7.1 GROUPING CONSUMERS

Market segmentation analysis based on consumer data is exploratory and involves unstructured datasets. Consumer preferences are diverse and do not exhibit clear groups when plotted. Instead, preferences are scattered across the entire plot. Extracting market segments from such data heavily relies on the assumptions made about the segment structure. Thus, the outcome of a segmentation analysis is influenced by both the data itself and the chosen extraction algorithm. Segmentation methods play a significant role in shaping the resulting segmentation solution.

FLOW OF GROUPING :

1. Market segmentation analysis is exploratory.
2. Consumer datasets are often unstructured.
3. Consumer preferences do not exhibit clear groups when plotted.
4. Preferences are spread across the entire plot.
5. Extracting market segments depends on assumptions about segment structure.
6. Segmentation results are influenced by data and extraction algorithms.
7. Segmentation methods shape the segmentation solution.
8. Results depend on data, assumptions, and algorithms.
9. Consider limitations and biases when interpreting segmentation results.

Data set and segment characteristics informing extraction algorithm selection

Data set characteristics	• Size (number of consumers, number of segmentation variables) • Scale level of segmentation variables (nominal, ordinal, metric, mixed) • Special structure, additional information
Segment characteristics	• Similarities of consumers in the same segment • Differences between consumers from different segments • Number and size of segments

7.2 DISTANCE BASED METHODS

	beach	action	culture
Anna	100	0	0
Bill	100	0	0
Frank	60	40	0
Julia	70	0	30
Maria	80	0	20
Michael	0	90	10
Tom	50	20	30

PROBLEM STATEMENT
Grouping tourists based on similar activities and behavioral patterns.

DATASET
Artificial dataset was generated for seven tourists.

Market segmentation aims to group consumers based on their similar needs or behavior. In this case, the goal is to identify groups of tourists with similar vacation activity patterns. Anna and Bill have identical profiles and should be in the same segment since they exhibit the same characteristics. However, Michael stands out as he is the only one not interested in going to the beach, which sets him apart from the other tourists. To find these groups of similar tourists, a measure of similarity or dissimilarity is needed, typically represented mathematically as a distance measure. This distance measure helps quantify the differences between tourists and enables the identification of distinct segments based on their preferences and behaviors.

<u>DISTANCE MEASURES:</u>

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

In the given example, Anna's vacation activity profile is represented by the vector x1 = (100, 0, 0), while Tom's vacation activity profile is represented by the vector x7 = (50, 20, 30). These vectors represent observations in a matrix X.

To measure the distance between two vectors, various approaches exist and are commonly used in cluster analysis and market segmentation. A distance is a function, denoted as d(·, ·), with two arguments: the vectors x and y for which the distance is being calculated. The output is a nonnegative value representing the distance between the two vectors.

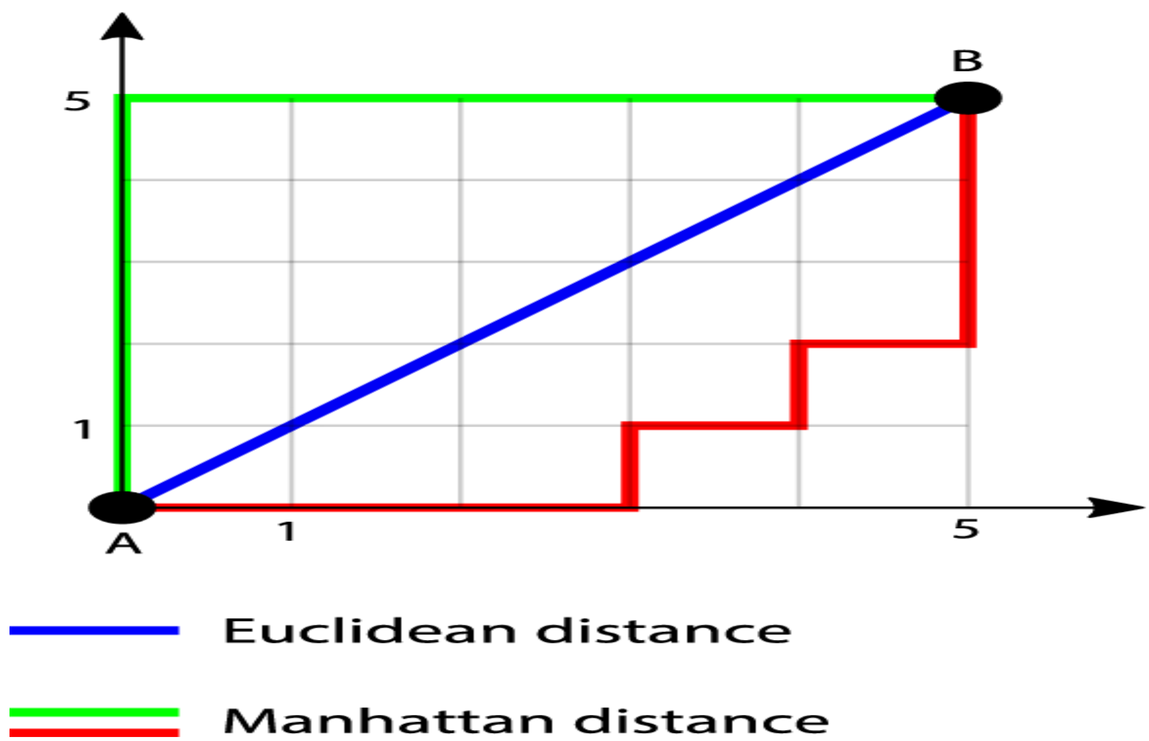
An analogy to understand distance is through geography. If we are interested in the distance between two cities, the vectors represent the locations of the cities, and the distance could be measured as the length of the air route between them in kilometers. However, there are other valid measures of natural distance between cities, such as the distance a car has to drive on roads to travel from one city to the other.

FEW CRITERIA:

- **d(x, y) = d(y, x)**
- **d(x, y) = 0 ⇔ x = y.**
- **d(x, z) ≤ d(x, y) + d(y, z).**

COMMON DISTANCE MEASURES USED:

- Euclidean distance
- Manhattan distance
- Asymmetric binary distance



- The asymmetric binary distance is a distance measure used in market segmentation analysis.
- It only considers dimensions where at least one of the two vectors has a value of 1, ignoring dimensions with 0s in both vectors.
- It treats 0s and 1s differently, considering similarity only if the vectors share 1s but not if they share 0s.
- Dissimilarity between vectors increases if one vector has a 1 and the other does not.
- The asymmetric binary distance has implications for market segmentation analysis.
- Unusual activities, such as horseback riding and bungee jumping, may not contribute much to segment extraction if their overall proportions in the population are low.
- Commonalities in not engaging in certain activities are not informative for segment extraction.
- The proportion of common 1s over dimensions where at least one vector contains a 1 represents the asymmetric binary distance.
- In the tourist example, it is the number of common vacation activities divided by the number of vacation activities at least one of the two tourists engages in.
- A symmetric binary distance measure, treating 0s and 1s equally, can be obtained by using the Manhattan distance between the two vectors.
- In this case, the distance is equal to the number of vacation activities where the values differ.

HIERARCHICAL METHODS:

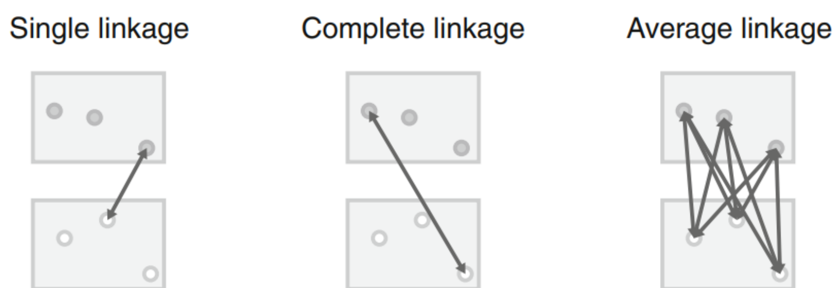
Hierarchical clustering methods closely resemble how humans would naturally group data. They mimic the process of dividing a set of observations (consumers) into a desired number of groups (segments). When aiming for a single large market segment ($k = 1$), the only solution is to have one big segment that includes all consumers in the dataset. On the other hand, if the goal is to have as many market segments as there are consumers in the dataset ($k = n$), each consumer becomes their own individual segment, resulting in n segments. Market segmentation analysis typically falls between these two extremes, finding an optimal number of segments that strike a balance between homogeneity within segments and heterogeneity across segments.

- Divisive hierarchical clustering starts with the complete dataset and splits it into two segments, then continues to split each segment into further segments until each consumer has their own segment.
- Agglomerative hierarchical clustering starts with each consumer representing their own segment and gradually merges the two closest segments until the entire dataset forms one large segment.
- Both approaches result in a sequence of nested partitions, ranging from one group (segment) to n groups (segments).
- The partitions are nested because each partition with $k+1$ groups is obtained by splitting one group from the partition with k groups.

- Numerous algorithms have been proposed for both divisive and agglomerative clustering, with Lance and Williams' framework serving as a unifying approach for agglomerative clustering.
- Standard implementations of hierarchical clustering perform optimal steps in each iteration, resulting in a deterministic algorithm without random components.
- Each application of hierarchical clustering to the same dataset will yield the exact same sequence of nested partitions.

Assuming two sets X and Y of observations (consumers), the following linkage methods can be used :

- Single Linkage
- Complete Linkage
- Average Linkage



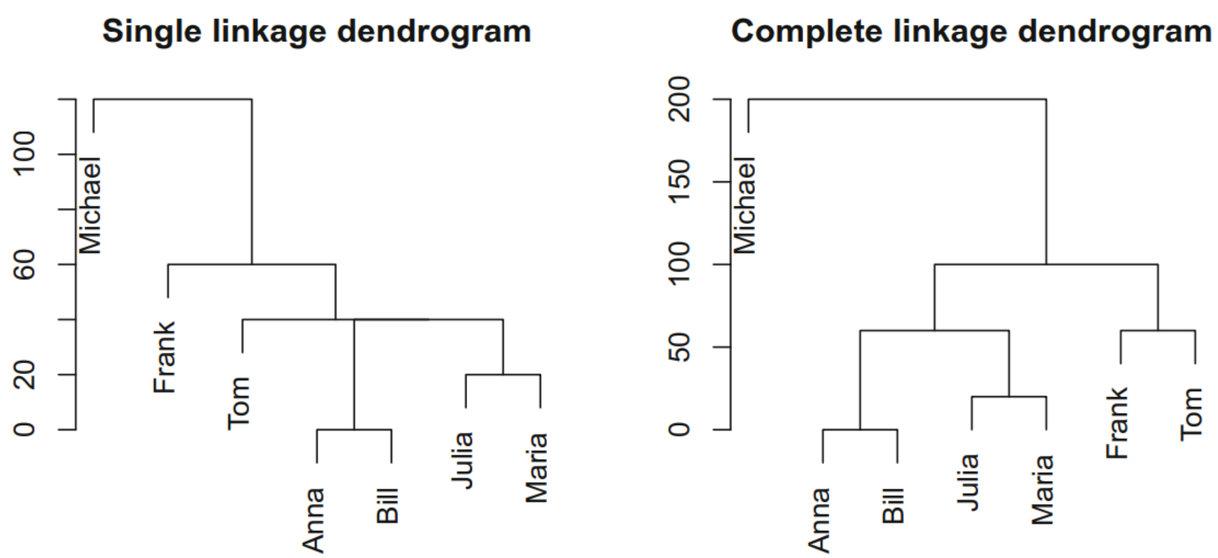
Both divisive and agglomerative clustering rely on a distance measure between groups of observations or segments. This distance measure is determined by specifying two factors: (1) a distance measure, denoted as d(x, y), between individual observations (consumers) x and y, and (2) a linkage method.

The distance measure, d(x, y), quantifies the dissimilarity or similarity between two individual observations. It provides a way to compare and assess the differences or similarities in their characteristics or attributes.

The linkage method, on the other hand, generalizes how distances between groups of observations are calculated based on the pairwise distances between individual observations within those groups. It determines how the distance between two groups is computed from the distances of their constituent observations.

By combining the distance measure and linkage method, both divisive and agglomerative clustering algorithms can determine the distances between groups of observations and iteratively merge or split segments based on these distances.

The result of hierarchical clustering is typically presented as a dendrogram. A dendrogram is a tree diagram. The root of the tree represents the one-cluster solution where one market segment contains all consumers. The leaves of the tree are the single observations (consumers), and branches in-between correspond to the hierarchy of market segments formed at each step of the procedure. The height of the branches corresponds to the distance between the clusters. Higher branches point to more distinct market segments. Dendrograms are often recommended as a guide to select the number of market segments.

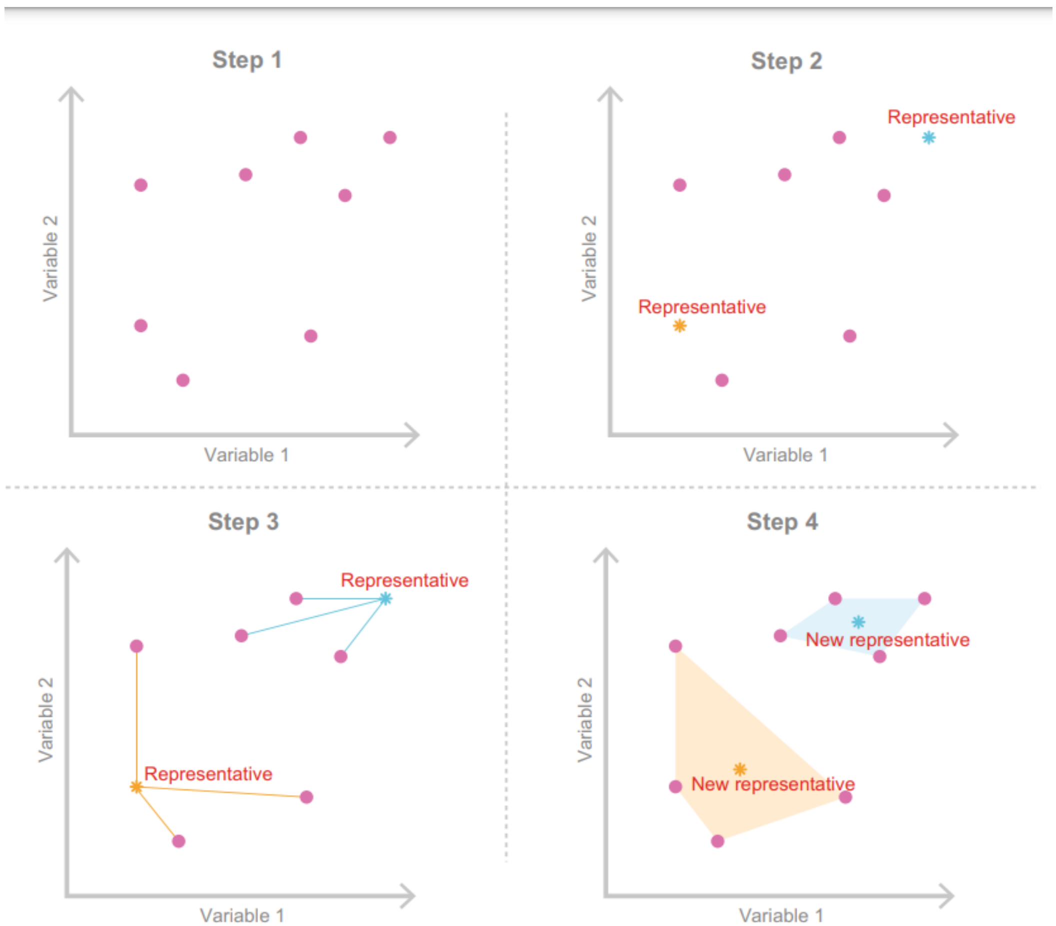


The observations or customers' order as tree leaves is not particular.
Every time there was a split into two branches, the left and right branches may be switched, giving rise to 2n alternative dendrograms for the exact same clustering, where n is the number of splits of consumers included in the data collection. Dendrograms generated by various software programmes may therefore appear different even though they are identical in terms of the market segmentation they reflect. When numerous groups have exactly the same distance, how ties are broken—that is, which two groups are linked first—could be another possible source of variance among software packages.

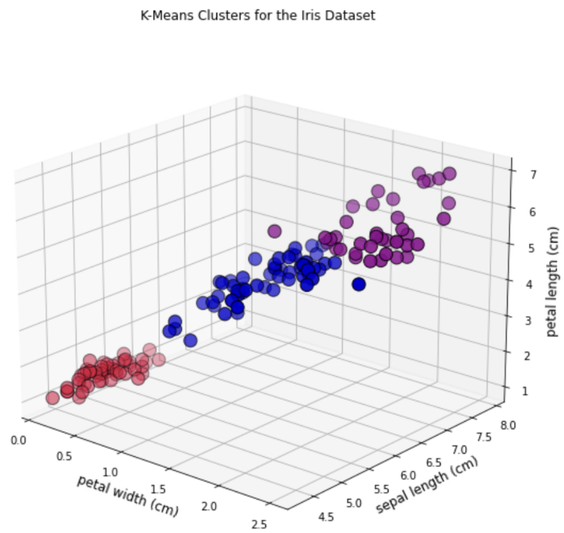
PARTITIONING METHODS:

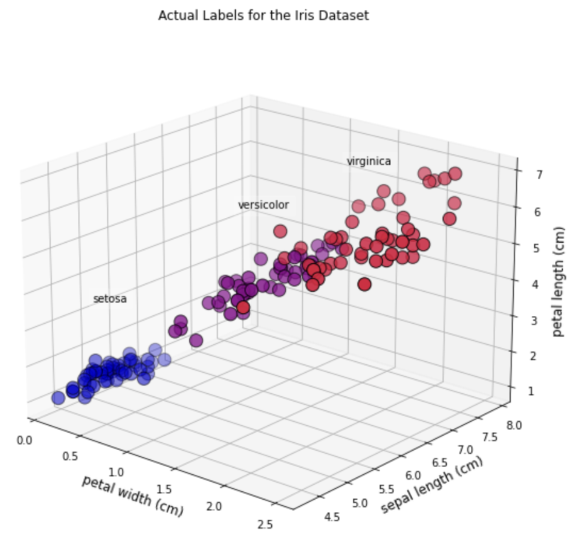
A partitioning clustering algorithm aiming to extract five market segments, in contrast, would only have to calculate between 5 and 5000 distances at each step of the iterative or stepwise process (the exact number depends on the algorithm used). In addition, if only a few segments are extracted, it is better to optimize specifically for that goal, rather than building the complete dendrogram and then heuristically cutting it into segments.

- **<u>K-Means Clustering</u>**



- **K Centroid Clustering**



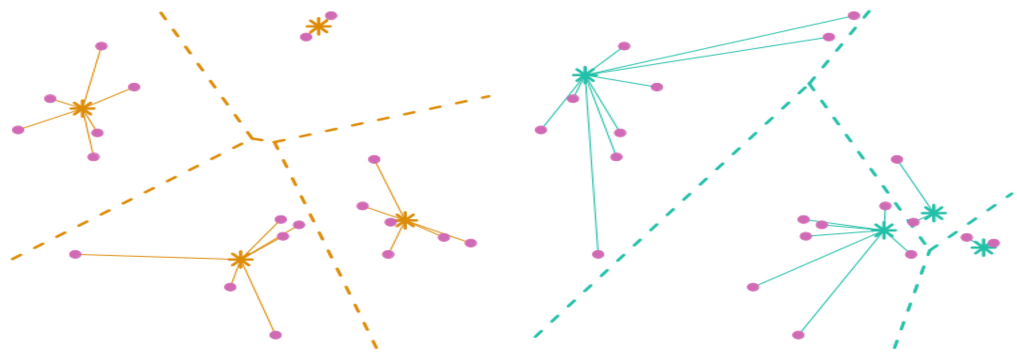


Several improvements have been proposed to enhance the k-means clustering algorithm. One straightforward improvement is to use "smart" initial values instead of randomly selecting k consumers from the dataset as starting points. Randomly selecting initial points can be suboptimal because some of them may end up being very close to each other, which does not represent the overall data space effectively. This increases the risk of the k-means algorithm getting stuck in a local optimum, a good solution but not the best possible one.

To avoid the issue of local optima, one approach is to initialize the algorithm with starting points evenly distributed across the entire data space. By using such evenly spread starting points, the algorithm can better capture the representation of the entire dataset and improve the chances of finding a globally optimal solution.

HARD COMPETITIVE LEARNING

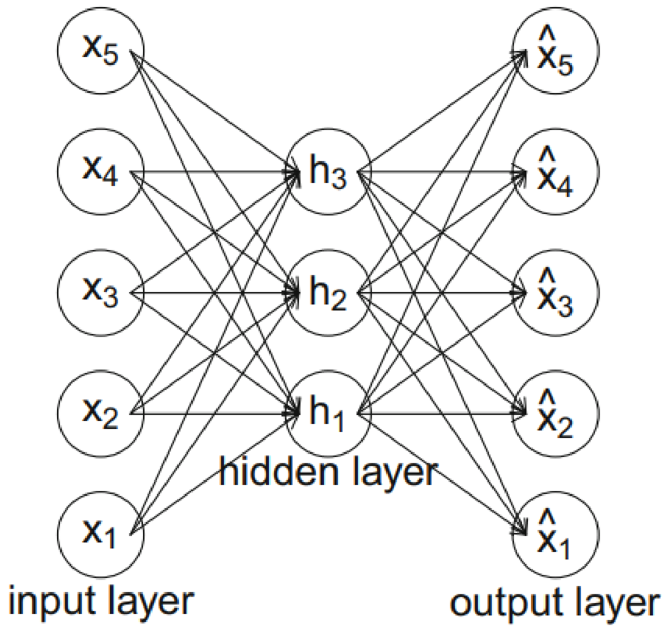
Hard competitive learning, also known as learning vector quantisation (e.g. Ripley 1996), differs from the standard k-means algorithm in how segments are extracted. Although hard competitive learning also minimizes the sum of distances from each consumer contained in the data set to their closest representative (centroid), the process by which this is achieved is slightly different.



Starting point good and bad examples respectively.

<u>NEURAL NETWORKS</u>

- Auto-encoding neural networks are a unique approach to cluster analysis.
- They utilize a single hidden layer perceptron as the main method.
- The network consists of three layers: input, hidden, and output.
- The hidden layer is named so because it has no connections to the outside of the network.
- Each node in the hidden layer is a weighted linear combination of the inputs.
- The weights are represented by arrows connecting the input and hidden layers.
- Non-linear functions are used to determine the values of the hidden nodes.
- The outputs of the network are weighted combinations of the hidden nodes.
- The network is trained by adjusting the parameters to minimize the squared Euclidean distance between inputs and outputs.
- The training process aims to predict the inputs as accurately as possible.
- The number of hidden nodes is usually less than the number of inputs, forcing the network to learn how to represent the data using segment representatives.
- Once trained, the parameters connecting the hidden layer to the output layer are interpreted as segment representatives.
- The parameters connecting the input layer to the hidden layer indicate membership in different segments.
- Consumers with similar values in the hidden layer nodes are considered part of the same segment.



HYBRID APPROACHES

- Two Step Clustering
- Bagged Clustering

7.3 MODEL BASED METHODS :

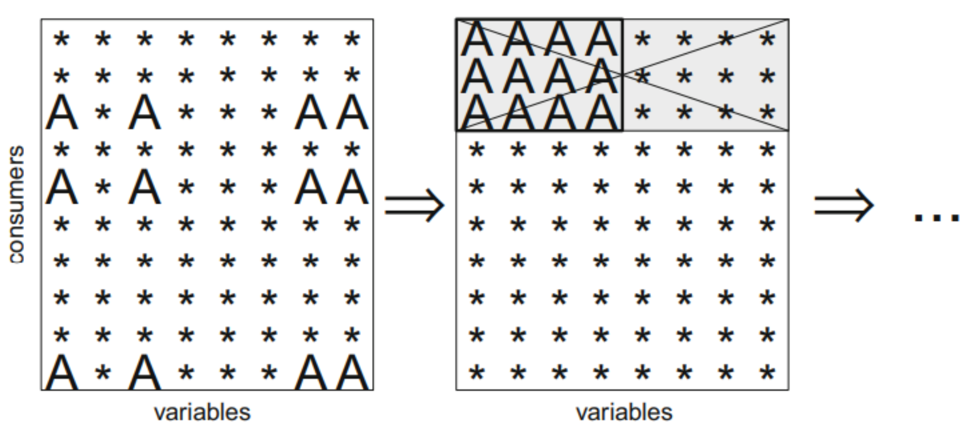
1. **FINITE MIXTURES OF DISTRIBUTIONS:**
    - Normal Distributions
    - Binary Distributions

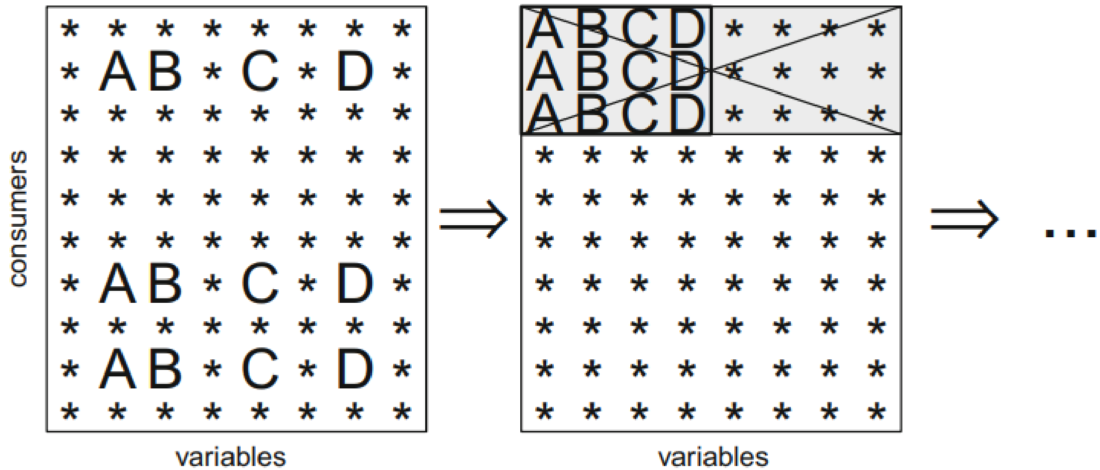
2. **FINITE MIXTURES OF REGRESSIONS:**
    - Linear Regression
    - Lasso Regression
3. **EXTENSIONS AND VARIABLES:**
   Finite mixture models are more complex than distance-based methods. They offer flexibility by allowing the use of any statistical model to describe a market segment.

- Different data characteristics can be accommodated, such as metric data with mixtures of normal distributions and binary data with mixtures of binary distributions.
- Nominal variables can be handled with mixtures of multinomial distributions or multinomial logit models.
- Ordinal variables require special attention due to response styles, and mixture models can disentangle response style effects from content-specific responses.
- Mixture models combined with conjoint analysis can account for differences in preferences.
- There is a debate in the segmentation literature about whether to model differences between consumers using a continuous distribution or distinct market segments.
- Mixture models can reconcile these positions by recognizing the existence of distinct segments while allowing variation within each segment.
- Mixture of mixed-effects models or heterogeneity models can be used to model demand and capture both distinct segments and within-segment variation.
- Mixture models can also be applied to time series data, clustering the time series and extracting groups of similar consumers.
- Dynamic latent change models, such as Markov chains, can track changes in brand choice and buying decisions over time.
- Mixture models can incorporate both segmentation and descriptor variables, with segmentation variables used for grouping and descriptor variables used to model differences in segment sizes.

7.4 ALGORITHMS WITH INTEGRATED VARIABLE SELECTION

- **BICLUSTERING ALGORITHMS**





- **VARIABLE SELECTION PROCEDURE FOR CLUSTERING BINARY DATA(VSBD):**
  The algorithm works as follows:
  - **Step 1:** Select only a subset of observations with size φ ∈ (0, 1] times the size of the original data set. Brusco (2004) suggests to use φ = 1 if the original data set contains less than 500 observations, 0.2 ≤ φ ≤ 0.3 if the number 7.4 Algorithms with Integrated Variable Selection 149 of observations is between 500 and 2000 and φ = 0.1 if the number of observations is at least 2000.
  - **Step 2:** For a given number of variables V , perform an exhaustive search for the set of V variables that leads to the smallest within-cluster sum-of-squares criterion. The value for V needs to be selected small for the exhaustive search to be computationally feasible. Brusco (2004) suggests using V = 4, but smaller or larger values may be required depending on the number of clusters k, and the number of variables p. The higher the number of clusters, the larger V should be to capture the more complex clustering structure. The higher p, the smaller V needs to be to make the exhaustive search computationally feasible.
  - **Step 3:** Among the remaining variables, determine the variable leading to the smallest increase in the within-cluster sum-of-squares value if added to the set of segmentation variables.
  - **Step 4:** Add this variable if the increase in within-cluster sum-of-squares is smaller than the threshold. The threshold is δ times the number of observations in the subset divided by 4. δ needs to be in [0, 1]. Brusco (2004) suggests a default δ value of 0.5.

- **VARIABLE FACTOR ANALYSIS:**

  In summary, the main points regarding factor-cluster analysis for market segmentation are as follows:

  - Two-step procedure: Factor-cluster analysis is a data-driven market segmentation approach. In the first step, segmentation variables are factor analyzed, and the raw data is discarded. In the second step, market segments are extracted using factor scores obtained from the factor analysis.

  - Conceptual legitimacy: Factor-cluster analysis can be conceptually legitimate when the original variables are replaced with factor scores that represent validated psychological test batteries or other variables designed to load onto factors. However, the factor scores should be determined simultaneously or separately from the data, not in a data-driven manner when groups are suspected.

  - Sample size limitations: Factor-cluster analysis is often used when the number of segmentation variables is high compared to the sample size. Simulation studies suggest that the sample size should be at least 100 times the number of segmentation variables for reliable results. However, many market segmentation studies have a smaller sample size, making it challenging to meet this criterion.

  - Information loss: Factor analysis leads to a substantial loss of information. The percentage of explained variance after factor analysis represents the amount of information retained. Using factor scores for segment extraction means sacrificing a significant portion of the information contained in the original segmentation variables.

  - Data transformation: Factor analysis transforms the data, resulting in segments extracted from a modified version of the consumer data. This differs from extracting segments directly from the original segmentation variables, which may affect the representation and interpretation of market segments.

  - Interpretation challenges: Factor-cluster results are more difficult to interpret compared to results based on the original segmentation variables. Factors contain partial information from multiple variables, making it challenging to translate segment profiles into practical recommendations for the marketing mix.

- Preference for raw data: Cluster analysis using raw data is often considered to produce more accurate or detailed segmentation results as it preserves a greater degree of the original data. Factor-cluster analysis is discouraged for market segmentation purposes, except when developing an instrument for the entire population assuming homogeneity among consumers.

- Performance compared to raw data clustering: Empirical evidence suggests that factor-cluster analysis does not consistently outperform clustering of raw data in terms of identifying the correct market segment structure. Even when the data is generated following a factor-analytic model, factor-cluster analysis may fail to capture the underlying structure effectively.

Overall, factor-cluster analysis has conceptual, information loss, data transformation, interpretation, and performance limitations compared to clustering using raw data for market segmentation purposes.