PEYTON VINES

STA4702

CHAPTER 12.1-12.4

CLUSTERING, DISTANCE METHODS, NONHIERARCHICAL CLUSTERING METHODS, AND SIMILARITY MEASURES

Grouping, or clustering, is distinct from the classification methods discussed in the previous chapter. Classification pertains to a *known* number of groups, and the operational objective is to assign new observations to one of these groups. Cluster analysis is a more primitive technique in that no assumptions are made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities). The inputs required are similarity measures or data from which similarities can be computed.

The basic objective in cluster analysis is to discover natural groupings of the items (or variables). In turn, we must first develop a quantitative scale on which to measure the association (similarity) between objects.

[1] The number of ways of sorting n objects into k nonempty groups is a Stirling number of the second kind given by $(1/k!)\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$. (See [1].) Adding these numbers for $k = 1, 2, \ldots, n$ groups, we obtain the total number of possible ways to sort n objects into groups.

Ex: sorting cards based on suit (hearts, spades, clubs, and diamonds), sorting cards based on whether they are face cards or not

SIMILARITY MEASURES

Most efforts to produce a rather simple group structure from a complex data set require a measure of "closeness," or "similarity.

Important considerations include the nature of the variables (discrete, continuous, binary), scales of measurement (nominal, ordinal, interval, ratio), and subject matter knowledge.

When *items* (units or cases) are clustered, proximity is usually indicated by some sort of distance. By contrast, *variables* are usually grouped on the basis of correlation coefficients or like measures of association.

Distances and Similarity Coefficients for Pairs of Items

We discussed the notion of distance in Chapter 1, Section 1.5. Recall that the Euclidean (straight-line) distance between two p-dimensional observations (items) $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$ and $\mathbf{y}' = [y_1, y_2, \ldots, y_p]$ is, from (1-12),

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \\ &= \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})} \end{aligned} \tag{12-1}$$

The statistical distance between the same two observations is of the form [see (1-23)]

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'\mathbf{A}(\mathbf{x} - \mathbf{y})} \tag{12-2}$$

Ordinarily, $\mathbf{A} = \mathbf{S}^{-1}$, where $\mathbf{S}$ contains the sample variances and covariances. However, without prior knowledge of the distinct groups, these sample quantities cannot be computed. For this reason, Euclidean distance is often preferred for clustering.

Another distance measure is the Minkowski metric

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m} \tag{12-3}$$

For $m = 1$, $d(\mathbf{x}, \mathbf{y})$ measures the "city-block" distance between two points in p dimensions. For $m = 2$, $d(\mathbf{x}, \mathbf{y})$ becomes the Euclidean distance. In general, varying m changes the weight given to larger and smaller differences.

DISTANCE MEASURE FOR NONNEGATIVE VARIABLES

Canberra metric: $$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}$$

Czekanowski coefficient: $$d(\mathbf{x}, \mathbf{y}) = 1 - \frac{2 \sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i)}$$

When items cannot be represented by meaningful p-dimensional measurements, pairs of items are often compared on the basis of the presence or absence of certain characteristics.

The presence or absence of a characteristic can be described mathematically by introducing a *binary variable*, which assumes the value 1 if the characteristic is present and the value 0 if the characteristic is absent. For $p = 5$ binary variables, for instance, the "scores" for two items i and *k* might be arranged as follows:

	Variables				
	1	2	3	4	5
Item *i*	1	0	0	1	1
Item *k*	1	1	0	1	0

In this case, there are two 1–1 matches, one 0–0 match, and two mismatches.

Let x_{ij} be the score (1 or 0) of the *j*th binary variable on the *i*th item and x_{kj} be the score (again, 1 or 0) of the *j*th variable on the *k*th item, $j = 1, 2, \ldots, p$. Consequently,

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & \text{if } x_{ij} = x_{kj} = 1 \quad \text{or} \quad x_{ij} = x_{kj} = 0 \\ 1 & \text{if } x_{ij} \neq x_{kj} \end{cases} \tag{12-6}$$

and the squared Euclidean distance, $\sum_{j=1}^{p} (x_{ij} - x_{kj})^2$, provides a count of the number of mismatches. A large distance corresponds to many mismatches—that is, dissimilar items. From the preceding display, the square of the distance between items *i* and *k* would be

$$\sum_{j=1}^{5} (x_{ij} - x_{kj})^2 = (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2$$
$$= 2$$

In some cases, a 1-1 match is a stronger indication of similarity than a 0-0 match. To allow for the differential treatment of the 1-1 matches and the 0-0 matches, several schemes for defining similarity coefficients have been suggested.

		Item k 1	Item k 0	Totals
Item i	1	a	b	$a + b$
	0	c	d	$c + d$
Totals		$a + c$	$b + d$	$p = a + b + c + d$

(12-7)

In this table, a represents the frequency of 1–1 matches, b is the frequency of 1–0 matches, and so forth. Given the foregoing five pairs of binary outcomes, $a = 2$ and $b = c = d = 1$.

Monotonicity is important, because some clustering procedures are not affected if the definition of similarity is changed in a manner that leaves the relative orderings of similarities unchanged. The single linkage and complete linkage hierarchical procedures discussed in section 12.3 are not affected. For these methods, any choice of the coefficients 1, 2, and 3 in table 12.1 will produce the same groupings. Similarly, any choice of the coefficients 5, 6, and 7 will yield identical groupings.

Table 12.1 Similarity Coefficients for Clustering Items*

Coefficient	Rationale
1. $\dfrac{a+d}{p}$	Equal weights for 1–1 matches and 0–0 matches.
2. $\dfrac{2(a+d)}{2(a+d)+b+c}$	Double weight for 1–1 matches and 0–0 matches.
3. $\dfrac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4. $\dfrac{a}{p}$	No 0–0 matches in numerator.
5. $\dfrac{a}{a+b+c}$	No 0–0 matches in numerator or denominator. (The 0–0 matches are treated as irrelevant.)
6. $\dfrac{2a}{2a+b+c}$	No 0–0 matches in numerator or denominator. Double weight for 1–1 matches.
7. $\dfrac{a}{a+2(b+c)}$	No 0–0 matches in numerator or denominator. Double weight for unmatched pairs.
8. $\dfrac{a}{b+c}$	Ratio of matches to mismatches with 0–0 matches excluded.

*[p binary variables; see (12-7).]

Example 12.1 (Calculating the values of a similarity coefficient) Suppose five individuals possess the following characteristics:

	Height	Weight	Eye color	Hair color	Handedness	Gender
Individual 1	68 in	140 lb	green	blond	right	female
Individual 2	73 in	185 lb	brown	brown	right	male
Individual 3	67 in	165 lb	blue	blond	right	male
Individual 4	64 in	120 lb	brown	brown	right	female
Individual 5	76 in	210 lb	brown	brown	left	male

Define six binary variables $X_1, X_2, X_3, X_4, X_5, X_6$ as

$$X_1 = \begin{cases} 1 & \text{height} \geq 72 \text{ in.} \\ 0 & \text{height} < 72 \text{ in.} \end{cases} \qquad X_4 = \begin{cases} 1 & \text{blond hair} \\ 0 & \text{not blond hair} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{weight} \geq 150 \text{ lb} \\ 0 & \text{weight} < 150 \text{ lb} \end{cases} \qquad X_5 = \begin{cases} 1 & \text{right handed} \\ 0 & \text{left handed} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{brown eyes} \\ 0 & \text{otherwise} \end{cases} \qquad X_6 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

The scores for individuals 1 and 2 on the $p = 6$ binary variables are

		X_1	X_2	X_3	X_4	X_5	X_6
Individual	1	0	0	0	1	1	1
	2	1	1	1	0	1	0

and the number of matches and mismatches are indicated in the two-way array

		Individual 2: 1	0	Total
Individual 1	1	1	2	3
	0	3	0	3
	Totals	4	2	6

Employing similarity coefficient 1, which gives equal weight to matches, we compute

$$\frac{a+d}{p} = \frac{1+0}{6} = \frac{1}{6}$$

Continuing with similarity coefficient 1, we calculate the remaining similarity numbers for pairs of individuals. These are displayed in the 5×5 symmetric matrix

$$\begin{array}{cc} & \text{Individual} \\ & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \text{Individual} \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 1 & & & & \\ \frac{1}{6} & 1 & & & \\ \frac{4}{6} & \frac{3}{6} & 1 & & \\ \frac{4}{6} & \frac{3}{6} & \frac{2}{6} & 1 & \\ 0 & \frac{5}{6} & \frac{2}{6} & \frac{2}{6} & 1 \end{bmatrix} \end{array}$$

Based on the magnitudes of the similarity coefficient, we conclude that individuals 2 and 5 are most similar while individuals 1 and 5 are least similar.

When the variables are binary, the data can again be arranged in the form of a contingency table. This time, however, the variables, rather than the items, delineate the categories. For each pair of variables, there are n items categorized in the table. With the usual 0 and 1 coding, the table becomes as follows:

		Variable k 1	0	Totals
Variable i	1	a	b	$a + b$
	0	c	d	$c + d$
	Totals	$a + c$	$b + d$	$n = a + b + c + d$

(12-10)

For instance, variable i equals 1 and variable k equals 0 for b of the n items.

The usual product moment correlation formula applied to the binary variables in the contingency table of (12-10) gives (see Exercise 12.3)

$$r = \frac{ad - bc}{[(a + b)(c + d)(a + c)(b + d)]^{1/2}} \tag{12-11}$$

This number can be taken as a measure of the similarity between the two variables.

Table 12.2 Numerals in 11 Languages

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neljä
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

Table 12.3 Concordant First Letters for Numbers in 11 Languages

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du	3	5	4	10							
G	4	6	5	5	10						
Fr	4	4	4	1	3	10					
Sp	4	4	5	1	3	8	10				
I	4	4	5	1	3	9	9	10			
P	3	3	4	0	2	5	7	6	10		
H	1	2	2	2	1	0	0	0	0	10	
Fi	1	1	1	1	1	1	1	1	1	2	10

The words for 1 in French, Spanish, and Italian all begin with *u*. For illustrative purposes, we might compare languages by looking at the *first letters* of the numbers. We call the words for the same number in two different languages *concordant* if they have the same first letter and *discordant* if they do not. From Table 12.2, the table of concordances (frequencies of matching first initials) for the numbers 1–10 is given in Table 12.3. We see that English and Norwegian have the same first letter for 8 of the 10 word pairs. The remaining frequencies were calculated in the same manner.

The results in Table 12.3 confirm our initial visual impression of Table 12.2. That is, English, Norwegian, Danish, Dutch, and German seem to form a group. French, Spanish, Italian, and Polish might be grouped together, whereas Hungarian and Finnish appear to stand alone. ■

12.3 Hierarchical Clustering Methods

Agglomerative hierarchical methods are proceeded by a series of successive mergers. Initially regard each object as a cluster, the most similar objects are first grouped, continually merge until all of the subgroups are fused into a single cluster.

Divisive hierarchical methods work in the opposite direction. They are proceeded by a series of successive divisions. An initial single group of all objects is divided into 2 subgroups such that the objects in one subgroup are “far from” the objects in the other. Continually divide until each object forms a group.

Agglomerative Hierarchical Clustering for Grouping N Objects.

1. Start with *N* clusters, each containing a single entity and an $N \times N$ symmetric matrix of distances (or similarities) $\boldsymbol{D} = \{d_{ik}\}$.
2. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters ***U*** and ***V*** be d_{UV}.
3. Merge clusters ***U*** and ***V***. Label the newly formed cluster (***UV***). 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 cluster (***UV***) and the remaining clusters.
4. Repeat Steps 2 and 3 a total of *N-1* times. (All objects will be in a single cluster after the algorithm terminates.) Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

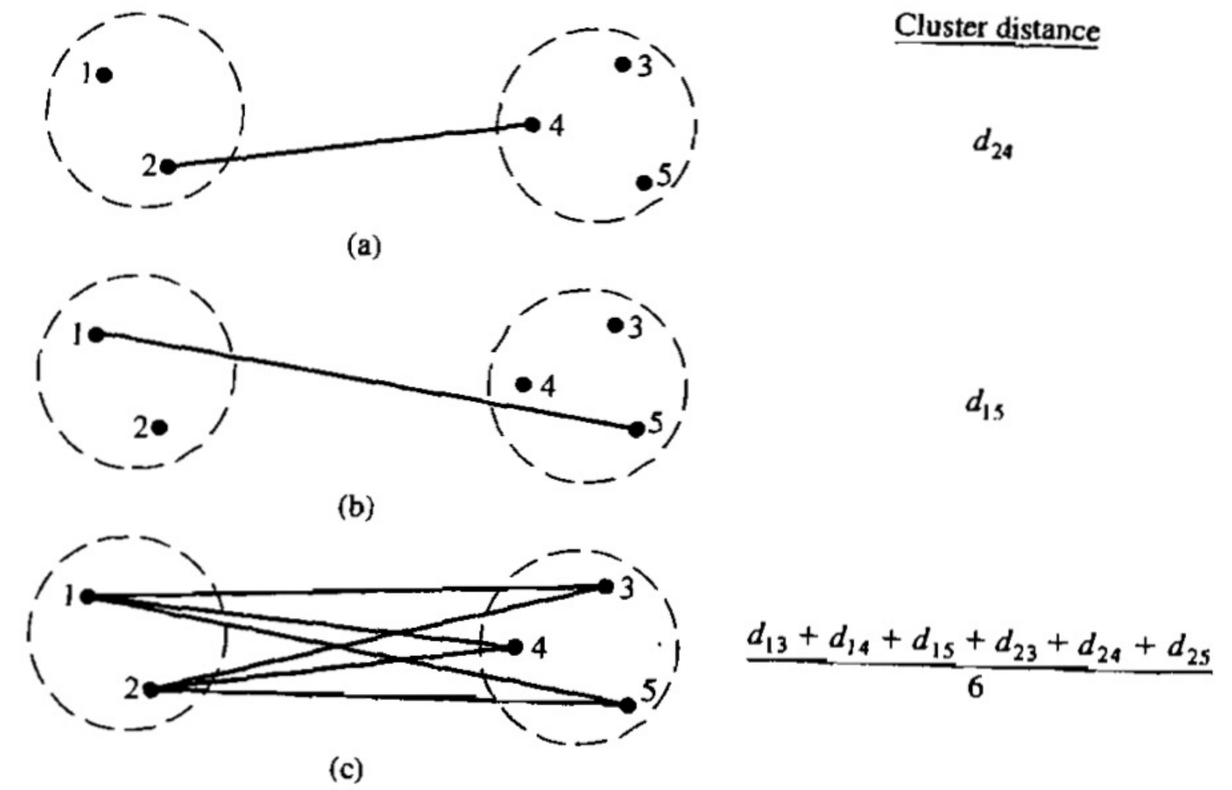


Figure 12.2 Intercluster distance (dissimilarity) for (a) single linkage, (b) complete linkage, and (c) average linkage.

From the figure, we see that single linkage results when groups are fused according to the distance between their nearest members.

Complete linkage occurs when groups are fused according to the distance between their farthest members.

For average linkage, groups are fused according to the average distance between pairs of members in the respective sets.

SINGLE LINKAGE $d_{(UV)W} = \min\{d_{UW}, d_{VW}\}$

Example 12.3 (Clustering using single linkage) To illustrate the single linkage algorithm, we consider the hypothetical distances between pairs of five objects as follows:

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array}$$

The nearest neighbor distances are

$$d_{(35)1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3$$
$$d_{(35)2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7$$
$$d_{(35)4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8$$

New distance matrix

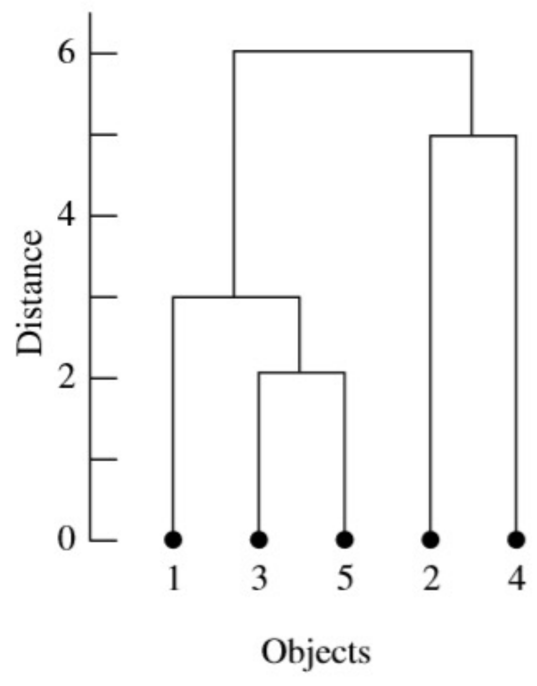
$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

Now repeat until all of the smallest pairs have been clustered.

$$d_{(135)2} = \min\{d_{(35)2}, d_{12}\} = \min\{7, 9\} = 7$$
$$d_{(135)4} = \min\{d_{(35)4}, d_{14}\} = \min\{8, 6\} = 6$$

$$\begin{array}{c} \\ (135) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

$$\begin{array}{c} \\ (135) \\ (24) \end{array} \begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{array}$$



Single linkage dendrogram

COMPLETE LINKAGE (most distant) $d_{(UV)W} = max\{d_{UW}, d_{VW}\}$

Example 12.5 (Clustering using complete linkage) Let us return to the distance matrix introduced in Example 12.3:

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \left[\begin{array}{ccccc} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{array}\right] \end{array}$$

$$d_{(35)1} = \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11$$
$$d_{(35)2} = \max\{d_{32}, d_{52}\} = 10$$
$$d_{(35)4} = \max\{d_{34}, d_{54}\} = 9$$

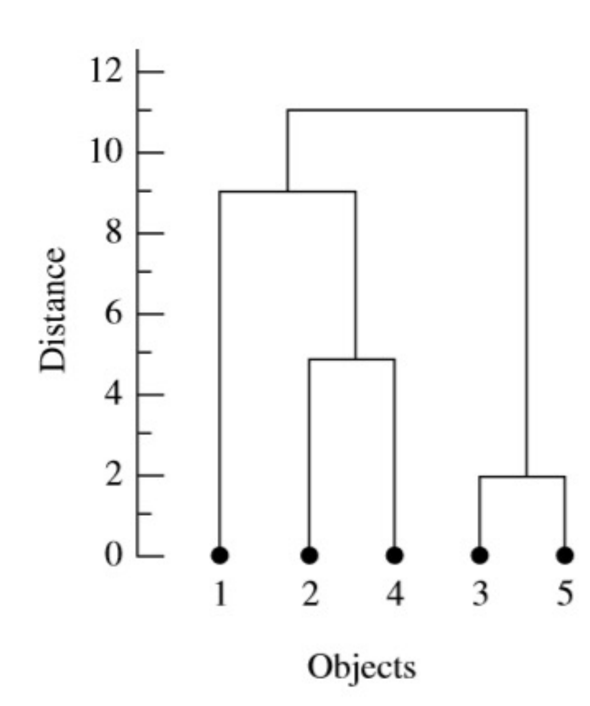
Modified distance matrix

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \left[\begin{array}{cccc} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & 5 & 0 \end{array}\right] \end{array}$$

Continue to the next merger

$$d_{(24)(35)} = \max\{d_{2(35)}, d_{4(35)}\} = \max\{10, 9\} = 10$$
$$d_{(24)1} = \max\{d_{21}, d_{41}\} = 9$$

$$\begin{array}{c} \\ (35) \\ (24) \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} (35) & (24) & 1 \end{array} \\ \left[\begin{array}{ccc} 0 & & \\ 10 & 0 & \\ 11 & 9 & 0 \end{array}\right] \end{array}$$



Complete linkage dendrogram for distances between five objects.

AVERAGE LINKAGE $d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$

Average linkage treats the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster.

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)	Romanian (Ro)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi	unu
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi	doi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme	trei
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	nelja	patru
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi	cinci
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi	sase
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman	sapte
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan	opt
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan	noua
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen	zece

Measure the similarities of 12 languages

Concordant First Letters for Numbers in 12 European Languages

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi	Ro
E	10											
N	8	10										
Da	8	9	10									
Du	3	5	4	10								
G	4	6	5	5	10							
Fr	4	4	4	1	3	10						
Sp	4	4	5	1	3	8	10					
I	4	4	5	1	3	9	9	10				
P	3	3	4	0	2	5	7	6	10			
H	1	2	2	2	1	0	0	0	0	10		
Fi	1	1	1	1	1	1	1	1	1	2	10	
Ro	4	4	5	1	4	7	8	8	5	0	1	10

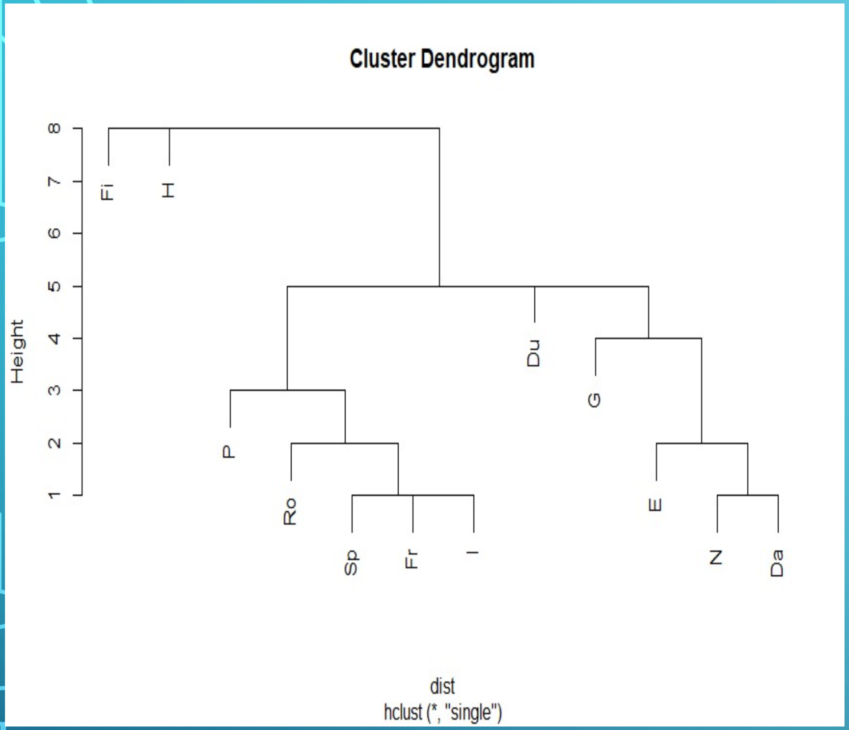
Matrix of distances:
Subtract the concordances from the perfect agreement figure of 10

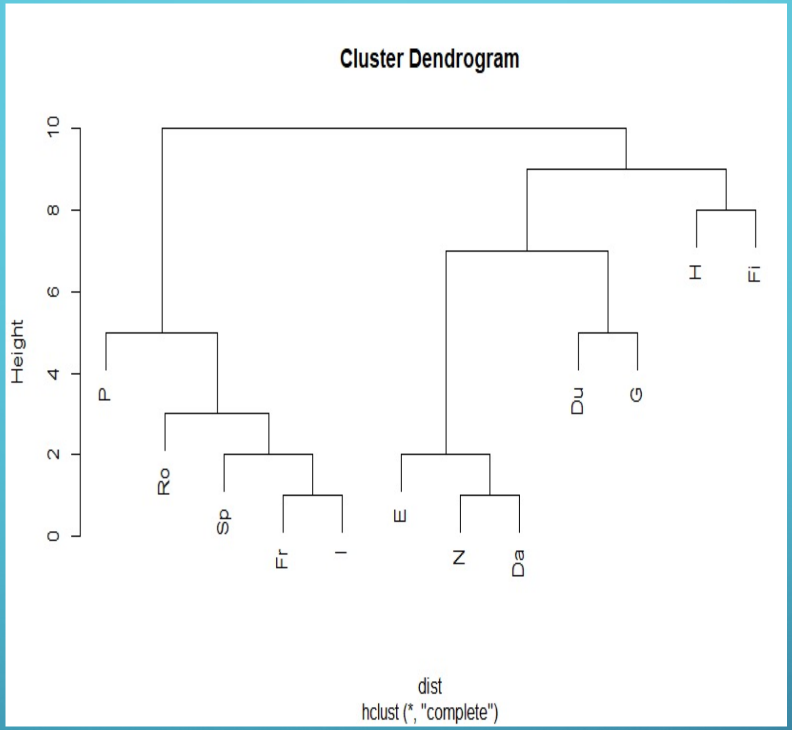
R software

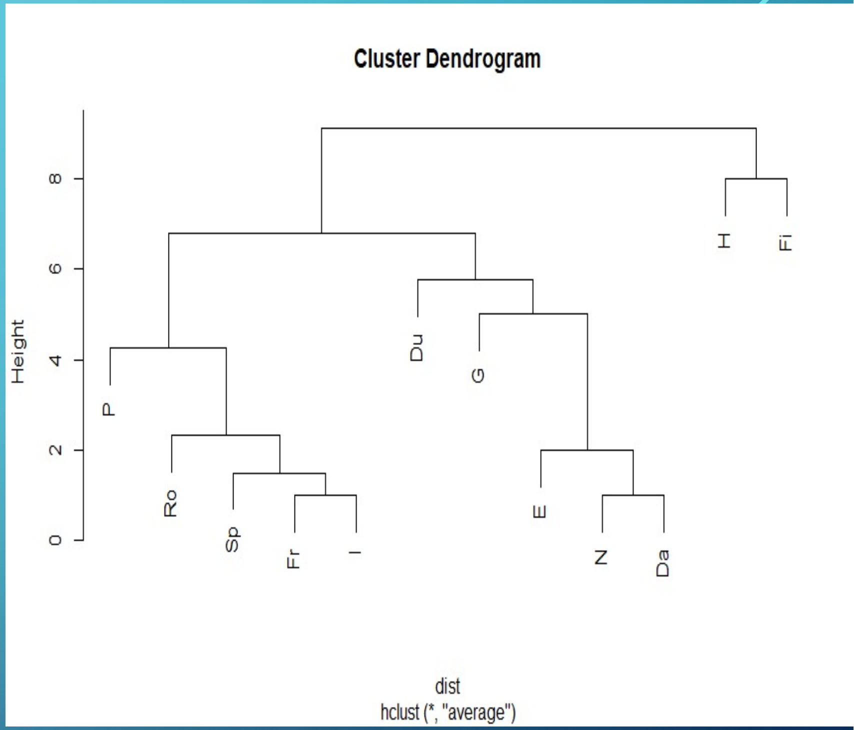
```
> dist
```

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi	Ro
E	0											
N	2	0										
Da	2	1	0									
Du	7	5	6	0								
G	6	4	5	5	0							
Fr	6	6	6	9	7	0						
Sp	6	6	5	9	7	2	0					
I	6	6	5	9	7	1	1	0				
P	7	7	6	10	8	5	3	4	0			
H	9	8	8	8	9	10	10	10	10	0		
Fi	9	9	9	9	9	9	9	9	9	8	0	
Ro	6	6	5	9	6	3	2	2	5	10	9	0

```
> fit1<-hclust(dist,method="single")
> plot(fit1)
> fit2<-hclust(dist,method="complete")
> plot(fit2)
> fit3<-hclust(dist,method="average")
> plot(fit3)
```







From these we can see that French, Italian, Spanish, and Romanian are very similar, and that Finish and Hungarian are both quite different from other languages.

Average linkage leads to a dendrogram that is much like the single linkage configuration.

Ward's Hierarchical Clustering Method

- Ward considered hierarchical clustering procedures based on minimizing the 'loss of information' from joining two groups. This method is usually implemented with loss of information taken to be an increase in an error sum of squares criterion, ESS.

Use Ward's method to cluster the four items whose measurements on a single variable X are given in the following table.

Item	Measurements x
1	2
2	1
3	5
4	8

(a) Initially, each item is a cluster and we have the clusters

$$\{1\} \quad \{2\} \quad \{3\} \quad \{4\}$$

Show that ESS = 0, as it must.

(b) If we join clusters $\{1\}$ and $\{2\}$, the new cluster $\{12\}$ has

$$\text{ESS}_1 = \sum (x_j - \bar{x})^2 = (2 - 1.5)^2 + (1 - 1.5)^2 = .5$$

and the ESS associated with the grouping $\{12\}$, $\{3\}$, $\{4\}$ is ESS = .5 + 0 + 0 = .5. The *increase* in ESS (loss of information) from the first step to the current step in .5 − 0 = .5. Complete the following table by determining the increase in ESS for all the possibilities at step 2.

Clusters			Increase in ESS
{12}	{3}	{4}	.5
{13}	{2}	{4}	
{14}	{2}	{3}	
{1}	{23}	{4}	
{1}	{24}	{3}	
{1}	{2}	{34}	

(c) Complete the last two algamation steps, and construct the dendrogram showing the values of ESS at which the mergers take place.

$$\text{ESS} = \sum_{j=1}^{N} (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})$$

(a) $\text{ESS}_1 = (2-2)^2 = 0$, $\text{ESS}_2 = (1-1)^2 = 0$, $\text{ESS}_3 = (5-5)^2 = 0$, and $\text{ESS}_4 = (8-8)^2 = 0$.

(b) At step 2

Clusters			Increase in ESS
{12}	{3}	{4}	.5
{13}	{2}	{4}	4.5
{14}	{2}	{3}	18.0
{1}	{23}	{4}	8.0
{1}	{24}	{3}	24.5
{1}	{2}	{34}	4.5

(c) At step 3

Clusters		Increase in ESS
{12}	{34}	5.0
{123}	{4}	8.7

Finally all four together have

$$\text{ESS} = (2-4)^2 + (1-4)^2 + (5-4)^2 + (8-4)^2 = 30$$

Thank You!