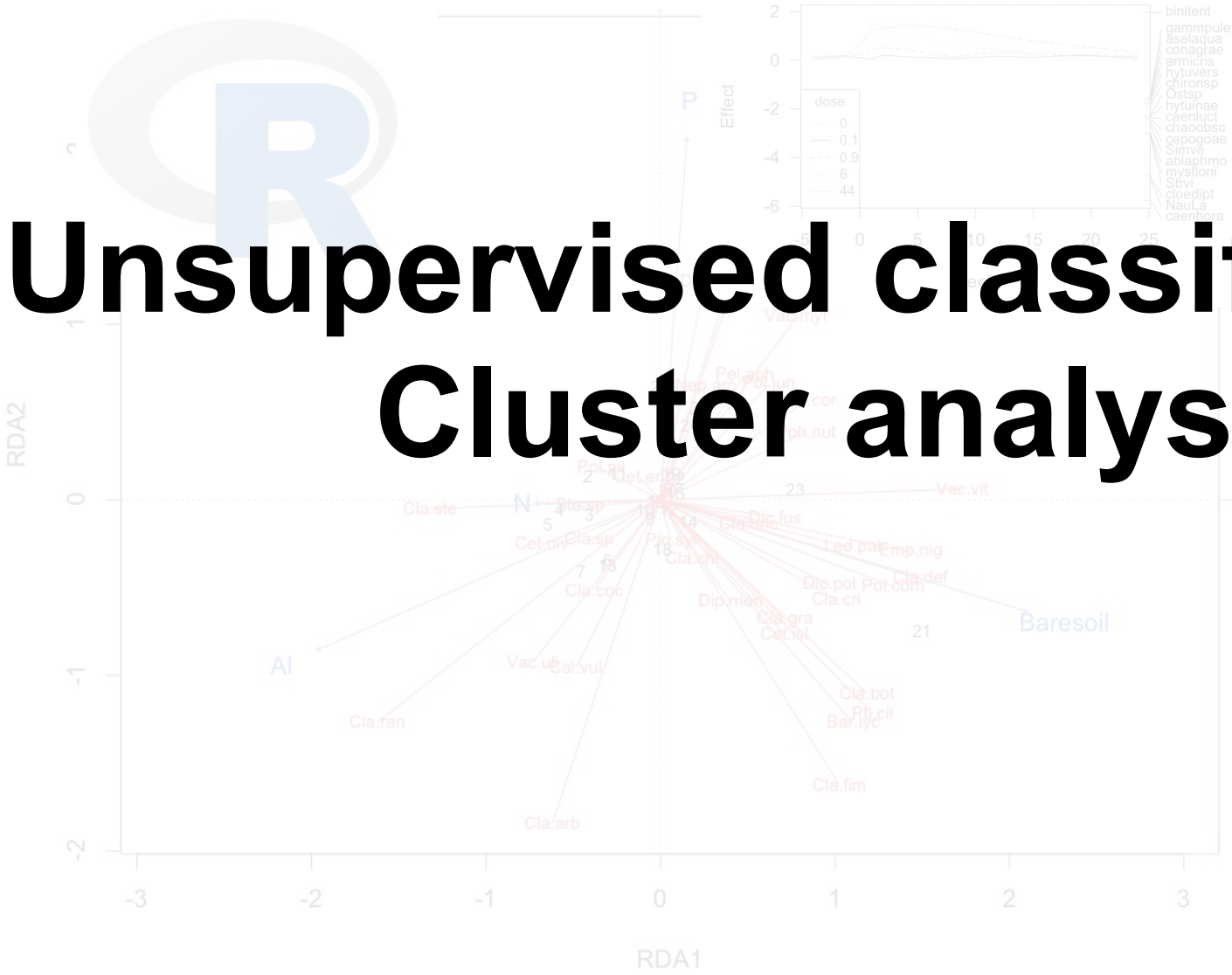
Tools for complex data analysis

University of Koblenz-Landau 2018/19



Unsupervised classification: Cluster analysis

Ralf B. Schäfer

Unsupervised classification: Cluster analysis

Contents

Learning targets

- Knowledge on the aims and methods of cluster analysis
- Understanding hierarchical and non-hierarchical cluster analysis

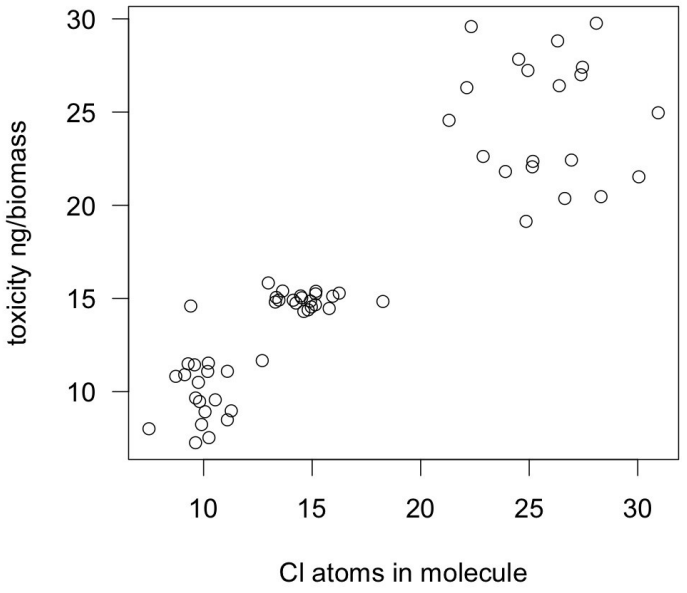
Learning targets and study questions

- Knowledge on the aims and methods of cluster analysis
  - What is the aim of cluster analysis?
  - What is the difference between hierarchical and non-hierarchical cluster analysis?
- Understanding hierarchical and non-hierarchical cluster analysis.
  - Outline the algorithms for hierarchical clustering and *k*-means.
  - Describe the calculation of distances between clusters for single, average and complete linkage. How does the choice of the method influence the interpretation of results?
  - Explain the analogy between *k*-means and ANOVA.
  - List cluster validity indices. What is the difference between external and internal validation?
  - Discuss limitations of cluster analysis.

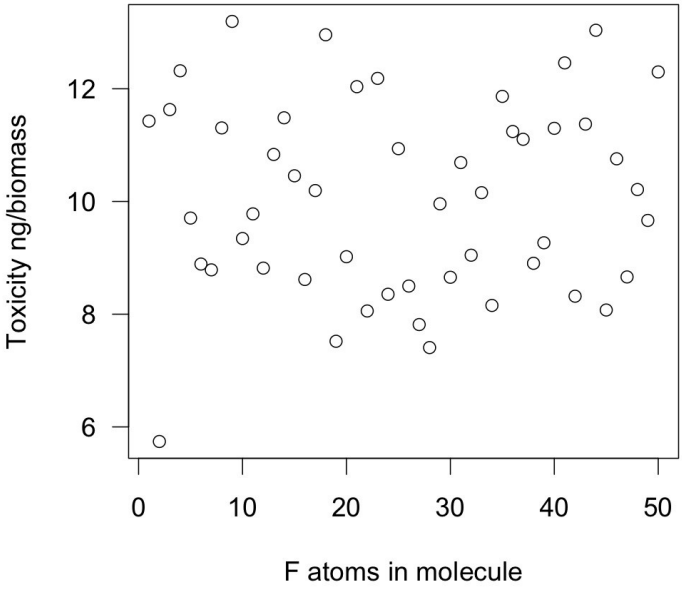
Unsupervised classification

- Group structure not known *a priori*
- Aim: identification of hidden structures and grouping of similar observations
- Methods include cluster analysis and self-organising maps

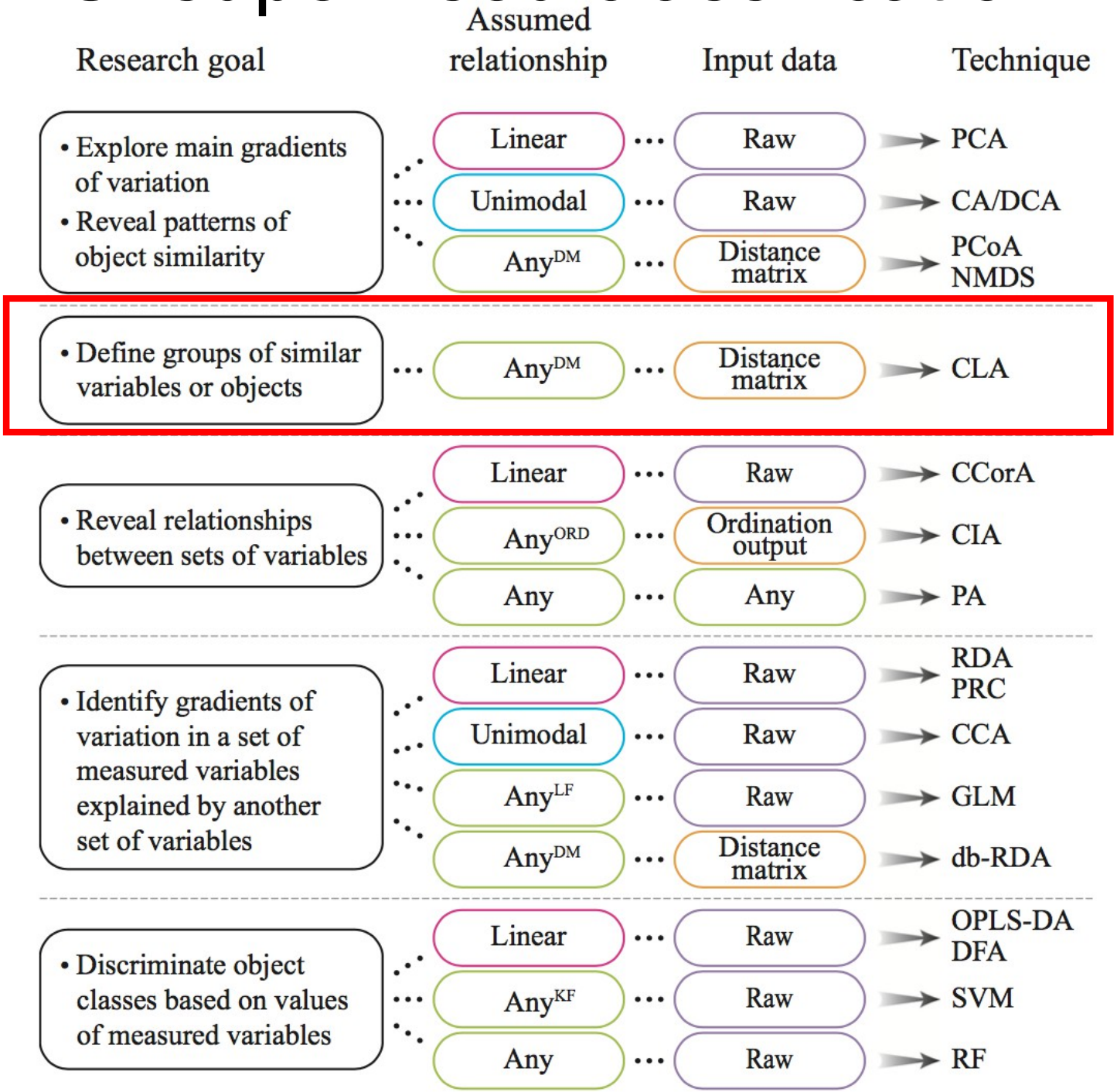
Identifiable groups



No groups identifiable



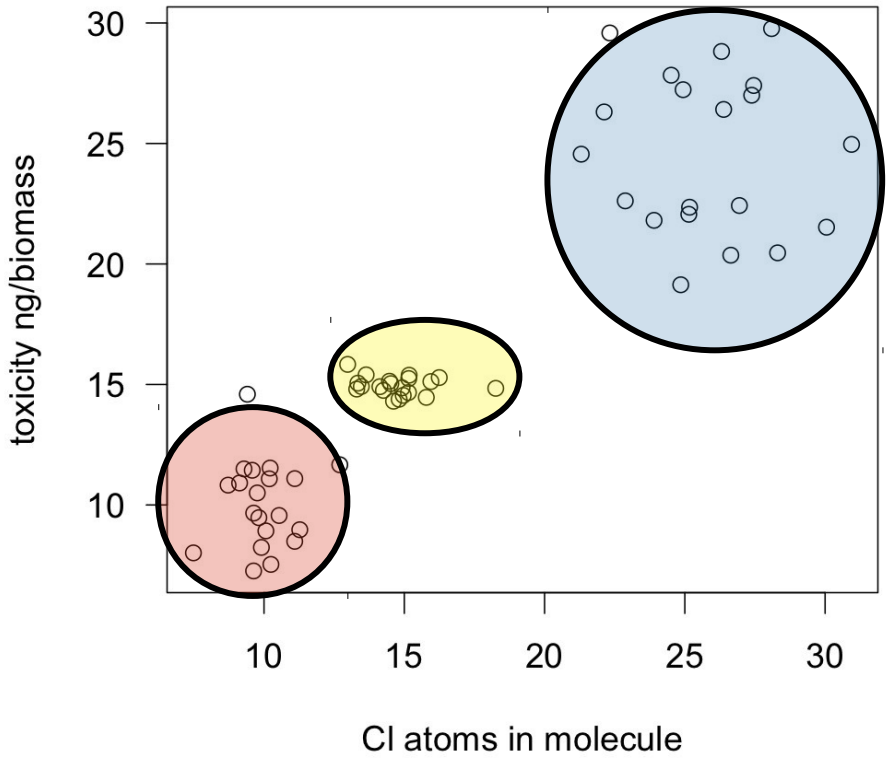
Unsupervised classification



Paliy & Shankar 2016 *Mol Ecol*:1032

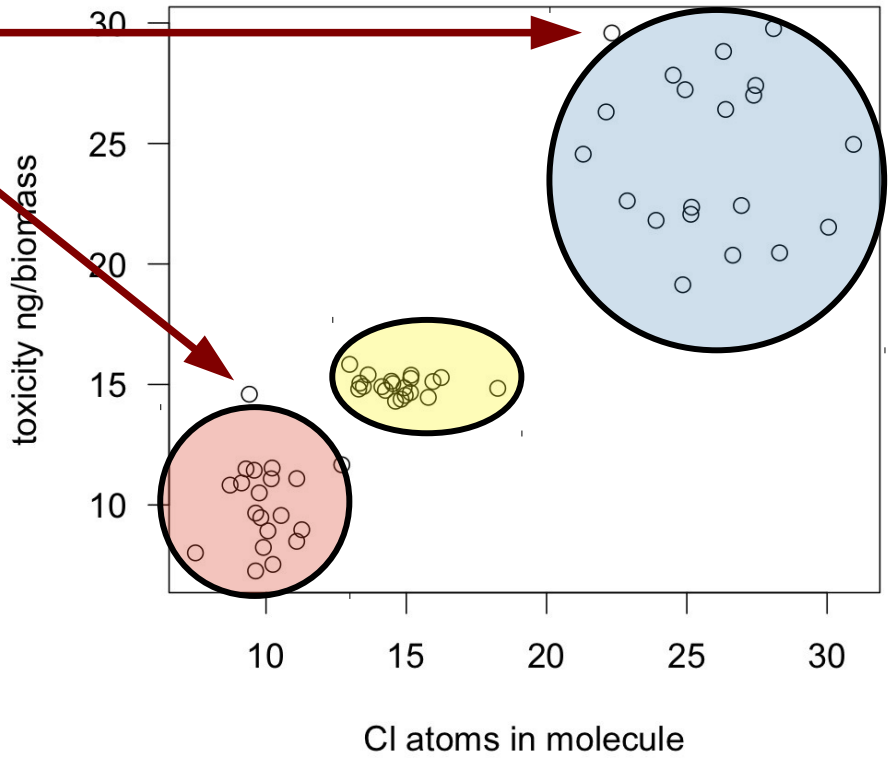
Cluster analysis: Application

- Identification of groups
- Data aggregation → Reduction of dimensions or „noise“
- Visualising similarity/distance of objects



Cluster analysis: Application

- Identification of groups
- Data aggregation → Reduction of dimensions or „noise“
- Visualising similarity/distance of objects
- Identification of outliers



Unsupervised classification: Cluster analysis

Contents

1. Introduction to unsupervised classification and cluster analysis
2. **Hierarchical agglomerative clustering**
3. Linkage methods for hierarchical clustering
4. Overview and *k*-means clustering
5. Cluster validity indices and number of clusters
6. Discussion of cluster analysis and further clustering techniques

Cluster analysis: Intro

How does clustering work?

Cluster analysis relies on similarity or distance matrix

1. maximise within-group similarity of objects in cluster

2. minimise between group similarity

➡ Results depend on similarity/distance measure

Cluster analysis: Intro

How does clustering work?

Cluster analysis relies on similarity or distance matrix

1. maximise within-group similarity of objects in cluster

2. minimise between group similarity

 Results depend on similarity/distance measure

Hierarchical Agglomerative Clustering

- Widely used
- Starts with all objects as single clusters
- Objects merged into joint cluster based on distance or similarity

Hierarchical agglomerative clustering

1. Search for shortest distance between pairs of objects
→ merge into cluster

	1	2	3	4	5	6
1	0	1.414	2.000	4.472	5.657	6.708
2		0	1.414	3.162	4.243	5.385
3			0	4.000	4.472	5.000
4				0	2.000	4.123
5					0	2.236
6						0

Izenman 2008:415f

Hierarchical agglomerative clustering

1. Search for shortest distance between pairs of objects
→ merge into cluster

2. Re-calculation of distances and repetition of step 1

	1	2	3	4	5	6
1	0	1.414	2.000	4.472	5.657	6.708
2		0	1.414	3.162	4.243	5.385
3			0	4.000	4.472	5.000
4				0	2.000	4.123
5					0	2.236
6						0

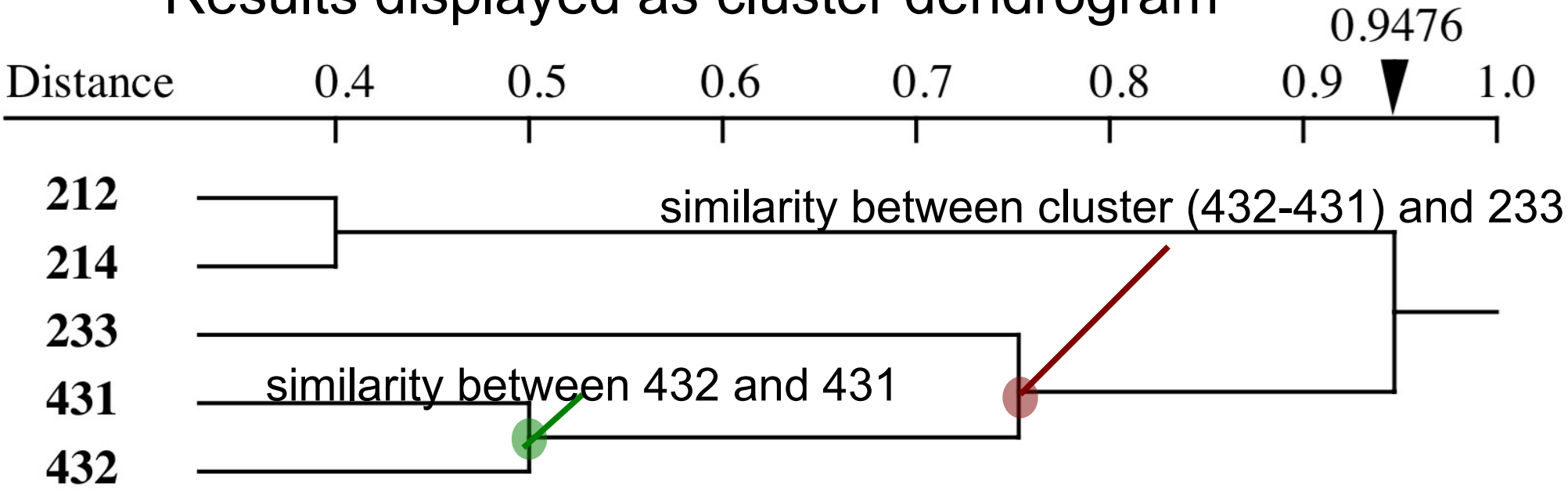
	1	23	4	5	6
1	0	1.414	4.472	5.657	6.708
23		0	3.162	4.243	5.000
4			0	2.000	4.123
5				0	2.236
6					0

Ends when all objects merged into one cluster

Izenman 2008:415f

Hierarchical agglomerative clustering

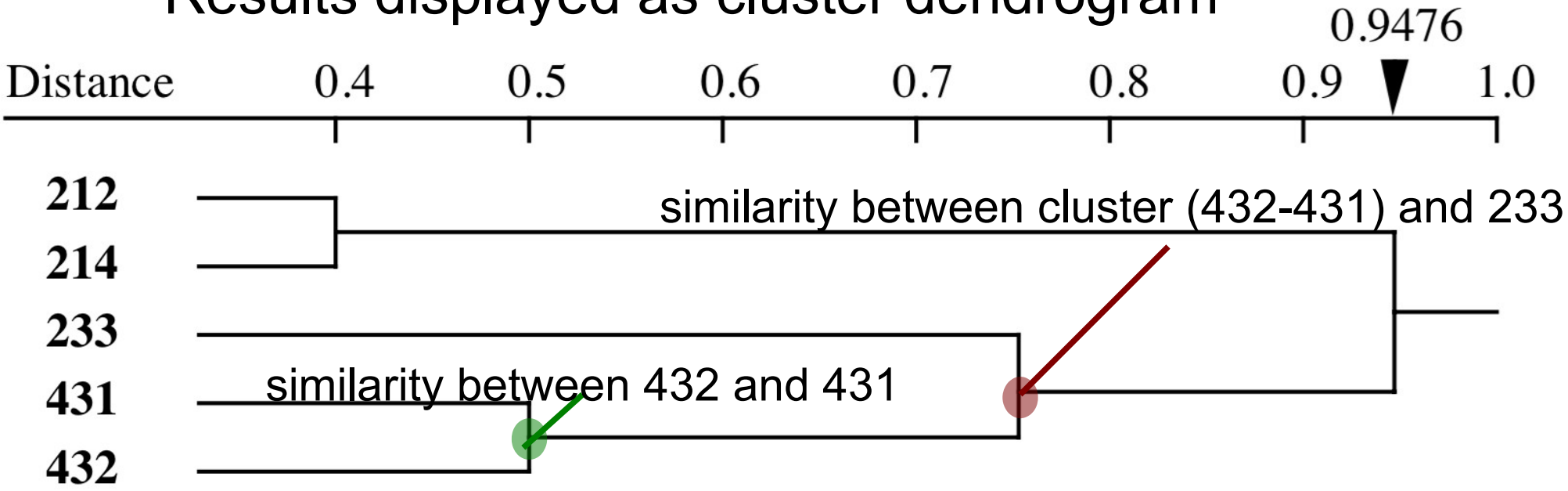
Results displayed as cluster dendrogram



Legendre & Legendre 2012: 356

Hierarchical agglomerative clustering

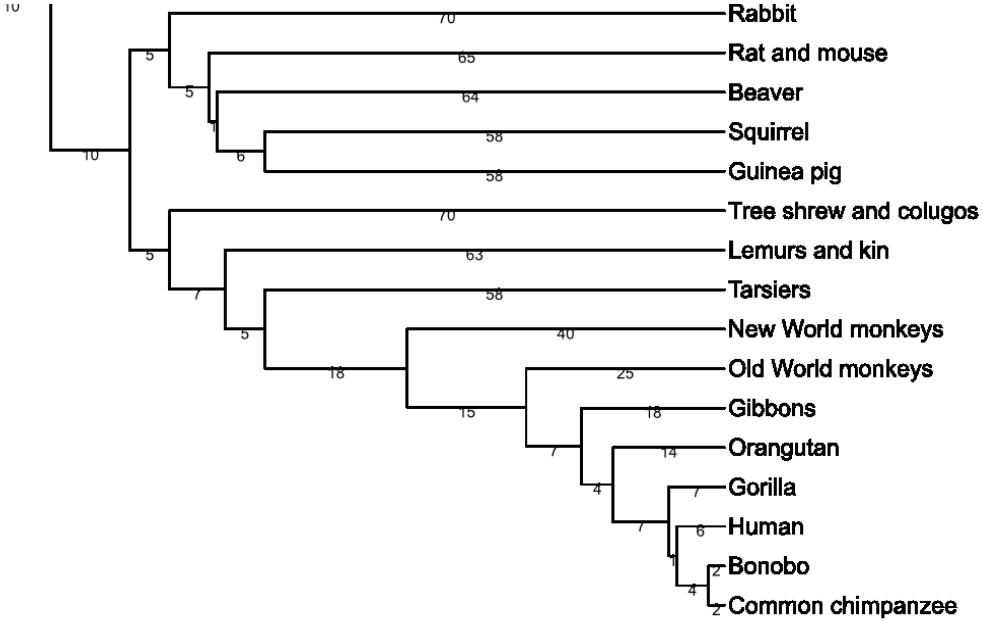
Results displayed as cluster dendrogram



Legendre & Legendre 2012: 356

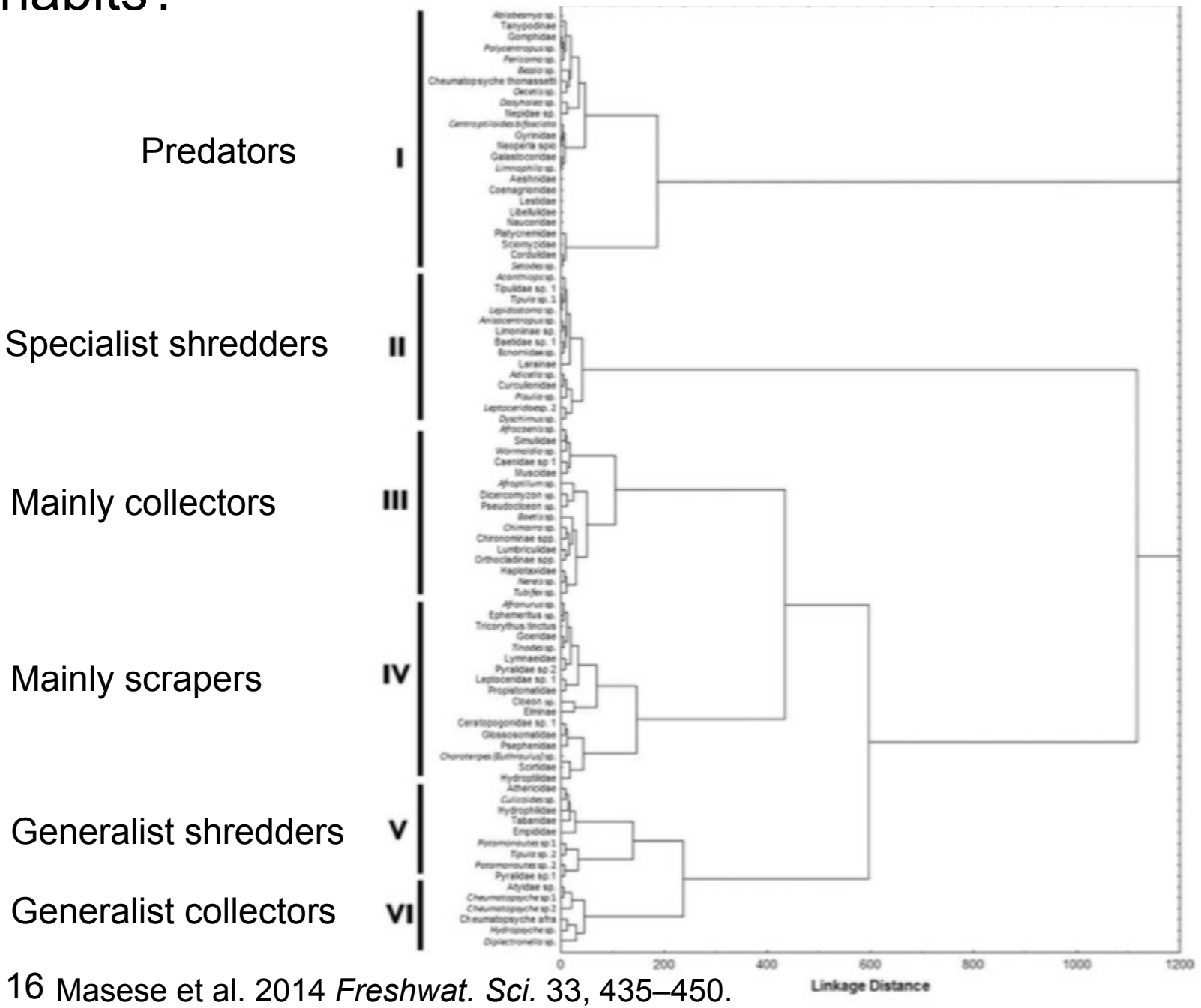
Note the similarity in interpretation to a phylogenetic tree!

For example, humans are not directly related to rabbits, but they share a common ancestor



https://commons.wikimedia.org/

Example: What are the functional feeding groups in Kenyan freshwater invertebrates? Which species have similar feeding habits?



16 Masese et al. 2014 *Freshwat. Sci.* 33, 435–450.

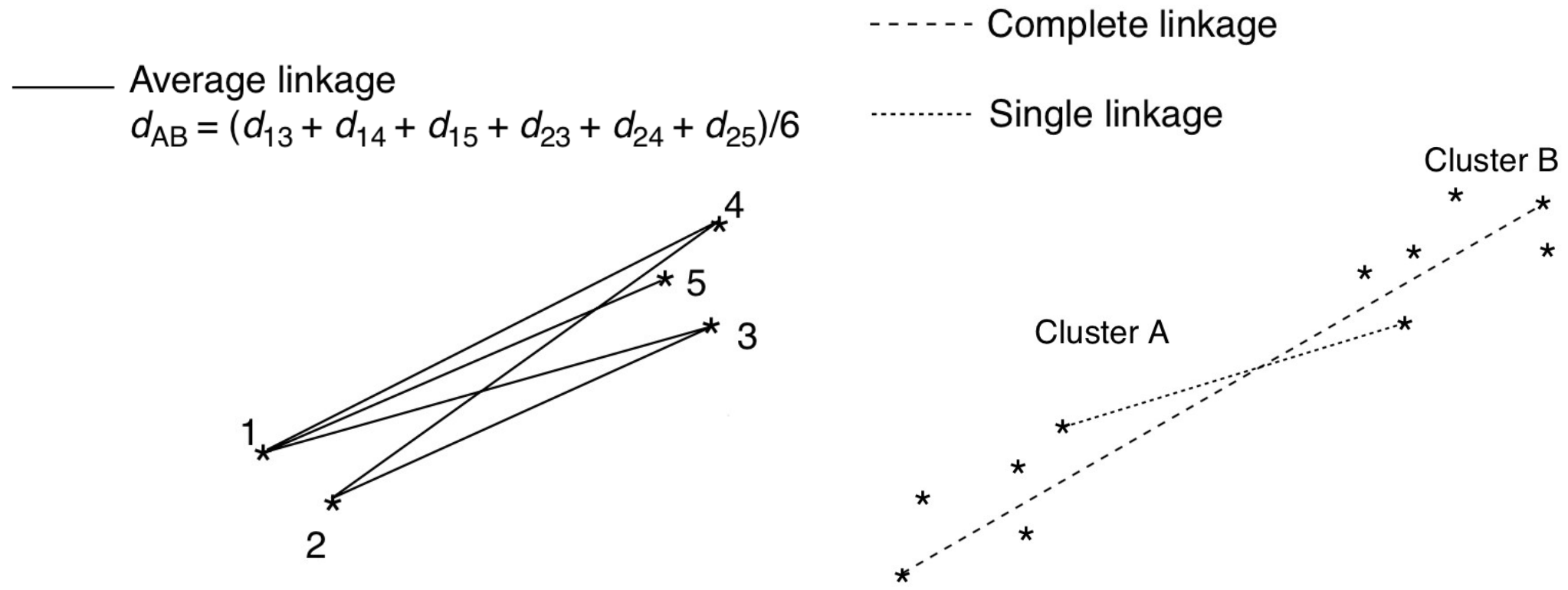
Unsupervised classification: Cluster analysis

Contents

1. Introduction to unsupervised classification and cluster analysis
2. Hierarchical agglomerative clustering
3. **Linkage methods for hierarchical clustering**
4. Overview and *k*-means clustering
5. Cluster validity indices and number of clusters
6. Discussion of cluster analysis and further clustering techniques

Methods for calculating cluster distances

Examples of methods for calculating between-cluster distances



Everitt 2011: 77

Methods for calculating cluster distances

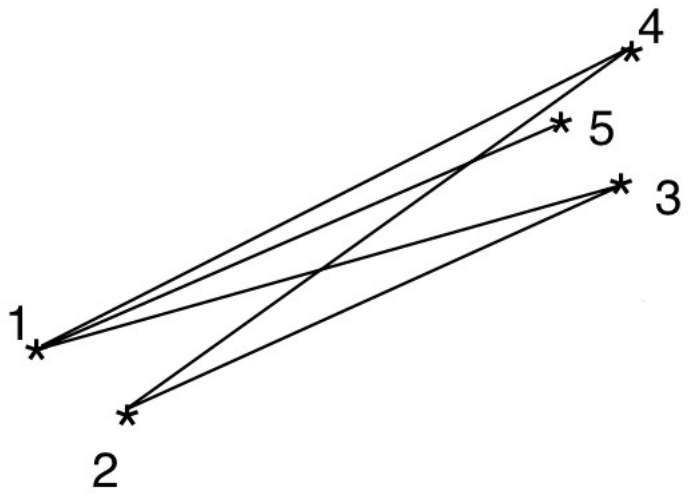
Examples of methods for calculating between-cluster distances

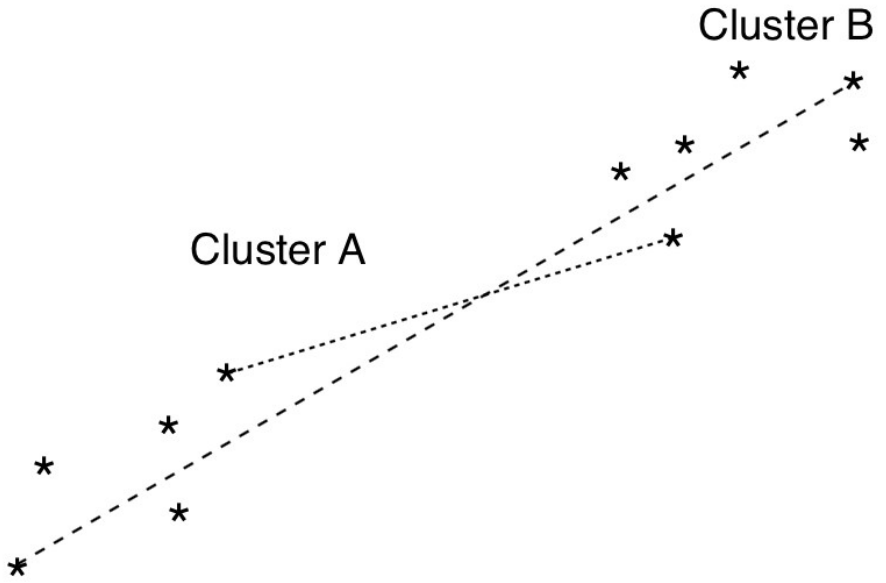
Complete linkage

Average linkage

$d_{AB} = (d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25})/6$

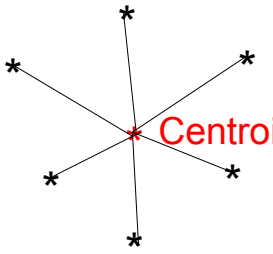
Single linkage





Wards method

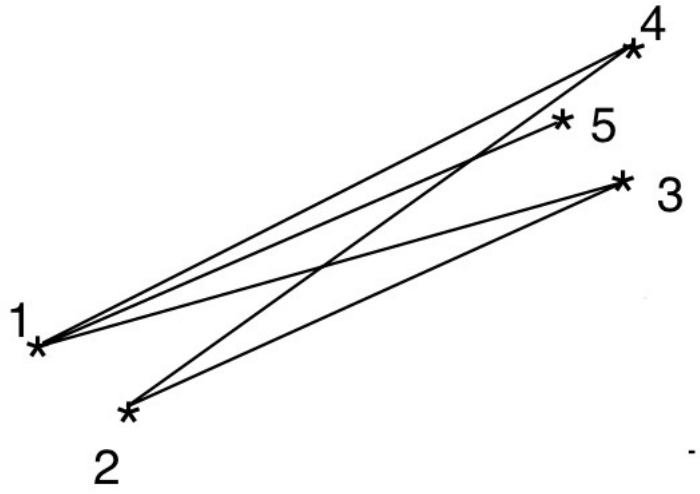
Minimize Within-cluster Sum of Squares (SSW)

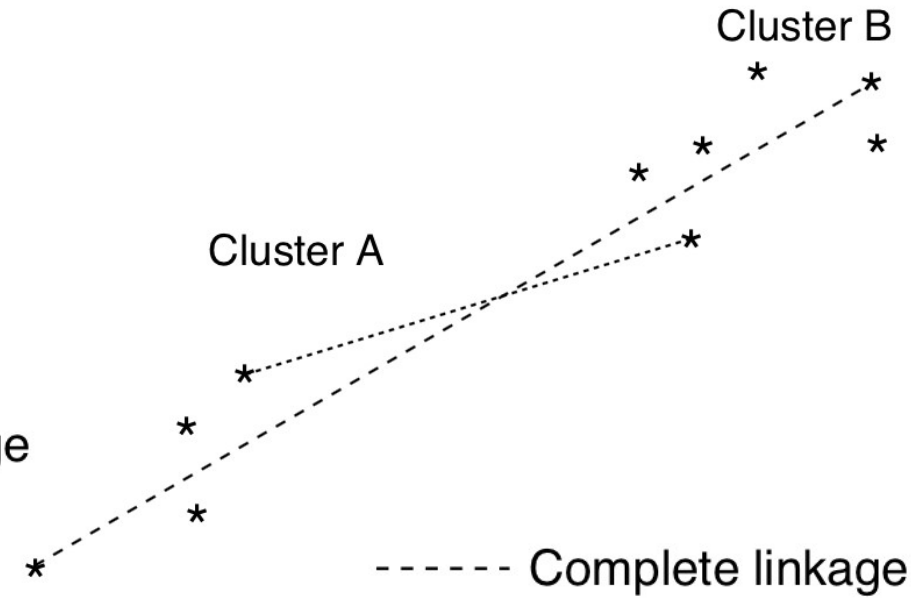


Centroid (calculated from cluster members)

Everitt 2011: 77

Influence of methods on clustering





Single linkage

Average linkage

$d_{AB} = (d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25})/6$

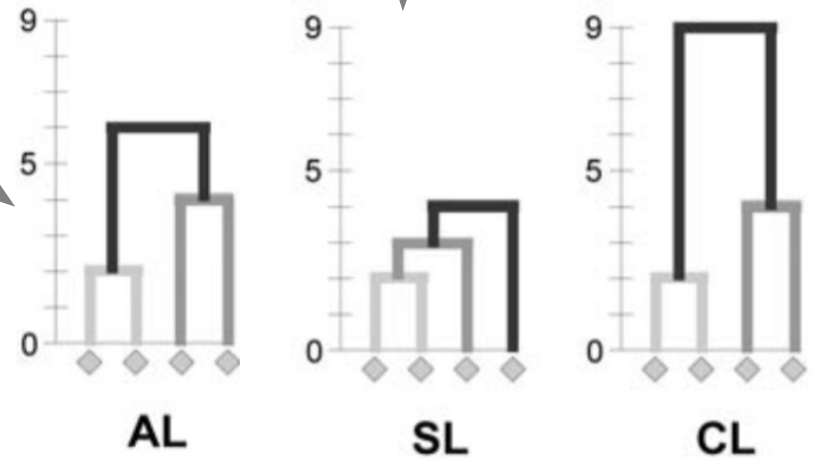
Complete linkage

Methods can be classified as

Space Conserving

Space Contracting

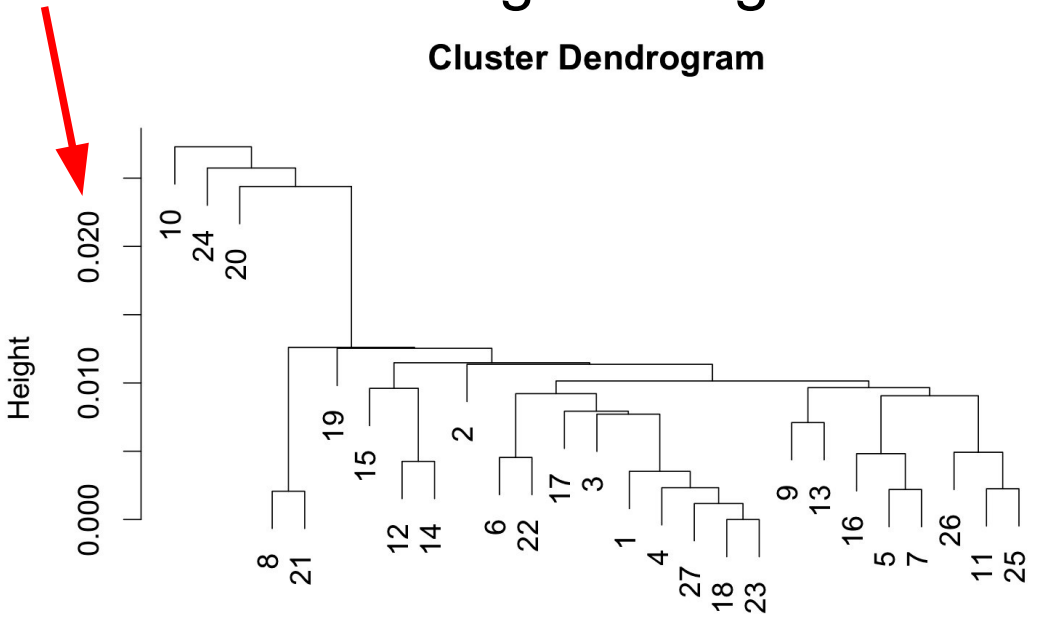
Space Dilating



Wildi 2010: 63

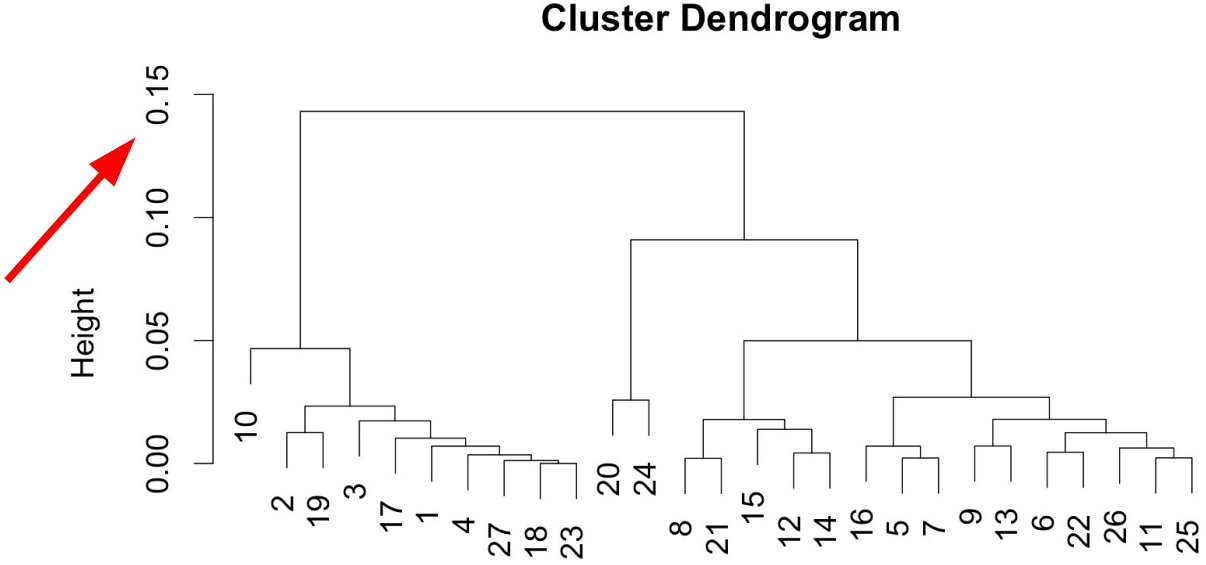
Example: Complete and single linkage

Single linkage



Large and close clusters

Complete linkage



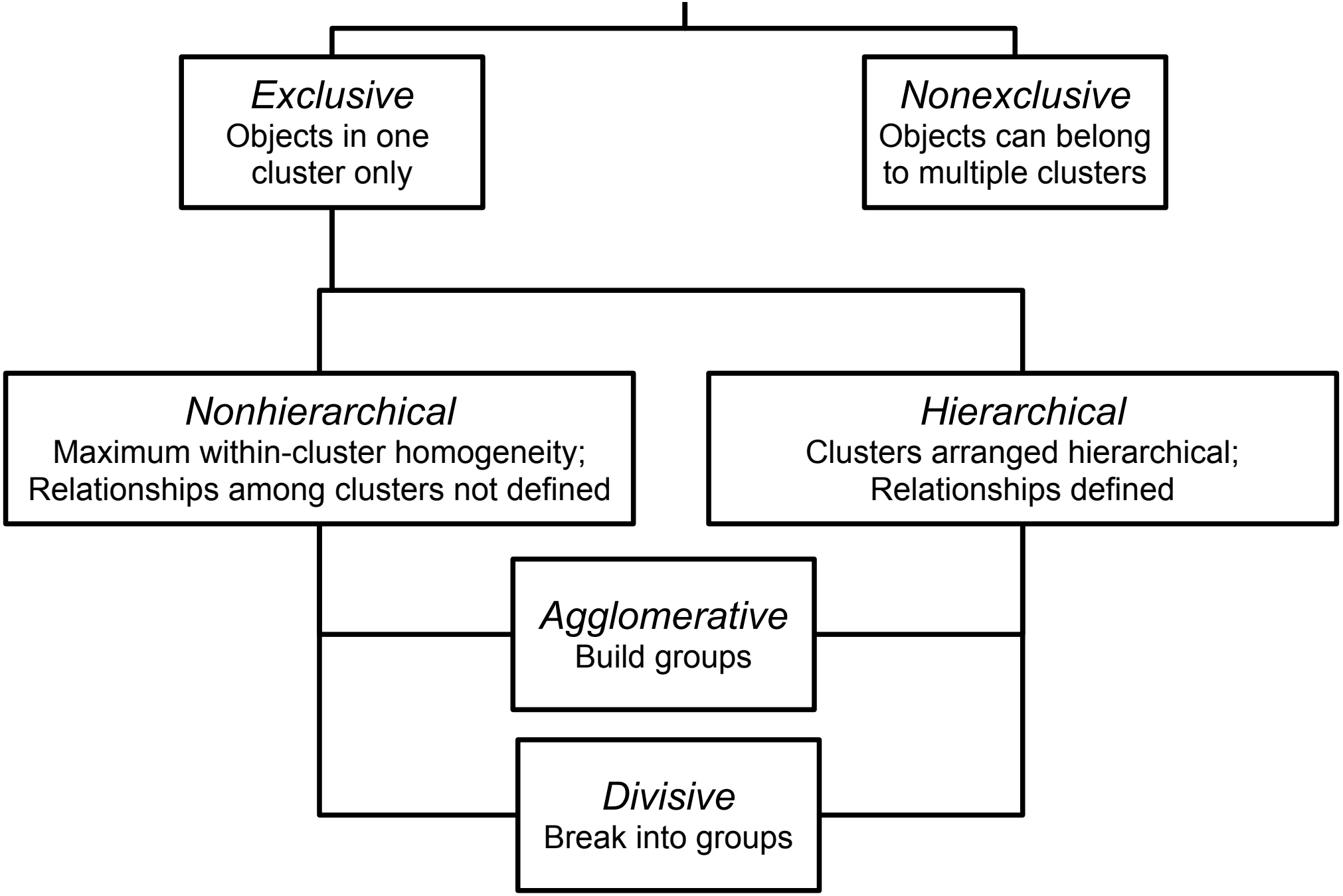
Small and distant clusters

Unsupervised classification: Cluster analysis

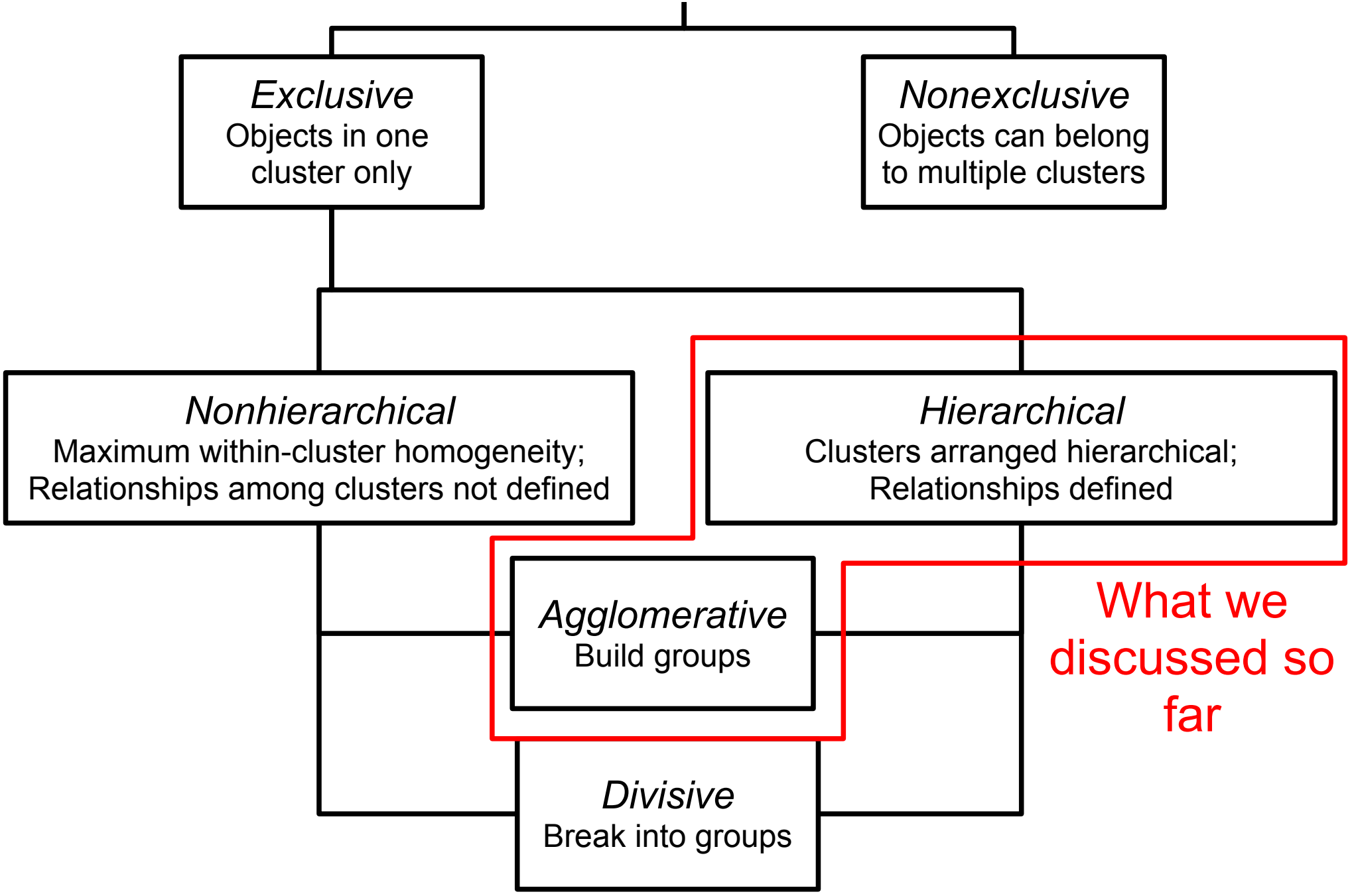
Contents

1. Introduction to unsupervised classification and cluster analysis
2. Hierarchical agglomerative clustering
3. Linkage methods for hierarchical clustering
4. **Overview and *k*-means clustering**
5. Cluster validity indices and number of clusters
6. Discussion of cluster analysis and further clustering techniques

Overview cluster analysis



Overview cluster analysis



Non-hierarchical cluster analysis: k-means clustering

Assigns objects to a pre-defined number of clusters k

Problem:

n	k	**Number of possible partitions**
15	3	2,375,101
20	4	45,232,115,901
25	8	690,223,721,118,368,580
100	5	10^{68}

Calculation of all possible partitions becomes unsurmountable even for relatively small sample sizes (n)

Algorithm required

Everitt & Hothorn 2011: 175

k-means clustering

Algorithm

1. Partition objects into *k* groups
2. Move objects and calculate change
3. Choose best solution
4. Repeat 2-3 until cluster criterion does not improve

Everitt & Hothorn 2011: 175

k-means clustering

Algorithm

1. Partition objects into *k* groups
2. Move objects and calculate change
3. Choose best solution
4. Repeat 2-3 until cluster criterion does not improve

Which criterion is used to assess best solution?

Minimisation of Within-cluster SSQ (SSW)
→ Note analogy to Wards method and ANOVA

k-means clustering implicitly relies on euclidean distances
→ Ecological data require data preparation
(see Borcard et al. 2011: 80 for details)

Everitt & Hothorn 2011: 175

Unsupervised classification: Cluster analysis

Contents

1. Introduction to unsupervised classification and cluster analysis
2. Hierarchical agglomerative clustering
3. Linkage methods for hierarchical clustering
4. Overview and *k*-means clustering
5. **Cluster validity indices and number of clusters**
6. Discussion of cluster analysis and further clustering techniques

Cluster validity indices (CVIs)

- How many clusters?
- Many criteria, differing scopes, e.g. homogeneity, separation, uniformity, representation and stability
  → Do not overinterpret individual criteria

Cluster validity indices (CVIs)

- How many clusters?
- Many criteria, differing scopes, e.g. homogeneity, separation, uniformity, representation and stability
  → Do not overinterpret individual criteria

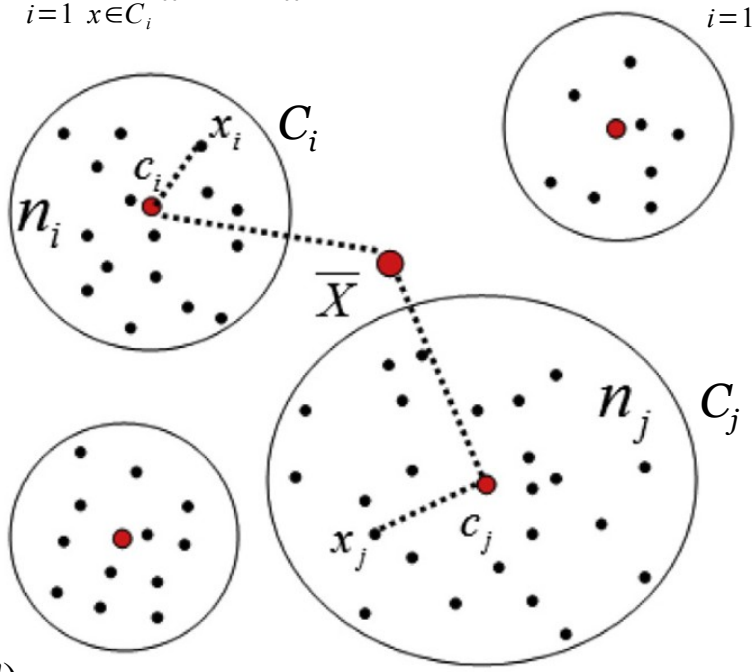
Selected internal CVIs:

Within-cluster SSQ

$$SSW = \sum_{i=1}^{M} \sum_{x \in C_i} \|x - c_i\|^2$$

Between-cluster SSQ

$$SSB = \sum_{i=1}^{M} n_i \|c_i - \overline{X}\|^2$$



Caliniski-Harabsz

$$CH = \frac{SSB/(M-1)}{SSW/(N-M)}$$

with

N = no. of observations

M = no. of clusters

x_i = observation in C_i

n_i = no. of observations in cluster C_i

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$ ('overall center')

c_i = centroid of cluster C_i ('cluster center')

Modified from Zhao & Fränti 2014

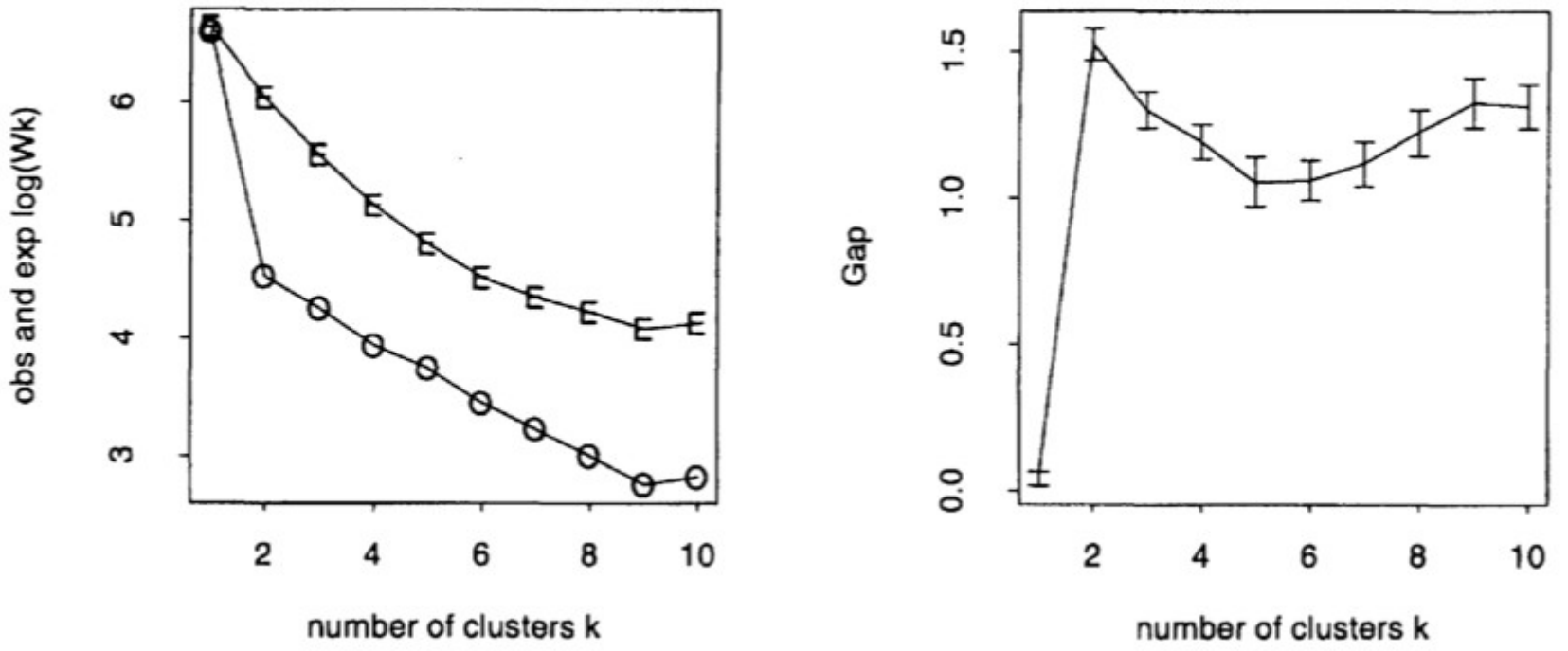
Cluster validity indices (CVIs)

Selected internal CVIs:

GAP index

$$GAP_n(k) = E_n^*\{\log(SSW_k)\} - \log(SSW_k)$$

- Determines uniform reference distribution E^* via bootstrapping using range of original data for same sample size n
- Optimal no. of clusters *k:* maximum difference in cluster criterion (*SSW*)



Tibshirani *et al.* 2001

Cluster validity indices (CVIs)

Selected internal CVIs:

Silhouette width *s*

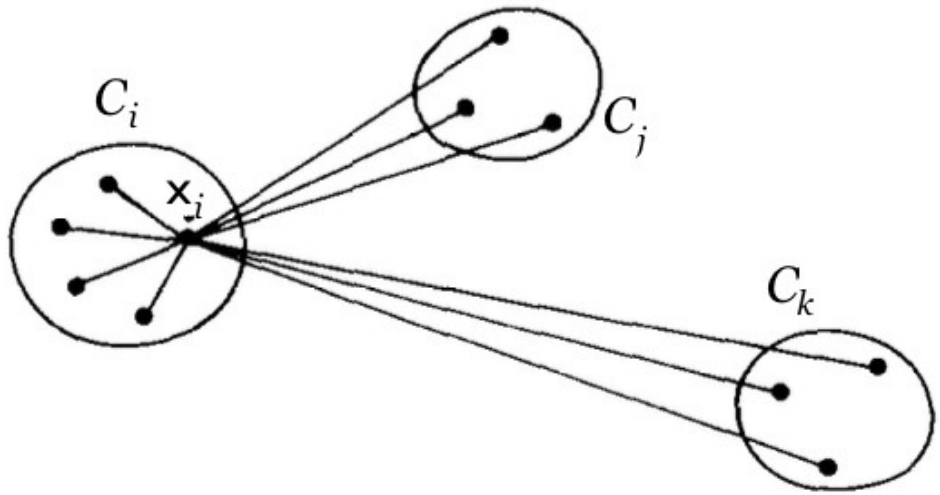
$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(b(x_i), a(x_i))}$$

with

$$b(x_i) = \frac{1}{n_j} \sum_{b \in C_j} dist(b, x_i) \quad \text{('average dissimilarity of } x_i \text{ to all other objects of cluster } C_j\text{')}$$

$$a(x_i) = \frac{1}{n_i} \sum_{a \in C_i} dist(a, x_i) \quad \text{('average dissimilarity of } x_i \text{ to all other objects of cluster } C_i\text{')}$$

Cluster C_j is the nearest neighbour to C_i



Modified from Rousseeuw 1987

Cluster validity indices (CVIs)

Selected internal CVIs:

Silhouette width *s*

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(b(x_i), a(x_i))}$$

with

$$b(x_i) = \frac{1}{n_j} \sum_{b \in C_j} dist(b, x_i) \quad \text{('average dissimilarity of } x_i \text{ to all other objects of cluster } C_j\text{')}$$

$$a(x_i) = \frac{1}{n_i} \sum_{a \in C_i} dist(a, x_i) \quad \text{('average dissimilarity of } x_i \text{ to all other objects of cluster } C_i\text{')}$$

Evaluation of individual *x*

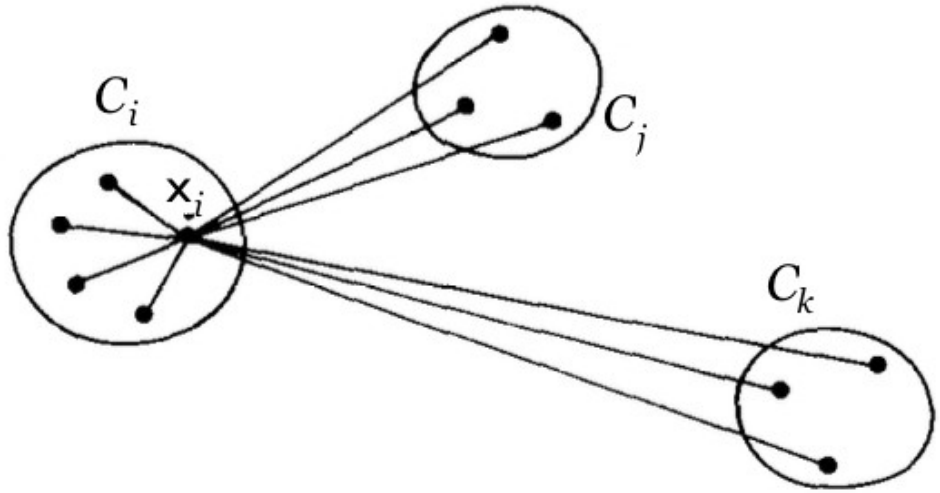
$-1 \le s(x_i) \le 1$

$s(x_i) \rightarrow 1 \Rightarrow$ well classified

$s(x_i) \rightarrow -1 \Rightarrow$ misclassified

$s(x_i) \approx 0 \Rightarrow$ unclear

Cluster C_j is the nearest neighbour to C_i



Modified from Rousseeuw 1987

Cluster validity indices (CVIs)

Selected internal CVIs:

Silhouette width *s*

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(b(x_i), a(x_i))}$$

with

$$b(x_i) = \frac{1}{n_j} \sum_{b \in C_j} dist(b, x_i)$$ ('average dissimilarity of x_i to all other objects of cluster C_j')

$$a(x_i) = \frac{1}{n_i} \sum_{a \in C_i} dist(a, x_i)$$ ('average dissimilarity of x_i to all other objects of cluster C_i')

Cluster C_j is the nearest neighbour to C_i

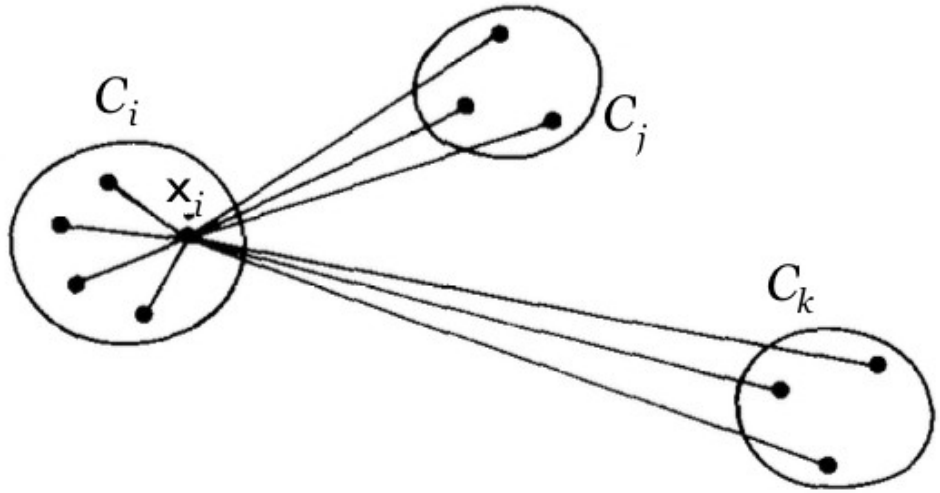
Evaluation of individual x

$-1 \le s(x_i) \le 1$

$s(x_i) \rightarrow 1 \Rightarrow$ well classified

$s(x_i) \rightarrow -1 \Rightarrow$ misclassified

$s(x_i) \approx 0 \Rightarrow$ unclear



Evaluation of cluster solution

$$\overline{S} = \frac{1}{N} \sum_{l=1}^{N} s(x_l)$$

('average silhouette width over the N observations')

$\overline{S} > 0.5$ suggests reasonable clustering

$\overline{S} < 0.2$ suggests lack of cluster structure

Modified from Rousseeuw 1987

Cluster validity indices (CVIs)

Selected internal CVIs:

Cluster stability via bootstrapping

- Draws *B* bootstrapped samples from original data
- Computes clustering for original data and bootstrapped data
- Calculates Jaccard index to compare each clustering for *B* with the clustering for the original data. Mean of Jaccard index used to evaluate stability (> 0.75 good stability)

Cluster validity indices (CVIs)

Selected internal CVIs:

Cluster stability via bootstrapping

- Draws *B* bootstrapped samples from original data
- Computes clustering for original data and bootstrapped data
- Calculates Jaccard index to compare each clustering for *B* with the clustering for the original data. Mean of Jaccard index used to evaluate stability (> 0.75 good stability)

Selected external CVI:

Rand index

- Computes pairs of true (a) and false (b) positives, true (d) and false (c) Negatives
- Same formula as Simple matching coefficient
- Adjusted Rand index: adjusted to yield 0 for two random partitions

		Cluster results	
		Same cluster	Different cluster
External data	Same class	a	b
	Different class	c	d

$$\mathrm{RI} = \frac{a+d}{a+b+c+d}$$

Unsupervised classification: Cluster analysis

Contents

1. Introduction to unsupervised classification and cluster analysis
2. Hierarchical agglomerative clustering
3. Linkage methods for hierarchical clustering
4. Overview and *k*-means clustering
5. Cluster validity indices and number of clusters
6. **Discussion of cluster analysis and further clustering techniques**

Critical discussion of cluster analysis

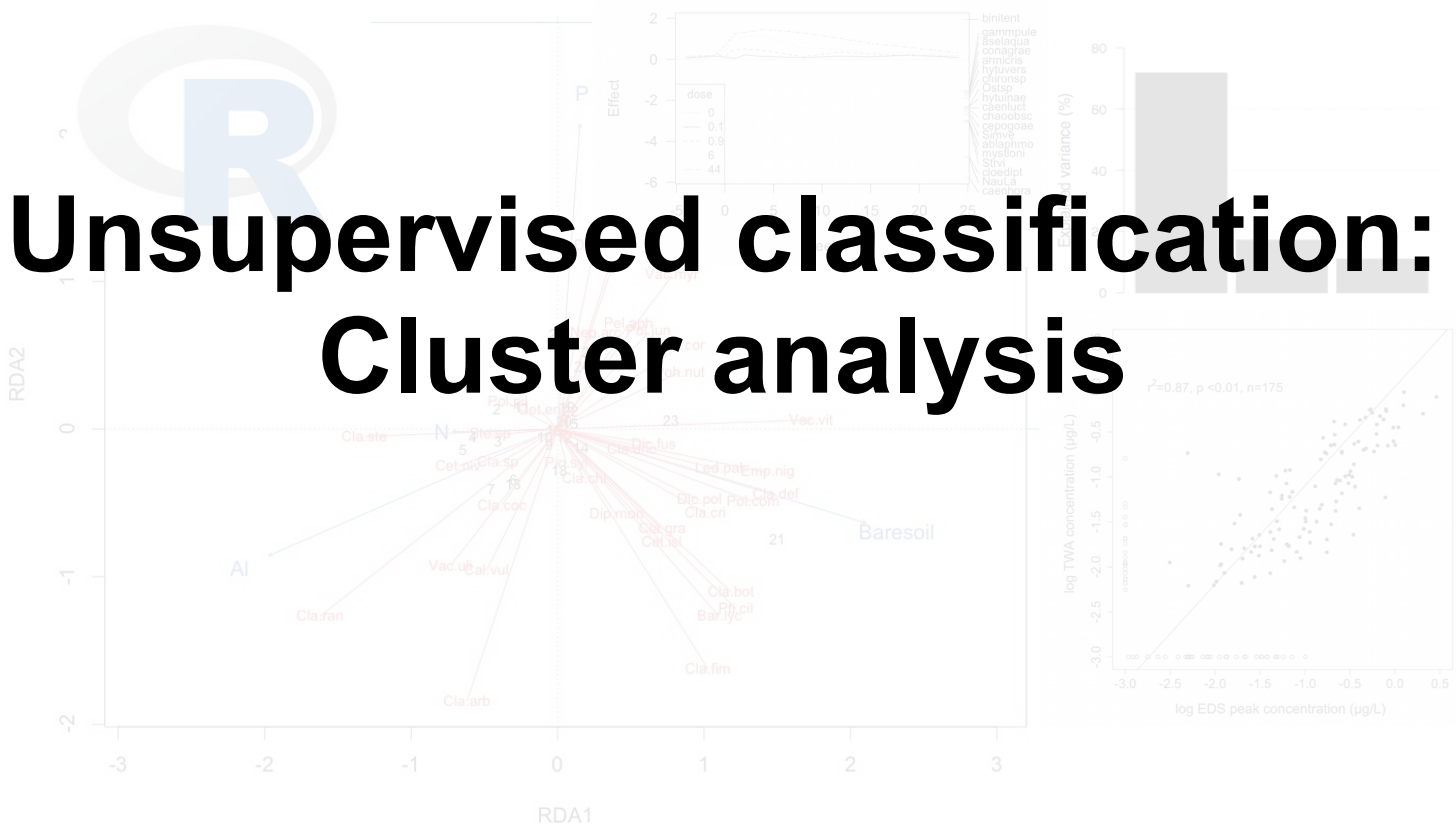
- Lack of formalisation, many choices (particularly in hierarchical clustering) that influence outcomes
  → results can be ambiguous, rather exploratory technique
- Tendency for specific (spherical) clusters
- *k*-means or hierarchical clustering (HC)?
  - *k* (or narrow range of *k*) has been fixed or is known
    → *k*-means
  - Dendrogram and cluster steps are of interest → HC
  - HC is more flexible (distance measure, clustering method), *k*-means restricted to euclidean distance and SSQs
  - large data sets → *k*-means (more efficient)

Further techniques in Cluster analysis

- Partitioning around medoids: similar to *k*-means but is non-euclidean, open to alternative distance metrics
  `pam() {cluster}`
- Non-hierarchical clustering for large data sets
  `clara() {cluster}`
- Model-based clustering: Estimates model parameters from data, assumes specific cluster structure (spherical, diagonal, ellipsoid)
  `Mclust() {mclust}`
- Clustering for non-spherical shapes
  `dbscan() {fpc}`
- Variable clustering: Useful to identify multicollinearity and surrogate variables
  `varclus() {Hmisc}` `hclustvar() {ClustOfVar}`

Tools for complex data analysis

University of Koblenz-Landau 2018/19



Unsupervised classification: Cluster analysis

Ralf B. Schäfer

These slides and notes complement the lecture with exercises “Tools for complex data analysis” for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R demonstration associated with the lecture.

Unsupervised classification: Cluster analysis

Contents

2

Learning targets

- Knowledge on the aims and methods of cluster analysis
- Understanding hierarchical and non-hierarchical cluster analysis

3

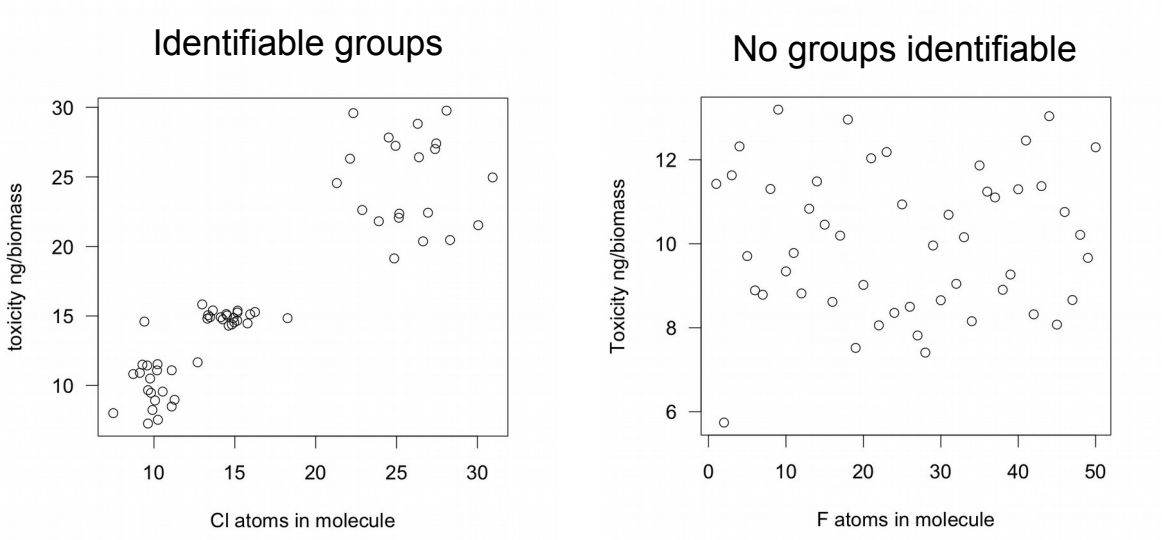
Learning targets and study questions

- Knowledge on the aims and methods of cluster analysis
  - What is the aim of cluster analysis?
  - What is the difference between hierarchical and non-hierarchical cluster analysis?
- Understanding hierarchical and non-hierarchical cluster analysis.
  - Outline the algorithms for hierarchical clustering and *k*-means.
  - Describe the calculation of distances between clusters for single, average and complete linkage. How does the choice of the method influence the interpretation of results?
  - Explain the analogy between *k*-means and ANOVA.
  - List cluster validity indices. What is the difference between external and internal validation?
  - Discuss limitations of cluster analysis.

4

Unsupervised classification

- Group structure not known *a priori*
- Aim: identification of hidden structures and grouping of similar observations
- Methods include cluster analysis and self-organising maps

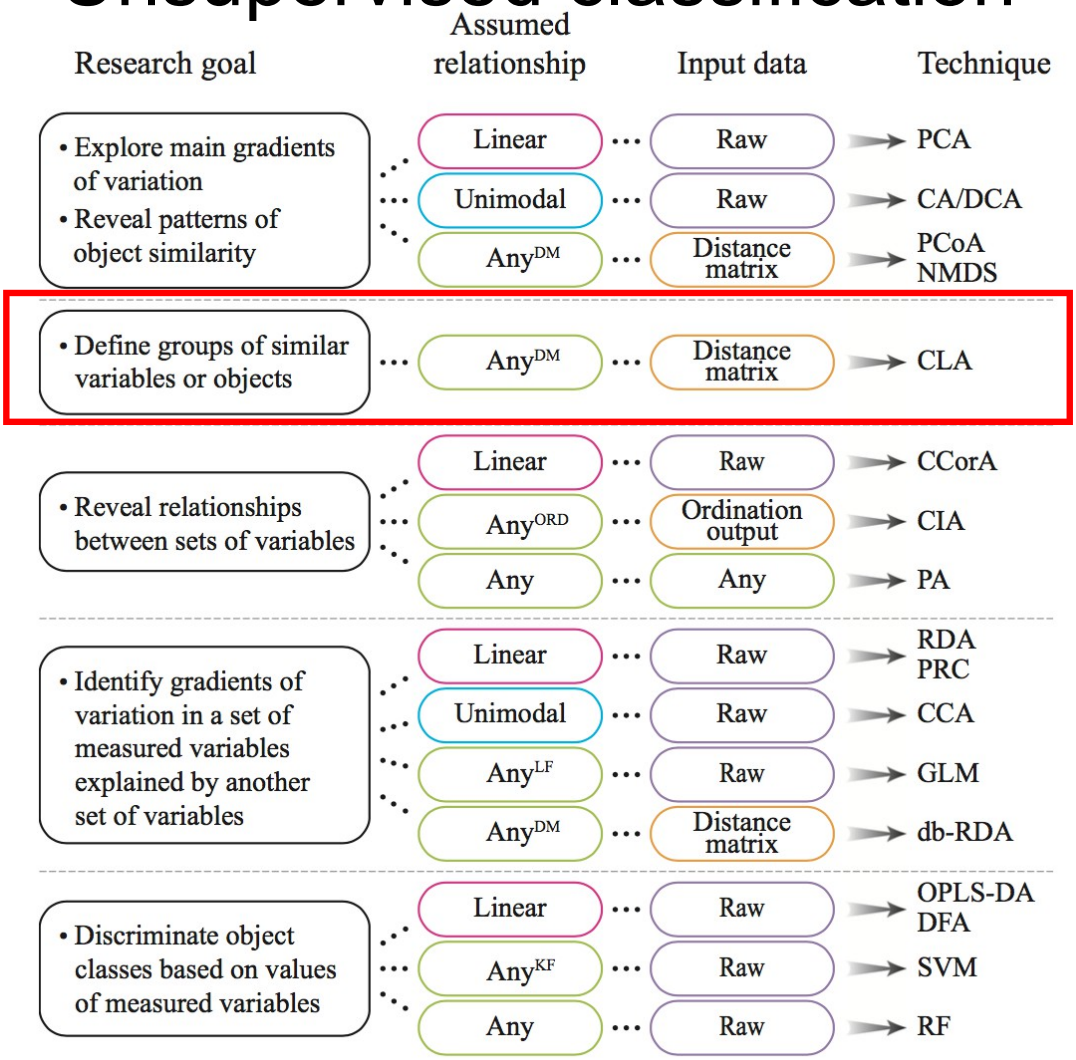


5

The given example of the relation between toxicity and specific ions in the molecule serves only the purpose of illustration and is not based on real data.

We limit our discussion to Cluster analysis.

Unsupervised classification

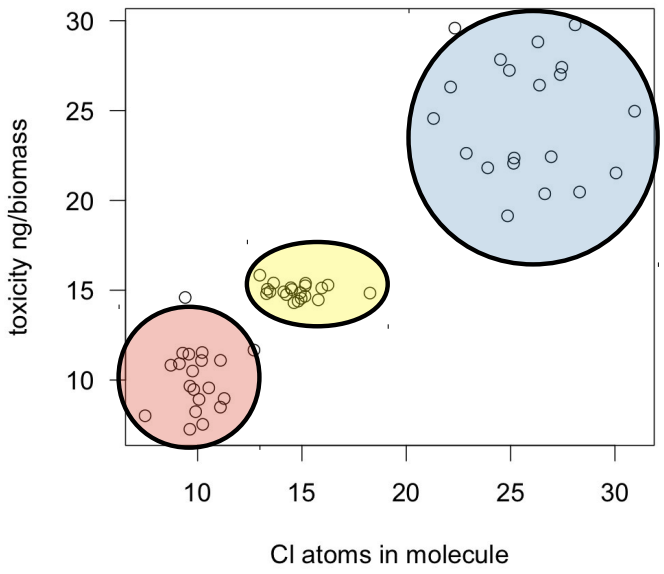


Paliy & Shankar 2016 *Mol Ecol*:1032

6

Cluster analysis: Application

- Identification of groups
- Data aggregation → Reduction of dimensions or „noise“
- Visualising similarity/distance of objects



7

Cluster analysis is used to identify groups (clusters), for example, groups of plants or animals that are similar in some sense and different in this sense to other groups, or likewise to identify genes with a similar expression pattern.

Data aggregation refers to the replacement of the individual observations by their group centroid or group membership in categorical analysis.

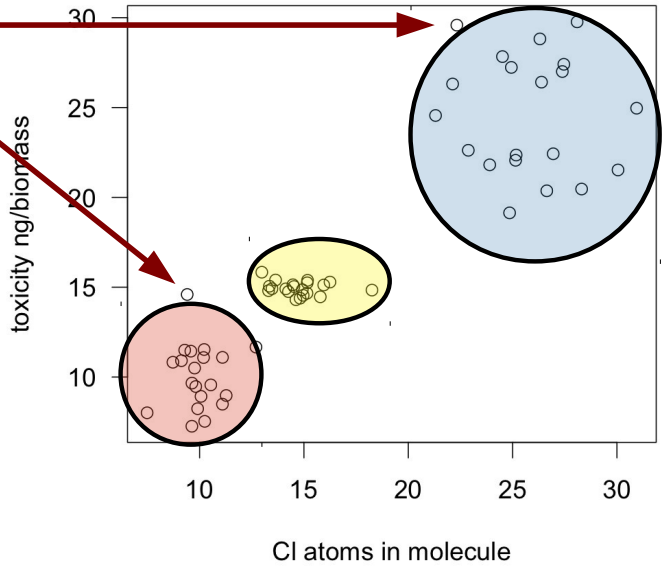
In addition, cluster analysis can be used to visualise the differences between objects in the multidimensional space in two dimensions (dendrogram) and to identify outliers from groups. For example, the point at x = 10; y = 15 could constitute a cluster with only one observation and consequently represents an outlier.

Wildi (2010) includes a very readable, short introduction into cluster analysis for ecological data. However, chapters on cluster analysis can be found in almost any textbook on multivariate methods in the environmental sciences. A modern treatment of cluster analysis is given in Everitt et al. (2011).

Everitt B.S. (2011) Cluster analysis, 5th edn. Wiley, Chichester.
Wildi O. (2010) Data analysis in vegetation ecology. Wiley-Blackwell, Chichester, West Sussex, UK ; Hoboken, NJ.

Cluster analysis: Application

- Identification of groups
- Data aggregation → Reduction of dimensions or „noise“
- Visualising similarity/distance of objects
- Identification of outliers



8

Cluster analysis is used to identify groups (clusters), for example, groups of plants or animals that are similar in some sense and different in this sense to other groups, or likewise to identify genes with a similar expression pattern.

Data aggregation refers to the replacement of the individual observations by their group centroid or group membership in categorical analysis.

In addition, cluster analysis can be used to visualise the differences between objects in the multidimensional space in two dimensions (dendrogram) and to identify outliers from groups. For example, the point at x = 10; y = 15 could constitute a cluster with only one observation and consequently represents an outlier.

Wildi (2010) includes a very readable, short introduction into cluster analysis for ecological data. However, chapters on cluster analysis can be found in almost any textbook on multivariate methods in the environmental sciences. A modern treatment of cluster analysis is given in Everitt et al. (2011).

Everitt B.S. (2011) Cluster analysis, 5th edn. Wiley, Chichester.
Wildi O. (2010) Data analysis in vegetation ecology. Wiley-Blackwell, Chichester, West Sussex, UK ; Hoboken, NJ.

Unsupervised classification: Cluster analysis

Contents

9

Cluster analysis: Intro

How does clustering work?

Cluster analysis relies on similarity or distance matrix

1. maximise within-group similarity of objects in cluster

2. minimise between group similarity

➡ Results depend on similarity/distance measure

10

The main challenge of cluster analysis lies in the quantification of the similarity or distance between objects and groups. The results of the cluster analysis vary with the measure selected to quantify the similarity/distance between objects. Thus, cluster analysis should be considered as a tool for exploratory analysis rather than for inference. However, clustering can also be conducted based on a statistical model (see Everitt et al. 2011: Chapter 6 and Kassambara 2017: Chapter 18).

The opposite of agglomerative clustering is divisive clustering, which starts with all objects in one large cluster and breaks the cluster into smaller clusters.

Everitt B.S. (2011) Cluster analysis, 5th edn. Wiley, Chichester.
Kassambara A. (2017) Practical guide to cluster analysis in R: unsupervised machine learning, Edition 1. STHDA.

Cluster analysis: Intro

How does clustering work?

Cluster analysis relies on similarity or distance matrix

1. maximise within-group similarity of objects in cluster

2. minimise between group similarity

➡ Results depend on similarity/distance measure

Hierarchical Agglomerative Clustering

- Widely used
- Starts with all objects as single clusters
- Objects merged into joint cluster based on distance or similarity

11

The main challenge of cluster analysis lies in the quantification of the similarity or distance between objects and groups. The results of the cluster analysis vary with the measure selected to quantify the similarity/distance between objects. Thus, cluster analysis should be considered as a tool for exploratory analysis rather than for inference. However, clustering can also be conducted based on a statistical model (see Everitt et al. 2011: Chapter 6 and Kassambara 2017: Chapter 18).

The opposite of agglomerative clustering is divisive clustering, which starts with all objects in one large cluster and breaks the cluster into smaller clusters.

Everitt B.S. (2011) Cluster analysis, 5th edn. Wiley, Chichester.
Kassambara A. (2017) Practical guide to cluster analysis in R: unsupervised machine learning, Edition 1. STHDA.

Hierarchical agglomerative clustering

1. Search for shortest distance between pairs of objects
→ merge into cluster

	1	2	3	4	5	6
1	0	1.414	2.000	4.472	5.657	6.708
2		0	1.414	3.162	4.243	5.385
3			0	4.000	4.472	5.000
4				0	2.000	4.123
5					0	2.236
6						0

Izenman 2008:415f

Izenman A.J. (2008) Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer, New York, NY.

12

Hierarchical agglomerative clustering

1. Search for shortest distance between pairs of objects → merge into cluster

2. Re-calculation of distances and repetition of step 1

	1	2	3	4	5	6
1	0	1.414	2.000	4.472	5.657	6.708
2		0	1.414	3.162	4.243	5.385
3			0	4.000	4.472	5.000
4				0	2.000	4.123
5					0	2.236
6						0

	1	23	4	5	6
1	0	1.414	4.472	5.657	6.708
23		0	3.162	4.243	5.000
4			0	2.000	4.123
5				0	2.236
6					0

Ends when all objects merged into one cluster

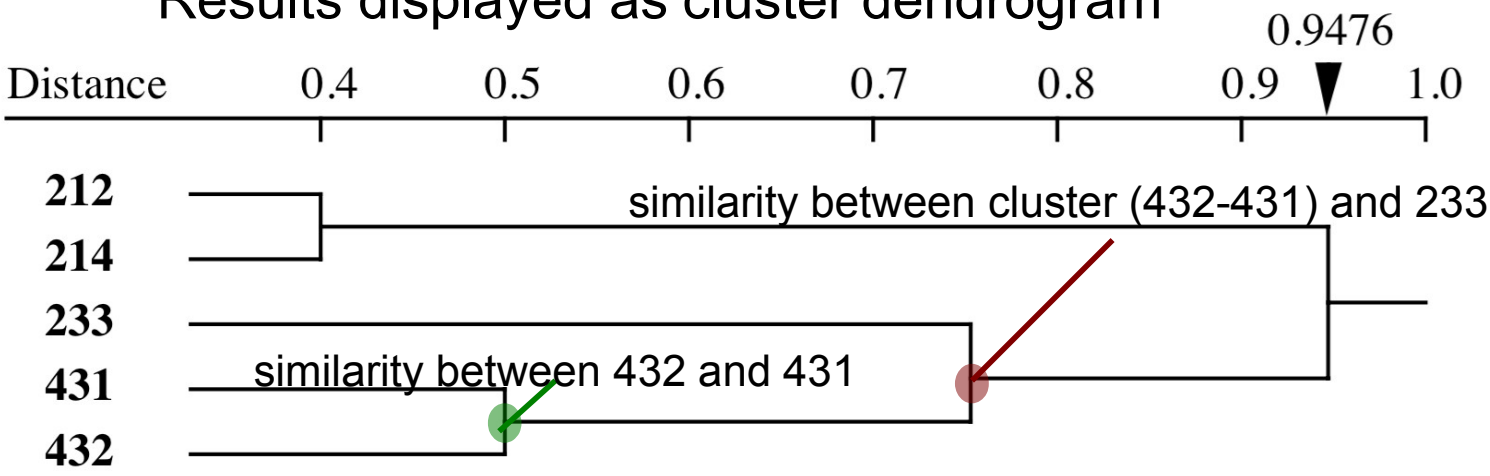
13

Izenman 2008:415f

Izenman A.J. (2008) Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer, New York, NY.

Hierarchical agglomerative clustering

Results displayed as cluster dendrogram

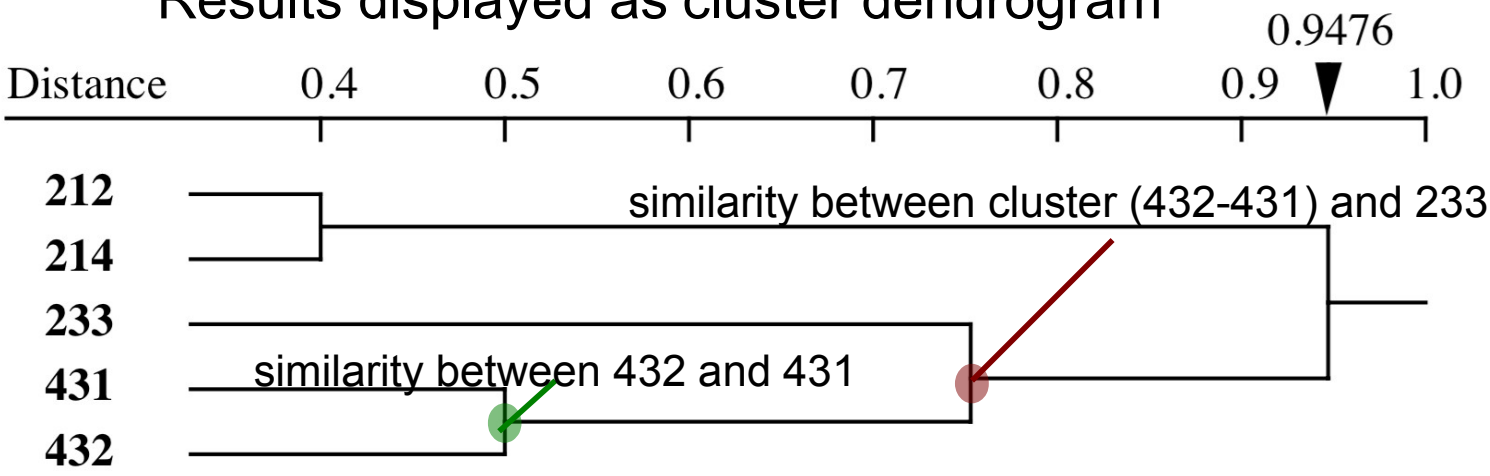


Legendre & Legendre 2012: 356

Phylogenetic tree for mammals taken from:
https://commons.wikimedia.org/wiki/File:The_Ancestors_Tale_Mammals_Phylogenetic_Tree_in_mya.png

Hierarchical agglomerative clustering

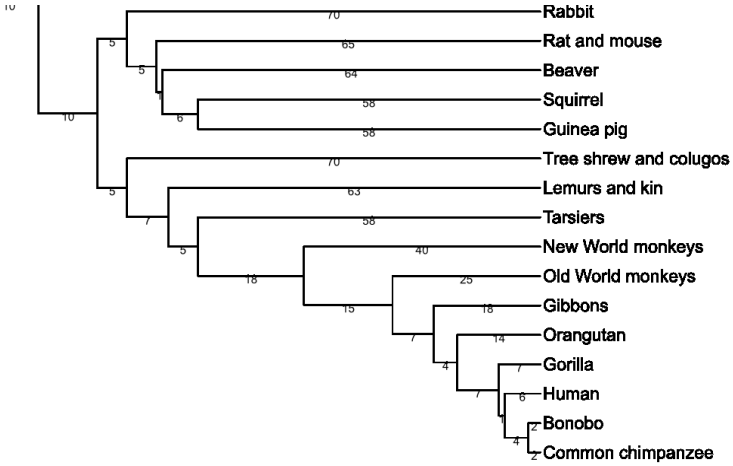
Results displayed as cluster dendrogram



Legendre & Legendre 2012: 356

Note the similarity in interpretation to a phylogenetic tree!

For example, humans are not directly related to rabbits, but they share a common ancestor

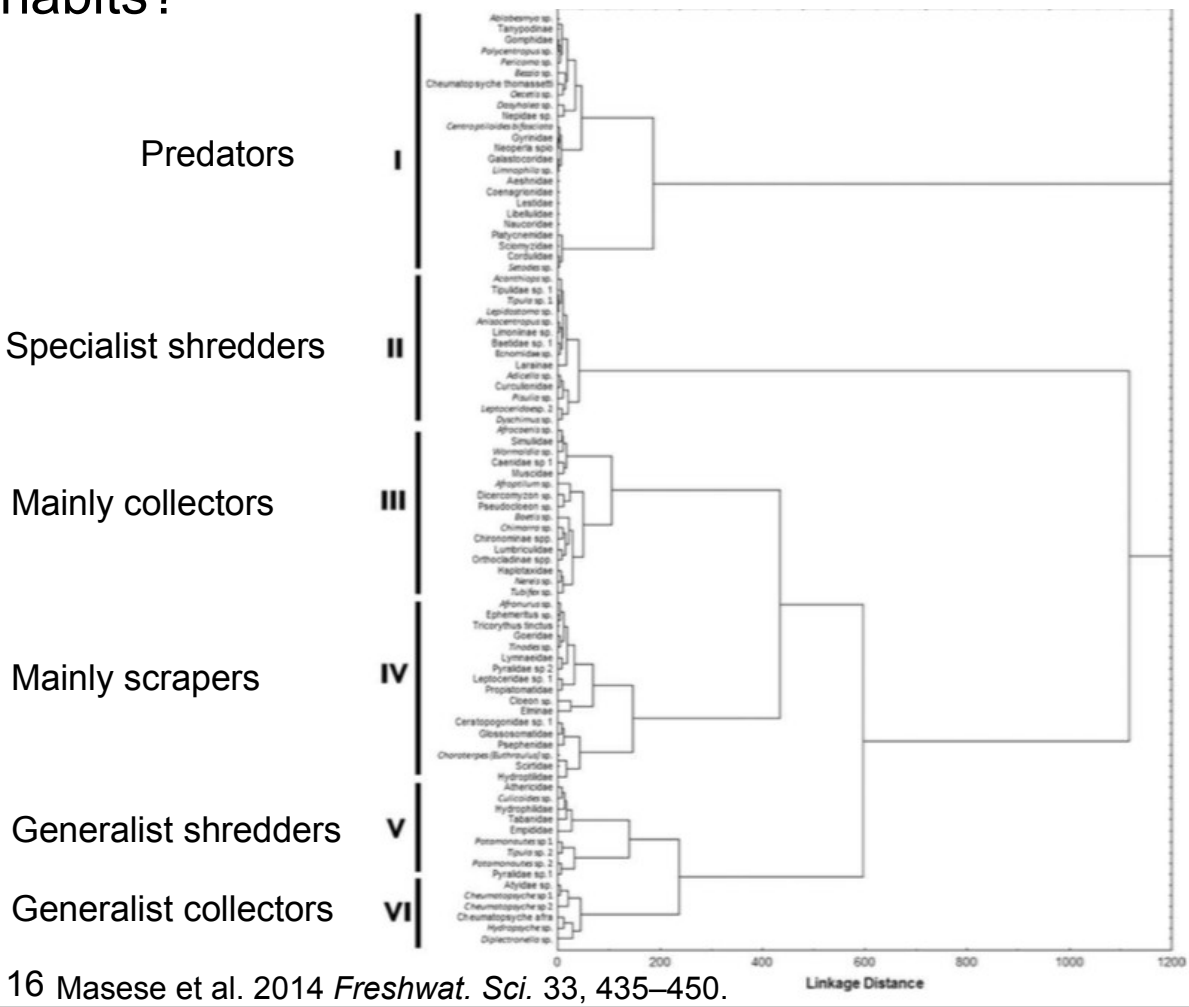


https://commons.wikimedia.org/

15

Phylogenetic tree for mammals taken from: https://commons.wikimedia.org/wiki/File:The_Ancestors_Tale_Mammals_Phylogenetic_Tree_in_mya.png

Example: What are the functional feeding groups in Kenyan freshwater invertebrates? Which species have similar feeding habits?



16 Masese et al. 2014 *Freshwat. Sci.* 33, 435–450.

Masese F.O., Kitaka N., Kipkemboi J., Gettel G.M., Irvine K. & McClain M.E. (2014) Macroinvertebrate functional feeding groups in Kenyan highland streams: evidence for a diverse shredder guild. Freshwater Science 33, 435–450.

The clustering of invertebrates was based on the gut content of the organisms.

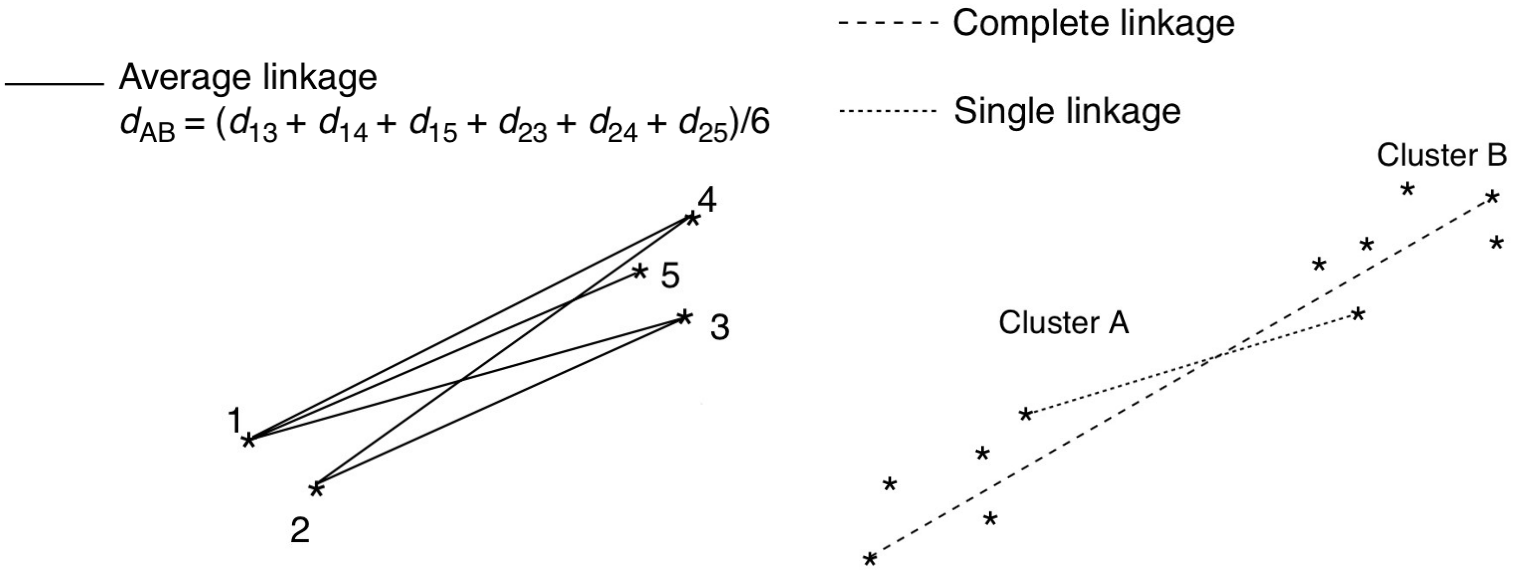
Unsupervised classification: Cluster analysis

Contents

17

Methods for calculating cluster distances

Examples of methods for calculating between-cluster distances



18

Everitt 2011: 77

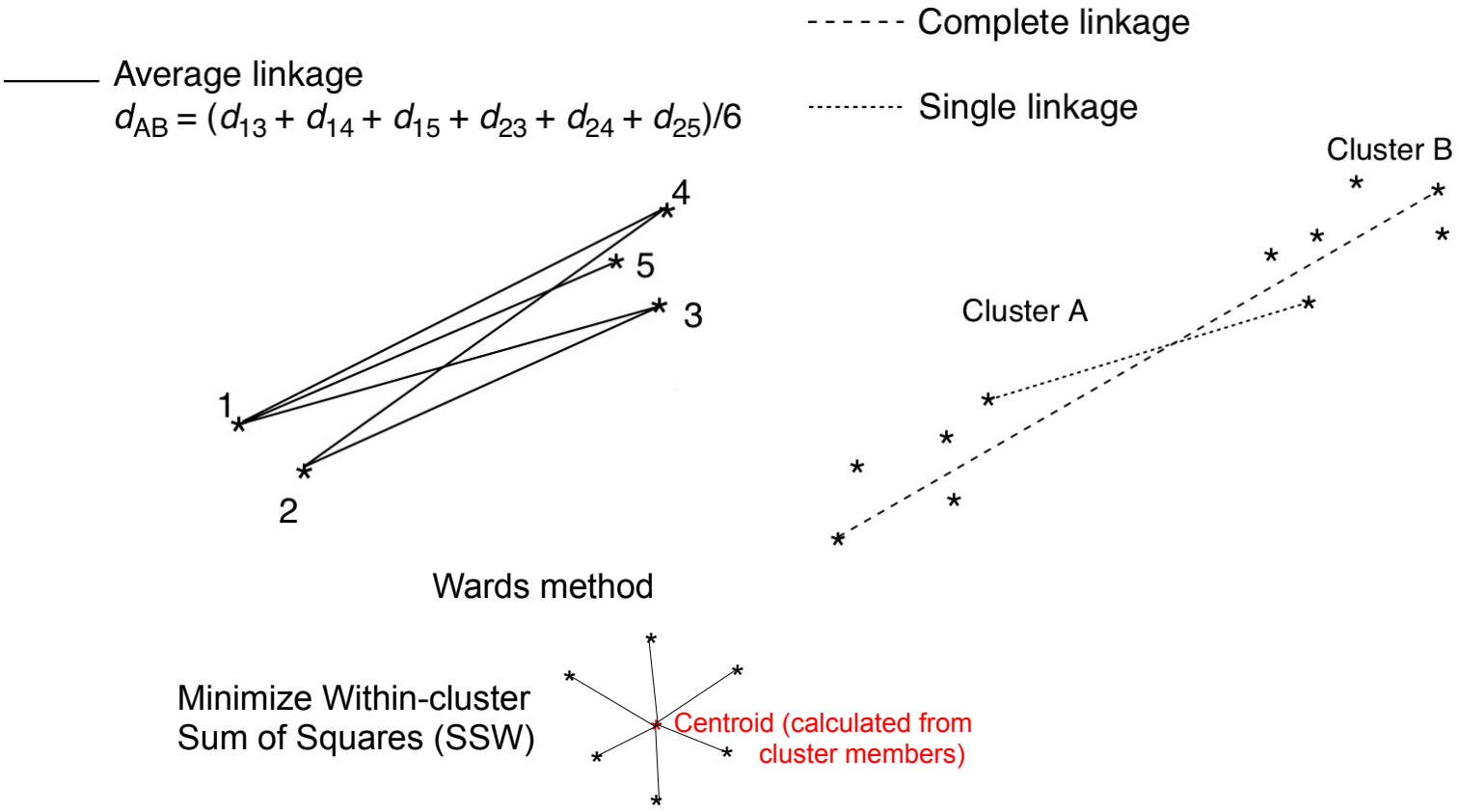
For more information on these and other methods see Everitt et al. (2011): 79.

Note that the selected method influences the result of the cluster analysis, just as the selection of the distance measure does. The R package clusterSim allows to simulate cluster results for different methods and distance measures.

Everitt et al. (2011) Cluster analysis. 5th ed. Wiley: Chichester.

Methods for calculating cluster distances

Examples of methods for calculating between-cluster distances



19

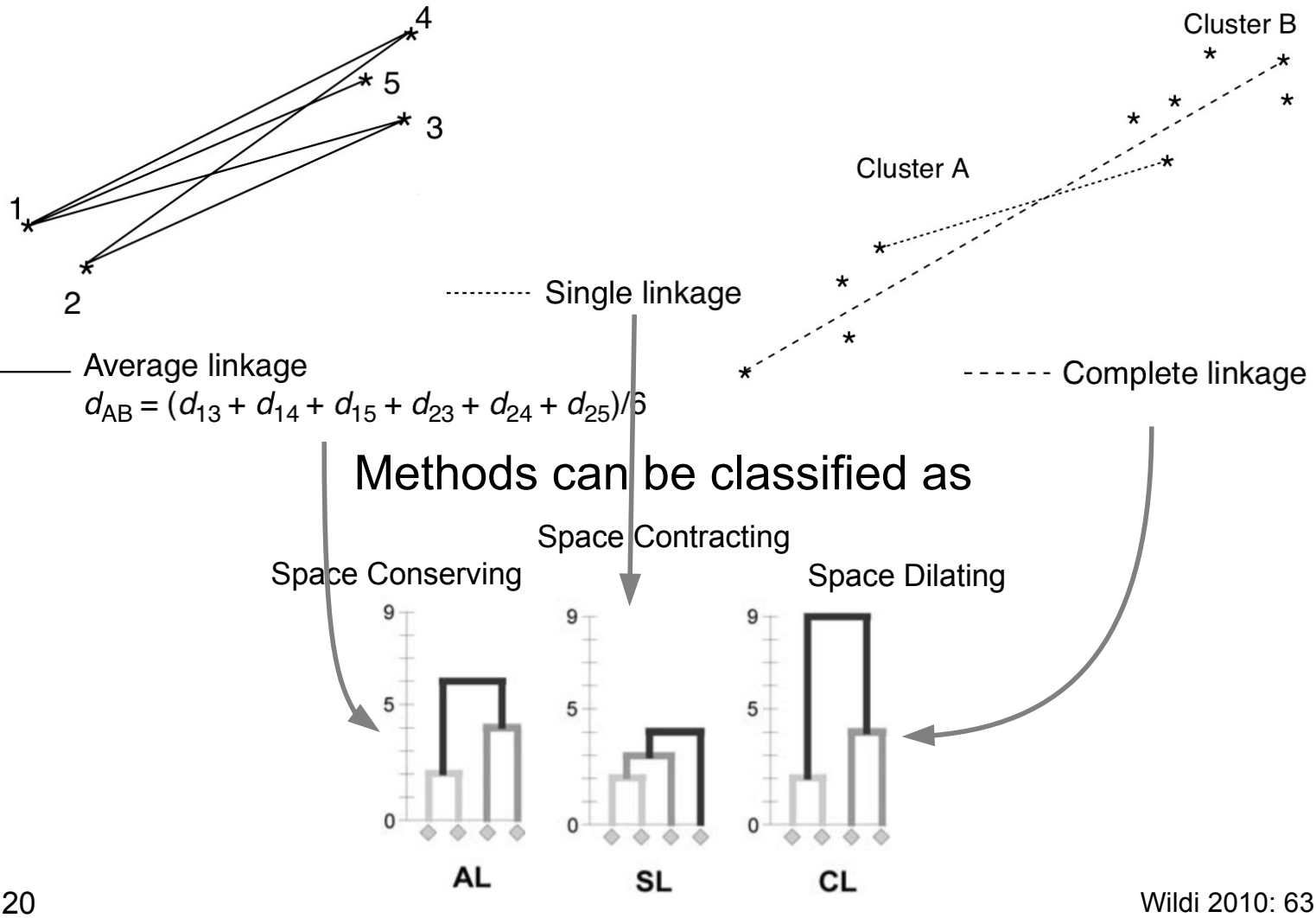
Everitt 2011: 77

For more information on these and other methods see Everitt et al. (2011): 79.

Note that the selected method influences the result of the cluster analysis, just as the selection of the distance measure does. The R package clusterSim allows to simulate cluster results for different methods and distance measures.

Everitt et al. (2011) Cluster analysis. 5th ed. Wiley: Chichester.

Influence of methods on clustering



20

Wildi 2010: 63

Three groups of agglomeration methods can be distinguished:

Space-conserving: The original distances (as given in the distance matrix) are preserved → high correlation between the original distances and the distances after clustering, which are given in the so-called cophenetic matrix. We can also compute STRESS 1 (see NMDS) for the cophenetic matrix, i.e. a measure to evaluate the preservation of original distances. This approach should be employed when the major aim is to preserve the characteristics of the input data. Related methods include “average linkage” or “Wards method”.

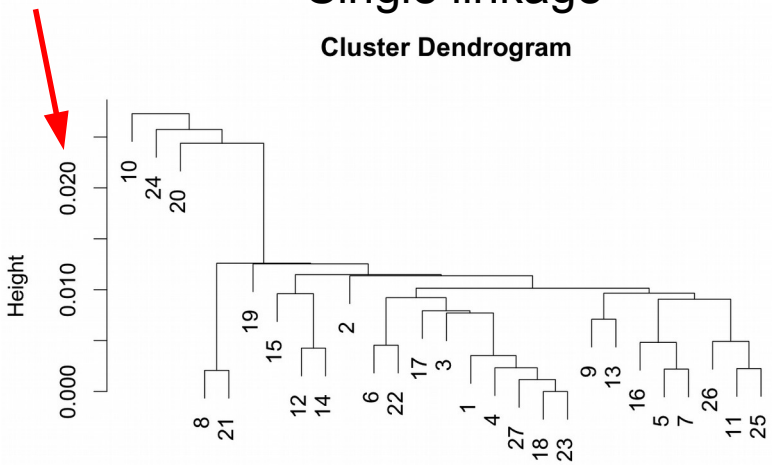
Space-contracting: This approach adds objects to clusters even if this leads to shorter distances between clusters than between objects in the raw data. It is employed to find discontinuities in the data. Related methods include the “single linkage” clustering.

Space-dilating: This approach minimises the distances of the objects within the clusters i.e. constructs clusters as homogeneous as possible. It can be used to combine objects with very similar characteristics. Related methods include the “Complete linkage” clustering.

Wildi O. (2010) Data analysis in vegetation ecology. Wiley-Blackwell, Chichester, West Sussex, UK ; Hoboken, NJ.

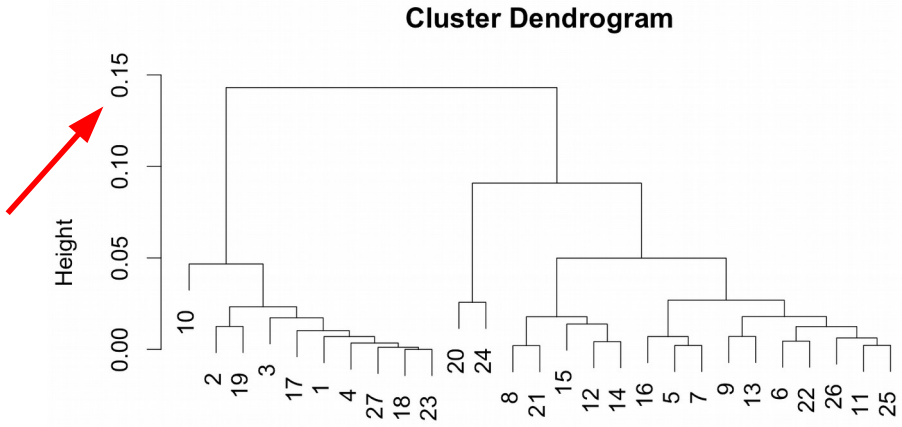
Example: Complete and single linkage

Single linkage



Large and close clusters

Complete linkage



Small and distant clusters

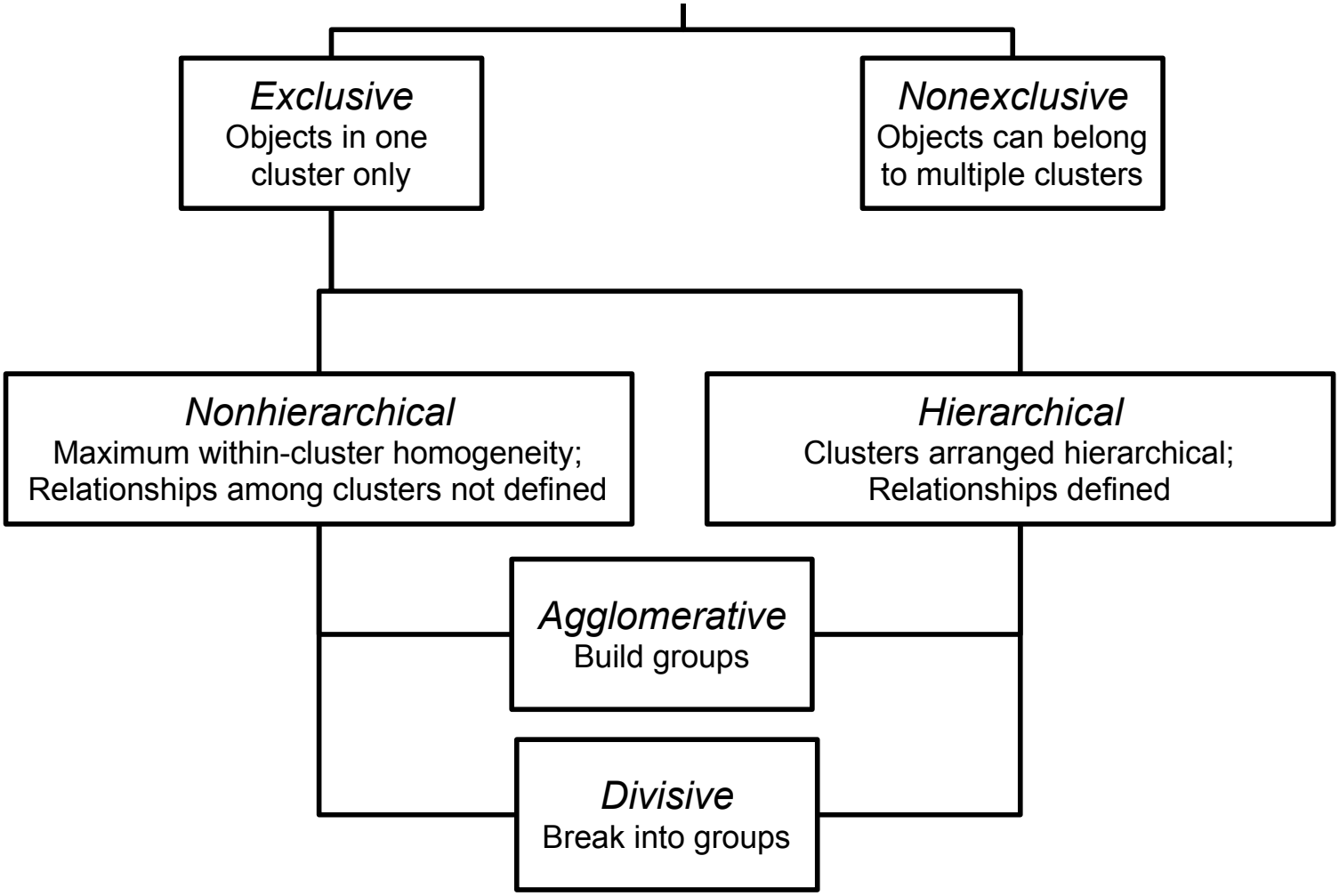
The figures show the results of a clustering with the Bray Curtis index for the single linkage (left figure) and for the complete linkage (right figure). The complete linkage yields rather small and distant clusters, whereas the single linkage technique results in larger and less distant clusters (note the red arrows that highlight differences in the scaling of the axes).

Unsupervised classification: Cluster analysis

Contents

22

Overview cluster analysis

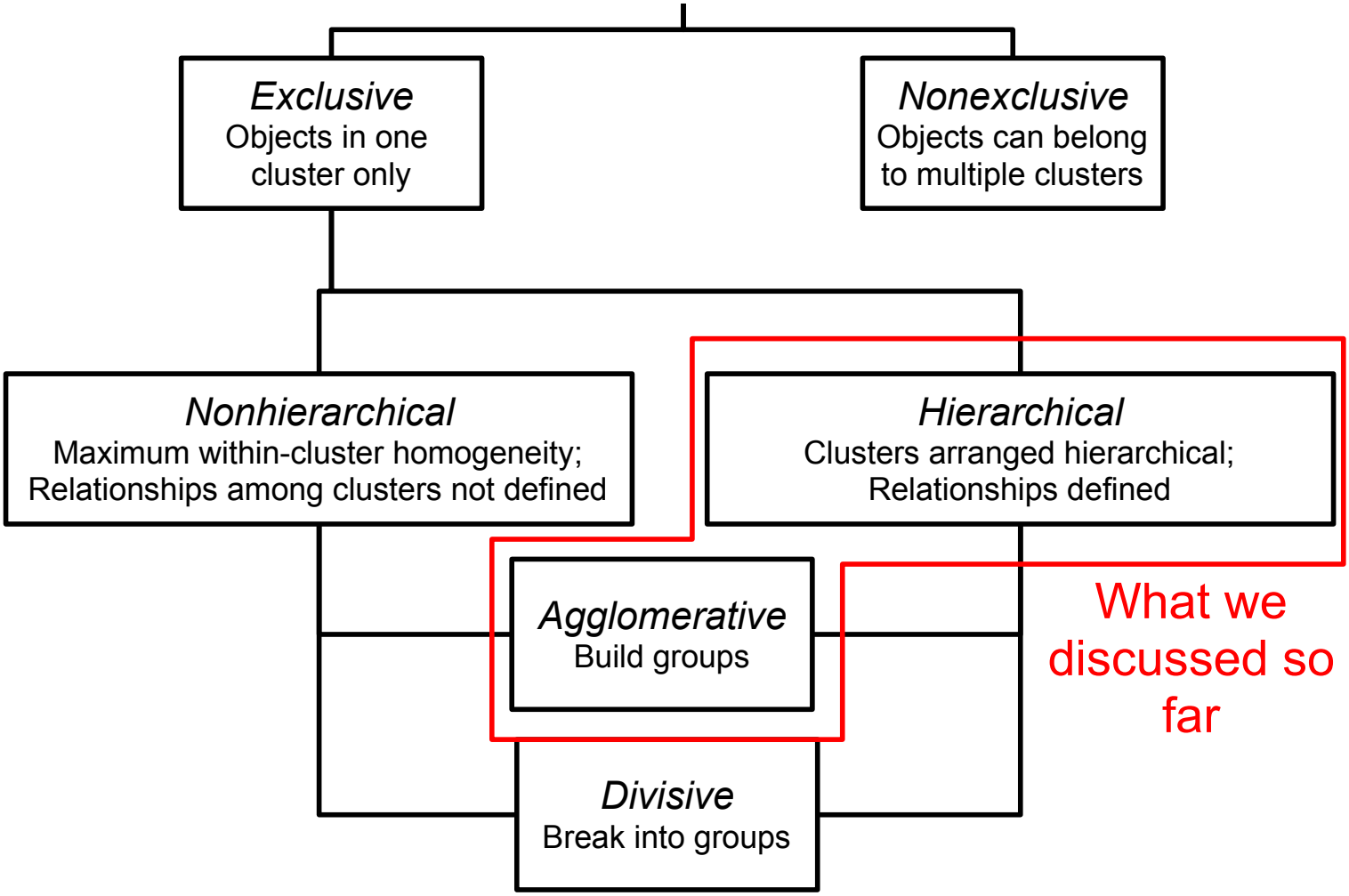


23

Nonexclusive techniques assign cluster membership values to objects. They produce overlapping and continuous clusters. Details are described in Everitt et al. 2011, see “Overlapping Clustering” and ”Fuzzy-based Clustering”. The technique is implemented in the R package cluster – function: funny, see also Kassambara 2017: 168-170.

Everitt B.S. (2011) Cluster analysis, 5th edn. Wiley, Chichester.
Kassambara A. (2017) Practical guide to cluster analysis in R: unsupervised machine learning, Edition 1. STHDA.

Overview cluster analysis



24

Nonexclusive techniques assign cluster membership values to objects. They produce overlapping and continuous clusters. Details are described in Everitt et al. 2011, see “Overlapping Clustering” and ”Fuzzy-based Clustering”. The technique is implemented in the R package cluster – function: funny, see also Kassambara 2017: 168-170.

Everitt B.S. (2011) Cluster analysis, 5th edn. Wiley, Chichester.
Kassambara A. (2017) Practical guide to cluster analysis in R: unsupervised machine learning, Edition 1. STHDA.

Non-hierarchical cluster analysis: k-means clustering

Assigns objects to a pre-defined number of clusters k

Problem:

n	k	**Number of possible partitions**
15	3	2,375,101
20	4	45,232,115,901
25	8	690,223,721,118,368,580
100	5	10^{68}

Calculation of all possible partitions becomes unsurmountable even for relatively small sample sizes (n)

Algorithm required

Everitt & Hothorn 2011: 175

Everitt B. & Hothorn T. (2011) An introduction to applied multivariate analysis with R. Springer, New York.

k-means clustering

Algorithm

1. Partition objects into *k* groups
2. Move objects and calculate change
3. Choose best solution
4. Repeat 2-3 until cluster criterion does not improve

Everitt & Hothorn 2011: 175

Given the vast amount of possibilities for partitioning, the optimal solution is not necessarily found in a single run. Thus, the algorithm starts multiple times with a random partitioning into the desired number of groups. Alternatively, the result of a preceding hierarchical cluster analysis can be provided as starting point. Moreover, the so-called *k*-means++ algorithm augments *k*-means with a different initial partitioning, potentially resulting in more accurate partitioning (Arthur & Vassilvitskii 2007).

Arthur D. & Vassilvitskii S. (2007) k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035. Society for Industrial and Applied Mathematics, New Orleans, Louisiana. https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf

Borcard D., Gillet F. & Legendre P. (2011) Numerical ecology with R, 1. Springer, New York, NY.

Everitt B. & Hothorn T. (2011) An introduction to applied multivariate analysis with R. Springer, New York.

k-means clustering

Algorithm

1. Partition objects into *k* groups
2. Move objects and calculate change
3. Choose best solution
4. Repeat 2-3 until cluster criterion does not improve

Which criterion is used to assess best solution?

Minimisation of Within-cluster SSQ (SSW)
→ Note analogy to Wards method and ANOVA

k-means clustering implicitly relies on euclidean distances
→ Ecological data require data preparation
(see Borcard et al. 2011: 80 for details)

Everitt & Hothorn 2011: 175

Given the vast amount of possibilities for partitioning, the optimal solution is not necessarily found in a single run. Thus, the algorithm starts multiple times with a random partitioning into the desired number of groups. Alternatively, the result of a preceding hierarchical cluster analysis can be provided as starting point. Moreover, the so-called *k*-means++ algorithm augments *k*-means with a different initial partitioning, potentially resulting in more accurate partitioning (Arthur & Vassilvitskii 2007).

Arthur D. & Vassilvitskii S. (2007) k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035. Society for Industrial and Applied Mathematics, New Orleans, Louisiana. https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf

Borcard D., Gillet F. & Legendre P. (2011) Numerical ecology with R, 1. Springer, New York, NY.

Everitt B. & Hothorn T. (2011) An introduction to applied multivariate analysis with R. Springer, New York.

Unsupervised classification: Cluster analysis

Contents

1. Introduction to unsupervised classification and cluster analysis
2. Hierarchical agglomerative clustering
3. Linkage methods for hierarchical clustering
4. Overview and *k*-means clustering
5. **Cluster validity indices and number of clusters**
6. Discussion of cluster analysis and further clustering techniques

28

Cluster validity indices (CVIs)

- How many clusters?
- Many criteria, differing scopes, e.g. homogeneity, separation, uniformity, representation and stability
  → Do not overinterpret individual criteria

Internal CVIs are used to compare different cluster solutions, whereas external CVIs compare the cluster solution to a known grouping structure. A complete overview on CVIs is beyond the scope of this course, we only discuss a few indices with different scopes. The scopes include achieving a high within-cluster homogeneity, between-cluster separation, uniformity of the clusters, reliable representation of the original data or stability of the cluster solution. These scopes depend on the research question. For example, in a study on the feeding types of insects, we may want to achieve a high between-cluster separation to assign groups, whereas if we want to identify surrogate species for experiments, we may rather want a high within-cluster homogeneity.

Given the many (subjective) decisions that are involved in cluster analysis, Hennig (2014) warns that “using supposedly “objective” criteria in a more traditional fashion does not solve these problems [lack of formalisation and ambiguity of cluster requirements, modified by RBS] but rather hides them”.

Several indices rely on the sum of squares (SSQ) within and between clusters. We only discuss the Calinski-Harabsz index (below), which is widely used, but other indices based on SSQ may perform better for specific data sets (for a comparison see Zhao & Fränti (2014)). Generally, these indices work best for Gaussian-type data, i.e. data that follow a bell-shaped form.

Some test statistics (not discussed in detail here) enable the assessment of hypotheses related to the cluster solutions such as the cubic cluster, the pseudo R^2 and the pseudo F statistic. Given that these statistics require multivariate normality of the cluster solution, they should only be used supportively and not as objective decision criterion. In addition, the statistics would be applied to each step of cluster formation, which is computationally intensive and leads to large clusters, if accounting for multiple inference. Note that testing a resulting cluster solution for statistical significance with ANOVA is meaningless, as even random data will produce a statistical significant clustering (see R script).

The *CH* index is based on SSQs, note the similarity to ANOVA. The index considers homogeneity (SSW) and separation (SSB). According to this index, the maximum *CH* relates to the *k* that gives the optimal number of clusters.

Hennig C. (2014) How Many Bee Species? A Case Study in Determining the Number of Clusters. In: Data Analysis, Machine Learning and Knowledge Discovery. (Eds M. Spiliopoulou, L. Schmidt-Thieme & R. Janning), pp. 41–49. Springer International Publishing.

Zhao Q. & Fränti P. (2014) WB-index: A sum-of-squares based index for cluster validity. Data & Knowledge Engineering 92, 77–89.

Cluster validity indices (CVIs)

- How many clusters?
- Many criteria, differing scopes, e.g. homogeneity, separation, uniformity, representation and stability
  → Do not overinterpret individual criteria

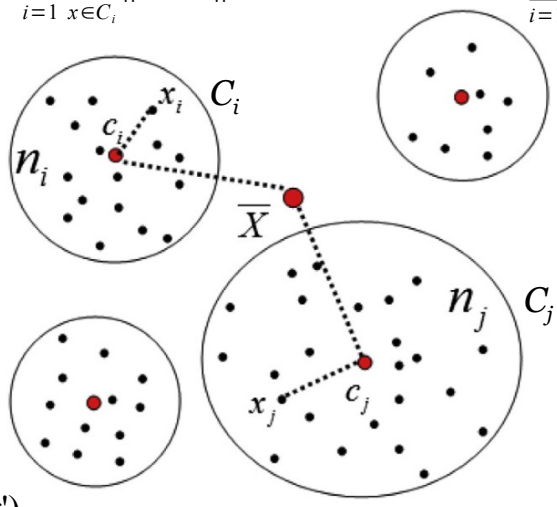
Selected internal CVIs:

Within-cluster SSQ

$$SSW = \sum_{i=1}^{M} \sum_{x \in C_i} \|x - c_i\|^2$$

Between-cluster SSQ

$$SSB = \sum_{i=1}^{M} n_i \|c_i - \overline{X}\|^2$$



Calinski-Harabsz

$$CH = \frac{SSB/(M-1)}{SSW/(N-M)}$$

with

N = no. of observations

M = no. of clusters

x_i = observation in C_i

n_i = no. of observations in cluster C_i

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$ ('overall center')

c_i = centroid of cluster C_i ('cluster center')

Modified from Zhao & Fränti 2014

Internal CVIs are used to compare different cluster solutions, whereas external CVIs compare the cluster solution to a known grouping structure. A complete overview on CVIs is beyond the scope of this course, we only discuss a few indices with different scopes. The scopes include achieving a high within-cluster homogeneity, between-cluster separation, uniformity of the clusters, reliable representation of the original data or stability of the cluster solution. These scopes depend on the research question. For example, in a study on the feeding types of insects, we may want to achieve a high between-cluster separation to assign groups, whereas if we want to identify surrogate species for experiments, we may rather want a high within-cluster homogeneity.

Given the many (subjective) decisions that are involved in cluster analysis, Hennig (2014) warns that “using supposedly “objective” criteria in a more traditional fashion does not solve these problems [lack of formalisation and ambiguity of cluster requirements, modified by RBS] but rather hides them”.

Several indices rely on the sum of squares (SSQ) within and between clusters. We only discuss the Calinski-Harabsz index (below), which is widely used, but other indices based on SSQ may perform better for specific data sets (for a comparison see Zhao & Fränti (2014)). Generally, these indices work best for Gaussian-type data, i.e. data that follow a bell-shaped form.

Some test statistics (not discussed in detail here) enable the assessment of hypotheses related to the cluster solutions such as the cubic cluster, the pseudo R^2 and the pseudo F statistic. Given that these statistics require multivariate normality of the cluster solution, they should only be used supportively and not as objective decision criterion. In addition, the statistics would be applied to each step of cluster formation, which is computationally intensive and leads to large clusters, if accounting for multiple inference. Note that testing a resulting cluster solution for statistical significance with ANOVA is meaningless, as even random data will produce a statistical significant clustering (see R script).

The *CH* index is based on SSQs, note the similarity to ANOVA. The index considers homogeneity (SSW) and separation (SSB). According to this index, the maximum *CH* relates to the *k* that gives the optimal number of clusters.

Hennig C. (2014) How Many Bee Species? A Case Study in Determining the Number of Clusters. In: Data Analysis, Machine Learning and Knowledge Discovery. (Eds M. Spiliopoulou, L. Schmidt-Thieme & R. Janning), pp. 41–49. Springer International Publishing.

Zhao Q. & Fränti P. (2014) WB-index: A sum-of-squares based index for cluster validity. Data & Knowledge Engineering 92, 77–89.

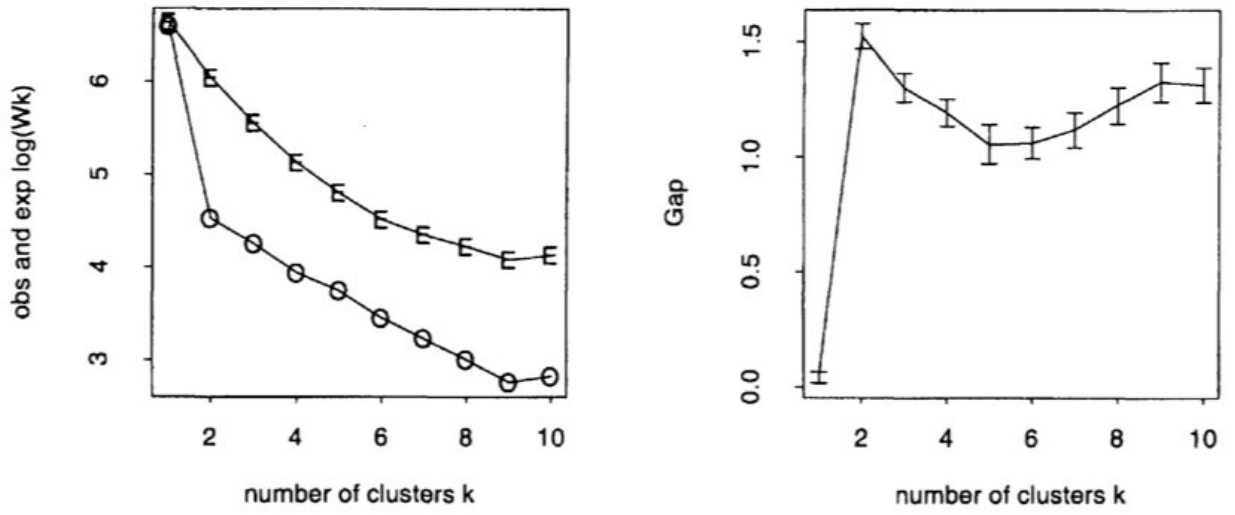
Cluster validity indices (CVIs)

Selected internal CVIs:

GAP index

$$GAP_n(k) = E_n^*\{\log(SSW_k)\} - \log(SSW_k)$$

- Determines uniform reference distribution E^* via bootstrapping using range of original data for same sample size n
- Optimal no. of clusters *k:* maximum difference in cluster criterion (*SSW*)



Tibshirani *et al.* 2001

The GAP index allows to evaluate the single cluster solution, i.e. whether the data forms a single homogeneous group.

Tibshirani R., Walther G. & Hastie T. (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 411–423.

Cluster validity indices (CVIs)

Selected internal CVIs:

Silhouette width *s*

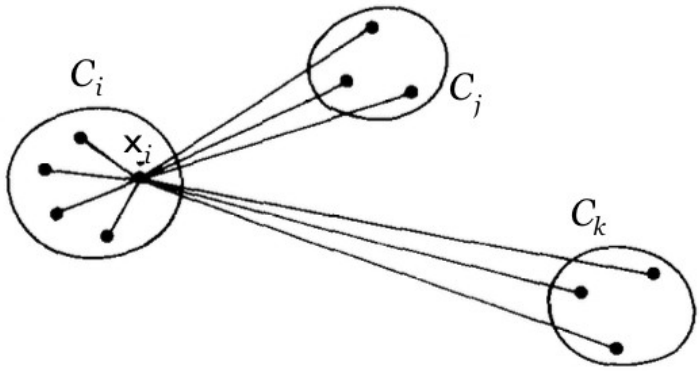
$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(b(x_i), a(x_i))}$$

with

$b(x_i) = \frac{1}{n_j} \sum_{b \in C_j} dist(b, x_i)$ ('average dissimilarity of x_i to all other objects of cluster C_j')

$a(x_i) = \frac{1}{n_i} \sum_{a \in C_i} dist(a, x_i)$ ('average dissimilarity of x_i to all other objects of cluster C_i')

Cluster C_j is the nearest neighbour to C_i



Modified from Rousseeuw 1987

The Silhouette index $s(x)$ evaluates the goodness of the clustering for each individual observation. It approaches its maximum (of 1) for high homogeneity within the cluster of x and strong separation (i.e. high average distance to observations from the nearest neighbour cluster). $s(x)$ values < 0 indicate that an observation would better be moved to the neighbouring cluster, which is mathematically evident because for this case the average distance to other observations from the cluster of x is higher than from the neighbouring cluster.

For comparison of cluster results for a different number of clusters k, the average silhouette width can be used. According to this index, the optimal number of clusters *M** is identified by the k related to the maximum S.

Rousseeuw P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65.

Cluster validity indices (CVIs)

Selected internal CVIs:

Silhouette width *s*

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(b(x_i), a(x_i))}$$

with

$$b(x_i) = \frac{1}{n_j} \sum_{b \in C_j} dist(b, x_i)$$ ('average dissimilarity of x_i to all other objects of cluster C_j')

$$a(x_i) = \frac{1}{n_i} \sum_{a \in C_i} dist(a, x_i)$$ ('average dissimilarity of x_i to all other objects of cluster C_i')

Cluster C_j is the nearest neighbour to C_i

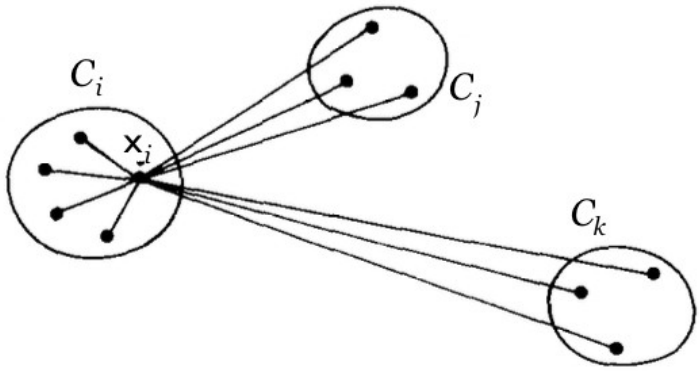
Evaluation of individual *x*

$-1 \le s(x_i) \le 1$

$s(x_i) \rightarrow 1 \Rightarrow$ well classified

$s(x_i) \rightarrow -1 \Rightarrow$ misclassified

$s(x_i) \approx 0 \Rightarrow$ unclear



33

Modified from Rousseeuw 1987

The Silhouette index $s(x)$ evaluates the goodness of the clustering for each individual observation. It approaches its maximum (of 1) for high homogeneity within the cluster of x and strong separation (i.e. high average distance to observations from the nearest neighbour cluster). $s(x)$ values < 0 indicate that an observation would better be moved to the neighbouring cluster, which is mathematically evident because for this case the average distance to other observations from the cluster of x is higher than from the neighbouring cluster.

For comparison of cluster results for a different number of clusters k, the average silhouette width can be used. According to this index, the optimal number of clusters *M** is identified by the k related to the maximum S.

Rousseeuw P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65.

Cluster validity indices (CVIs)

Selected internal CVIs:

Silhouette width *s*

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(b(x_i), a(x_i))}$$

with

$$b(x_i) = \frac{1}{n_j} \sum_{b \in C_j} dist(b, x_i)$$ ('average dissimilarity of x_i to all other objects of cluster C_j')

$$a(x_i) = \frac{1}{n_i} \sum_{a \in C_i} dist(a, x_i)$$ ('average dissimilarity of x_i to all other objects of cluster C_i')

Cluster C_j is the nearest neighbour to C_i

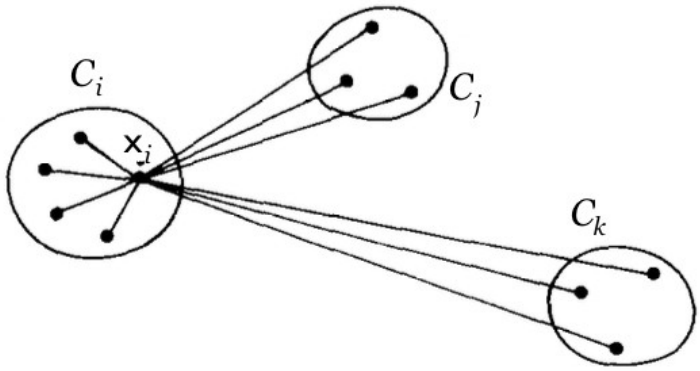
Evaluation of individual x

$-1 \le s(x_i) \le 1$

$s(x_i) \rightarrow 1 \Rightarrow$ well classified

$s(x_i) \rightarrow -1 \Rightarrow$ misclassified

$s(x_i) \approx 0 \Rightarrow$ unclear



Evaluation of cluster solution

$$\overline{S} = \frac{1}{N} \sum_{l=1}^{N} s(x_l)$$

('average silhouette width over the N observations')

$\overline{S} > 0.5$ suggests reasonable clustering

$\overline{S} < 0.2$ suggests lack of cluster structure

Modified from Rousseeuw 1987

The Silhouette index $s(x)$ evaluates the goodness of the clustering for each individual observation. It approaches its maximum (of 1) for high homogeneity within the cluster of x and strong separation (i.e. high average distance to observations from the nearest neighbour cluster). $s(x)$ values < 0 indicate that an observation would better be moved to the neighbouring cluster, which is mathematically evident because for this case the average distance to other observations from the cluster of x is higher than from the neighbouring cluster.

For comparison of cluster results for a different number of clusters k, the average silhouette width can be used. According to this index, the optimal number of clusters *M** is identified by the k related to the maximum S.

Rousseeuw P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65.

Cluster validity indices (CVIs)

Selected internal CVIs:

Cluster stability via bootstrapping

- Draws *B* bootstrapped samples from original data
- Computes clustering for original data and bootstrapped data
- Calculates Jaccard index to compare each clustering for *B* with the clustering for the original data. Mean of Jaccard index used to evaluate stability (> 0.75 good stability)

35

For further details on the cluster stability via bootstrapping see Hennig C. (2007) Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis* 52, 258–271.

Cluster validity indices (CVIs)

Selected internal CVIs:

Cluster stability via bootstrapping

- Draws *B* bootstrapped samples from original data
- Computes clustering for original data and bootstrapped data
- Calculates Jaccard index to compare each clustering for *B* with the clustering for the original data. Mean of Jaccard index used to evaluate stability (> 0.75 good stability)

Selected external CVI:

Rand index

- Computes pairs of true (a) and false (b) positives, true (d) and false (c) Negatives
- Same formula as Simple matching coefficient
- Adjusted Rand index: adjusted to yield 0 for two random partitions

		Cluster results	
		Same cluster	Different cluster
External data	Same class	a	b
	Different class	c	d

$$\mathrm{RI} = \frac{a+d}{a+b+c+d}$$

For further details on the cluster stability via bootstrapping see Hennig C. (2007) Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis* 52, 258–271.

Unsupervised classification: Cluster analysis

Contents

1. Introduction to unsupervised classification and cluster analysis
2. Hierarchical agglomerative clustering
3. Linkage methods for hierarchical clustering
4. Overview and *k*-means clustering
5. Cluster validity indices and number of clusters
6. **Discussion of cluster analysis and further clustering techniques**

37

Critical discussion of cluster analysis

- Lack of formalisation, many choices (particularly in hierarchical clustering) that influence outcomes → results can be ambiguous, rather exploratory technique
- Tendency for specific (spherical) clusters
- *k*-means or hierarchical clustering (HC)?
  - *k* (or narrow range of *k*) has been fixed or is known → *k*-means
  - Dendrogram and cluster steps are of interest → HC
  - HC is more flexible (distance measure, clustering method), *k*-means restricted to euclidean distance and SSQs
  - large data sets → *k*-means (more efficient)

38

Cluster analysis remains at least partially subjective. Given the multiple techniques and methods available, it can easily be „misused“. Nevertheless, it represents a valuable tool for exploratory data analysis in different scientific fields. Similarly, Everitt (2011: 287) states: “The methods of cluster analysis can be valuable tools in the exploration of multivariate data. […] Applying the methods in practice, however, requires considerable care if over-interpretation […] is to be avoided. [...] Simply applying a particular method of cluster analysis [...] and accepting the solution at face value is in general not adequate.“

Partitioning Around Medoids represents a more flexible alternative to *k*-means and allows for the use of different distance measures. Still, it is not as flexible as hierarchical clustering.

All methods discussed focus on spherical clusters (see next slide for an alternative). For an example on the poor performance of k-means for a non-spherical clusters based on an R simulation see:
http://enhancedatascience.com/2017/10/24/machine-learning-explained-kmeans/

Further techniques in Cluster analysis

- Partitioning around medoids: similar to *k*-means but is non-euclidean, open to alternative distance metrics
  `pam() {cluster}`
- Non-hierarchical clustering for large data sets
  `clara() {cluster}`
- Model-based clustering: Estimates model parameters from data, assumes specific cluster structure (spherical, diagonal, ellipsoid)
  `Mclust() {mclust}`
- Clustering for non-spherical shapes
  `dbscan() {fpc}`
- Variable clustering: Useful to identify multicollinearity and surrogate variables
  `varclus() {Hmisc}` `hclustvar() {ClustOfVar}`

The textbook by Kassambara (2017) describes the implementation and visualisation in R of all these techniques.

Partitioning Around Medoids represents an alternative to *k*-means for non-hierarchical cluster analysis that allows for the use of alternative distance metrics. The sum of distances is minimised instead of the squared euclidian distances, leading to more robust results compared to *k*-means.

For large data sets the function „clara“ should be used. What is considered „small“ and „large“, depends on the number of variables and objects as well as the processing power of the computer.

The approaches considered so far were rather exploratory. However, model-based cluster analysis is also available, but requires rather a large sample size. For details refer to the package „mclust” or to the chapters in Everitt 2011, Kassambara 2017 as well as Everitt & Hothorn 2011.

Everitt B.S. (2011) Cluster analysis, 5th edn. Wiley, Chichester.

Everitt B. & Hothorn T. (2011) An introduction to applied multivariate analysis with R. Springer, New York.

Kassambara A. (2017) Practical guide to cluster analysis in R: unsupervised machine learning, Edition 1. STHDA.