

Agglomerative Hierarchical Clustering Methods

Hierarchical Clustering Methods

- Agglomerative hierarchical methods
 - Begin with as many clusters as objects. Clusters are successively merged until only one cluster remains.
- Divisive hierarchical methods
 - Begin with all objects in one cluster. Groups are continually divided until there are as many clusters as objects.

Steps in Agglomerative Hierarchical Clustering

1. Start with N clusters, each containing a single entity, and an $N \times N$ symmetric matrix of distances (or similarities)

Let

d_{ij} = distance between item i and item j .

Steps in Agglomerative Hierarchical Clustering

2. Search the distance matrix for the nearest pair clusters (i.e., the two clusters that are separated by the smallest distance).

Denote the distance between these most similar clusters U and V by d_{UV} .

Steps in Agglomerative Hierarchical Clustering

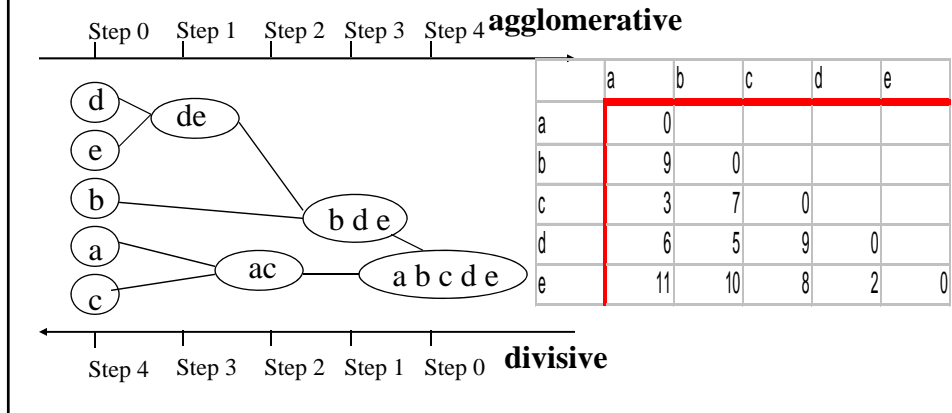
3. Merge clusters U and V into a new cluster, labeled T. Update the entries in the distance matrix by
 - a. Deleting the rows and columns corresponding to clusters U and V, and
 - b. Adding a row and column giving the distances between the new cluster T and all the remaining clusters.

Steps in Agglomerative Hierarchical Clustering

4. Repeat steps (2.) and (3.) a total of N-1 times.

Hierarchical Clustering

Use distance matrix as clustering criteria. This method does not require the number of clusters k as an input, but needs a termination condition.

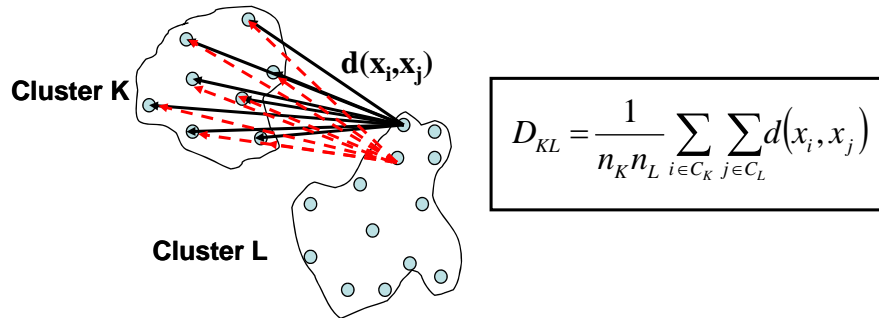


Linkage Methods for Hierarchical Clustering

Method	Multivariate Data	Distance Data
Average Linkage	√	√
Centroid Linkage	√	√
Complete Linkage		√
Density Linkage		√
Single Linkage		√
Two-Stage Linkage		√

Average Linkage Method

The distance between clusters is the average distance between pairs of observations.

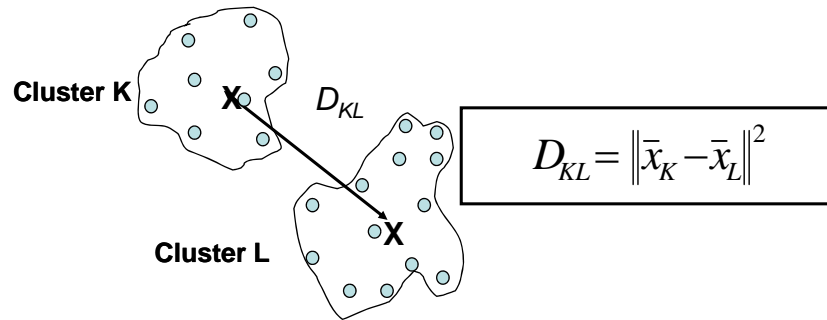


Average Linkage Method

- Average linkage tends to join clusters with small variances, and it is slightly biased toward producing clusters with the same variance.
- Because it considers all members in the cluster rather than just a single point, however, average linkage tends to be less influenced by extreme values than other methods.

Centroid Linkage Method

The distance between clusters is defined as the (squared) Euclidean distance between cluster centroids \bar{x}_K and \bar{x}_L .

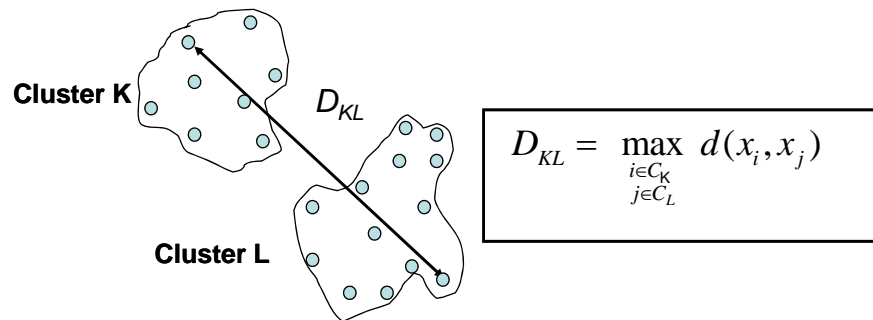


Centroid Linkage Method

- Because the centroid method compares cluster means, outliers affect it less than most other hierarchical clustering methods.
- In other respects, however, it may not perform as well as Ward's method or average linkage (Milligan 1980).
- The larger of two unequally sized groups merged using centroid linkage tends to dominate the merged cluster.

Complete Linkage Method

The distance between two clusters is based on the points in each cluster that are furthest apart.

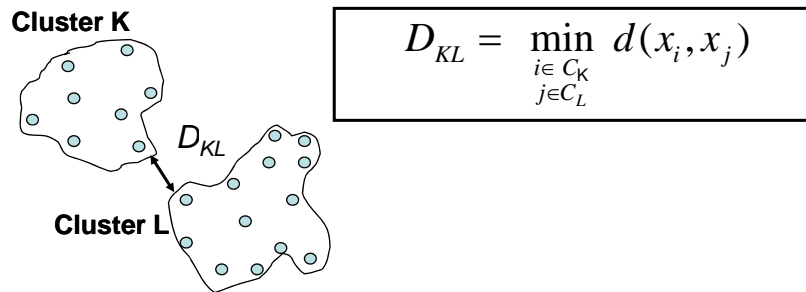


Complete Linkage Method

- Complete linkage is strongly biased toward producing compact clusters with roughly equal diameters, and it can be severely distorted by moderate outliers.
- Complete linkage ensures that all items in a cluster are within some maximum distance of one another.

Single Linkage Method

The distance between two clusters is based on the points in each cluster that are nearest together.



Single Linkage Method

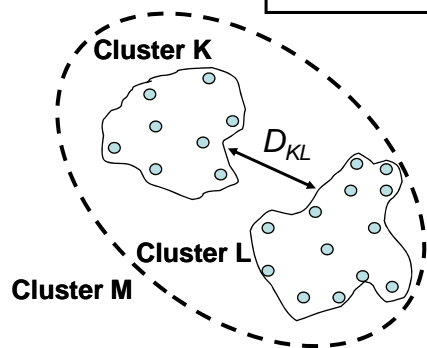
- Single linkage clustering is also known as *nearest-neighbor clustering*.
- Single linkage has many desirable theoretical properties but has fared poorly in Monte Carlo studies.
- By imposing no constraints on the shape of clusters, single linkage sacrifices performance in the recovery of compact clusters in return for the ability to detect elongated and irregular clusters.
- Also, single linkage tends to chop off the tails of distributions before separating the main clusters.
- The notorious chaining tendency of single linkage can be alleviated by specifying the TRIM= option.

Other Methods for Hierarchical Clustering

Method	Multivariate Data	Distance Data
EML	√	
Flexible-Beta		√
McQuitty's		√
Median		√
Ward's	√	√

Equal-Variance Maximum Likelihood (EML) Method

$$D_{KL} = n\gamma \ln \left(1 + \left[\frac{B_{KL}}{P_G} \right] \right) - p(n_M \ln(n_M) - n_K \ln(n_K) - n_L \ln(n_L))$$



Unique to SAS

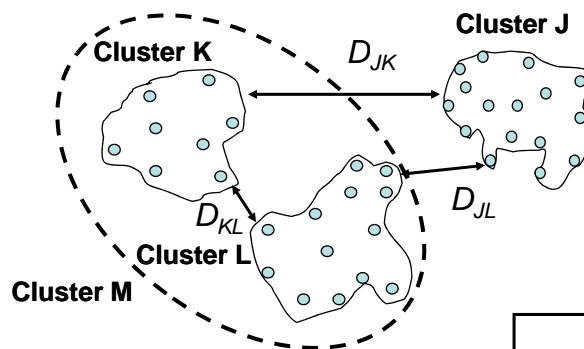
Equal-Variance Maximum Likelihood (EML) Method

The EML method joins clusters to maximize the likelihood at each level of the hierarchy under the following assumptions.

- multivariate normal mixture
- equal spherical covariance matrices
- unequal sampling probabilities

The EML method is similar to Ward's minimum-variance method but removes the bias toward equal-sized clusters. Practical experience has indicated that EML is somewhat biased toward unequal-sized clusters.

Median Method

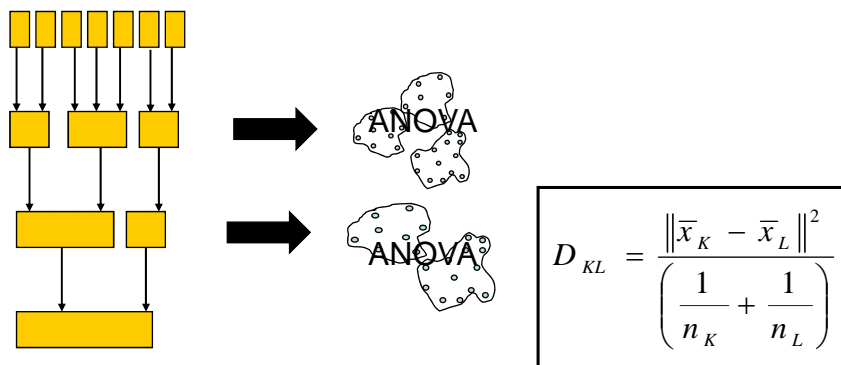


$$D_{JM} = \frac{D_{JK} + D_{JL}}{2} - \frac{D_{KL}}{4}$$

Median Method

- The median method has performed poorly in simulation studies.
- It offers few (if any) advantages over the other methods.
- The median method assumes that the observations can be represented in Euclidean space. When true, the new cluster can be interpreted as in an intermediate position between the merged groups.
- The median method is subject to cluster assignment reversals.

Ward's Minimum-Variance Method



Biased toward equal-sized clusters

Ward's Minimum-Variance Method

For the k^{th} cluster, define the Error Sum of Squares as

ESS_k = sum of squared deviations from the cluster centroid

If there are C clusters, define the Total Error Sum of Squares as

ESS = Sum of ESS_k , for $k=1, \dots, C$

Ward's Minimum-Variance Method

Consider the union of every possible pair of clusters.

Combine the 2 clusters whose combination results in the smallest increase in ESS .

Ward's Minimum-Variance Method

In Ward's minimum-variance method, the distance between two clusters is the *ANOVA* sum of squares between the two clusters added up over all the variables.

At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation.

The sums of squares are easier to interpret when they are divided by the total sum of squares to give proportions of variance (squared semipartial correlations).

Ward's Minimum-Variance Method

Ward's method joins clusters to maximize the likelihood at each level of the hierarchy under the following assumptions:

- multivariate normal mixture
- equal spherical covariance matrices
- equal sampling probabilities

Ward's Minimum-Variance Method

Ward's method tends to join clusters with a small number of observations, and it is strongly biased toward producing clusters with the same shape and with roughly the same number of observations.

It is also very sensitive to outliers.

Problems with Hierarchical Clustering

- There is no particular hierarchical clustering method that can be recommended (Everitt et al., 2001).
- Hierarchical methods do not scale up well with the number of observations.
- After they are made, divisions are irrevocable. As Kaufman and Rousseeuw (1990) put it, "A hierarchical method suffers from the defect that it can never repair the damage that was done in previous steps."

The TREE Procedure

Alternatively, the diagram can be oriented horizontally, with the root at the left.

Any numeric variable in the output data set can be used to specify the heights of the clusters.

PROC TREE can also create an output data set containing a variable to indicate the disjoint clusters at a specified level in the tree.

Sarle's Cubic Clustering Criterion

Sarle's cubic clustering criterion (CCC) tests the following hypothesis:

H_0 = the data has been sampled from a uniform distribution on a (hyper)box

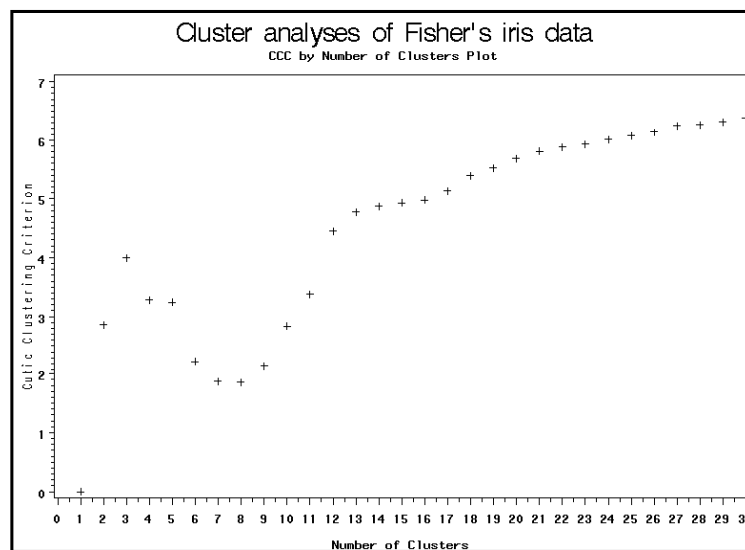
H_1 = the data has been sampled from a mixture of spherical multivariate normal distributions, with equal variances and sampling probabilities.

Positive CCC values mean that the obtained R^2 value is greater than would be expected if the sampling was from a uniform distribution (therefore, reject H_0).

Sarle's Cubic Clustering Criterion

- Plot CCC vs. # of clusters.
- Look for peaks where $CCC > 3$.

Graphically Interpreting Sarle's CCC



Interpreting the Cluster History CCC

Cluster History								
NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	T i e
10	CL12	CL44	25	0.0099	.911	.891	2.84	
9	CL11	CL15	45	0.0129	.898	.882	2.16	
8	CL18	CL27	18	0.0130	.885	.870	1.86	
7	CL17	CL22	29	0.0137	.872	.854	1.87	
6	CL8	CL20	26	0.0149	.857	.834	2.24	
5	CL10	CL16	30	0.0150	.842	.806	3.24	
4	CL19	CL7	49	0.0364	.805	.764	3.28	
3	CL9						4.00	
2	CL3	CL5	101	0.1331	.619	.552	2.86	
1	CL4	CL2	150	0.6188	.000	.000	0.00	

Recommended Solution



The Pseudo- F Statistic

- The pseudo- F statistic (or PSF) measures the separation among the clusters at the current level in the hierarchy.
- Large values indicate that the mean vectors of all clusters are different.
- Look for peaks in the PSF value, and choose cluster solutions corresponding to the peaks.
- It is **not** distributed as an F random variable.

The Pseudo- F Criterion

Cluster History										
NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Time
15	Oman	CL37	5	0.0039	.957	.933	6.03	132	12.1	
14	CL31	CL22	13	0.0040	.953	.928	5.81	131	9.7	
13	CL41	CL17	32	0.0041	.949	.922	5.70	131	13.1	
12	CL19	CL21	10	0.0045	.945	.916	5.65	132	6.4	
11	CL39	CL15	9	0.0052	.940	.909	5.60	134	6.3	
10	CL76	CL27	6	0.0075	.932	.900	5.25	133	18.1	
9	CL23	CL11	15	0.0130	.919	.890	4.20	125	12.4	
8	CL10	Afghanistan	7	0.0134	.895	.878	3.55	122	7.3	
7	CL9	CL25	11	0.0134	.895	.878	3.55	114	11.6	
6	CL8	CL20	14	0.0239	.860	.846	1.42	112	10.5	
5	CL14	CL13	45	0.0307	.829	.822	0.65	112	59.2	
4	CL16	CL7	28	0.0323	.797	.788	0.57	122	14.8	
3	CL12	CL6	24	0.0323	.765	.732	1.84	153	11.6	
2	CL3	CL4	52	0.1782	.587	.613	-.82	135	48.9	
1	CL5	CL2	97	0.5866	.000	.000	0.00	.	135	

Potential Solutions

The Pseudo- T^2 Statistic

- The pseudo- T^2 statistic is a variant of Hotelling's T^2 test.
- If the pseudo- T^2 statistic value is large, then the two clusters being considered should **not** be combined, since the mean vectors of these two clusters can be regarded as different.
- If the value is small, the clusters can safely be combined.
- Move down the column until you find the first value markedly larger than the previous value, then choose as the cluster solution the one corresponding to the previous value.

The Pseudo- T^2 Criterion

Cluster History										
NGL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Time
15	Oman	CL37	5	0.0039	.957	.933	6.03	132	12.1	
14	CL31	CL22	13	0.0040	.953	.928	5.81	131	9.7	
13	CL41	CL17	32	0.0041	.949	.922	5.70	131	13.1	
12	CL19	CL21	10	0.0045	.945	.916	5.65	132	6.4	
11	CL39	CL15	9	0.0052	.940	.909	5.60	134	6.3	
10	CL76	CL27	6	0.0075	.932	.900	5.25	173	18.1	
9	CL23	CL11	15	0.0100	.915	.888	4.99	125	12.4	
8	CL10	Afg	12	0.0100	.915	.888	4.99	122	7.3	
7	CL9	CL25	17	0.0217	.884	.864	2.26	114	11.6	
6	CL8	CL20	14	0.0239	.860	.846	1.42	112	10.5	
5	CL14	CL13	45	0.0307	.829	.822	0.65	112	59.2	
4	CL16	CL7	28	0.0323	.797	.788	0.57	122	14.8	
3	CL12	CL6	24	0.0323	.765	.732	1.84	153	11.6	
2	CL3	CL4	52	0.1782	.587	.613	-.82	135	48.9	
1	CL5	CL2	97	0.5866	.000	.000	0.00	.	135	

Potential Solutions

Beale's F -Type Statistic

$$\frac{(w_2 - w_1)}{w_1} \cdot \frac{(n - c_1)k_1}{(n - c_2)k_2 - (n - c_1)k_1}$$

- Calculates the uncorrected sum of squares for the distance each observation is from its cluster mean.
- If Beale's F -type statistic is greater than the critical F statistic, choose the cluster solution with more clusters.
- Otherwise, the cluster solution with the smaller number of clusters is preferable.

R² Criterion

For a given level of the hierarchy

$$R^2 = 1 - \frac{\sum_{k \text{ clusters}} W_k}{T}$$

Plot R² vs. # of clusters.

Look for where the curve levels off.

Semi-Partial R² Criterion

For joining clusters C_K and C_L

$$R^2 = \frac{B_{KL}}{T}$$

= proportional reduction in variance
due to joining clusters C_K and C_L

Look for small values (indicating that the
2 clusters can be regarded as one).