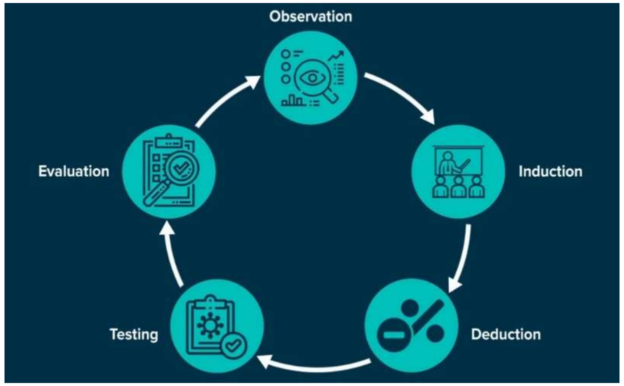
Empirical research in management and economics

Cluster analysis



Thorsten Pachur

Technical University of Munich
School of Management
Chair of Behavioral Research Methods

Exam

Please register on TUMonline for the exam!

Deadline: 15 January 2026

Statistical software for next lecture

- For the exercise, please download & install R and RStudio

  https://www.rstudio.com/products/rstudio/download/#download

- Please make yourself familiar with R

→ Free tutorials and help files
  - http://www.r-tutor.com/r-introduction
  - https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf
  - You can take the free DataCamp online tutorial! (https://www.datacamp.com/courses/free-introduction-to-r)

Recap of last lecture

- What are the goals of factor-analytic techniques?
- What do eigenvalues in a principle component analysis (PCA) represent?
- What is the key idea underlying the scree test? What is the purpose of parallel analysis?
- What are factor loadings? What is meant by uniqueness and how is it related to communality?
- Why is it usually useful to conduct a rotation of the extracted factor solution?
- What is the difference between orthogonal and oblique rotation?
- What are factor scores?
- What is measured with Bartlett‘s test and the KMO test, and what results of these tests are desirable?
- Give two rules of thumb when planning the sample size for a PCA

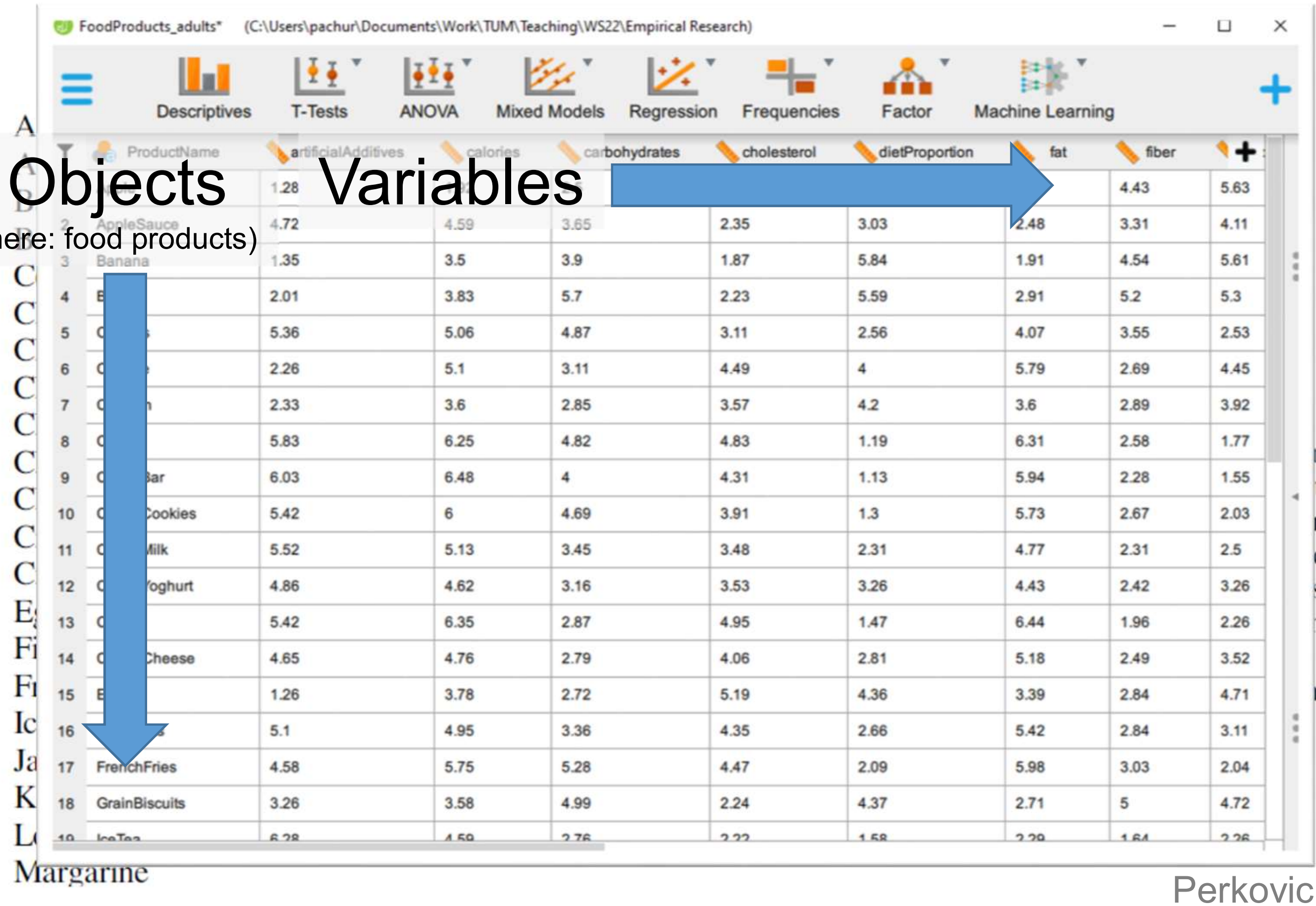
Agenda for the semester

Session	Date	Topic
1	13 October	Introduction
2	20 October	Descriptive data analysis
3	27 October	Hypothesis development and measurement
4	3 November	Inferential data analysis I
5	10 November	Inferential data analysis II
6	17 November	Simple regression
7	24 November	Multiple regression
8	1 December	Logistic regression
9	8 December	Factor analysis
10	**15 December**	**Cluster analysis**
11	12 January	Conjoint analysis
12	19 January	The replication crisis and open science
13	26 January	Summary and questions
	11 February	Exam

Goals for this week

- You know the goals and principles of cluster analysis
- You understand different ways to quantify the similarity of objects
- You understand what is meant by a hierarchical approach to clustering
- You are familiar with different methods for creating clusters
- You know indices for evaluating cluster solutions
- You have experience with interpreting and characterizing clusters

Food ecology



	ProductName	artificialAdditives	calories	carbohydrates	cholesterol	dietProportion	fat	fiber	
1	[illegible]	1.28	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.43	5.63
2	AppleSauce	4.72	4.59	3.65	2.35	3.03	2.48	3.31	4.11
3	Banana	1.35	3.5	3.9	1.87	5.84	1.91	4.54	5.61
4	B[illegible]	2.01	3.83	5.7	2.23	5.59	2.91	5.2	5.3
5	C[illegible]s	5.36	5.06	4.87	3.11	2.56	4.07	3.55	2.53
6	C[illegible]e	2.26	5.1	3.11	4.49	4	5.79	2.69	4.45
7	C[illegible]n	2.33	3.6	2.85	3.57	4.2	3.6	2.89	3.92
8	C[illegible]	5.83	6.25	4.82	4.83	1.19	6.31	2.58	1.77
9	C[illegible]Bar	6.03	6.48	4	4.31	1.13	5.94	2.28	1.55
10	C[illegible]Cookies	5.42	6	4.69	3.91	1.3	5.73	2.67	2.03
11	C[illegible]Milk	5.52	5.13	3.45	3.48	2.31	4.77	2.31	2.5
12	C[illegible]Yoghurt	4.86	4.62	3.16	3.53	3.26	4.43	2.42	3.26
13	C[illegible]	5.42	6.35	2.87	4.95	1.47	6.44	1.96	2.26
14	C[illegible]Cheese	4.65	4.76	2.79	4.06	2.81	5.18	2.49	3.52
15	E[illegible]	1.26	3.78	2.72	5.19	4.36	3.39	2.84	4.71
16	[illegible]s	5.1	4.95	3.36	4.35	2.66	5.42	2.84	3.11
17	FrenchFries	4.58	5.75	5.28	4.47	2.09	5.98	3.03	2.04
18	GrainBiscuits	3.26	3.58	4.99	2.24	4.37	2.71	5	4.72
19	IceTea	6.28	4.59	2.76	2.22	1.58	2.29	1.64	2.26

Objects
(here: food products)

Variables

Perkovic et al. (2022)

Factor analysis

Extracted factors

Variables

- Artificial additives
- Calorie content
- Good fat
- Level of packaging
- Level of processing
- Mineral content
- Natural production
- Local origin
- Sugar content
- Vitamin content
- Recommended proportion of a healthy diet

→ **Naturalness**

- Cholesterol content
- Fat content
- Protein content
- Salt content

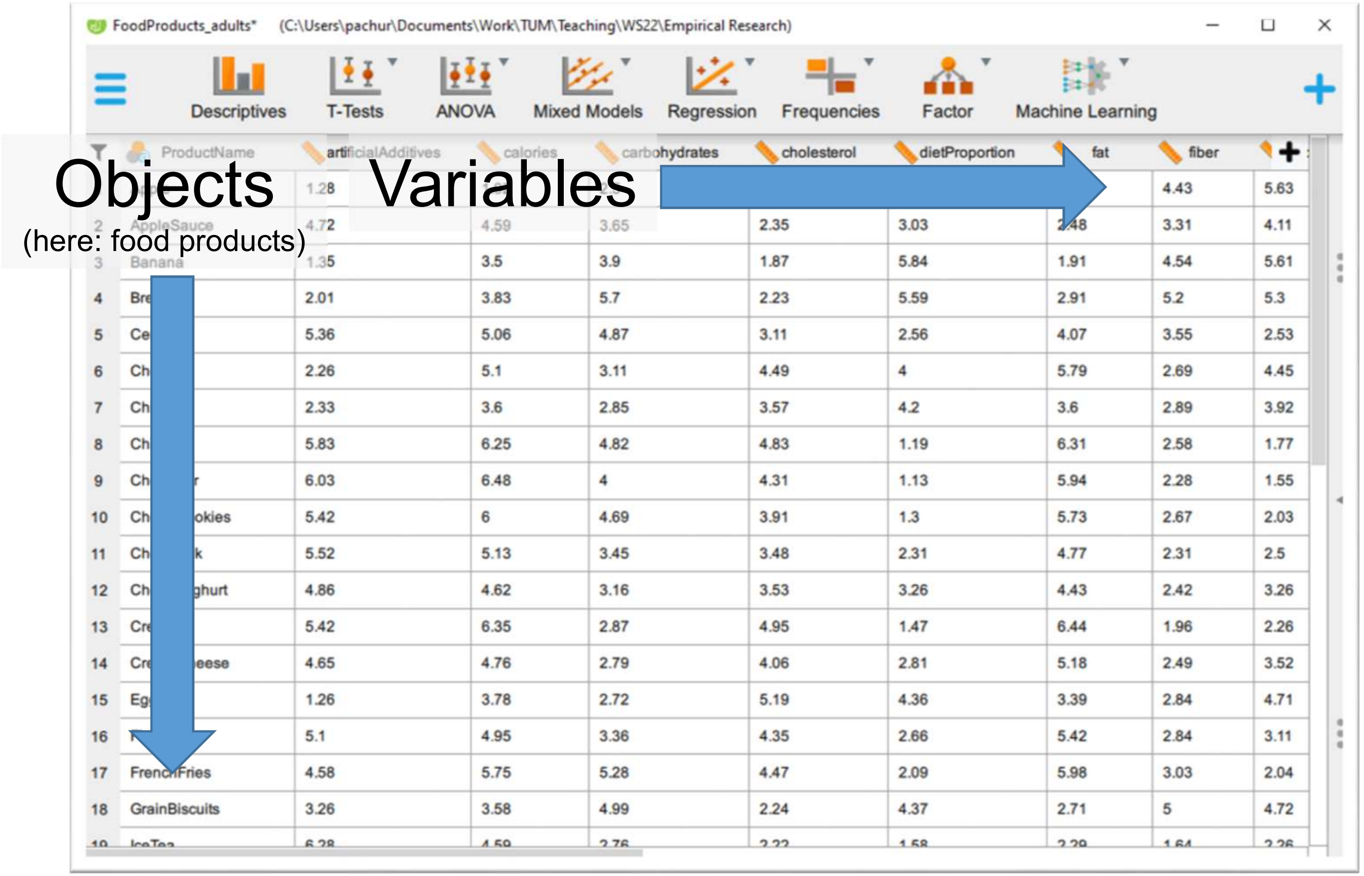
→ **Cholesterol and protein**

- Carbohydrate content
- Fiber content

→ **Good carbs**

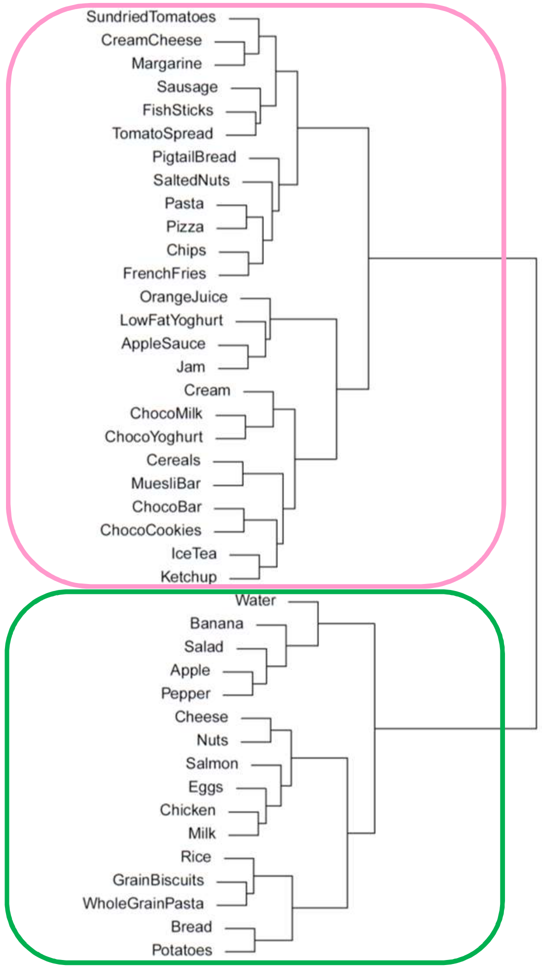
Perkovic et al. (2022)

Food ecology



Cluster analysis

Objects



Cluster 1 (“Processed food products“)

Cluster 2 (“Nonprocessed food products“)

Perkovic et al. (2022)

Why can it be useful to identify clusters?

Some examples

- Differentiate groups among dairy products based on characteristics (e.g., fat content, sugar content, popularity, image, price, customers) → identify market gaps
- Differentiate groups among consumers of beverage products based on socio-demographic variables (e.g., age, gender, income, etc.) → targeted communication
- Group the students of a university based on their characteristics (e.g., field of study, professional goals, gender, age, number of semesters, nationality) → develop tailored information events

Cluster analysis

Goal

Find clusters such that within a cluster the objects are as similar as possible (*internal homogeneity*) while at the same time between the clusters the objects are as distinct from each other as much as possible (*external heterogeneity*)

Procedure

1) Quantify **similarity** of objects (based on their values on a set of variables)
2) Group objects into **clusters** according to similarity
3) Determine the **optimal number** of clusters
4) **Interpret** the obtained clusters

A working example

Rating of five chocolate flavors on three characteristics

Object	Variable		
	Crunchy	Exotic	Sweet
Cookie	1	2	1
Nuts	2	3	3
Nougat	3	2	1
Cappuccino	5	4	7
Espresso	6	7	6

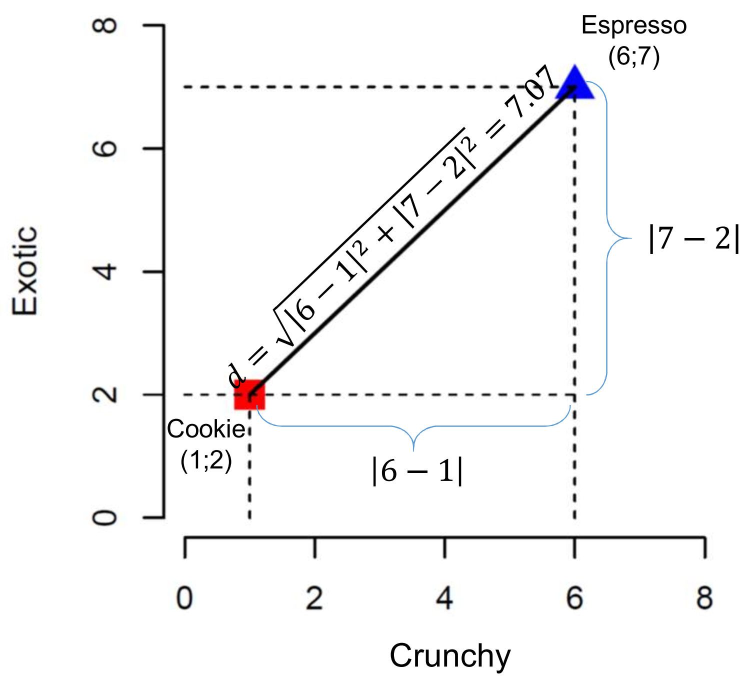
Variables need to be on the same scale
→ If they are not, z-standardize them !

Quantifying (dis)similarity: Distance

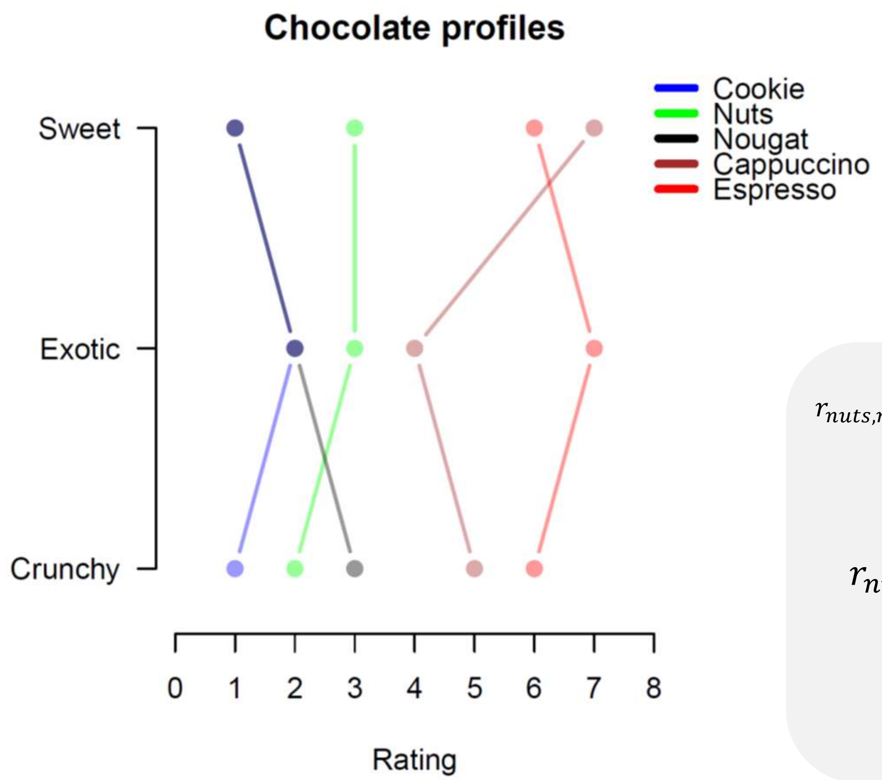
Euclidean distance

$$d_{a,b} = \sqrt{\sum_{j=1}^{J} \left| x_{aj} - x_{bj} \right|^2}$$

→ Higher values of *d* indicate lower similarity



Quantifying similarity: Correlation



$$r_{a,b} = \frac{\sum_{j=1}^{J}\left(x_j^a - \bar{x}^a\right) \times \left(x_j^b - \bar{x}^b\right)}{\sqrt{\sum_{J=1}^{J}\left(x_j^a - \bar{x}^a\right)^2} \times \sqrt{\sum_{J=1}^{J}\left(x_j^b - \bar{x}^b\right)^2}}$$

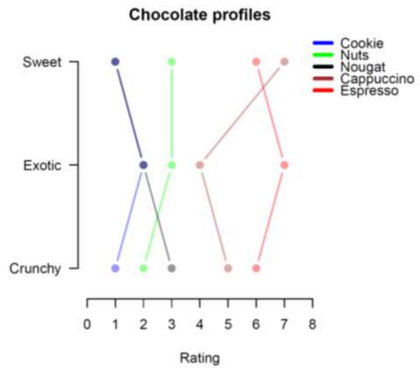
→ Higher values of *r* indicate higher similarity

$$r_{nuts,nugat} = \frac{\left(x_{sweet}^{nuts} - \bar{x}^{nuts}\right) \times \left(x_{sweet}^{nugat} - \bar{x}^{nugat}\right) + \left(x_{exotic}^{nuts} - \bar{x}^{nuts}\right) \times \left(x_{exotic}^{nugat} - \bar{x}^{nugat}\right) + \cdots}{\sqrt{\left[\left(x_{sweet}^{nuts} - \bar{x}^{nuts}\right)^2 + \left(x_{exotic}^{nuts} - \bar{x}^{nuts}\right)^2 + \cdots\right] \times \left[\left(x_{sweet}^{nugat} - \bar{x}^{nugat}\right)^2 + \left(x_{exotic}^{nugat} - \bar{x}^{nugat}\right)^2 + \cdots\right]}}$$

$$r_{nuts,nugat} = \frac{(3-2.67) \times (1-2) + (3-2.67) \times (2-2) + \cdots}{\sqrt{[(3-2.67)^2 + (3-2.67)^2 + \cdots]} \times \sqrt{[(1-2)^2 + (2-2)^2 + \cdots]}}$$

$$= \frac{-1}{1.15} = -.87$$

Similarity matrix



Euclidean distance (i.e., *dis*similarity)

	Cookie	Nuts	Nougat	Cappuccino	Espresso
Cookie	0				
Nuts	2.45	0			
Nougat	2	2.45	0		
Cappuccino	7.48	5.1	6.63	0	
Espresso	8.66	6.4	7.68	3.32	0

Correlation

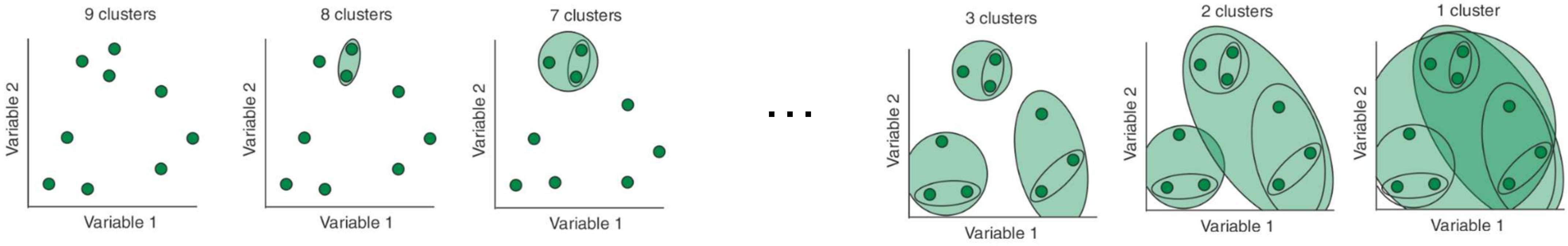
	Cookie	Nuts	Nougat	Cappuccino	Espresso
Cookie	1.00				
Nuts	.50	1.00			
Nougat	.00	-.87	1.00		
Cappuccino	-.76	.19	-.66	1.00	
Espresso	1.00	.50	.00	-.76	1.00

Similarity as distance or as correlation?
→ Depends on what should count as “similar“ in the given case
(e.g., companies with similar financial dynamics across the years might be considered as similar, even if they performed on different levels)

Creation of clusters: Hierarchical clustering

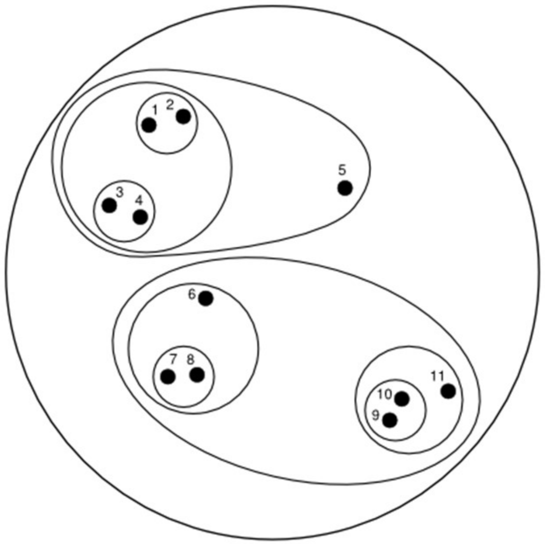
→ Clusters are created in a step-wise fashion

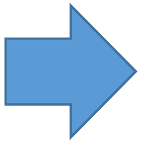
Agglomerative hierarchical clustering



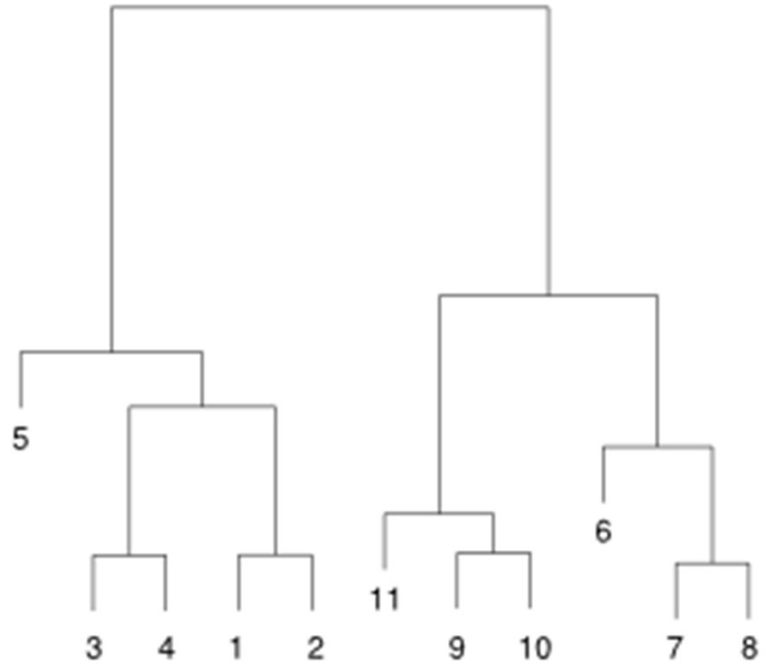
Divisive hierarchical clustering

Creation of clusters: Hierarchical clustering





Dendrogram



Procedure for creating clusters

Agglomerative hierarchical approach

Starting point

Calculate pairwise similarity between objects (based on distance or correlation)

Step 1

Merge the objects with the highest similarity into a cluster

Step 2

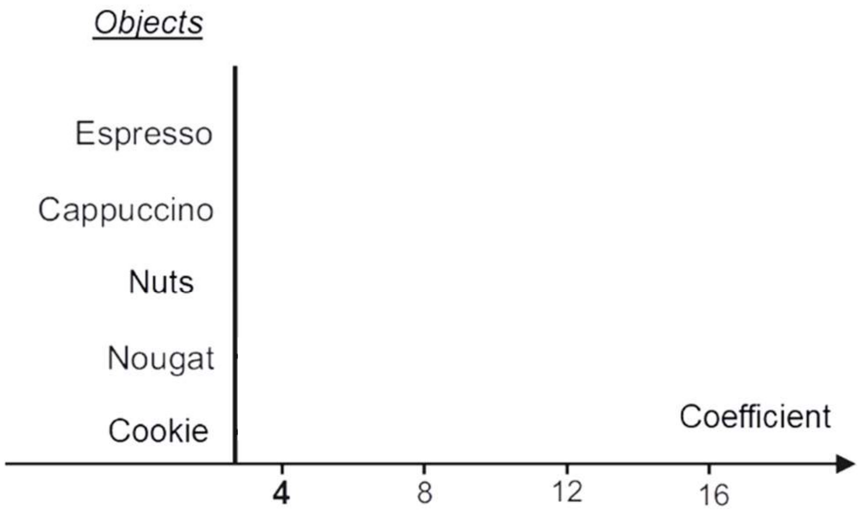
Calculate *linkage criterion* for new candidate clusters (based on the new cluster and the other objects or the already existing clusters)

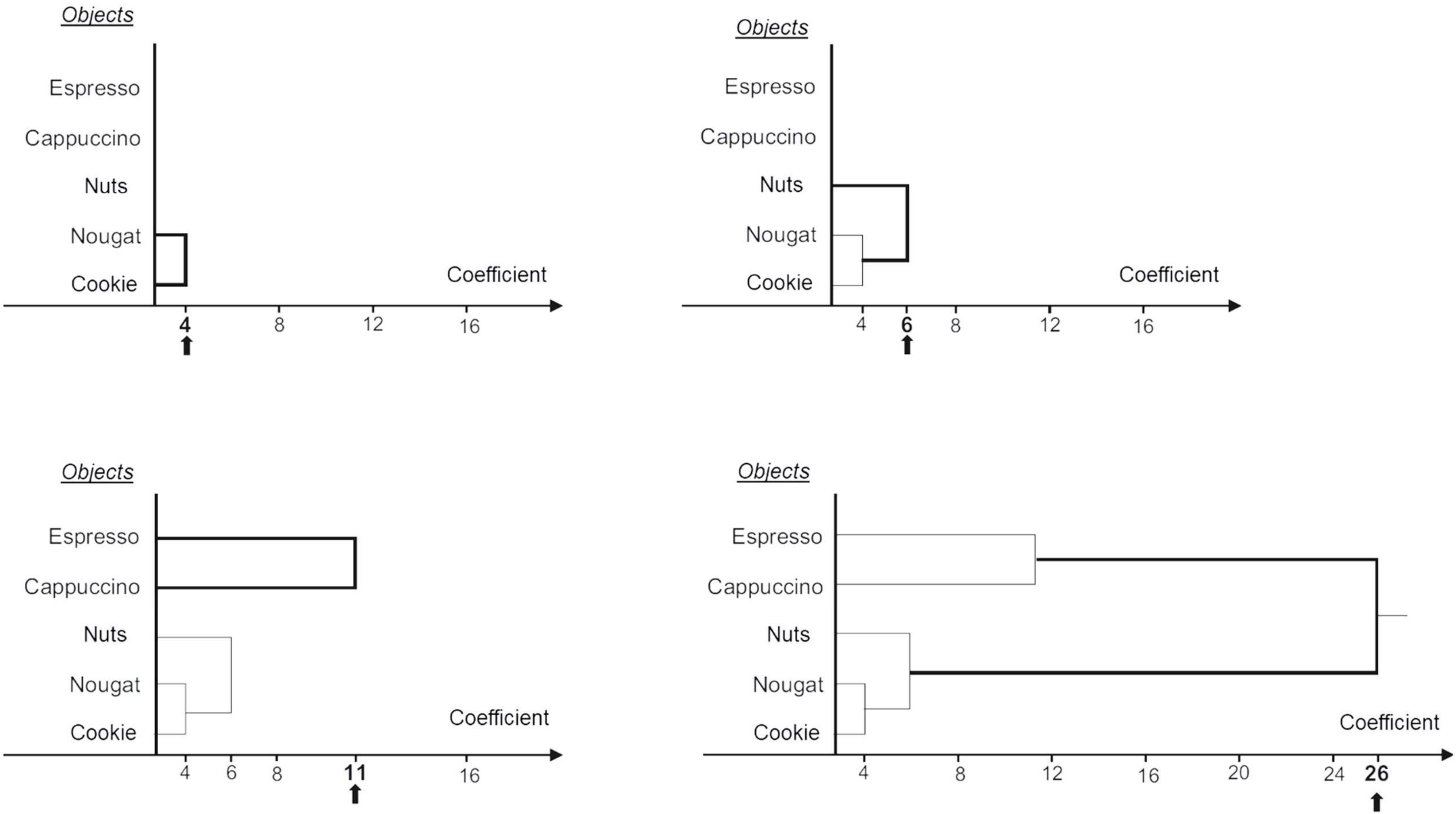
Step 3

Choose the candidate cluster that minimizes the linkage criterion

Step 4

Repeat until there is a single cluster


Objects
Espresso
Cappuccino
Nuts
Nougat
Cookie
Coefficient
4
8
12
16



Objects
Espresso
Cappuccino
Nuts
Nougat
Cookie
Coefficient
4
8
12
16
Objects
Espresso
Cappuccino
Nuts
Nougat
Cookie
Coefficient
4
6
8
12
16
Objects
Espresso
Cappuccino
Nuts
Nougat
Cookie
Coefficient
4
6
8
11
16
Objects
Espresso
Cappuccino
Nuts
Nougat
Cookie
Coefficient
4
8
12
16
20
24
26

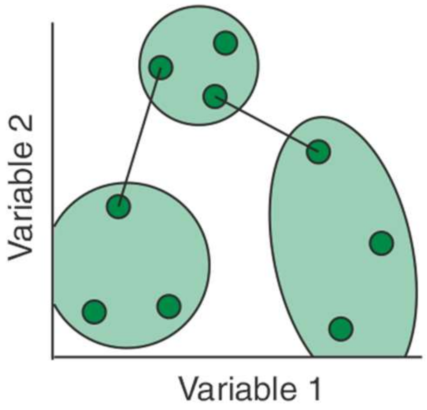
Linkage methods

→ Criteria for deciding which clusters to merge next

Single linkage

$$D(c_1, c_2) = \min_{x_1 \in c_1,\, x_2 \in c_2} (D(x_1, x_2))$$

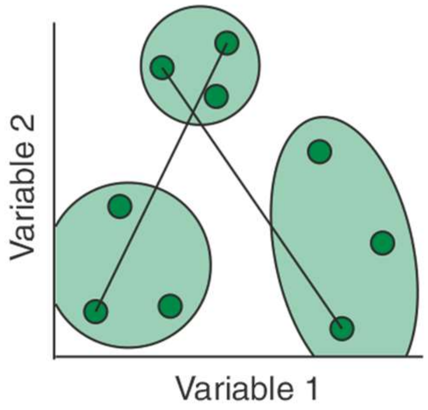
"Nearest neighbor"



Complete linkage

$$D(c_1, c_2) = \max_{x_1 \in c_1,\, x_2 \in c_2} (D(x_1, x_2))$$

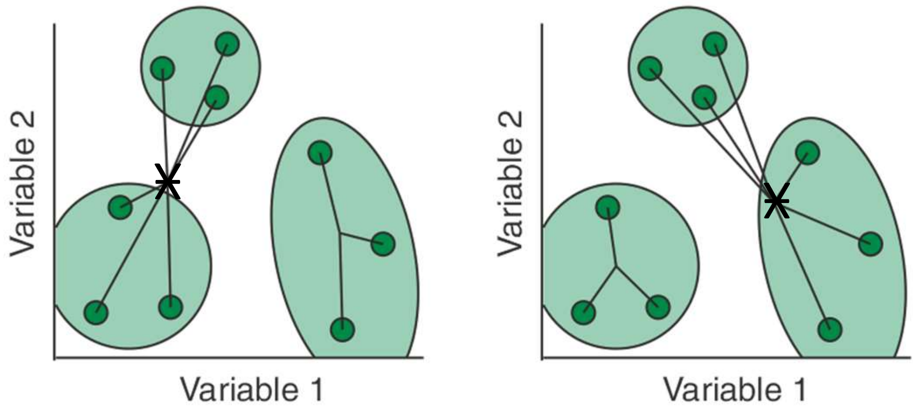
"Farthest neighbor"



Ward's method

(only when similarity is measured as distance)

$$TD_{c_1 \cup c_2} = \sum_{i=1}^{N_{x \in c_1 \cup c_2}} D\left(x_i, \bar{x}_{c_1 \cup c_2}\right)^2$$



Ward's method

Approach: Total distance (variance) of objects within a candidate new cluster *k* is minimized

Number of objects in candidate cluster *k*

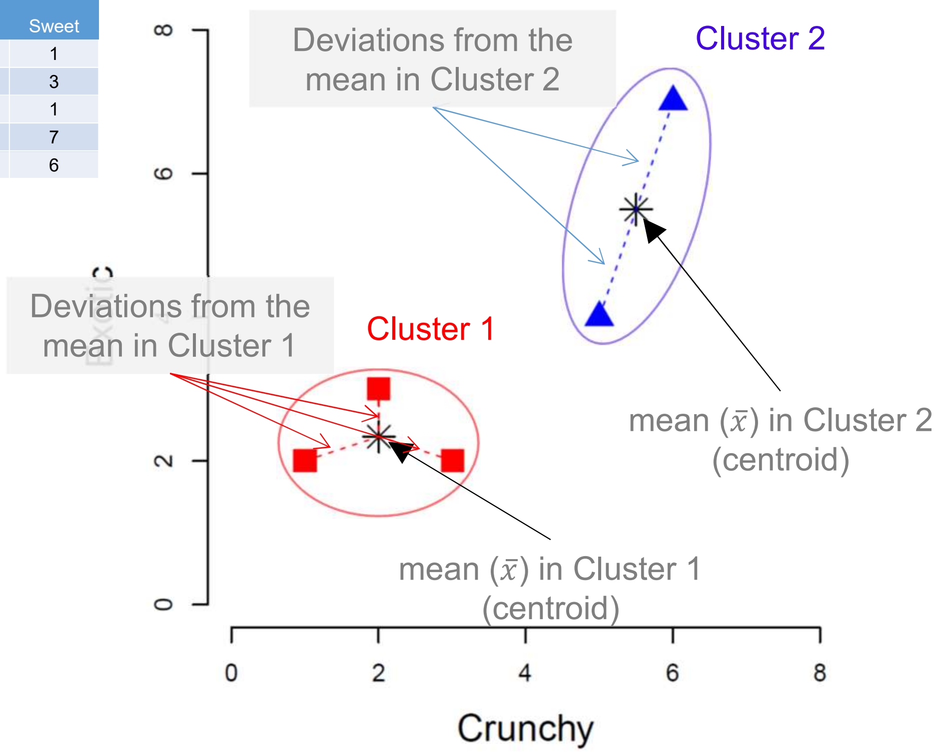
Number of variables

$$TD_{c_1 \cup c_2} = s_k^2 = \sum_{i=1}^{I_k} \sum_{j=1}^{J} (x_{ijk} - \bar{x}_{jk})^2$$

"Centroid"

Indices for model evaluation

Object	Variable		
	Crunchy	Exotic	Sweet
Cookie	1	2	1
Nuts	2	3	3
Nougat	3	2	1
Cappuccino	5	4	7
Espresso	6	7	6



Within-cluster sum of squares (WSS)

$$WSS = \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{j=1}^{J} (x_{ij} - \bar{x}_{jk})^2$$

"Centroid"

Sum across …

all clusters

all objects in a cluster

all variables

Indices for model evaluation

- Within-cluster sum of squares (WSS)

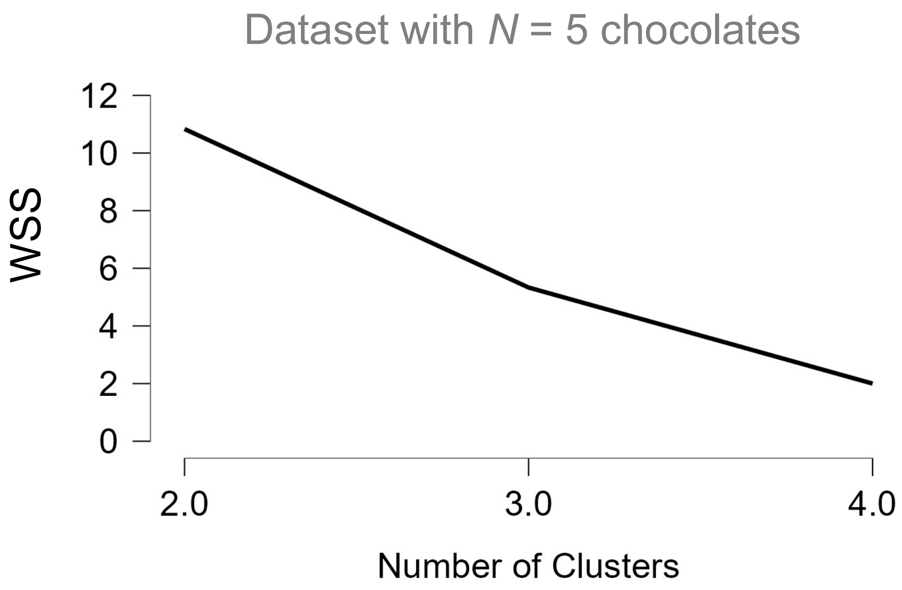
  → Pure measure of model fit: WSS is lower when there are more clusters (i.e., higher model complexity)

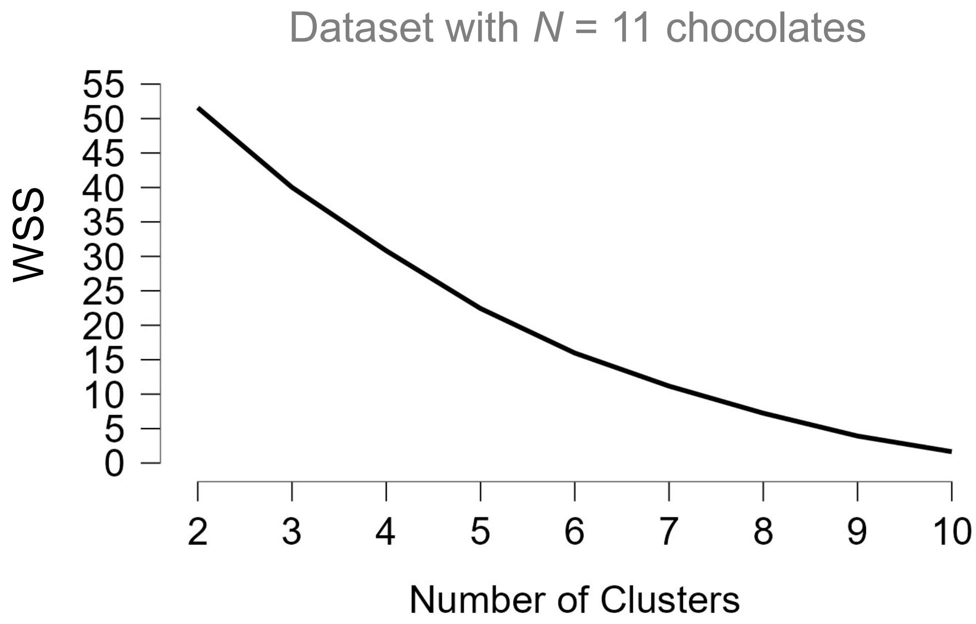
- Information criteria
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)

  → Both criteria trade off model fit against model complexity

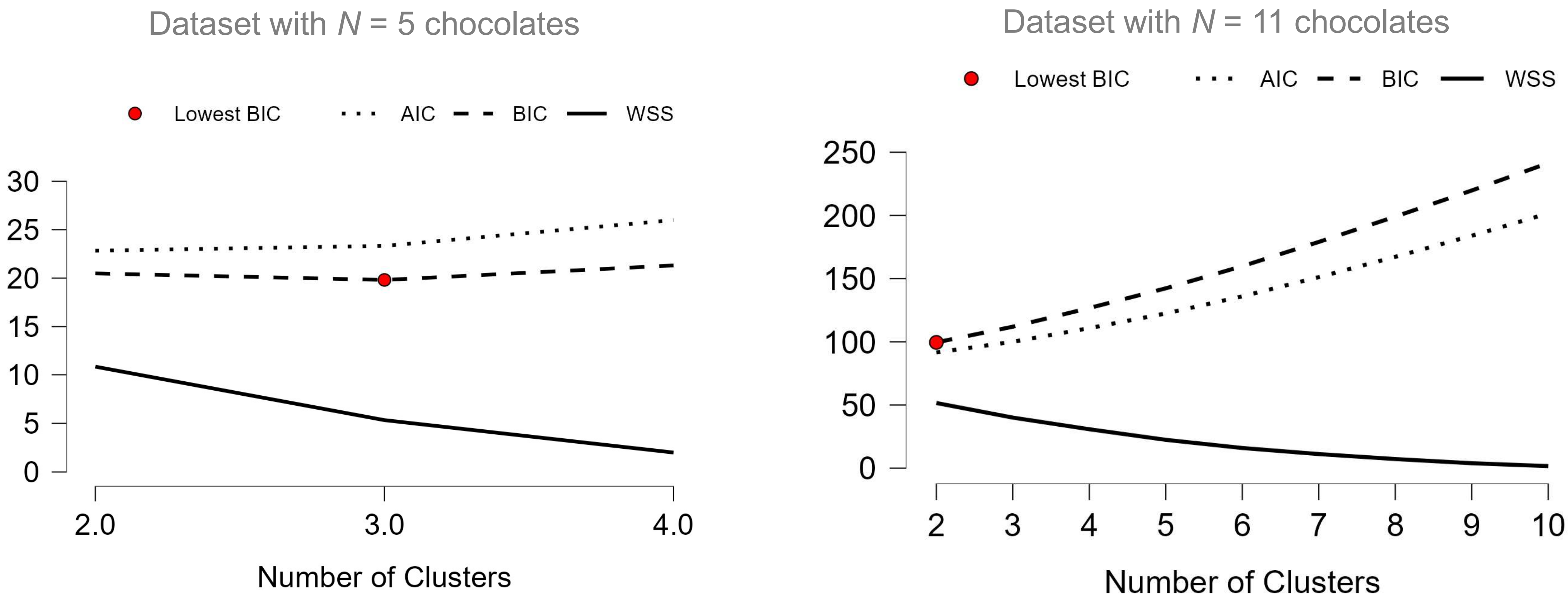
Which cluster solution should be retained?

Scree plot





Which cluster solution should be retained?



Silhouette coefficient

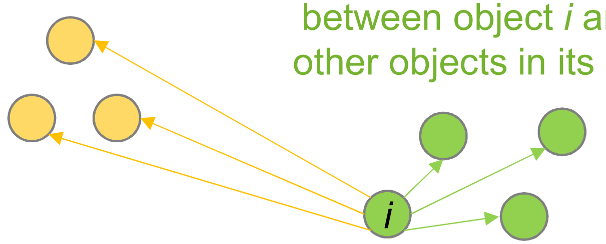
Indicates how clearly the clusters are separated

It measures how similar an object is to its own cluster (cohesion) relative to other clusters (separation)

- Range: −1 to +1. A high positive value indicates that the object is homogeneous with the other objects of its cluster and distinct from the neighboring cluster.
- If most objects have a **high positive Silhouette coefficient**, then the clustering configuration is **appropriate**. If many objects have a low or negative coefficient, then the clustering configuration may have too many or too few clusters.

***b*(*i*)**: Average distance between object *i* and the objects in the nearest other cluster

***a*(*i*)**: Average distance between object *i* and the other objects in its cluster



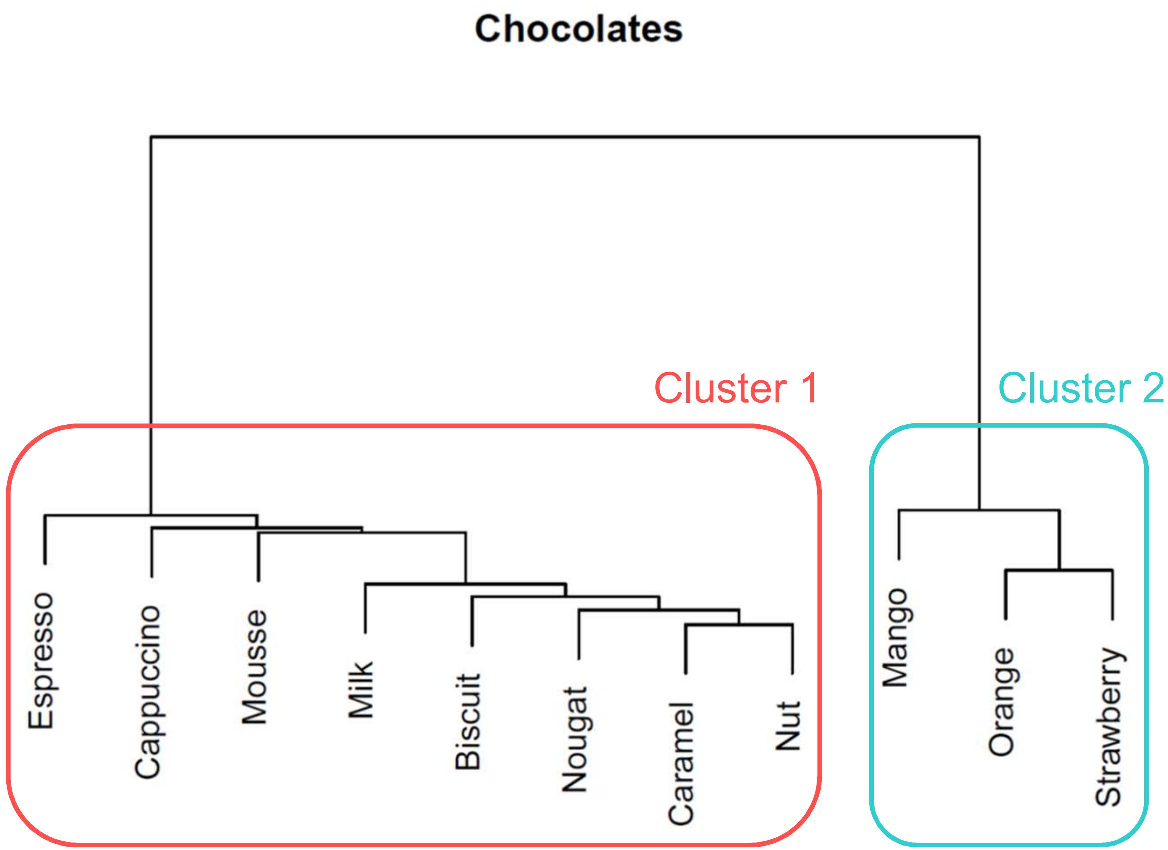
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$-1 \leq s(i) \leq 1$$

Overlapping clusters

Well-separated clusters

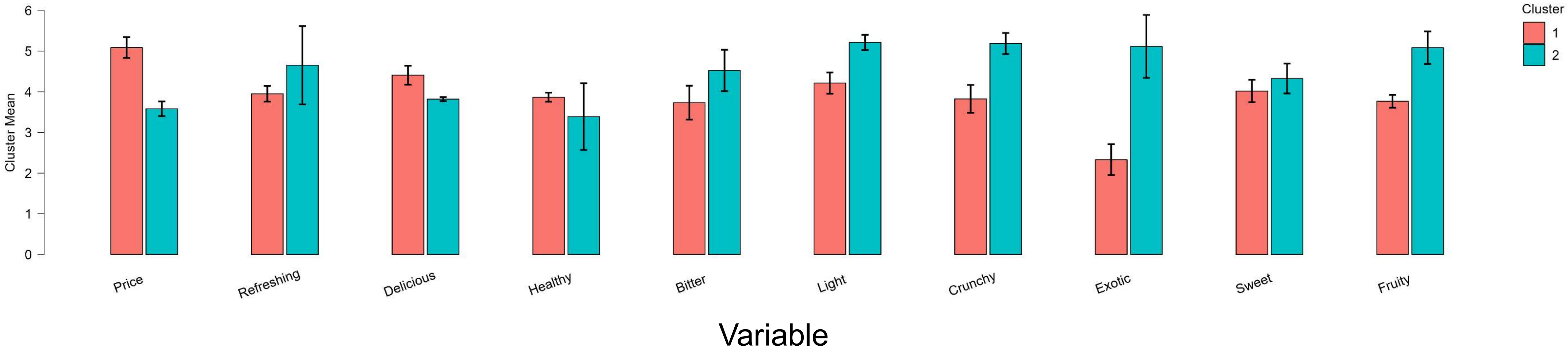
Interpreting the clusters



Interpreting the clusters

→ Average value of the objects in each cluster on the variables

Dataset with *N* = 11 chocolates (and 10 variables)



Self-quiz questions

- What is the goal of cluster analysis, and how does it differ from factor analysis?
- Give two ways to measure similarity and describe how they differ from each other
- What‘s the purpose of linkage methods, and how do single linkage, complete linkage, and Ward‘s method differ from each other?
- Give a measure to quantify the model fit of a cluster solution
- Give two measures of model performance of a cluster solution that trade off model fit against model complexity
- What do high and low values on the Silhouette score mean?

Background reading for next lecture

Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2021). Conjoint analysis. In: K. Backhaus, B. Erichson, S. Gensler, R. Weiber, & T. Weiber, *Multivariate analysis: An application-oriented introduction* (p. 531–598). Springer.