Segmentation I

Classical cluster analysis

Morten Berg Jensen

Department of Economics and Business Economics

April 8, 2024

Outline

Outcome

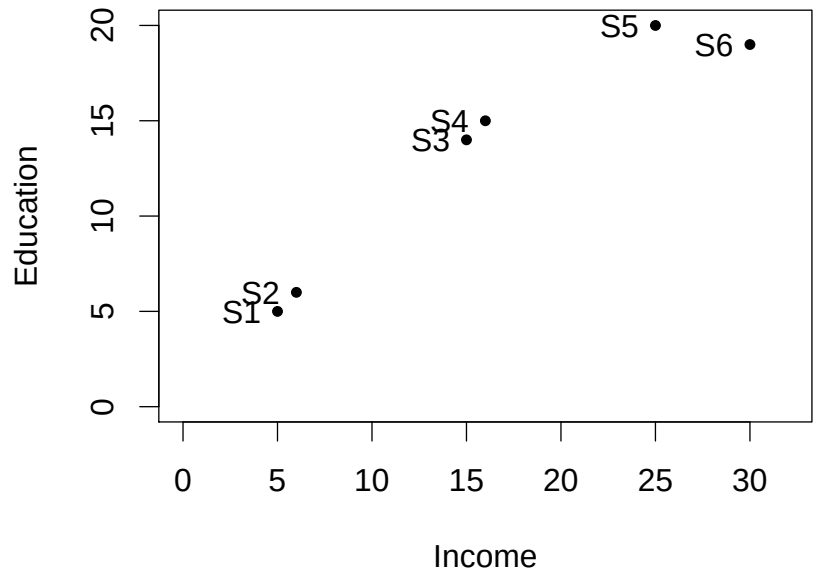
This lecture will help you to understand

- The purpose of cluster analysis
- Research design and similarity measures
- Clustering algorithms
- The number of clusters, profiling, validation, and description

Income/Education example

- Consider this hypothetical sample comprising Income and Education

Subject Id	Income ($ thous)	Education (years)
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19



- From the figure it is quite obvious how to form three groups if the purpose is to have groups with similar members within the group and dissimilar members across the groups

Income/Education example (cont'd)

- Notice also that if we had only access to data on Education we would have a hard time discriminating between groups S3/S4 and S5/S6
- This is a prime example of what multivariate analyses can achieve over repeated use of univariate analyses
- We might be able to carry out a similar analysis in three dimensions but beyond that the ability of the eye to detect patterns fails

Objectives

- Cluster analysis addresses one or a combination of the following research objectives:
  - Taxonomy description - that is an empirically based classification of objects. The objects can be customers, markets, products, firms etc.
  - Data simplification - in this situation the clustering is a means to an end. Instead of looking at an entire population management can focus on the profiles of a few groups
  - Relationship identification - using additional multivariate analyses (for instance logistic regression or ANOVA) allows the researcher to identify relationships among observations that typically isn't possible with the individual observations

Objectives (cont'd)

- The selection of variables is linked to the research objectives
- Therefore, conceptual as well as practical considerations must be taken into account
  - Practical considerations - the inclusion of undifferentiated variables should be avoided. If a given variable doesn't differ across the clusters it should be eliminated
  - Conceptual considerations - the inclusion of irrelevant variables should be avoided. All variables entering the cluster analysis must characterize the objects being clustered and relate specifically to the objectives of the analysis

Research design

- Sample size:
  - As (classical) cluster analysis isn't a statistical inference technique sample size considerations are solely related to the possibility of representing the structure and small groups in the population
  - Thus, given the objective of the analysis the researcher must make sure that the sample is large enough to represent the desired groups
- Outlier detection
  - An object which doesn't fit into a given pattern may represent: a true outlier, a small but insignificant segment of the population or an actual and relevant group which is underrepresented in the sample
  - In the first two cases deletion of the object is the appropriate action. However, in the latter case this could mean disaster
  - A profile diagram/snake plot is a graphical approach to detect outliers. However, this approach isn't suitable for large samples
  - Another approach is to calculate a measure of deviance from the average observation and then rank the objects according to this measure

Assumptions

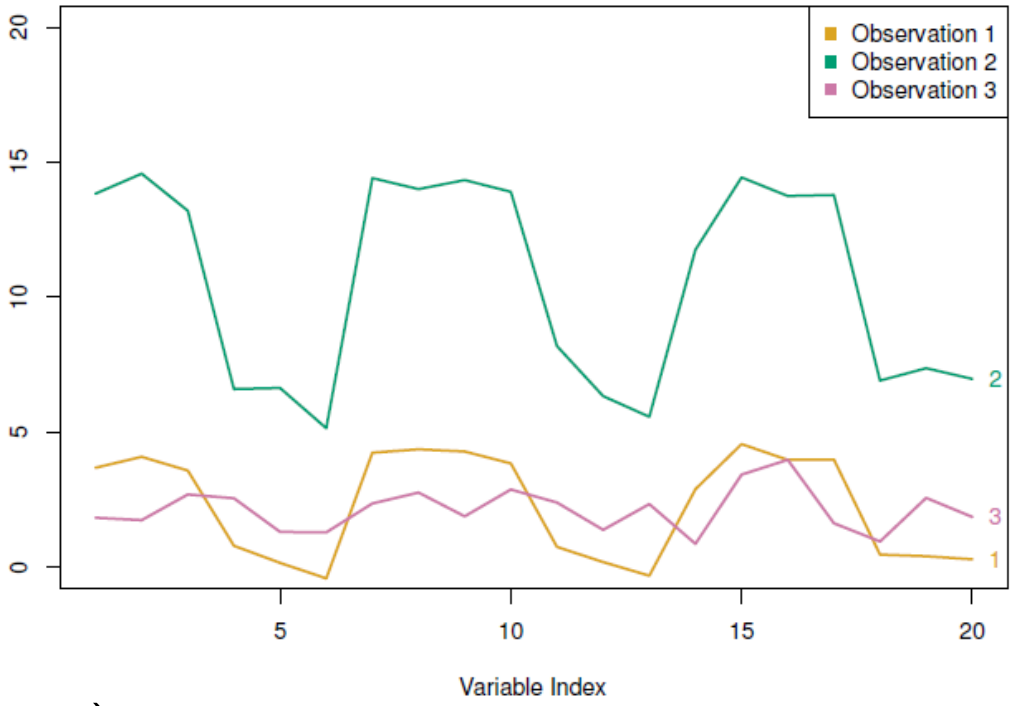
- No distributional assumptions!
- Representativeness - along with considerations governing sample size the researcher must make all efforts possible to ensure that the sample is representative
- Multicollinearity - the effect of multicollinearity is an implicit weighting
  - Thus, when two variables more or less measure the same their joint impact will be twice the impact of only one variable
  - A correlation analysis of the variables selected will reveal whether multicollinearity is an issue

Measures of similarity

- One of the key characteristics in connection with the previous example is that both variables were measured using an interval scale
- For data measured using an interval scale we can determine similarity using either **distance measures** or **correlation measures**
- Using a correlation measure of two objects both measured on several variables corresponds to a transposition of the data matrix and then calculating the usual Pearson correlation between the two objects (columns)
- However, due to the unequivocal focus on patterns/shape correlation measures are seldom used in cluster analysis
- Instead, distance measures are used thus focusing on magnitudes/levels (notice the inverse relationship between similarity and distance)

Measures of similarity (cont'd)

- The difference between distance- and correlation-based distances



(Ref: ISLR p. 529)

- Observations 1 and 3 are close to each other – small Euclidean distance
- Observations 1 and 2 behave similarly – high correlation(small distance = 1 - correlation)

Measures of similarity (cont'd)

- Standardization
  - In the case of using distance measures the researcher must be aware of the influence of the scales or magnitudes of the chosen variables
  - Variables with larger dispersion have more impact on the distance measure
  - The solution to this is to standardize the variables
  - Assessing the standard deviations from a descriptive analysis of the selected variables will tell you whether standardization is required

Measures of similarity (cont'd)

- For the purpose of measuring the distance between two of our objects (say i and j) several measures are available
- Euclidean distance: $\sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2}$
- Squared Euclidean distance: $(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2$
- City-block/Manhattan distance: $|X_{i1} - X_{j1}| + |X_{i2} - X_{j2}|$
- For the general case where we want to measure the distance between object i and j with each object consisting of p variables we have these generalizations:
- Euclidean distance: $\sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2}$
- Squared Euclidean distance: $\sum_{k=1}^{p}(X_{ik} - X_{jk})^2$
  This measure is the recommended measure for the centroid and Ward's method (see later)
- City-block/Manhattan distance: $\sum_{k=1}^{p}|X_{ik} - X_{jk}|$

Income/Education example – distance measure

- We calculate the similarity matrix using the squared Euclidean distance for our six objects:

	S1	S2	S3	S4	S5	S6
S1	0.00	2.00	181.00	221.00	625.00	821.00
S2	2.00	0.00	145.00	181.00	557.00	745.00
S3	181.00	145.00	0.00	2.00	136.00	250.00
S4	221.00	181.00	2.00	0.00	106.00	212.00
S5	625.00	557.00	136.00	106.00	0.00	26.00
S6	821.00	745.00	250.00	212.00	26.00	0.00

Measures of similarity (cont'd)

- For binomial data association or matching measures are used - for example count or percentage of times agreement occurs
- For binary variables $x_1, \ldots, x_p$ we can calculate the number of occurrences of (1,1), (1,0), (0,1), and (0,0) for two observations
- These numbers are generically referred to as *a*, *b*, *c* and *d*
- One measure of similarity is then given as

$$r = \frac{a + d}{p}$$

with the Euclidean distance given as

$$\delta = \sqrt{p(1 - r)}$$

- Be aware of limitations of this 0/1 coding – in some situations only (1,1) will be considered a match

Measures of similarity (cont'd)

- For nominal data with more than two categories dummy variables might be constructed and used subsequently
- In such a situation Jaccard's coefficient is often used as a similarity measure

$$\frac{a}{a+b+c}$$

- Also purely **subjective** assessments might be used for constructing a distance/similarity matrix (but this is rarely done)

Hierarchical methods

- The hierarchical methods comprise the agglomerative and divisive methods
- In the agglomerative methods all objects start out as their own cluster and objects are the successively joined
- In the divisive methods all objects start out as a single cluster and objects are then successively divided - these methods are seldom used
- There is a number of different hierarchical agglomerative methods available
- They all seek to join clusters which are as close to each other as possible
- They differ in the way distance is measured between clusters (one of the clusters can be a single object)

Hierarchical methods (cont'd)

- Single-linkage/nearest-neighbor measures the distance between clusters as the minimum of the distance between all possible pairs of objects in the two clusters – tends to produce unbalanced and "straggly" clusters
- Complete-linkage/farthest-neighbor measures the distance between clusters as the maximum of the distance between all possible pairs of objects in the two clusters – tends to find compact clusters with equal diameters
- These methods only depend on the ordinal properties of the distances

Hierarchical methods (cont'd)

- Average-linkage measures the distance between clusters as the average distance between all possible pairs of objects in the two clusters
- Centroid measures the distance between clusters by forming average objects for each cluster and then measuring the distance between these average objects
- Ward's method forms clusters by minimizing the within-cluster sums of squares – tends to find same-size clusters
- These methods depend on metrical properties of the data matrix and hence the distances

Hierarchical methods (cont'd)

- It is recommended that the robustness of the cluster solution is assessed by calculating different cluster solutions for different combinations of distance measure and method
- Complete-linkage and Wards methods are generally preferred
- Regarding the assumptions – "Its success or otherwise is to be judged by whether or not it produces 'meaningful' clusters rather than the means used for their construction" – Bartholomew, Steele, Moustaki, and Galbraith, 2008

Income/Education example – single-linkage

- For **single-linkage** the original similarity matrix is the point of departure. The distance between S1 and S2 is the smallest for which reason S1 and S2 are joined in the first step. We can now calculate the resulting similarity matrix:

	S1&S2	S3	S4	S5	S6
S1&S2	0.00	145.00	181.00	557.00	745.00
S3	145.00	0.00	2.00	136.00	250.00
S4	181.00	2.00	0.00	106.00	212.00
S5	557.00	136.00	106.00	0.00	26.00
S6	745.00	250.00	212.00	26.00	0.00

- In the second step we can see that S3 and S4 should be joined. After this a new similarity matrix must be calculated

Income/Education example – single-linkage (cont'd)

- Here is the updated similarity matrix:

	S1&S2	S3&S4	S5	S6
S1&S2	0.00	145.00	557.00	745.00
S3&S4	145.00	0.00	106.00	212.00
S5	557.00	106.00	0.00	26.00
S6	745.00	212.00	26.00	0.00

- In the third step we can see that S5 and S6 should be joined. After this a new similarity matrix must be calculated

Income/Education example – single-linkage (cont'd)

- Here is the updated similarity matrix:

	S1&S2	S3&S4	S5&S6
S1&S2	0.00	145.00	557.00
S3&S4	145.00	0.00	106.00
S5&S6	557.00	106.00	0.00

- We join S3&S4 and the S5&S6 and update the similarity matrix:

	S1&S2	S3&S4&S5&S6
S1&S2	0.00	145.00
S3&S4&S5&S6	145.00	0.00

- Last step is then to join the two remaining groups

Income/Education example – Ward's

- **Ward's method** calculates the within-cluster sums of squares for all possible five-cluster solutions:

	Members in Cluster					
Solution	1	2	3	4	5	SS
1	S1,S2	S3	S4	S5	S6	1.0
2	S1,S3	S2	S4	S5	S6	90.5
3	S1,S4	S2	S3	S5	S6	110.5
4	S1,S5	S2	S3	S4	S6	312.5
5	S1,S6	S2	S3	S4	S5	410.5
6	S2,S3	S1	S4	S5	S6	72.5
7	S2,S4	S1	S3	S5	S6	90.5
8	S2,S5	S1	S3	S4	S6	278.5
9	S2,S6	S1	S3	S4	S5	372.5
10	S3,S4	S1	S2	S5	S6	1.0
11	S3,S5	S1	S2	S4	S6	68.0
12	S3,S6	S1	S2	S4	S5	125.0
13	S4,S5	S1	S2	S3	S6	53.0
14	S4,S6	S1	S2	S3	S5	106.0
15	S5,S6	S1	S2	S3	S4	13.0

- The solution with the lowest sums of squares is chosen. Ties are broken randomly. Hence we choose solution 1

Nonhierarchical methods

- The general purpose of nonhierarchical clustering is to classify objects into k clusters such that
  1. The objects within the same cluster are as similar as possible – **high intra-class similarity**
  2. The objects from different clusters are as dissimilar as possible – **low inter-class similarity**
- There are two key differences between nonhierarchical and hierarchical clustering
  1. The number of clusters, k, must be known a priori
  2. Once an object has been assigned a particular cluster it can later in the process be assigned a different cluster - thus the treelike construction process doesn't apply

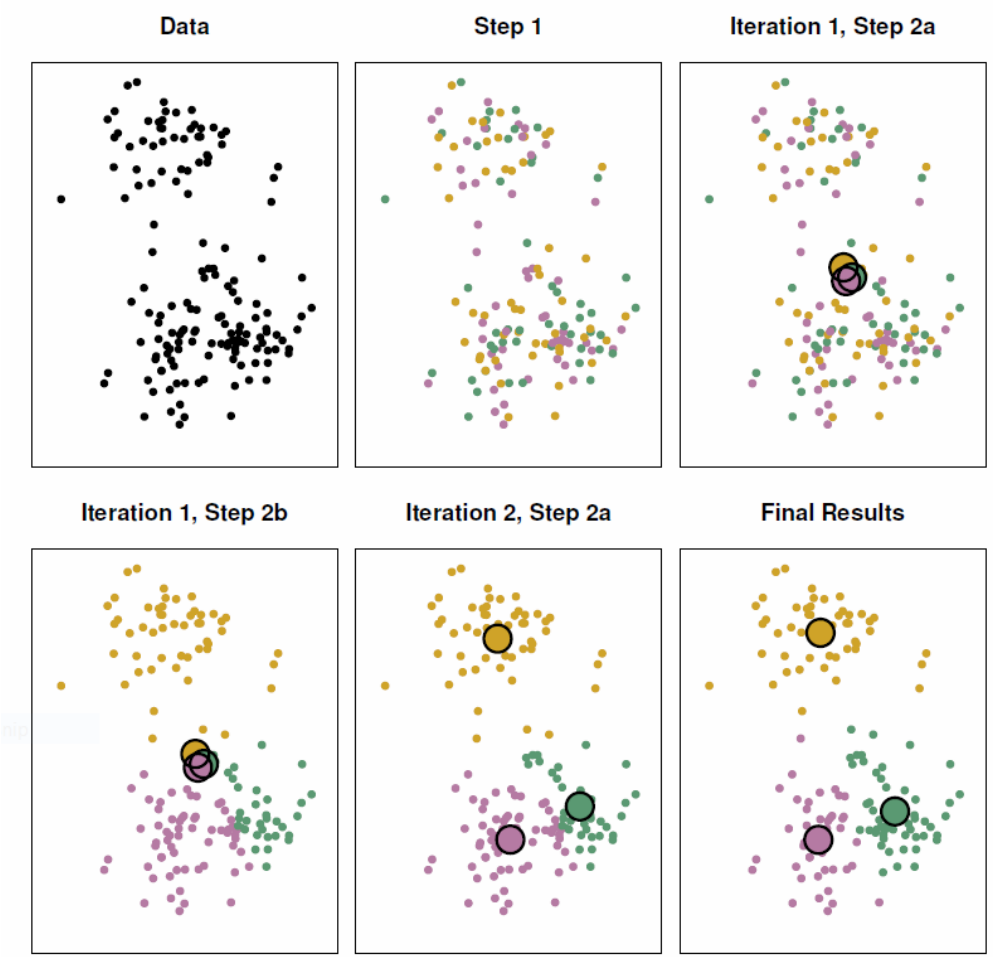
Nonhierarchical methods (cont'd)

- For a k-cluster analysis the nonhierarchical method comprises the following steps
  1. Select k initial cluster seeds (centroids)
  2. Assign each of the observations to the nearest cluster
  3. Reassign each object to one of the k clusters according to a stopping rule
  4. Stop if there is no reassignment

  This description also covers situations where no reassignment takes place

The dynamics of the K-means algorithm

- Progress of the K-means algorithm for $K = 3$ with random allocation to the clusters in step 1



(Ref: ISLR p. 520)

Selecting the cluster seeds

- There are a number of possible ways to select the initial cluster centroids (cluster seeds)
  - The researcher supplies the seeds - this may be relevant if prior research has already defined segment profiles and the purpose of clustering is solely allocation
  - Select the first k objects with nonmissing data
  - Select the first seed as the first object with nonmissing data. The second seed is the next observation with nonmissing data that is separated from the first seed by a specified distance ...
  - Randomly select k nonmissing objects
- A key issue in this respect is the fact that the order of the observations may have implications for the solution found
- This latter issue can typically be handled using a number of restarts

Nonhierarchical algorithms

- There are also a number of different algorithms for reassigning the objects
  - Compute the centroids of each cluster and reassign objects to the closest cluster centroid. The centroids are only updated after every object has been reassigned
  - Compute the centroids of each cluster and reassign objects to the closest cluster centroid. The centroids (receiving and ceding clusters only) are updated after each reassignment
- This group of algorithms is known as K-means algorithms

Comparing hierarchical and nonhierarchical methods

- It is recommended that a combination of the hierarchical and nonhierarchical methods is used
  - Use the hierarchical methods to get a qualified estimate on the number of clusters
  - Use the nonhierarchical methods to determine the cluster affiliation for each object

Deciding on the number of clusters

- The decision on the number of clusters is a key decision in connection with the hierarchical methods and it is a part of the input in the non-hierarchical procedures
- The researcher is advised to use the hierarchical procedures to come up with a limited number of possible cluster sizes and then utilize these cluster sizes in the nonhierarchical analysis
- No standard objective selection procedure exists which is why the researcher is advised to consider several possible solutions
- A number of ad hoc rules have been developed, often these rules are tied to a particular software package
- The measures can either be categorized as measures of heterogeneity change or direct measures of heterogeneity

NbClust

- This R package calculates a vast number of measures informing on the number of clusters
- Still - no unanimous answer to the question of how many clusters
- However, you will get some quantitative input to address the problem

Dendrogram

- A key tool for determining the number of clusters in association with hierarchical clustering is the dendrogram
- It is a tree-like figure, where one dimension holds the distances between observations/clusters and the other dimension identifies the observations
- A tendency for a group of branches to come together at the same point and then not be involved in further amalgamations for a while indicates a cluster
- We will look at "big jumps in the dendrogram" and the percentage increase in agglomeration coefficients to inform on the number of clusters

Profiling

- The profiling is carried out by looking at the centroids of the cluster
- The idea is to look for characteristics on one or several of the clustering variables that identify a cluster
- In this way one can name each of the clusters based on the information just identified
- The profiling may also be helpful in choosing from different potential cluster solutions

Validation and description

- The purpose of the validation is to ensure that results are representative of the population and thus generalizable
  - One possible technique is to divide the sample into subsamples and assess the relationship between the solutions
  - Another possibility is to use a set of different variables known to vary across clusters and then test for differences – criterion validity
- Description
  - This stage involves the use variables not previously used in the analysis
  - The idea is to describe characteristics of the clusters, typically in form of demographics, psychographic profiles, consumption etc.

Background

- HBAT is a manufacturer of paper products who sells products to two market segments: the newsprint industry and the magazine industry. The data used in this example is based on a survey of HBAT customers who completed a questionnaire on a website. 100 customers (purchasing managers from firms) buying from HBAT completed the questionnaire.
  - A first type of information is available from HBAT's data warehouse and includes information, such as, size of the customer and length of purchase relationship
  - The second type of information was consumers perceptions of HBAT's performance on 13 attributes using a 0-10 scale with 10 being "Excellent" and 0 being "Poor"
  - The third type of information relates to purchase outcomes and business relationships (e.g. satisfaction with HBAT, and whether the firm would consider a strategic alliance/partnership with HBAT)

Data base description

Variable	Description	Measurement scale
Data Warehouse Classification Variables		
X_1	Customer Type	nonmetric
X_2	Industry Type	nonmetric
X_3	Firm Size	nonmetric
X_4	Region	nonmetric
X_5	Distribution System	nonmetric
Performance Perceptions Variables		
X_6	Product Quality	metric
X_7	E-Commerce Activities/Website	metric
X_8	Technical Support	metric
X_9	Complaint Resolution	metric
X_{10}	Advertising	metric
X_{11}	Product Line	metric
X_{12}	Salesforce Image	metric
X_{13}	Competitive Pricing	metric
X_{14}	Warranty & Claims	metric
X_{15}	New Products	metric
X_{16}	Ordering & Billing	metric
X_{17}	Price Flexibility	metric
X_{18}	Delivery Speed	metric
Outcome/Relationship Measures		
X_{19}	Satisfaction	metric
X_{20}	Likelihood of Recommendation	metric
X_{21}	Likelihood of Future Purchase	metric
X_{22}	Current Purchase/Usage Level	metric
X_{23}	Consider Strategic Alliance/Partnership in Future	nonmetric

Objective

- The **objective** of the analysis is to develop a customer segmentation (taxonomy development) based on the customers perceptions of HBAT's performance on the variables X_6, X_8, X_{12}, X_{15} and X_{18})
- If HBAT succeed then the basis for a market segmentation is provided

Variables and sample

- Cluster analysis is an interdependence technique, hence we do not distinguish between dependent and independent variables
- The variables are:

X_6	Product Quality	metric
X_8	Technical Support	metric
X_{12}	Salesforce Image	metric
X_{15}	New Products	metric
X_{18}	Delivery Speed	metric

Analysis

- We start out by taking a look at the data and try to identify outliers
- We can sort the data by "dev" to identify possible outliers
- We decide to delete cases 6 and 87
- We do the hierarchical clustering using squared Euclidean distance and Ward's method
- Based on the appearance of the dendrogram, the percentage increase and the interpretability we decide on a 4 cluster solution

Profiling

- The profiling is most easily done using a set of ANOVA analyses and some descriptives
  - There are significant differences between the averages of the four clusters for each of the five variables X_6, X_8, X_{12}, X_{15}, and X_{18}
  - All clusters contain more than 10 % of the observations
  - Cluster 1 is characterized by the low average of X_{15}
  - Cluster 2 is characterized by the low average of X_8
  - Cluster 3 is characterized by the low average of X_6 and the high average of X_{12}
  - Cluster 4 is characterized by the low average of X_{12} and X_{18}
- A "snake plot" can also be useful – we illustrate this for the nonhierarchical solution

Analysis (cont'd)

- We proceed with the nonhierarchical analysis where the researcher has to decide on the number of clusters in advance
- We assess the criterion validity of the solution by computing separate ANOVA's for the variables X_{19}, X_{20}, X_{21}, and X_{22}
- Finally, we will profile our solution. Hence, we do cross-table analysis for the variables X_1, X_2, X_3, X_4, and X_5 against cluster membership

Profiling (cont'd)

- This profiling is also done most easily using a set of ANOVA analyses and some descriptives
  - There is a more even distribution of observations across clusters compared to the hierarchical solution
  - Apart from X_{18} there are significant differences between the averages of the four clusters

Profiling (cont'd)

- Cluster 1 has 17 observations and is distinguished by a relatively high mean for new products (X_{15}) and also to a smaller extent for product quality (X_6). Thus, this segment suggests that HBAT offers new and innovative products and of good quality. Despite scoring HBAT high on new products there is still room for improvement (mean 6.75)
- Cluster 2 has 22 observations and is distinguished by a rather high mean for product quality (X_6) and technical support (X_8). The means of the other cluster variables are relatively average
- Cluster 3 has 28 observations and is distinguished by the lowest mean for product quality (X_6). Moreover, the means of the other cluster variables are relatively average except for salesforce image (X_{12})
- Cluster 4 has 31 observations and is distinguished by a relatively low mean for new products (X_{15}). Only for product quality (X_6) is the mean relatively high. This is the largest cluster

Validation and description

- Clearly, the means of X_{19}, (X_{20}), X_{21}, and X_{22} differ across the different clusters supporting criterion validity – satisfaction (X_{19}), likelihood to recommend (X_{20}), likelihood to purchase (X_{21}), and purchase level (X_{22}) all have the largest average for cluster 1,2 and 4
- There seems to be a significant relationship between customer and industry type (X_1 and X_2) and cluster category as well as between region (X_4) and cluster category