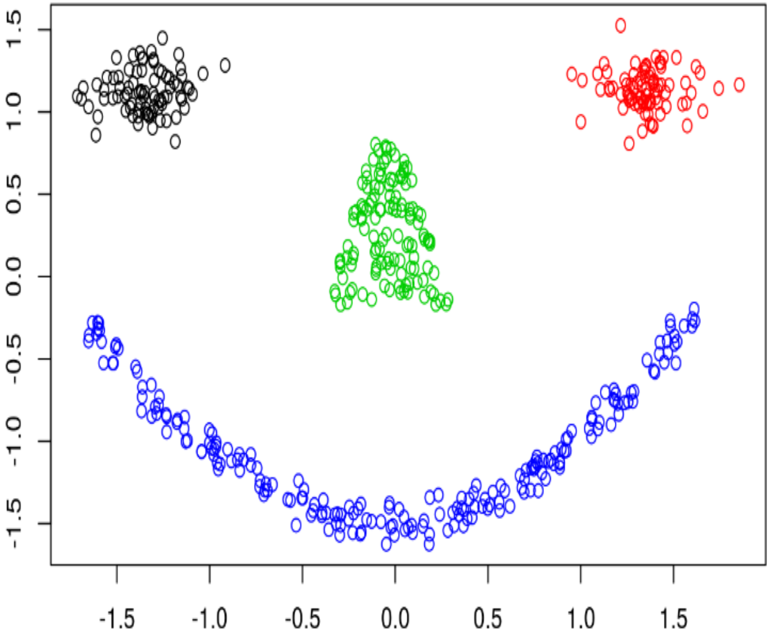


The k-means produced too many clusters based off of the dataset.

Hierarchical clustering

- Now lets use the Hierarchical clustering method:

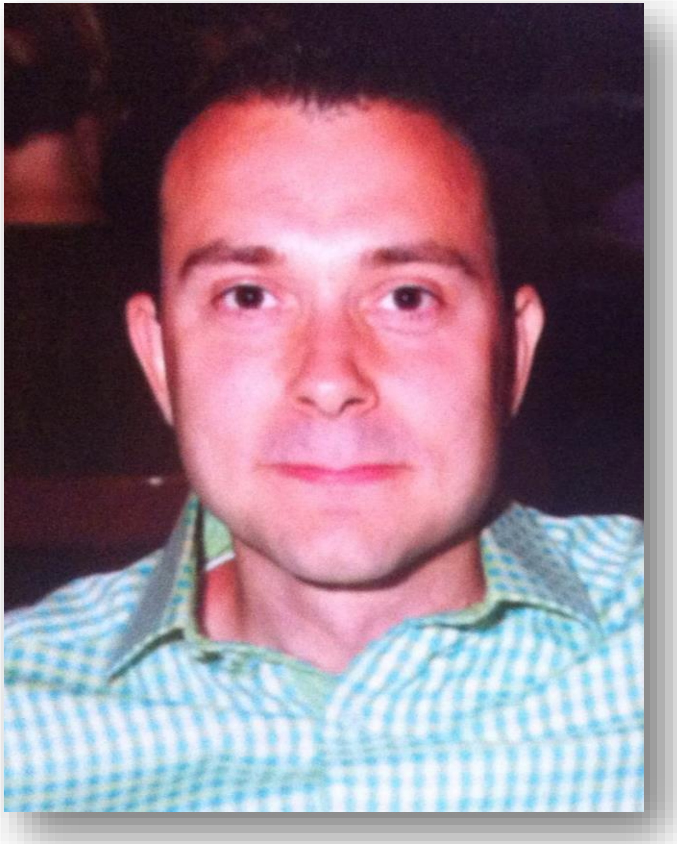




K = 4 Clusters

About Me

- Reside in Wayne, Illinois
- Active Semi-Professional Classical Musician (Bassoon).
- Married my wife on 10/10/10 and been together for 10 years.
- Pet Yorkshire Terrier / Toy Poodle named Brunzie.
- Pet Maine Coons' named Maximus Power and Nemesis Gul du Cat.
- Enjoy Cooking, Hiking, Cycling, Kayaking, and Astronomy.
- Self proclaimed Data Nerd and Technology Lover.



Acknowledgements

- http://www.stat.washington.edu/mclust/
- http://www.stats.gla.ac.uk/glossary/?q=node/451
- http://www.norusis.com/pdf/SPC_v13.pdf
- http://en.wikipedia.org/wiki/Cluster_analysis
- http://michael.hahsler.net/SMU/EMIS7332/R/chap8.html
- http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html
- http://www.autonlab.org/tutorials/gmm14.pdf
- http://en.wikipedia.org/wiki/Distance_matrix

Fine