



Aalto University
School of Business

Cluster Analysis Tutorial

Pekka Malo

Assist. Prof. (statistics)

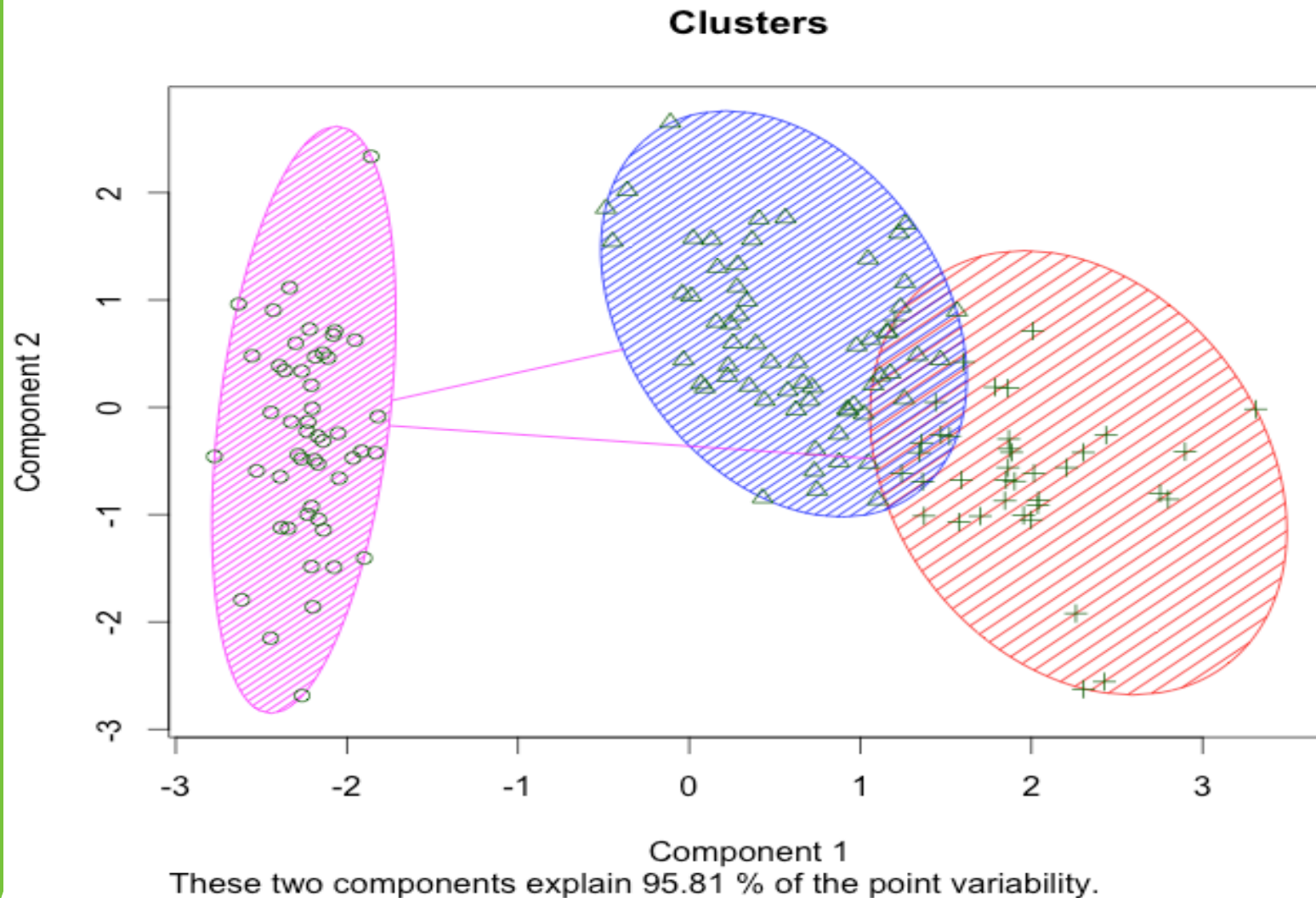
Business Intelligence (57E00500)
Autumn 2015



Learning objectives

- Understand the concept of cluster analysis
- Explain situations where cluster analysis can be applied
- Describe assumptions used in the analysis
- Know the use of hierarchical clustering and K-means cluster analysis
- Know how to use cluster analysis in SPSS
- Learn to interpret various outputs of cluster analysis

What is cluster analysis?



Cluster analysis is known by many names ...

Segmentation

Q-analysis

Numerical taxonomy

Classification

Unsupervised learning

Taximetrics

Learning without a teacher



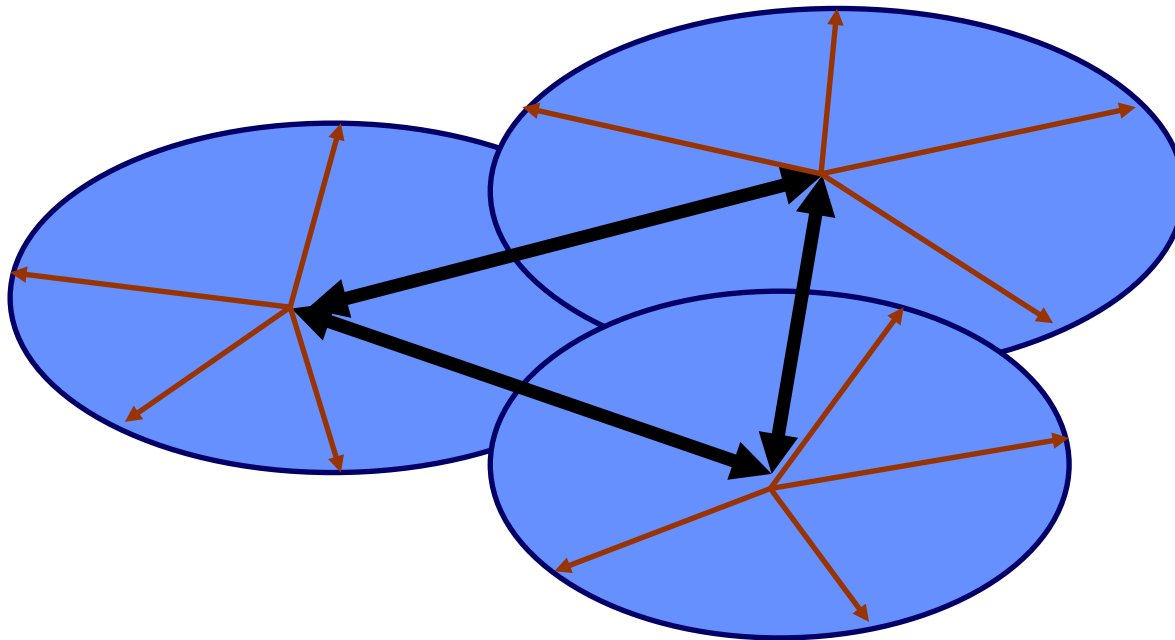
Purpose: Find a way to group data in a meaningful manner

Cluster Analysis (CA) ~ method for organizing data (people, things, events, products, companies, etc.) into meaningful groups or taxonomies based on a set of variables that describe the key features of the observations

Cluster ~ a group of observations, which are similar to each other and different from observations in other clusters

Objectives in Cluster Analysis

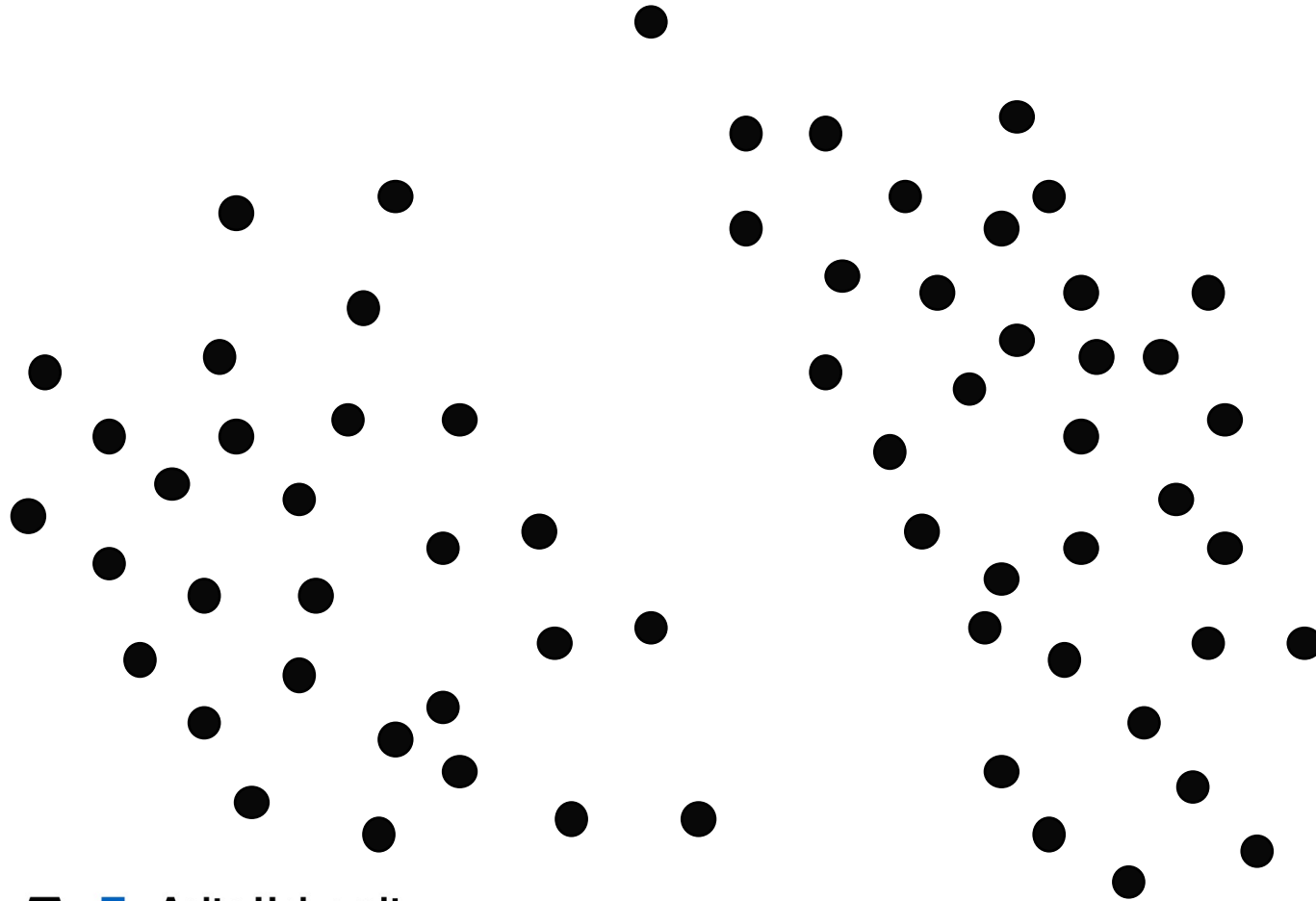
-  Between-Cluster Variation = Maximize
-  Within-Cluster Variation = Minimize



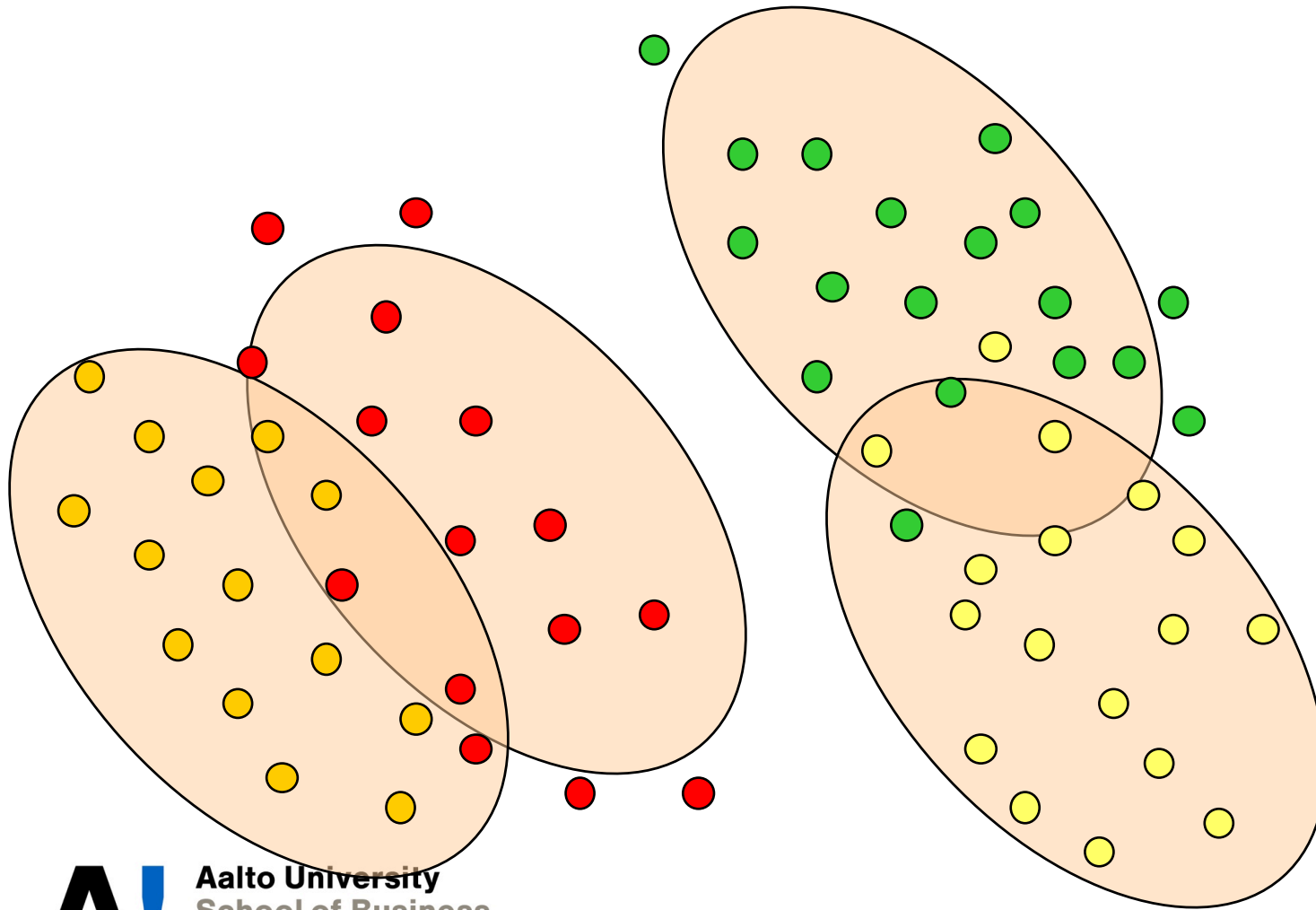
Within-groups vs. Between-groups

- **Within-groups property:** Each group is homogenous with respect to certain characteristics, i.e. observations in each group are similar to each other
- **Between-groups property:** Each group should be different from other groups with respect to the same characteristics, i.e. observations of one group should be different from the observations of other groups

How many clusters and how do you cluster ?



How many clusters and how do you cluster?



What we can do with cluster analysis?

- Detect groups which are statistically significant
 - Taxonomy description: Natural groups in data
 - Simplification of data: Groups instead of individuals
- Identify meaning for the clusters
 - Which relationships can be identified?
- Explain and find ways how they can be used

Clustering vs. classification

- Classification
 - We know the “groups” for at least some of the observations
 - Objective is to find a rule / function which correctly assigns observations into groups
 - Supervised learning procedure
- Clustering
 - We don't know the groups a priori
 - Objective group together points “which are similar”
 - Identify the underlying “hidden” structure in the data
 - Unsupervised learning procedure (i.e. no labeled data for training)

Clustering ~ “post hoc” segmentation

- Any discrete variable is a segmentation
 - E.g., gender, geographical area, etc.
- A priori segmentation
 - Use existing discrete variables to create segments
- Post hoc segmentation
 - Collect data on various attributes
 - Apply statistical technique to find segments



Cluster Analysis with SPSS

Techniques used:

- Hierarchical Clustering with Ward's method
- k-Means Clustering
- ANOVA and cross-tabulations

Example data: Luxury consumption and customer satisfaction



Data View in SPSS

Visible: 16 of 16 Variables

	experience	brand	price	quality	exclusivity	selection	spending	recommend	age	income	channel	gender	return	influence	QCL_1	LN
1	2	4	2	3	4	2	836	2	23	2	2	2	1	4	1	
2	2	2	3	2	2	4	1321	3	22	1	2	2	3	6	3	
3	3	4	2	2	3	1	798	2	23	2	2	2	3	2	1	
4	1	1	4	1	2	3	1255	3	56	2	2	2	2	5	3	
5	1	1	3	1	3	2	1270	2	45	2	2	2	2	1	3	
6	2	2	3	2	2	4	1197	3	51	3	1	1	3	6	3	
7	4	4	3	3	5	2	596	3	22	2	2	2	2	6	1	
8	4	4	4	3	4	5	2158	1	44	5	2	2	1	6	2	
9	1	1	3	2	2	2	1407	3	51	3	1	2	1	3	3	
10	3	3	3	4	4	3	2112	1	43	3	1	1	1	1	2	
11	4	4	2	4	4	1	777	1	29	4	3	2	1	2	1	
12	3	3	1	2	3	2	702	2	25	4	2	2	2	1	1	
13	1	1	3	2	2	3	1247	1	47	4	2	1	3	5	3	
14	3	3	3	3	3	2	848	1	34	4	2	2	2	2	1	
15	2	2	3	1	2	3	1220	2	59	4	2	2	2	5	3	
16	3	4	4	3	4	4	1933	2	46	4	2	2	3	3	2	
17	4	3	5	3	5	4	2033	3	33	5	3	1	1	3	2	
18	2	1	3	3	2	3	1235	3	64	5	3	2	3	5	3	
19	1	1	3	1	1	3	1245	3	54	1	2	2	1	4	3	
20	2	1	2	2	2	1	1278	3	52	5	1	1	1	4	3	

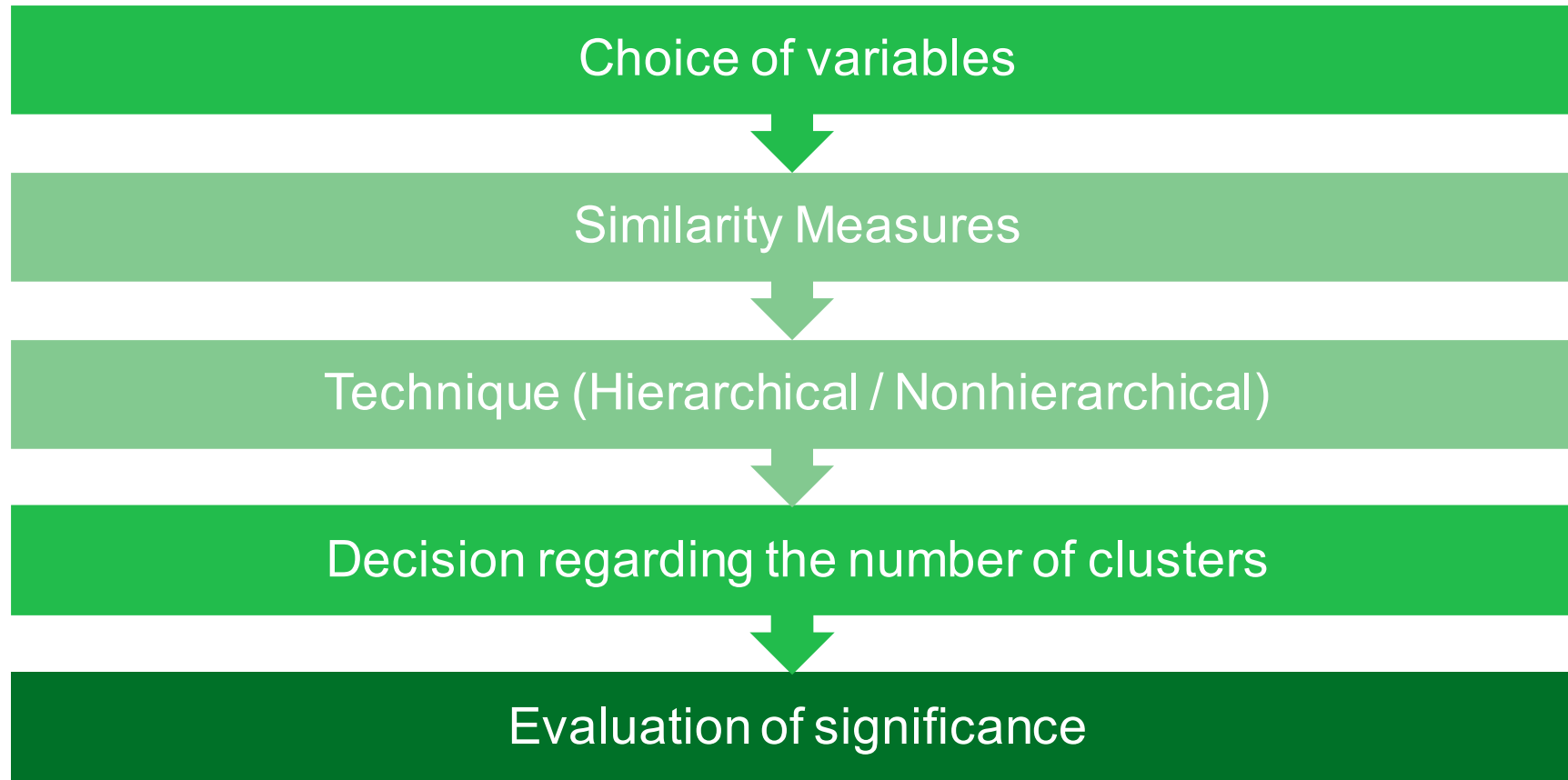
Data View Variable View

IBM SPSS Statistics Processor is ready

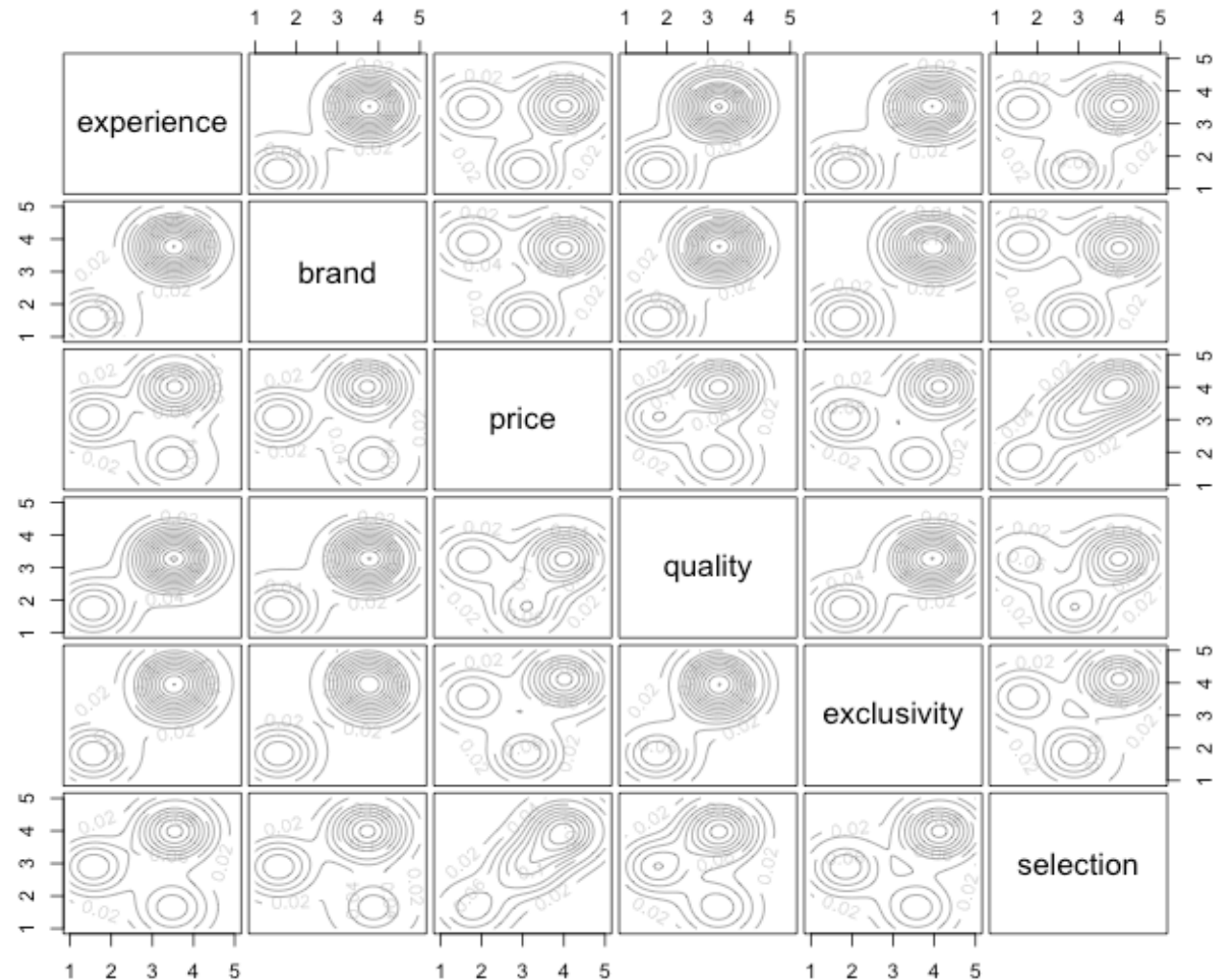
Sample size considerations

- Representativeness: The sample used for obtaining the cluster analysis should be representative of the population and its underlying structure (in particular the potential groups of interest)
- Minimum group sizes based on relevance to research question and confidence needed in characterization of the groups

Phases of Clustering



Step 1. Goals and choice of variables: Cluster by customer satisfaction



General note on choice of variables

- No theoretical guidelines
- Driven by the problem and practical significance
 - Do the variables help to characterize the objects?
 - Are the variables clearly related to the objectives?
- Warning:
 - Avoid including variables “just because you can”
 - Results are dramatically affected by inclusion of even one or two inappropriate or undifferentiated variables

Step 2: Choice of similarity measure

Interobject similarity is an empirical measure of correspondence, or resemblance, between objects to be clustered.

How close or similar are two observations?

Types of similarity measures

- Distance (or dissimilarity) Measures
 - Euclidean Distance
 - Minkowski Metric
 - Euclidean Distance for Standardized Data
 - Mahalanobis Distance
- Association Coefficient
- Correlation Coefficient
- Subjective Similarity

Distance Measures

- *Minkowski metric between cases i and j:*

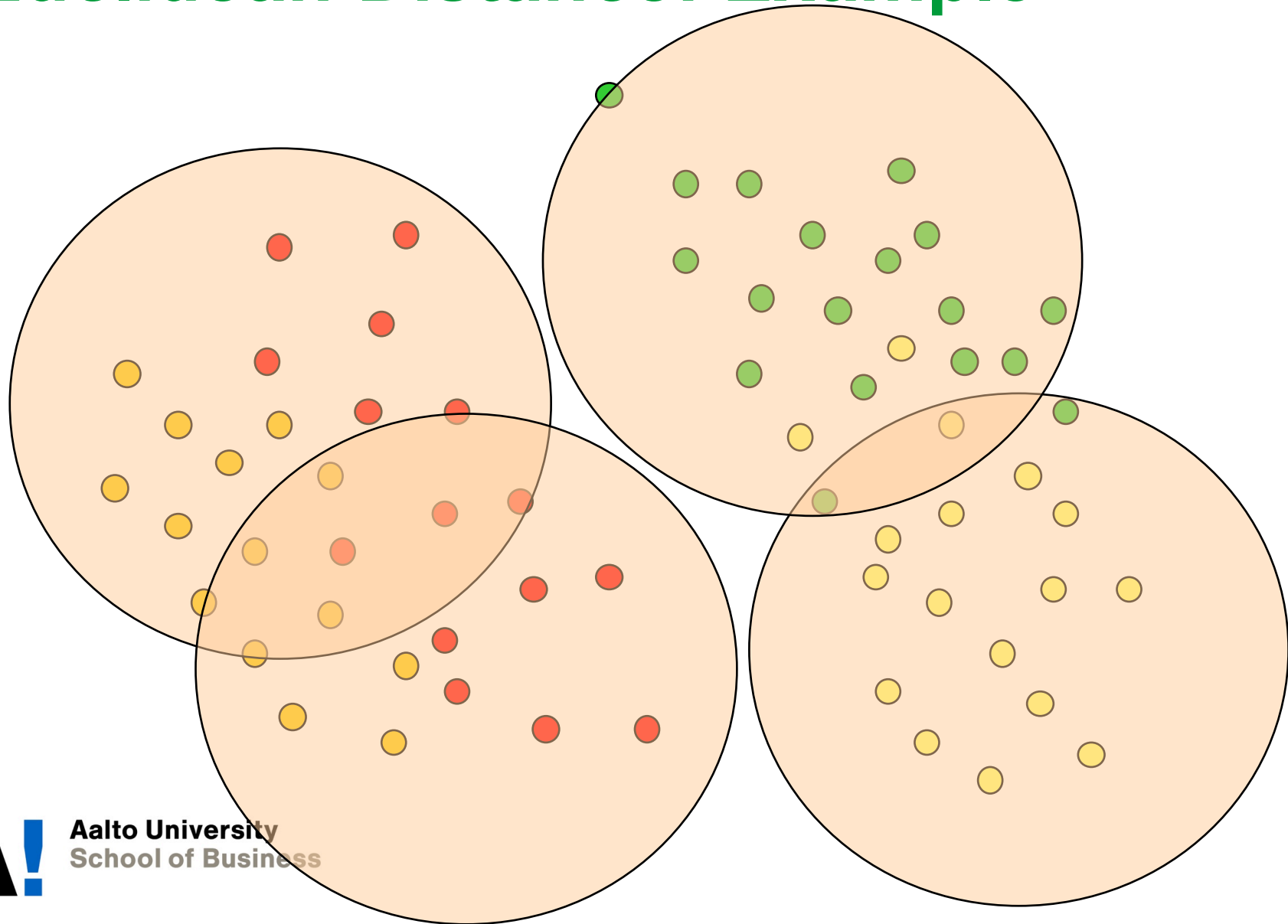
$$D_{i,j} = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^s \right)^{1/s}$$

- X_{ik} = measurement of ith case on kth variable
- $s = 2$: Euclidean Distance
- $s = 1$: City-block Distance
- p = number of variables

	Age	Income	Qualification
Employee 1	2.5	2.4	2.4
Employee 2	2.3	2.1	1.9
Difference	0.2	0.3	0.5
Squared difference	0.04	0.09	0.25

	Age	Income	Qualification
Employee 1	2.5	2.4	2.4
Employee 2	2.3	2.1	1.9
Absolute difference	0.2	0.3	0.5

Euclidean Distance: Example



Standardization of variables

- Note: Euclidean distance depends on the scale of the variables! Variables with large values will contribute more to distance
- Standardization of variables is commonly preferred to avoid problems due to different scales
- Most commonly done using Z-scores
- If groups are to be formed based on respondents' response styles, then within-case or row-centering standardization can be considered

Distance Measures ...

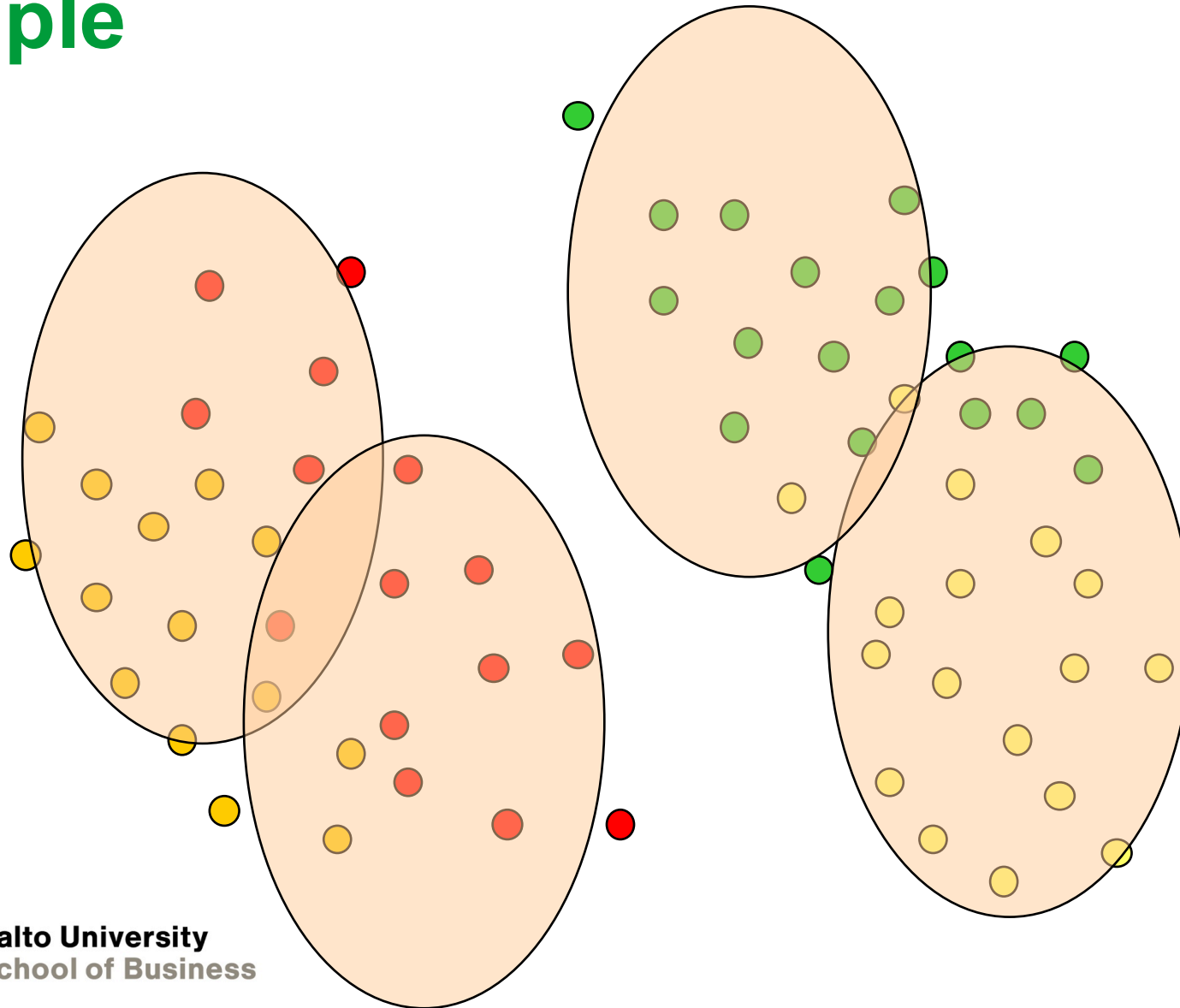
- C = covariance matrix between variables
- *Euclidean Distance for Standardized Data:*

$$\begin{aligned}SD_{ij} &= (X_i - X_j) \text{diag}(C)^{-1/2} (X_i - X_j)^T \\ &= \left(\sum_{k=1}^p \left(\frac{x_{ik} - x_{jk}}{\sqrt{c_{kk}}} \right)^2 \right)^{1/2}\end{aligned}$$

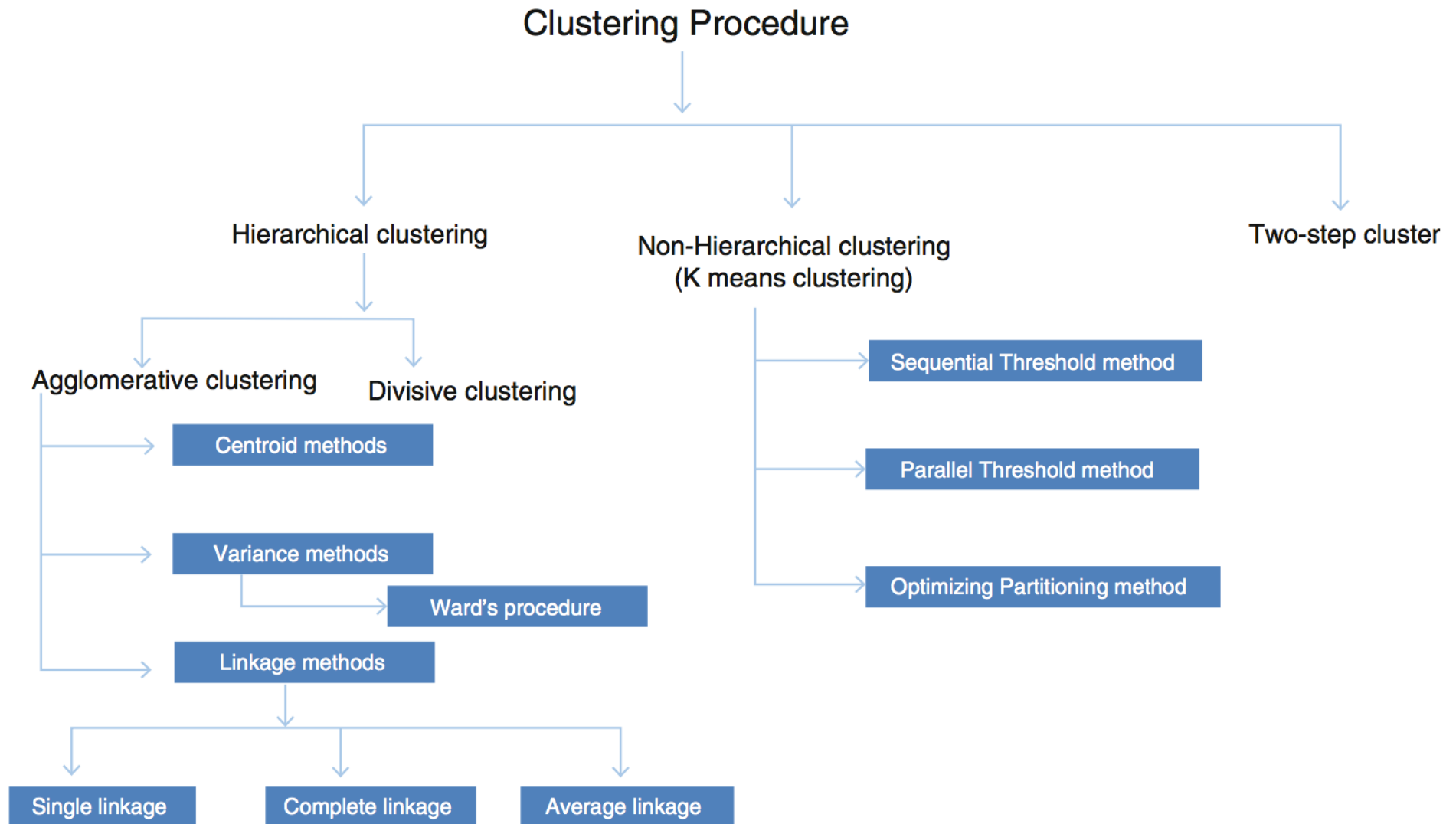
- *Mahalanobis (or Correlation) Distance*

$$MD_{ij}^2 = (X_i - X_j) C^{-1} (X_i - X_j)^T$$

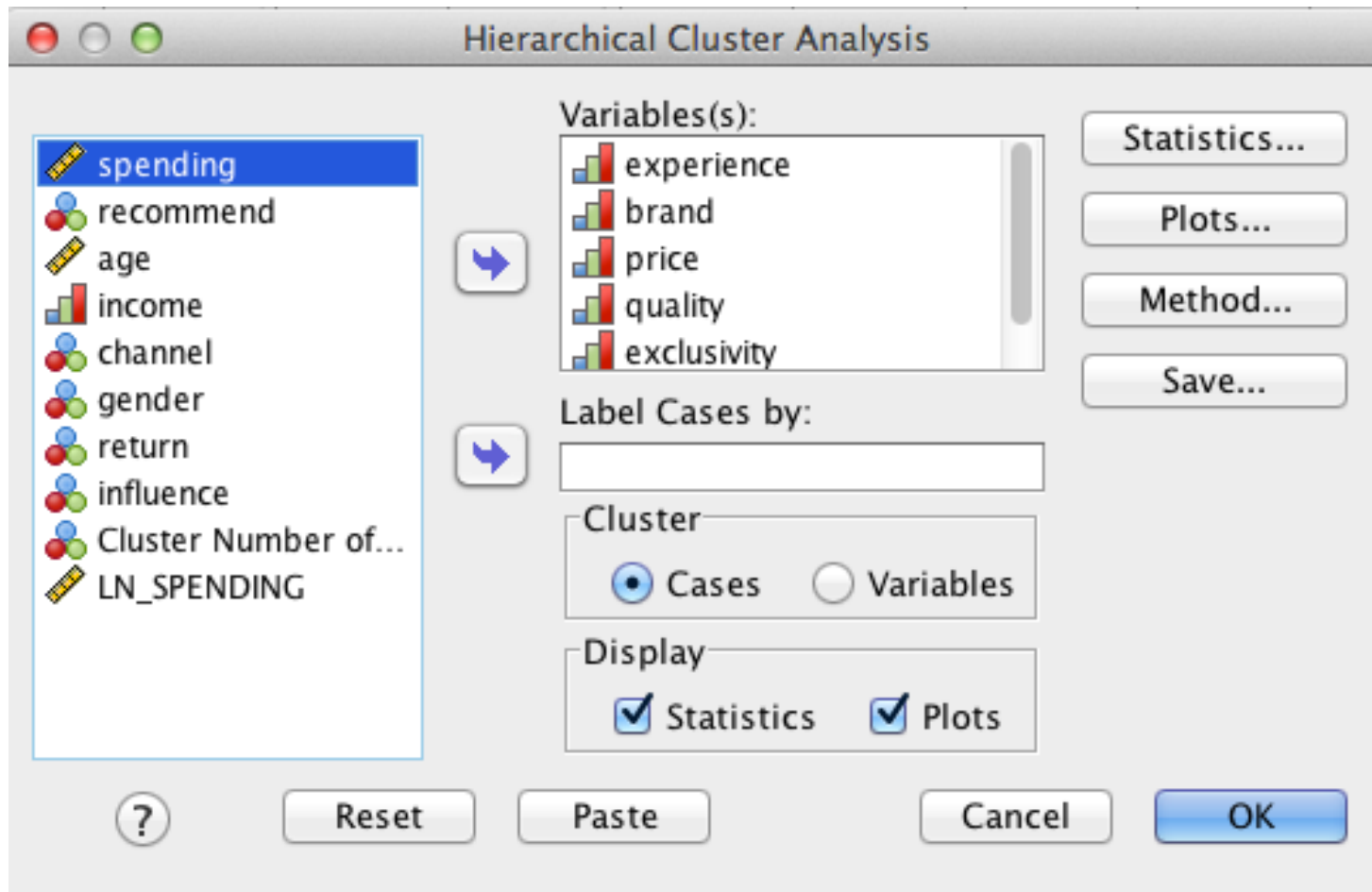
Standardized Euclidean Distance: Example



Step 3: Choice of clustering procedure



Hierarchical Clustering with SPSS



Agglomerative vs. Divisive

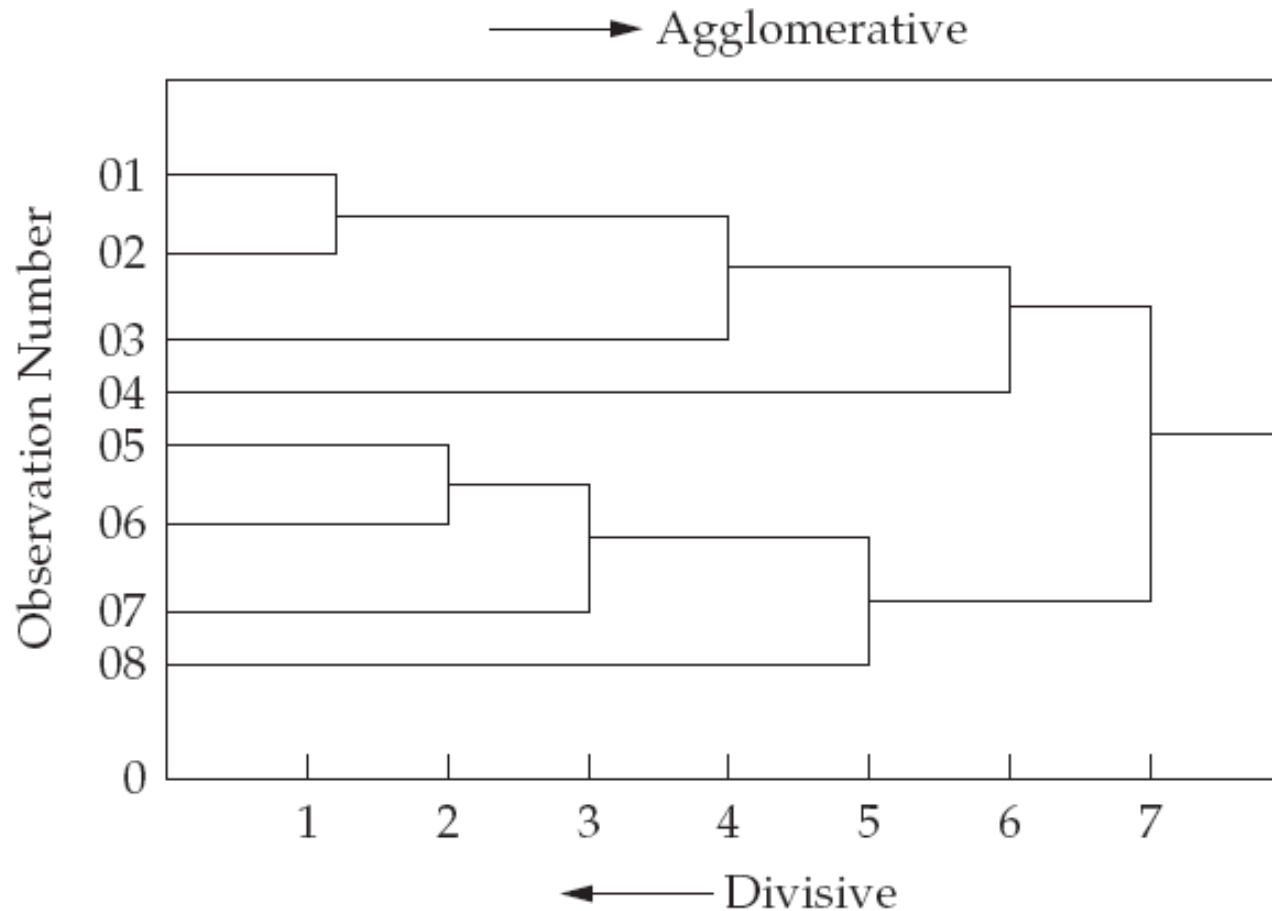


FIGURE 9-8 Dendrogram Illustrating Hierarchical Clustering

Source: Hair et al. (2010)

Hierarchical Clustering

- Centroid method
- Linkage methods
 - Nearest-neighbor or single-linkage method
 - Farthest-neighbor or complete-linkage method
 - Average linkage method
- Variance methods
 - Ward method

How agglomerative approaches work?


Start with all observations as their own cluster



Use selected similarity measure to combine two most similar observations into a new cluster of two observations



Repeat the procedure using the similarity measure to group together the most similar observations or combinations of observations into another new cluster

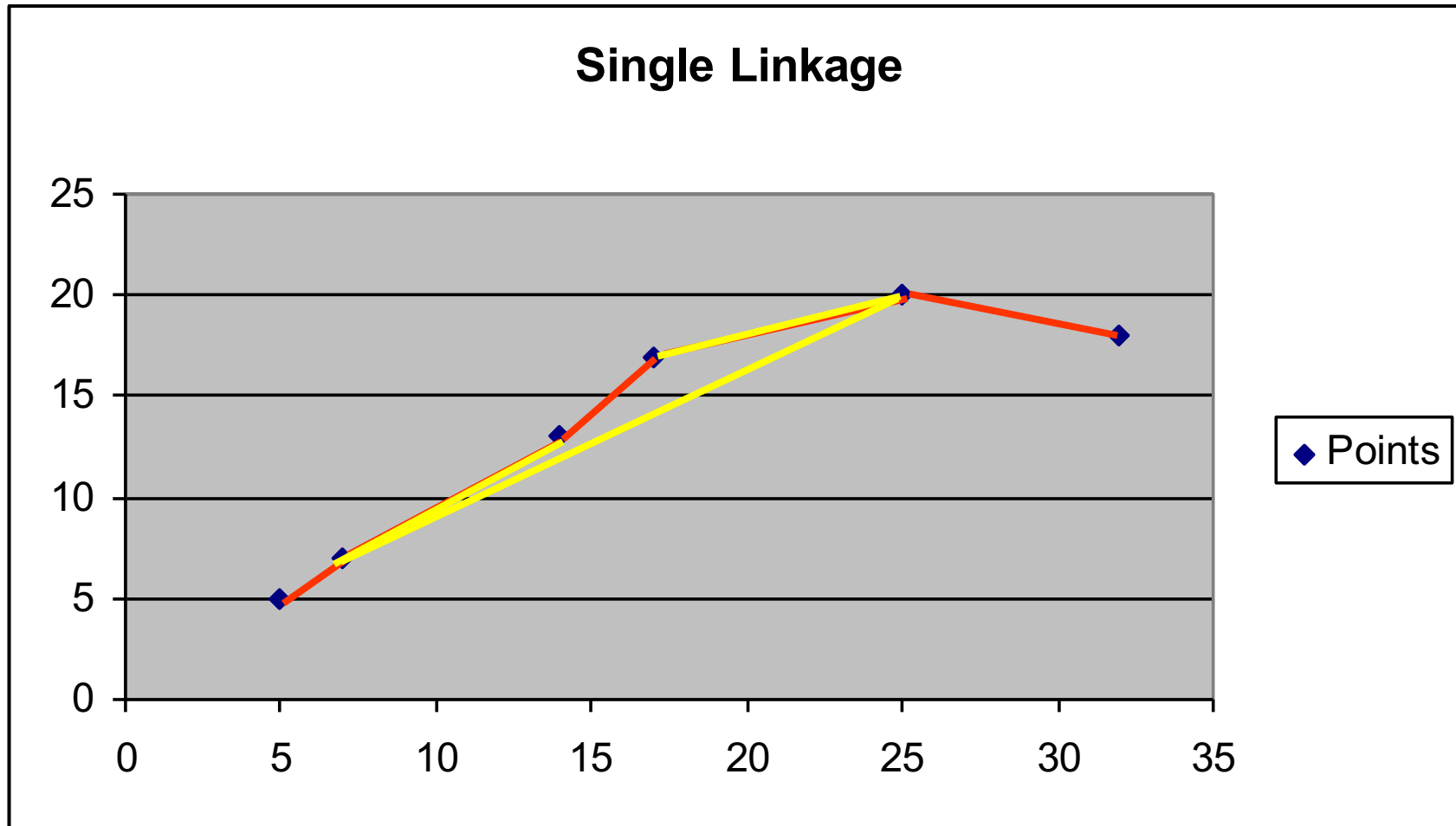


Continue until all observations are in a single cluster

Example: Single Linkage Method

- Principle
 - The distance between two-clusters is represented by the **minimum** of the distance between all possible pairs of subjects in the two groups

Example: Single-Linkage Method



Linkage methods

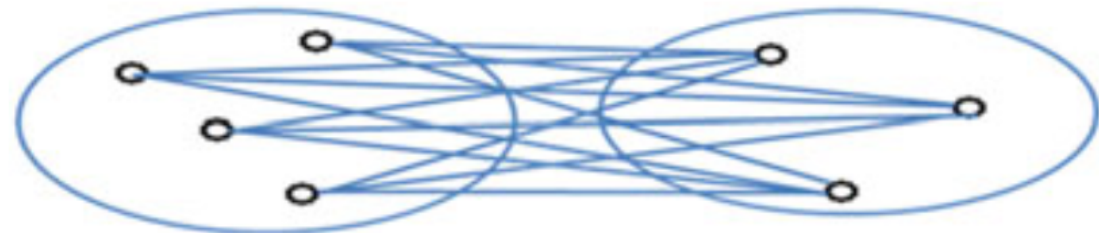
Single-linkage



Complete-linkage

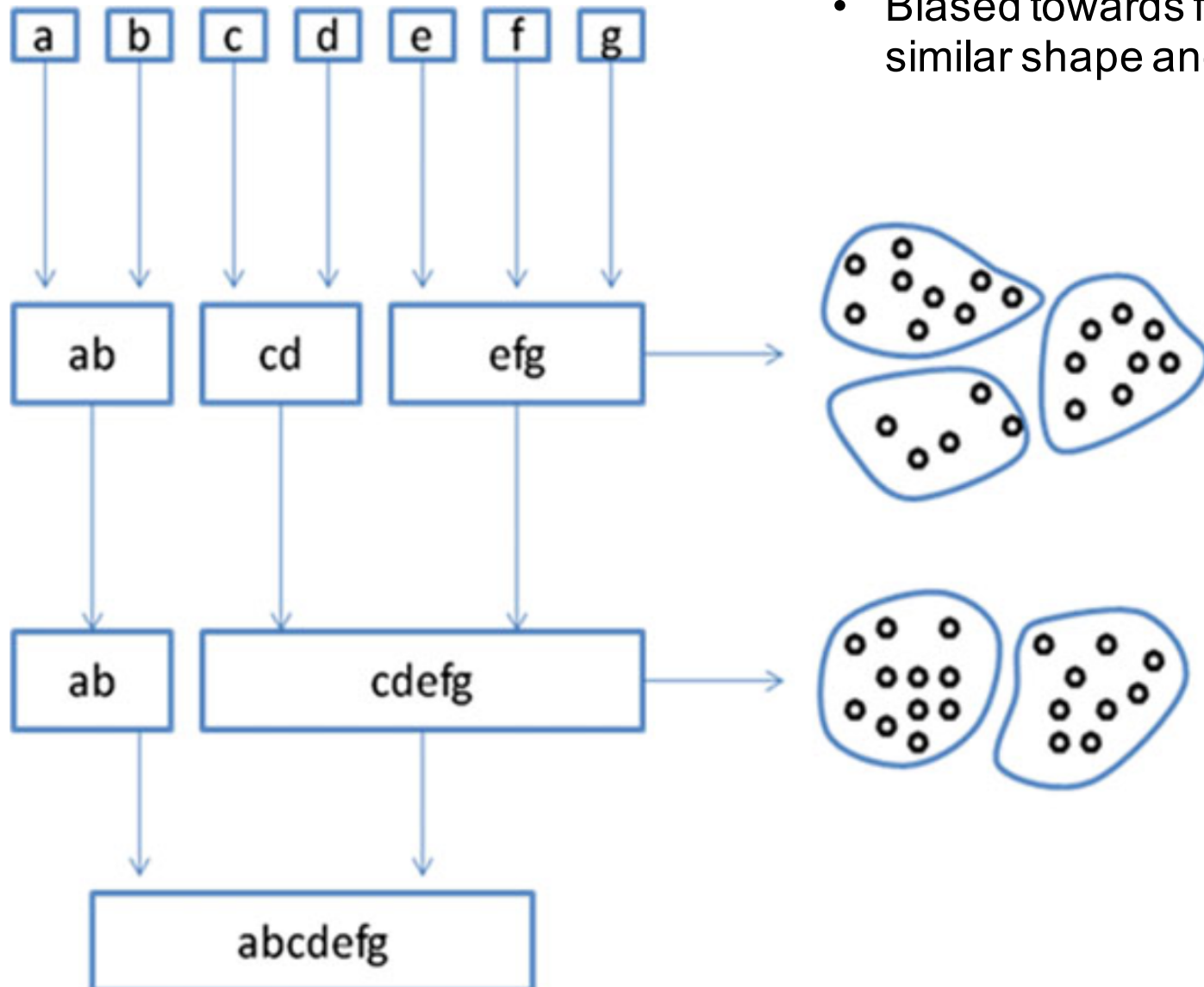


Average linkage



Example: Ward's method (variance linkage)

- Minimize variance within cluster
- Biased towards forming clusters of similar shape and size



Choice of hierarchical approach

Pros and cons

- Single-linkage
 - Most versatile, but poorly delineated cluster structures in a dataset may lead to snakelike cluster-chains
- Complete-linkage
 - No chaining, but impacted by outliers
- Average linkage
 - Considers average similarity of all individuals in a cluster
 - Tends to generate clusters with small within-cluster variation
 - Less affected by outliers

Choice of hierarchical approach (cont'd)

Pros and cons

- Ward's method
 - Uses total sum of squares within clusters
 - Most appropriate when equally sized clusters are expected
 - Easily distorted by outliers
- Centroid linkage
 - Considers difference between cluster centroids
 - Less affected by outliers

Hierarchical Cluster Analysis: Method

Cluster Method: **Ward's method**

Measure

Interval: **Squared Euclidean distance**
Power: **2** Root: **2**

Counts: **Chi-squared measure**


Binary: **Squared Euclidean distance**
Present: **1** Absent: **0**

Transform Values

Standardize: **None**
 By variable
 By case:

Transform Measure

Absolute values
 Change sign
 Rescale to 0-1 range

 **Cancel** **Continue**

Choosing the number of clusters

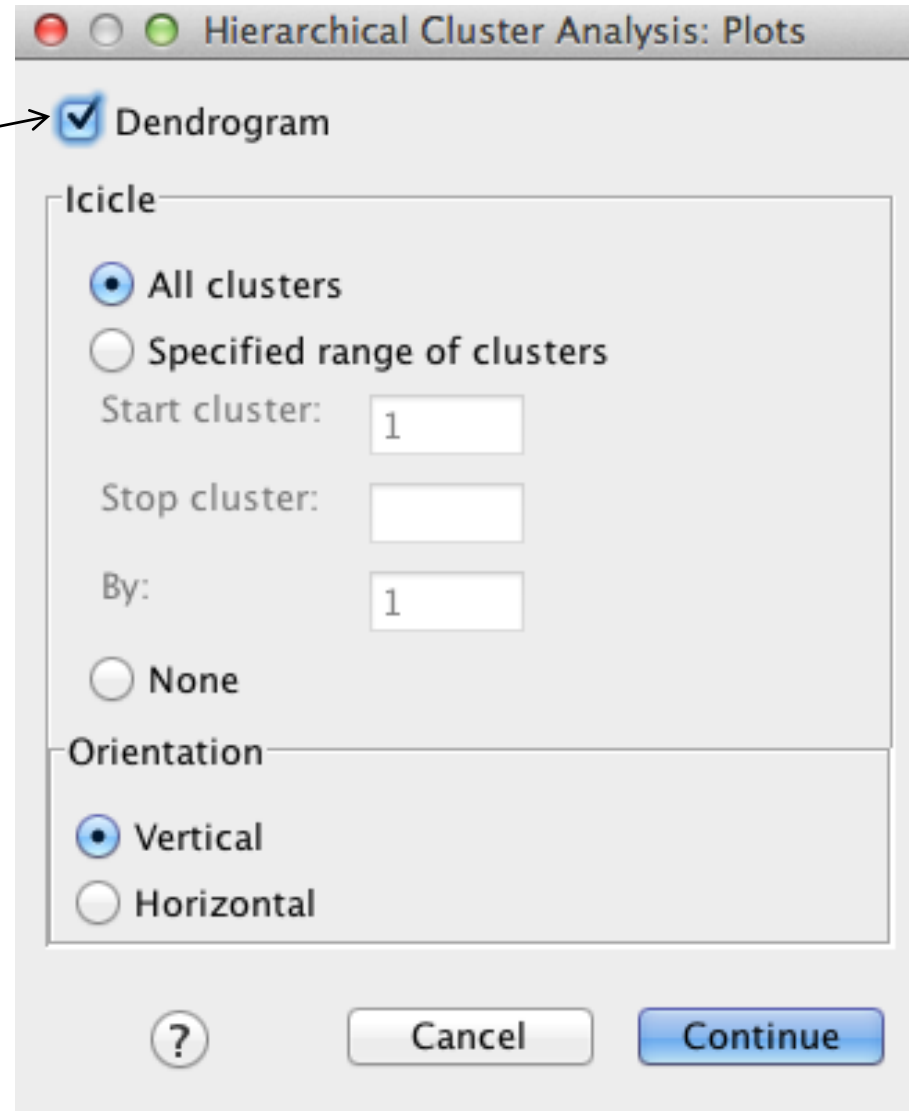
- No single objective procedure
- Evaluation based on following considerations:
 - Occurrence of single-member or extremely small clusters is not acceptable and should be eliminated
 - Ad-hoc stopping rules in hierarchical methods based on the rate of change in total similarity measure as the number of clusters increases or decreases
 - Clusters should be significantly different across the set of variables
 - Solutions must have theoretical validity based on external validation

Measures of heterogeneity change

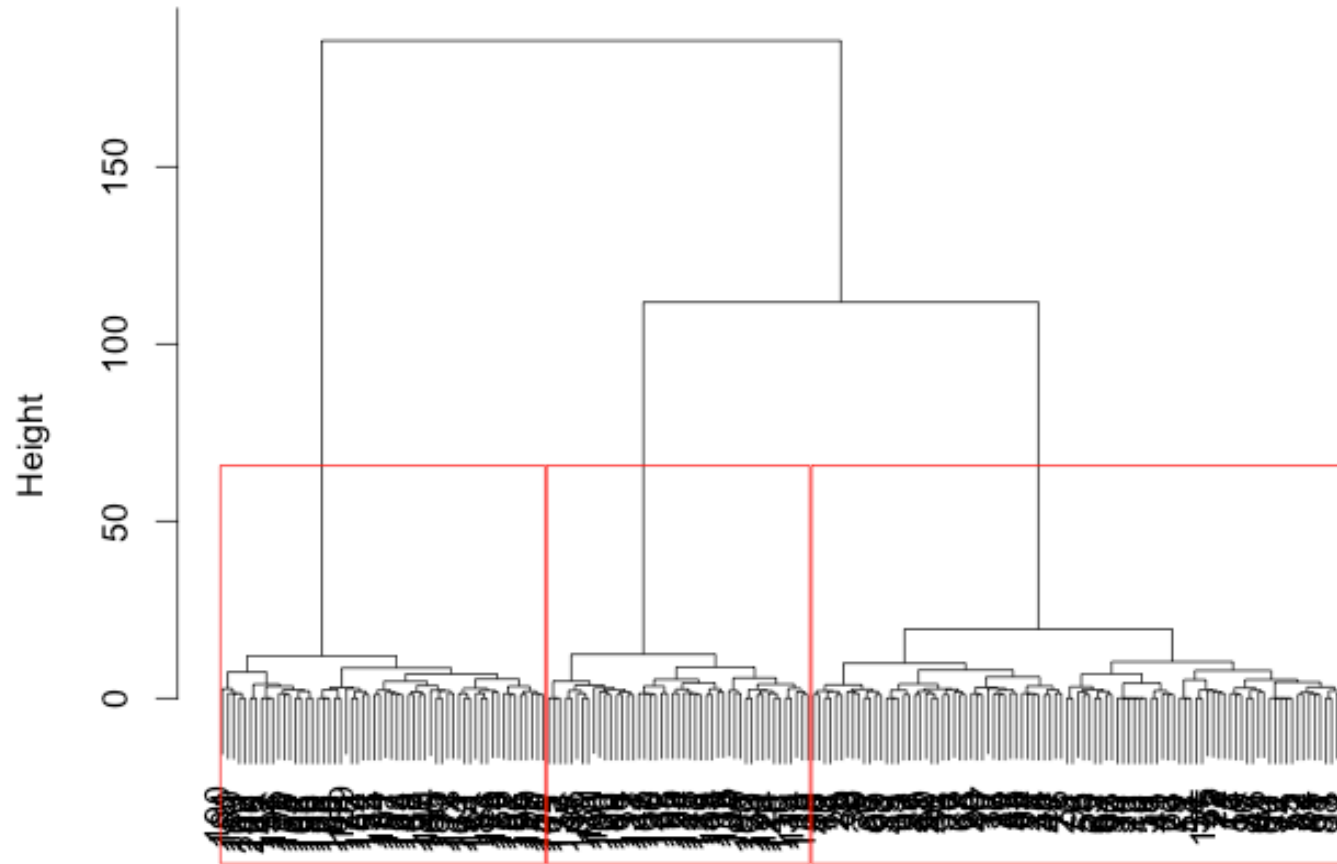
- Percentage changes in heterogeneity
 - E.g. use of agglomeration coefficient in SPSS, which measures heterogeneity as distance at which clusters are formed
 - E.g. within-cluster sum of squares when Ward method is considered
- Measures of variance change
 - Root mean square standard deviation (RMSSTD) ~ square root of the variance of the new cluster formed by joining two clusters, where the variance is computed across all clustering variables
 - Large increase in RMSSTD indicates joining of two dissimilar clusters

Visualization of solution

Dendrogram is convenient, when number of observations is not very high



Cluster Dendrogram



A!

d
hclust (*, "ward")

Use agglomeration schedule to decide number of clusters

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	99	200	.000	0	0	27
2	191	199	.000	0	0	4
3	173	196	.000	0	0	9
4	90	191	.000	0	2	153
5	31	189	.000	0	0	76

195	2	4	423.744	163	193	199
196	8	10	444.453	190	181	197
197	7	8	476.955	191	196	198
198	1	7	816.106	194	197	199
199	1	2	1467.140	198	195	0

Seek for demarcation point

Step 4: Refine solution with Nonhierarchical Clustering Procedures

- Sometimes a combination of hierarchical and nonhierarchical methods is considered:
 - Use hierarchical method (e.g., Ward) to choose number of clusters and profile cluster centers that serve as initial seeds
 - Use nonhierarchical method (e.g., k-Means) to cluster all observations using the seed points to provide more accurate cluster membership

Hierarchical vs. non-hierarchical

- Choose hierarchical method when
 - Wide range (possibly all) cluster solutions are to be examined
 - Sample size is moderate (under 300-400), no more than 1000
- Choose nonhierarchical method when
 - Number of clusters is known
 - Initial seed points can be specified by practical, objective or theoretical basis
 - Results are less susceptible to outliers, distance measure or inclusion of irrelevant variables
 - Works on large datasets

Simple k-Means algorithm

Given an initial seed, the algorithm alternates between the following steps:

1. Assignment step:

- Add each observation to the cluster, whose mean leads to the least within-group sum of squares (Squared Euclidean distance)

1. Update step:

- Compute new cluster means and use them as centroids for observations in the updated cluster

K-Means Cluster Analysis

- spending
- recommend
- age
- income
- channel
- gender
- return
- influence
- Cluster Number of Ca



Variables:

- experience
- brand
- price
- quality
- exclusivity
- selection

Iterate...

Save...

Options...

Label Cases by:

Number of Clusters:

3

Method

- Iterate and classify
- Classify only

Cluster Centers

Read initial:

Open dataset

External data file

Write final:

New dataset

Data file



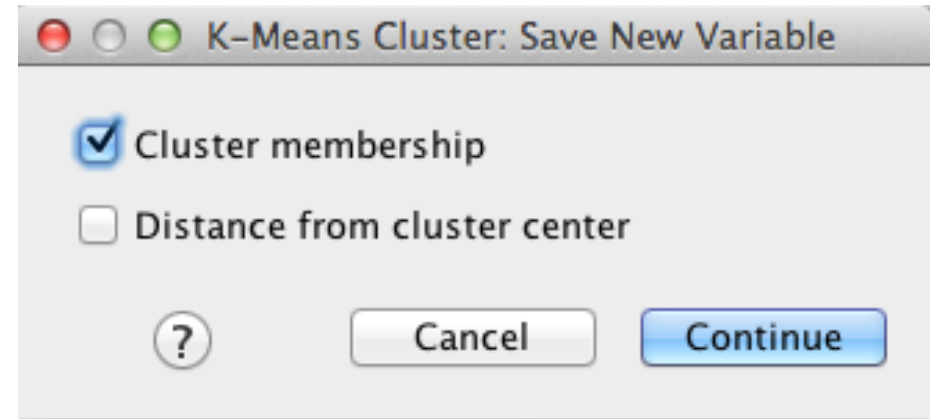
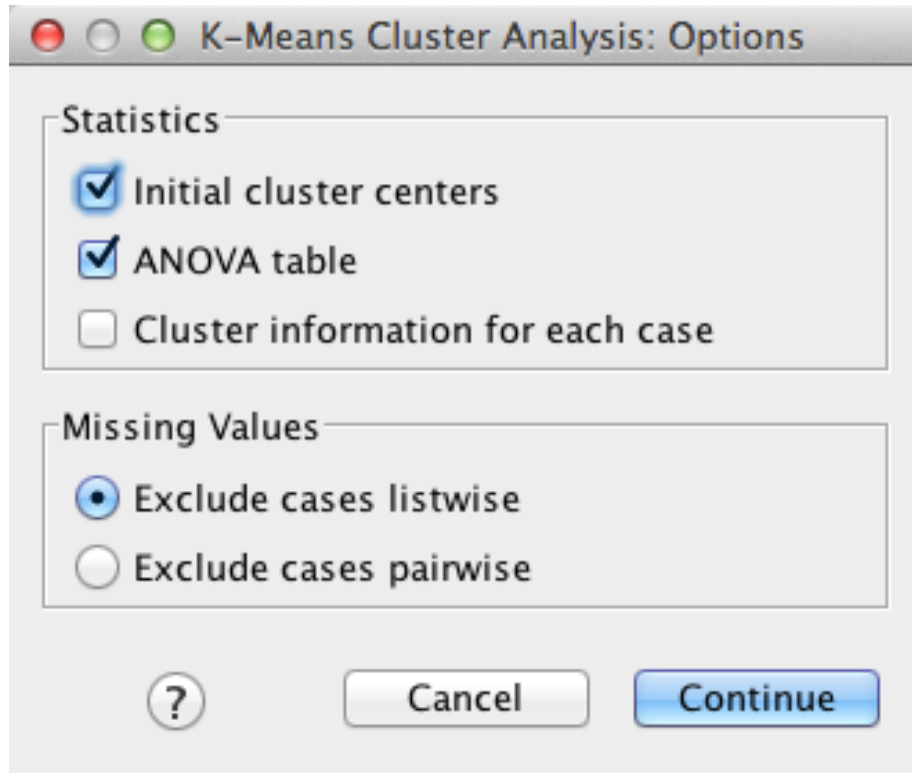
Reset

Paste

Cancel

OK

Save solution and examine output



Assumptions

- Variables should be independent of each other
- Data needs to be standardized if the range or scale of one variable is much larger or different from others
- In case of non-standardized data, Mahalanobis distance is preferred

Step 5: Evaluation of cluster solutions



How informative is your solution?

Segmentation is information compression. Good segmentation conveys key information about the important variables or attributes.

- **Generalizability:** Are the segments identifiable in a larger population?
- **Substantiality:** How sizeable are the segments when compared to each other?
- **Accessibility and actionability:** How easily can the segments be reached? Can we execute strategies using the solution?
- **Stability:** Is the solution repeatable (if new measurements are done)?

Statistical vs. practical criteria

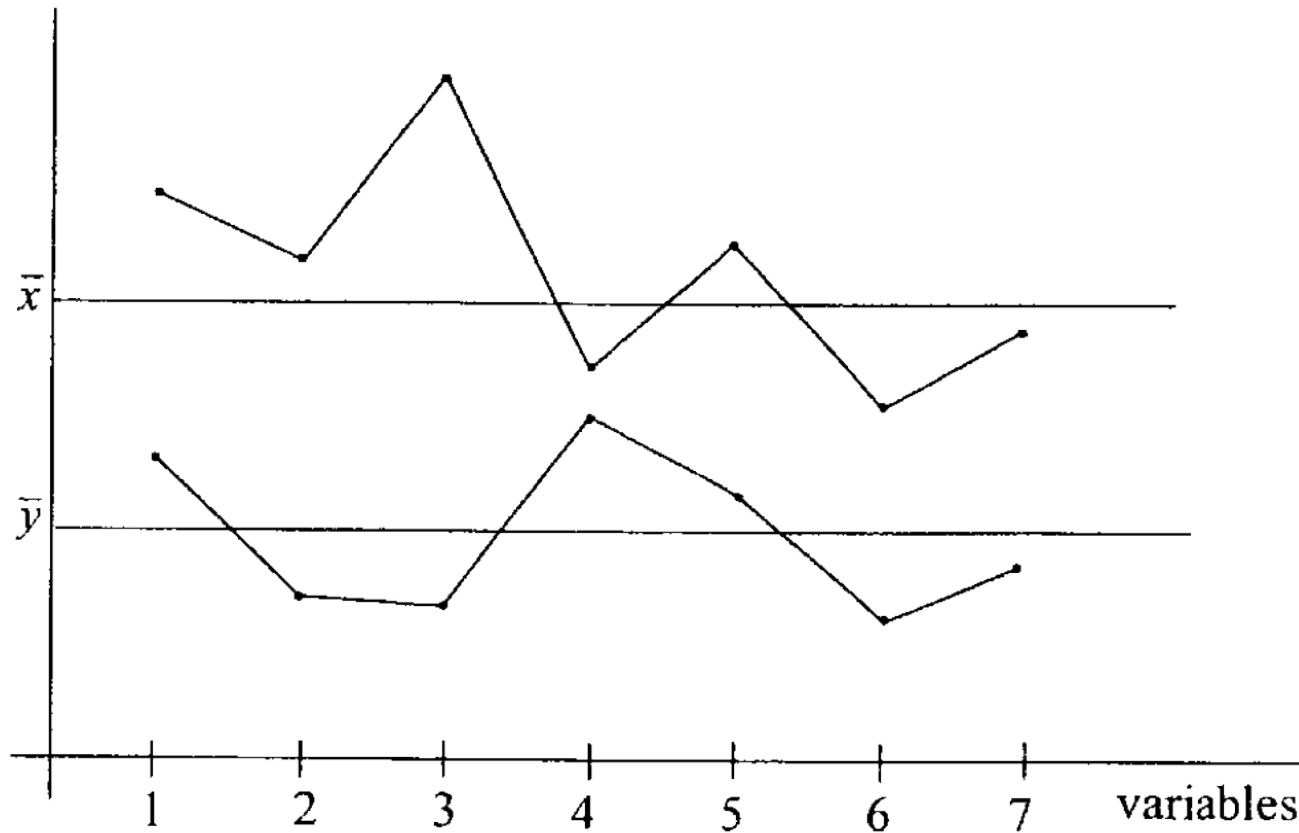
- Statistical:
 - Do the segment profiles differ in a statistically significant manner?
 - What attributes contribute most to the group differences?
 - Are the groups internally homogeneous and externally heterogeneous?
- Practical:
 - Are the segments substantial enough for making profit?
 - Is the solution stable?
 - Can we reach the segment in a cost-effective manner?
 - Is it useful for decision making purposes?
 - Do the segments respond consistently to stimulus?



Dash of criticism

- Conceptual vs. empirical support
- Descriptive, atheoretical, non-inferential?
- Clusters always produced regardless of empirical structure?
- Solution not generalizable due to dependence on variables used for defining similarity measure?

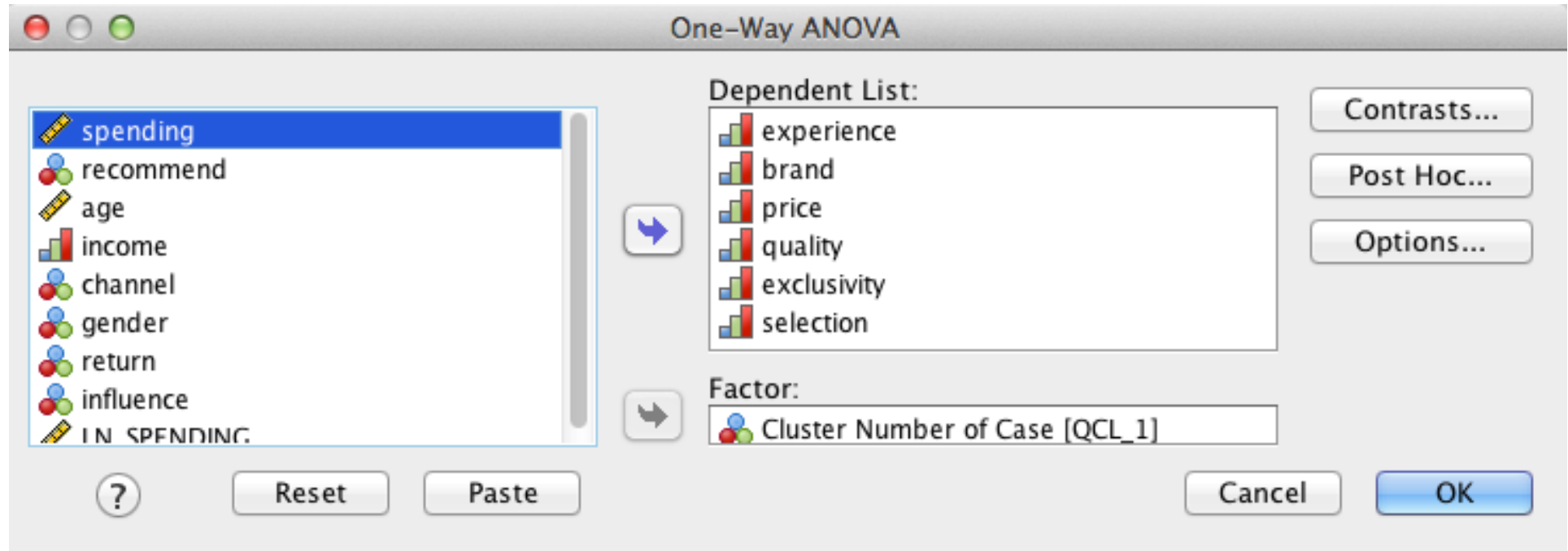
Comparison of profiles



Profiling the cluster solutions

- Once clusters are identified, objective is to describe the characteristics of each cluster and how they differ on relevant dimensions
- Utilize data not included in the cluster procedure to profile the characteristics of each cluster
 - Demographics, psychographics, consumption patterns, etc.
- Often done using Discriminant Analysis to compare average score profiles for the clusters
 - Dependent variable (categorical) = cluster membership
 - Independent variables = Demographics + Psychographics + ...

Analysis of Variance in SPSS



Analysis of Variance

A special case of multiple regression, where the objective is to compare differences between two or more groups for single metric dependent variable.

Example:

- Consumers shown different advertising messages: Which message is more likely to lead to purchase?
- A company has several customer segments: Do the segments differ in terms of customer satisfaction?

Univariate One-Way ANOVA

	Sample 1 from $N(\mu_1, \sigma^2)$	Sample 2 from $N(\mu_2, \sigma^2)$...	Sample k from $N(\mu_k, \sigma^2)$
	y_{11}	y_{21}	...	y_{k1}
	y_{12}	y_{22}	...	y_{k2}
	\vdots	\vdots		\vdots
	y_{1n}	y_{2n}	...	y_{kn}
Total	$y_{1.}$	$y_{2.}$...	$y_{k.}$
Mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$...	$\bar{y}_{k.}$
Variance	s_1^2	s_2^2	...	s_k^2

$$y_{i.} = \sum_{j=1}^n y_{ij}, \quad \bar{y}_{i.} = \sum_{j=1}^n \frac{y_{ij}}{n}.$$

Univariate One-Way ANOVA -model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$= \mu_i + \varepsilon_{ij}, \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, n;$$

Do the means between the different groups 1 to k differ?

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

H_1 : one or more of the groups has a different mean

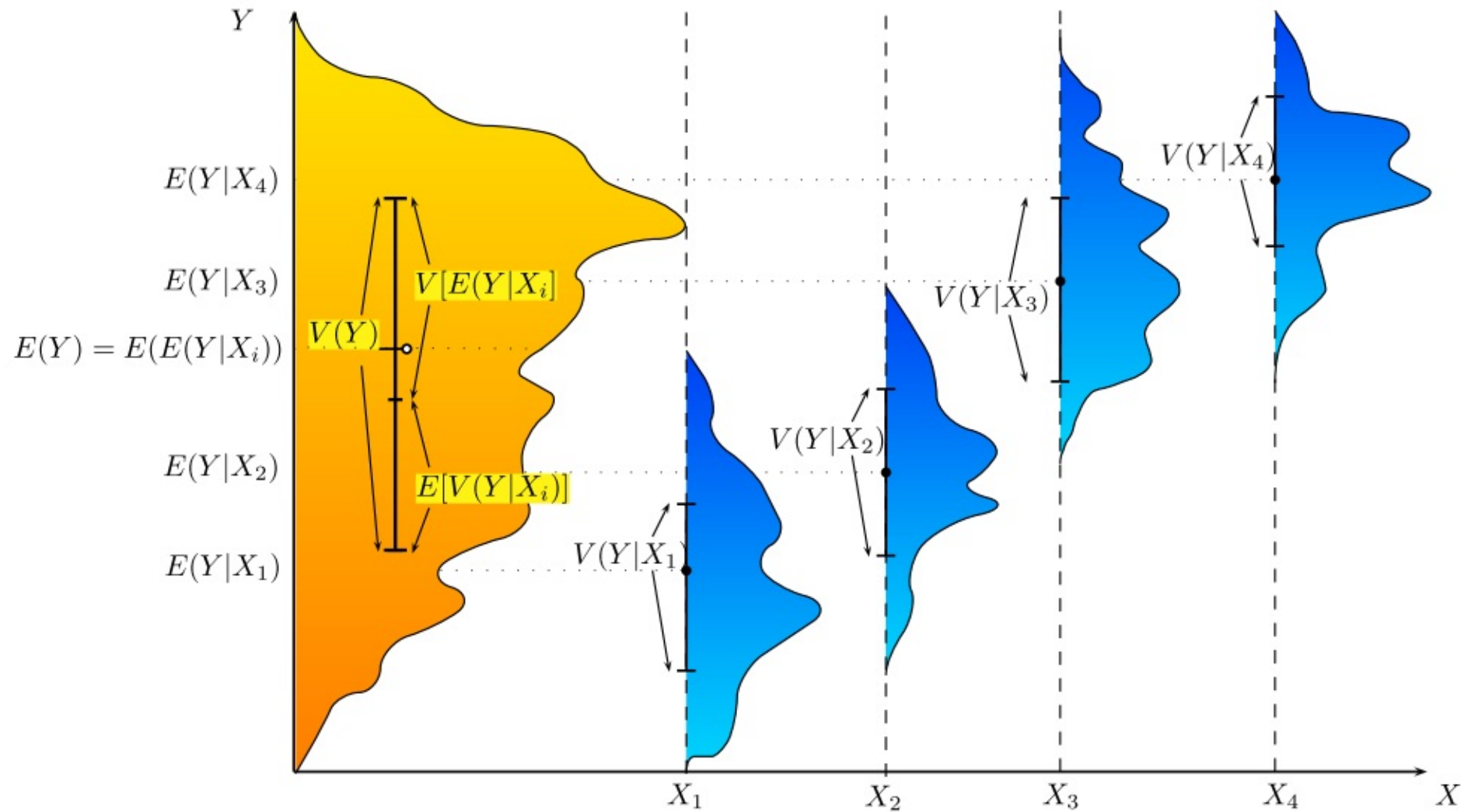


Figure 1: ANOVA : Fair fit

Source: Wikipedia

Validation

- Vary similarity measure, clustering procedure
- Cross-validation:
 - Create sub-samples of the dataset (random splitting)
 - Compare cluster solutions for consistency (number of clusters and profiles)
 - Very stable solution would be produced with less than 10 % of observations assigned differently
 - Stable solution is when 10-20% of observations are assigned to a different group
 - Somewhat stable solution when 20-25% are assigned to a different cluster
- Using relevant external variables:
 - Examine differences on variables not included in the cluster analysis but for which there is a theoretical and relevant reason to expect variation across the clusters

Review Questions

- What is the objective of cluster analysis?
- What is the difference between hierarchical and non-hierarchical clustering techniques (e.g., Ward's method vs. K-means)?
- What criteria can you use when choosing the number of clusters?
- How does an agglomerative approach work?
- Why do you use ANOVA after assigning cases to clusters?

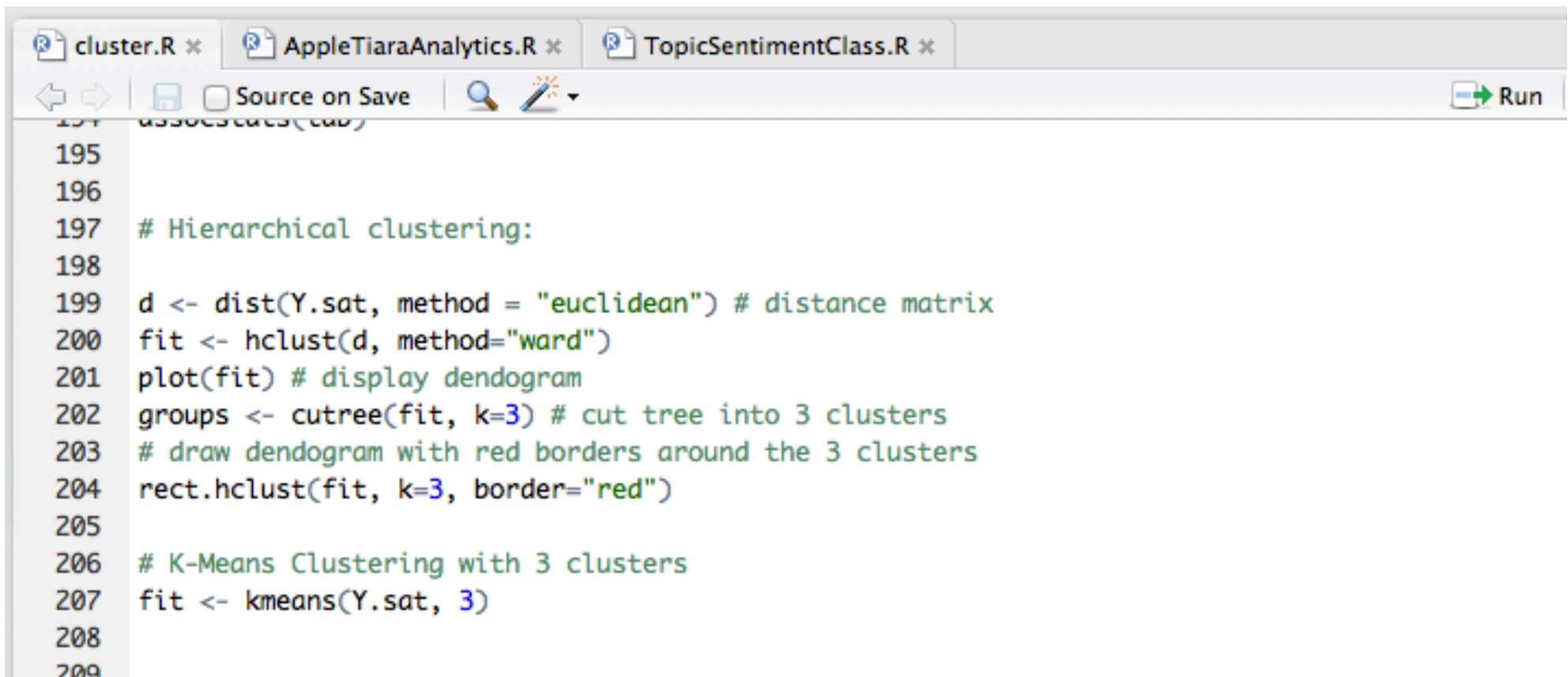


Aalto University
School of Business

Thank you!



R – give it a spin!



The image shows a screenshot of the RStudio interface. The top toolbar includes a 'Run' button with a green arrow icon. The editor window contains the following R code:

```
195  
196  
197 # Hierarchical clustering:  
198  
199 d <- dist(Y.sat, method = "euclidean") # distance matrix  
200 fit <- hclust(d, method="ward")  
201 plot(fit) # display dendrogram  
202 groups <- cutree(fit, k=3) # cut tree into 3 clusters  
203 # draw dendrogram with red borders around the 3 clusters  
204 rect.hclust(fit, k=3, border="red")  
205  
206 # K-Means Clustering with 3 clusters  
207 fit <- kmeans(Y.sat, 3)  
208  
209
```