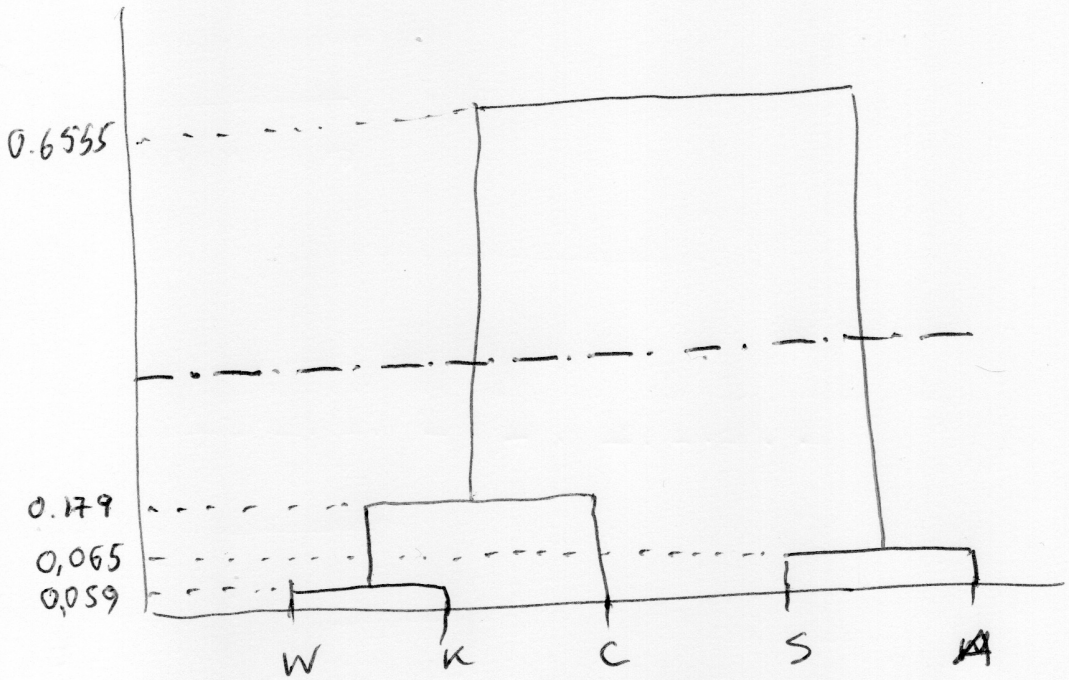
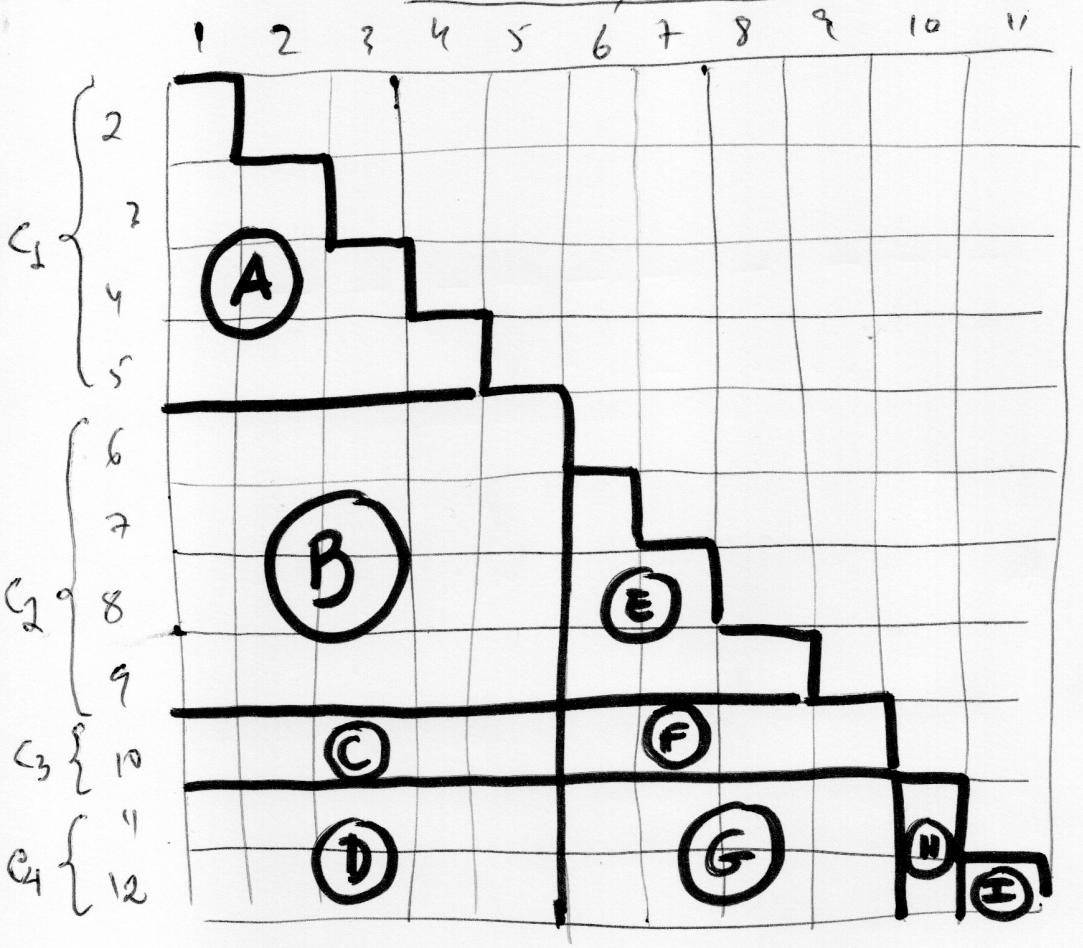


The dendrogram clearly suggests two groups of clustering methods: {Ward, k-means, complete} and {single, average}

The fact that both methods Ward and k-means share the same optimization criterion (minimizing the intra-cluster variance) explains why they are the most similar methods. These two methods tend two form compact and ~~spherical~~ rounded shape clusters, which are also characteristics of the complete method. On the other cluster we have the single and average linkage methods that can produce unbalanced clusters that are not very homogeneous (particularly the single linkage method).

9. Dissimilarity matrix



- Diam $(C_1) = \max_{x,y \in C_1} d(x,y) =$ largest value $= 0.189$ in (A)

- Diam $(C_2) = \max_{x,y \in C_2} d(x,y) =$ largest value $= 0.124$ in (E)

- Diam $(C_3) =$ ~~MAX~~ 0 since $C_3 = \{10\}$!

- Diam $(C_4) = d_{11,12} = 0.252$ (I)

Partition diameter $= \max \{$ Diam (C_1), Diam (C_2), diam (C_3), Diam $(C_4) \} =$

$= \max \{0, 189, 0.124, 0, 0.252\} = 0.252$ //

b) By definition of the complete-linkage method

- $D(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x,y) = 0.268$ (largest value of (B)) B
- $D(C_1, C_3) = \max_{x \in C_1, y \in C_3} d(x,y) = 0.284$ (largest value of (E)) C
- $D(C_1, C_4) = \max_{x \in C_1, y \in C_4} d(x,y) = 0.322$ (largest value of (A)) D
- $D(C_2, C_3) = \max_{x \in C_2, y \in C_3} d(x,y) = 0.293$ (largest value of (F)) F
- $D(C_2, C_4) = \max_{x \in C_2, y \in C_4} d(x,y) = 0.319$ (largest value of (G)) G
- $D(C_3, C_4) = \max_{x \in C_3, y \in C_4} d(x,y) = 0.320$ (largest value of (H)) H

	C_1	C_2	C_3
C_2	0.268		
C_3	0.284	0.293	
C_4	0.322	0.319	0.320

smallest value

⇒

	$C_1 \cup C_2$	C_3
C_3	0.293	
C_4	0.322	0.320

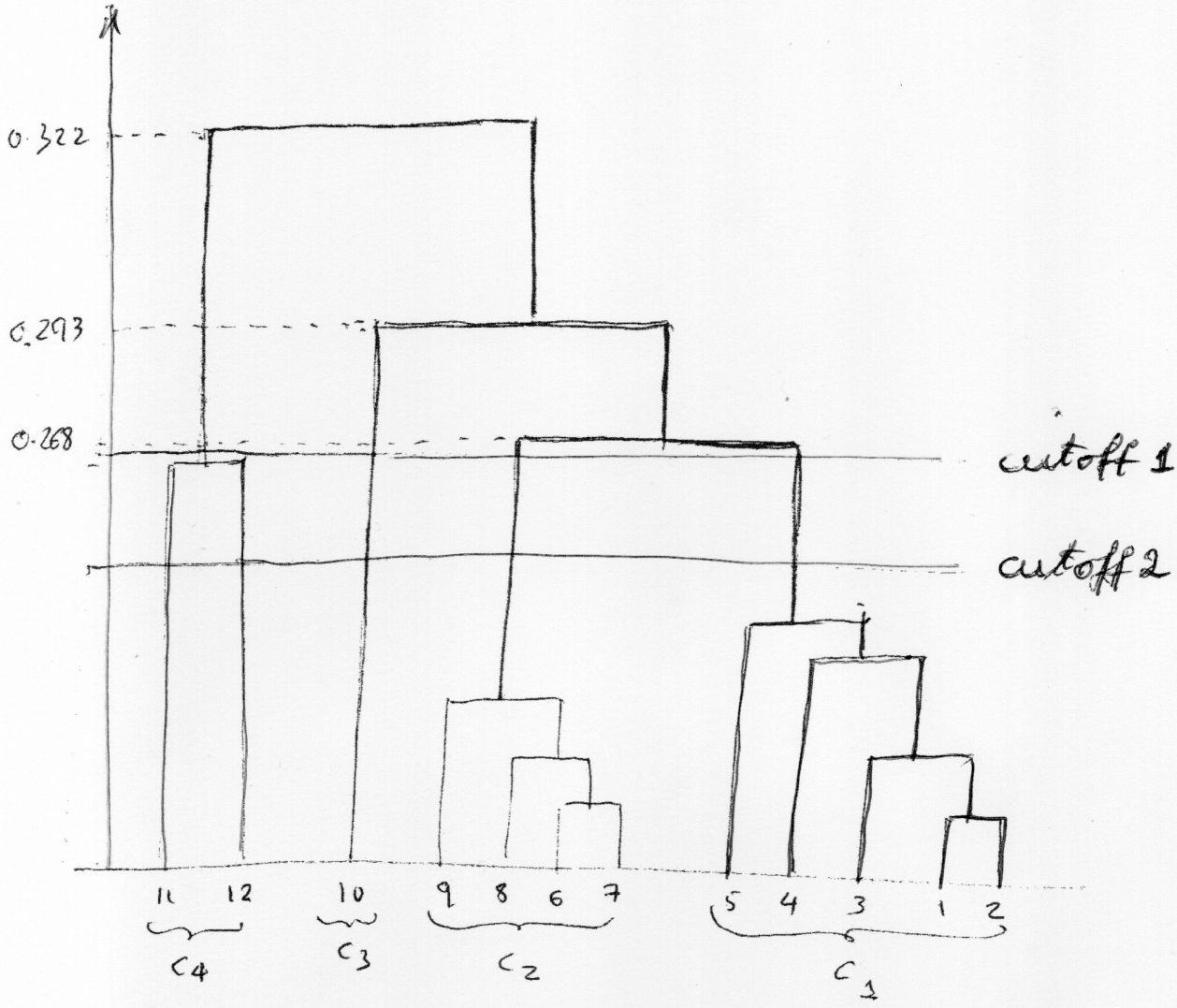
smallest value

⇒

	$C_1 \cup C_2 \cup C_3$
C_4	0.322

$D(C_3, C_1 \cup C_2) = \max\{D(C_3, C_1), D(C_3, C_2)\}$ etc...

$= \max\{0.284, 0.293\} = 0.293$



The option of forming 4 groups (cutoff 1) is not correct since it will produce a very heterogeneous group $C_1 = \{11, 12\}$, whose dissimilarity between its elements is approximately equal to the dissimilarity between the groups $C_3 = \{6,7,8,9\}$ and $C_4 = \{1,2,3,4,5\}$.

The correct option would be to form 5 clusters $C_1 = \{11\}$, $C_2 = \{12\}$, $C_3 = \{10\}$, $C_4 = \{6,7,8,9\}$ and $C_5 = \{1,2,3,4,5\}$ (cutoff 2). - creating separate and compact groups

d) Partition into 4 groups

1 2 3 4 5	6 7 8 9	10	11 12

Partition into 5 groups:

1 2 3 4 5	6 7 8 9	10	11	12

We have:

i) There is only one pair for which both partitions don't agree → (11, 12)

$\Rightarrow B + C = 1$ (see slide 170)

ii) The Total number of pairs is

$$\binom{N}{2} = \binom{12}{2} = \frac{12 \times 11}{2} = 66$$

iii) The number of pairs classified in the same way by both partitions is

Therefore $A + D = 66 - 1 = 65$

Hence, $RI = \frac{A+D}{\binom{N}{2}} = \frac{65}{66}$ //

10. a) Both methods seek to minimize the statistic SSQw (see slide) ie, to minimize the intra-cluster variance. Since the k-means clustering algorithm was applied using as initial seeds the centroids of the groups obtained with Ward's clustering method and in each step of the k-means algorithm the intra-cluster variance (SSQw) cannot increase, the value of SSQw is smaller than or equal to the value of the SSEw statistic of the partition obtained with Ward's method. Actually if there are misplaced elements, ie belonging to a cluster but closer to the centroid of other cluster the value of SSQw will be lower with the k-means clustering. Otherwise both methods produce the same value of the SSQw statistic. Since this statistic is a measure of the groups heterogeneity, the

16

The groups of the partition obtained with the K-means algorithm are more homogeneous (or equally homogeneous if there are not necessity to reassign elements)

b) $$RI = \frac{A+D}{\binom{N}{2}} = \frac{\binom{N}{2} - (B+C)}{\binom{N}{2}} = \frac{21945 - 1358}{21945}$$

$$\approx 0.9381\ldots$$

(see slide 170/171)

c)

d_{ij}	c_1	c_2
c_2	29.44	
c_3	21.55	41.80

$\Longrightarrow$

d_{ij}^2	c_1	c_2
c_2	866.71	
c_3	464.40 (smallest value)	1747.24

Next fusion: $C_{13} = C_1 \cup C_3$ with squared fusion cost

$$d^2_{13} = 464.4$$

Last fusion: $C_{13} \cup C_2 = C_{13,2}$

$$d_{13,2} = ?$$

By the Lance-Williams updating formula for the Ward's method, with $i=1, j=3, k=2$ $n_1 = 73$, $n_2 = 70$ and $n_3 = 67$ we get,

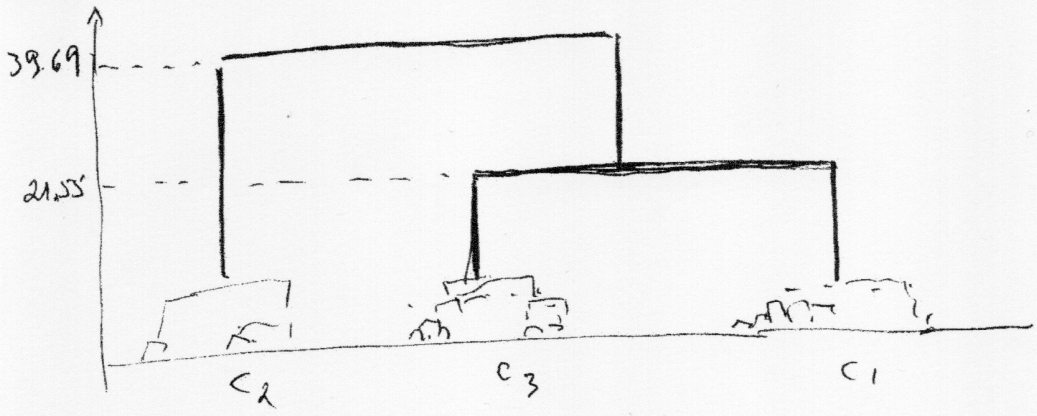
$$d^2_{13,2} = \frac{n_1+n_2}{n_1+n_2+n_3} d^2_{1,2} + \frac{n_3+n_2}{n_1+n_2+n_3} d^2_{3,2} - \frac{n_2}{n_1+n_2+n_3} d^2_{1,3}$$

$$= \frac{73+70}{210} 866.71 + \frac{67+70}{210} 1747.24 - \frac{70}{210} 464.40 = 1575.25$$

Hence $d^2_{13,2} = 1575.25 \Longrightarrow d_{13,2} = D(C_{13}, C_2) =$

$$= \sqrt{1575.25} = 39.69 //$$

Final dendrogram:

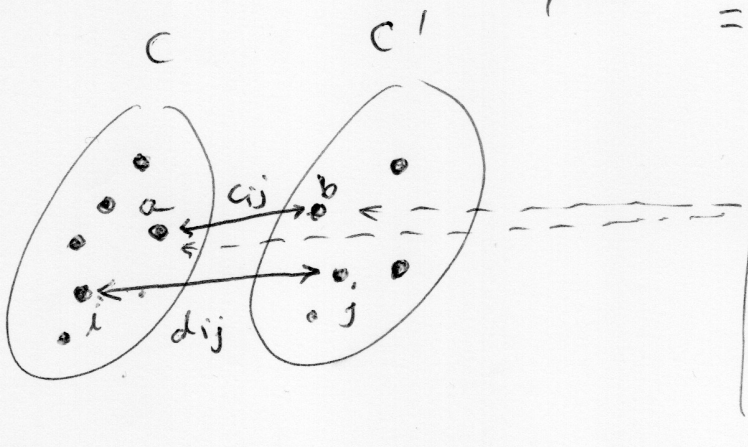


13. By difinition, the <u>cophenetic distance</u> between two individuals i and j is the fusion cost at which i and j became members of the same cluster for the first time (see slide 138).

If in the step previous to the fusion of the groups C and C', we have $i \in C$ and $j \in C'$ then,

$$h_{\$} = c_{ij} = D(C, C') = \min_{\substack{x \in C \\ y \in C'}} d(x,y) \leq d_{ij}$$

$$= \underbrace{d(a,b)}_{c_{ij}} \leq d_{ij}$$



a, b: nearest neighbors, one in C and the other in C'

The other inequality, $d_{ij} \leq h_c$ is left to the students.

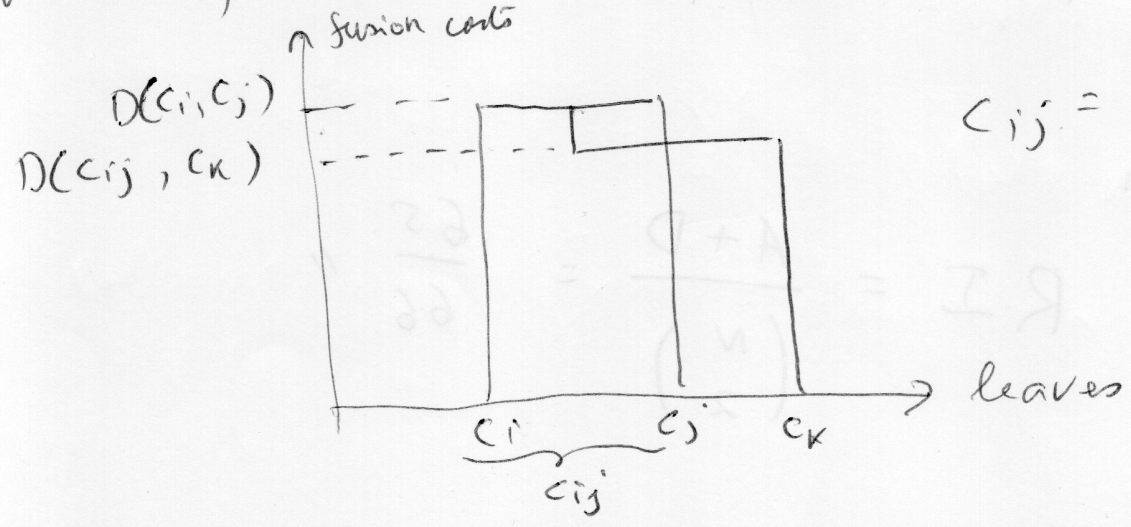
14. 19

We shall prove the result only for the complete-linkage method (~~for~~ the single-linkage method is similar and left to the students)

Let C_{ij} be a cluster that was ~~obtained~~ obtained by merging C_i and C_j at some step. If this cluster is going to be merged with a cluster C_k we have to prove that

$$D(C_{ij}, C_k) \geq D(C_i, C_j) \quad (\#\#\#)$$

That is, we cannot have an inversion:



Since C_i and C_j were merged together we know that,

$$D(C_i, C_j) \leq D(C_r, C_s) \quad \forall\, r \neq s,$$

for all clusters C_r, C_s present at the step before the fusion of C_i and C_j (why?). In particular,

$$D(C_i, C_j) \leq D(C_i, C_k),\ D(C_j, C_k)$$

and therefore

$$d_{ij} = D(C_i, C_j) \leq \max\{D(C_i, C_k), D(C_j, C_k)\}$$
$$= \max\{d_{ik}, d_{jk}\} = d_{(ij),k}$$

(see slide 113)

Hence the complete-linkage method is monotonic ie doesn't have inversions.