Cluster Analysis

Today . . .

- Additional clustering methods
- Evaluating cluster quality
- Mixed measures

Additional clustering methods attempt to improve upon limitations in classical approaches:

- Partitioning Around Medoids (PAM)
- Minimax Clustering
- Density-Based Clustering (DBSCAN)
- Fuzzy Sets
- Minimum Spanning Trees

The Partitioning Around Medoids (PAM, Kaufman & Rousseeuw (1990) attempts to overcome limitations in the K-Means method. The K-Means method operates on a Euclidean space. By comparison, PAM operates with any of a broad range of dissimilarities or distances and is generally viewed as being more robust than the K-Means procedure.

Kaufman, L. & P.J. Rousseeuw (1990). *Finding groups in data.* New York: John Wiley & Sons.

A medoid is the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. It is the most centrally located or representative object in the cluster.

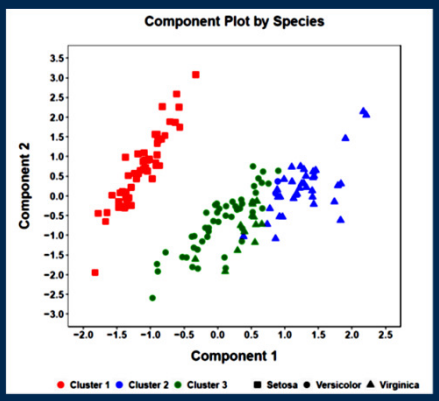
The goal of PAM is to find K medoids that minimize the sum of the dissimilarities of the observations to their closest medoid. The clustering operation identifies the observations (one per cluster) that can be considered representative examples or prototypes.

```
Iris_P <- pam(Iris[, 1:4], k = 3, diss = FALSE, metric = "euclidean")
```

```
Iris_P$clusinfo

##      size max_diss av_diss diameter separation
## [1,]   50    12.37   4.846    24.29     16.401
## [2,]   38    17.23   7.260    24.19      2.646
## [3,]   62    18.38   7.470    26.78      2.646

Iris_P$medoids

##      Sepal_Length Sepal_Width Petal_Length Petal_Width
## [1,]           50          34           15           2
## [2,]           68          30           55          21
## [3,]           60          29           45          15
```

```
Iris_P$id.med

## [1]  96 133  70
```

```
             1  2  3
Setosa      50  0  0
Versicolor   0  2 48
Virginica    0 36 14
```



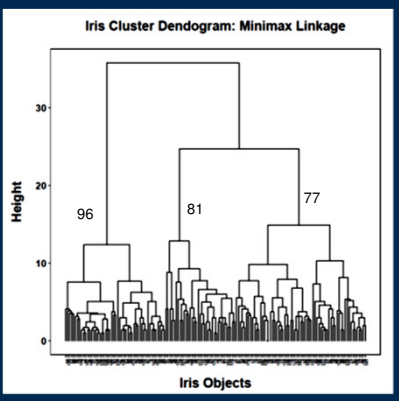
Minimax clustering (Bien & Tibshirani, 2011) is an alternative hierarchical method that resembles PAM in that at each step the method identifies the most highly representative object (the prototype) for the cluster that has been formed.

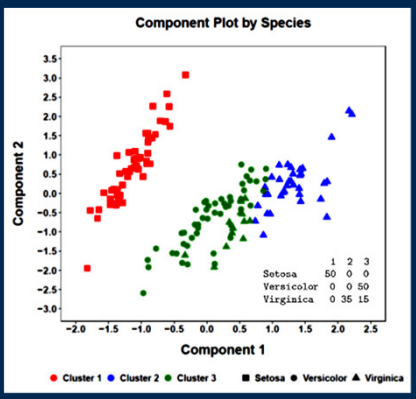
Bien, J., & Tibshirani, R. (2011). *Prototype selection for interpretable classification.* Annals of Applied Statistics, 5, 2403-2424.

The linkage in this method is the radius of the smallest enclosing ball, centered at a point chosen from the two clusters being considered for joining. This is done by identifying the object whose farthest distance from another object (max) is the closest (min). This central object is the prototype for the newly formed cluster.

```
Iris_Proto <- protoclust(Iris_Dist)
```

```
protocut(Iris_Proto, k = 3)[[2]]

## [1] 96 81 77
```

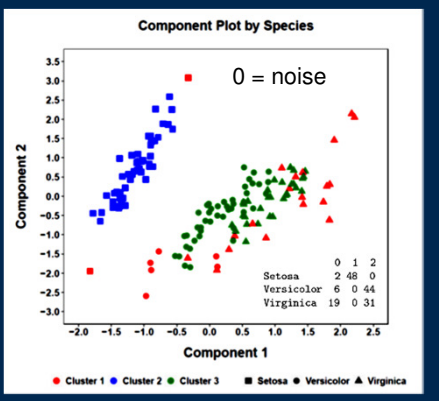


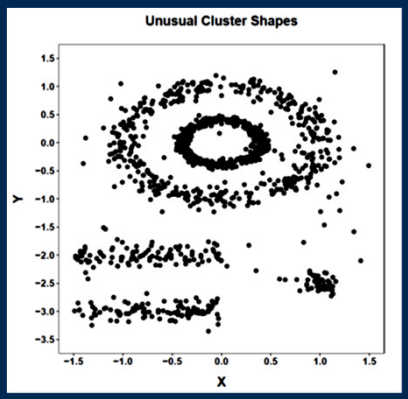


One limitation to hierarchical and partitioning methods is that they work best for circular, spherical, or convex clusters. When clusters have more unusual shapes, then these traditional methods struggle.

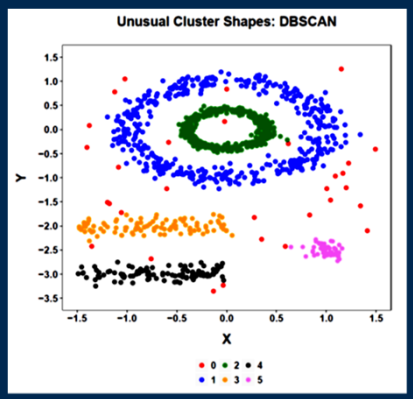
Density-based clustering can provide a solution for finding unusual cluster shapes. The most popular algorithm is called DBSCAN, which stands for density-based spatial clustering and application with noise. The goal of DBSCAN is to identify dense regions.

```
Iris_Density <- fpc::dbscan(Iris_Dist, eps = 5, MinPts = 10, method = "dist")
```

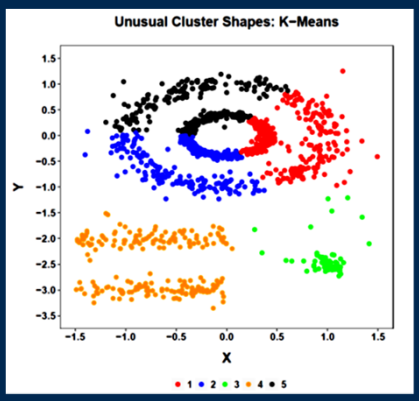




DBSCAN is particularly suited for finding unusual shapes.



The unusual shapes are recovered. Noise is recognized as such.



The K-Means method is able to recover just one of the original shapes, a relatively circular one.

In fuzzy clustering, data objects are treated as members of all clusters with varying degrees of fuzzy membership, indexed by a probability between 0 and 1. Objects closer to the centers of clusters have higher degrees of membership than objects nearer the borders of clusters.

This provides a way to gauge the certainty or confidence with which objects can be classified into categories.

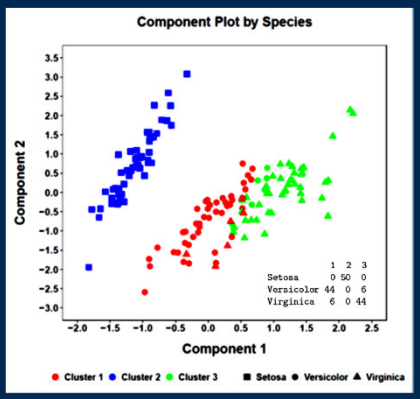
In the Fuzzy C-Means (FCM) clustering algorithm (Bezdek,1974, 1981) the parameter, m, specifies the amount of "fuzziness" of the clustering result. The default value is 2. As m increases, fuzzier clusters are produced. When m is 1, FCM produces the same results as the K-Means procedure.

Bezdek, J.C. (1974). Cluster validity with fuzzy sets. Journal of Cybernetics, 3, 58-73.
Bezdek J.C. (1981). Pattern recognition with fuzzy objective function algorithms. New York: Plenum.

```
Iris_Fuzzy <- fcm(Iris[, 1:4], centers = 3, m = 2, dmetric = "euclidean",
    iter.max = 1000, nstart = 10)
```

```
Iris_Fuzzy$u

##      Cluster 1 Cluster 2 Cluster 3
## 1      0.05260   0.88099   0.06641
## 2      0.66366   0.07755   0.25879
## 3      0.37006   0.07690   0.55304
## 4      0.63413   0.09280   0.27308
## 5      0.55678   0.06521   0.37801
## 6      0.08821   0.80102   0.11077
## 7      0.62755   0.08893   0.28351
## 8      0.32668   0.10714   0.56618
## 9      0.40255   0.08488   0.51257
## 10     0.11935   0.73428   0.14637
```

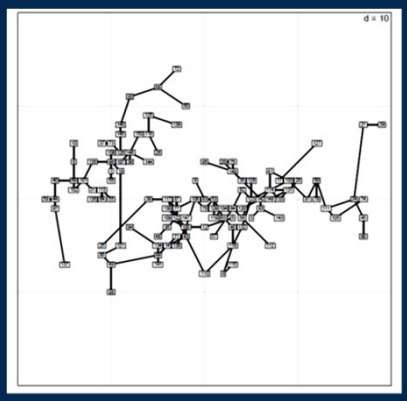
```
Iris_Fuzzy$d

##      Cluster 1 Cluster 2 Cluster 3
## 1       44.389     2.650    35.1548
## 2        5.426    46.436    13.9159
## 3        7.414    35.677     4.9609
## 4        7.018    47.958    16.2970
## 5        4.599    39.265     6.7738
## 6       45.616     5.023    36.3252
## 7        6.268    44.230    13.8743
## 8       11.558    35.244     6.6690
## 9        7.626    36.166     5.9891
## 10      49.558     8.055    40.4093
```



A dissimilarity matrix can be represented in graph theory as an undirected graph with objects as vertices (or nodes) and distances as edge weights. A minimum spanning tree (MST) is then the subset of the edges that connect all the vertices together without any cycles and with the minimum possible sum of edge weights.

In simple terms, if the data have a cluster structure, a minimum spanning tree will have numerous interconnected paths within a cluster but few paths between clusters. The restriction that there be no cycles can be relaxed and successive layers of minimum edges added to the graph. This can highlight clustering in the data to the extent that within-cluster edges outnumber between-cluster edges.

```
Iris_MST <- ade4::mstree(dist(Iris[, 1:4], method = "euclidean"),
    ngmax = 1)
s.label(Iris[, 1:4], xlim = c(50, 70), ylim = c(10, 50), addaxes = TRUE,
    neig = Iris_MST, clabel = 0.5, pch = 16)
```



The "quality" of a clustering solution can be assessed in a number of ways. Here are some of the more popular ones:

- Silhouette coefficient
- Cophenetic correlation
- Pseudo F
- Hopkins statistic
- Duda-Hart statistic
- Gap statistic
- Rand coefficient

The silhouette score is calculated using two values for each object, a_i and b_i. The value, a_i, is the average distance between $Object_i$ and all other objects in the same cluster. The value, b_i, is the smallest average distance of $Object_i$ to all objects in another cluster. These two values are sometimes called *cohesion* and *separation*, respectively.

The silhouette score (sometimes called its width), s_i, is defined as:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

Silhouette scores can take on values between -1 and 1, with higher values indicating that an object is well matched to its cluster and a poor match to any neighboring clusters.

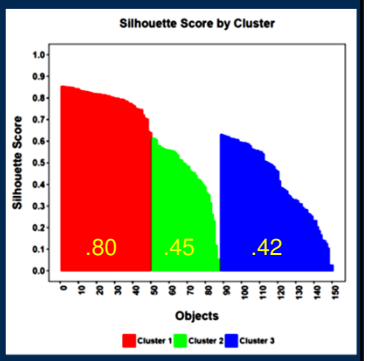
The average silhouette score is provided for each cluster and for the overall solution. The overall average, sometimes called the silhouette coefficient, is an index of cluster quality.

Coefficients that approach 1 represent very clear evidence that the chosen cluster number produces a good cluster solution.

Some common benchmarks for the average cluster silhouette coefficient:

.71 to 1.00	A strong structure has been found
.51 to .70	A reasonable structure has been found
.26 to .50	The structure is weak and could be artificial
< .25	No substantial structure has been found

The average silhouette for the three-cluster solution (.55) with the Iris data suggest that a reasonable structure has been found but some cases (other than *Setosa* members) are clearly easily moved.



The cophenetic distance between two observations that have been clustered hierarchically is defined to be the intergroup dissimilarity at which the two observations are first combined into a single cluster.

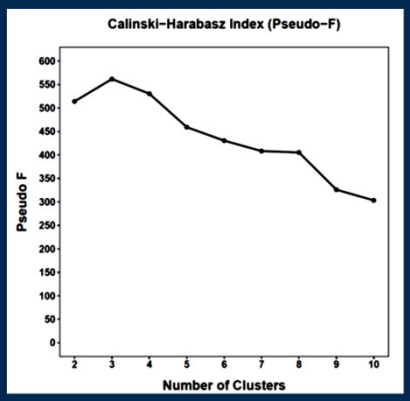
The correlation between the original distances and the cophenetic distances is an index of how well a dendrogram preserves the pairwise distances between the original objects.

```
Iris_HC <- hclust(d1, "ward.D2")
d2 <- cophenetic(Iris_HC)
cor(d1, d2)

## [1] 0.8728
```

The Calinski-Harabasz (Pseudo F) statistic closely resembles the F test in ANOVA:

$$Pseudo\ F = \frac{\frac{SS_{BC}}{C-1}}{\frac{SS_{WC}}{N-C}}$$

These values can be displayed for different numbers of clusters, with the maximum indicating the best number of clusters in the sense of maximizing their separation and cohesion.



Three clusters in the K-Means method are optimal.

The Hopkins statistic examines whether objects in a data set differ significantly from the assumption that they are uniformly distributed in the multidimensional space. It compares the distances, x_i, between the real objects and their nearest neighbors to the distances, y_i, between artificial objects and their nearest neighbors, with the artificial objects uniformly generated over the data space.

$$H = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i}$$

Values close to .5 indicate data are uniformly distributed. As H approaches 0, the data exhibit increasing clustering.

```
Iris_M <- as.matrix(Iris[, 1:4])
hopkins(Iris_M, n = 149, byrow = FALSE, header = FALSE)

## $H
## [1] 0.1649
```

The Duda-Hart test indicates if a data set should be split into two clusters. It thus represents a very basic test that indicates if clustering is justified.

Variants exist for different clustering methods. The one used here is suitable for interval level data for which sums of squares would be an appropriate calculation. The dh value calculated here is the ratio of the within-cluster sum of squares for two clusters to the overall sum of squares.

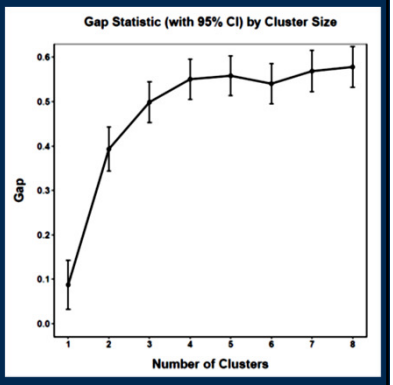
```
Iris_K <- kmeans(Iris[, 1:4], centers = 2)
Duda_Hart <- dudahart2(Iris[, 1:4], Iris_K$cluster, alpha = 0.001)
Duda_Hart

## $p.value
## [1] 0
##
## $dh
## [1] 0.2236
##
## $compare
## [1] 0.6815
##
## $cluster1
## [1] FALSE
##
## $alpha
## [1] 0.001
##
## $z
## [1] 3.09
```

The data depart significantly from a single cluster solution.

The gap statistic (Tibshirani et al., 2000) compares the within-cluster dispersion to that expected under an appropriate reference null distribution, which assumes random dispersion (e.g., uniform or Gaussian on the range of the original variables or a simplified space [e.g., PC]).

The latter is defined by bootstrapping from the null reference distribution. As the obtained WSS curve departs ("gap") from that expected under the reference curve, there is evidence for non-random lumpiness in the data.

In the Iris data there is no significant improvement beyond three clusters and the evidence for three is reasonably strong.



The Rand coefficient can be used to compare two clustering methods. The simple Rand coefficient is given by:

$$R = \frac{a + b}{\binom{N}{2}}$$

in which **a** is the number of times a pair of objects is classified together across the two methods and **b** is the number of times a pair of objects is classified in different clusters across two methods. The denominator is the number of unique pairs of objects. A corrected version of the Rand Coefficient that takes chance agreement into account.

```
Iris_HC_1 <- hclust(d1, "single")
C1 <- cutree(Iris_HC_1, k = 3)
Iris_HC_2 <- hclust(d1, "ward.D2")
C2 <- cutree(Iris_HC_2, k = 3)
Iris_HC_3 <- hclust(d1, "average")
C3 <- cutree(Iris_HC_2, k = 3)
CS_1_2 <- cluster.stats(Iris_Dist, C2, C1, silhouette = TRUE, G2 = FALSE,
    G3 = FALSE, wgap = TRUE, sepindex = TRUE, sepprob = 0.1, sepwithnoise = TRUE,
    compareonly = FALSE, aggregateonly = FALSE)
CS_1_2$corrected.rand
## [1] 0.6069

CS_2_3 <- cluster.stats(Iris_Dist, C2, C3, silhouette = TRUE, G2 = FALSE,
    G3 = FALSE, wgap = TRUE, sepindex = TRUE, sepprob = 0.1, sepwithnoise = TRUE,
    compareonly = FALSE, aggregateonly = FALSE)
CS_2_3$corrected.rand
## [1] 1
```

Ward's method agrees poorly with the single linkage method but very well with the average linkage method.

For all of the data sets examined in past examples, all measures shared the same level of measurement (e.g., interval). When measures with different levels of measurement, the Gower distance provides a way to combine them.

For each variable type, a particular distance metric that works well for that type is used and resulting distances scaled to fall between 0 and 1. The distances for each pair of objects are then averaged to create the final distance matrix.

The Gower distance is always a number between 0 (identical) and 1 (maximally dissimilar).

Simply knowing the number and size of clusters in a data set is an important goal. But, sometimes we want to know more than that. The next step beyond identification of clusters is prediction of cluster membership using additional variables.

Next time . . .

Discriminant analysis