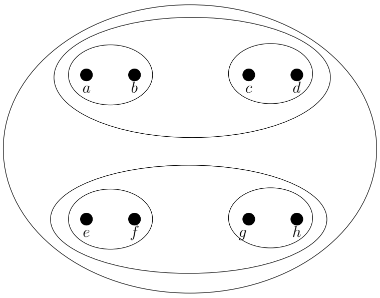
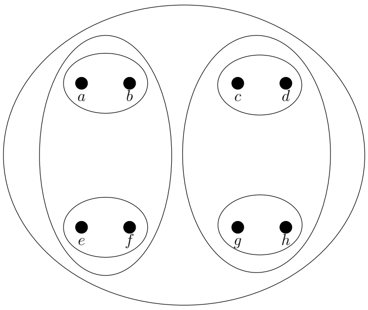
Clustering analysis - exercises (22/23)

The exercises marked with a (*) were partially inspired or modified from exercises appearing in the bibliography cited in the slides. Some exercises arose or were modified from previous evaluation tests and exams.

1. (*) Two hierarchical clustering analyses were performed on the set of 8 points $X = \{a, b, c, d, e, f, g, h\}$, with the euclidean distance, yielding the following two nested partitions of X i) and ii).



i



ii

   Indicate, justifying, two hierarchical clustering methods that could have produced these nested partitions and represent the respective (approximated) dendrograms.

2. (*) Consider the set of points in the real line,

$$X = \{0.2, 3, 4.2, 5, 5.9\}.$$

   (a) Perform a partition of this set into 2 groups using the complete method with the euclidean distance and represent the respective dendrogram. Comment.

   (b) Indicate the respective cophenetic distances matrix.

   (c) Compute the respective cophenetic Pearson and Spearman correlation coefficients.

3. (*) Perform a hierarchical classification of the set of points

$$X = \{(1, 2), (2, 2), (4.5, 3), (6, 3)\},$$

   using the average hierarchical method and the Manhattan distance.

4. (*) Apply the hierarchical centroid clustering method to the set of points in the plane,

$$X = \{(0, 0), (8, 0), (4, 7.5)\}$$

   and represent the respective dendrogram. Comment.

5. The following table contains the presences (1) / absences (0) records with respect to 10 species of fishes in 4 river basins located in Africa.

	SP1	SP2	SP3	SP4	SP5	SP6	SP7	SP8	SP9	SP10
OUEME	1	0	0	1	0	1	1	1	0	1
GAMBIE	1	0	1	0	1	1	0	0	0	1
GEBA	0	1	1	1	0	1	0	0	0	0
CRUBAL	0	1	0	0	1	1	0	0	0	0

Investigate if these river basins can be aggregated into homogeneous groups regarding the presences of the 10 species, using the complete method and an appropriate dissimilarity measure.

6. The following table contains the components of 5 binary vectors, a, b, c, d, e:

a:	1	0	0	1	1
b:	0	1	1	0	1
c:	1	0	0	0	1
d:	1	1	0	1	0
e:	0	1	1	0	0

   Perform classification analyses on the set of these binary vectors using the Manhattan distance with the single- and complete-linkage methods. Coment.

7. Considere a tabela de contingência do slide 77. Consider the contingency table of slide 77.

   (a) Classify the 5 countries according to the primary spoken language using the complete method and an appropriate distance.

   (b) Investigate if the 5 spoken languages can be aggregated into homogeneous groups with regard to their distribution by the countries, using Ward's method and an appropriate distance.

8. A clustering analysis of a set of $N = 178$ wines was performed with the k-means algorithm, considering the number of clusters k ranging from 1 to 10. The values obtained for the total within-groups inertia (SSQ_w) are reported in the table below, as a function of the number of clusters. It was decided to classify the wines into 3 groups.

k	1	2	3	4	5	6	7	8	9	10
SSQ_w	2301	1649	1271	1174	1116	1064	992	930	921	895

   Posteriorly, it were also performed classifications in 3 groups with the hierarchical *single*, *complete*, *average* and *Ward's* aggregagtion methods. The classifications were then pairwise compared with the RAND index. The results are reported below.

	single	complete	average	Ward
complete	0.3467			
average	0.9346	0.3495		
Ward	0.3445	0.8302	0.3448	
k−means	0.3460	0.8202	0.3467	0.9407097

(a) Justify that the total sum-of-squares (SSQ_t) of the distances of each point (wine) to the cloud's center of gravity is equal to the total within-groups inertia (SSQ_w) if $k = 1$.

(b) Give support to the decision of forming 3 groups of wines using two distinct criteria.

(c) Knowing that Ward's and k-means methods assign the same class to 4767 pairs of wines, determine the number of pairs of wines for which both clustering methods did not agree.

(d) Perform a hierarchical clustering analysis with the complete method using an appropriate dissimilarity measure, to aggregate into homogeneous groups the partitions obtained applying the clustering methods of table above. Represent the respective dendrogram and comment.

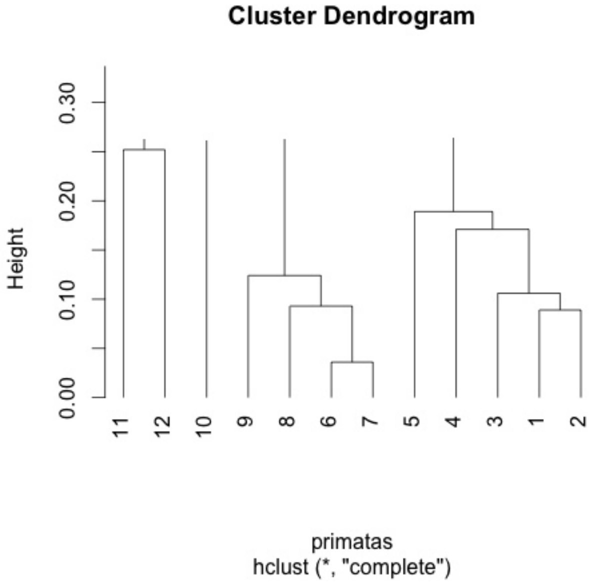
9. We call *diameter* of a set C to the largest pairwise dissimilarity between elements of C, i.e., $\text{diam}(C) = \max_{x,y \in C} d(x, y)$. We call *diameter* of a partition $X = C_1 \cup \ldots \cup C_k$, to the largest of the diameters of its clusters, i.e., to $\max\{\text{diam}(C_1), \ldots, \text{diam}(C_k)\}$.

In the "DIMACS Workshop on Reticulated Evolution" organized by the Rutgers University in september 2004, the researchers P. Legendre and V. Makarenkov illustrate a method to define dissimilarities between species. An example was presented by the authors regarding the dissimilarities between 12 species of primates:

1. Homo sapiens	7. Macaca mulatta
2. Pan	8. Macaca fascicular.
3. Gorila	9. Macaca sylvanus
4. Pongo	10. Saimiri sciureus
5. Hylobatas	11. Tarsius syrichta
6. Macaca fuscata	12. Lemur catta

Based on this data a hierarchical classification of the set of primates into the 4 groups $C_1 = \{1, 2, 3, 4, 5\}$, $C_2 = \{6, 7, 8, 9\}$, $C_3 = \{10\}$ and $C_4 = \{11, 12\}$ was obtained, using the complete method and the dissimilarity matrix below (the respective parcial dendrogram is depicted in the next page).

```
       1     2     3     4     5     6     7     8     9    10 11
2 0.089
3  0.104 0.106
4  0.161 0.171 0.166
5  0.182 0.189 0.189 0.188
6  0.232 0.243 0.237 0.244 0.247
7  0.233 0.251 0.235 0.247 0.239 0.036
8  0.249 0.268 0.262 0.262 0.257 0.084 0.093
9  0.256 0.249 0.244 0.241 0.242 0.124 0.120 0.123
10 0.273 0.284 0.271 0.284 0.269 0.289 0.293 0.287 0.287
11 0.322 0.321 0.314 0.303 0.309 0.314 0.316 0.311 0.319 0.320
12 0.308 0.309 0.293 0.293 0.296 0.282 0.289 0.298 0.287 0.285 0.252
```



(a) Determine the diameters of the groups C_1, C_2, C_3 and C_4. What is the partition diameter?

(b) Complete the dendrogram and comment the option of forming 4 groups.

(c) Consider the partition of the set of primates into 5 groups defined by the dendrogram.

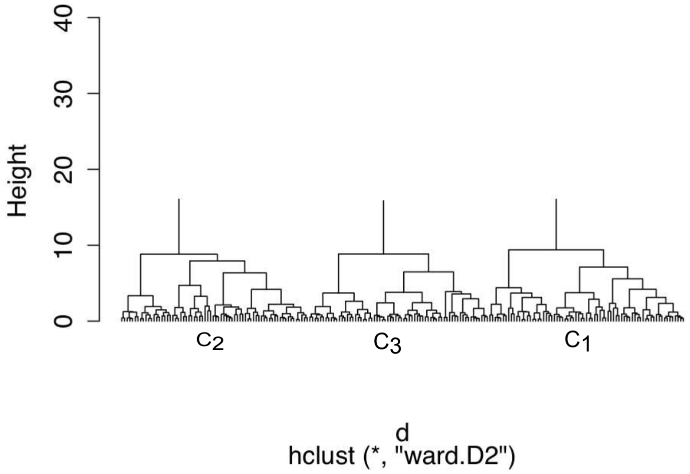
   i. Write the groups and determine how many pairs of primates would be classified distinctly by the new partition into 5 groups and the partition into 4 groups referred before.

   ii. Derive from the previous response the RAND index between the two partitions.

10. A study was conducted on seeds of three distinct varieties of wheat, Kama, Rosa and Canadiano. For this study, 70 seeds of each one of the varieties were randomly chosen and seven variables were observed for each seed:

Name	Description	Units
`Area`	Area (A)	mm^2
`Perimeter`	Perimeter (P)	mm
`Compacteness`	$\frac{4\pi A}{P^2}$	–
`Kernel_length`	Length	mm
`Kernel_width`	Width	mm
`asym_coeff`	Coefficient of assimetry	–
`length_kernel_groove`	Length of the groove	mm

A hierarchical agglomerative clustering was applied to the standardized variables of the wheat seeds using Ward's method and the euclidean distance. Then a cut in the dendrogram was performed yielding a partition of the dataset into 3 groups, C_1, C_2 and C_3, containing 73, 70 and 67 elements, respectively, as indicated in the partial dendrogram depicted below.



The previous classification was then consolidated with the k-means clustering method using the centroids of the groups C_1, C_2 and C_3 as initial seeds. It turned out that both classifications did not agree with respect to 1358 pairs of wheat seeds.

(a) Which of the two classifications produce the more homogeneous clusters? Justify.

(b) Determine the RAND index between the two partitions.

(c) Knowing that the pairwise distances between clusters C_1, C_2 and C_3 are given by,

$$d(C_1, C_2) = 29.44, \qquad d(C_1, C_3) = 21.55, \qquad d(C_2, C_3) = 41.80,$$

complete the dendrogram, indicating the fusion costs between the groups that are aggregated.

11. A study involved the observation of 4177 marine gastropod molluscs (abalones) of the *Haliotis rubra* species picked at random. For each individual were measured 8 numerical variables and determined the sex within 3 categories, male (`M`), female (`F`) and juvenile (`I`).

One of the 8 numerical variables, `Rings`, takes integer values and indicates the age of the individual throw the counting of the rings. The remaining are continuous variables: (`Length`); (`Diameter`) and (`Height`) - all in mm - ; the overall weight of the organism (`Whole`); the weight of the organism without the shell (`Shucked`); the visceras weight (`Viscera`); and the dry shell weight (`Shell`) - these in g.

The correlations matrix between the numerical variables and an image of an abalone is presented below.

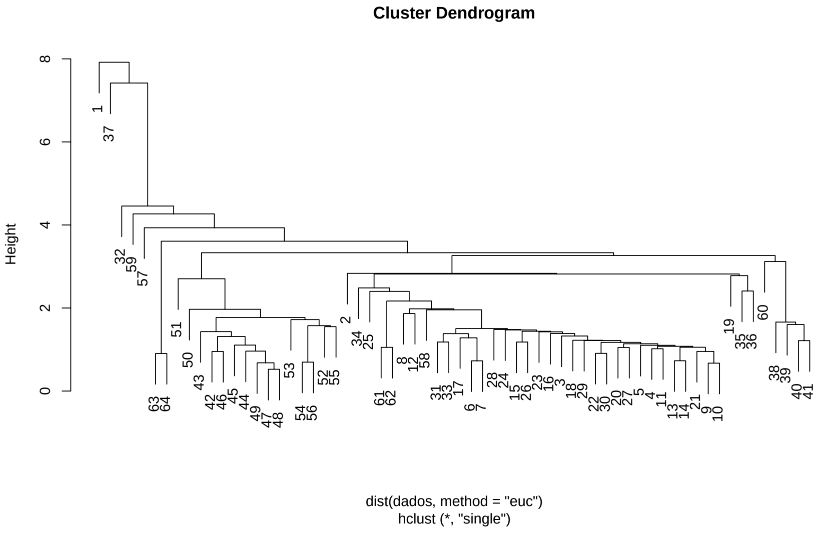
```
         Length Diameter Height Whole Shucked Viscera Shell Rings
Length    1.000    0.987  0.828 0.925   0.898   0.903 0.898 0.557
Diameter  0.987    1.000  0.834 0.925   0.893   0.900 0.905 0.575
Height    0.828    0.834  1.000 0.819   0.775   0.798 0.817 0.557
Whole     0.925    0.925  0.819 1.000   0.969   0.966 0.955 0.540
Shucked   0.898    0.893  0.775 0.969   1.000   0.932 0.883 0.421
Viscera   0.903    0.900  0.798 0.966   0.932   1.000 0.908 0.504
Shell     0.898    0.905  0.817 0.955   0.883   0.908 1.000 0.628
Rings     0.557    0.575  0.557 0.540   0.421   0.504 0.628 1.000
```

(a) Using an appropriate dissimilarity measure and the single-linkage method, perform a hierarchical classification of the set of the 6 continuous variables, `Length`, `Diameter`, `Height`, `Whole`, `Shucked` and `Viscera`, into homogeneous groups and comment the result. Write also the respective cophenetic distances matrix.

(b) Appying Ward's aggregation method to the 6 standardized variables with the euclieedean distance it was obtained a partition of the set of gastropods into 2 groups. This partition was then compared with the partition into the 2 groups, *juvenile* and *no juvenile* given by the variable `Sex`, using the RAND index, yielding a value of 0.6712376.

(c) What was the number of pairs of gastropods classified distinctly by both clustering methods?

12. A cluster analysis was performed on a set of 64 standardized observations using the hierarchical single-linkage method with the euclidean distance, yielding the dendrogram below. It is known that the cophenetic distances matrix associated with this dendrogram contains the values 7.42 and 7.92.



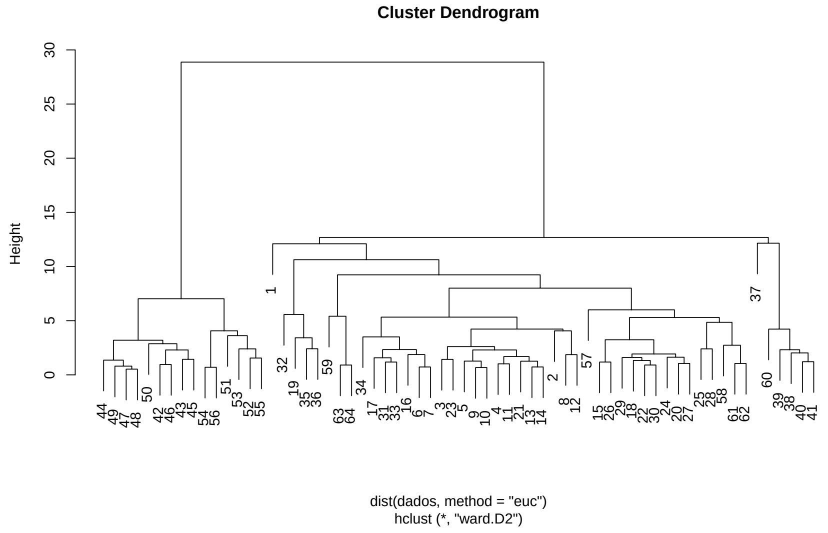
According to the available information, justify which of the following sentence(s) you can assure that it is correct:

(a) The distance between observations 1 and 37 is equal to 7.92.

(b) The distance between observations 1 and 37 is not inferior to 7.92.

(c) The distance between observations 37 and 32 smaller than or equal to the distance between observations 1 and 37.

Posteriorly, it was performed a clustering analysis on the same set of the 64 standardized observations using the hierarchical Ward's method with the euclidean distance and a cut in the respective dendrogram applied to obtain a partition into 5 groups. The dendrogram is depicted below and the pairwise distances between the 5 groups presented in the next table (rounded to 2 decimal places), where the designations of the groups follow the order the groups in the dendrogram from left to right.



	C_1	C_2	C_3	C_4
C_2	15.58			
C_3	28.88	12.10		
C_4	16.81	12.13	12.28	
C_5	14.57	13.04	12.48	12.15

(a) Determine the cophenetic distance between observations 1 and 37 for the Ward's method. What is the meaning of this distance?

(b) A consolidation procedure was performed to the partition into 5 groups applying the k-means clustering method with initial seeds given by the centers of gravity of the 5 groups. It turned out that we get the same classes previously obtained with Ward's method. What do you conclude?

(c) The label (tag) of each one of the 64 observations began with a two-symbol code, as follows:

```
> groups
 [1] 4C 4C 4C 3A 4C 2A 4C 2A 2A 2A 2A 2A 2A 2A 2A 2A
 [17] 2A 2A 2A  2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A
[33] 2A 2A 2A 2A 2A 2A 2A 4B 4B 2A 2A 2A 2A 2A 2A 2A
[49] 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A
> table(groups)
2A 3A 4A 4B 4C
40  1 16  2  5
```

Computation of the RAND index between the partition into 5 groups defined by these tags **2A**, **3A**, **4A**, **4B** and **4C**, and the partition into 5 groups defined by Ward's method yielded and random equal to 0.5729167, with 642 pairs of observations assigned to distinct groups by the two classification procedures. Determine the number of pairs that are assigned in the same group by both classifications.

(d) It turned out later that one of the observations with the tag **4B** was misclassified, and the new tag **4D** was assigned to it. Recompute the RI between the classification into the 6 groups defined by the tags and the previous classification into 5 groups given by Ward's method.

13. Consider a hierarchical agglomerative classification and denote by d_{ij} the dissimilarity between two individuals i and j. Let h_c denote the cophenetic distance between i and j for the clustering with the complete method and h_s denote the cophenetic distance between i and j for the clustering with the single-linkage method. Prove that $h_s \leq d_{ij} \leq h_c$.

14. Prove that in the single- and complete-linkage methods the fusion costs are monotonically increasing. (without using Lance-Williams updating formula).

15. (a) Prove that, if in Lance-Williams's formula we have $\alpha_i, \alpha_j, \gamma \geq 0$ with $\alpha_i + \alpha_j + \beta \geq 1$, the fusion costs increase monotonically, i.e., the respective dendrogram does not have inversions.

    (b) Conclude that the fusion costs obtained by the average and Ward's hierarchical agglomerative methods increase monotonically.

16. Prove the Lance-Williams's updating formula for the average-linkage method.

17. In a classification with the k-means clustering method and 3 initial seeds, one of the final groups is empty. In which conditions the obtained solution is optimal?

18. The adjusted Rand index (ARI) between two partitions $\mathcal{P}$ and $\mathcal{Q}$ of the same finite set X is defined as

$$ARI(\mathcal{P}, \mathcal{Q}) = \frac{RI(\mathcal{P}, \mathcal{Q}) - \mathrm{E}[RI]}{\max(RI) - \mathrm{E}[RI]},$$

where $\max(RI) = 1$ and $\mathrm{E}[RI]$ is the expected value for the Rand index (RI) between randomly chosen independent partitions $\mathcal{P}'$ and $\mathcal{Q}'$ of X with the same cluster sizes of $\mathcal{P}$ and $\mathcal{Q}$, respectively.

Use the above formula to compute $ARI(ab|cd, ab|c|d)$.