Lecture 14: Clustering

IST5573

統計方法 Statistical methods

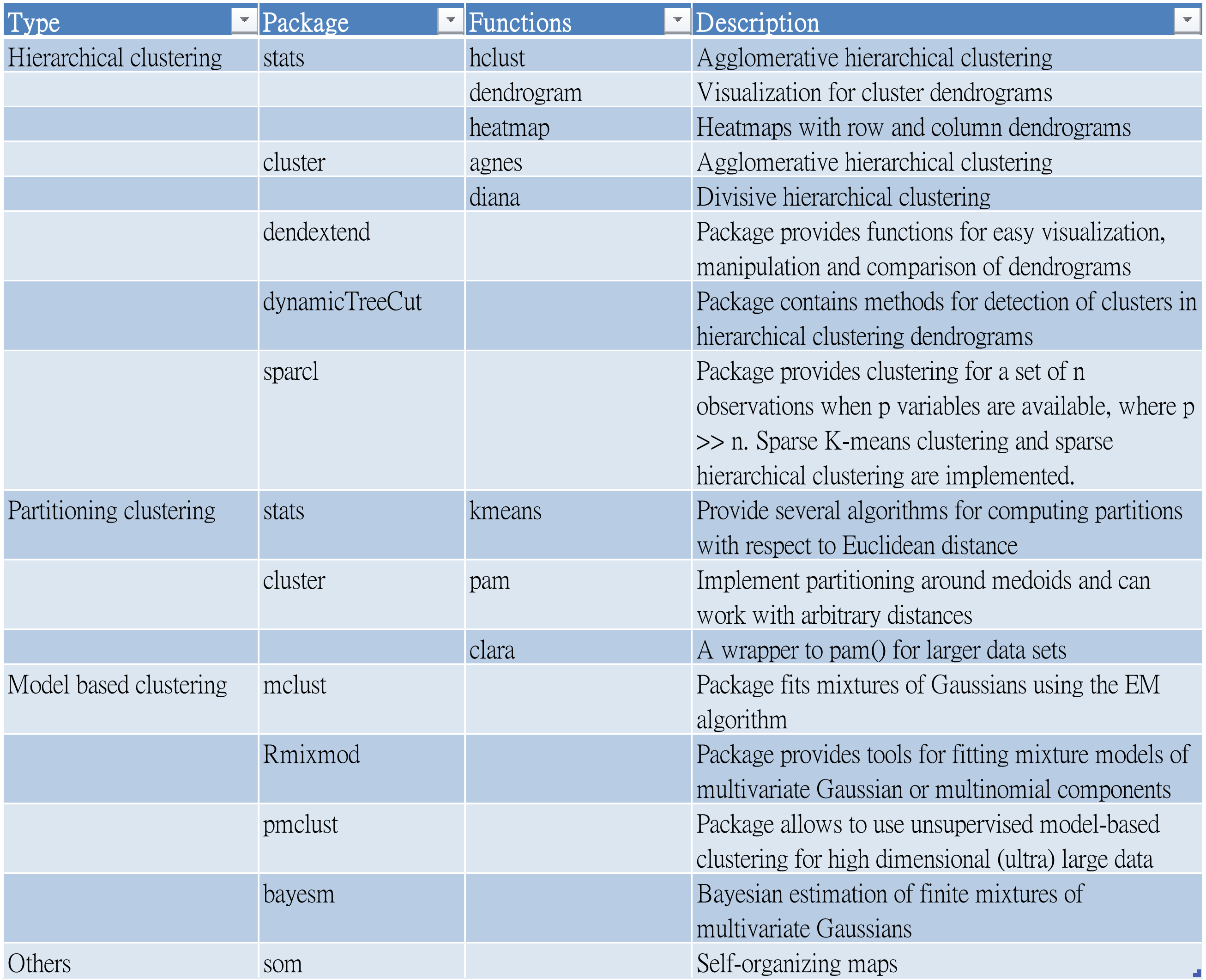
2016/12/14

Clustering vs. classification

- **Task**. Assign observational units to classes on the basis of variables describing/characterizing these observations.
- Clustering. The classes are unknown a priori and need to be "discovered" from the data.
- Classification. The classes are predefined and the task is to understand the basis for the classification from a set of labeled observations (learning set). This information is then used to predict the class of future observations.

Clustering

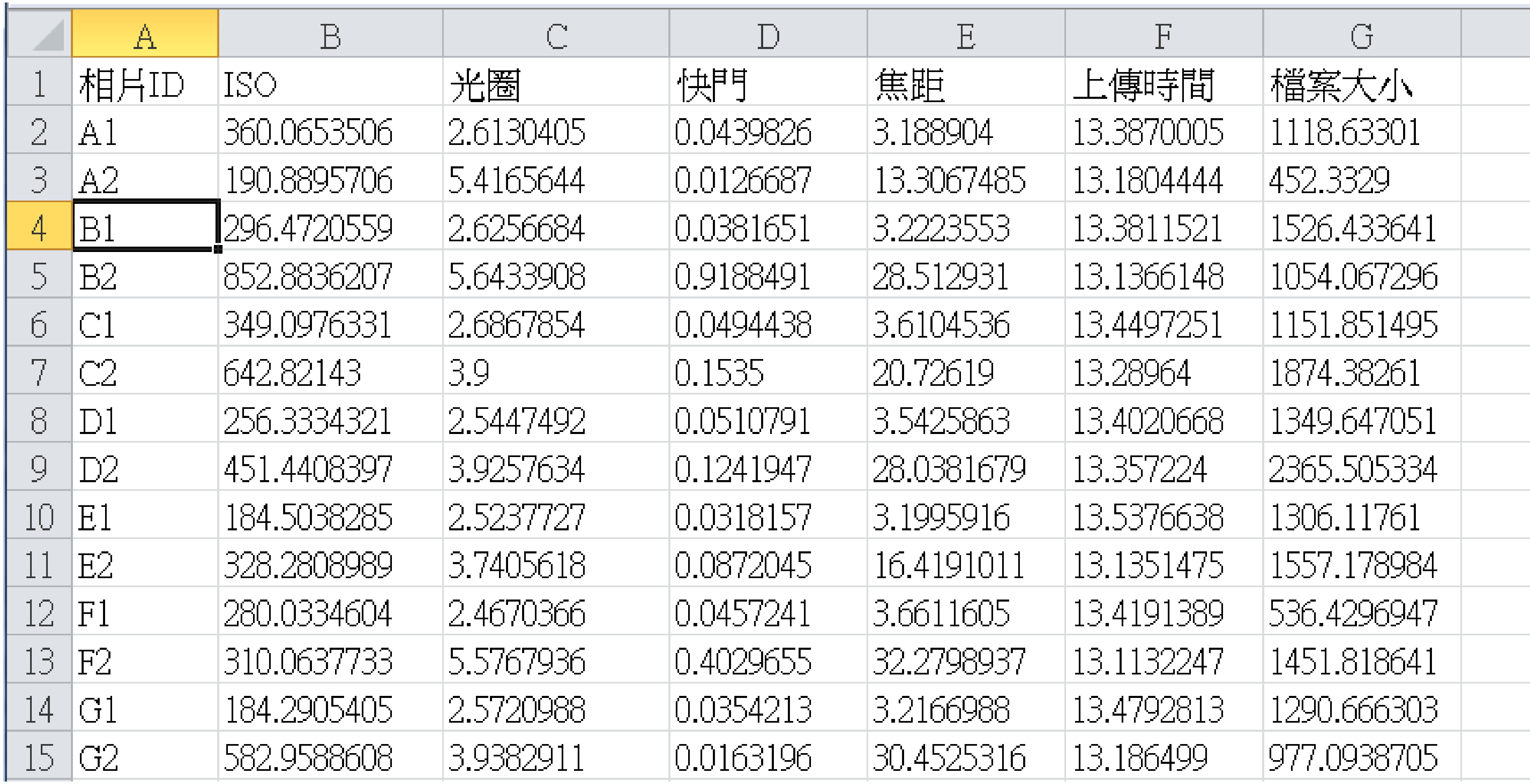
Cluster analysis: R software



Type	Package	Functions	Description
Hierarchical clustering	stats	hclust	Agglomerative hierarchical clustering
		dendrogram	Visualization for cluster dendrograms
		heatmap	Heatmaps with row and column dendrograms
	cluster	agnes	Agglomerative hierarchical clustering
		diana	Divisive hierarchical clustering
	dendextend		Package provides functions for easy visualization, manipulation and comparison of dendrograms
	dynamicTreeCut		Package contains methods for detection of clusters in hierarchical clustering dendrograms
	sparcl		Package provides clustering for a set of n observations when p variables are available, where p >> n. Sparse K-means clustering and sparse hierarchical clustering are implemented.
Partitioning clustering	stats	kmeans	Provide several algorithms for computing partitions with respect to Euclidean distance
	cluster	pam	Implement partitioning around medoids and can work with arbitrary distances
		clara	A wrapper to pam() for larger data sets
Model based clustering	mclust		Package fits mixtures of Gaussians using the EM algorithm
	Rmixmod		Package provides tools for fitting mixture models of multivariate Gaussian or multinomial components
	pmclust		Package allows to use unsupervised model-based clustering for high dimensional (ultra) large data
	bayesm		Bayesian estimation of finite mixtures of multivariate Gaussians
Others	som		Self-organizing maps

http://ghuang.stat.nctu.edu.tw/course/statmethods16/files/lectures/clustering_in_R.xlsx

照片分群



	A	B	C	D	E	F	G
1	相片ID	ISO	光圈	快門	焦距	上傳時間	檔案大小
2	A1	360.0653506	2.6130405	0.0439826	3.188904	13.3870005	1118.63301
3	A2	190.8895706	5.4165644	0.0126687	13.3067485	13.1804444	452.3329
4	B1	296.4720559	2.6256684	0.0381651	3.2223553	13.3811521	1526.433641
5	B2	852.8836207	5.6433908	0.9188491	28.512931	13.1366148	1054.067296
6	C1	349.0976331	2.6867854	0.0494438	3.6104536	13.4497251	1151.851495
7	C2	642.82143	3.9	0.1535	20.72619	13.28964	1874.38261
8	D1	256.3334321	2.5447492	0.0510791	3.5425863	13.4020668	1349.647051
9	D2	451.4408397	3.9257634	0.1241947	28.0381679	13.357224	2365.505334
10	E1	184.5038285	2.5237727	0.0318157	3.1995916	13.5376638	1306.11761
11	E2	328.2808989	3.7405618	0.0872045	16.4191011	13.1351475	1557.178984
12	F1	280.0334604	2.4670366	0.0457241	3.6611605	13.4191389	536.4296947
13	F2	310.0637733	5.5767936	0.4029655	32.2798937	13.1132247	1451.818641
14	G1	184.2905405	2.5720988	0.0354213	3.2166988	13.4792813	1290.666303
15	G2	582.9588608	3.9382911	0.0163196	30.4525316	13.186499	977.0938705

(RMD_example 14.1)

Variable	Description
相片ID	E.g., A1 → A：相簿分類，1：相機型號。共16張相片。
ISO	照片的感光度，數值越高，對光越敏感
光圈	光圈值
快門	快門速度 (秒)
焦距	毫米 (mm)
上傳時間	UNIX 時間戳，與1970年1月1日00:00:00的秒差$\times 10^{-8}$
檔案大小	Bytes

Distance and similarity

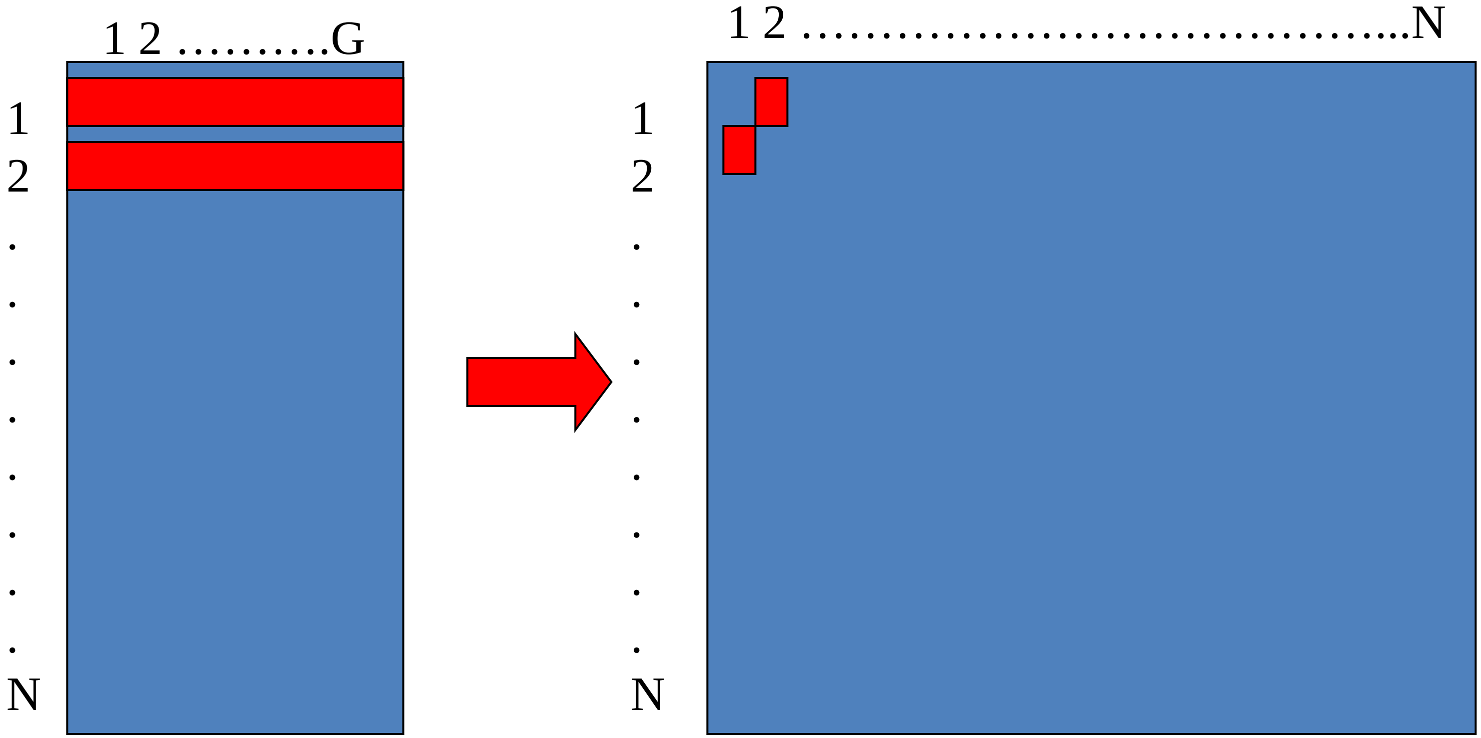
- Clustering organizes **points** that are close into groups.
- What does it mean for two **variables** to be close?
- What does it mean for two **samples** to be close?
- Points: E_{ig}= value of sample i, variable g
  - Variable1=(E_{11}, E_{21}, …, E_{N1})
  - Variable2=(E_{12}, E_{22}, …, E_{N2})

  - Sample1=(E_{11}, E_{12}, …, E_{1G})
  - Sample2=(E_{21}, E_{22}, …, E_{2G})

Distance and similarity

- Close: two points have a small distance or large similarity
- Every clustering method is based **solely** on the measure of distance or similarity.
- Distance
  - Euclidean distance
- Similarity
  - Correlation
  - Spearman correlation
  - Categorical measures

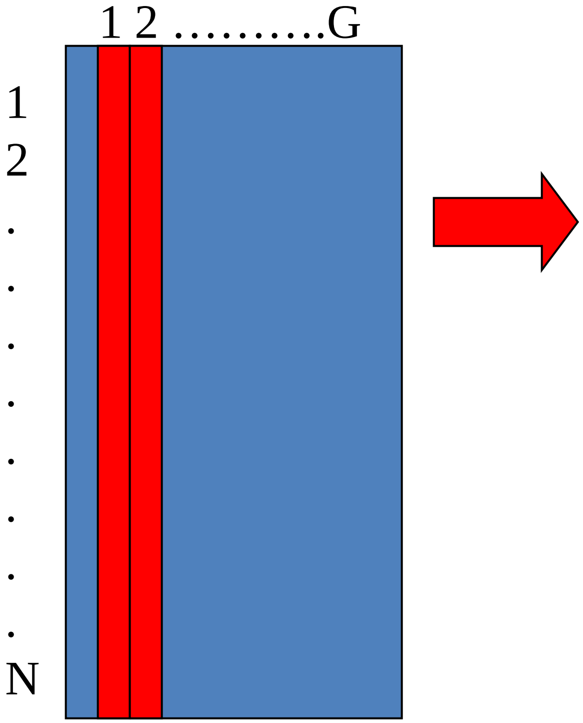
The similarity/distance matrices



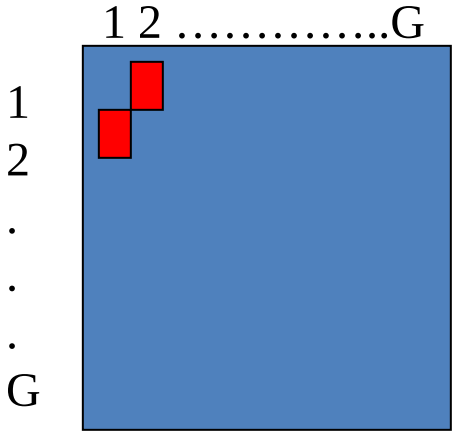
Data matrix

Sample similarity matrix

The similarity/distance matrices



Data matrix



Variable similarity matrix

Two common clustering approaches

- Hierarchical methods: provide a hierarchy of clusters, from the smallest, where all observations are in one cluster, through to the largest set, where each observation is in its own cluster.
    - either divisive or agglomerative
- Partitioning methods: partition the observations into disjoint clusters and usually require specification of the number of clusters.
    - K-means/K-medoids

Hierarchical clustering

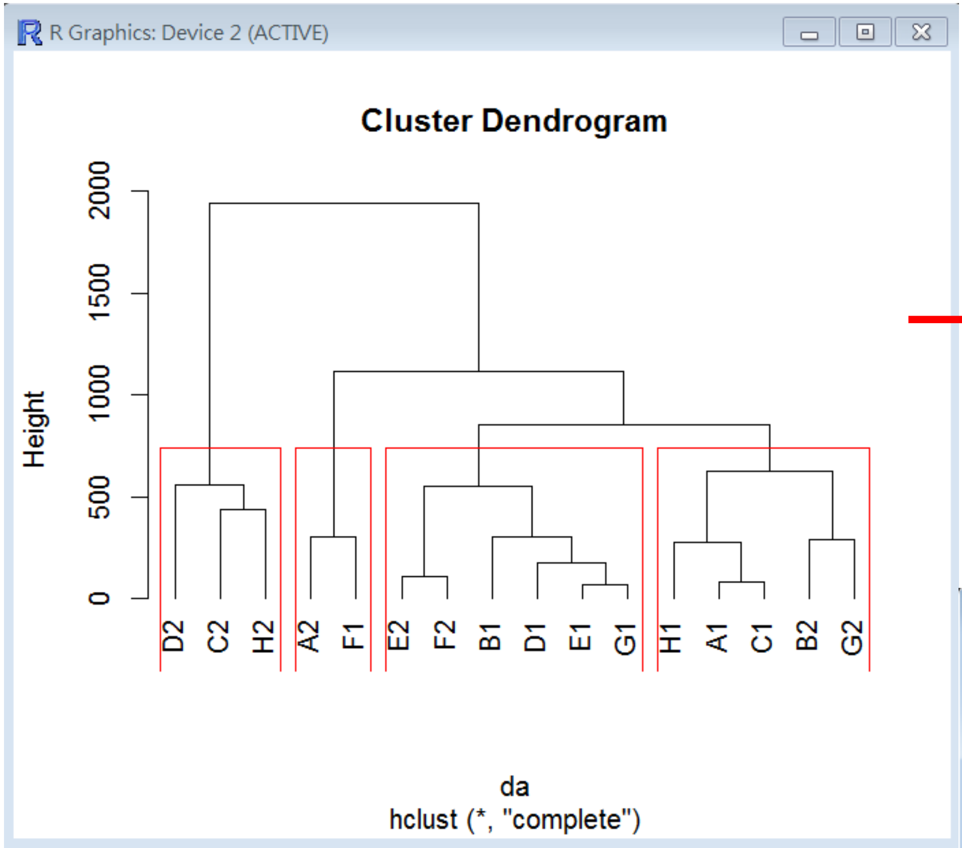
- Agglomerative clustering (bottom-up)
  - Starts with as each sample in its own cluster
  - Joins the two most similar clusters
  - Then, joins next two most similar clusters
  - Continues until all samples are in one cluster
- Divisive clustering (top-down)
  - Starts with all samples in one cluster
  - Choose split so that samples in the two clusters are most similar (maximize "distance" between clusters)
  - Find next split in same manner
  - Continue until all samples are in single clusters

Dendrograms

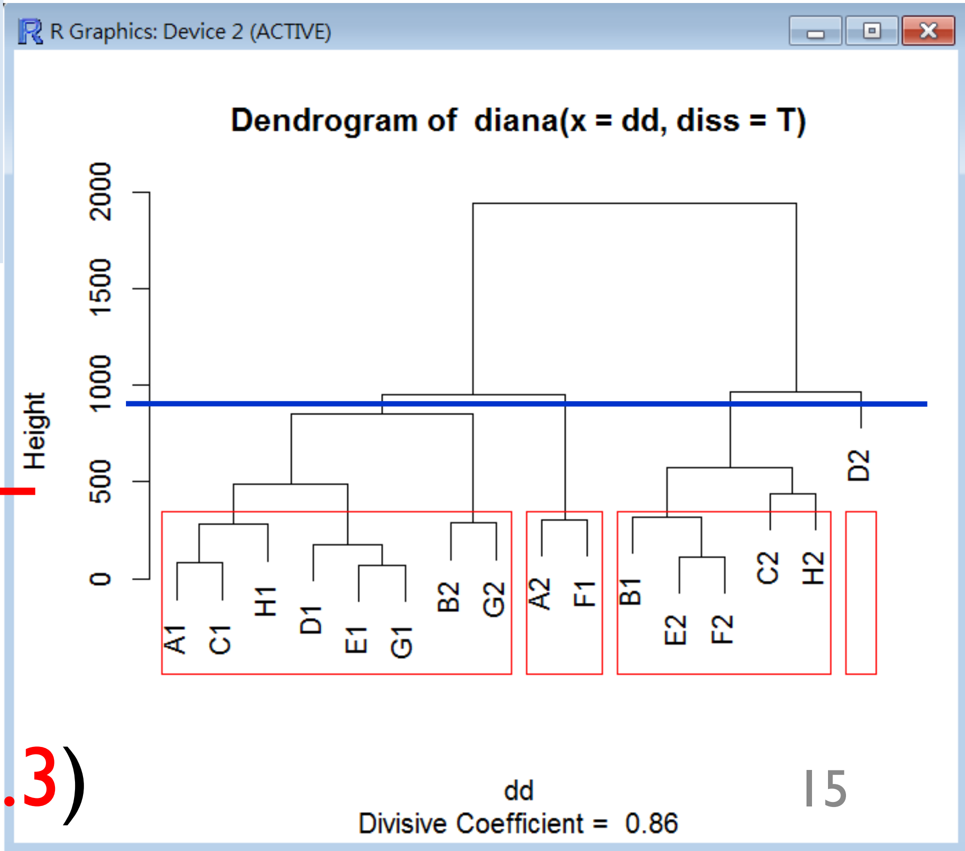
- Hierarchical clustering provides with clusters of every size: where to "cut" the "dendrogram" is user-determined
- We can then make dendrograms showing divisions or merging.
- The y-axis represents the distance between the groups divided at that point.

照片分群

- Sample：共16張相片
- Variable：ISO、光圈、快門、焦距、上傳時間、檔案大小



Agglomerative dendrogram


R Graphics: Device 2 (ACTIVE)
Dendrogram of diana(x = dd, diss = T)
Height
0
500
1000
1500
2000
A1
C1
H1
D1
E1
G1
B2
G2
A2
F1
B1
E2
F2
C2
H2
D2
dd
Divisive Coefficient = 0.86

Divisive dendrogram

(RMD_examples 14.2, 14.3)

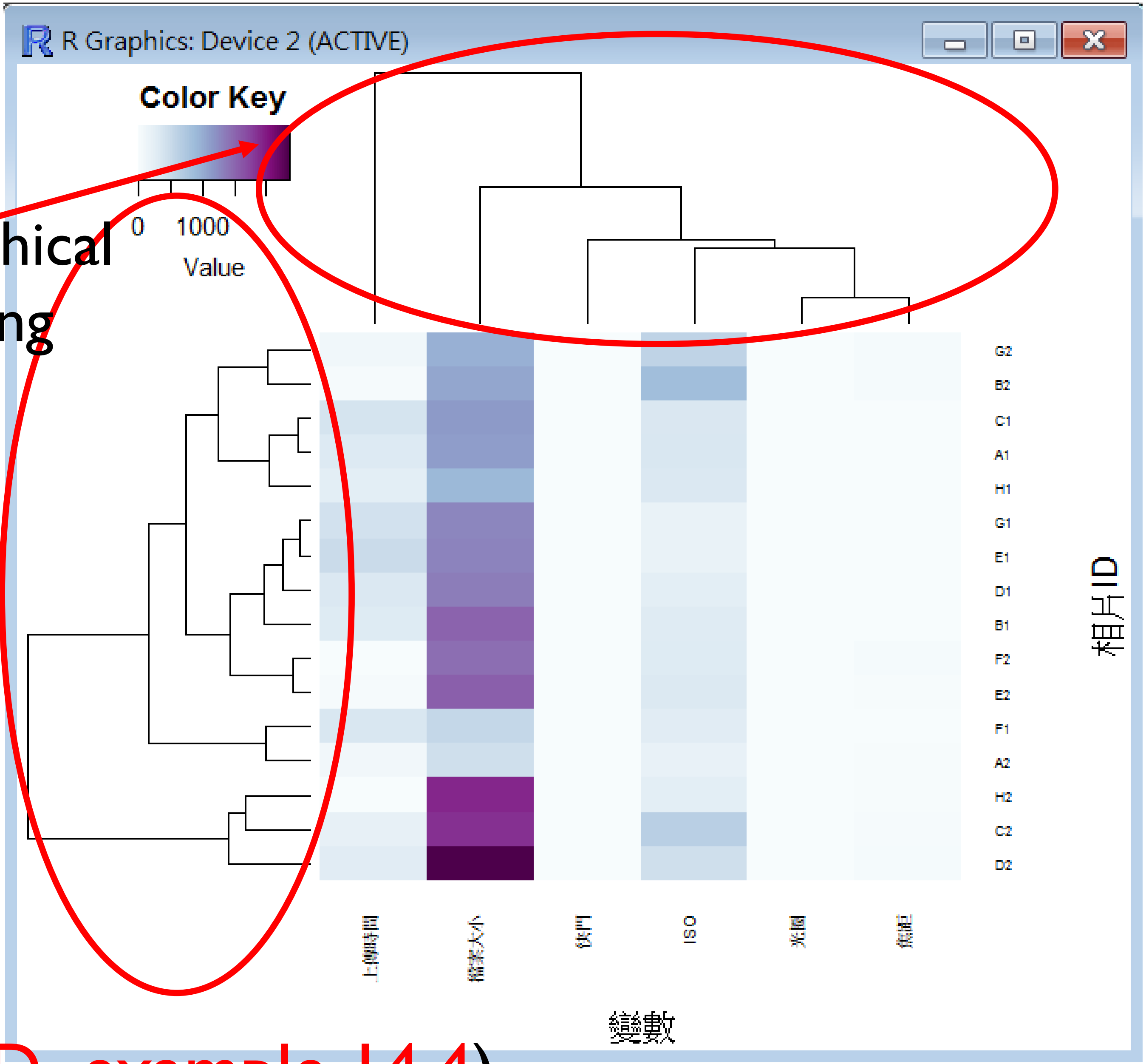
Heatmaps

- A two-dimensional, rectangular, colored grid.
- Displays the data matrix themselves.
- The color of each grid is determined by the value of the corresponding entry in the data matrix.
- The rows and columns of the matrix are reordered independently-- similar rows and columns are placed next to each other.

Heatmaps

- The orderings can be derived from a hierarchical clustering.
- Aid in determining which rows (the samples) have similar values within which subgroups of columns (the variables).

Heatmaps



Hierarchical clustering

(RMD_example 14.4)

How to make a hierarchical clustering

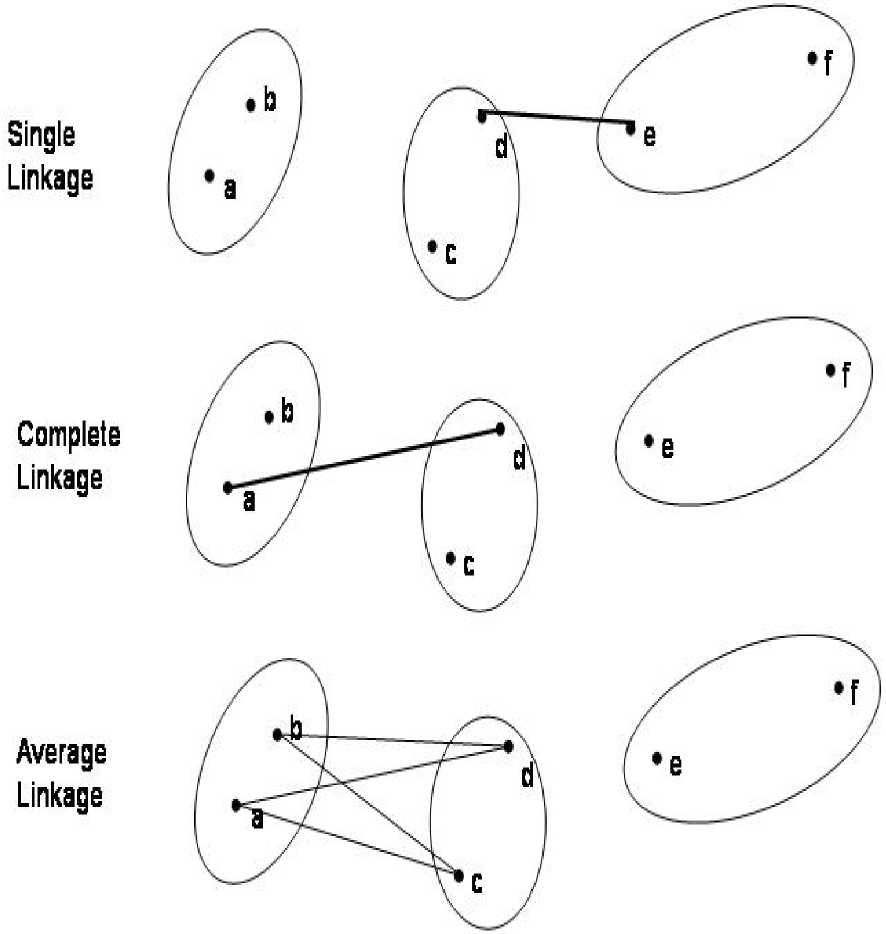
1. Choose samples and variables to include in cluster analysis
2. Choose similarity/distance metric
3. Choose clustering direction (top-down or bottom-up)
4. Choose linkage method (if bottom-up)
5. Calculate dendrogram
6. Choose height/number of clusters for interpretation
7. Assess cluster fit
8. Interpret resulting cluster structure

3. Choose clustering direction

- Both are only “step-wise” optimal: at each step the optimal split or merge is performed
- This does not imply that the final cluster structure is optimal!
- Agglomerative/Bottom-up
  - Computationally simpler, and more available
  - More “precision” at bottom of tree
  - When looking for small and/or many clusters, use agglomerative
- Divisive/Top-down
  - More “precision” at top of tree
  - When looking for large and/or few clusters, use divisive

4. Choose linkage method (if bottom-up)

- Single linkage: join clusters whose distance between closest samples is smallest
- Complete linkage: join clusters whose distance between furthest samples is smallest
- Average linkage: join clusters whose average distance is the smallest.



Partitioning clustering: k-means

- MUST choose number of clusters K a priori
- More of a “black box” because output is most commonly looked at purely as assignments
- Each object (variable or sample) gets assigned to a cluster
- Begin with initial partition
- Iterate so that objects within clusters are most similar
- Not unique solution: clustering can depend on initial partition

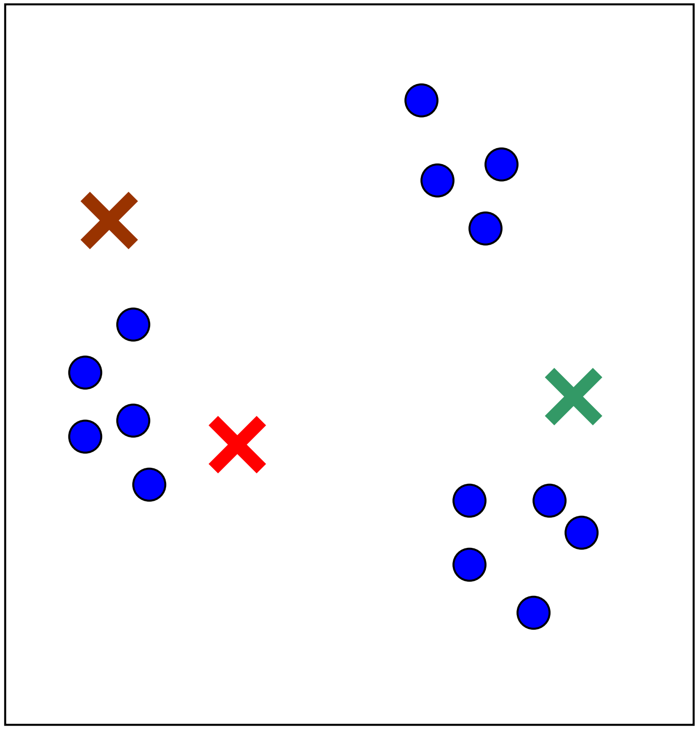
K-means algorithm

1. Choose K centroids at random
2. Make initial partition of objects into k clusters by assigning objects to closest centroid
3. Calculate the centroid (mean) of each of the k clusters.
4. a) For object i, calculate its distance to each of the centroids.

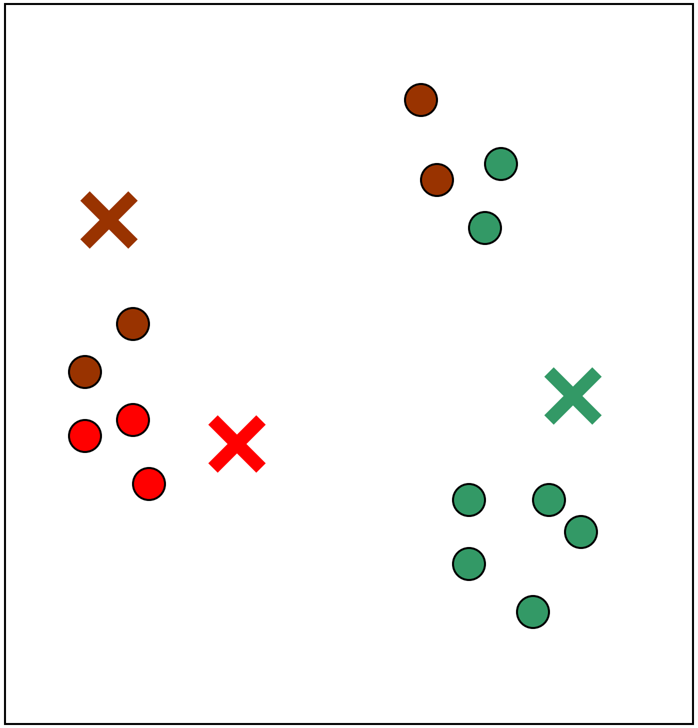
   b) Allocate object i to cluster with closest centroid.

   c) If object was reallocated, recalculate centroids based on new clusters.
5. Repeat 4 for object i = 1,….N.
6. Repeat 3 and 4 until no reallocations occur.
7. Assess cluster structure for fit and stability

K-means algorithm

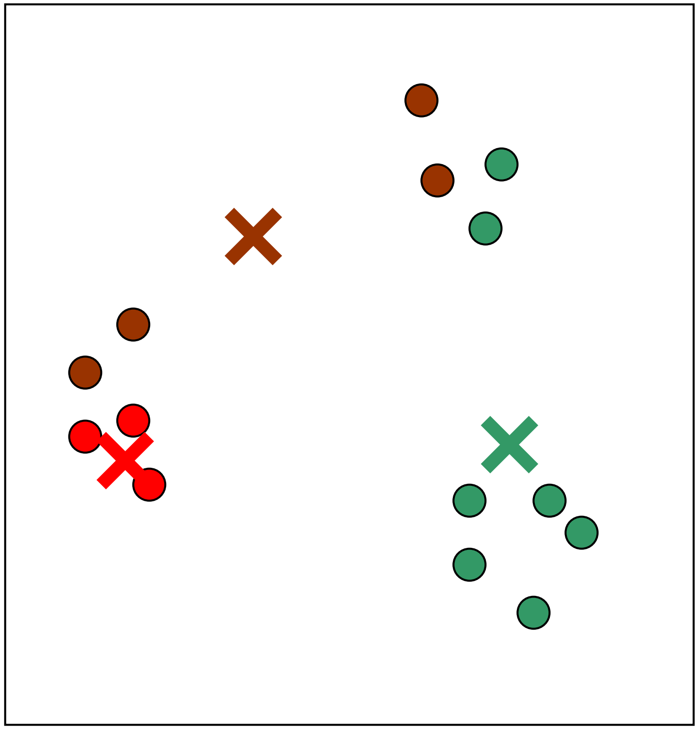


Iteration = 0

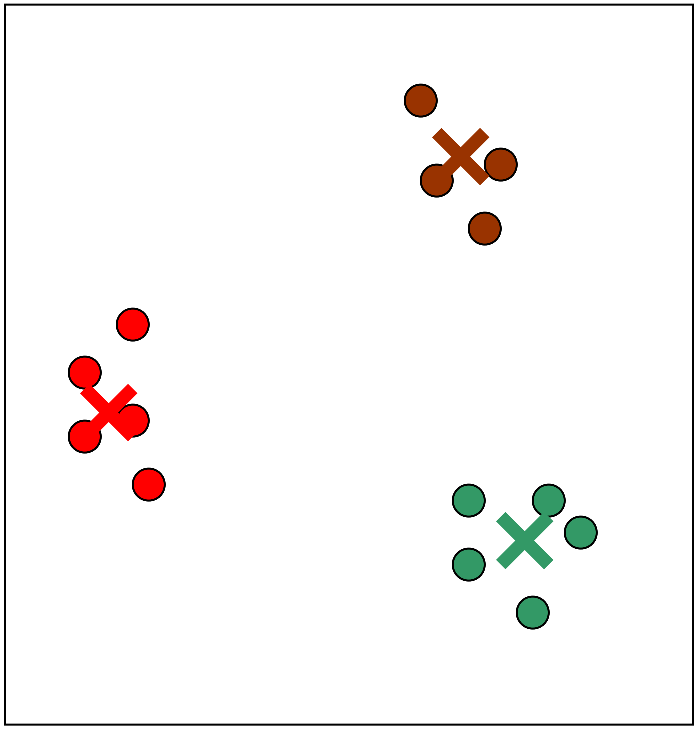


Iteration = 1

K-means algorithm



Iteration = 2

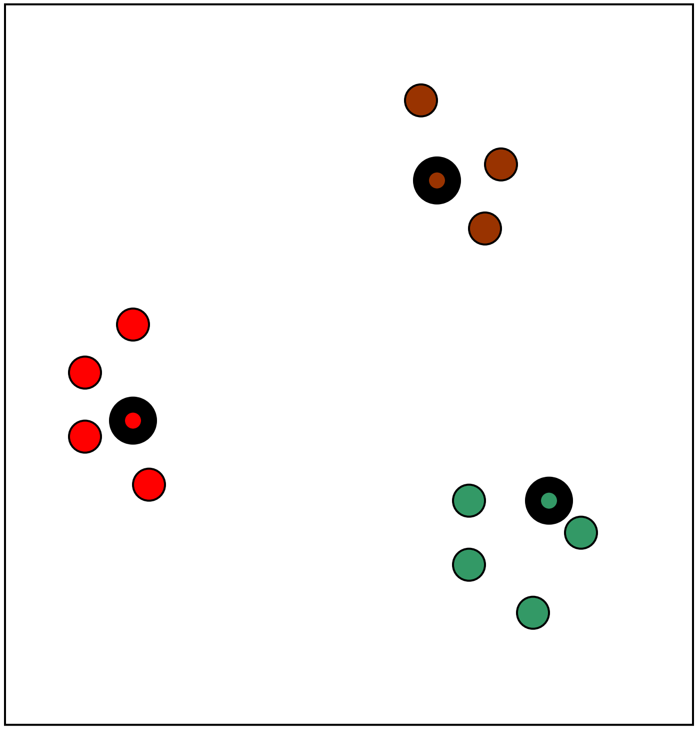


Iteration = 3

(RMD_example 14.5)

K-medoids

- A little different
- Centroid: The average of the samples within a cluster
- Medoid: The “representative object” within a cluster.
- Initializing requires choosing medoids at random.



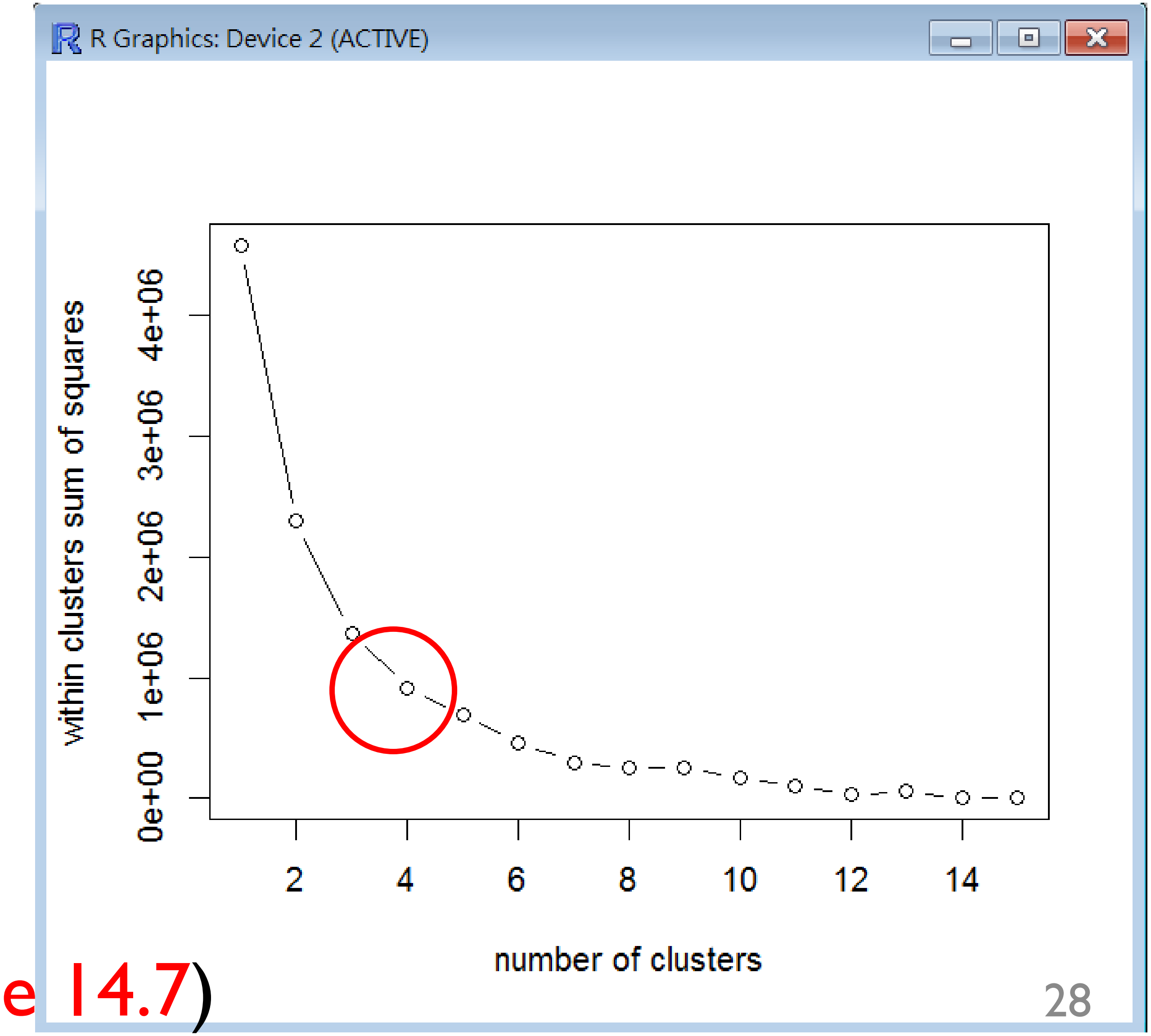
(RMD_example 14.6)

Determine number of clusters

- Plot the value of the clustering criterion against the number of clusters
- Large change of levels in the plot are taken as suggestive of a particular number of clusters

Determine number of clusters

Clustering criterion: within clusters sum of squares



(RMD_example 14.7)

Assess cluster fit

- Most often ignored
- Usually the cluster structure is rather unstable, at least at the bottom
- Can be VERY sensitive to noise and to outliers
- Cluster silhouettes and silhouette coefficient: how similar samples within a cluster are to samples in other clusters (composite separation and homogeneity) (Rousseeuw, 1987)

Silhouette

- If clustering samples
- Silhouette of sample *i* is defined as:

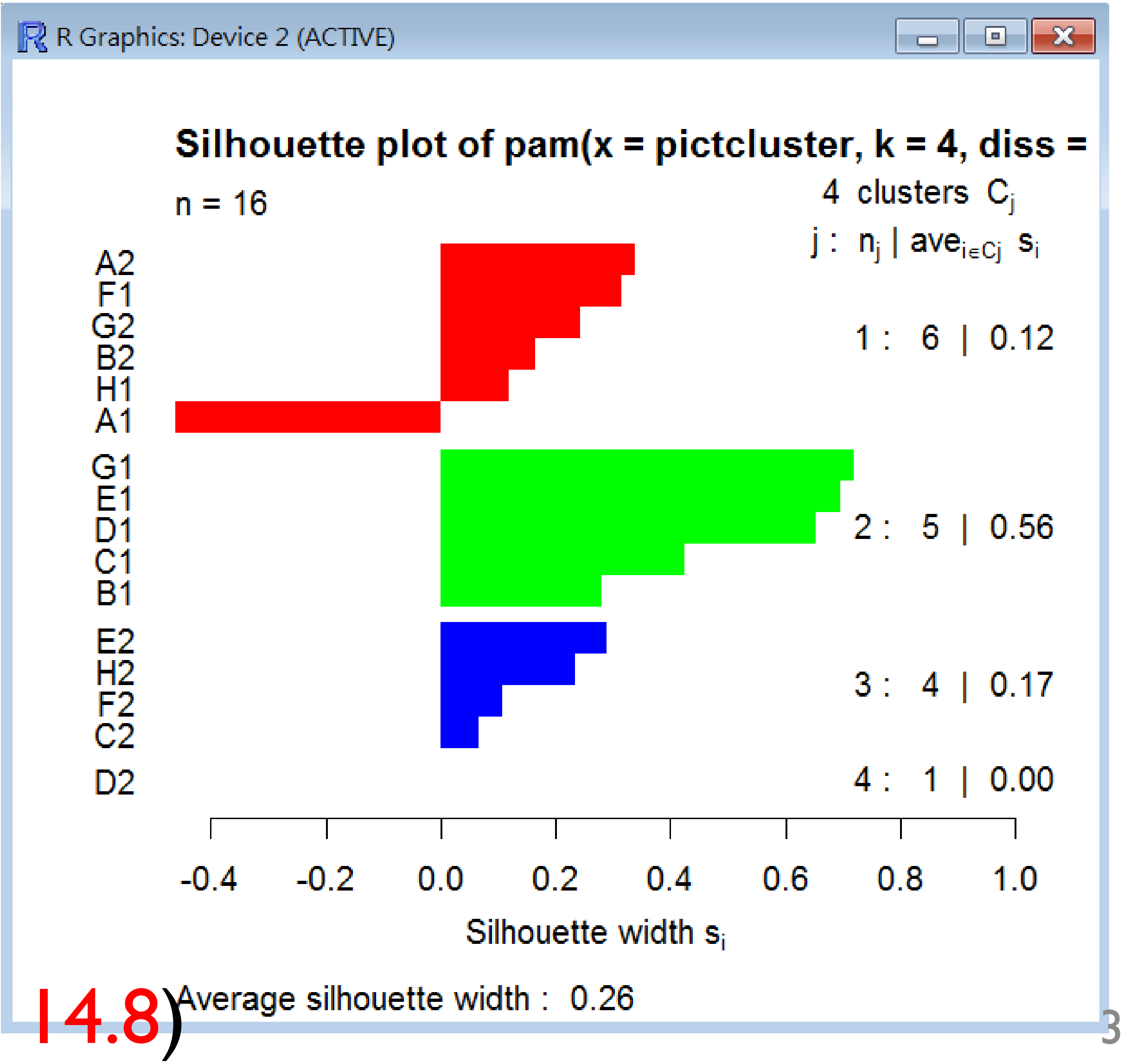
$$s(i) = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- a_i = average distance of sample *i* to other samples in same cluster
- b_i = average distance of sample *i* to samples in its nearest neighbor cluster

Silhouette plot

From

4-medoids



(RMD_example 14.8)

Some take-home points

- Clustering can be a useful exploratory tool.
- Cluster results are very sensitive to noise in the data.
- It is crucial to assess cluster structure to see how stable your result is.
- Different clustering approaches can give quite different results.
- For hierarchical clustering, interpretation is almost always subjective.

Multidimensional scaling

Multidimensional scaling (MDS)

- A method for displaying (transformed) multivariate data in low-dimensional space.
  - Similar method: plotting scores on the first two principal components or factors.
- Multidimensional scaling techniques deal with the following problem: for a set of observed similarities (or distances) between every pair of *N* samples, find a representation of samples in few dimensions such that inter-sample proximities "nearly match" the original similarities (or distances).

Graphical illustration of MDS

Original data:

$$\begin{array}{c|ccccc} \text{Sample\textbackslash Variable} & O_1 & O_2 & \cdots & O_p \\ \hline 1 & & & & \\ 2 & & & & \\ 3 & & \underset{\sim}{O} & & \\ \vdots & & & & \\ N & & & & \end{array} \Rightarrow \begin{array}{c|cccc} & 1 & 2 & 3 \cdots & N \\ \hline 1 & & & & \\ 2 & & & & \\ 3 & & \underset{\sim}{a} & & \\ \vdots & & & & \\ N & & & & \end{array}$$

distance matrix

MDS configuration:

$$\begin{array}{c|cccc} \text{Sample\textbackslash Variable} & X_1 & X_2 & \cdots & X_q \\ \hline 1 & & & & \\ 2 & & \underset{\sim}{X} & & \\ \vdots & & & & \\ N & & & & \end{array} \Rightarrow \begin{array}{c|ccc} & 1 & 2 \cdots & N \\ \hline 1 & & & \\ 2 & & \underset{\sim}{b} & \\ \vdots & & & \\ N & & & \end{array}$$

distance matrix

MDS finds the configuration that $\underset{\sim}{a}$ and $\underset{\sim}{b}$ are as closed as possible.

- Consequently, scaling techniques attempt to find configurations in $q \leq N - 1$ dimensions such that the match is as close as possible.

Glossary and steps for MDS

- The input target set of samples $\boldsymbol{O}$: a set of *N* samples among which the distances are to be studied. Only degree of relationship (distances) can be observed among the samples.
- The proximity of a pair of samples $(O_r, O_s), \delta_{rs}$: a distance or similarity measurement used to represent the relationship between two samples.
- The dimensionality of output configuration space q: a pre-specified number for the desired MDS dimensions.

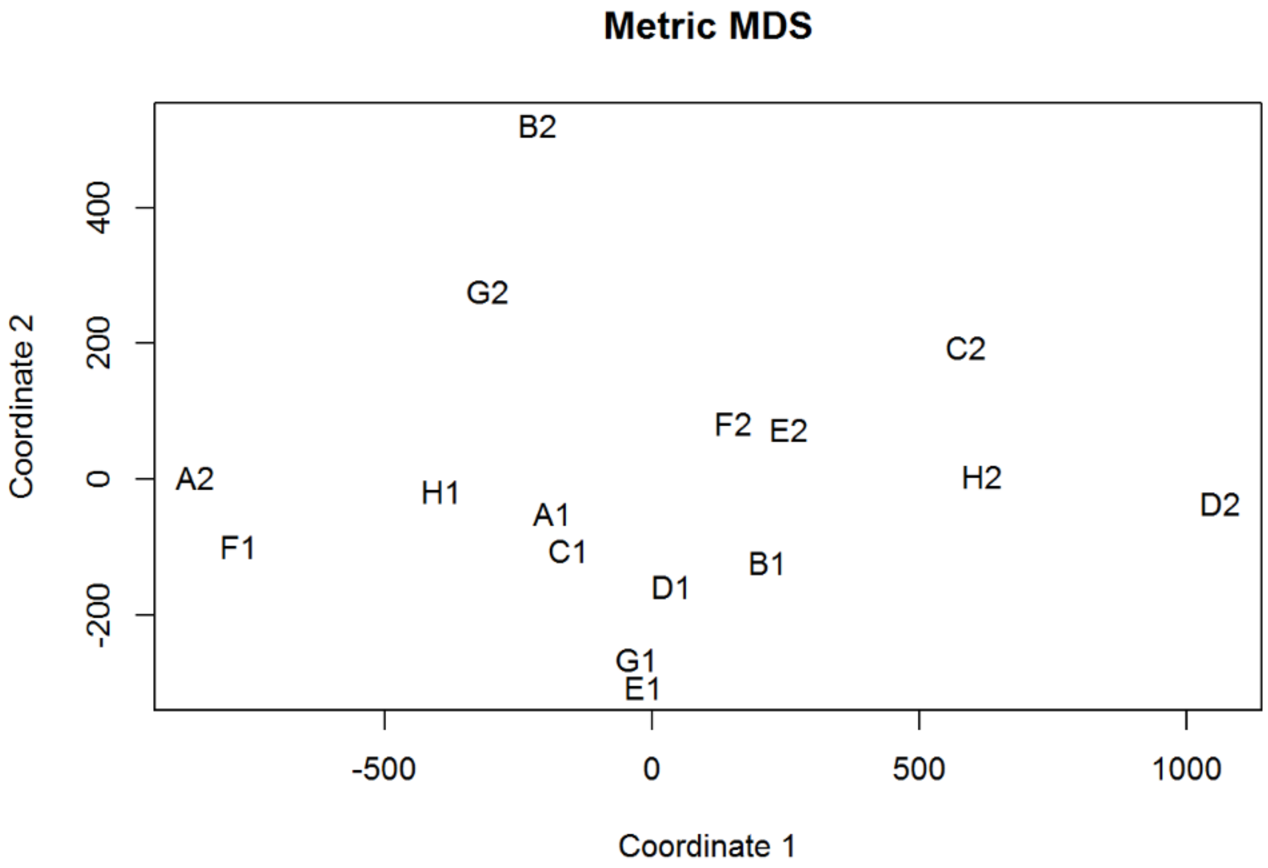
- The output configuration set $\boldsymbol{X} = \boldsymbol{X}(q)$: a set of *N* *q*-dimensional MDS solution configuration points, usually in Euclidean space, to represent the set of input objects $\boldsymbol{O}$.
- The distance $d_{rs} = d_{rs}(\boldsymbol{X})$ of a pair of configuration points (r, s) in $\boldsymbol{X}$: the distance between points representing samples *r* and *s* in the configuration space to represent δ_{rs} in the input set.

- The transformation function, $f\colon\ \delta_{rs} \rightarrow d_{rs}$: a function for specifying how the proximities should be related to the distances. Due to the existence of noise in the observed proximities, f is usually only used to map proximities to approximate distance.

  Only the type (ratio, interval, log, exp, ...) of f is pre-specified, the exact form (parameters) of f is part of the MDS solution.
- The disparity of a pair of samples (O_r, O_s), $\hat{d}_{rs} = f(\delta_{rs})$

- $\text{Stress}(q) = \sigma^2\left(d, \hat{d}\right) = \left\{\frac{\sum\sum_{r<s}(d_{rs} - \hat{d}_{rs})^2}{\sum\sum_{r<s} d_{rs}^2}\right\}^{1/2}$: a loss function for measuring the closeness of the mapping from proximities to distance.
- MDS procedure attempts to find an appropriate transformation f and a set of points in configuration space $\boldsymbol{X}$, such that the stress is as small as possible.

MDS for 照片分群

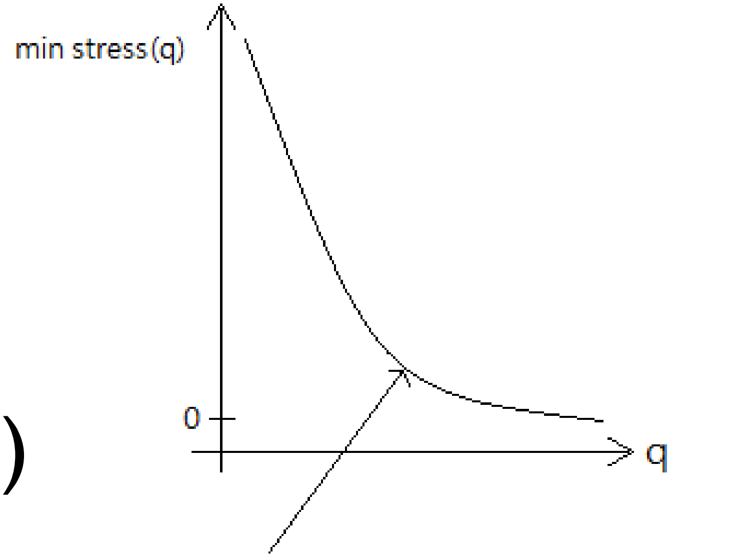


MDS q=2

(RMD_example 14.9)

Determine the *q*

- The Stress measure is a function of q. For each q, the configuration $\boldsymbol{X}$ leading to minimum Stress can be obtained. As q increases, minimum Stress will, within rounding error, decrease and will be zero for $q = N - 1$
- A plot of minimum Stress(q) versus q:



looking for an "elbow" in the plot for the best q