Impact of linkage on hierarchical clustering and covariance type on GMM clustering

MVE440 Statistical learning for big data

Project aim

- Used a suitable real dataset
  - Shopping dataset

- Discuss the effects of:
  - Hierarchical clustering using different linkage types.
    - Centroid linkage
    - Ward linkage
  - GMM clustering using different covariance types.
    - Diagonal covariance matrix
    - Spherical covariance matrix

Methods

- Standardized the datasets
  - Used StandardScaler from sklearn.preprocessing
  - The chosen features were standardized so that each data type has the same format.
- Setup:
  - Python
  - Library scikit-learn for machine learning
    - scipy.cluster.hierarchy: For hierarchical clustering and dendrograms .
    - sklearn.cluster.AgglomerativeClustering: Performing hierarchical clustering using Ward linkage.
    - sklearn.mixture.GaussianMixture: For GMM and to estimate the parameters of a GMM distribution.
    - sklearn.metrics.silhouette_score: Computing silhouette scores.
- Methods for analysing different covariance matrices:
  - Silhouette strength/width
  - Plot of datapoints with mixture densities with specific covariance matrix
- Methods for analysing linkage:
  - Dendrogram with a “cut”/threshold
  - Silhouette strength/width
  - Plot of clustered datapoints

Dataset

Shopping dataset

- Dataset that segment customers into different groups based on their shopping trends.
- The dataset contains five columns (features) : CustomerID, Gender, Age, Annual Income, and Spending Score and 200 observations, i.e observed on 200 customers .
- To analyse the result in two-dimensional feature space, two of these five columns are chosen.
- The features **Annual Income** (in thousands of dollars) and **Spending Score** (1-100) was chosen due an obvious connection to shopping trends.
- Since the features has different units, they were standardized.

Choosing cluster count – silhouette score

- Used the silhouette width as cluster validation to asses the “goodness” off a given cluster count value.
- The silhouette score measures how similar an object is to its own cluster compared to other clusters.
- The score ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster, i.e. less of a good match to the neighbouring clusters.

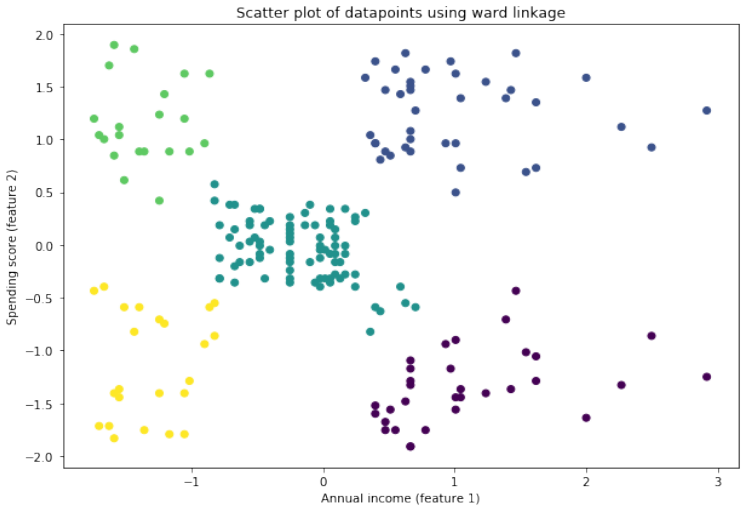
- **Algorithm:**
  - Loop over different number of clusters
  - Compute the cluster center and predict the index of the cluster for each sample
  - Compute the silhouette score with silhouette_score(data, prediction) for each cluster count
  - Plot the number of cluster against the silhouette score and visualise

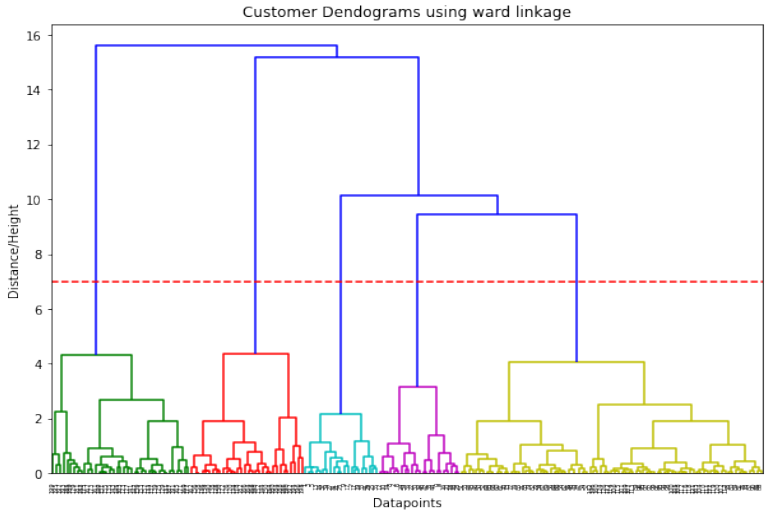
Hierarchical clustering with **ward** linkage

- The Ward linkage uses the **Ward variance minimization algorithm** to minimize the total within-cluster variance.
- The method is implemented by finding the pair of clusters that at each step leads to a minimum increase in the total within-cluster variance after merging.
- This increase is a **weighted squared distance** between cluster centers.

Result:

- Can see clearly from the scatter plot that there are **5 clusters** (indicated by different colours), which are predicted from the silhouette score as well as in the dendrogram, where the threshold/"cut" indicates that there are 5 clusters.



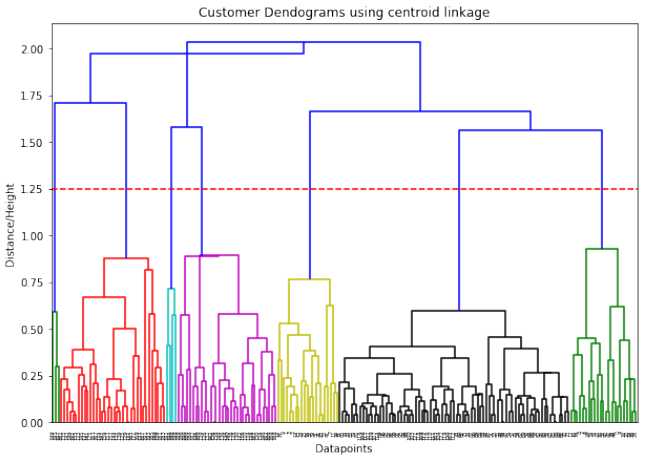


Hierarchical clustering with **centroid** linkage

- The centroid-linkage (or UPGMC) is the distance (Euclidean) **between the two centroids** (mean vectors of length p variables) of two clusters.
- Centroid linkage merges the groups whose means are closest.
- A disadvantage of centroid clustering is that it is a **nonmonotonic hierarchical clustering**:
  - Inversions can occur, which contradicts the assumption that the best merge available is chosen at each step.
  - inversions is presented in the dendrogram as crossing lines.

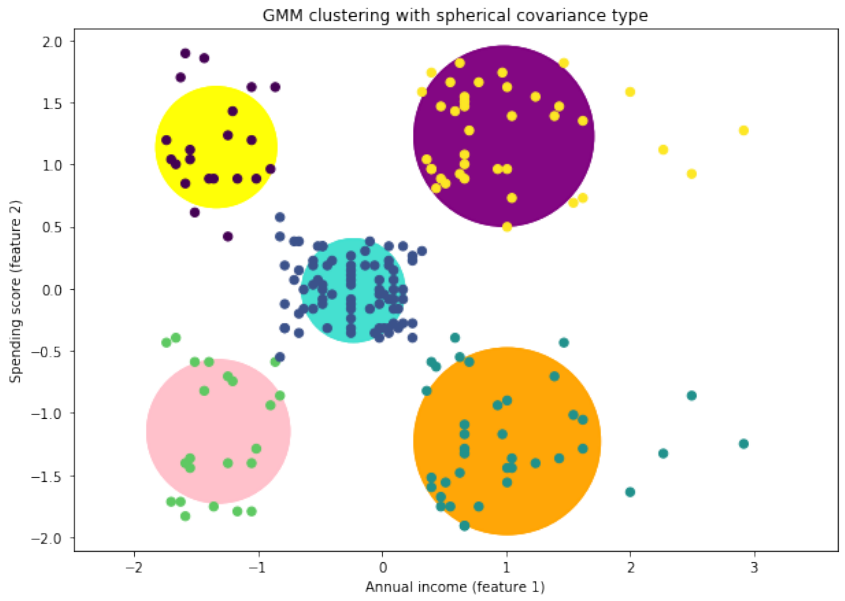
Result:

- One can analyse the **cluster count from the dendrogram to be 7** (Wards linkage predicted 5 clusters)
- It is also possible to visualise inversion from the dendrogram as the crossing of the line at distance/height approximately 1.9.
- Was not able to produce a scatterplot or silhouette plot due to restrains of the function **AgglomerativeClustering**, that was used for clustering.



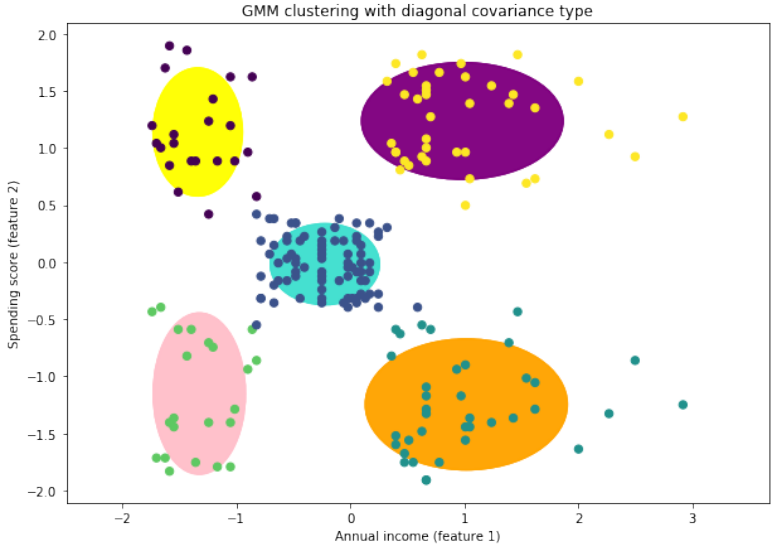
GMM with a **spherical** covariance matrix

- A covariance matrix $\mathbf{C}$ is spherical, if it is proportional to the identity matrix: $\mathbf{C}=\lambda\mathbf{I}$, i.e. if its diagonal and all elements on the diagonal are equal.
- Each component has its own single variance
- The equal values of the variances along the diagonal results in a probability distribution with spherical symmetry , since the spread along each one of the dimensions is exactly the same.
- The number of clusters (5) are obtained from the silhouette plot



GMM with a **diagonal** covariance matrix

- A covariance matrix $\mathbf{C}$ is diagonal if it is diagonal and all elements on the diagonal are variances of each variable, i.e. the diagonal elements are not equal and the off-diagonal elements are zero.
- Each component has its own diagonal covariance matrix
- In contrast to the spherical covariance matrix, the diagonal covariance matrix has different variances on its diagonal, i.e. the spread in each dimension is different, resulting in an elliptic shape of the probability distribution.
- The elliptic shape might enable the GMM with the diagonal covariance matrix to be able to capture spread out data better than the GMM with the spherical covariance matrix.
- The number of clusters (5) are obtained from the silhouette plot.



Conclusion

- Ward and centroid linkage predicts different number of clusters for the shopping dataset.
  - Ward might be a more general method and commonly used when clusters are expected to have a round shape, i.e. for round and dense data.
  - Inversion can be visualised from the dendrogram with centroid linkage, which contradicts the purpose of the linkage criteria.

- Diagonal covariance matrices has different variances on their diagonal resulting in an elliptic shape of the probability distribution, in contrast to the spherical shape that the spherical covariance matrices result in.
  - The diagonal covariance matrices might enable more spread data to be captured than the spherical covariance matrices.
  - For this specific dataset, the diagonal covariances are better at capturing the structure of the data.

- Silhouette width is a good measure for cluster validation as well as validation of cluster count together in addition to dendrograms.

- Performance on high dimensional data
  - Since distance based clustering performs worse in high dimension, ward and centroid linkage that has Euclidean distance as a distance measure, might not always be a good choice due to pairwise distances grow with dimensions.
  - GMM might perform better in higher dimension, given that some assumption is made on the covariance matrix, perhaps that all covariance matrices are the same for all components, to simplify complexity in estimation of parameters.