DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DETERMINING THE OPTIMAL NUMBER OF CLUSTERS IN CLUSTERING ANALYSIS: ADJUSTED SILHOUETTE COEFFICIENT BASED ON MEDIAN

by

Fatih Emre ÖZTÜRK

July, 2023

İZMİR

DETERMINING THE OPTIMAL NUMBER OF CLUSTERS IN CLUSTERING ANALYSIS: ADJUSTED SILHOUETTE COEFFICIENT BASED ON MEDIAN

A Thesis Submitted to the
Graduate School of Natural And Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science in Statistics, Data Science Program

by

Fatih Emre ÖZTÜRK

July, 2023
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**DETERMINING THE OPTIMAL NUMBER OF CLUSTERS IN CLUSTERING ANALYSIS: ADJUSTED SILHOUETTE COEFFICIENT BASED ON MEDIAN**" completed by **Fatih Emre ÖZTÜRL** under supervision of **Prof.Dr. NESLIHAN DEMIREL** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....

Prof.Dr. NESLIHAN DEMIREL

Supervisor

.....

Doç.Dr. Engin YILDIZTEPE

Jury Member

.....

Prof. Dr. Gözde ULUTAGAY

Jury Member

Prof.Dr. OKAN FISTIKOGLU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

"*Improbable oft will hap to men.*"
- Aristotle

In 2019, when I was completing my philosophy degree, I never thought I would be writing an acknowledgment to such a thesis. As I conclude this improbable oft today, I would like to thank my dear advisor Prof.Dr. Neslihan Demirel, who made it very easy for me to reach this ending with her infinite patience and understanding. I would also like to express my gratitude to everyone at the Dokuz Eylül University's Department of Statistics especially to Assoc. Prof. Engin Yıldıztepe, whose perspective and approach I will always take as an example.

Finally, Tuğba Öztürk, the best sister in the world, who is my biggest supporter in this life and allows me to stand behind every decision I make, I am so glad I have you!

Fatih Emre OZTURK

DETERMINING THE OPTIMAL NUMBER OF CLUSTERS IN CLUSTERING ANALYSIS: ADJUSTED SILHOUETTE COEFFICIENT BASED ON MEDIAN

ABSTRACT

Clustering is an unsupervised learning that divides observations into groups based on their similarity. The most widely used clustering algorithm is k-means. However, in this clustering algorithm, the number of clusters needs to be determined in advanced. There are some methods to determine the optimal number of clusters. The most widely used methods fails in some specific situations. In this thesis, we propose Adjusted Silhouette Based Coefficient Based on Median as an alternative method to determine the optimal number of clusters. Success of the method is compared with Average Silhouette, Calinski-Harabasz, and Davies-Bouldin both on simulated and benchmark data sets. Methods were compared with cluster validation metrics (Success Rate, Affinity Rate, Rand Index, and Meila's Variation of Information). Adjusted Silhouette Coefficient Based on Median is the most successful method of all metrics and studies.

Keywords: Cluster Analysis, Average Silhouette, Calinski-Harabasz, Davies-Bouldin, Adjusted Silhouette Coefficient Based on Median

KÜMELEME ANALİZİNDE OPTİMAL KÜME SAYISININ BELİRLENMESİ: MEDYANA GÖRE DÜZELTİLMİŞ SİLÜET KATSAYISI

ÖZ

Kümeleme, gözlemleri benzerliklerine göre gruplara ayıran denetimsiz bir öğrenmedir. En yaygın kullanılan kümeleme algoritması k-ortalamalardır. Ancak bu kümeleme algoritmasında küme sayısının önceden belirlenmesi gerekmektedir. Optimal küme sayısını belirlemek için bazı yöntemler vardır. En yaygın kullanılan yöntemler bazı durumlarda başarısız olur. Bu tezde, optimum küme sayısını belirlemek için alternatif bir yöntem olarak Medyana Dayalı Düzeltilmiş Siluet Katsayı yöntemini önerilmiştir. Yöntemin başarısı, hem simüle edilmiş hem de gerçek veri setlerinde Ortalama Silüet, Calinski-Harabasz ve Davies-Bouldin yöntemleri ile karşılaştırılmıştır. Yöntemler, küme doğrulama metrikleri (Başarı Oranı, Yakınlık Oranı, Rand Endeksi ve Meila'nın Bilgi Varyasyonu) ile karşılaştırılmıştır. Medyana Dayalı Düzeltilmiş Silüet Katsayısı, tüm metrikler ve çalışmalarda en başarılı yöntem olarak saptanmıştır.

Anahtar kelimeler: Kümeleme Analizi, Average Silhouette, Calinski-Harabasz, Davies-Bouldin, Medyana Göre Düzeltilmiş Silüet Katsayısı

CONTENTS

Page

M.Sc THESIS EXAMINATION RESULT FORM ii
ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZ v
LIST OF FIGURES viii
LIST OF TABLES x

CHAPTER ONE – INTRODUCTION 1

CHAPTER TWO – BACKGROUNDS 5

2.1 Data Preparation for Clustering 5
2.1.1 Standardize 6
2.1.2 Principle Component Analysis 7
2.2 k-means Clustering Algorithm 8
2.3 Methods to Determine the Optimal Number of Clusters 10
2.3.1 Average Silhouette Method 10
2.3.1.1 Cluster Tightness 11
2.3.1.2 Cluster Separation 12
2.3.2 Calinski-Harabasz Method 13
2.3.2.1 Within Cluster Sum of Squares 14
2.3.2.2 Between Cluster Sum of Squares 15
2.3.3 Davies-Bouldin Method 16
2.3.3.1 Intra-Cluster Dispersion 17
2.3.3.2 Separation Criteria 17
2.3.3.3 Cluster Similarity and The Most Similar Clusters 18
2.4 Validation Metrics 19
2.4.1 Rand Index 20

2.4.2 Meila's Variation of Information 21

CHAPTER THREE – ADJUSTED SILHOUETTE COEFFICIENT BASED ON MEDIAN 23

3.1 Within Cluster Variance 24
3.2 Between Cluster Variance 25

CHAPTER FOUR – SIMULATION STUDY 28

4.1 Results of the Same Sized Data Sets 28
4.2 Results of the Different Sized Data Sets 34

CHAPTER FIVE – REAL DATA APPLICATIONS 39

5.1 Wine Data Set 39
5.2 Column3c Data Set 40
5.3 Ecoli Data Set 42
5.4 Iris Data Set 43
5.5 Haberman Data Set 45
5.6 Breast Cancer Wisconsin Data Set 46
5.7 Breast Tissue Data Set 48
5.8 Glass Data Set 50
5.9 Yeast Data Set 51

CHAPTER SIX – CONCLUSION 53

REFERENCES 56

LIST OF FIGURES

Page

Figure 2.1 Illustration of the cluster tightness calculation 12
Figure 2.2 Illustration of the cluster separation calculation 13
Figure 2.3 Illustration of the WCSS calculation 15
Figure 2.4 Illustration of the BCSS calculation 16
Figure 2.5 Illustration of separation criteria calculation 18
Figure 3.1 Illustration of within cluster calculation 25
Figure 3.2 Illustration of between cluster variance calculation's first step 26
Figure 3.3 Illustration of between cluster variance calculation 27
Figure 4.1 An example scatter plot of the data sets created for the first part of the simulation study..... 30
Figure 4.2 An example scatter plot of the data sets created for the data sets with the different cluster sized 36
Figure 5.1 Scatter plot of column3c data set..... 42
Figure 5.2 Scatter plot of iris data set..... 45
Figure 5.3 Scatter plot of breast cancer wisconsin data set 48
Figure 5.4 Breast Tissue data set 50

LIST OF TABLES

Page

Table 4.1 Information about cluster centers, cluster covariance matrices, and observation numbers for two clustered data sets. 31

Table 4.2 Average Silhouette method's results on the same cluster sized data sets with four clusters. 32

Table 4.3 Calinski-Harabasz method's results on the same cluster sized data sets with four clusters. 32

Table 4.4 Davies-Bouldin method's results on the same cluster sized data sets with four clusters. 33

Table 4.5 Adjusted Silhouette Coefficient Based on Median's results on the same cluster sized data sets with four clusters. 34

Table 4.6 Validation metrics of the first part 35

Table 4.7 Average Silhouette method's results on the different cluster sized data sets with four clusters. 37

Table 4.8 Calinski-Harabasz method's results on the different cluster sized data sets with four clusters. 37

Table 4.9 Davies-Bouldin method's results on the different cluster sized data sets with four clusters. 38

Table 4.10 Adjusted Silhouette Coefficient Based on Median's results on the different cluster sized data sets with four clusters. 39

Table 4.11 Validation metrics of the second part 39

Table 5.1 Coefficients of each method for wine data set 41

Table 5.2 Coefficients of each method for column3c data set 43

Table 5.3 Coefficients of each method for the ecoli data set 44

Table 5.4 Coefficients of the iris data set for each method 46

Table 5.5 Coefficients of each method for the haberman data set 47

Table 5.6 Coefficients of each method for the breast cancer wisconsin data set 49

Table 5.7 Coefficients of each method for the breast tissue data set 51

Table 5.8 Coefficients of each method for the glass data set 52

Table 5.9 Coefficients of the yeast data set for each method 53
Table 5.10 Validation metrics of the real data set study 53

CHAPTER ONE
INTRODUCTION

Computer systems are designed to perform functions related to human intelligence, such as perception, thinking, learning and problem solving. Artificial intelligence is a branch of science that tries to imitate that human intelligence (Benko, 2009). Machine learning is a sub-branch in the field of artificial intelligence that enables computer systems to learn on their own to perform a specific task by analyzing data (El Naqa & Murphy, 2015). This learning process is accomplished through the analysis of data sets and the use of statistical models.

Machine learning is broadly divided into three categories: Supervised Learning, Unsupervised Learning and Semi - Supervised Learning. Supervised learning uses a machine learning model fed with labeled data (Learned-Miller, 2014). Data set consists of two parts: features and labels. Using this data, the model learns to match features to a specific label and can then apply this knowledge on new data. Semi-Supervised Learning is a machine learning approach that includes both supervised learning and unsupervised learning techniques. This approach is used when a part of data set is labeled (supervised) and the rest is unlabeled (unsupervised). Semi-supervised learning is useful for many applications where labeled data is scarce or labeling process is difficult or costly (Hady, 2013). Unsupervised learning, unlike supervised learning, learns structure of features without any labels in the data. This method is used to identify patterns of data and reveal their hidden structure (Hinton, 1999). Unsupervised learning has been used in many fields including biology (Sokal, 1966), psychology (Holzinger, 1941), computer security (Barbará, 2002), image processing (Chou, 2004), and pattern recognition (Mirkin, 2005).

Since unsupervised learning is often used to identify structures or patterns in the data set, these structures are identified using dimensionality reduction and clustering techniques in particular. Clustering is used to group data points with similar characteristics in a data set (Arbelaitz, 2013). In this way, structures and patterns in

the data set become easier to understand.

In general, there are three main components of a clustering analysis (Jain & Dubes, 1979). One component is to define cluster similarity measurement. In the case of clustering, similarity is calculated by the distances of observations from each other. There are various methods used in distance calculation (Duda & Hart, 1973). In general, the most widely preferred distance metric is Euclidean distance (Berthold & Höppner, 2016). Another component is grouping observations based on their similarity which also can be called as clustering. Likewise, just as there are various distance metrics, there are also many clustering algorithms (Kassambara, 2017). One of the most widely used clustering algorithms is the k-means algorithm which is also used in this thesis (Marutho, 2018). This algorithm creates a predetermined number of clusters (k) and determines a center (centroid) for each cluster. Then, it performs the clustering process by assigning each observation to the nearest centroid (Hartigan, 1979). Whether the number of clusters needs to be predetermined depends on the clustering algorithm used and the characteristics of the data set. Some clustering algorithms like density-based clustering and model-based clustering do not require the number of clusters to be predetermined and automatically determine the natural number of clusters in the data set. Such algorithms use various methods to learn more about the structure and distribution of the data set. Other clustering algorithms, like k-means and k-medoids, require a predetermined number of clusters (Löster, 2016). Finally, the last component of the clustering analysis is to check cluster validation metrics to assess the success of the clustering.

Determining the number of clusters is important for conducting an accurate and effective cluster analysis. When the number of clusters is chosen incorrectly, some clustering algorithms may create too many clusters or too few clusters. When too many clusters are formed, data with similar characteristics may be located in different clusters and data analysis cannot be done correctly. When too few clusters are formed, data with different characteristics may be in the same cluster and data analysis may be misleading. Thus, determining the cluster number correctly enables a more accurate and effective cluster analysis which can lead to better decisions to be

made in businesses and industry.

However, determining the number of clusters is a difficult step, just as it is very important. First of all, a proper understanding of the features and structures in the data set is critical for an appropriate number of clusters. Determining the number of clusters without a full understanding of the structure and characteristics of the data set can lead to inaccurate results. Size of the data set is also an important challenge in determining the optimal number of clusters. Large data sets can make it difficult to determine the appropriate number of clusters. In high-dimensional data sets, determining the number of clusters can be more difficult because the structures and features in the data set can become more complex. Lastly, identifying differences between clusters is not always an easy topic. Determining the number of clusters requires an accurate understanding of the differences between clusters in the data set since without a full understanding of the differences between clusters can lead to incorrect conclusions. In this sense, variance is an important feature (Schneider & Xhafa, 2022).

The difference between the variances of clusters is important when clustering. Variance measures the variability in a data set and the differences between the variances of different clusters provide information about the homogeneity or heterogeneity of the clusters. The similarity of samples within a cluster is measured by the low variance of that cluster. Likewise, the degree of similarity between different clusters is measured by the differences between the variances of the clusters. That is, the clustering algorithm works to maximise similarity within a cluster and dissimilarity between different clusters (Schneider & Xhafa, 2022).

Considering the importance and difficulties of the number of clusters, many methods have been proposed, and many papers have been wrote to determine the number of clusters in order to make this stage easier and more successful. Generally, papers propose a method to determine the optimal number of clusters and compare their method with other methods. In some studies, researchers conducted extensive experiments and arrived ambiguous results (Hardy, 1996) (Dimitriadou et al., 2002) (Milligan & Cooper, 1985). Also, in some studies success of the methods are

compared on several data sets and their limitations was found (Öztürk & Demirel, 2023).

Methods for determining the number of clusters which are more than thirty (Kassambara, 2017) may vary depending on the characteristics of the data set and the clustering algorithm. Among some common methods for determining the number of clusters, Average Silhouette method (Rousseeuw, 1987) is most widely used method. In this thesis, we propose a modified version of average silhouette method based on median. In the remainder of this thesis which consists of four sections, we firstly explain theoretical background of the methods and algorithms used. Then, the proposed Adjusted Silhouette Based on Median is explained. In Chapter 4, simulation study is explained and its results are given. Finally, in Chapter 5, results of the study done in benchmark data sets are given.

CHAPTER TWO
BACKGROUNDS

This chapter contains theoretical information about the four different methods (Average Silhouette, Calinski-Harabasz, and Davies-Bouldin) used in this thesis. Firstly, the procedures performed in the data preparation part and the theoretical knowledge behind those procedures are explained. Then, a detailed information about how k-means works as a clustering algorithm is given. After explaining the Average Silhouette, Calinski-Harabasz, and Davies-Bouldin methods, which are used to compare the Adjusted Silhouette based on Median method, information about the metrics to be used for validation is given in the last part.

2.1 Data Preparation for Clustering

Data preparation phase entails collecting, storing, wrangling, pre-processing and, if necessary, splitting data sets into smaller data sets. These steps ensure accurate and reliable results (Savage & Huffman, 2018). Missing or incorrect data in the data set may result in inaccurate results. Therefore, it is extremely important to check the data sets correctly and make the necessary adjustments during the data preparation phase (Hruschka et al., 2003). If the data are not prepared properly, the similarities or differences between the data may not be analysed correctly.

The data preparation phase for clustering is crucial to obtain accurate results. There are many things that need to be done to prepare the data set for clustering algorithms. The most important ones are data scaling, dimension reduction and data cleaning. Impute any missing or incorrect data in the data set. Since missing or incorrect data can cause misleading results in cluster analysis, it is very important that there is no missing or incorrect data in the data set. In this thesis, the data cleaning steps were only applied on real data sets.

The next two subsections describe the other two methods applied in the data preparation part: standardization and principal component analysis.

2.1.1 Standardize

Standardization is the transformation of a variable into a standard form by subtracting its population mean and dividing by its population standard deviation (Milligan & Cooper, 1988). Yet, it can be done both when population standard deviation is known, and not known. This standardization facilitates the comparison of variables with different units of scale and also makes it possible to more accurately analyze the relationship between variables.

As mentioned earlier, clustering algorithms allow data points to be divided into groups using similarity measures. Standardization is important for clustering because clustering algorithms use similarity measures, especially based on distance measures. Therefore, converting data points into a standardized form can prevent erroneous clustering results that can be caused by scale differences. To standardize observations, z-score is calculated and it is calculated as follows:

$$z_x = \frac{x - mean(x)}{sd(x)} \tag{2.1}$$

Standardization is important for clustering analysis because when variables in the data set have different scales, the clustering results can be affected. For example, if the scale of a variable is much larger than other variables, the clustering algorithm may weight this variable more and make other variables less effective in clustering. Therefore, it is important that all variables in the data set have the same scale for the clustering analysis to give accurate results (Mohamad & Usman, 2013).

Standardization is particularly important for clustering algorithms that work with distance measures such as the Euclidean distance (Milligan & Cooper, 1988). As a result, the scales of variables in the data set becomes the same in clustering analysis, ensuring that clustering results are accurate and consistent.

2.1.2 Principle Component Analysis

Principal Component Analysis (PCA) is a statistical technique that examines the relationships between variables in a data set and uses these relationships to reduce the data set to a smaller number of new variables. These new variables are called principal components and are used to explain a large proportion of the variance of the original variables.

Before performing PCA, each variable in the data set may have a different scale. Therefore, standardization of the data is necessary to remove the scale differences of the variables. Therefore, the first step is to standardize the data set (Abdi & Williams, 2010). Then the covariance matrix is calculated. Covariance is a statistical measure of the relationship between two variables. The covariance matrix contains the covariances of all variables in the data set with each other. The covariance matrix helps to understand how the variables in the data set are related to each other.

After calculating the covariance matrix, the eigenvalues and eigenvectors in the covariance matrix are calculated. The eigenvalues and eigenvectors of the covariance matrix are one of the most important steps of the PCA process (Ringnér, 2008). Eigenvalues are numbers that indicate how important each eigenvector is. Eigenvalues measure the strength of the relationships between variables in the data set. Eigenvectors are vectors that correspond to each eigenvalue and determine the directions of the new variables that will be created during the PCA process.

After the eigenvalues and eigenvectors are calculated, they are ordered from largest to smallest (Bro & Smilde, 2014). Ranking the eigenvalues from largest to smallest helps us to rank the importance of the variables in the data set according to the strength of their relationships with each other. The largest eigenvalue represents the most important variables. Then, using the largest eigenvalue and its corresponding eigenvectors, a transformation matrix is created to represent the variables in the original data set in a new space. This transformation matrix contains eigenvectors ordered by the order of magnitude of the eigenvalues. The transformation matrix

creates a vector space that can be transformed into new components of the original variables in the data set. The original data set is transformed into principal component values using the newly created transformation matrix.

PC values are calculated as the product of the values of each principal component for each observation in the original data set. To determine how much of the variance of the variables in the original data set each principal component explains, the sum of the eigenvalues is calculated. Each eigenvalue indicates how much of the variance of the variables in the data set is explained.

This completes the calculation of PCA. The resulting principal components are the original data set reduced to a smaller number of variables. These new variables help us to more easily understand the relationships between variables in the original data set.

Dimensionality reduction aims to reduce the size of data sets, remove unnecessary features, divide data sets into more meaningful and manageable parts, and reduce the complexity of the data set (Lee & Verleysen, 2009). When working with large data sets, data sets may contain a large number of variables. In this case, using every single feature in the data set in cluster analysis can be difficult to achieve efficient results (Ben-Hur & Guyon, 2003). Furthermore, high-dimensional data sets often cause the data to become more complex and thus can reduce the accuracy of cluster analysis (Ding & He, 2004). This thesis aims to avoid these problems by applying Principle Component Analysis to data sets with highly correlated variables.

2.2 k-means Clustering Algorithm

k-means algorithm is one of the most widely used algorithms for clustering data in a data set according to a given number of clusters. It firstly introduced in 1967 by James Macqueen (MacQueen et al., 1967). However, there are many variants. The most widely used algorithm with the best results is the Hartigan-Wong algorithm (Almanza-Ortega et al., 2022). In this thesis, k-means algorithm was used in all clustering calculations.

To group data, k-means measures how observations are similar to each other and assigns similar observations to the same cluster (Nazeer & Sebastian, 2009). The k-means algorithm operates on cluster centers (centroids) pre-selected by the user, and data is assigned to the closest centroid. Each centroid is then recalculated by the mean of the observations assigned to that cluster. This process ends until the centroids remain unchanged or after the algorithm has been repeated a specified number of times. As a result, each data point is assigned to a clustering class and the data set is divided into clusters according to a given number of clusters.

Unlike the k-means algorithm, the Hartigan-Wong algorithm takes into account the diversity among the data in each cluster and provides a more efficient clustering (Hartigan, 1979). The Hartigan-Wong algorithm is very popular in clustering problems and gives very fast and efficient results, especially for large data sets.

As mentioned earlier, the k-means algorithm continues until the centroids do not change. In which case the centroids will not change is an important point. Basically, the k-means algorithm works to minimize the total within cluster sum of squares. Therefore, the iteration continues until the total within cluster sum of squares finds its minimum value. Equation 2.2 explains how the total within cluster sum of squares we are trying to minimize is calculated:

$$tot.withinss = \sum_{k=1}^{K} \sum_{x_i \epsilon c_k} (x_i - \mu_k)^2 \tag{2.2}$$

where,

k is the cluster number,

x_i is the observation,

c_k is the cluster that observation x_i is assigned to,

μ_k is the mean of c_k. There are many reasons why the k-means algorithm is preferred over other algorithms. These advantages include the algorithm's simplicity and speed, scalability, compatibility with high-dimensional data, flexibility and wide range of applications (Nazeer & Sebastian, 2009). The k-means algorithm is known as a

simple and fast algorithm (Tang et al., 2017). Therefore, it takes a few steps to sort data into clusters.

However, the algorithm has also some disadvantages. First of all, the results of the algorithm depend on the number of clusters being chosen in advance (Kassambara, 2017) (Löster, 2016). Therefore, if the number of clusters is initially chosen incorrectly, the results may be incorrect. Also, when determining the number of data clusters, in the absence of expert knowledge or prior knowledge, it can be difficult to determine the correct number. Furthermore, the algorithm is based on the assumption that all data have the same variance value. This assumption is rarely valid in real life and the performance of the algorithm may be poor for data with different variances. Finally, the algorithm's solution may not guarantee global optimality and the results may vary depending on the starting point. Therefore, the k-means algorithm may not give accurate results in some cases and alternative methods may need to be used.

2.3 Methods to Determine the Optimal Number of Clusters

Determining the number of clusters is one of the most important decision processes in data clustering problems. In this process, it is decided how many different clusters to group the data into. The right number of clusters provides a better understanding of the data and more accurate results. On the contrary, identifying the wrong number of clusters can lead to misinterpretation of results and wrong decisions. There are more than thirty methods for determining the number of clusters. Since Average Silhouette, Calinski-Harabasz and Davies-Bouldin methods are used to compare the method developed in this thesis, these methods will be explained in detail in the following sections.

2.3.1 Average Silhouette Method

Average Silhouette score(AS) is the average of the Silhouette scores calculated for each observation. Silhouette is based on calculating a score that measures how well

each observation stands in its own cluster and how well it is related to other clusters (Rousseeuw, 1987). It is calculated in two steps, which will be explained in the next subsections. The resulting score takes a value between -1 and 1, with values closer to 1 indicating better clustering results.

Average Silhouette method applies the k-means algorithm for different cluster numbers to calculate Silhouette scores for each cluster number. Then, by averaging the Silhouette scores, the number of clusters with the highest score is determined as the optimal number of clusters.

2.3.1.1 Cluster Tightness

The first step in calculating the Silhouette score is the calculation of cluster tightness. Cluster tightness measures the similarity of an observation to other observations in its cluster. For each observation, it is calculated as the average of the distances of that observation from the other observations in its cluster. Equation 2.3 explains how cluster tightness is calculated. For a concrete illustration, please see Figure 2.1.

$$a_i = \frac{1}{n_k - 1} \sum_{x_i, x_i' \epsilon c_k} d(x_i, x_i') \tag{2.3}$$

where,

c_k is the cluster that observation x_i is assigned to,

n_k is the total observation number of c_k,

x_i' are all observations other than x_i in c_k,

$d(x_i, x_i')$ Euclidean distance between x_i and x_i'.

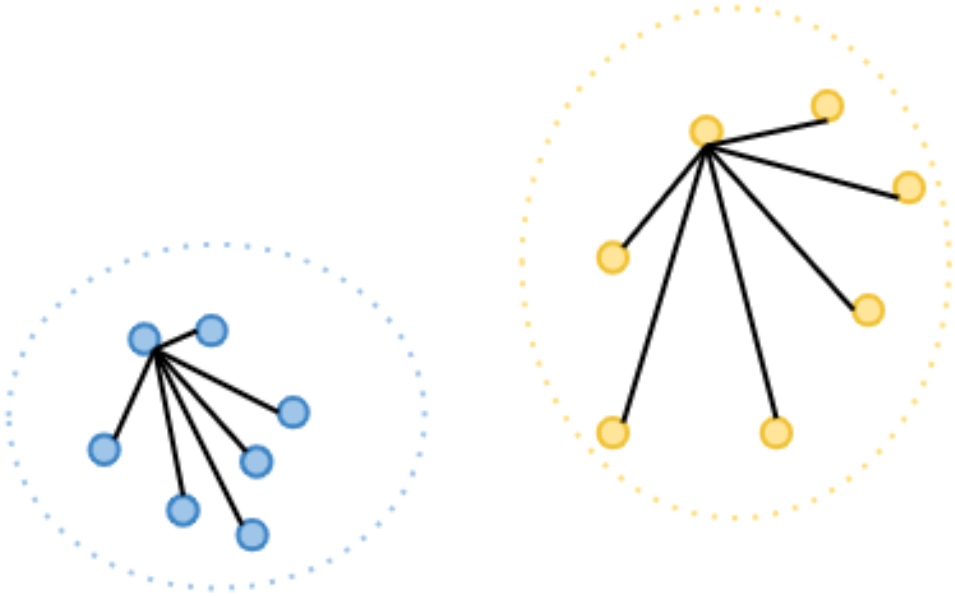


Figure 2.1 Illustration of the cluster tightness calculation

2.3.1.2 Cluster Separation

The second step in calculating the Silhouette score is the calculation of cluster separation. Cluster separation measures the similarity of an observation to other clusters that it does not belong to. For each observation, it is calculated as the minimum of the distances of that observation from other observations in other clusters. Equation 2.4 explains how cluster separation is calculated. For a concrete illustration, please see Figure 2.2.

$$b_i = min\left(d(x_i, x_i')\right) \tag{2.4}$$

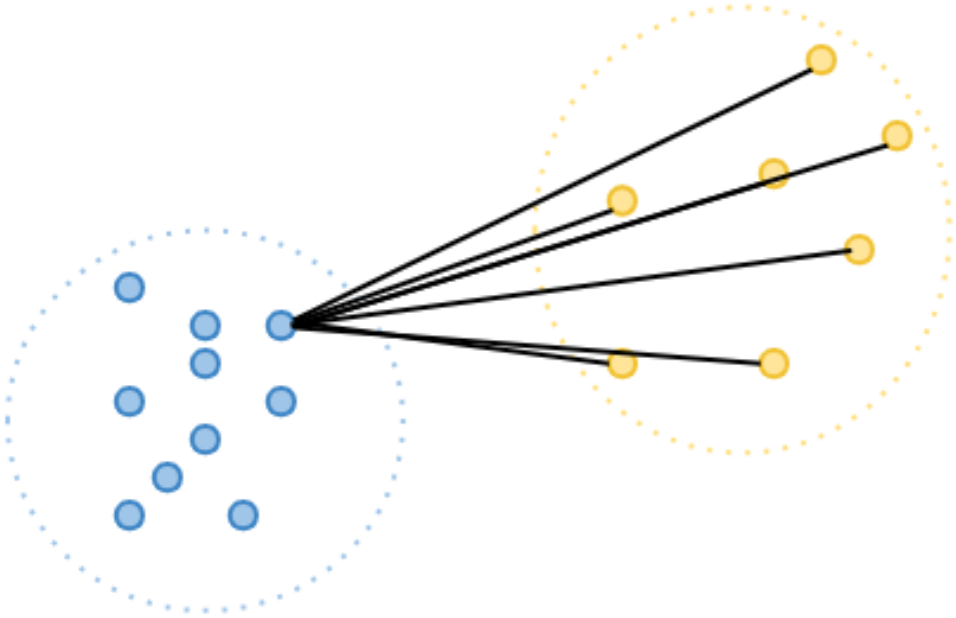


Figure 2.2 Illustration of the cluster separation calculation

After calculating cluster tightness and cluster separation, the silhouette score is calculated for each observation using Equation 2.5.

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{2.5}$$

2.3.2 Calinski-Harabasz Method

Calinski-Harabasz (CH) method is a method used to determine the optimal number of clusters in a data set. This method uses a performance metric to measure clustering performance and determines the optimal number of clusters by comparing performance metrics corresponding to different numbers of clusters.

CH method considers how homogeneous each cluster is and how large the differences between clusters are (Caliński & Harabasz, 1974). As the number of clusters increases, homogeneity increases and differences between clusters decrease. The CH measure is based on these measures of homogeneity and dissimilarity.

To use the CH measure to determine the number of clusters, first create cluster models for different numbers of clusters. Then the CH measure is calculated for each model. Since the CH measure is based on the ratio of cluster homogeneity and cluster dissimilarity, a large CH measure indicates a clear distinction between homogeneous clustering and dissimilar clustering. Therefore, when the CH measure is calculated for different numbers of clusters, the number of clusters with the highest value is selected. This is considered the optimal number of clusters for the data set.

In order to calculate the CH score, Between Cluster Sum of Squares and Within Cluster Sum of Squares metrics should be calculated. The next subsection explains how these two metrics are calculated.

2.3.2.1 Within Cluster Sum of Squares

Within-cluster sum of squares (WCSS) is used to measure within-cluster homogeneity in the CH calculation. WCSS is the sum of the squares of the distances of all samples in a cluster from the cluster center. To calculate the WCSS, the center of each cluster is first calculated. The distance of each observation from the center of its cluster is calculated and squared. These distances are then summed. Since the WCSS value measures how homogeneous the data within the cluster is, a higher WCSS value indicates heterogeneity. Equation 2.6 explains how the WCSS is calculated for each cluster, while Equation 2.7 explains how the final value of the WCSS for CH is calculated. For a concrete illustration, please see Figure 2.3.

$$WCSS_k = \sum_{i=1}^{n_k} (x_i - c_{ck})^2 \tag{2.6}$$

$$WCSS = \sum_{k=1}^{K} WCSS_k \tag{2.7}$$

where,

$WCSS_k$ is within cluster sum of squares of the cluster k,

n_k is the total observation number of cluster k,

x_i is ith observation of cluster k

c_{ck} is the centroid of cluster k

K is the total number of clusters.

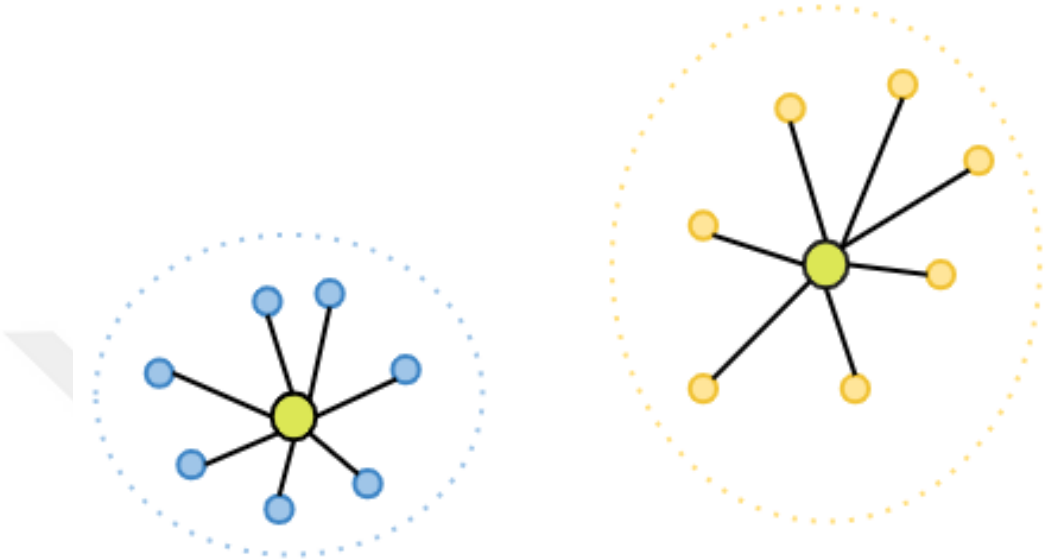


Figure 2.3 Illustration of the WCSS calculation

2.3.2.2 *Between Cluster Sum of Squares*

Between cluster sum of squares (BCSS) is a metric used to measure the difference between clusters in the CH calculation. BCSS is the sum of squares of the distances between the centroids of each cluster and the barycenter of the data set. For BCSS calculation, the barycenter, the center point of all observations, is first calculated. Then the centroid of each cluster is calculated. The sum of the squares of the distances between the centroids of the clusters and the barycenter gives BCSS. It decreases as the number of clusters increases, which may indicate the presence of heterogeneous clusters. Equation 2.8 Explain how the BCSS is calculated. For a concrete illustration, please see Figure 2.4.

$$BCSS = \sum_{k=1}^{K} n_k \times (c_{ck} - C)^2 \tag{2.8}$$

where,

C is barycenter of the data set,

n_k is the total observation number of cluster k,

c_{ck} is the centroid of cluster k

K is the total number of clusters.

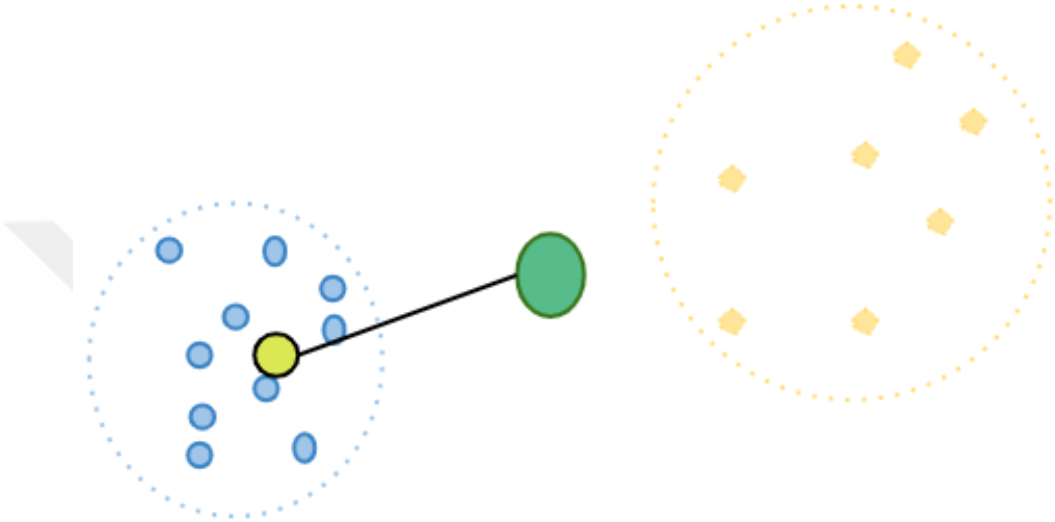


Figure 2.4 Illustration of the BCSS calculation

2.3.3 Davies-Bouldin Method

Davies-Bouldin method (DB) is an evaluation metric for clustering algorithms and is used to measure the quality of clustering results. This index is calculated by dividing the within-cluster similarities by the between-cluster differences (Davies & Bouldin, 1979). An index with a lower score indicates more homogeneous clustering results and a better clustering structure.

In order for DB to be used to determine the number of clusters, the index is calculated for different cluster numbers and a DB score is obtained for each cluster number. Then, the cluster number with the lowest DB score is selected as the best cluster number.

2.3.3.1 Intra-Cluster Dispersion

DB index is calculated by computing the Intra-Cluster Dispersion for each cluster. Intra-Cluster Dispersion is a measure of the distance of observations within a cluster from the cluster centroid and indicates how homogeneous the cluster is. It is usually calculated as sum of squared errors. This calculation method squares the distance of each point from the cluster center and calculates the sum of all these squares. Equation 2.9 explains how the Intra-Cluster Dispersion is calculated. For a concrete illustration, please see Figure 2.3.

$$S_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_i - c_{ck})^2 \tag{2.9}$$

2.3.3.2 Separation Criteria

Another important measure used in the calculation of DB is the separation criteria used to determine cluster separation. This criterion measures the similarity between different clusters and determines how well cluster separation is achieved. Separation criteria is calculated based on the difference in the similarities of two clusters. To measure this, the distance between the cluster centroids is measured. The larger value means the greater the difference between the two clusters and the better the clustering result. Equation 2.10 explains how the Separation Criteria is calculated. For a concrete illustration, please see Figure 2.5.

$$M_{kk'} = (c_{ck} - c_{ck'})^2 \tag{2.10}$$

where,

c_{ck} is the centroid of cluster, k

$c_{ck'}$ is the centroid of cluster k'

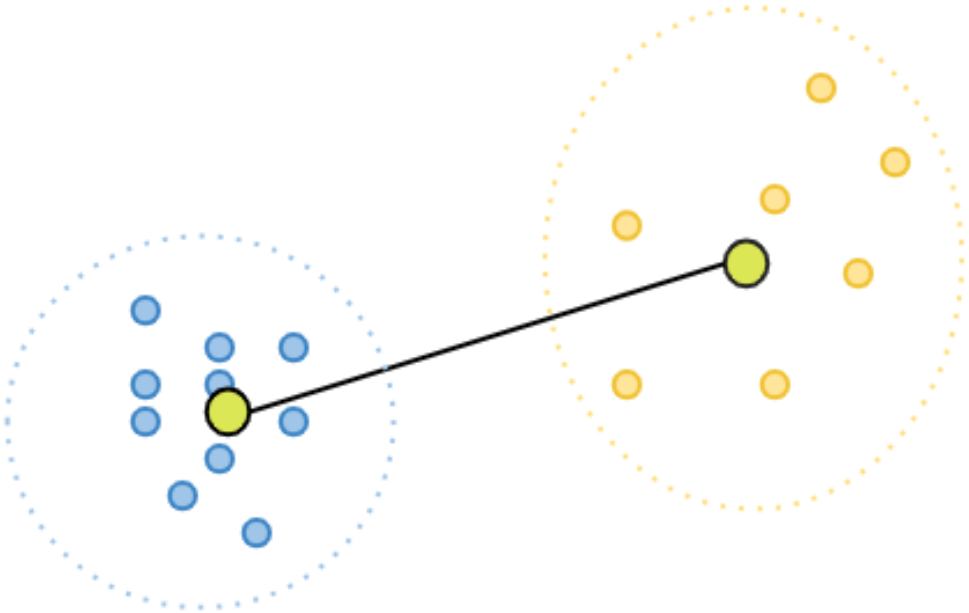


Figure 2.5 Illustration of separation criteria calculation

Since the separation criterion is used to measure how different each cluster is from the other clusters, it needs to be calculated separately for each cluster. Let k to be three. The first cluster is A, the second is B and the third is C. The separation criterion for cluster A is the distance from the centroid of cluster A to the centroid of cluster B and cluster C.

2.3.3.3 Cluster Similarity and The Most Similar Clusters

The cluster similarity concept used in DB index is a value that measures how similar two clusters are. This similarity is calculated by considering the distance between cluster centers and the similarity of points within the cluster to each other. The similarity of two clusters is calculated as the sum of the Intra-Cluster Dispersion of the two clusters divided by the Separation Criteria of the two clusters. The most similar clusters are found by determining the maximum of cluster similarity values for each cluster. Equation 2.11 and 2.12 explain how the Cluster Similarity and the most similar clusters are calculated.

$$R_{kk'} = \frac{S_k + S_{k'}}{M_{kk'}} \tag{2.11}$$

where,

S_k is Intra-Cluster Dispersion of c_k,

$S_{k'}$ is Intra-Cluster Dispersion of not c_k,

$M_{kk'}$ is Separation criteria of c_k and not c_k.

$$R_k = max(R_{kk'}) \tag{2.12}$$

For instance, to find the most similar cluster to cluster A, first the Intra Cluster Dispersion values of cluster A and cluster B are summed and divided by the separation criteria values of cluster A and cluster B. Then the same process is performed for cluster A and cluster B. Whichever of the two values obtained is larger is selected as the cluster that is most similar to cluster A. This is performed for each cluster.

2.4 Validation Metrics

Validation metrics are numerical metrics that measure the quality of clustering results. When making decisions such as which clustering algorithm and the number of clusters to use in clustering, various validation metrics can be used to make the right decisions and evaluate the clustering results. These metrics help to measure the accuracy, performance and success of clustering algorithms.

Validation metrics are divided into internal validation and external validation metrics. The methods described in the previous step, which are used to determine the number of clusters in this thesis, can all be used as internal validation metrics

(Rendón et al., 2011).

In order to check the success of the methods to determine the optimal number of clusters, three criteria were examined in this thesis. The first one is the Success Rate, which is the rate at which methods to determine the optimal number of clusters successfully determine the correct number of clusters. For the calculation of the second and third metrics, Rand Index and Meila's Variation of Information, the cluster assignments obtained as a result of k-means clustering with the number of clusters suggested by each method were used.

In this section, Rand Index and Meila's Variation of Information metrics, which are external validation metrics to test the success of the methods used to determine the number of clusters, are explained.

2.4.1 Rand Index

Rand Index (RI) is a metric used to evaluate the quality of a clustering algorithm. It measures the similarity between data set labels and clustering results. It was introduced by William M. Rand in 1971 and is calculated as the ratio of the number of correctly clustered observations to the total number of observations (Warrens & van der Hoef, 2020).

RI is also used to measure the similarity between different clustering techniques applied to the same data set. To calculate it, confusion matrix is first defined. The confusion matrix, which has four entries (a, b, c and d), determines the number of correctly and incorrectly clustered observations by comparing clustering labels with the true labels. a is the number of pairs of observations that are in the same cluster in both the true label and clustering label. b is the number of pairs of observations that are in different clusters in both the true label and clustering label. c is the number of pairs of observations that are in the same cluster in the clustering label but are in different clusters in the true label. Finally, d is the number of pairs of observations that are in different clusters in clustering label but are in the same cluster in the true labels.

Equation 2.13 explains how the RI is calculated.

$$RI = \frac{a+d}{a+b+c+d} \tag{2.13}$$

Rand Index values range from 0 to 1, with 1 indicating that the data set is perfectly clustered and 0 indicating that it is completely mis-clustered.

2.4.2 Meila's Variation of Information

Meila's Variation of Information (MVI) is a validation metric that measures how close or far the clustering result is from the true labels.

MVI is calculated in three steps (Meilă, 2007). As the first step, Mutual Information (MI), is a measure of the amount of information between the clustering result X and the data set label Y, is calculated. Mutual Information is used to measure the dependency between X and Y and calculates the amount of common information between X and Y. That is, as a given level of dependency between X and Y increases, the value of Mutual Information increases. Equation 2.14 explains how the Mutual Information is calculated.

$$MI = \sum_{k=1}^{n_k} \sum_{k'=1}^{n_{k'}} p(k,k') log \frac{p(k,k')}{(p(k)p'(k'))} \tag{2.14}$$

The second, Entropy(H), is a measure of uncertainty calculated as the negative logarithm of the probabilities of all possible states in a probability distribution - or all possible outcomes in a probability distribution. Equation 2.15 explains how the Entropy is calculated.

$$H = -\sum_{k=1}^{n_k} p(k)log(p(k)) \tag{2.15}$$

where,

$p(k)$ and $p'(k')$ are random variables associated with the clustering c_k and $c_{k'}$,

$p(k,k')$ is the probability that an observation belong to c_k in the clustering result and $c_{k'}$ in the label.

As the last step, Equation 2.16 explains how MVI is calculated:

$$MVI = H(k) + H(k') - 2(MI) \tag{2.16}$$

where,

$H(k)$ and $H(k')$ are the entropies of the clustering result and the data set label,

MI is the mutual information between clustering result and the data set label.

MVI takes values between 0 and infinity. A value of 0 means that the clustering results are perfectly matched, while an infinite value means that the clustering results are not matched at all.

CHAPTER THREE
ADJUSTED SILHOUETTE COEFFICIENT BASED ON MEDIAN

In this section we describe in detail the idea behind the Adjusted Silhouette Coefficient based on median (ASCM) method, which is proposed as an alternative to the methods used to determine the optimal number of clusters described theoretically in Chapter 2.

ASCM, as the name suggests, is basically a modification of Average Silhouette method. Although it is one of the most widely used methods among the existing methods, Average Silhouette Method does not respond well to specific data sets in some situations. Considering such cases, some modifications were considered to make the Average Silhouette method more robust.

Two criteria were taken into account when modifying the Average Silhouette method. The first one was to transform the Average Silhouette Method, which performs observation-based calculation, into cluster-based calculation. As explained before, in the Average Silhouette method, the Silhouette coefficient measures how well an observation fits into a cluster. This calculation takes into account, for each observation, both its similarity to its own cluster and its similarity to other clusters. These properties become even clearer when cluster-based calculation is used. Cluster-based calculation is a more specific and descriptive measure, as only the similarity of each observation to its own cluster is calculated. For example, while an observation may be very similar to other observations in its cluster, it may also receive similarity scores from observations in other clusters. In this case, a coefficient calculated by cluster gives a clearer indication of how well this observation actually fits into its cluster only.

The second modification to the Average Silhouette method is the use of a calculation based on the median. Since the coefficient calculation is based on clusters instead of observations, clusters need to be represented. According to the classical silhouette approach, representation by averaging may seem logical. However, in the case of ASCM, the median is preferred to represent the clusters. The reason behind

this is to reduce the effect of outliers in the data set and make the method more robust. Calculating according to the median of the clusters is also intended to make the calculation process faster and more applicable to larger data sets.

Two steps are required to calculate the ASCM. The first step measures the homogeneity of clusters, while the second step focuses on the separation of clusters. These two steps are explained in below.

3.1 Within Cluster Variance

The first step in the calculation of the ASCM method is the calculation of Within Cluster Variance (WCV). Simply put, to calculate WCV, the median of the cluster is first calculated. This median value is considered as the centroid of the cluster. Then, the distances of all observations in the cluster to this cluster centroid are calculated. Finally, these distances are averaged to obtain the WCV value. Equation 3.1 explain how WCV is calculated. For a concrete illustration, please see Figure 3.1.

$$WCV = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} (x_i - c_{ck})^2 \tag{3.1}$$

where,

n_k is the total observation number of c_k,

K is the total cluster number,

x_i are all observations in c_k,

c_{ck} is the cluster centroid that observation x_i is assigned to.

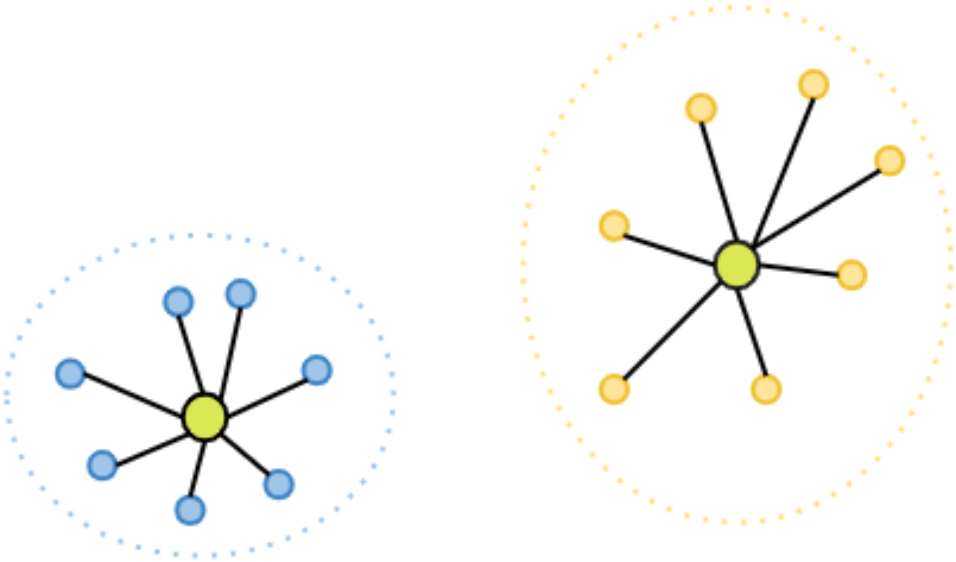


Figure 3.1 Illustration of within cluster calculation

Each observation differs from each other by its distance from the cluster centroid. Therefore, the average distance of the observations within the cluster to the centroid gives information about the homogeneity of the cluster. If the WCV takes a large value, it means that the observations within the cluster are farther from the centroid. This indicates that the homogeneity of the cluster is low and the clustering algorithm needs to be improved. On the contrary, if WCV takes a small value, it means that the homogeneity of the cluster is high and gives a better clustering result.

3.2 Between Cluster Variance

The second step in calculating the ASCM method is to calculate the Between Cluster Variance (BCV). Simply put, BCV is the average distance of a cluster to its nearest cluster. This measurement assesses how similar each cluster is to its closest neighbor. The larger the BCV value, the more different the clustering result will be. This indicates that the clustering result needs to be improved to reduce its unreliability.

BCV is calculated in two steps for each cluster. In the first step, the Euclidean

distance between each cluster center and the centers of other clusters is calculated. The centroids of the clusters are medians of the cluster. By calculating BCV, the most similar cluster for each cluster is calculated. For a concrete illustration of the first step, please see Figure 3.2.

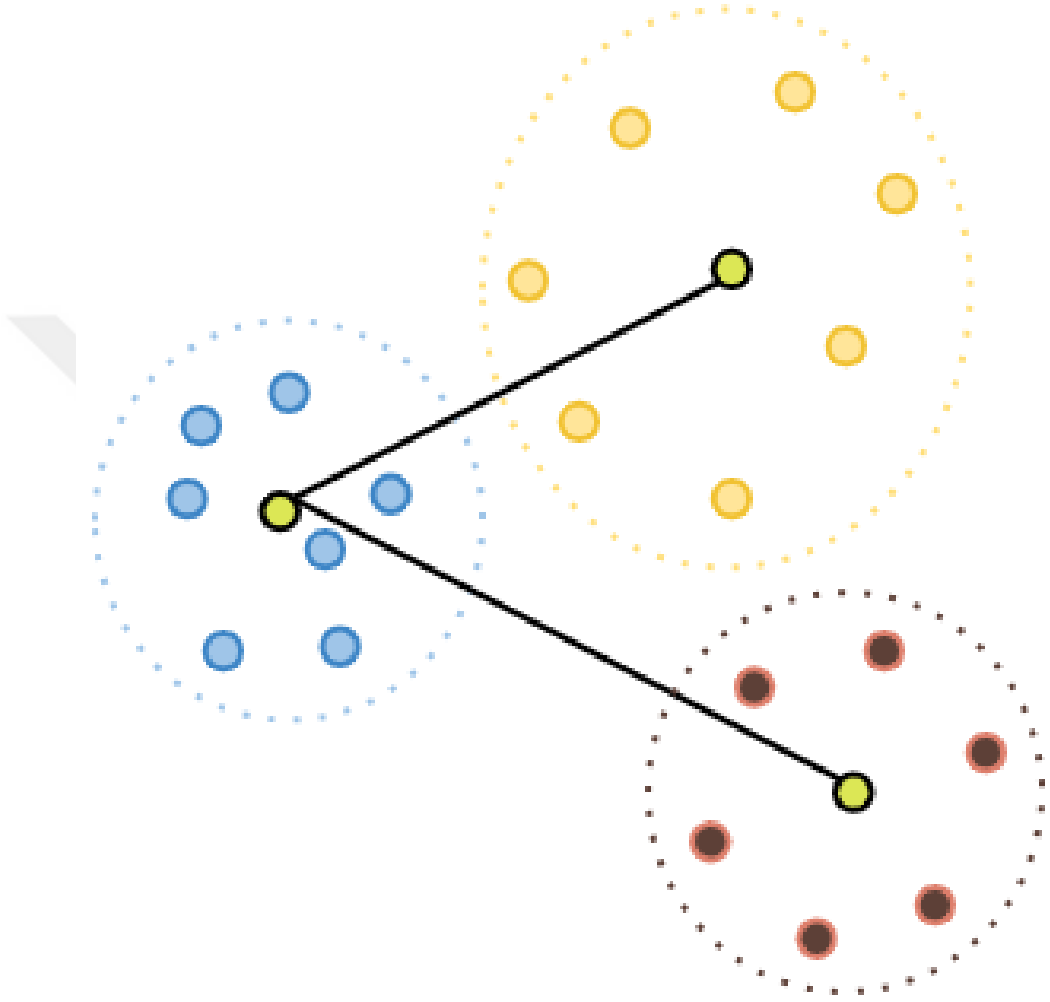


Figure 3.2 Illustration of between cluster variance calculation's first step

To calculate BCV, the second step for each cluster is to find the average Euclidean distance of all observations in the cluster to the centroid of the nearest cluster which we find in the first step. Again, the centroids of the clusters are medians of the cluster. Equation 3.2 explains how BCV is calculated. For a concrete illustration of how shared variance is calculated, please see Figure 3.3.

$$BCV = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}(x_i - c_{clk})^2 \tag{3.2}$$

where,

K is the total cluster number,

n_k is the total observation number of c_k,

x_i are all observations in c_k,

c_{clk} is the centroid of the cluster closest to the centroid of the cluster which observation x_i is assigned to.

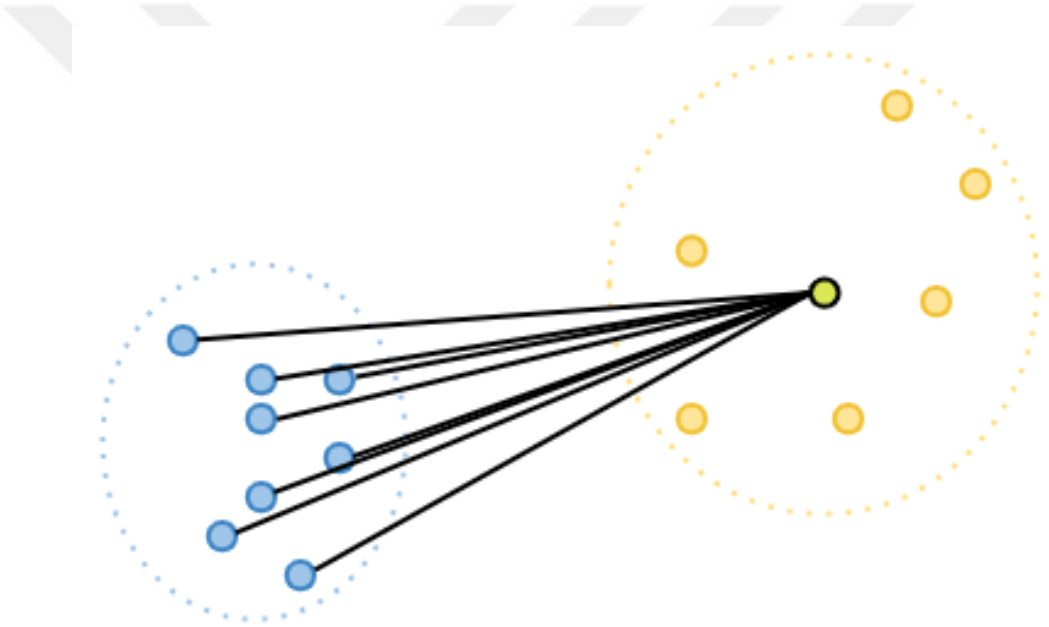


Figure 3.3 Illustration of between cluster variance calculation

In the last stage, the ASCM coefficient for the clustering result is obtained with Equation 3.3.

$$ASCM = \frac{(BCV - WCV)}{max(WCV, BCV)} \tag{3.3}$$

At this point, it needs to be noted that similar to the Average Silhouette method, higher ASCM coefficient means the data set is clustered well.

CHAPTER FOUR
SIMULATION STUDY

This section provides details on the simulation studies conducted to measure the success of ASM method. All applications were carried out using R 4.0.3 with the following packages: factoextra (Kassambara & Mundt, 2017), fpc (Hennig & Imports, 2015), and (Walesiak et al., 2011), fossil (Vavrek, 2011), and (Mi et al., 2009).

The simulation study conducted within the scope of this thesis is divided into two parts. For each part 42, in total 84 bi-variate normally distributed data sets which in the range of two and eight cluster numbers were created. In the first part, the number of observations of each simulated cluster is equal (50), while in the second part it is different. Information about observation numbers for each part can be seen in Table 4.1. Data sets created in both parts were again created with two different cases. In the first case, the Euclidean distance between the centers of the clusters is quite large, while in the second case the cluster centers are close to each other. Three different scenarios were created for both scenarios (near cluster centers and far cluster centers). In these scenarios, the variance of each cluster was simulated as equal and small, equal and large, and finally different and large. All of the simulated data sets, codes, and distribution plots are avaliable in the following Github repository[https://github.com/ozturkfemre/masters-thesis].

In the remainder of this section, the results of the studies conducted with data sets consisting of clusters with equal number of observations and clusters with different number of observations are interpreted respectively.

4.1 Results of the Same Sized Data Sets

A total of 42 data sets were created in this study, consisting of data sets in which the number of observations of each cluster was simulated equally and repeated 1000 times. The plots of the distributions of all data sets by cluster are available on Github repository. However, as an example, information about centers, covariance matrixes

(CM), and observation numbers (OBS for same sized clusters, OBD for different sized clusters) of the clusters can be found in Table4.1. Just like the plots, only tables containing the method coefficients of the data sets simulated as four clusters are included in the thesis. Other tables can also be found on Github profile.

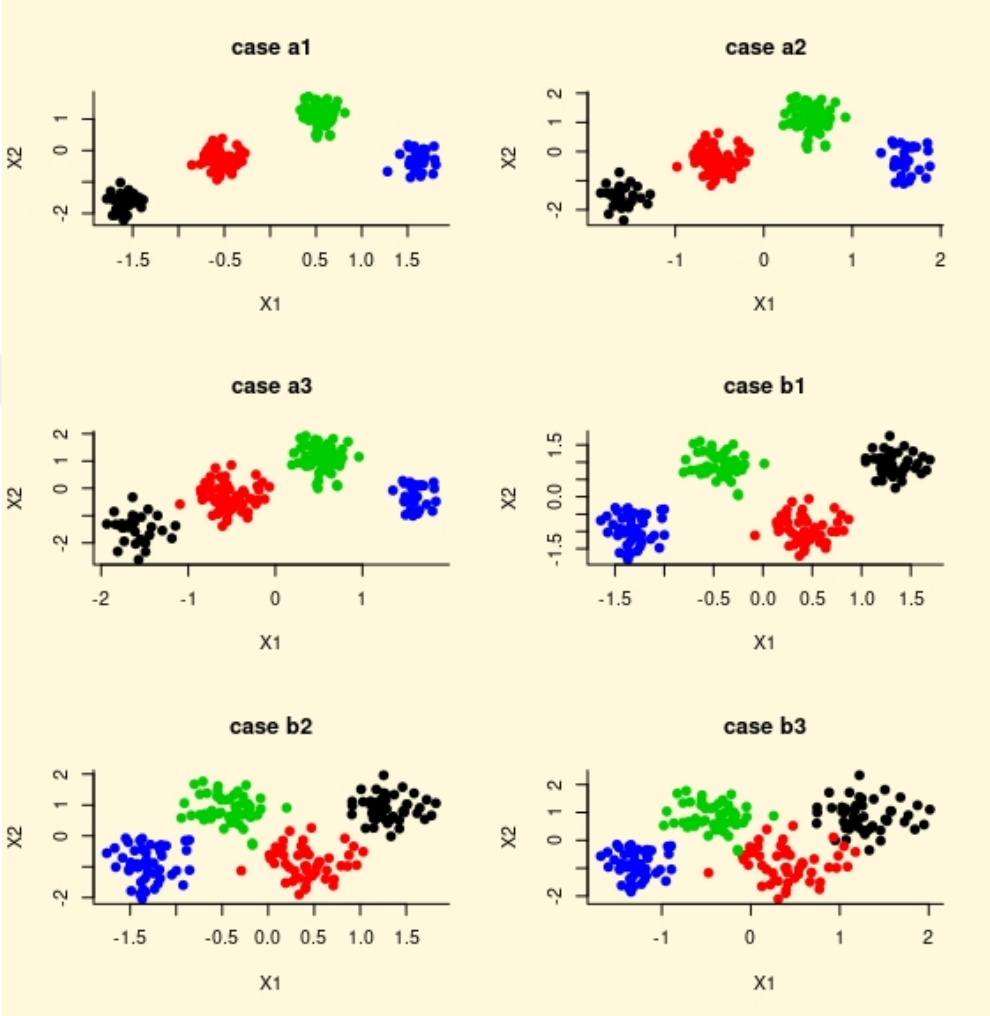


Figure 4.1 An example scatter plot of the data sets created for the first part of the simulation study

Table 4.1 Information about cluster centers, cluster covariance matrices, and observation numbers for two clustered data sets.

Cases	**Centers**	**CM of Cluster 1**	**CM of Cluster 2**	**OBS***	**OBD****
A1	(50,50) , (0,0)	(25,0,0,25)	(25,0,0,25)	(50, 50)	(30,70)
A2	(50,50) , (0,0)	(50,0,0,50)	(50,0,0,50)	(50, 50)	(30,70)
A3	(50,50) , (0,0)	(50,0,0,50)	(75,0,0,75)	(50, 50)	(30,70)
B1	(50,50), (25,25)	(25,0,0,25)	(25,0,0,25)	(50, 50)	(30,70)
B2	(50,50), (25,25)	(50,0,0,50)	(50,0,0,50)	(50, 50)	(30,70)
B3	(50,50), (25,25)	(50,0,0,50)	(75,0,0,75)	(50, 50)	(30,70)

In order to test and compare the success of the methods, each data set simulated in this part was clustered with the k-means clustering algorithm in the range of 2 to 10 clusters. For each clustering, the coefficients of the methods were calculated. For Average Silhouette, the number of clusters with the maximum coefficient in the coefficient calculation from 2 to 10 clusters was considered as the recommended number of clusters. This was again the maximum coefficient for AS, CH and ASCM, while the minimum coefficient value for DB was considered as the determined number of clusters.

AS was able to successfully determine the number of clusters in 18 of the 42 data sets in the first part of the simulation study. While it was able to correctly determine the number of clusters in all of the data sets with 2 and 3 clusters, it was noticed that its success gradually decreased, as the number of clusters of the data set increased. Precisely, it could not successfully determine the correct cluster number for the data sets with five, six, seven, and eight clusters. An example of AS method's results on the data sets with four clusters can be seen in Table 4.2.

Table 4.2 Average Silhouette method's results on the same cluster sized data sets with four clusters.

k	A1	A2	A3	B1	B2	B3
2	0.52	0.61	0.68	0.89	0.98	0.97
3	0.58	0.62	0.61	0.63	0.72	0.80
4	0.73	0.80	0.81	0.39	0.55	0.59
5	0.41	0.53	0.61	0.57	0.63	0.73
6	0.58	0.67	0.68	0.70	0.71	0.72
7	0.71	0.70	0.73	0.75	0.82	0.98
8	0.73	0.75	0.79	0.79	0.84	0.78
9	0.69	0.75	0.75	0.77	0.85	0.77
10	0.73	0.79	0.76	0.84	0.87	0.79

CH method was successful on 22 data sets. Similar to Average Silhouette method, as the number of clusters of the data sets increased, the success of the CH method decreased. On the contrary to the AS, CH was successful in all data sets simulated with only 2 clusters. An example of CH method's results on the data sets with four clusters can be seen in Table 4.3.

Table 4.3 Calinski-Harabasz method's results on the same cluster sized data sets with four clusters.

k	A1	A2	A3	B1	B2	B3
2	0.52	0.61	0.68	0.89	0.98	0.97
3	0.58	0.62	0.61	0.63	0.72	0.80
4	0.73	0.80	0.81	0.39	0.55	0.59
5	0.41	0.53	0.61	0.57	0.63	0.73
6	0.58	0.67	0.68	0.70	0.71	0.72
7	0.71	0.70	0.73	0.75	0.82	0.98
8	0.73	0.75	0.79	0.79	0.84	0.78
9	0.69	0.75	0.75	0.77	0.85	0.77
10	0.73	0.79	0.76	0.84	0.87	0.79

DB achieved worse results than the AS and CH. It correctly determined the number of clusters for 16 data sets. It failed in scenarios where cluster centers were close and cluster variance was high. While it successfully detected the number of clusters in all 2 and 3-clustered data sets, it was less successful as the number of clusters increased. In Table 4.4, an example of DB method's results on the data sets with four clusters can be seen.

Table 4.4 Davies-Bouldin method's results on the same cluster sized data sets with four clusters.

k	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**
2	0.52	0.61	0.68	0.89	0.98	0.97
3	0.58	0.62	0.61	0.63	0.72	0.80
4	0.73	0.80	0.81	0.39	0.55	0.59
5	0.41	0.53	0.61	0.57	0.63	0.73
6	0.58	0.67	0.68	0.70	0.71	0.72
7	0.71	0.70	0.73	0.75	0.82	0.98
8	0.73	0.75	0.79	0.79	0.84	0.78
9	0.69	0.75	0.75	0.77	0.85	0.77
10	0.73	0.79	0.76	0.84	0.87	0.79

ASCM achieved better results than all the other methods. In the all data sets, the method was able to correctly determine the number of clusters in all data sets. In Table 4.5, an example of ASCM's results on the data sets with four clusters can be seen.

Table 4.5 Adjusted Silhouette Coefficient Based on Median's results on the same cluster sized data sets with four clusters.

k	A1	A2	A3	B1	B2	B3
2	0.64	0.63	0.66	0.61	0.58	0.58
3	0.73	0.73	0.72	0.71	0.66	0.64
4	0.85	0.80	0.76	0.81	0.74	0.72
5	0.66	0.76	0.72	0.76	0.69	0.71
6	0.80	0.71	0.71	0.72	0.65	0.67
7	0.70	0.66	0.62	0.69	0.62	0.60
8	0.73	0.69	0.69	0.58	0.59	0.64
9	0.63	0.65	0.65	0.61	0.58	0.65
10	0.62	0.59	0.61	0.61	0.60	0.59

In addition to whether the methods correctly determine the number of clusters, two validation metrics (Rand Index, Meila's Variation of Information) were used to test how successful the methods are in placing observations in the correct cluster. For this purpose, the validation metrics were calculated as a result of clustering with the number of clusters proposed by each method and the validity of the clusters was also compared. In addition to these two metrics, Table 4.6 shows the validation results including the Success Rate (SR), which is the proportion of the total number of data sets for which the number of clusters was correctly determined, and Affinity Rate(AR), which is calculated to test how close proposed cluster number are to true label. AR is calculated as the sum of the differences between the number of clusters incorrectly proposed and true label divided by the total number of data sets. Small Affinity Rate means method proposed the cluster number more closely than higher affinity rate. At this point, it needs to be noted that since in thesis results of the lots of data sets are compared, affinity rate is calculated. Similar to the success rate, affinity rate cannot be calculated for a clustering results. It requires more than two clustering results. For all metrics in the Table 4.6, ASCM was the most succesful method.

Table 4.6 Validation metrics of the first part

Methods	**RI**	**MVI**	**SR**	**AR**
AS	0.96	0.23	0.43	0.95
CH	0.97	0.19	0.52	0.95
DB	0.97	0.19	0.38	0.71
ASCM	**0.99**	**0.08**	**1.00**	**0.00**

4.2 Results of the Different Sized Data Sets

In this study, 42 data sets were created with different numbers of observations for each cluster. Just like the first same sized data sets, the scatter plot (see Figure 4.2) of the data sets simulated as only four clusters and Tables containing the coefficient information of the methods are given.

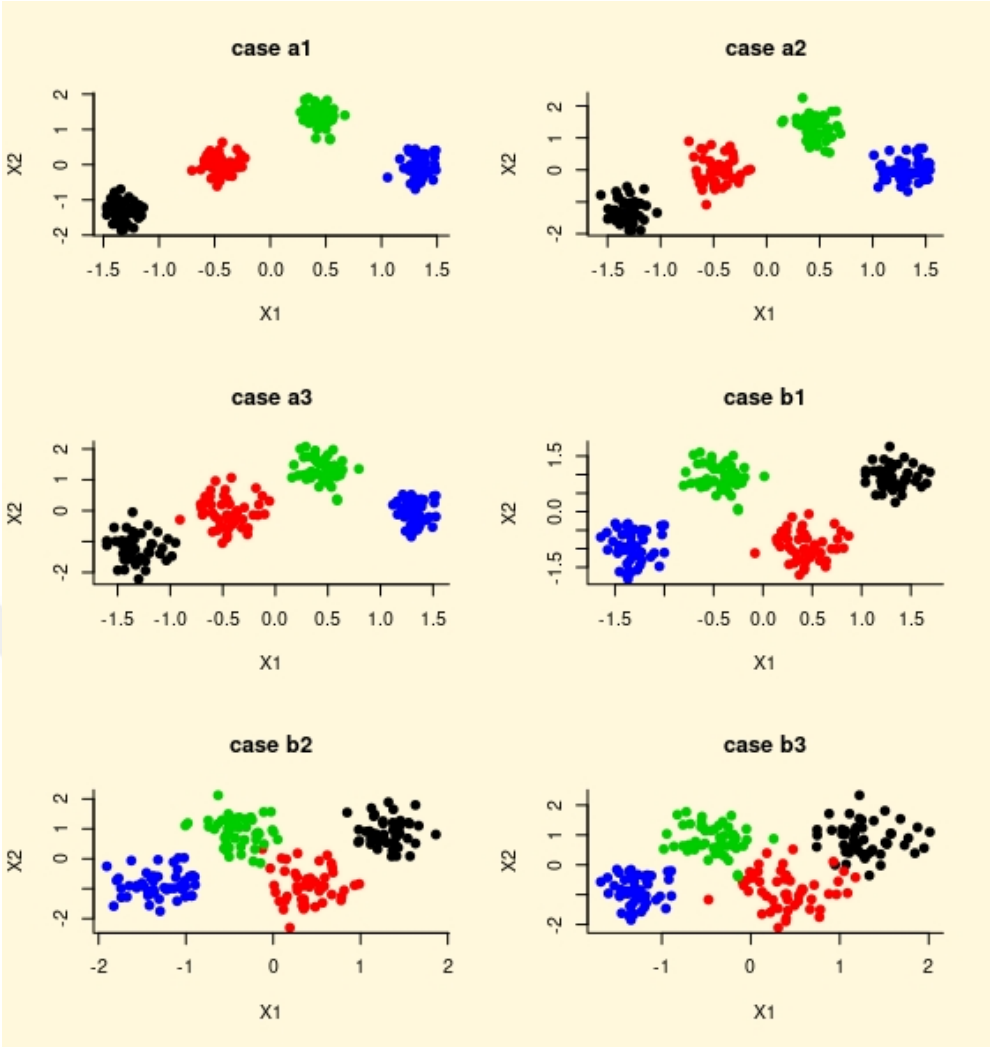


Figure 4.2 An example scatter plot of the data sets created for the data sets with the different cluster sized

AS correctly determined the number of clusters in 15 data sets. It successfully determined the number of clusters in all the data sets simulated with two and three clusters. It was not successful in any of the data sets with five, six, seven, and eight clusters. An example of AS's results on the data sets with four clusters can be seen in Table 4.7.

Table 4.7 Average Silhouette method's results on the different cluster sized data sets with four clusters.

k	A1	A2	A3	B1	B2	B3
2	0.65	0.65	0.64	0.49	0.44	0.39
3	0.67	0.63	0.62	0.63	0.54	0.47
4	0.61	0.57	0.55	0.69	0.59	0.53
5	0.73	0.63	0.57	0.59	0.54	0.46
6	0.63	0.56	0.49	0.51	0.48	0.45
7	0.60	0.52	0.44	0.49	0.44	0.40
8	0.48	0.42	0.43	0.47	0.42	0.39
9	0.47	0.42	0.43	0.45	0.39	0.38
10	0.46	0.41	0.42	0.45	0.37	0.37

CH was correctly determined cluster numbers of 20 data sets. Compared to the first part, it was less successful in correctly determining the number of clusters. While it correctly determined the number of clusters in all the data sets with two and eight clusters, it did not successfully determine the number of clusters in any of the data sets simulated as five and seven clusters. An example of CH's results on the data sets with four clusters can be seen in Table 4.8.

Table 4.8 Calinski-Harabasz method's results on the different cluster sized data sets with four clusters.

k	A1	A2	A3	B1	B2	B3
2	569.76	602.28	568.03	197.92	176.81	154.51
3	597.67	604.45	553.24	365.48	267.34	211.96
4	2054.84	1100.46	763.44	602.66	352.28	266.55
5	1939.24	1031.24	781.40	231.90	338.77	267.78
6	1763.95	926.66	784.36	484.83	302.36	276.71
7	1979.30	975.63	704.70	450.15	345.19	272.86
8	1736.14	897.98	609.21	504.44	331.27	247.90
9	1188.97	959.74	767.08	496.59	343.39	259.23
10	1712.99	866.87	719.65	457.62	295.62	260.57

DB was able to correctly determine the number of clusters in 24 of the 42 data sets simulated in this part. It correctly determined the number of clusters in all simulated data sets with two, three and eight clusters. An example of DB's results on the data sets with four clusters can be seen in Table 4.9.

Table 4.9 Davies-Bouldin method's results on the different cluster sized data sets with four clusters.

k	A1	A2	A3	B1	B2	B3
2	0.48	0.44	0.49	0.89	0.93	0.95
3	0.50	0.59	0.68	0.61	0.71	0.76
4	0.71	0.73	0.77	0.72	0.82	0.64
5	0.49	0.56	0.59	0.79	0.65	0.73
6	0.59	0.60	0.63	0.66	0.72	0.79
7	0.57	0.73	0.70	0.68	0.77	0.81
8	0.72	0.74	0.74	0.83	0.83	0.82
9	0.74	0.75	0.77	0.79	0.82	0.83
10	0.75	0.81	0.81	0.88	0.85	0.87

ASCM was the most successful method, as it was able to correctly determine the number of clusters for 40 data sets. ASCM was successful in all data sets simulated with two, three, four, five, and seven clusters, but failed to correctly determine the number of clusters for only one data set in the data sets with six and eight clusters. An example of ASCM method's results on the data sets with four clusters can be seen in Table 4.10.

Table 4.10 Adjusted Silhouette Coefficient Based on Median's results on the different cluster sized data sets with four clusters.

k	A1	A2	A3	B1	B2	B3
2	0.78	0.79	0.77	0.60	0.57	0.55
3	0.83	0.79	0.77	0.67	0.60	0.62
4	0.86	0.80	0.76	0.79	0.74	0.71
5	0.82	0.66	0.71	0.54	0.69	0.65
6	0.61	0.60	0.59	0.74	0.68	0.65
7	0.72	0.65	0.61	0.73	0.64	0.62
8	0.77	0.71	0.67	0.66	0.64	0.62
9	0.67	0.63	0.67	0.68	0.63	0.62
10	0.68	0.64	0.59	0.64	0.61	0.59

As in the first simulation study, in addition to comparing whether the methods correctly determine the number of clusters, RI, MVI and AR metrics were also compared in Table 4.11. As stated earlier, ASCM were the most successful method on the success rate. In the case of RI,MVI, and AR, ASCM method was the most successful method. Even though success rate of AS was less than CH method, it is more successful in the case of RI and MVI. From this, it can be inferred that in scenarios where the number of clusters cannot be accurately determined, AS determines the correct number of clusters closer than CH.

Table 4.11 Validation metrics of the second part

Methods	RI	MVI	SR	AR
AS	0.96	0.22	0.36	0.80
CH	0.95	0.28	0.48	1.57
DB	0.96	0.20	0.57	0.64
ASCM	**0.99**	**0.06**	**0.95**	**0.04**

CHAPTER FIVE
REAL DATA APPLICATIONS

In this section, ASCM and other methods compared on nine benchmark data sets. All of the data sets are taken from UCI Machine Learning Repository (Dua & Graff, 2017). Data sets used in this study are wine, column3c, ecoli, iris, haberman, wdbc, breast tissue, glass, yeast. In the remainder of this section, the results obtained for each data set is explained separately. Scatter plots were drawn with PC1 on the x-axis and PC2 on the y-axis, for only PCA applied data sets. Scatter plots of standardized data sets were not plotted because they are multidimensional.

5.1 Wine Data Set

Wine data set from the UCI Machine Learning Repository contains information on the chemical analysis of wines grown in a particular region in Italy. There are 178 observations in the data set with 13 variables. One of these variables is cultivar, which represents the grape variety that the wine was made from. Cultivar is the label of the data set which it has three classes that methods needs to be determined. In data preparation process, data set is standardized because of the differences between scales of the variables. All of the methods successfully determined the correct cluster number, as it can be seen from Table 5.1.

Table 5.1 Coefficients of each method for wine data set

k	AS	CH	DB	ASCM
2	0.27	69.49	1.53	0.25
3	0.28	70.94	1.39	0.45
4	0.26	56.20	1.79	0.35
5	0.23	46.11	1.94	0.31
6	0.24	41.20	1.57	0.32
7	0.22	36.59	1.57	0.37
8	0.19	36.31	1.75	0.29
9	0.13	33.83	1.72	0.34
10	0.14	31.56	1.55	0.37

5.2 Column3c Data Set

This data set, used in the medical diagnosis of spine tumors, has 310 observations and six variables. In addition to these variables, there is a variable called Class, which contains information about the three classes in the data set. Therefore, the number of clusters that the methods should determine is three. In Figure 5.1, it can easily be seen that there is an overlap between two clusters. It should be difficult for methods to determine these clusters separately.

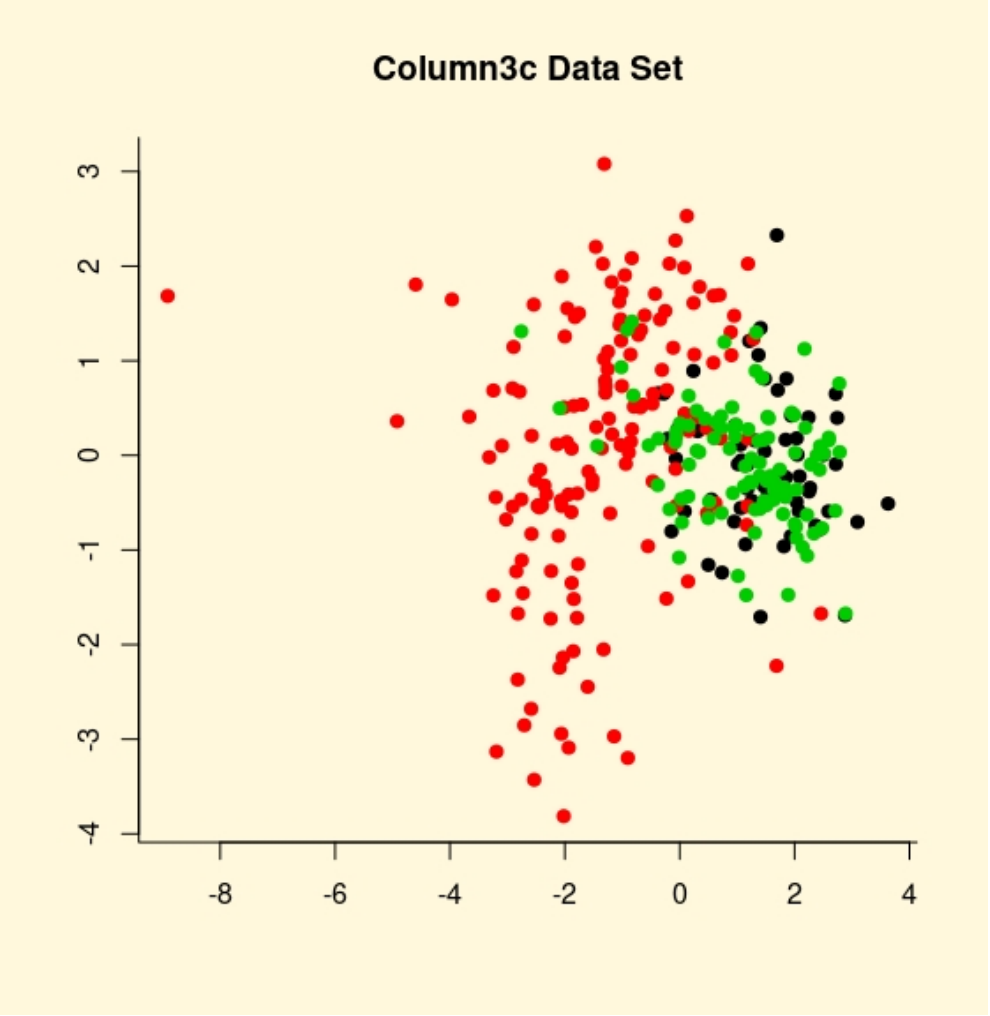


Figure 5.1 Scatter plot of column3c data set

In the data preparation process, PCA is applied to the data set because of the high correlation between variable pairs. Two principle components with 0.7401 cumulative proportion is selected as the principle component number. None of the methods successfully determined the correct cluster number, as it can be seen from Table 5.2.

Table 5.2 Coefficients of each method for column3c data set

k	AS	CH	DB	ASCM
2	0.47	325.31	0.84	0.61
3	0.42	313.24	0.87	0.58
4	0.36	275.56	0.86	0.55
5	0.35	248.09	0.90	0.56
6	0.36	268.89	0.84	0.57
7	0.34	257.29	0.89	0.59
8	0.33	229.47	0.82	0.60
9	0.36	271.44	0.82	0.60
10	0.38	274.77	0.89	0.58

5.3 Ecoli Data Set

Ecoli data set contains information on the characteristics of Escherichia coli bacteria, a commonly studied bacterium in biology and microbiology. There are 9 variables and 336 observations in the data set. One of these variables contains label information. There are 8 different labels in this variable, so the optimal number of clusters that the methods need to determine is eight. There are some clusters that observation number they contain are less than 5. Therefore, this data set provides a useful test for the detection of small clusters. In the data preparation process, PCA was not performed as it was not found to be highly correlated between pairs of variables. However, the data set was standardized due to the high difference between the scales of the variables. As it can be seen from Table 5.3, ASCM and DB successfully determined the cluster number correctly.

Table 5.3 Coefficients of each method for the ecoli data set

k	AS	CH	DB	ASCM
2	0.32	105.40	1.41	0.49
3	0.38	112.72	1.27	0.51
4	0.30	83.70	1.54	0.56
5	0.34	112.87	1.56	0.36
6	0.24	102.25	1.35	0.54
7	0.21	93.00	1.27	0.50
8	0.24	91.60	1.00	0.77
9	0.22	155.45	1.32	0.48
10	0.22	145.62	1.28	0.74

5.4 Iris Data Set

Iris data set contains 150 observations and five variables including the species of the iris. In this data set, methods need to determine three species of the Iris as the optimal number of cluster. In Figure 5.2, it can easily be seen that there is an overlap between the clusters represented with the green and red. Therefore, it might be difficult for methods to separate these clusters.

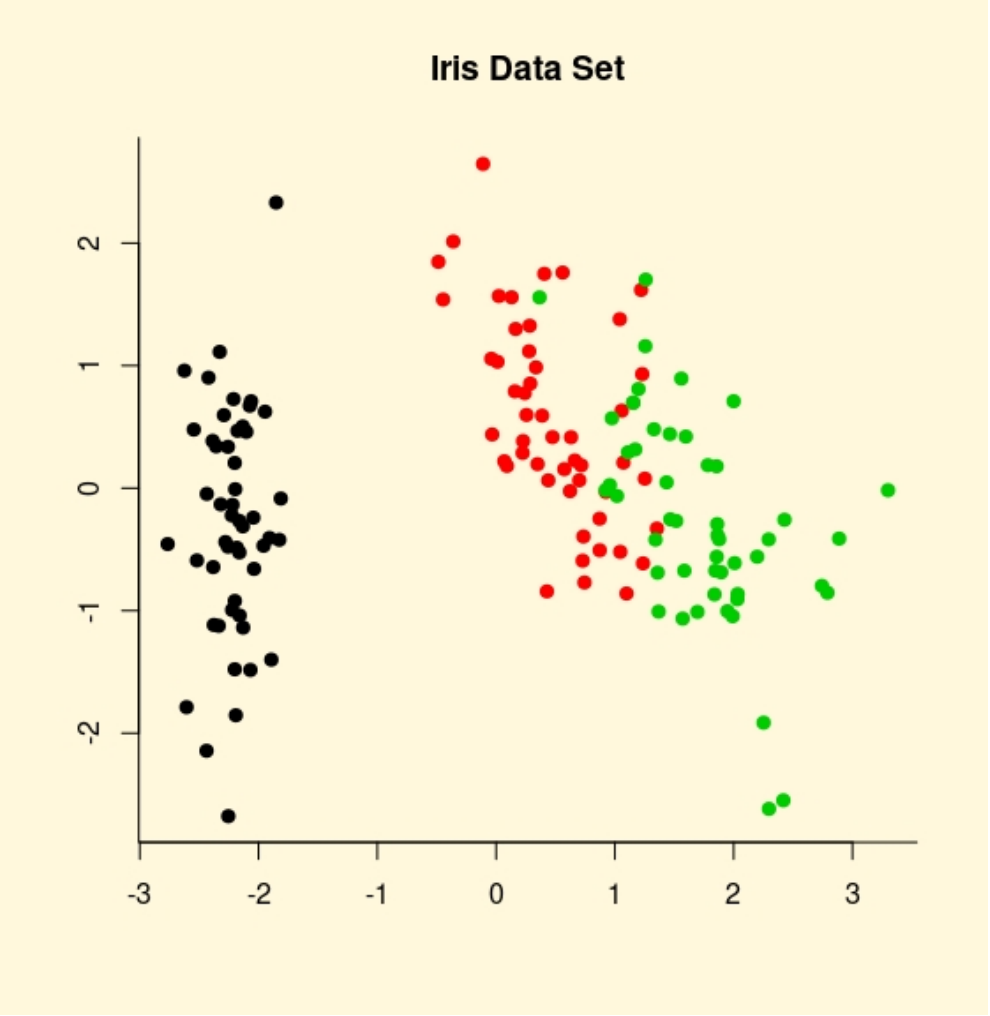


Figure 5.2 Scatter plot of iris data set

Since the correlation between pairs of variables is high, PCA was applied to the data set at the data preparation process. Two principle components are chosen with cumulative proportion of 0.9581. As it can be seen from Table 5.4, no method could successfully determine the number of clusters. All methods determined cluster number as two. Thus, it can be inferred that methods could not detect the overlap between clusters represented with green and red.

Table 5.4 Coefficients of the iris data set for each method

k	AS	CH	DB	ASCM
2	0.61	283.01	0.54	0.74
3	0.51	180.98	0.71	0.67
4	0.48	262.42	0.74	0.66
5	0.41	214.44	0.89	0.71
6	0.42	259.65	0.82	0.64
7	0.43	176.18	0.70	0.68
8	0.40	272.24	0.70	0.66
9	0.44	248.39	0.82	0.66
10	0.41	195.63	0.66	0.61

5.5 Haberman Data Set

Haberman data set contains clinical data from patients treated for breast cancer. The data set is based on a study conducted at the University of Wisconsin Cancer Center between 1958 and 1970. There are a total of 306 observations and four variables which are patient age, year treated, number of lymph nodes, and label. The variable containing the label information has two classes, meaning that the methods must determine two as the optimal number of clusters. In the data preparation process, data set is standardized since there was no high correlation between variables pairs and number of variables is less. As it can be seen from Table 5.5, no methods successfully determine the correct cluster number. AS, CH, and ASCM determined cluster number as five, while DB determined it as four.

Table 5.5 Coefficients of each method for the haberman data set

k	AS	CH	DB	ASCM
2	0.26	93.47	1.57	0.38
3	0.26	130.00	1.13	0.46
4	0.27	142.51	0.99	0.53
5	0.31	151.22	1.19	0.54
6	0.28	135.43	1.10	0.51
7	0.30	138.82	1.00	0.50
8	0.28	125.96	1.05	0.50
9	0.28	127.35	1.06	0.51
10	0.29	128.51	1.00	0.51

5.6 Breast Cancer Wisconsin Data Set

Breast Cancer Wisconsin is a data set of clinical characteristics of patients diagnosed with breast cancer. Data was collected at the University of Wisconsin Hospital in 1992 and it contains a total of 569 observations. There are 30 variables in the data set, however, in this analysis ten variables are used: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These attributes are standard histopathological features used to describe the characteristics of the tumor. Another variable is diagnosis which contains label information. There are two classes which methods must determine. In Figure 5.3, it can be observed that there is an overlap between clusters, yet, clusters can be detected.

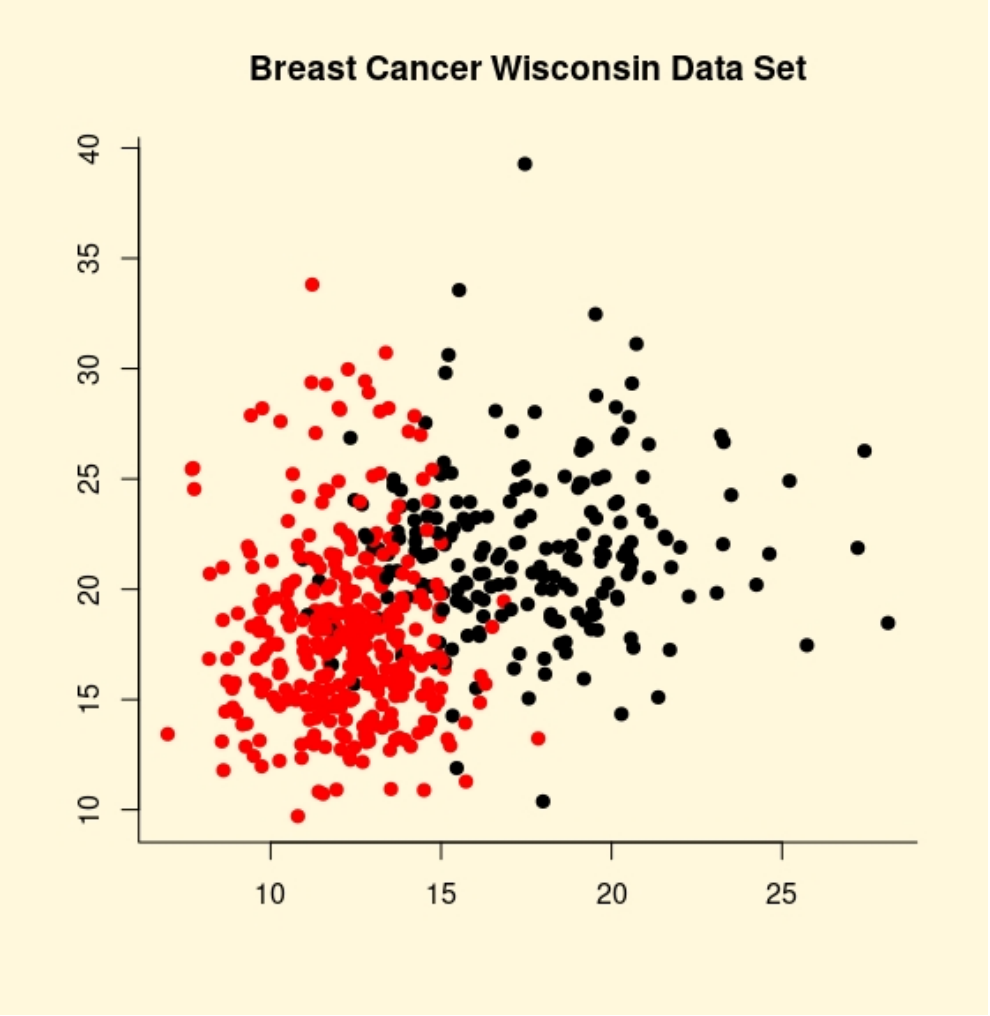


Figure 5.3 Scatter plot of breast cancer wisconsin data set

Since the correlation between pairs of variables is high, PCA was applied to the data set at the data preparation process. Two principle components is chosen with cumulative proportion of 0.7997. As it can be seen from Table 5.6, all methods successfully determined the correct cluster number.

Table 5.6 Coefficients of each method for the breast cancer wisconsin data set

k	AS	CH	DB	ASCM
2	0.49	534.47	0.87	0.58
3	0.44	468.31	0.93	0.57
4	0.34	425.54	1.02	0.53
5	0.35	453.34	0.91	0.55
6	0.35	390.21	0.95	0.54
7	0.33	429.38	0.90	0.53
8	0.32	421.38	0.87	0.55
9	0.31	410.69	0.91	0.53
10	0.32	397.64	0.91	0.54

5.7 Breast Tissue Data Set

Breast Tissue data set contains characteristics of different types of breast tissue and can be used to understand breast tissue characteristics of patients diagnosed with breast cancer or to determine cancer risk based on breast tissue type. The data set contains a total of 106 observations and nine variables. Each variable refers to a degree of differentiation feature. Each observation in the data set is paired with a breast tissue sample and contains a label that identifies the type of breast tissue. Label has six different classes. Therefore, for this data set, it is expected for methods to determine six clusters as the optimal number of clusters. It can be observed that clusters are so nested that it could nearly impossible for methods to detect in Figure 5.4.

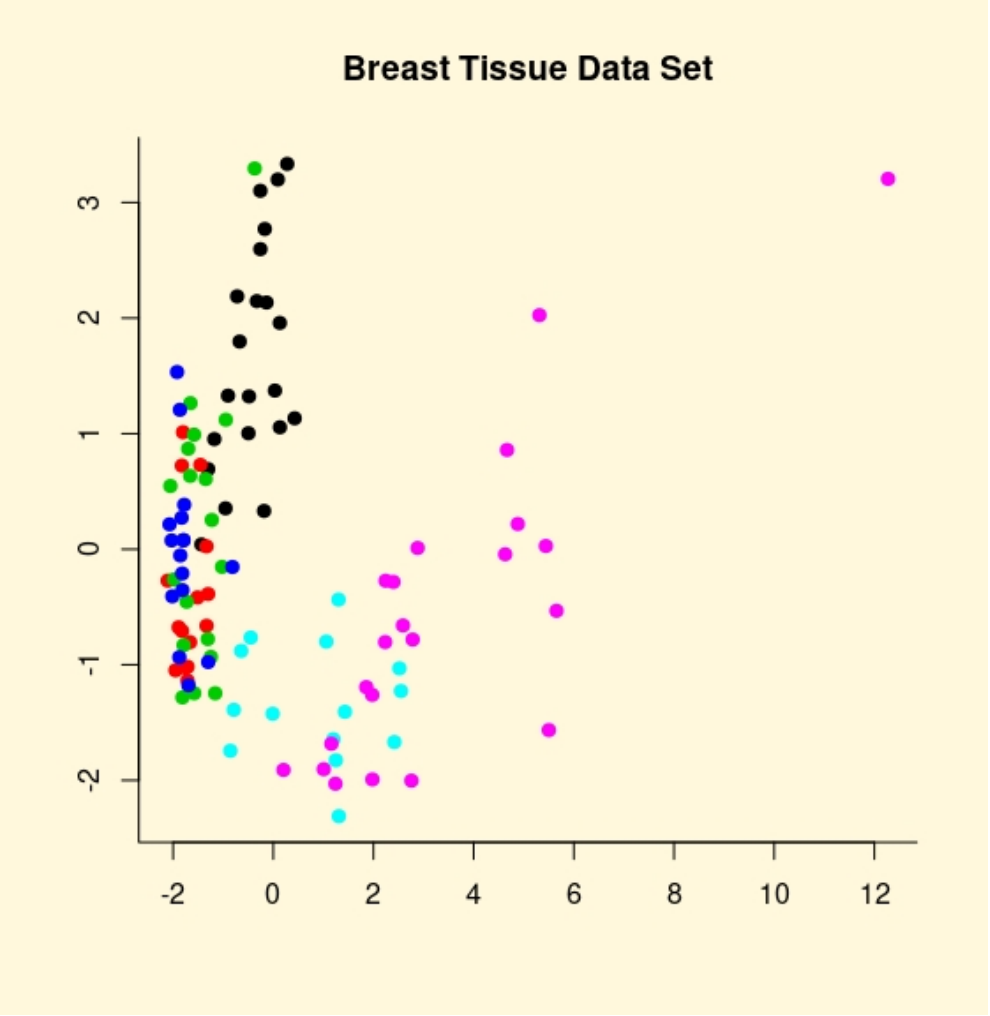


Figure 5.4 Breast Tissue data set

In the data preparation process, it is noticed that there are high correlation between variables. Thus, PCA is applied to the data set. Two principle component is chosen with cumulative proportion of 0.8082. As it can be observed from Table 5.7, only ASCM method successfully determined the correct cluster number. AS incorrectly determined the cluster number as two, CH as five, and DB as three.

Table 5.7 Coefficients of each method for the breast tissue data set

k	AS	CH	DB	ASCM
2	0.57	117.68	0.73	0.65
3	0.48	96.94	0.65	0.63
4	0.42	108.23	0.67	0.68
5	0.47	138.25	0.82	0.62
6	0.46	119.53	0.71	0.82
7	0.42	103.37	0.74	0.80
8	0.42	99.29	0.83	0.66
9	0.43	95.30	0.85	0.63
10	0.45	86.76	0.73	0.62

5.8 Glass Data Set

Glass data set contains features of different types of glass. There are a total of 214 observations and nine variables in the data set. Variables are glass number, aluminum, silicon, potassium, calcium, sodium, magnesium, iron and glass type. Glass types are divided into 7 different classes: auto glass, building glass, container glass, vehicle glass, decorative glass, lamp glass and other glass. Therefore, methods to determine the optimal number of clusters must determine seven as the number of clusters for this data set. Since there are no high correlation between variable pairs, it is decided in the data preparation process that there is no need for PCA. Rather, differences between scales leads to the standardization. Thus, the data set is standardized. As it can be seen from Table 5.8, only ASCM successfully determine the correct cluster number.

Table 5.8 Coefficients of each method for the glass data set

k	AS	CH	DB	ASCM
2	0.45	55.52	1.67	0.39
3	0.37	58.29	1.52	0.40
4	0.27	60.06	1.07	0.40
5	0.31	58.29	1.03	0.66
6	0.32	56.19	1.20	0.66
7	0.34	56.26	1.18	0.63
8	0.34	59.82	1.15	0.42
9	0.26	53.57	1.17	0.59
10	0.32	52.20	1.25	0.57

5.9 Yeast Data Set

Yeast is a data set containing features of proteins of yeast cells. The data set contains a total of 1484 observations nine variables. Variables include various protein properties derived from proteomic measurements. Each observation in the data set is paired with a yeast cell sample and contains a label that identifies which protein group the sample belongs to. In total, the data set contains ten different protein groups. Therefore methods must determine the number of clusters as ten. In the data preparation process, data set is standardized since the scales between variables different. As it can be observed from Table 5.9 that none of the methods successfully determined the cluster number. CH and ASCM's suggestions were closer to the correct cluster number.

Table 5.9 Coefficients of the yeast data set for each method

k	AS	CH	DB	ASCM
2	0.27	325.43	1.66	0.34
3	0.16	286.49	1.78	0.30
4	0.28	252.28	1.52	0.38
5	0.23	243.30	1.40	0.60
6	0.23	363.51	1.25	0.58
7	0.20	325.91	0.95	0.57
8	0.24	560.05	1.04	0.65
9	0.22	529.33	1.38	0.64
10	0.21	504.18	1.10	0.63

Overall, out of a total of nine real data sets, ASCM successfully determined the number of clusters for five data sets. AS and CH successfully determined two data sets, while the DB method successfully determined only one data set. ASCM was also able to successfully determine the data sets that were successfully determined by the other methods. No other method was able to successfully determine any data set that the ASCM was not able to successfully determine. Therefore, ASCM, which achieved a success rate of 0.55 was the most successful method in the real data sets test. As it can be seen from Table 5.10, ASCM was also the most successful method in AR.

Table 5.10 Validation metrics of the real data set study

Methods	AR	SR
AS	2.6	0.22
CH	1.2	0.22
DB	1.6	0.33
ASCM	**0.7**	**0.55**

CHAPTER SIX
CONCLUSION

One of the most important problems in clustering analysis is the determination of the number of clusters, since the number of clusters must be known for many algorithms to work. In this thesis, the Average Silhouette method, which is one of the most widely used methods, has been adjusted to be more robust and to work more successfully in some special cases, and this method is called Adjusted Silhouette Coefficient Based on Median.

Two criteria were taken into account while making improvements. As opposed to Average Silhouette method, which calculates silhouette value of each observation and measures how successful the clustering result of that observation is, Adjusted Silhouette Coefficient Based on Median is cluster-based. It may be more efficient to explain the reason behind this adjustment with an example. For example, in an observation-based calculation, an observation may be very close to an observation of the closest cluster. In this case, it will receive a high similarity value and affect the clustering result. In Adjusted Silhouette Based on Median, it is aimed to prevent such situations by calculating the distance to the center of the cluster (median) and to calculate how an observation fits with the cluster it belongs to.

The second criterion taken into account when making improvements also gives the method its name. Since the calculations are performed on clusters instead of observations, clusters need to be represented. When calculating the Adjusted Silhouette Coefficient Based on Median, the median of each cluster was used as the center representing each cluster. This adjustment makes the method more robust, increases the computational speed and makes it applicable for large data sets.

The performance of Adjusted Silhouette Coefficient Based on Median was compared with three different methods (Average Silhouette, Calinski-Harabasz, Davies-Bouldin) on both synthetic and benchmark data sets. In the simulation study with synthetic data sets, the success of the methods was tested with a study consisting of two parts of the same and different sized clusters. For each part, 42 data sets were

simulated with 2 to 8 clusters. For each number of clusters, the simulated data sets were subjected to six different scenarios depending on the distance between cluster centers and cluster variances. Adjusted Silhouette Coefficient Based on Median method was the most successful method in both parts of these simulation studies on a total of 84 data sets. While the method correctly determined the number of clusters for 82 data sets, it failed to correctly determine the number of clusters for the case of two data sets simulated with different sized clusters. It was also the most successful method in cluster validity calculations with Rand Index and Meila's Variation of Information metrics in both parts.

In the real data application part, a total of nine benchmark data sets were used. Data sets from different fields were either standardized or PCA was applied before clustering and the transformed versions were used for clustering. Adjusted Silhouette Based on Median correctly identified the clusters of five data sets, which is more successful than Davies-Bouldin, which correctly identified the clusters of three data sets. In this section, we also compared how close the clusters were to the correct cluster number using a metric called Affinity Rate. Adjusted Silhouette Coefficient Based on Median was the most successful in this metric.

Due to the simulation study and real data applications, Adjusted Silhouette Coefficient Based on Median is not only more successful than Average Silhouette, but also better than all other methods compared in this study. Thus Adjusted Silhouette Coefficient Based on Median may offer a solution to the cluster number determination problem. Results of this study can be a reference for researchers who are looking for a new method for cluster number determination in the field of data science.

Although Adjusted Silhouette Coefficient Based on Median is more successful any other methods, it also has limitations. First of all, the study was conducted only for k-means clustering algorithm. Its success in other clustering algorithms may also need to be tested. In addition, especially in data sets with different shaped like aggregation, compound, pathbased, flame, and spiral, it can be said that its success is quite low. Therefore, it cannot be said that the method works well in all kind of data

sets. Future work could be to test Adjusted Silhouette Coefficient Based on Median on other clustering algorithms and make modifications to make it more successful for data sets with different shapes.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

Almanza-Ortega, N. N., Pérez-Ortega, J., Zavala-Díaz, J. C., & Solís-Romero, J. (2022). Comparative analysis of k-means variants implemented in r. *Computación y Sistemas*, *26*(1), 125–133.

Arbelaitz, O., G. I. M. J. P. J. M. . P. I. (2013). An extensive comparative study of cluster validity indices. *Pattern recognition*, 243–256.

Barbará, D., . J. S. (2002). Applications of data mining in computer security. *Springer*.

Ben-Hur, A., & Guyon, I. (2003). Detecting stable clusters using principal component analysis. *Functional Genomics: Methods and Protocols*, 159–182.

Benko, A. Lányi, C. S. (2009). History of artificial intelligence. *IGI GLOBAL*, 1759–1762.

Berthold, M. R., & Höppner, F. (2016). On clustering time series using euclidean distance and pearson correlation. *arXiv preprint arXiv:1601.02213*.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical methods*, *6*(9), 2812–2831.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27.

Chou, C. H., S. M. C. . L. E. (2004). A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications*, 205–220.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224–227.

Dimitriadou, E., Dolničar, S., & Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, *67*(1), 137–159.

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, 29.

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. from http://archive.ics.uci.edu/ml.

Duda, R. O., & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. in machine learning in radiation oncology.

Hady, M. F. A., . S. F. (2013). Semi-supervised learning. *Handbook on Neural Information Processing*, 215–239.

Hardy, A. (1996). On the number of clusters, computational statistics and data analysis, 23, 83-96.

Hartigan, J. A., . W. M. A. (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 100–108.

Hennig, C., & Imports, M. (2015). Package 'fpc'. *Flexible Procedures for Clustering*.

Hinton, G., . S. T. J. (1999). Unsupervised learning: foundations of neural computation.

Holzinger, K. J., . H. H. H. (1941). Factor analysis; a synthesis of factorial methods.

Hruschka, E. R., Hruschka Jr, E. R., & Ebecken, N. F. (2003). A nearest-neighbor method as a data preparation tool for a clustering genetic algorithm. In *SBBD*, *Citeseer*, 319–327.

Jain, A. K., & Dubes, R. (1979). Algorithms for clustering data.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*.

Kassambara, A., & Mundt, F. (2017). Package 'factoextra'. *Extract and visualize the results of multivariate data analyses*, *76*(2).

Learned-Miller, E. G. (2014). Introduction to supervised learning. i: Department of computer science.

Lee, J. A., & Verleysen, M. (2009). Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, *72*(7-9), 1431–1443.

Löster, T. (2016). Determining the optimal number of clusters in cluster analysis. *The 10th international days of statistics and economics, Prague*.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, *Oakland, CA, USA*, 281–297.

Marutho, D., H. S. H. . W. E. (2018). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. *International seminar on application for technology of information and communication*, 533–538.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of multivariate analysis*, *98*(5), 873–895.

Mi, X., Miwa, T., & Hothorn, T. (2009). mvtnorm: New numerical algorithm for multivariate normal probabilities. *R Journal 1 (2009), Nr. 1*, *1*(1), 37–39.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*, 159–179.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of classification*, *5*, 181–204.

Mirkin, B. (2005). Clustering for data mining: a data recovery approach. *Chapman and Hall*.

Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303.

Nazeer, K. A., & Sebastian, M. (2009). Improving the accuracy and efficiency of the k-means clustering algorithm. In *Proceedings of the world congress on engineering*, vol. 1, *Association of Engineers London London, UK*, 1–3.

Öztürk, F. E., & Demirel, N. (2023). Comparison of the methods to determine optimal number of cluster. *Veri Bilimi*, *6*(1), 34–45.

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of computers and communications*, *5*(1), 27–34.

Ringnér, M. (2008). What is principal component analysis? *Nature biotechnology*, *26*(3), 303–304.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Savage, N. J., & Huffman, J. A. (2018). Evaluation of a hierarchical agglomerative clustering method applied to wibs laboratory data for improved discrimination of biological particles by comparing data preparation techniques. *Atmospheric Measurement Techniques*, *11*(8), 4929–4942.

Schneider, P., & Xhafa, F. (2022). Chapter 2 - data stream processing: models and methods: The complexity of data stream processing. In Schneider, P., & Xhafa, F., editors, *Anomaly Detection and Complex Event Processing over IoT Data Streams*, 29–47.

Sokal, R. R. (1966). Numerical taxonomy. *Scientific American*, 106–117.

Tang, J., Wang, D., Zhang, Z., He, L., Xin, J., & Xu, Y. (2017). Weed identification based on k-means feature learning combined with convolutional neural network. *Computers and electronics in agriculture*, *135*, 63–70.

Vavrek, M. J. (2011). Fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontologia electronica*, *14*(1), 16.

Walesiak, M., Dudek, A., & Dudek, M. A. (2011). *clustersim package*.

Warrens, M. J., & van der Hoef, H. (2020). Understanding the rand index. In *Advanced Studies in Classification and Data Science*, *Springer*, 301–313.