Multivariate Data Analysis

(MGT513, BAT531, TIM711)

Lecture 6

Cluster Analysis

What is Cluster Analysis?

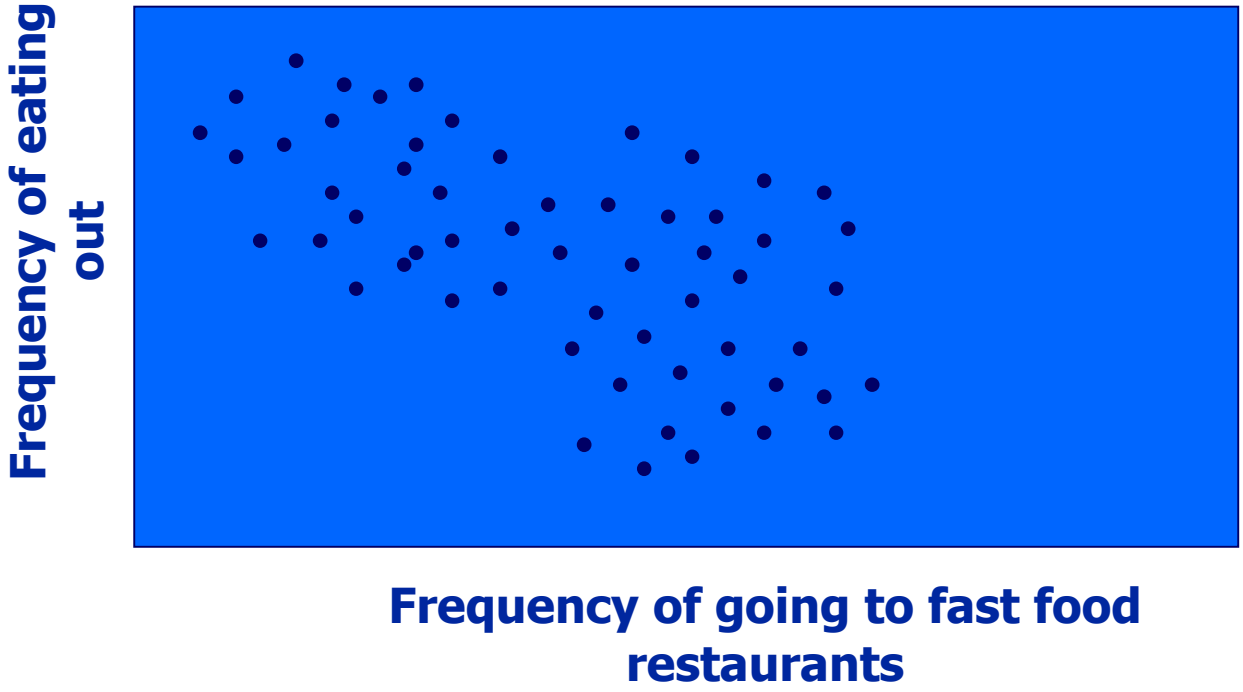
- Cluster analysis is a group of multivariate techniques whose primary purpose is to group objects based on the characteristics they possess
- Cluster analysis involves categorization: dividing a large group of observations into smaller groups so that the observations within each group are relatively similar (i.e., they possess largely the same characteristics) and the observations in different groups are relatively dissimilar - *Closely related to MDS*

What is Cluster Analysis?

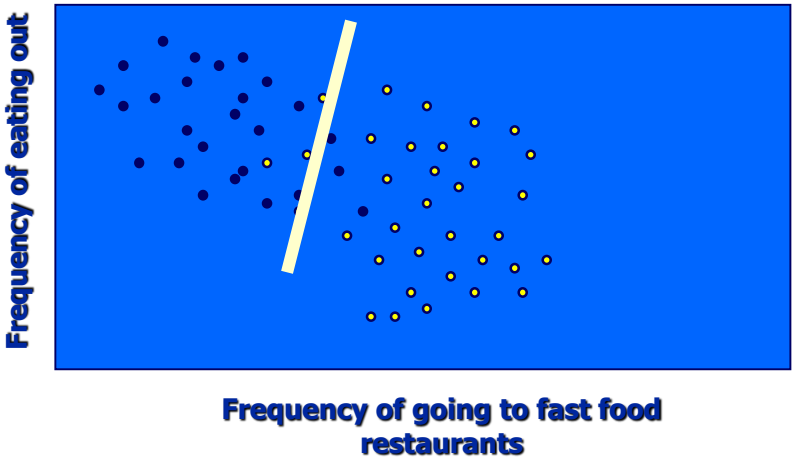
- Main Objectives:
  - To partition a set of objects into two or more groups based on the similarity of the objects for a set of specified characteristics (the cluster variate)
  - To addressing the heterogeneity in the data with hopefully, a small (manageable) number of groups
  - To determine whether the data contain naturally occurring, homogeneous subsets of observations

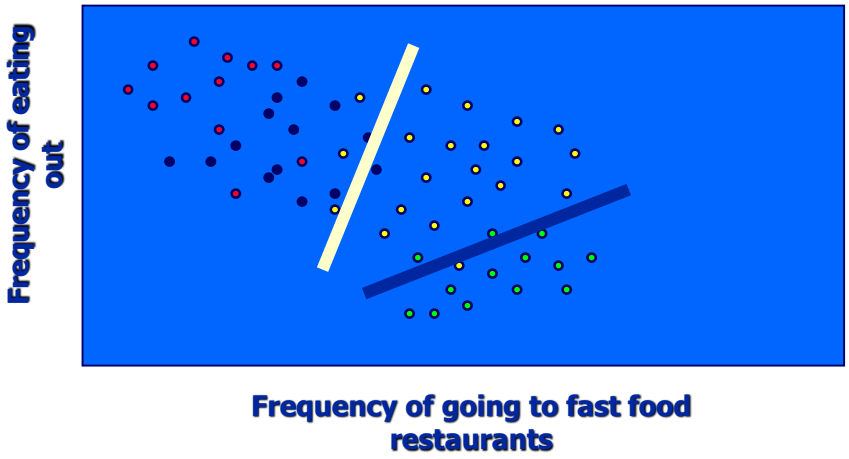
Scatter Diagram for Cluster Observations

Fundamental Question: How Many Clusters?

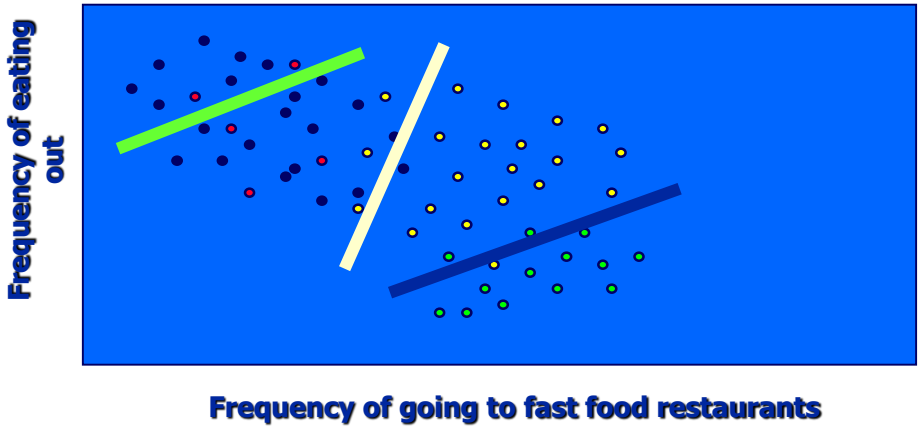


Potential Two, Three and Four Cluster Solutions

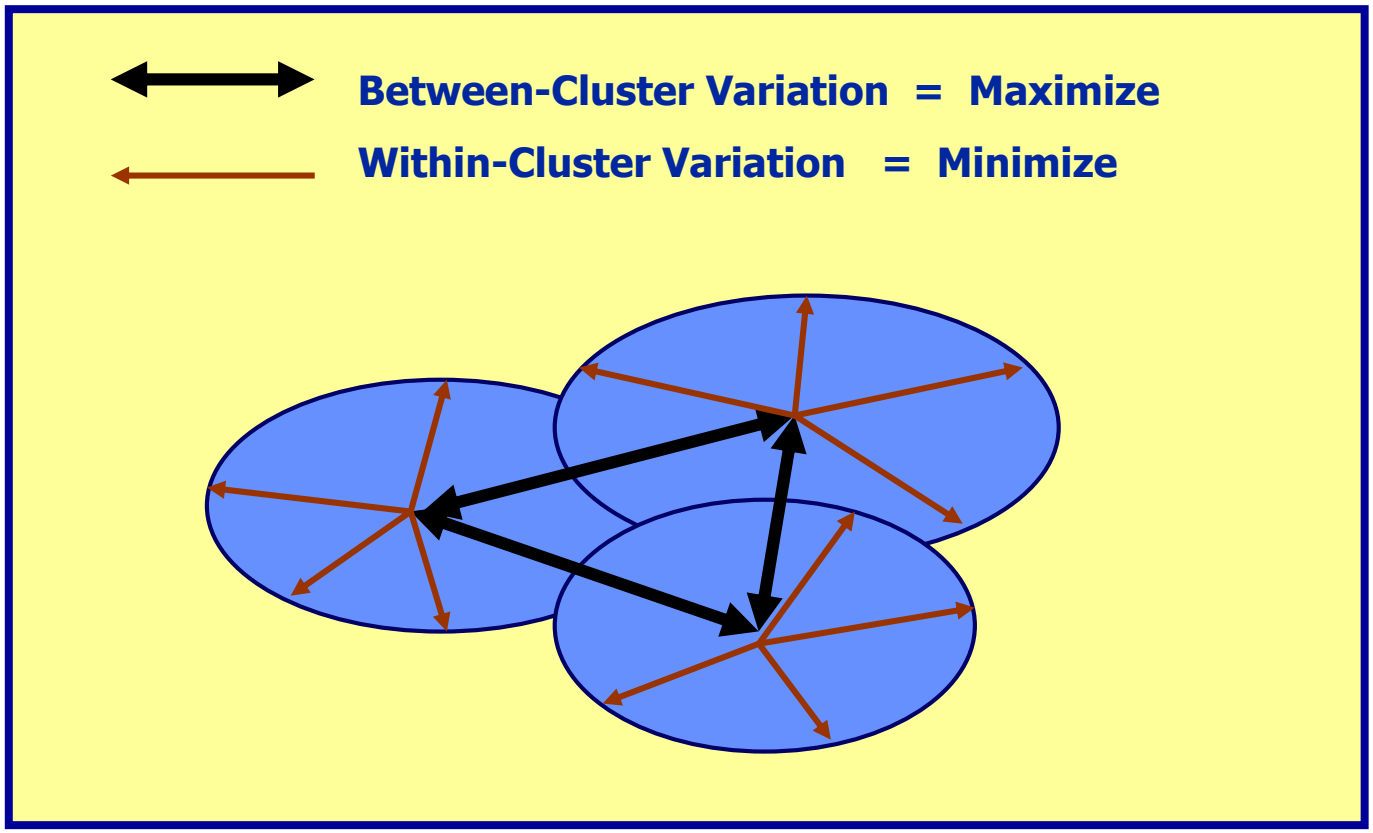




Which one is correct?



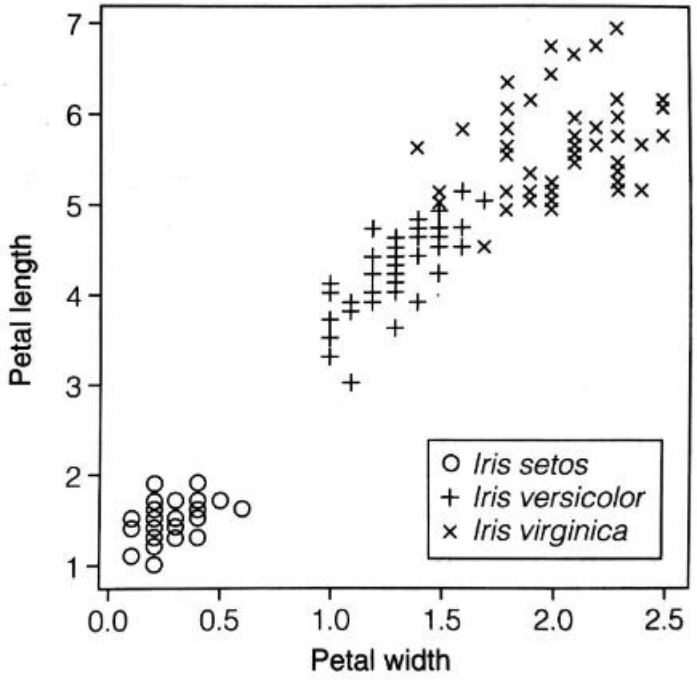
Principal of Cluster Analysis



Potential Applications

Numerical Taxonomy: Three different types of *Iris*: Iris setosa, Iris versicolor, Iris Virginica

FIGURE 8.1
Plot of Fisher's iris data (petal width versus petal length)



Potential Applications

Market Segmentation

Aim: Finding relative preferences [*K*-means method]

TABLE 8.1 Preferences expressed by 32 student subjects for 10 different brands of beer (measured on a 9-point scale)

	Anchor Steam	Bass Ale	Beck's	Corona	Gordon-Biersch	Guinness	Heineken	Pete's Wicked Ale	Sam Adams	Sierra Nevada
S001	5	9	7	1	7	6	6	5	9	5
S008	7	5	6	8	8	4	8	8	7	7
S015	7	7	5	6	6	1	8	4	7	5
S022	7	7	5	2	5	8	4	6	8	9
S029	9	7	3	1	6	8	2	7	6	8
S036	7	6	4	3	7	6	6	5	4	9
S043	5	5	5	6	6	4	7	5	5	6
S050	5	3	1	5	5	5	3	5	5	9
S057	9	3	2	6	4	6	1	5	3	6
S064	2	6	6	5	6	4	8	4	4	3
S071	7	7	7	5	7	8	6	7	7	8
S078	8	3	3	9	9	2	1	9	7	8
S085	6	5	3	7	6	5	8	6	7	5
S092	5	6	3	8	6	7	6	7	6	7
S099	4	7	2	8	5	9	8	3	8	8
S106	3	3	4	5	6	5	9	7	5	5
S113	2	4	5	7	6	6	8	1	7	4
S120	9	3	7	4	2	4	6	3	8	6
S127	5	3	4	7	7	7	6	6	6	6
S134	2	4	4	8	5	5	5	4	6	6
S141	5	7	6	7	5	8	8	7	5	7
S148	8	9	6	7	7	8	6	8	8	8
S162	5	6	6	7	5	3	7	3	4	3
S169	5	5	6	7	7	4	6	3	7	6
S176	5	5	7	8	7	6	7	5	4	7
S183	3	5	4	7	3	1	2	6	6	5
S190	4	3	6	8	6	1	8	2	7	7
S197	3	8	4	8	6	2	8	4	6	1
S204	3	5	1	5	5	3	4	6	7	5
S211	3	8	5	8	7	5	5	3	7	8
S218	8	8	5	7	9	9	7	7	6	8
S225	7	6	2	2	6	6	2	7	5	5

Potential Applications

Market Segmentation

TABLE 8.2 Cluster profiles from *K*-means cluster analysis of beer preference data (for $K = 2$)

	Cluster Profiles	
	Cluster 1 ($n = 20$)	Cluster 2 ($n = 12$)
Anchor Steam	4.2	7.4
Bass Ale	5.3	5.9
Beck's	4.6	4.3
Corona	7.0	4.3
Gordon-Biersch	5.9	6.2
Guinness	4.5	6.3
Heineken	6.8	4.2
Pete's Wicked Ale	4.7	6.2
Sam Adams	6.1	6.3
Sierra Nevada	5.5	7.4

Potential Applications

Market Segmentation

- On the Left: lighter beers and lager
- On the Right: darker and ale

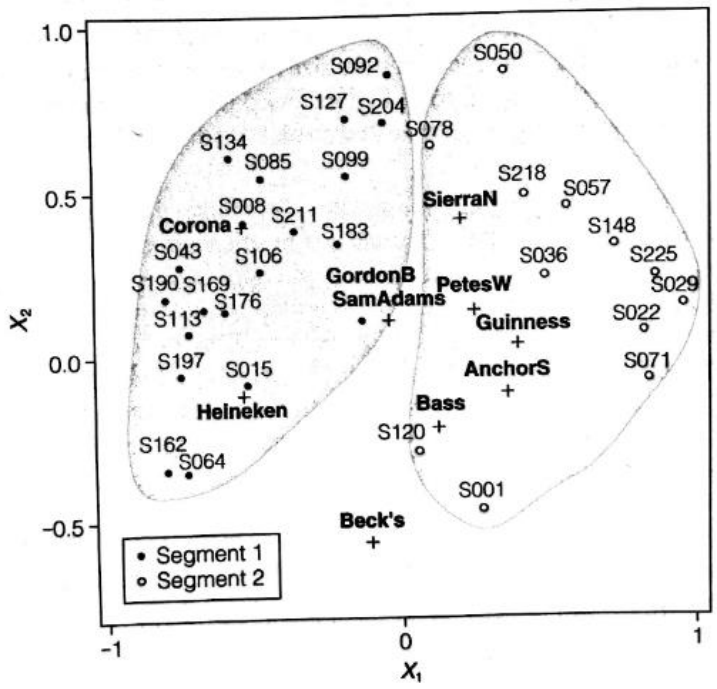


FIGURE 8.2
MDPREF map of student beer preferences showing two-cluster segmentation

Potential Applications

Market Structure Analysis

TABLE 8.3 Store-switching matrix shows patterns of switching (from row store to column store) for 300 households across 13 stores in Sioux Falls, ND

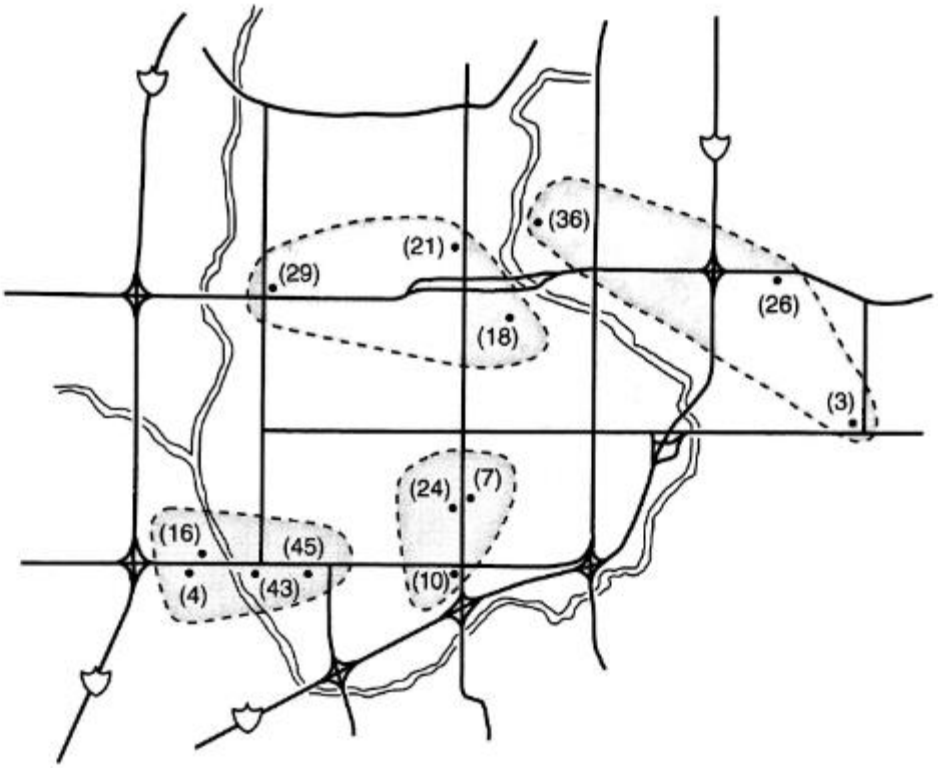
	S03	S04	S07	S16	S18	S21	S24	S26	S29	S36	S10	S43	S45
S03	1428	52	199	5	55	12	19	422	14	167	37	32	14
S04	15	1001	256	227	20	26	33	33	162	23	76	185	95
S07	84	96	2553	124	292	214	655	165	227	122	329	206	130
S16	1	80	29	256	23	27	37	21	82	2	67	206	44
S18	34	11	164	11	910	112	105	98	50	42	102	50	20
S21	13	19	123	10	63	1294	70	59	187	44	60	88	61
S24	17	47	364	7	39	30	803	55	119	1	248	236	126
S26	506	29	125	10	47	37	15	1983	67	132	121	175	30
S29	22	162	219	44	48	87	43	49	1590	35	73	268	139
S36	78	3	29	14	93	100	12	220	48	868	29	20	15
S10	73	82	330	17	52	34	76	52	13	34	441	370	117
S43	113	299	509	102	42	49	123	99	113	36	70	1040	450
S45	62	243	358	54	24	74	97	28	116	15	47	100	455

Source: Bucklin and Lattin, 1992.

Potential Applications

Market Structure Analysis

FIGURE 8.3
Clustering of stores based on store-switching behavior of consumers (Source: Lattin and Bucklin, 1992). Reprinted by permission of the authors.



Three Basic Questions in a Cluster Analysis

1. How do we measure similarity?
   - We require a method of simultaneously comparing observations on the clustering variables. Several methods are possible, including the correlation between objects or perhaps a measure of their proximity in two-dimensional space such that the distance between observations indicates similarity.
2. How do we form clusters?
   - No matter how similarity is measured, the procedure must group those observations that are most similar into a cluster, thereby determining the cluster group membership of each observation for each set of clusters formed.
3. How many groups do we form?
   - The final task is to select one cluster solution (i.e., set of clusters) as the final solution. In doing so, the researcher faces a trade-off: fewer clusters and less homogeneity within clusters versus a larger number of clusters and more within-group homogeneity.

Measuring Similarity: Distance Measures

- Minkowski distance
  - L_p norm distance
  - The p th root of the sum of the p th powers of the differences of the components
  - $d_{ij}(p) = (\sum_{k=1}^{n} |x_{ik} - x_{jk}|^{\mathrm{p}})^{1/p}$
  - Manhattan distance (p=1), Euclidean distance (p=2), Maximum distance (p=∞)
- Manhattan distance (city-block distance)
  - Sum of the absolute differences of the variables
  - Minkowski distance of order 1 or L_1 norm
  - $d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$

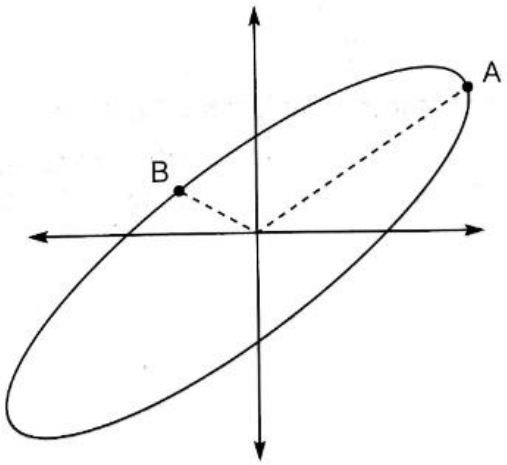
Measuring Similarity: Distance Measures

- Euclidean distance
  - Straight-line distance
  - Minkowski distance of order 2 or L_2 norm
  - $d_{ij} = (\sum_{k=1}^{n} |x_{ik} - x_{jk}|^2)^{1/2}$
  - For metric variables such as ratio or interval
  - Useful to measure distance for the same units
- Maximum distance (Chebyshev distance)
  - The largest distance on any one dimensions
  - $d_{ij} = \max_k(|x_{ik} - x_{jk}|)$
  - Minkowski distance of order ∞ or L_∞ metric/sup-metric

Measuring Similarity: Distance Measures

- Mahalanobis distance
  - The squared generalized Euclidean distance: accounts for the standard deviation of each variable and the correlations among variables
  - $D_{ij}^2 = \left(\mathbf{x}_\mathrm{i} - \mathbf{x}_j\right)^T \Sigma^{-1} \left(\mathbf{x}_\mathrm{i} - \mathbf{x}_j\right)$ where Σ is the population covariance matrix of $\mathbf{X}$

FIGURE 8.8
Diagram demonstrating Mahalanobis distance: An isodistance contour in two dimensions



Measuring Similarity: Matching Measures

- Matching Measures
  - For nominal scale data (non-metric clustering variables)
  - A simple matching measure for object pair (i, j) is obtained by counting the number of matches and dividing by the total number of attributes
  - The larger the value, the greater the similarity between the two objects

TABLE 8.4 Profiles of four soft drink brands across four attributes

	Cola Flavor	Caffeine	Diet	Manufactured by Coke
Coke	1	1	0	1
Pepsi	1	1	0	0
Diet Coke	1	1	1	1
Caffeine-Free Diet Coke	1	0	1	1

TABLE 8.5 Similarity measures of four soft drink brands based on attribute matching

	Coke	Pepsi	Diet Coke	Caffeine-Free Diet Coke
Coke				
Pepsi	3/4			
Diet Coke	3/4	2/4		
Caffeine-Free Diet Coke	2/4	1/4	3/4	

Measuring Similarity: Correlational Measures

- Correlation coefficient
  - Note that a correlation coefficient is not always an appropriate measure of similarity.
  - Correlation is an Index to measure linear relationship between two continuous(ratio, interval scale) variables
  - (1,2,1,2) and (9,10,9,10): correlation is 1 but not close
  - (1,2,1,2) and (1,1,2,2): correlation is 0 but closer
  - Before using a correlation matrix as input to a clustering routine, it is important to verify that the data are scaled in such a way (e.g., standardized by observation) that the results of the analysis can be interpreted appropriately.

Data Standardization

- Clustering variables should be standardized whenever possible to avoid problems resulting from the use of different scale values among clustering variables.
  - Standardizing the variables
    - Most common standardization is Z scores
  - Using a standardized distance measure
    - Use Mahalanobis distance
  - Standardizing by observation
    - If groups are to be identified according to an individual's response style, then within-case or row-centering standardization is appropriate.

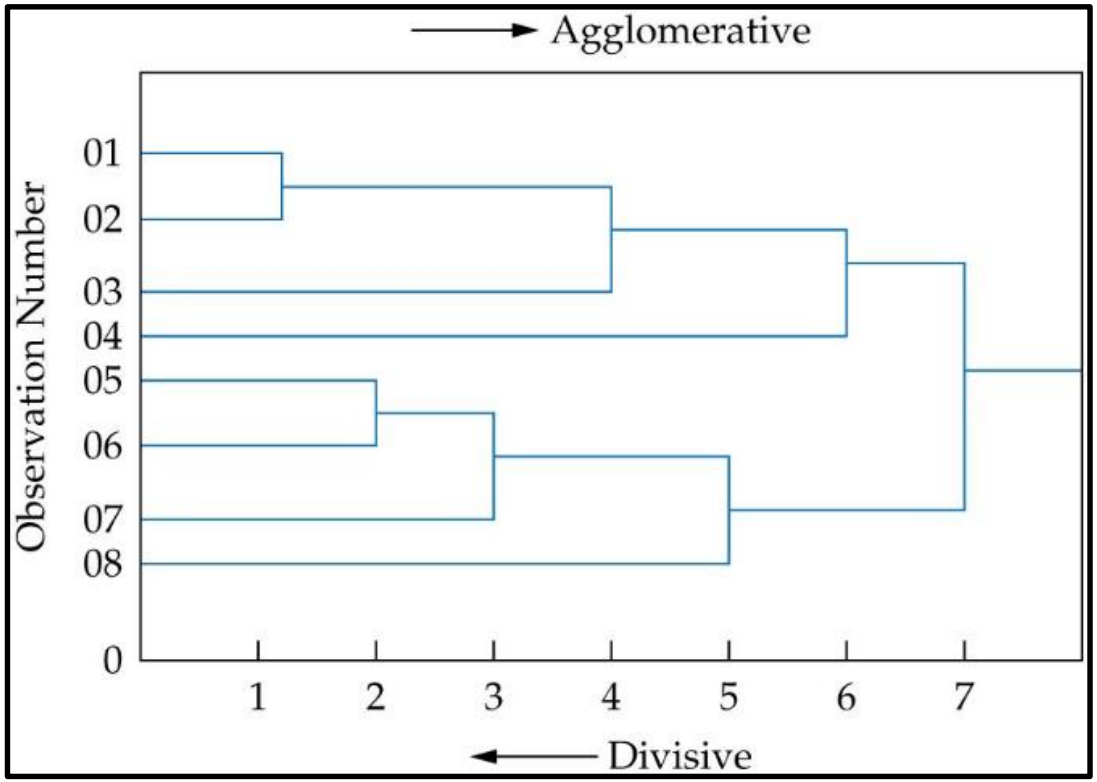
Two Approaches

- Hierarchical
  1. Agglomerative Methods: Most common approach is where all objects start as separate clusters and then are joined sequentially such that each step forms a new cluster joining by two clusters at a time until only a single cluster remains
  2. Divisive Methods

- Non-hierarchical (Partitioning)
  - The number of clusters is specified by the analyst and then the set of objects are formed into that set of groupings.

Two Types of Hierarchical Clustering Procedures

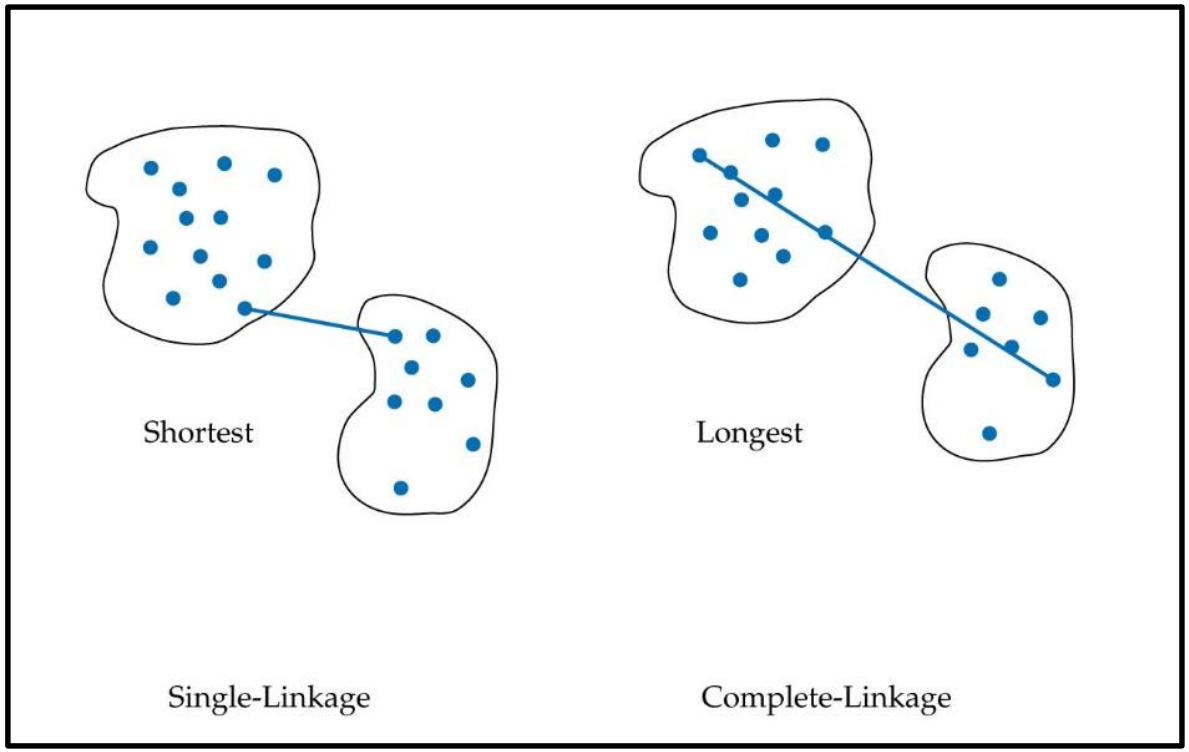
- **Agglomerative Methods**
  - Buildup: all observations start as individual clusters, join together sequentially.
  - AGNES (Agglomerative Nesting)
- **Divisive Methods**
  - Breakdown: initially all observations in a single cluster, then divided into smaller clusters.
  - DIANA (Divisive Analysis)



Clustering Algorithms - Agglomerative

- Provides a method for combining clusters which have more than one observation
- Most widely used algorithms:
  1. Single Linkage (nearest neighbor) – shortest distance from any object in one cluster to any object in the other.
  2. Complete Linkage (farthest neighbor) – based on maximum distance between observations in each cluster.
  3. Average Linkage – based on the average similarity of all individuals in a cluster.
  4. Centroid Method – measures distance between cluster centroids.
  5. Ward's Method – based on the total sum of squares within clusters.

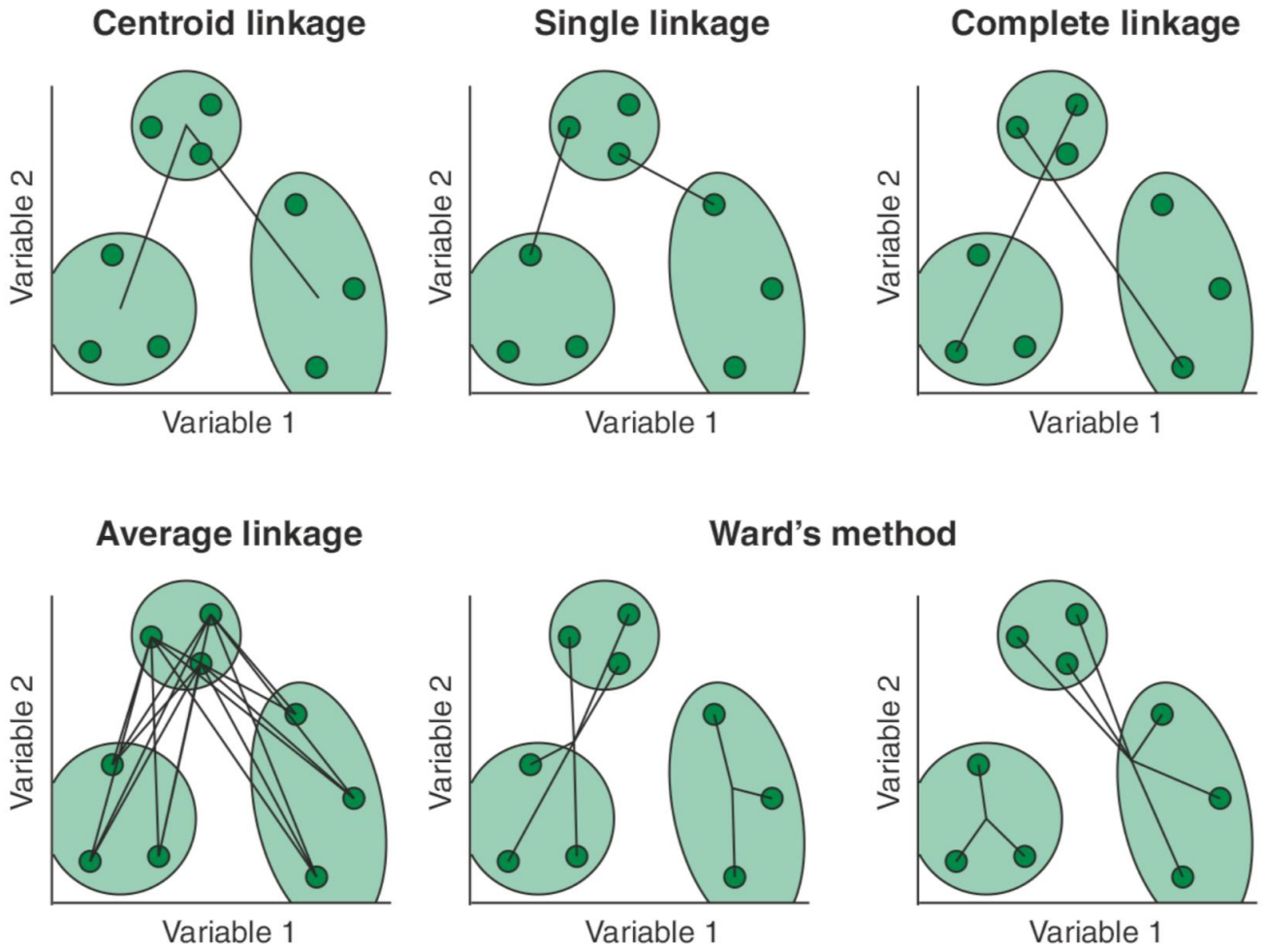
Single-Linkage Versus Complete Linkage



Single linkage: Similarity based only on on two closest observations.

Complete linkage: Similarity based only two farthest observations.

Agglomerative Clustering Methods



Sourece: https://livebook.manning.com/book/machine-learning-for-mortals-mere-and-otherwise/chapter-17/50

Agglomerative Clustering: Single Linkage Clustering

1. Start with all objects in separate clusters (i.e., n clusters with one object in each). Denote these clusters $C_1, C_2, C_3, \ldots, C_n$. In this initial step, the distance between two clusters is defined to be the distance between the two objects they contain; that is, $d_{C_i C_j} = d_{ij}$. Let $t = 1$ be an index of the iterative process.
2. Find the *smallest* distance between any two clusters. Denote these two closest clusters C_i and C_j.
3. Amalgamate clusters C_i and C_j to form a new cluster denoted C_{n+t}.
4. Define the distance between the new cluster C_{n+t} and all remaining clusters C_k as follows:

$$\boldsymbol{d_{C_{n+t}C_k} = \min\left\{d_{C_iC_k}, d_{C_jC_k}\right\}.}$$

5. Add cluster C_{n+t} as a new cluster and remove clusters C_i, and C_j. Let $t = t + 1$.
6. Return to 2 and continue until one cluster remains.

Agglomerative Clustering algorithm: Alternatives

- Complete Linkage
  - Define the distance between the new cluster C_{n+t} and all remaining clusters C_k as follows:

$$d_{C_{n+t}C_k} = \max\left\{d_{C_iC_k}, d_{C_jC_k}\right\}$$

- Average Linkage
  - The new distance is defined as the average distance between cluster C_k and the new cluster C_{n+t} as follows:

$$d_{C_{n+t}C_k} = \frac{n_i d_{C_iC_k} + n_j d_{C_jC_k}}{n_i + n_j}$$

Agglomerative Clustering algorithm: Alternatives

- Centroid Method
  - Let d_{ij}^2 represent the squared Euclidean distance between objects i and j. If $C = \{i, j\}$, then the squared distance between object k and the centroid of C can be written as

$$d_{kC}^2 = \frac{d_{ik}^2 + d_{jk}^2}{2} - \frac{d_{ij}^2}{4}$$

  - In general, the squared distance between any cluster C_k and the new cluster C_{n+t} created by joining C_i and C_j can be written as

$$d^2(C_k, C_i \cup C_j) = \frac{n_{C_i} d_{C_k,C_i}^2 + n_{C_j} d_{C_k,C_j}^2}{n_{C_i} + n_{C_j}} - \frac{n_{C_i} n_{C_j} d_{C_i,C_j}^2}{{n_{C_i} + n_{C_j}}^2}$$

Agglomerative Clustering algorithm: Alternatives

- Ward's Method
  - Minimum Variance Method: Finding the Smallest Within-Cluster Sum of Squares (Minimum Within-Group Variance)
  - Different from *Pair Group Methods*
  - Often, produces equal-sized clusters
  - Often produces a clustering solution — if the tree is "cut" in the right place — that is similar to the partitioning methods (K-means clustering)

Example: 1. Single Linkage

- 5 data points: (1,1), (2,1), (2,4), (4,3), (5,4)

Sourece: https://rfriend.tistory.com/

Example: 2. Complete Linkage

- 5 data points: (1,1), (2,1), (2,4), (4,3), (5,4)

Sourece: https://rfriend.tistory.com/

Example: 3. Average Linkage

- 5 data points: (1,1), (2,1), (2,4), (4,3), (5,4)

Sourece: https://rfriend.tistory.com/

Example: 4. Centroid Method

- 5 data points: (1,1), (2,1), (2,4), (4,3), (5,4)

Sourece: https://rfriend.tistory.com/

Example: 5. Ward's method

Sums-of-Squared Distances

First Stage:	A = 2	B = 5	C = 9	D = 10	E = 15
Second Stage:		AB = 4.5	BD = 12.5		
		AC = 24.5	BE = 50.0		
		AD = 32.0	CD = 0.5		
		AE = 84.5	CE = 18.0		
		BC = 8.0	DE = 12.5		
Third Stage:	CDA = 38.0	CDB = 14.0	CDE = 20.66	AB = 5.0	
	AE = 85.0	BE = 50.5			
Fourth Stage:		ABCD = 41.0	ABE= 93.17	CDE = 25.18	
Fifth Stage:			ABCDE = 98.8		

Comparing the Agglomerative Algorithms

1. Single linkage
   - probably the most versatile algorithm, but poorly delineated cluster structures within the data produce unacceptable snakelike “chains” for clusters.
2. Complete linkage
   - eliminates the chaining problem, but only considers the outermost observations in a cluster, thus impacted by outliers.
3. Average linkage
   - generates clusters with small within-cluster variation and less affected by outliers.
4. Centroid linkage
   - like average linkage, is less affected by outliers.
5. Ward’s method
   - most appropriate when the researcher expects somewhat equally sized clusters, but easily distorted by outliers.

How to Select a Clustering Method? Rough (and Practical) Guidelines

- **Cluster metaphor**: "I preferred this method because it constitutes clusters such (or such a way) which meets with my concept of a cluster in my particular project“
- **Data/method assumptions**: "I preferred this method because my data nature or format predispose to it“
- **Internal validity**: "I preferred this method because it gave me most clear-cut, tight-and-isolated clusters“
- **External validity**: "I preferred this method because it gave me clusters which differ by their background or clusters which match with the true ones I know“
- **Cross-validity:** "I preferred this method because it is giving me very similar clusters on equivalent samples of the data or extrapolates well onto such samples“
- **Interpretation:** "I preferred this method because it gave me clusters which, explained, are most persuasive that there is meaning in the world“
- **Gregariousness:** "I preferred this method because it gave with my data similar results with a number of other methods among all those I probed"

Sourece: https://stats.stackexchange.com/questions/195456/how-to-select-a-clustering-method-how-to-validate-a-cluster-solution-to-warran

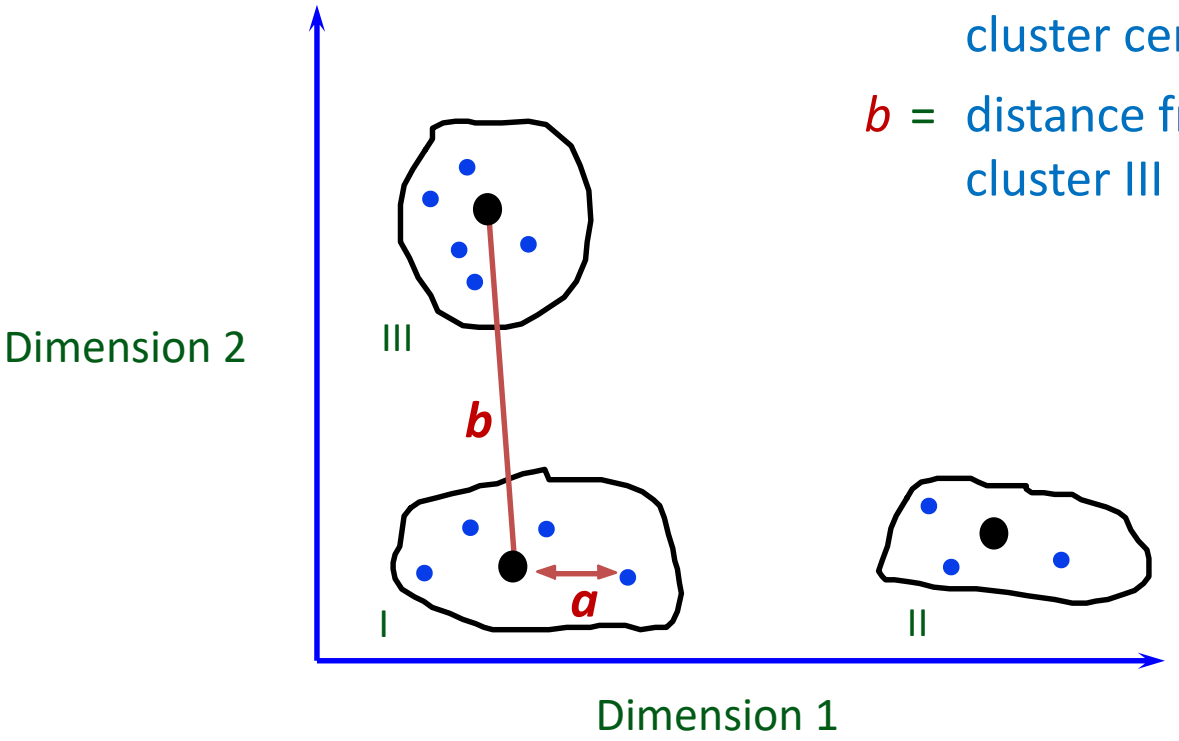
Non-Hierarchical Clustering (Partitioning)

- Goal: To divide the sample into a predetermined number K of non-overlapping groups
- Different from agglomerative clustering
- To measure within-group similarity and between-group difference
- To find the best of these partitions (at least locally optimal if not globally optimal)

K-Means Clustering

a = distance from member to cluster center

b = distance from cluster I to cluster III



K-Means Clustering

1. Select an initial partition of the data into K clusters
2. Calculate the centroid for each cluster C, $\bar{\mathbf{x}}_C$
3. Calculate the sum of squared distances of each object to its cluster centroid (ESS):

$$\text{ESS} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_{C(i)})^{\mathrm{T}} (\mathbf{x}_i - \bar{\mathbf{x}}_{C(i)})$$

   where $C(i)$ is the cluster for object i. Goal is to make ESS as small as possible since it means the within-group distances among observations.
4. Reassign each object i to the cluster whose centroid is closest. If cluster membership remains unchanged, the process has converged to al least a local minimum. If the cluster membership of at least one object has changed, then return to step 2 with the new partition.

K-Means Clustering

FIGURE 8.15
Initial two-cluster partition of seven observations on two variables

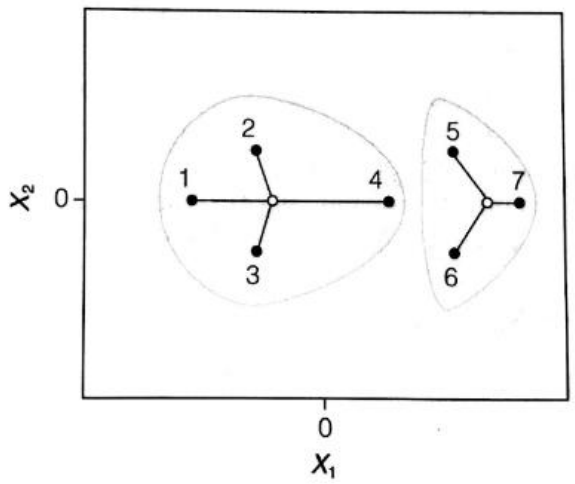
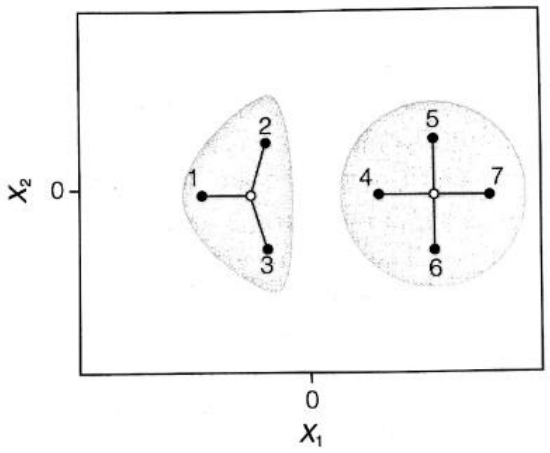


FIGURE 8.16
Final two-cluster partition after reassignment of object 4



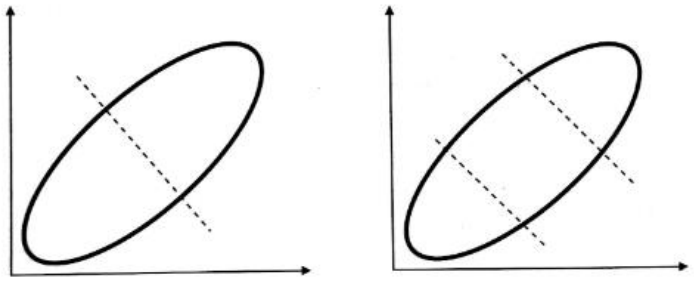
First iteration: ESS=7.836

Second iteration: ESS=6.775

K-Means Clustering

- Selecting the Initial Partition
  - Important to choose an initial partition
  - Most software - Heuristic for selecting the initial partition
- Properties of the Solution
  - Not Hierarchical in nature unlike the agglomerative procedures
  - Each partition is arrived at separately; the three-cluster partition does not take the two-cluster partition as a starting

FIGURE 8.17
Clustering solutions from partitioning methods are not necessarily hierarchical



(a) $K = 2$-cluster solution (b) $K = 3$-cluster solution

K-Means Clustering

How Many Clusters?

- Depend on primary concern of the problem:
  - Simplicity: smaller clusters is better
  - Adequacy: if reducing within-group heterogeneity is the object, then more clusters are better
- Weakness of ESS: ESS decreases as clusters increases
- Calinski and Harabasz (1974):

$$\text{pseudo} - F = \frac{\text{tr}[\mathbf{B}/(K-1)]}{\text{tr}[\mathbf{W}/(n-K)]}$$

  where **B** is the between-clusters sum of squares matrix and **W** is the within-clusters sum of squares matrix
- In general, the larger the pseudo-*F*, the more ”efficient” the partition is in reducing within-group heterogeneity

K-Means Clustering

Interpreting the Final Solution

- Location: Cluster Centroid
- But, cluster centroid does not capture degree of overlapping
- Decompose total sum of squares(TSS) of X into the within-cluster sum of squares(WSS) and between-cluster sum of squares(BSS)

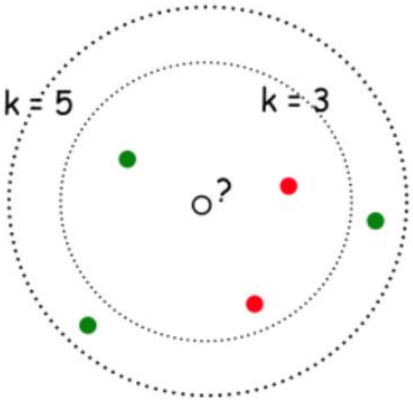
$$TSS = WSS + BSS$$

- The tighter the clusters, the smaller WSS and the larger BSS

$$R^2 = \frac{BSS}{TSS} \quad \text{or} \quad \frac{R^2}{1 - R^2}$$

K-Nearest Neighbor Algorithm vs. K-Means Clustering

- K-Nearest Neighbors (K-NN) Algorithm
  - A supervised algorithm used for classification. We have some labeled data.
  - In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors



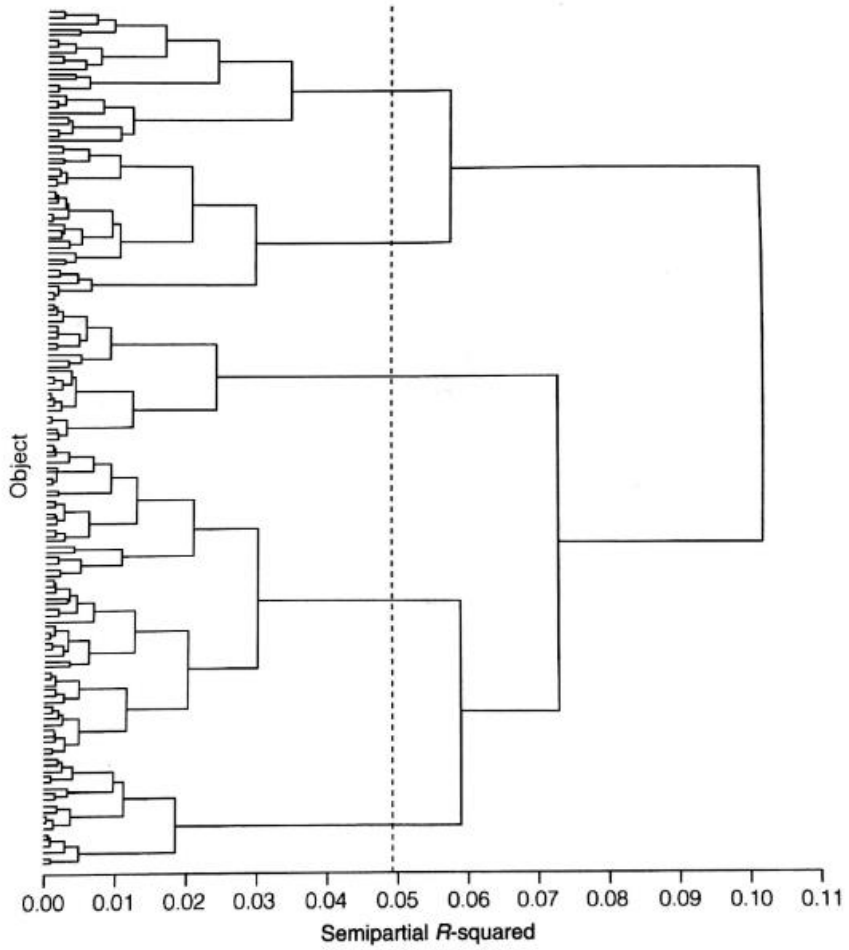
Sample Problem: Preference Segmentation

- 303 MBA students indicating preference about 10 cars: for each car, 10 scale (1=low, 10=high)
- Goal: segmentation
- Clustering method: Ward's Method

Sample Problem: Preference Segmentation

- Ward's Method – 5 clusters with larger pseudo-*F* (than 4 or 6)

FIGURE 8.18
Dendrogram from Ward's method applied to car preference data



Sample Problem: Preference Segmentation

TABLE 8.8 Summary of *K*-means cluster analysis of car preference data for $K = 5$ cluster solution

Cluster	Frequency
1	27
2	35
3	34
4	39
5	17

	Statistics for Variables			
Variable	Total STD	Within STD	R^2	$R^2/(1 - R^2)$
BMW	1.880	1.577	0.314	0.459
Ford	2.172	1.898	0.257	0.346
Infiniti	1.795	1.692	0.136	0.157
Jeep	2.068	1.691	0.349	0.535
Lexus	1.917	1.747	0.192	0.237
Chrysler	1.348	1.268	0.139	0.161
Mercedes	1.818	1.645	0.203	0.255
Saab	2.150	1.989	0.167	0.201
Porsche	2.480	1.483	0.652	1.870
Volvo	2.237	1.590	0.508	1.032
Overall	2.008	1.669	0.327	0.487

Sample Problem: Preference Segmentation

	Cluster Means				
Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
BMW	6.93	6.74	7.24	5.90	3.71
Ford	2.93	6.37	4.59	4.74	4.88
Infiniti	4.37	4.31	4.50	2.97	3.18
Jeep	3.26	6.66	5.35	5.67	7.06
Lexus	5.74	5.66	5.44	3.67	4.82
Chrysler	1.11	2.46	1.41	1.54	2.24
Mercedes	7.11	6.17	6.06	4.87	4.88
Saab	5.85	5.94	4.24	3.95	4.41
Porsche	7.33	6.97	2.97	7.15	2.82
Volvo	2.93	5.37	2.74	2.08	6.53

Pseudo F-statistic = 17.89

Pros and Cons of Hierarchical Methods

- Pros
  - Simplicity – generates tree-like structure which is simplistic portrayal of process
  - Measures of similarity – multiple measures to address many situations
  - Speed – generate entire set of cluster solutions in single analysis

- Cons
  - Permanent combinations – once joined, clusters are never separated
  - Impact of outliers – outliers may appear as single object or very small clusters
  - Large samples – not amenable to very large samples, may require samples of large populations

Pros and Cons of Nonhierarchical Methods

- Pros
  - Results are less susceptible to:
    - outliers in the data
    - the distance measure used, and
    - the inclusion of irrelevant or inappropriate variables
  - Can easily analyze very large data sets

- Cons
  - Best results require knowledge of seed points
  - Difficult to guarantee optimal solution
  - Generates typically only spherical and more equally sized clusters
  - Less efficient in examining wide number of cluster solutions

Determining the Number of Clusters

Stopping rules

- Foundational principle – a natural increase in heterogeneity comes from the reduction in number of clusters
- Common to all stopping rules:
  - evaluating the trend in heterogeneity across cluster solutions to identify marked increases
  - substantive increases in this trend indicate relatively distinct clusters were joined and that the cluster structure before joining is a potential candidate for the final solution

Determining the Number of Clusters

1. ***Elbow method***: total within-cluster sum of square is minimized
2. ***Average Silhouette method***: the average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering
3. ***Gap statistic method***: The gap statistic compares the total intra-cluster variation for different values of *k* with their expected values under null reference distribution of the data (i.e. a distribution with no obvious clustering). The estimate of the optimal clusters will be the value that maximizes the gap statistic. This means that the clustering structure is far away from the uniform distribution of points
4. ***Pseudo-F***: *K*-means clustering

https://uc-r.github.io/kmeans_clustering

Deriving Clusters

Selection of hierarchical or nonhierarchical methods is based on:

- Hierarchical clustering solutions are preferred when:
  - A wide range of alternative clustering solutions is to be examined
  - The sample size is moderate (under 300–400, not exceeding 1,000) or a sample of the larger data set is acceptable (not very large sample)
- Nonhierarchical clustering methods are preferred when:
  - The number of clusters is known and/or initial seed points can be specified according to some practical, objective, or theoretical basis
  - Outliers cause concern, because nonhierarchical methods generally are less susceptible to outliers

Deriving Clusters

A combination approach using a hierarchical approach followed by a nonhierarchical approach is often advisable

- A hierarchical approach is used to select the number of clusters and profile cluster centers that serve as initial cluster seeds in the nonhierarchical procedure
- A nonhierarchical method then clusters all observations using the seed points to provide more accurate cluster memberships

Deriving the Final Cluster Solution

No single objective procedure is available to determine the correct number of clusters; rather the researcher must evaluate alternative cluster solutions on the following considerations to select the optimal solution:

- Single-member or extremely small clusters are generally not acceptable and should be eliminated
- For hierarchical methods, ad hoc stopping rules, based on the rate of change in a total heterogeneity measure as the number of clusters increases or decreases, are an indication of the number of clusters
- All clusters should be significantly different across the set of clustering variables
- Cluster solutions ultimately must have theoretical validity assessed through external validation

Interpreting, Profiling, and Validating Clusters

The cluster centroid, a mean profile of the cluster on each clustering variable, is particularly useful in the interpretation stage:

- Interpretation involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters
- Cluster solutions failing to show substantial variation indicate other cluster solutions should be examined
- The cluster centroid should also be assessed for correspondence with the researcher's prior expectations based on theory or practical experience