Grouping **Clustering analysis**

Contents

IV.B. Clustering analysis

1 Clustering analysis

An unsupervised learning algorithm (term used by AI community).

1.1 Primary objective

Identify and classify objects (objects can be either observations or variables) into one of a (unspecified) number of clusters (groups) so that clusters exhibit high within-cluster homogeneity (similarity) and high between-cluster heterogeneity (dissimilarity).

1.2 Other names for clustering analysis

Q analysis, typology, classification analysis, numerical taxonomy, ...

1.3 Need for algorithms

Seek for “natural” groupings ⇒ preclassification. For example, in grouping for 16 face cards (A, K, Q, J of 4 suits), there are 32,767 ways of partitioning them into 2 groups of varying sizes, and 7,141,686 ways of partitioning them into 3 groups of varying sizes. TOO MANY NATURAL GROUPS. Algorithms are therefore in store.

1.4 Three stages of clustering analysis

Stage 1 : Partitioning (forming clusters)

1. What variables are used to compute similarity (or distance) among objects?
2. How to define similarity (or distance) measure?
3. What algorithm should be used to place similar objects into clusters?
4. How many clusters should be formed?

Stage 2 : Interpretation.

Stage 3 : Validation and Profiling.

1.5 Variable selection

The selection of variables to be included in clustering analysis must

- be done with regard to both theoretical/conceptual and practical considerations;
- include only those variables that characterize the objects being clustered, and relate specifically to the objectives of the cluster analysis;
- avoid variables that produce small range across objects.

2 Similarity measures and distance measures

1. Between observations

   (a) **Quantitative variables**
   It is advisible to use "true" distance measure (or metric) whenever possible. A metric satisfies the followig properties. Consider objects denoted by $p \times 1$ vector

   $$\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^t,$$

   where t is the matrix transpose. The distance between two objects $\boldsymbol{x}$ and $\boldsymbol{y}$, denoted by $d(\boldsymbol{x}, \boldsymbol{y})$, should satisfies

   $< a >$ $d(\boldsymbol{x}, \boldsymbol{x}) = 0$, and $d(\boldsymbol{x}, \boldsymbol{y}) > 0$ if $\boldsymbol{x} \neq \boldsymbol{y}$;
   $< b >$ $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$;
   $< c >$ $d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y})$.

   Some distance measures:

   i. <u>Minkowski</u>

   $$d(\boldsymbol{x}, \boldsymbol{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{\frac{1}{m}}.$$

   When $m = 1$, we have 'city-block' distance which works best if the variables are uncorrelated and unit scales of the variables are compatible. When $m = 2$, it is the usual Euclidean distance and can be written in the form:

   $$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^t(\boldsymbol{x} - \boldsymbol{y})}.$$

   The Euclidean distance is normally used in statistics after data are standardized. The distance after standardization can be shown as

   $$d_A(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^t A(\boldsymbol{x} - \boldsymbol{y})}.$$

   If $A = \text{diag}(1/s_1^2, \ldots, 1/s_p^2)$, that is, a diagonal matrix with diagonal elements formed by the sample variances of the p variables, then the standardization was executed on individual variable (by subtracting mean from each observation and then divided by sample standard deviation). However, the potential intercorrelation among the variables may cause problem if this standardization procedure is employed. A more reasonable standardization is to adjust for intercorrelation and is given by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^t S^{-1} (\boldsymbol{x} - \boldsymbol{y})},$$

where S is the sample variance-covariance matrix in which the principal diagonal (NW-SE direction) elements are sample variances and the off principal diagonal elements are sample covariances (covariance between two variables in their original scale units quantifies the extent the two variables spread out together). This is the well-known '*Mahalanobis*' distance.

ii. Canberra

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i}.$$

iii. Czekanowski

$$d(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{2 \sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} x_i + y_i}.$$

Similarity can always be constructed from distance, for instance,

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{1 + d(\boldsymbol{x}, \boldsymbol{y})}.$$

However, "true" distance may not be able to be constructed for some similarity measure.

Cautions should be taken when defining/using distance measure:

i. Many metric are particularly sensitive to outliers (extreme observations). A preliminary screening for outliers is advisable. Then '*eliminate*' outliers before clustering.

ii. Different metrics may lead to different cluster solutions. Hence, it is advisable to use several measures and compare the results to theoretical or known patterns.

iii. When variables have different units (not compatible), one should standardize the data before running the cluster analysis.

iv. Mahalanobis metric is preffered if the variable are intercorrelated.

(b) **Qualitative variables**

Example: Consider two objects, and 5 binary variables:

	variable, k				
object	1	2	3	4	5
i	1	0	0	1	1
j	1	1	0	1	0

where a value of 1 or 0 for a variable (characteristic) denote the presence or absence of the charactersitic. Then

$$(x_{ik} - x_{jk})^2 = \begin{cases} 0, & \text{if } x_{ik} = x_{jk} = 1 \text{ (both present)} \\ & \quad \text{or } x_{ik} = x_{jk} = 0 \text{ (both absent)}, \\ 1, & \text{if } x_{ik} \neq x_{jk} \text{ (mis-match)} \end{cases}$$

and the Euclidean distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$ gives the mis-match counts of the two objects (here number of mis-matched is 2). Contingency table on counts of the matches or mis-matches in p variables for the two objects can then be constructed:

		object j		
		1	0	totals
object	1	a	b	$a+b$
i	0	c	d	$c+d$
	totals	$a+c$	$b+d$	$p=a+b+c+d$

(example above)

		object j		
		1	0	totals
object	1	2	1	3
i	0	1	1	2
	totals	3	2	5

Here are some similarity measures defined upon these:

	Similarity	[eg above]	Rationale
(a)	$\frac{a+d}{p}$	$[\frac{3}{5}]$	Equal weights for 1-1 and 0-0 matches
(b)	$\frac{2(a+d)}{2(a+d)+b+c}$	$[\frac{3}{4}]$	Double weights for 1-1 and 0-0 matches
(c)	$\frac{a+d}{a+d+2(b+c)}$	$[\frac{3}{7}]$	Double weights for mis-matches
(d)	$\frac{a}{p}$	$[\frac{2}{5}]$	Only 1-1 matches are considered similar
(e)	$\frac{a}{a+b+c}$	$[\frac{1}{2}]$	0-0 matches are irrelevant
(f)	$\frac{2a}{2a+b+c}$	$[\frac{2}{3}]$	Double weights for 1-1; 0-0's are irrelevant
(g)	$\frac{a}{a+2(b+c)}$	$[\frac{1}{3}]$	Double weights for mis-matches; 0-0's are irrelevant
(h)	$\frac{a}{b+c}$	1	Ratio of 1-1 matches against mis-matches

Example: Consider 5 individuals on 6 variables:

		X_1		X_2	X_3	X_4	X_5	X_6
individual	height	(height≥72in?)	weight	(weight≥150lb?)	Eye Color	Hair Color	Handedness	Gender
1	68(in.)	0	140(lb.)	0	0(green)	1(blond)	1(right)	1(female)
2	73	1	185	1	1(brown)	0(brown)	1	0(male)
3	67	0	165	1	0(blue)	1	1	0
4	64	0	120	0	1(brown)	0	1	1
5	76	1	210	1	1(brown)	0	0(left)	0

Using similarity measure (a), similarity matrix is

$$\tilde{S} = \begin{array}{cc} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 1 & & & & \\ \frac{1}{6} & 1 & & & \\ \frac{4}{6} & \frac{3}{6} & 1 & & \\ \frac{4}{6} & \frac{3}{6} & \frac{2}{6} & 1 & \\ 0 & \frac{5}{6} & \frac{2}{6} & \frac{2}{6} & 1 \end{bmatrix} \end{array}.$$

If the similarity matrix is nonnegative definite, i.e., all eigen values are nonnegative (in the example, the eigen values are 2.75, 1.36, 0.67, 0.16, and 0.06), and $s(\boldsymbol{x}_i, \boldsymbol{x}_i) = 1$, then one can define distance by

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{2(1 - s(\boldsymbol{x}_i, \boldsymbol{x}_j))}.$$

The distance matrix for the example using this transformation:

$$D = \begin{array}{cc} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & & \\ 1.29 & 0 & & & \\ 0.82 & 1 & 0 & & \\ 0.82 & 1 & 1.15 & 0 & \\ 1.41 & \boxed{0.58} & 1.15 & 1.15 & 0 \end{bmatrix} \end{array}.$$

2. Between variables

   (a) Quantitative variables: Use correlation$+1$ or —correlation— for similarity.

   (b) Qualitative variables:
   Example: Binary variables; n observations

		variable j		
		1	0	totals
variable	1	a	b	$a+b$
i	0	c	d	$c+d$
	totals	$a+c$	$b+d$	$n=a+b+c+d$

usual (product moment) correlation:

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

gives a similarity measure.

3 Clustering Algorithms

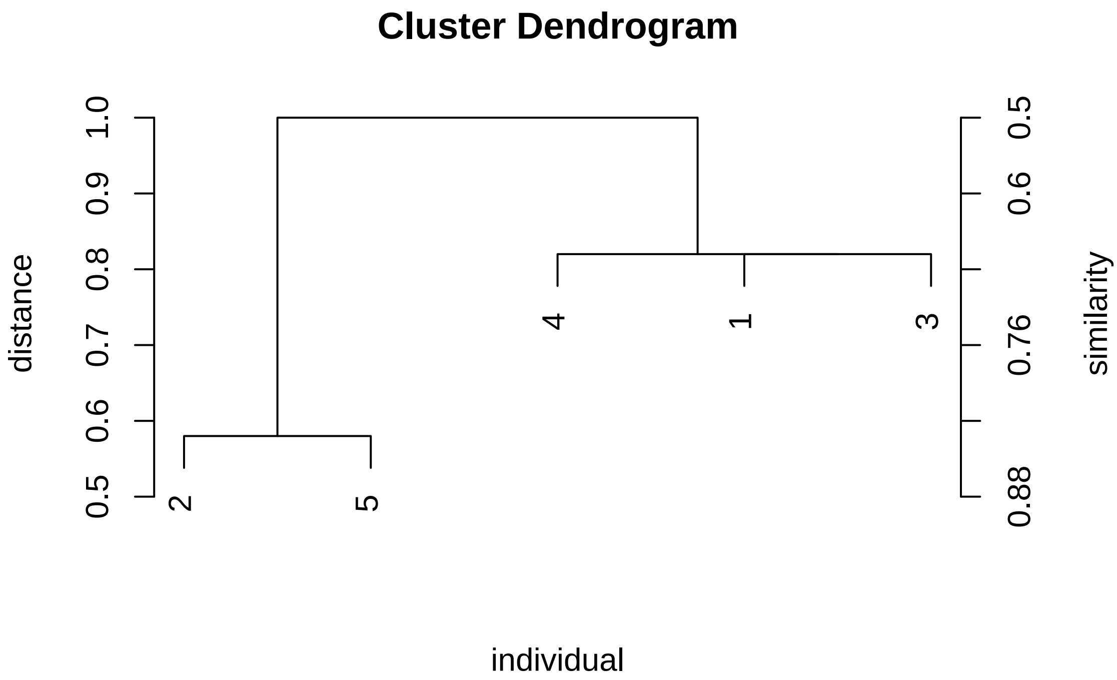
Any algorithm is used to maximize the ‘differences’ between clusters relative to the variation within the clusters.

1. Hierarchical clustering procedures.

2. Nonhierarchical procedures.

4 Hierarchical clustering procedures

There are two types of hierarchical clustering procedures—*agglomerative* and *divisive*. In a agglomerative procedure, each object starts out as its own cluster. In the subsequent steps, the two ‘closest’ clusters/objects are combined into a new aggregate cluster, thus reducing the number of clusters by one in each step. Eventually, all objects are grouped into one large cluster. In a divisive procedure, the process proceeds in the opposite direction: we start out with one large cluster containing all objects; in the subsequent steps, the objects that are most dissimilar are split off and turned into smaller clusters; this process continues until each object forms a cluster of itself. Most computer packages use agglomerative procedures. After clustering procedure is completed, the outcomes are best represented and visualized by a **dendrogram** as the one showing the above example:



4.1 Agglomerative Clustering Procedures

Five popular agglomerative procedures used to form clusters are

1. single linkage (closely related to the minimal spanning tree) adopts a'friends of friends' clustering strategy. It tends to create fewer clusters than the next method. The single linkage method is based on minimum distance (hence maximum 'connectedness' or similarity). It finds the two objects separated by the shortest distance and places them in the first cluster. Then the next shortest distance is found, and either a third object joins the previously formed cluster or a new two-object cluster is formed. This process continues until all objects are in on cluster. The dendrogram above shows the result of clustering when the single linkage was applied. The sequence of steps:

   Step 1 . (cluster $\boxed{2\text{-}5}$ is formed) the updated distance matrix is

$$\begin{array}{cccccc} & & 1 & \boxed{2\text{-}5} & 3 & 4 \\ 1 & \left[\right. & 0 & & & \\ \boxed{2\text{-}5} & & 1.29 & 0 & & \\ 3 & & \boxed{0.82} & 1 & 0 & \\ 4 & & 0.82 & 1 & 1.15 & 0 \left.\right] \end{array},$$

   Note that, for instance,

$$d(\boldsymbol{x}_1, \boxed{\boldsymbol{x}_2\text{-}\boldsymbol{x}_5}) = \min(d(\boldsymbol{x}_1, \boldsymbol{x}_2), d(\boldsymbol{x}_1, \boldsymbol{x}_5)) = \min(1.29, 1.41) = 1.29.$$

   Step 2 . (cluster $\boxed{1\text{-}3}$ is formed) the updated distance matrix is

$$\begin{array}{cccc} & \boxed{1\text{-}3} & \boxed{2\text{-}5} & 4 \\ \boxed{1\text{-}3} & 0 & & \\ \boxed{2\text{-}5} & 1 & 0 & \\ 4 & \boxed{0.82} & 1 & 0 \end{array}.$$

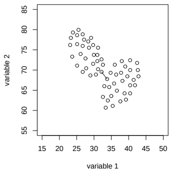
   Step 3 . (object 4 is joined to cluster $\boxed{1\text{-}3}$) the updated distance matrix is

$$\begin{array}{ccc} & \boxed{1\text{-}3\text{-}4} & \boxed{2\text{-}5} \\ \boxed{1\text{-}3\text{-}4} & 0 & \\ \boxed{2\text{-}5} & 1 & 0 \end{array}.$$

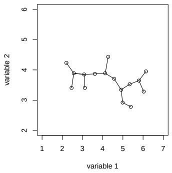
   Step 4 . (clusters $\boxed{1\text{-}3\text{-}4}$ and $\boxed{2\text{-}5}$ are joined).

   Single linkage has potential problem in two aspects:

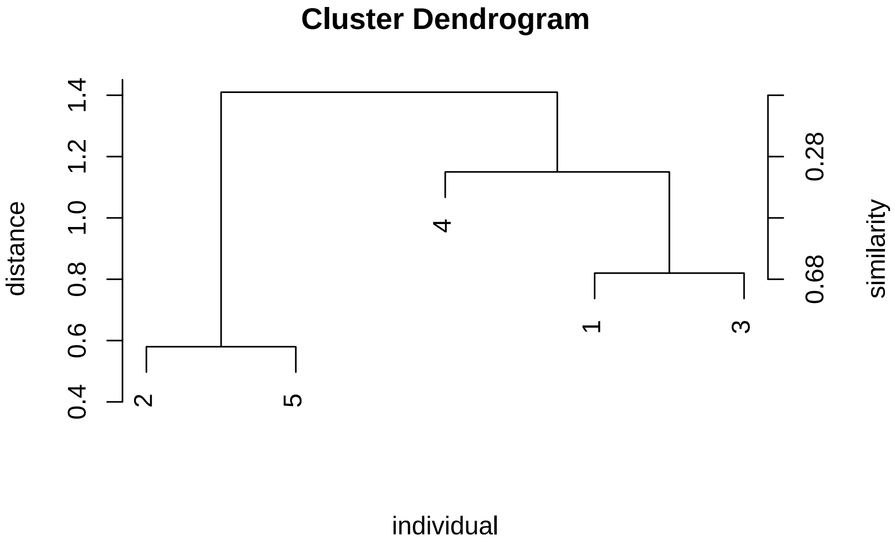
(a) (easily confused by two near overlapped clusters)



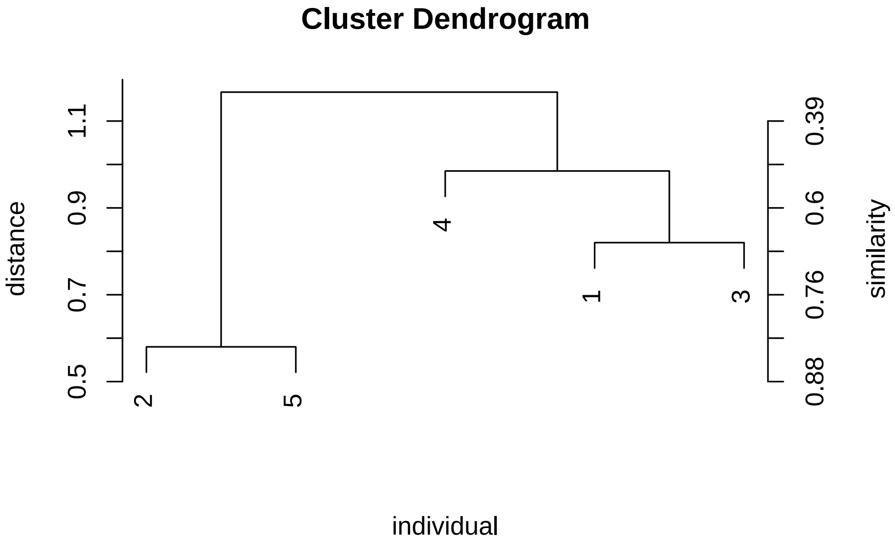
(b) (chaining effect)



2. complete linkage tends to find similar cluster which minimizes within cluster distance so that the resulting clusters are more 'compact'. For this reason, it is also called compact linkage. The between-cluster distance is based on maximum distance criterion. The dendrogram of the clustering result using complete linkage is given:



3. average linkage aims for clusters having characteristics somewhere between the single and complete linkage methods. The distance between two clusters is defined by the average of the distances pairing two objects each from a cluster. This method tends to combine clusters with small variances and tends to be biased toward the production of clusters with approximately the same variance. The dendrogram using average linkage is given:



4. centroid. The distance between two clusters is the distance between their centroids (means). Every time objects are grouped, a new centroid is computed. Cluster centroids migrate as cluster mergers take place. This method may produce confusing results that reversals may occur. That is, the distance between the centroids of a pair of clusters may be less than the distance between the controids of another pair merged at an earlier time. The advantage of

this method is that it is less affected by outliers than other hierarchical methods. It should be noted that this method is limited to qualitative data.

5. Ward's method. The distance between two clusters is the sum of squares between the two clusters summed over all variables. At each stage in the procedure, the within-cluster sum of squares is minimized over all partitions obtainable by combining two clusters from the previous stage. This method tends to combine clusters with a small number of objects and is biased toward the production of clusters with approximately the same number of objects.. Initially, $ESS_k = 0$ (Error Sum of Squares), $k = 1, \dots, N$ for the N objects. If there are currently K clusters, define $ESS = \sum_{k=1}^{K} ESS_k$, where

$$ESS_k = \sum_{i=1}^{c_k} d^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}}_k) = \sum_{i=1}^{c_k} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k)^t (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k),$$

$\overline{\boldsymbol{x}}_k$ is the centroid (mean) and c_k is the size of cluster k. At each step, union of every possible pair of clusters is considered and the two clusters whose combination produces smallest increase in ESS (minimum loss of information) are joined. Again, like centroid method, Ward's method requires quantitative data.

4.1.1 Example: Public Utility Data

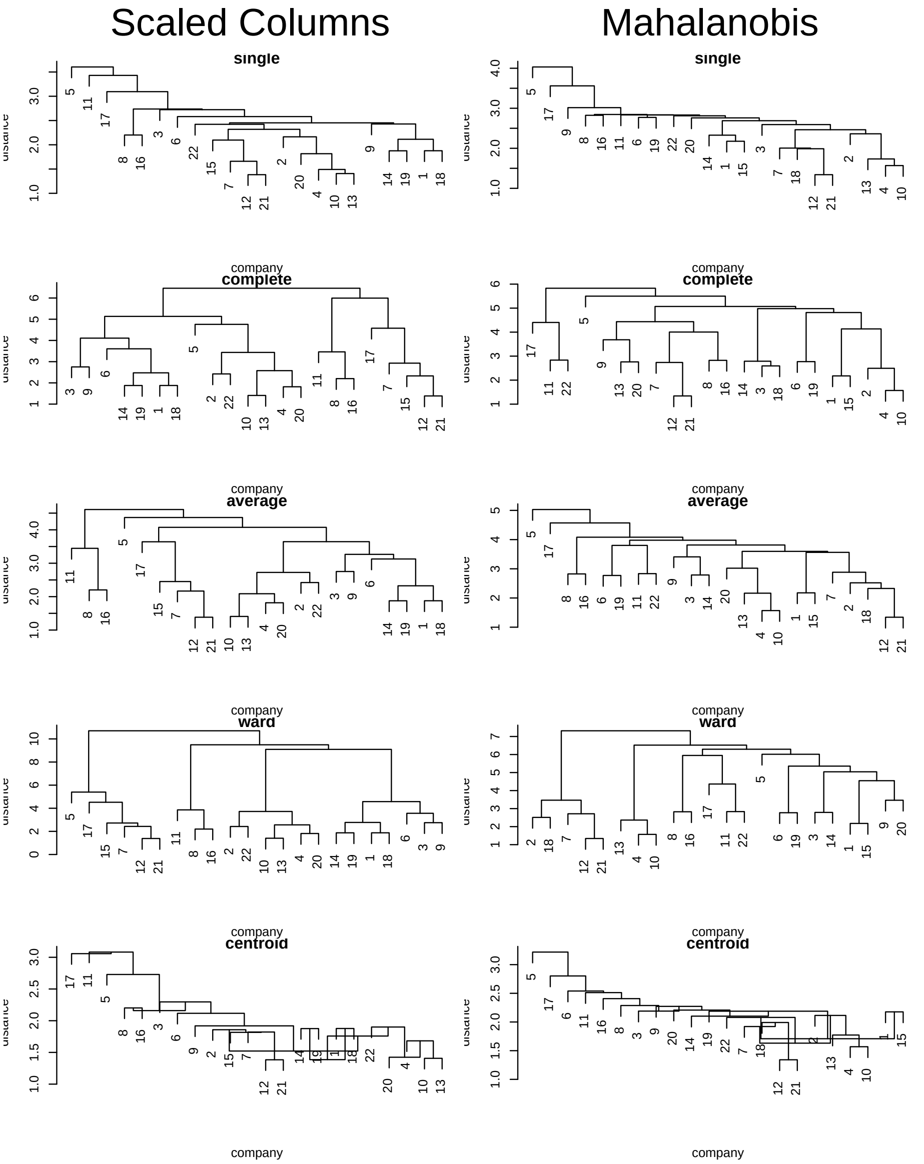
Public utility data (H. E. Thompson) on 22 US public utility componanies for the year 1975.

The correlation matrix is shown below:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.00	0.64	−0.10	−0.08	−0.26	−0.15	0.04	−0.01
X_2	0.64	1.00	−0.35	−0.09	−0.26	−0.01	0.21	−0.33
X_3	−0.10	−0.35	1.00	0.10	0.44	0.03	0.11	0.01
X_4	−0.08	−0.09	0.10	1.00	0.03	−0.29	−0.16	0.49
X_5	−0.26	−0.26	0.44	0.03	1.00	0.18	−0.02	−0.01
X_6	−0.15	−0.01	0.03	−0.29	0.18	1.00	−0.37	−0.56
X_7	0.04	0.21	0.11	−0.16	−0.02	−0.37	1.00	−0.19
X_8	−0.01	−0.33	0.01	0.49	−0.01	−0.56	−0.19	1.00

Note that the variables are incompatible and exhibit intercorrelation.

Hence it is desirable to perform clustering analysis on the standardized scale:



The five dendrograms on the left panel are the results on data in which columns are individually standardized before the Euclidean distance was applied. Those on the right are the results using Mahalanobis distance. Note that reversals occured when centroid method was employed. Disregard the two dendrograms produced by the centroid method. Note that the four methods produced fairly consistent results when standardization on individual columns was employed. It largely reveals that utility companies form natural clusters on similar locations (or types of locations): concentrating on intermediate-size clusters, companies 1, 14, 18, and 19 form a cluster; companies 7, 12, 15, and 21 form a cluster (coastal area companies); companies 4, 10, 13, and 20 form another cluster. Companies 5 and 17 stand by themselves until the final stages. The results produced by employing Mahalanobis distance lack consistent patterns across different method. The moderate correlations (for instance, correlation between X_1 and X_2 is 0.64) may be spurious due to the existence of several natural clusters.

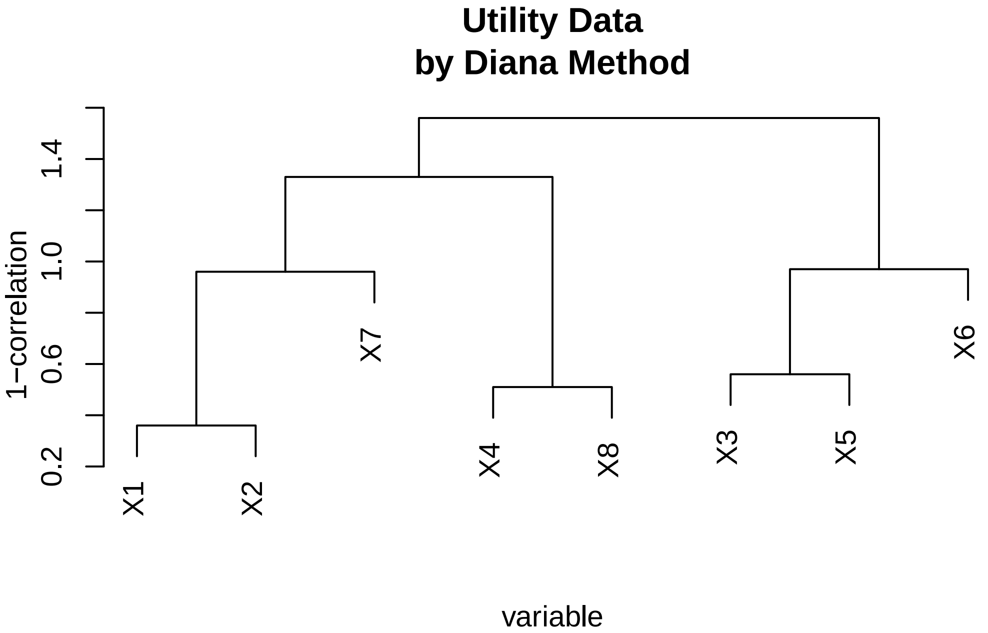
4.2 Divisive Clustering Procedure—Diana

A divisive clustering procedure, named *DiAna* (short for *Di*visive *Ana*lysis Clustering), was developed by Kaufman, L. and Rousseeuw, P.J. (*Finding Groups in Data: An Introduction to Cluster Analysis*, 1990, Wiley, New York.) The algorithm is as follows:

- It starts with a single cluster, the entire set of n objects.
- In each step, the cluster with largest diameter is selected and is to be divided (splintered) into two clusters. Here the *diameter* of a cluster is defined as the maximum distance or dissimilarity (i.e., minimum similarity) among all objects within the cluster. An object within this cluster having largest average dissimilarity to other objects within the cluster is identified. This object initiates '*splinter group*.' An object within this cluster is reassigned to the splinter group if it is closer to the splinter group than to the 'old party.' Consequently, at the end of the step, the cluster is divided into two new clusters.
- The above step is repeated until n clusters are formed.

4.2.1 Utility Example Revisited

The dendrogram produced by the method Diana is given below:



4.3 Comments on the hierarchical procedures

- Most clustering methods are sensitive to outliers due to the fact that sources of error and variation are not formally considered in most methods.

- No provision for a reallocation of "incorrectly" grouped objects at earlier stages. So it is advisable to examine the final configuration of clusters.

- It is a good idea to try several cluster methods and similarity (or distance) measures. If the outcomes from the several methods are (roughly) consistent with one another, perhaps a case for "natural" groupings can be advanced.

- It is advisable to check for the *stability* of a hierarchical solution by applying the clustering algorithm before and after small perturbations (errors) have been added to the data units. If the clusters are fairly well separated, the clusterings before and after perturbation should agree.

- Common values (i.e., ties) in the similarity (or distance) matrix can produce multiple solutions to a hierarchical clustering problem (differential results occur particularly at the lower levels due to different treatments of the tied similarities/distances). The users should know that this is not an inherent problem of any method.

5 Nonhierarchical Clustering Procedures

Contrasting to hierarchical procedures, nonhierarchical procedures do not involve the treelike construction process. Instead, its first step is to select a cluster center (or seed), and all objects within a prespecified threshold distance are included in the resulting cluster. Nonhierarchical clustering procedures are frequently referred to as K-means clustering. There are three different nonhierarchical procedures:

1. Sequential threshold procedure starts by selecting one cluster seed, and includes all objects within a prespecified distance. A second cluster seed (beyond the first cluster) is then selected, and all objects within the prespecified distance are included. Then a third seed is selected, and the process continues as before. An object is no longer considered for subsequent seeds if it is already included in a cluster.

2. Parallel threshold procedure selects several cluster seeds simultaneously in the beginning, and objects within the threshold distance are assigned to the nearest seed. As the process evolves, threshold distances can be adjusted to include fewer or more objects in the clusters. Also, in some methods, objects remain unclustered if they are beyond the prespecified distance from any cluster seed.

3. Optimizing procedure is similar to the other two except that it allows for reassignment of objects to another cluster from the original on the basis of some overall optimizing criterion.

The number of clusters, K, may either be specified in advance or determined as part of the procedure. A matrix of distances (or similarities) does not have to be determined, and the basic data do not have to be stored during the computational run, nonhierarchical procedures can therefore be applied much larger data sets than can hierarchical procedures.

Example: Suppose we measure two variables X_1 and X_2 for each of four objects A, B, C, and D with the following data:

	variable	
Object	X_1	X_2
A	5	4
B	1	-2
C	-1	1
D	3	1

The objective is to divide these objects into $K = 2$ clusters such that the objects within a cluster are closer to one another than they are to the objects in different clusters. Suppose an initial *arbitrary* partition groups (AB) and (CD). The clustering process is given as follows:

Step 1.

		distance^2 from cluster centroid for			
cluster	centroid	A	B	C	D
(AB)	$(\frac{5+1}{2}, \frac{4+(-2)}{2}) = (3, 1)$	13	13	16	0
(CD)	$(\frac{-1+3}{2}, \frac{1+1}{2}) = (1, 1)$	25	9	4	4

Note that objects B and D need to be reassigned. The algorithm will take two steps to reassign these objects. However, we present here the combined step:

Step 2.

		distance^2 from cluster centroid for			
cluster	centroid	A	B	C	D
(AD)	$(\frac{5+3}{2}, \frac{4+1}{2}) = (4, 2.5)$	3.25	29.25	27.25	3.25
(BC)	$(\frac{1+(-1)}{2}, \frac{(-2)+1}{2}) = (0, -0.5)$	45.25	3.25	3.25	11.25

Now, all objects are at their respective clusters. To check the stability of the clustering, it is advisable to rerun the algorithm with a new partition (and hence new intial cluster seeds).

5.1 Example: Utility data (revisited)

K-mean clustering method is used, columns are individually standardized.
Suppose we started with $K = 4$ and chose randomly companies 1, 6, 17, and 19 as initial cluster seeds. The results are:

clusters	standardized variables								
	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	members (companies)
1	−0.01	0.33	0.22	−0.36	0.17	−0.40	1.57	−0.60	4, 10, 13, 20, 22
2	−0.24	−0.66	0.26	0.80	−0.05	−0.86	−0.29	1.25	2, 5, 7, 12, 15, 17, 21
3	0.50	0.78	−0.99	−0.34	−0.49	0.35	−0.52	−0.41	1, 3, 6, 9, 14, 18, 19
4	−0.60	−0.83	1.34	−0.48	0.99	1.86	−0.71	−0.97	8, 11, 16

with the distance matrix for the cluster centers given by

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \left[\begin{array}{cccc} 0 & & & \\ 3.08 & 0 & & \\ 2.70 & 3.15 & 0 & \\ 3.75 & 4.07 & 3.75 & 0 \end{array}\right] \end{array}.$$

5.2 Comments on using nonhierarchical procedures

In general, the choice of a particular K is not clear cut and depends on subject matter under study, as well as data-based appraisals. Moreover, there are strong arguments for not fixing K in advance:

- When two or more cluster seeds inadvertantly lie within a single cluster, their resulting clusters will be poorly differentiated.
- If there exist outliers, then the algorithm might produce at least one cluster with very disperse objects within it.
- Even if it is known to have K groups in advance, the sampling method may cause the rarest group not appear in the sample. Consequently, forcing the data (sample) into K clusters would lead to nonsensical groups.

6 Deciding number of clusters

A major issue with all clustering methods is how to determine the number of clusters. Although many criteria and guidelines for approaching this problem exist, it lacks standard, objective selection procedure. The distances between clusters at successive steps may serve as a useful guideline, and the practioner may choose to stop when this distance exceeds a pre-determined value or when the successive distances between steps make a sudden jump. Also, some intuitive conceptual or theoretical relationship may suggest a natural number of clusters. In the final analysis, however, it is probably best to compute solutions for several different number of clusters (e.g., two, three, four) and then decide among the alternative solutions based upon *a priori* criteria, practical judgement, common sense, or theoretical foundations. A statistical procedure exists for deciding the number of clusters—namely *Beale's Pseudo F test*:

Consider clustering N objects. Suppose there are two clusterings with c_1, and c_2 clusters, respectively, and $c_1 > c_2$. The within-cluster sums of squares are computed by

$$W_r = \sum_{i=1}^{c_r} \sum_{j=1}^{n_i} d^2(\boldsymbol{x}_{ij}, \overline{\boldsymbol{x}}_i), \ r = 1, 2,$$

where $\overline{\boldsymbol{x}}_i$ is the ith cluster centroid under respective clustering and $\boldsymbol{x}_{ij}$ is the jth object within ith cluster (with cluster size n_i). If W_1 and W_2 are nearly the same, then the clustering with fewer number of clusters is just as good as that with larger number of clusters. For simplicity, one would then select the clustering with fewer number of clusters. If W_1 is much smaller than W_2, then the first clustering is an improvement over the second, and one would select the clustering with larger number of clusters. A pseudo F test is facilitated by

$$F^* = \frac{W_2 - W_1}{W_1} \cdot \frac{(N - c_1)k_1}{(N - c_2)k_2 - (N - c_1)k_1}, \quad \text{where } k_r = c_r^{-2/p}.$$

It has an approximate F distribution with $(N - c_2)k_2 - (N - c_1)k_1$ and $(N - c_1)k_1$ degrees of freedom. If F^* is greater than an F critical value, then we would choose the first clustering (the one with more clusters) over the second (the one with fewer clusters).

7 Stage Two: Interpretation

When starting the interpretation process, one measure that is used frequently is the cluster's centroid which provides a logical description when the clustering procedure was performed on the raw data. However, if the data were standardized or if the clustering was performed on principal components or factor components (some sorts of transformation being applied on the data), the practitioner would have to go back to the raw scores for the original variables and compute average profiles within clusters using these data.

8 Stage Three: Validation and Profiling

Validation includes the attempts by the practitioner to assure that the clustering solution is representative of the general population and is generalizable to other objects (in the population) and stable over time. One approach is to split the sample into two groups. Clusterings then are applied on these groups and results are compared. The profiling stage involves describing the characteristics of each cluster in order to explain how they may differ on relevant dimensions. The procedure begins with the clusters identified in stage two. The practitioner utilizes measurements/characteristics not previously included in the clustering procedure to profile the characteristics of each cluster.