Bachelor Project

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Katedra teorie obvodů

Cluster separability in multidimensional biomedical data

Ivana Klikarová

Supervisor: Ing. Eduard Bakštein, Ph.D.
Field of study: Medical Electronics and Bioinformatics
May 2024

Acknowledgements

I want to thank Ing. Eduard Bakštein, Ph.D., for the invaluable support and guidance throughout this project. Your assistance and patience are greatly appreciated.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of the university thesis.

In Praze, . May 2024

Abstract

This study presents a comprehensive evaluation of a number of cluster estimation methods applied to both simulated and real biomedical data. We use methodology for assessing clustering quality using synthetic datasets that mimic real-world biomedical data characteristics. These datasets vary in inter-individual variability, noise levels, and cluster separability, allowing for systematic evaluation of clustering methods' robustness. Each data point is assigned to a ground truth cluster, serving as reference labels for evaluating clustering accuracy. We characterize clusters using interclass-to-between-class ratios and analyze the performance of clustering algorithms across different dataset dimensions.

Additionally, we apply clustering methods to real biomedical data obtained from the National Institute of Mental Health, focusing on COVID-related variables. Exploratory data analysis, preprocessing, and principal component analysis are conducted before clustering estimation. The study aims to provide insights into the performance of clustering methods and their applicability to real-world biomedical data.

Keywords:

Supervisor: Ing. Eduard Bakštein, Ph.D.

Abstrakt

Tato studie představuje přehled metod pro odhad počtu shluků aplikovaných na simulovaná i reálná biomedicínská data. Byla měřena kvalita predikce počtu shluků pomocí syntetických datasetů, které napodobují vlastnosti reálných biomedicínských dat. Tyto soubory dat se liší z hlediska variability, úrovně šumu a separability shluků, což umožňuje systematické hodnocení robustnosti metod. Každý bod je přiřazen ke shluku a slouží jako reference pro hodnocení přesnosti shlukování. Shluky charakterizujeme pomocí poměrů vzdáleností ve shluku a mezi nimi.

Kromě toho aplikujeme metody shlukování na reálná biomedicínská data získaná z Národního ústavu duševního zdraví (NÚDZ) se zaměřením na proměnné související s COVID nákazou. Před odhadem shlukování je provedena explorační analýza dat, předzpracování dat a analýza hlavních komponent. Cílem studie je poskytnout poznatky o přesnosti metod pro odhad počtu shluků a zda je lze použít na reálná data.

Klíčová slova:

Překlad názvu: Separovatelnost shluků v mnoharozměrných biomedicinských datech

Contents

Figures

Tables

Chapter 1

Introduction

Over the past few years, biomedical research has experienced a significant increase in the production of complex, multidimensional data. These data include genomics, proteomics, medical imaging, and clinical data. However, analyzing such complex datasets is challenging, particularly when identifying meaningful patterns and structures that can lead to valuable insights.

In data analysis, clustering is a powerful method for uncovering hidden structures and patterns within datasets. At its core, clustering is a technique that groups similar data based on specific characteristics, thereby enabling the identification of natural groups within the data. Clustering is a fundamental unsupervised learning method, identifying subsets without predefined labels. Various methodological approaches have been developed to address clustering challenges across diverse datasets.
Data clustering algorithms can be categorized into partitioning, hierarchical, density-based, centroid-based, and distribution-based clustering.

Partitioning clustering methods, such as k-means and k-medoids, segment datasets into distinct partitions, assigning each data point to one cluster only. The k-means algorithm is, for example, often used as a prerequisite step for other algorithms. It uses partitioning-based clustering whereby data sets are divided into a predetermined number of clusters. This is enabled through the mechanism of iteratively assigning data points to the nearest cluster centroid and updating the centroids until convergence. However, it is sensitive to outliers and may not always yield an optimal solution. [1] The k-medoids algorithm, on the other hand, a variant of k-means, addresses some of its drawbacks by using the median instead of the mean as the centroid of each cluster. This makes it more robust to outliers. K-medoids iteratively assigns data points to the nearest median and updates the medoids until convergence.[2]

Hierarchical clustering algorithms operate by iteratively establishing clusters based on the experience of previously formed clusters, whereas partitional algorithms define all clusters simultaneously. Hierarchical algorithms can be further categorized into agglomerative, which build clusters from individual

elements by progressively merging them into larger groups (bottom-up), and divisive, which start at the level of the entire dataset with the goal of dividing it into smaller clusters in subsequent steps (top-down).[3]
Density-based clustering methods, presented by algorithms like DBSCAN, are useful for identifying clusters in datasets that exhibit arbitrary shapes and sizes. These algorithms primarily focus on grouping dense regions of points in the data space, which are then distinguished from the areas of low density that lie between them. [4]

Finally, distribution-based clustering methods, such as Gaussian Mixture Models, model clusters as probability distributions, assigning data points to clusters based on their likelihood of belonging to each distribution. GMMs are widely used in various systems to model the probability distribution of continuous measurements or features. These models are defined by parameters such as mean vectors, covariance matrices, and mixture weights, which are typically estimated from training data using methods like the Expectation-Maximization algorithm or Maximum A Posteriori estimation. [5] It is essential to understand that depending on the task, there will be a different method that yields the most effective solution.

Chapter 2
Background

Clustering techniques are widely used to group similar data points based on patterns and similarities, making it easier to understand complex data. However, determining the optimal number of clusters is one of the most significant challenges in cluster analysis. This decision significantly influences the interpretation and application of the results. This challenge becomes even more crucial when dealing with multi-collinear and noisy biological data, where identifying underlying clusters can be complex and uncertain due to the complexity of natural occurrences. Therefore, choosing the correct clustering number and carefully evaluating the results is essential to ensure accurate and reliable clustering outcomes.

2.1 Existing cluster evaluation methods

Choosing the correct number of clusters is a crucial step in clustering analysis as it directly influences the resulting insights. Different methods are available to estimate the optimal number of clusters in a dataset, each with its unique approach and assumptions, and careful consideration is required in selecting an appropriate technique. One approach uses internal validation metrics like the silhouette score or the Davies-Bouldin index, which evaluate the clusters' compactness and separation to identify the best number of clusters. Another popular technique is the gap statistic, which measures the dispersion within clusters and compares it to a reference distribution to determine the optimal number of clusters. Besides, hierarchical clustering methods can use dendrogram-based techniques like the cophenetic correlation coefficient to estimate the number of clusters by assessing the clustering hierarchy's stability. The suitability of each method depends on the dataset's characteristics and the desired outcomes.

2.1.1 Elbow method

The Elbow Method evaluates the proportion of explained variance concerning the number of clusters. It suggests that beyond a certain threshold, increasing the number of clusters does not significantly reduce variance. The method involves gradually increasing the cluster count, represented as k, and recording the Sum of Squared Errors (SSE):

$$SSE = \sum_{k-1}^{k} \sum_{x_i \in S_k} \|x_i - C_k\|_2^2$$

Here, S_k represents the cluster k. SSE represents the sum of the average Euclidean distances from each point to its centroid. The optimal number of clusters, denoted as k, is identified where the SSE experiences a significant drop, forming an "elbow" shape on the plot. This typically happens when adding a new cluster ($k = k + 1$), resulting in the largest decrease in SSE compared to the previous step. Further increases in k tend to yield diminishing returns, as the additional cluster structure may not substantially reduce errors or significantly improve the model fit. [6]

2.1.2 Gap Statistics

The Gap Statistic, introduced by Tibshirani et al. (2001) [7], is a method for estimating the number of clusters in a dataset. It compares the within-cluster dispersion to a null reference distribution to identify the number of clusters that provide the best fit to the data.The gap statistic is computed as follows:

Let $\{x_{ij}\}$ represent observations with $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, where p features are measured on n independent samples. These observations are clustered into k clusters $C_1, C_2, \ldots, C_k$, where C_r denotes the indices of samples in cluster r, and $n_r = |C_r|$. The distance between samples i and i', denoted by $d_{ii'}$, can be calculated, for example, as the squared Euclidean distance $d_{ii'} = \sum_j (x_{ij} - x_{i'j})^2$. The sum of the pairwise distances D_r for all points in cluster r is given by $D_r = \sum_{i,i' \in C_r} d_{ii'}$. Then, we define W_k as:

$$W_k := \sum_{r=1}^{k} \frac{1}{2n_r} D_r.$$

If d represents the squared Euclidean distance, then W_k represents the within-cluster sum of squared distances from the cluster means. For the calculation of the Gap function, Tibshirani et al. proposed using the difference between the expected value of $\log(W_k^*)$ of an appropriate null reference and $\log(W_k)$ of the dataset:

$$\mathrm{Gap}_n(k) := E_n^*[\log(W_k^*)] - \log(W_k).$$

The appropriate number of clusters for the given dataset is determined by finding the smallest value of k such that the Gap statistic satisfies the condition:

$$\mathrm{Gap}_n(k) \geq \mathrm{Gap}_n(k+1) - s_{k+1}$$

where s_k represents the simulation error calculated from the standard deviation $\mathrm{sd}(k)$ of B Monte Carlo replicates of $\log(W_k^*)$, given by $s_k = \sqrt{\frac{1}{1+1/B} \cdot \mathrm{sd}(k)}$.

The expected value $E_n^*[\log(W_k^*)]$ of within-dispersion measures W_{kb}^* is determined as:

$$E_n^*[\log(W_k^*)] = \frac{1}{B}\sum_b \log(W_{kb}^*)$$

where W_{kb}^* is obtained by clustering the B reference datasets. [8]

When using the gap statistic, selecting the right reference distribution is essential. It is suggested that a specific single-component reference distribution, namely the uniform distribution, be used. This ensures that the gap statistic can reliably detect multiple clusters in the data. Regarding univariate data, the uniform distribution U(0,1) is the best choice for this reference distribution. Among all unimodal distributions, it is the most likely to produce spurious clusters detected by the gap test. [7]

A comparison between Gap statistic definitions with and without the logarithm function reveals differences in their performance and interpretation. While the logarithm function is commonly used to standardize the within-cluster sum of squares, some formulations omit this transformation. Tibshirani et al. highlight that the logarithm function provides a more effective means of assessing the deviation of $\log(W_k)$ from its expected value. However, formulations without the logarithm function may still offer insights into the underlying structure of the data with different characteristics. The choice between these definitions depends on the specific requirements of the clustering task. [8]

2.1.3 Davies-Bouldin Index

The Davies–Bouldin index (DB) proposed by Davies and Bouldin in 1979 evaluates clustering quality based on the average similarity between each cluster and its most similar one. The index is formulated as follows:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j}\left(\frac{d_{i,j} + d_{j,i}}{d_{i,i}}\right)$$

Here, $d_{i,j}$ represents the Euclidean distance between the centroids of the i^{th} and j^{th} clusters, while $d_{i,i}$ and $d_{j,i}$ denote the average distances from each data point in the i^{th} and j^{th} clusters to their respective centroids. A lower DB value indicates a well-defined data partition, where clusters are both internally compact and well-separated from each other. [9]

2.1.4 Calinski-Harabasz Index

The Calinski–Harabasz index (CH) proposed by Caliński and Harabasz in 1974 assesses the quality of a clustering solution based on the average sum of squares of between and within clusters. It is calculated as:

$$CH = \frac{SSB}{SSW} \times \frac{n-k}{k-1}$$

Here, SSB represents the average between-cluster sum of squares, SSW is the average within-cluster sum of squares, k is the number of clusters, and n is the number of observations. The average between-cluster sum of squares, SSB, is computed as:

$$SSB = \frac{1}{k} \sum_{i=1}^{k} n_i \cdot \|m_i - \mu\|^2$$

where m_i is the centroid of cluster i, μ is the mean of all data points, and $\|\cdot\|$ represents the Euclidean distance between the centroid of a cluster and the mean of all data points. The average within-cluster sum of squares, SSW, is calculated as:

$$SSW = \sum_{i=1}^{k} \sum_{x \in P_i} \|x - m_i\|^2$$

Here, x represents a sample, P_i is the i^{th} cluster, m_i is the centroid of cluster P_i, and $\|\cdot\|$ denotes the Euclidean distance between a sample and the centroid of its cluster.

A higher CH value indicates a better data clustering result, with larger SSB and smaller SSW values indicating a more well-partitioned cluster. [10]

2.1.5 Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is a statistical measure used for estimating the number of clusters in a dataset. It balances model fit and complexity, penalizing models with a higher number of parameters. The BIC is calculated as:

$$BIC = -2 \log L + k \log n$$

where L is the maximized value of the likelihood function for the estimated model, k is the number of parameters in the model, and n is the sample size. The likelihood function L represents the probability of observing the given data under the assumed model parameters. A higher value of L indicates a better fit of the model to the data. The BIC penalizes models with more parameters, as indicated by the term $k \log n$, to prevent overfitting and favor simpler models. Therefore, the optimal number of clusters is determined by selecting the model with the lowest BIC value, which represents the best compromise between model fit and complexity. [11]

2.1.6 Akaike Information Criterion

The Akaike Information Criterion (AIC) is a statistical measure used for model selection among a set of candidate models. It is based on the principle of trading off the goodness of fit of the model and its complexity.

The formula for computing AIC is given by:

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{2.1}$$

where k is the number of parameters in the model, and L is the maximum likelihood of the model.

A lower AIC value indicates a better model fit, considering both the goodness of fit and the complexity of the model.

The AIC criterion is commonly used in clustering analysis to estimate the number of clusters. By fitting models with different numbers of clusters and computing their AIC values, one can determine the optimal number of clusters that balances model fit and complexity. [12]

2.1.7 Weighted consensus clustering

Weighted Consensus Clustering (WCC) is a technique used to combine multiple clustering solutions obtained from different algorithms or parameter settings into a single consensus clustering solution. The goal of WCC is to enhance the robustness and stability of the clustering results by integrating diverse perspectives from individual clustering solutions.

Let n be the number of data points and m be the number of clustering solutions. Each clustering solution C_i assigns data points to k_i clusters, where $i = 1, 2, ..., m$. The key idea behind WCC is to assign weights to each clustering solution based on its reliability or quality and iteratively update the cluster assignments to maximize the consensus across solutions.

The objective function for WCC can be formulated as follows:

$$\max_{\mathbf{X}} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{k_i} w_{ij}^{(k)} \delta_{\mathbf{x}_j, c_{ij}^{(k)}} \tag{2.2}$$

where $\mathbf{X}$ is the data matrix, $\mathbf{x}_j$ is the j-th data point, $c_{ij}^{(k)}$ is the cluster assignment of data point $\mathbf{x}_j$ in clustering solution C_i for cluster k, $w_{ij}^{(k)}$ is the weight assigned to clustering solution C_i for cluster k, and $\delta_{\mathbf{x}_j, c_{ij}^{(k)}}$ is an indicator function that equals 1 if $\mathbf{x}_j$ is assigned to cluster $c_{ij}^{(k)}$ and 0 otherwise.

The weights $w_{ij}^{(k)}$ are usually initialized to equal values and updated iteratively based on the agreement between clustering solutions. Common strategies for updating the weights include measuring the similarity or dissimilarity between clusters across solutions and adjusting the weights accordingly.

The consensus clustering solution obtained from WCC provides a robust representation of the underlying clusters in the data by integrating multiple clustering perspectives. It can help identify stable and reliable clusters that are consistent across different clustering algorithms or parameter settings. [13]

2.1.8 Hartigan's Rule of Thumb

Hartigan proposed a heuristic method to identify well-separated clusters in a dataset. The method suggests that if there are K* well-separated clusters, then an optimal K+1 cluster partition should be a K cluster partition with one of its clusters split in two. This split would drastically decrease W_K, as the split parts are well-separated. W_K is accumulative value of document dissimilarity level to the closest centroid. However, for K>=K*, W_K should stay the same. To compute Hartigan's statistic, denoted by H, increase K while keeping track of W_K and $W_K + 1$. H is given by

$$H = \left(\frac{W_K}{W_{K+1}} - 1 \right) (N - K - 1)$$

where N is the number of entities. The first K value at which H decreases to 10 or less is taken as the estimate of K*. Hartigan's rule was indirectly supported by the related Duda and Hart criterion and was found to perform well in experiments. Milligan and Cooper found that Hartigan's rule performed well, and Chiang and Mirkin found that it did surprisingly well in experiments involving non-spherical clusters generated with overlapping data. [14]

2.1.9 Silhouette Analysis

Silhouette Analysis is a technique used to assess the effectiveness of clusters produced by a clustering algorithm. It quantifies how well an object fits into its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, with a higher value indicating a strong match within its cluster and a poor match with neighboring clusters. The average silhouette score for all objects is used to gauge the overall clustering quality.

To calculate the silhouette score for each object i, the following formula is used:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Here, $a(i)$ represents the average distance from i to other points in the same cluster, and $b(i)$ is the smallest average distance from i to points in a different cluster.

A silhouette score near 1 suggests effective clustering, while a score near -1 indicates potential misclassification. A score around 0 implies overlapping clusters.

By computing the silhouette score for various k values (number of clusters), one can determine the number of clusters that optimizes the average silhouette score. [15]

2.1.10 Cross-Validation

Cross-validation is a technique used to assess how well a predictive model will generalize to an independent dataset. It involves partitioning the dataset into multiple subsets, training the model on a subset of the data, and then evaluating its performance on the remaining subset.

In k-fold cross-validation, the dataset is divided into k equal-sized subsets. The model is trained k times, each time using a different subset as the test set and the remaining subsets as the training set. The performance metrics are then averaged across all k iterations.

In leave-one-out cross-validation, each observation in the dataset is used as the test set once, with the remaining observations used as the training set. This process is repeated n times, where n is the number of observations in the dataset.

Cross-validation provides a more reliable estimate of model performance compared to traditional validation methods such as a single train-test split. It reduces the risk of overfitting by assessing the model's performance on multiple subsets of the data. The estimation method alone is implemented with k-means and silhouette index and as k-fold cross-validation. [16]

2.2 K-means

The k-means clustering algorithm is widely recognized as one of the most influential and commonly used data mining techniques. Despite its popularity, the algorithm faces certain challenges, including issues with the random initialization of centroids, which can lead to unexpected convergence. Additionally, the algorithm requires the number of clusters to be predefined, which can result in varying cluster shapes and outlier effects. Another fundamental limitation of k-means is its inability to handle different types of data effectively.[17]

Suppose we have a dataset $X = \{x_1, \ldots, x_N\}, x_n \in \mathbb{R}^d$. The M-clustering problem aims to divide this dataset into M disjoint subsets (clusters) $C_1, \ldots, C_M$, optimizing a clustering criterion. The most common criterion is the sum of squared Euclidean distances between each data point x_i and the centroid m_k (cluster center) of the subset C_k containing x_i. This criterion, known as clustering error, relies on the cluster centers $m_1, \ldots, m_M$:

$$E(m_1, \ldots, m_M) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(x_i \in C_k) ||x_i - m_k||^2,$$

where $I(X) = 1$ if X is true and 0 otherwise. The k-means algorithm seeks locally optimal solutions concerning the clustering error.[18]

The k-means clustering algorithm follows a simple iterative process:

1. Choose k initial cluster centers either randomly from the dataset or by defining them within the dataset's space.
2. Assign each data point to the nearest cluster center.
3. Update the cluster centers based on the current assignments.
4. Repeat steps 2 and 3 until convergence, typically defined by minimal changes in cluster assignments or cluster center positions.

Various variants of the k-means algorithm have been developed to enhance its performance. Some focus on improving the initial partition to increase the likelihood of finding the global minimum. Others allow for cluster splitting and merging based on predefined criteria, enabling the algorithm to converge to an optimal solution from any initial partition.[19]

Chapter 3

Methods

In this section, we outline the approach for evaluating the performance of clustering methods on simulated and real biomedical data. We describe the criteria and metrics used to assess clustering quality and discuss the methodology for comparing different methods.

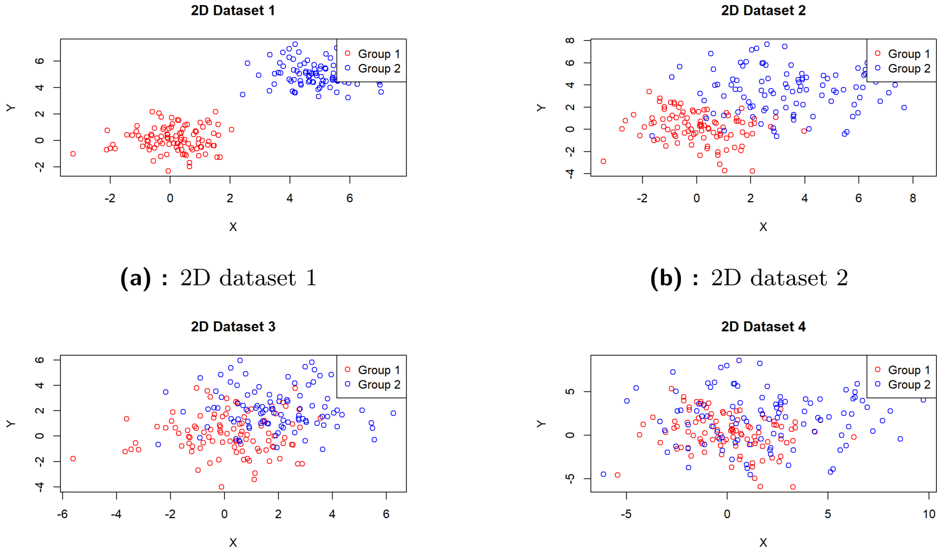
3.1 Evaluation using simulated data

To evaluate the clustering methods under controlled conditions, we generate synthetic datasets that mimic the characteristics of real biomedical data. These datasets are designed to vary in terms of inter-individual variability, noise levels, and cluster separability, allowing us to systematically assess the methods' robustness and sensitivity to different data scenarios.

The synthetic datasets consist of multidimensional feature vectors, where each feature represents a specific measurement or attribute. The dimensionality of the feature space varies across datasets, ranging from low-dimensional 2D representation to higher-dimensional datasets. This variation allows us to evaluate scalability and performance across different data complexities of clustering methods.

We assign each data point to a ground truth cluster based on the predefined cluster structures. These true cluster assignments serve as reference labels for evaluating clustering methods' accuracy and consistency in identifying underlying clusters.

We created four distinct datasets with two clusters in each dimension to evaluate clustering algorithms. We designed the clean dataset as benchmark 3.1a, featuring two well-separated clusters with distinct group means 0 and 5 and identical variance. The second dataset 3.1b has mean values of 0 and 3.5, with a variance of 2 for both clusters. The third dataset 3.1c has mean values of 0 and 2, each with a variance of 3. Finally, the fourth dataset 3.1d presents mean values of 0 and 2, with a variance of 5 for both clusters. Each dataset contains 100 observations per group, ensuring robust evaluations of algorithmic performance. To illustrate 10D datasets, the first 3.2a and the fourth datasets 3.2b were projected to 2D.



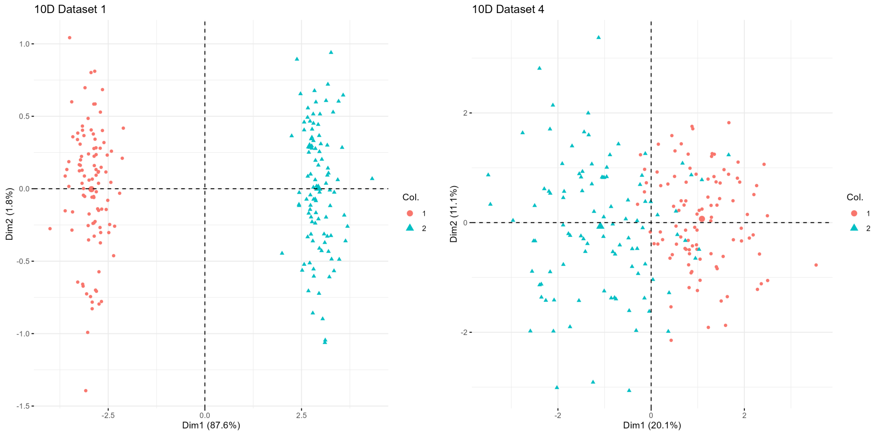
(a) : 2D dataset 1

(b) : 2D dataset 2

(c) : 2D dataset 3

(d) : 2D dataset 4

Figure 3.1: 2D datasets



(a) : 10D dataset 1

(b) : 10D dataset 4

Figure 3.2: PCA projections

3.1.1 Characterization of clusters

The evaluation of clustering algorithms often involves measuring how well the clusters are separated from each other and how compact the clusters are internally. One commonly used metric for evaluating clustering results is the interclass to between-class ratio, which measures the compactness of clusters relative to their separation. A lower ratio indicates better separability between the clusters, signifying that the clusters are more distinct and well-separated.

The ratio was calculated for each dataset as the sum of the average Euclidean distances from each data point to the centroid of its respective group, divided by the sum of distances between the centroids of the two groups.

The interclass to between-class ratio ($ratio$) can be calculated using the following equation:

$$ratio = \frac{\sum_{i=1}^{N_c} \sum_{j=1}^{N_i} \|x_{ij} - \mu_i\|}{\sum_{i=1}^{N_c} \|\mu_i - \mu\|}$$

where N_c is the number of clusters, N_i is the number of points in cluster i. x_{ij} represents the j-th point in cluster i and μ_i is the centroid of cluster i, μ is the centroid of all data points. $\|\cdot\|$ denotes the Euclidean distance.

The mean interclass between-class ratios were then calculated 3.1 over 100 iterations for each dataset type and dimensionality. These mean ratios summarize the separability of the clusters and describe how the datasets differ.

The parameters for variance and mean were carefully established to contain a wide range of ratios, ensuring comprehensive coverage of the data distribution.

Dimension	Dataset 1	Dataset 2	Dataset 3	Dataset 4
2D	0.24	0.46	0.76	0.83
3D	0.25	0.47	0.76	0.82
5D	0.26	0.48	0.76	0.83
10D	0.26	0.48	0.77	0.83

Table 3.1: Mean interclass between-class ratios for clean and noisy datasets

In the following analysis phase, the goal is to determine the optimal number of clusters for all datasets, considering that the ground truth number is two. Subsequently, the purpose was to compare the most effective estimation methods across all four datasets and dimensions.

Algorithm 1 Clustering Analysis

```
Define required functions
for each dimension $dim$ in $[2, 3, 5, 10]$ do
    for each method $m$ in methods do
        for $i$ from 1 to 1000 do
            Generate dataset $D_{\mathrm{x}}$ using $dim$
            Estimate clusters using method $m$: $k_{\mathrm{x}} \leftarrow method\_func(D_{\mathrm{x}})$
            Increment $C_{\mathrm{x}}[k_{\mathrm{x}}]$
        end for
    end for
end for
```

3.1.2 Simulation of more clusters

We generate datasets 3.3 with varying numbers of clusters to evaluate the performance of clustering methods. The purpose is to assess which methods consistently yield better results across datasets with increasing complexity in the number of clusters. The parameters include definitions for the centers of datasets with 3, 5, and 7 clusters. The datasets consist of 800 samples each, distributed among clusters located at specific coordinates in a two-dimensional space. For instance, the 3 clusters dataset comprises clusters located at (-3, -3), (3, 3), and (0, 0), while the 5 clusters dataset includes two additional clusters at (-5, 5) and (5, -5). The 7 clusters dataset adds clusters at (-7, 0) and (7, 0) to the existing five clusters.

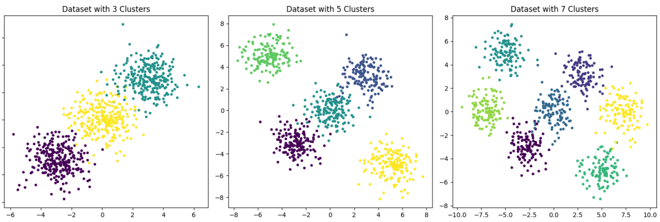


Figure 3.3: Datasets

The datasets exhibit similar inter-class to between-class ratios across varying numbers of clusters, it suggests a level of consistency in the dispersion of data points within and between clusters 3.2.

Number of clusters	Inter-class to Between-class Ratio
3	0.998
5	0.998
7	0.998

Table 3.2: Inter-class to Between-class Ratio for Datasets

3.2 Application on real data

The dataset was provided by the National Institute of Mental Health (NÚDZ) and contains detailed patient information. It includes demographic information such as unique identification codes, birth year, sex, nationality, marital status, and medical history. Additionally, it covers lifestyle factors like education, occupation, smoking, and alcohol consumption habits. The medical history section provides information on neurological, cardiovascular, respiratory, gastrointestinal, endocrine, and dermatological disorders and records related to COVID-19, such as confirmation methods, symptoms, hospitalization details, and long-term effects.

The following steps were taken for the analysis of COVID-related data: First, relevant data was selected from the dataset for exploratory analysis. Next, the data was preprocessed to handle missing values, and principal component analysis (PCA) was applied. PCA was chosen to reduce the dimensionality of the dataset while preserving its important features. Since the dataset includes binary variables, PCA components enable clustering by capturing the variability in the data. The variance of the PCA components was examined, and relevant components were selected for further analysis. Subsequently, cluster estimation was performed using selected methods based on the results obtained from simulated datasets. The clustering was conducted according to the outcomes of the analysis.

Chapter 4

Results

4.1 Simulated data

The results chapter for simulated data presents the outcomes of various methods used to predict the number of clusters when the true number is 2, across different dimensionalities. Each column corresponds to a different method.

For chosen dimensionalities as 2, 3, 5, and 10, the methods used include the Elbow Method, Silhouette Analysis (SA), Calinski-Harabasz Index (C-H), Davies-Bouldin Index (D-B), Weighted Consensus Clustering (WC), Hartigan's Rule of Thumb (HR), Cross-Validation (CV), Gap Statistics (Gap), Bayesian Information Criterion (BIC), and Akaike Information Criterion (AIC). The table displays the number of predicted clusters by each method for k (k pred) values ranging from 1 to 10.

The results show that different clustering methods provide varying predictions for the number of clusters in the datasets. The Elbow Method consistently suggested 10 clusters for all datasets, indicating a potential limitation in identifying the correct number of clusters, possibly due to its inability to handle noisy data and unclear cluster boundaries.

On the other hand, Silhouette Analysis (SA), Calinski-Harabasz Index (C-H), and Davies-Bouldin Index (D-B) consistently predicted two clusters for most datasets, suggesting their effectiveness in determining the correct number of clusters. However, their predictions varied for the last noisy dataset, particularly for SA and C-H.

Davies-Bouldin Index (D-B) performed well for datasets with well-separated clusters but showed increased variability in predicting cluster numbers for datasets with more complex distributions and higher dimensions.

Weighted Consensus Clustering (WC) showed poor results for all datasets, indicating its limitations in capturing the underlying data structure. Similarly, Hartigan's Rule of Thumb (HR) and Cross-Validation (CV) showed inadequate performance across all datasets, suggesting their ineffectiveness in determining our datasets' optimal number of clusters.

Gap Statistics (Gap) performed well for most datasets, except the last one, highlighting its robustness across various dimensionalities.

Dimensionality $ndim = 2$										
k pred	Elbow	SA	C-H	D-B	WC	HR	CV	Gap	BIC	AIC
1	0	0	0	0	0	1000	1000	0	0	0
2	0	1000	1000	1000	0	0	0	1000	1000	866
3	0	0	0	0	0	0	0	0	0	58
4	0	0	0	0	0	0	0	0	0	12
5	0	0	0	0	0	0	0	0	0	8
6	0	0	0	0	0	0	0	0	0	10
7	0	0	0	0	0	0	0	0	0	8
8	0	0	0	0	0	0	0	0	0	9
9	1	0	0	0	0	0	0	0	0	16
10	999	0	0	0	1000	0	0	0	0	13
Dimensionality $ndim = 3$										
1	0	0	0	0	0	1000	1000	0	0	0
2	0	1000	1000	1000	0	0	0	999	1000	818
3	0	0	0	0	0	0	0	1	0	32
4	0	0	0	0	0	0	0	0	0	12
5	0	0	0	0	0	0	0	0	0	19
6	0	0	0	0	0	0	0	0	0	28
7	0	0	0	0	0	0	0	0	0	35
8	0	0	0	0	0	0	0	0	0	38
9	1	0	0	0	0	0	0	0	0	45
10	999	0	0	0	1000	0	0	0	0	73
Dimensionality $ndim = 5$										
1	0	0	0	0	0	1000	1000	0	0	0
2	0	1000	1000	1000	0	0	0	949	1000	623
3	0	0	0	0	0	0	0	43	0	42
4	0	0	0	0	0	0	0	8	0	25
5	0	0	0	0	0	0	0	0	0	20
6	0	0	0	0	0	0	0	0	0	35
7	0	0	0	0	0	0	0	0	0	37
8	0	0	0	0	0	0	0	0	0	55
9	3	0	0	0	2	0	0	0	0	72
10	997	0	0	0	998	0	0	0	0	91
Dimensionality $ndim = 10$										
1	0	0	0	0	0	1000	1000	0	0	0
2	0	1000	1000	1000	0	0	0	772	999	504
3	0	0	0	0	0	0	0	130	1	15
4	0	0	0	0	0	0	0	45	0	2
5	0	0	0	0	0	0	0	23	0	2
6	0	0	0	0	0	0	0	15	0	19
7	0	0	0	0	0	0	0	9	0	57
8	0	0	0	0	0	0	0	3	0	61
9	6	0	0	0	3	0	0	3	0	129
10	994	0	0	0	997	0	0	0	0	211

Table 4.1: Cluster Evaluation For Dataset 1

k pred	Elbow	SA	C-H	D-B	WC	HR	CV	Gap	BIC	AIC
Dimensionality $ndim = 2$										
1	0	0	0	0	0	1000	942	0	60	0
2	0	1000	1000	1000	0	0	0	1000	940	864
3	0	0	0	0	0	0	0	0	0	44
4	0	0	0	0	0	0	0	0	0	13
5	0	0	0	0	0	0	0	0	0	8
6	0	0	0	0	0	0	0	0	0	12
7	0	0	0	0	0	0	0	0	0	14
8	0	0	0	0	0	0	0	0	0	11
9	6	0	0	0	2	0	0	0	0	12
10	994	0	0	0	998	0	58	0	0	22
Dimensionality $ndim = 3$										
1	0	0	0	0	0	1000	813	0	1	0
2	0	1000	1000	1000	0	0	0	998	999	653
3	0	0	0	0	0	0	0	2	0	53
4	0	0	0	0	0	0	0	0	0	18
5	0	0	0	0	0	0	0	0	0	20
6	0	0	0	0	0	0	0	0	0	30
7	0	0	0	0	0	0	0	0	0	49
8	0	0	0	0	0	0	0	0	0	42
9	2	0	0	0	2	0	0	0	0	59
10	998	0	0	0	998	0	187	0	0	76
Dimensionality $ndim = 5$										
1	0	0	0	0	0	1000	726	0	0	0
2	0	1000	1000	1000	0	0	0	939	1000	612
3	0	0	0	0	0	0	0	51	0	46
4	0	0	0	0	0	0	0	9	0	24
5	0	0	0	0	0	0	0	1	0	29
6	0	0	0	0	0	0	0	0	0	25
7	0	0	0	0	0	0	0	0	0	42
8	0	0	0	0	0	0	0	0	0	50
9	4	0	0	0	4	0	0	0	0	65
10	996	0	0	0	996	0	274	0	0	107
Dimensionality $ndim = 10$										
1	0	0	0	0	0	1000	760	0	500	0
2	0	1000	1000	1000	0	0	0	745	497	458
3	0	0	0	0	0	0	0	117	3	13
4	0	0	0	0	0	0	0	79	0	3
5	0	0	0	0	0	0	0	27	0	8
6	0	0	0	0	0	0	0	18	0	20
7	0	0	0	0	0	0	0	9	0	71
8	0	0	0	0	0	0	0	2	0	66
9	2	0	0	0	3	0	0	3	0	135
10	998	0	0	0	997	0	240	0	0	226

Table 4.2: Cluster Evaluation For Dataset 2

Dimensionality $ndim = 2$										
k pred	Elbow	SA	C-H	D-B	WC	HR	CV	Gap	BIC	AIC
1	0	0	0	0	0	1000	0	170	1000	809
2	0	662	477	1	0	0	0	829	0	96
3	0	83	67	1	0	0	0	1	0	19
4	0	51	116	28	0	0	0	0	0	12
5	0	28	54	40	0	0	0	0	0	7
6	0	36	56	87	0	0	0	0	0	7
7	0	31	38	101	0	0	0	0	0	8
8	0	34	60	166	0	0	0	0	0	10
9	2	29	46	223	4	0	0	0	0	17
10	998	46	86	353	996	0	1000	0	0	15
Dimensionality $ndim = 3$										
1	0	0	0	0	0	1000	0	10	1000	612
2	0	984	998	5	0	0	0	933	0	77
3	0	3	0	1	0	0	0	57	0	17
4	0	0	0	2	0	0	0	0	0	14
5	0	3	0	27	0	0	0	0	0	24
6	0	2	0	52	0	0	0	0	0	28
7	0	3	0	110	0	0	0	0	0	40
8	0	2	0	177	0	0	0	0	0	44
9	3	2	0	215	6	0	0	0	0	61
10	997	1	0	411	994	0	1000	0	0	83
Dimensionality $ndim = 5$										
1	0	0	0	0	0	1000	0	0	1000	374
2	0	1000	1000	518	0	0	0	831	0	111
3	0	0	0	0	0	0	0	138	0	27
4	0	0	0	0	0	0	0	22	0	26
5	0	0	0	0	0	0	0	7	0	33
6	0	0	0	3	0	0	0	1	0	53
7	0	0	0	16	0	0	1	0	0	68
8	0	0	0	36	0	0	0	0	0	78
9	10	0	0	101	4	0	0	0	0	98
10	990	0	0	326	996	0	1000	0	0	132
Dimensionality $ndim = 10$										
1	0	0	0	0	0	1000	0	0	1000	330
2	0	1000	1000	998	0	0	0	693	0	169
3	0	0	0	2	0	0	0	156	1	19
4	0	0	0	0	0	0	0	77	0	11
5	0	0	0	0	0	0	0	26	0	7
6	0	0	0	0	0	0	0	22	0	27
7	0	0	0	0	0	0	0	8	0	53
8	0	0	0	0	0	0	0	13	0	46
9	14	0	0	0	12	0	0	5	0	90
10	986	0	0	0	988	0	1000	0	0	249

Table 4.3: Cluster Evaluation For Dataset 3

Dimensionality $ndim = 2$										
k pred	Elbow	SA	C-H	D-B	WC	HR	CV	Gap	BIC	AIC
1	0	0	0	0	0	1000	0	529	1000	876
2	0	293	128	0	0	0	0	467	0	36
3	0	199	121	3	0	0	0	4	0	23
4	0	97	140	26	0	0	0	0	0	3
5	0	49	72	30	0	0	0	0	0	6
6	0	74	99	92	0	0	0	0	0	7
7	0	60	382	125	0	0	0	0	0	6
8	0	59	97	159	0	0	0	0	0	9
9	1	73	97	219	1	0	0	0	0	8
10	999	96	164	346	999	0	1000	0	0	26
Dimensionality $ndim = 3$										
1	0	0	0	0	0	1000	0	125	1000	667
2	0	744	904	0	0	0	0	751	0	42
3	0	30	56	0	0	0	0	120	0	14
4	0	25	31	8	0	0	0	4	0	9
5	0	35	8	39	0	0	0	0	0	17
6	0	37	1	64	0	0	0	0	0	23
7	0	24	0	118	0	0	0	0	0	36
8	0	29	0	146	0	0	0	0	0	34
9	3	32	0	203	8	0	0	0	0	75
10	997	44	0	422	992	0	1000	0	0	83
Dimensionality $ndim = 5$										
1	0	0	0	0	0	1000	0	4	1000	491
2	0	992	994	7	0	0	0	714	0	27
3	0	2	6	1	0	0	0	203	0	24
4	0	0	0	0	0	0	0	52	0	34
5	0	1	0	4	0	0	0	17	0	35
6	0	1	0	15	0	0	0	6	0	39
7	0	0	0	67	0	0	1	3	0	59
8	0	2	0	111	0	0	0	1	0	77
9	3	1	0	232	7	0	0	0	0	90
10	997	1	0	563	993	0	1000	0	0	124
Dimensionality $ndim = 10$										
1	0	0	0	0	0	1000	0	0	1000	606
2	0	1000	999	754	0	0	0	587	0	12
3	0	0	1	2	0	0	0	184	0	6
4	0	0	0	0	0	0	0	96	0	2
5	0	0	0	0	0	0	0	45	0	6
6	0	0	0	1	0	0	0	25	0	10
7	0	0	0	4	0	0	0	28	0	46
8	0	0	0	3	0	0	0	23	0	41
9	17	0	0	34	11	0	0	12	0	83
10	983	0	0	204	989	0	1000	0	0	188

Table 4.4: Cluster Evaluation For Dataset 4

Bayesian Information Criterion (BIC) was effective for datasets with well-separated clusters but showed limitations for more complicated distributions. Akaike Information Criterion (AIC) exhibited high variance in predicted cluster numbers but consistently performed well across all datasets.

The table 4.5 contains the results of the evaluation for datasets with more clusters for multiple clustering evaluation methods. Silhouette Analysis, Calinski-Harabasz Index, Davies-Bouldin Index, and Gap statistics consistently suggest the number of clusters that match the true number in the simulated datasets, indicating their reliability in estimating cluster numbers. However, Bayesian Information Criterion (BIC) occasionally diverges, particularly when faced with datasets containing 5 clusters, where it predicts one extra cluster. This discrepancy suggests potential sensitivity to dataset characteristics.

Method	**Predicted Number of Clusters**		
	3 Clusters	**5 Clusters**	**7 Clusters**
Silhouette Analysis	3	5	7
Calinski-Harabasz Index	3	5	7
Davies-Bouldin Index	3	5	7
Bayesian Information Criterion	3	6	7
Gap Statistics	3	5	7

Table 4.5: Predicted Number of Clusters by Evaluation Method

For real dataset evaluation, it's crucial to select cluster evaluation methods that demonstrate consistency and robustness across different datasets and dimensionalities. Based on the observations:

1. **Silhouette Analysis (SA):** SA consistently predicted two clusters for most datasets, indicating its stability and reliability in identifying the optimal number of clusters.

2. **Calinski-Harabasz Index (C-H):** Similar to SA, C-H consistently predicted two clusters for most datasets, suggesting its effectiveness in capturing the underlying data structure.

3. **Davies-Bouldin Index (D-B):** Despite its variability in predicting cluster numbers for datasets with complex distributions, D-B exhibited good performance for datasets with well-separated clusters. Its ability to measure the average similarity between each cluster and its most similar cluster makes it a valuable tool for evaluating real datasets, especially those with clear cluster boundaries.

4. **Bayesian Information Criterion (BIC):** BIC worked well for datasets with well-separated clusters, showcasing its ability to penalize complex models. However, it may exhibit high variance in predicting cluster

numbers across different datasets. Nonetheless, its stability across various dimensionalities makes it a valuable addition to the evaluation process for real datasets.

5. **Gap Statistics (Gap):** Gap Statistics performed well for most datasets, except the last one, but is robust across various dimensionalities.

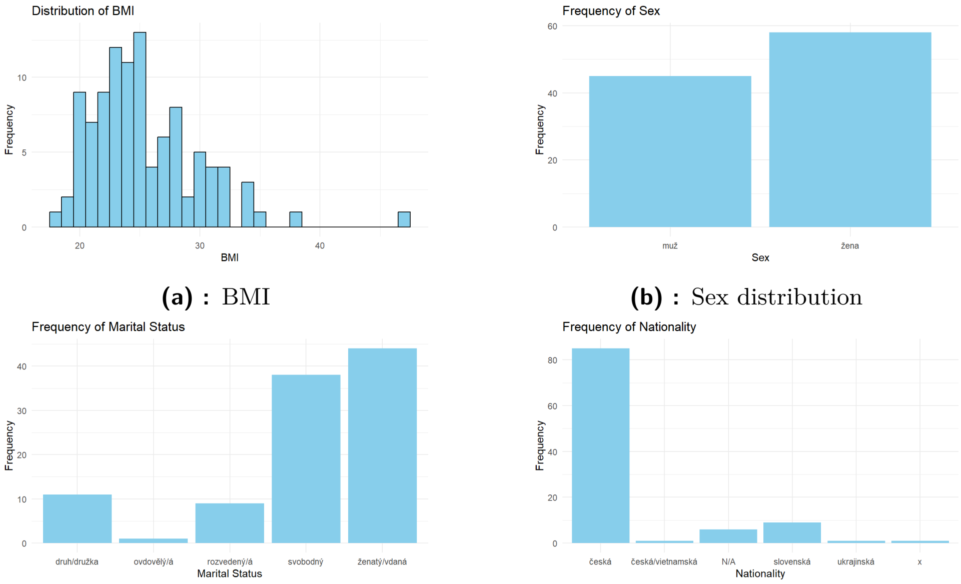
Based on their consistency, stability, and ability to capture the underlying data structure, Silhouette Analysis (SA), Calinski-Harabasz Index (C-H), Davies-Bouldin Index (D-B), Bayesian Information Criterion (BIC), and Gap Statistics (Gap) are chosen for real dataset evaluation.

4.2 Real Covid data

4.2.1 Exploratory Data Analysis

The EDA analysis was performed on the provided COVID dataset using various visualization techniques. This included boxplots and distribution plots to explore the distribution of age at the first visit, segmented by sex 4.2, marital status, and nationality 4.1. Then the correlation of COVID-related variables was calculated to avoid processing correlated data. A boxplot 4.2 provides information of the central tendency and spread of age within each sex category.

Distribution plot was created to explore the distribution of Body Mass Index (BMI) 4.1a values across the dataset, identifying trends and outliers in BMI values.



(a) : BMI

(b) : Sex distribution

(c) : Marital status

(d) : Nationality

Figure 4.1: EDA results

We plot the correlation graph 4.3 to visualize the correlation among individual categories. Except for the PCR and AG results, the highest correlation observed among all COVID-related categories in the dataset is 0.26. This suggests that the variables are not strongly correlated, which reduces the risk of multicollinearity. Multicollinearity leads to instability and unreliable estimates in clustering algorithms. Having moderate correlations helps to obtain more reliable clustering results.

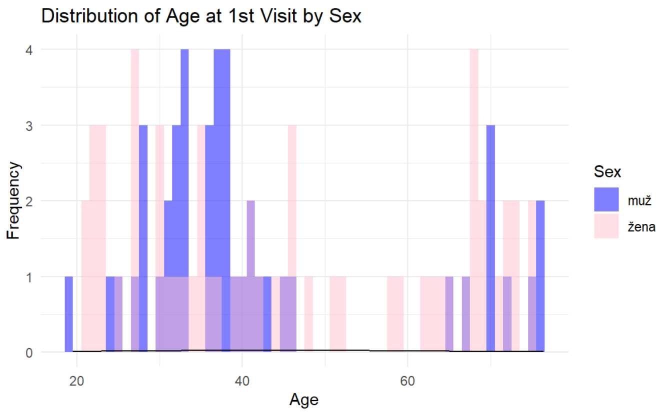


Figure 4.2: Age distribution plot

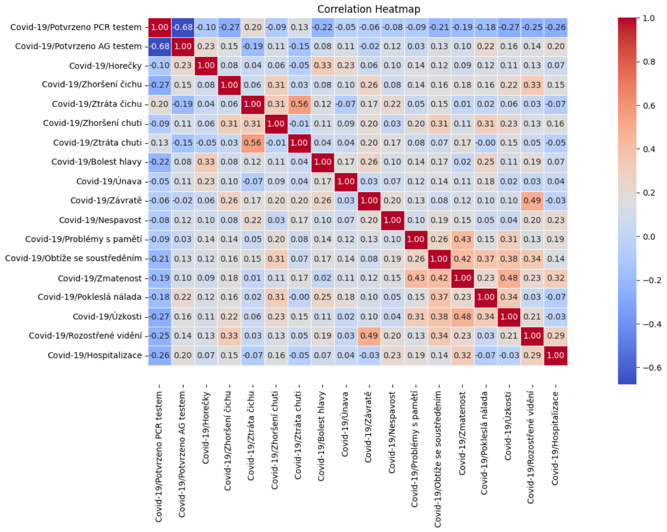


Figure 4.3: Correlation

4.2.2 Data preprocessing

The dataset was preprocessed to ensure it was ready for further analysis. This involved extracting columns related to COVID-19. Subsequently, any columns with missing values (highlighted in yellow on the heatmap 4.4 data graph) were removed to maintain data integrity. The yellow columns contained additional data that were filled in with questions just in case of hospitalization. Additionally, duplicate rows were carefully identified and eliminated to prevent redundancy in the dataset. One missing value in the column named "tinnitus" was replaced by zero.

During data preprocessing, we noticed that the sum of binary variables across each column identified one column representing hospitalization, which was positive for just one patient. As a result, this column and the corresponding patient were excluded from further analysis, and the entire ’Hospitalization’ column was not used. Additionally, only symptoms were included for PCA,

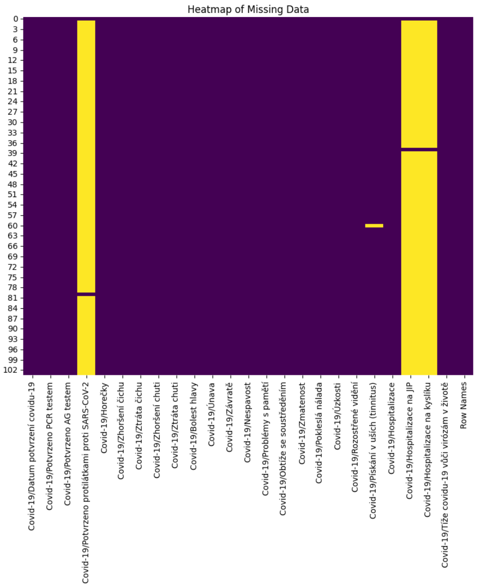


Figure 4.4: Missing values

omitting the columns 'Potvrzeno PCR testem' (Confirmed by PCR test) and 'Potvrzeno AG testem' (Confirmed by AG test) from the dataset. In total, we got 16 different symptoms.

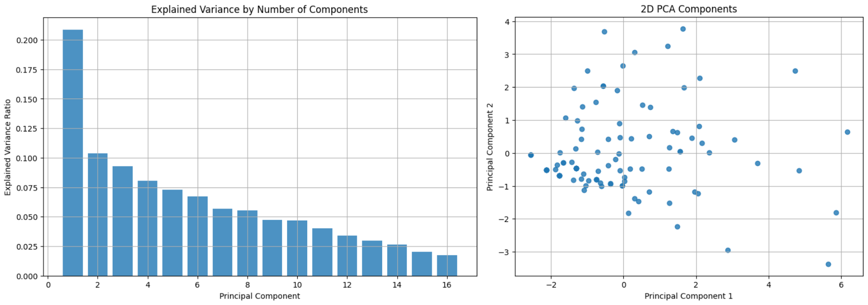
Principal component analysis (PCA) was then performed on the modified dataset, including only symptom variables. This preprocessing step ensured that the PCA analysis focused solely on symptom-related features, excluding variables related to hospitalization confirmation tests.

4.2.3 Principal component analysis

In the graph of PCA explained variance of the data 4.5b, it is clear that the first principal component capture the largest variance. Despite the dominance of the first component, the significant variance explained by other components capture the multidimensional structure. This observation highlights the importance of considering multiple dimensions when analyzing and interpreting the COVID-19 dataset.

The projection of the patient data onto the first two principal components reveals a representation of the dataset's structure. This projection preserves the most significant variability in the original multidimensional data while reducing its dimensionality for visualization and analysis. The visualization facilitates the exploration of patient similarities, differences, and groupings and shows factors of variability in the dataset. According to the explained variance plot, only eight components were included for further clustering, with each component explaining at least five percent of the variance.

The table 4.6 illustrates the PCA loadings, with each row representing a different Covid-19 symptom and each column corresponding to a principal



(a) : PCA Explained Variance

(b) : Projection of data

Figure 4.5: PCA results

component. The values in the table indicate the correlation between each symptom and principal component. Examining the loadings enables us to interpret the principal components concerning the original variables.

Component	1	2	3	4	5	6	7	8
Fever	0.164	-0.118	-0.327	-0.384	0.078	-0.351	-0.227	-0.058
Worsening of Smell	0.254	0.039	-0.177	0.366	-0.212	0.256	-0.183	-0.330
Loss of Smell	0.191	0.501	0.266	-0.279	-0.269	-0.001	-0.076	0.048
Worsening of Taste	0.267	-0.014	0.060	0.028	-0.691	-0.012	-0.044	-0.134
Loss of Taste	0.175	0.461	0.315	-0.219	0.173	0.223	-0.202	0.014
Headache	0.224	0.045	-0.402	-0.312	-0.019	-0.146	-0.120	0.483
Fatigue	0.143	-0.169	-0.264	-0.371	0.085	0.217	-0.026	-0.637
Dizziness	0.275	0.333	-0.250	0.296	0.065	-0.172	-0.117	-0.014
Insomnia	0.164	0.300	0.002	-0.174	0.253	-0.119	0.589	-0.256
Memory Problems	0.267	-0.191	0.180	-0.066	0.180	-0.095	-0.442	-0.064
Concentration Prob.	0.354	-0.173	0.152	-0.018	-0.043	-0.141	0.418	0.027
Confusion	0.315	-0.226	0.331	0.072	0.363	0.071	-0.067	-0.154
Depressed Mood	0.301	-0.278	-0.055	-0.145	-0.227	0.237	0.319	0.253
Anxiety	0.345	-0.237	0.256	0.128	0.089	-0.002	-0.068	0.176
Blurred Vision	0.280	0.173	-0.254	0.438	0.174	-0.250	0.098	-0.001
Tinnitus	0.111	0.091	-0.310	0.022	0.194	0.705	0.028	0.249

Table 4.6: Loadings for each PCA component

Component 1 primarily contains symptoms commonly associated with neurological and cognitive manifestations of COVID-19. It includes symptoms like worsening taste, headache, dizziness, memory problems, concentration problems, and anxiety. Component 2 demonstrates a distinct pattern characterized by a strong correlation with symptoms related to the loss of smell and taste. In Component 3, symptoms such as fever, headache, and confusion are prevalent. Component 4 displays notable positive correlations with symptoms, including fever, blurred vision, and fatigue. Symptoms related to cognitive function, particularly worsening of taste and confusion, are strongly associated with Component 5. Component 6 characterizes the less prevalent symptom of tinnitus. Component 7 includes symptoms such as insomnia, memory, and concentration problems, suggesting sleep disturbances and cognitive function issues. Lastly, Component 8 is characterized by symptoms of headache and fatigue, which are prevalent and often reported symptoms in COVID-19 cases.

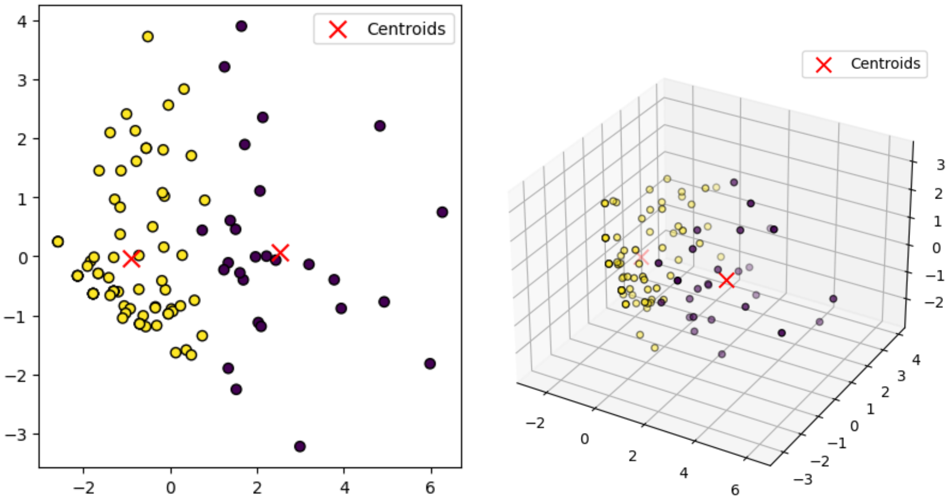
4.2.4 Number of cluster estimation

In clustering analysis, determining the optimal number of clusters is a crucial step that directly influences the quality of the resulting clusters. Here, we discuss the results obtained from different methods applied to a dataset of 102 patients.

The estimated number of clusters using various metrics is as follows:

The Silhouette Score and Calinski-Harabasz Score both suggest an optimal number of clusters as 2, indicating good cohesion and separation. Conversely, the Davies-Bouldin Score predicts a value of 10, suggesting poorer cluster quality. The Bayesian Information Criterion (BIC) indicates a higher optimal number of clusters at 9, while the Gap Statistics also suggest 2 clusters as optimal.

These results may indicate some ambiguity in determining the exact number of clusters, with variations across different metrics. However, the most reasonable choice seems to be 2 clusters. K-means was performed on the data 4.6.



(a) : Clustering with 2 Clusters (2D) **(b) :** Clustering with 2 Clusters (3D)

Figure 4.6: KMeans

4.2.5 Characterization of clusters

The sizes of the clusters were not the same; the first cluster contains 76 patients, whereas the second cluster sontains 26 patients. Despite this discrepancy in size, the interclass-to-between-class ratio, which serves as a measure of the compactness and separation of clusters, was calculated at 0.53. This value closely resembled the ratio observed in the second simulated dataset, indicating a comparable clustering structure in terms of cluster cohesion and dispersion.

Comparing this ratio to an alternative configuration with 10 clusters, where the interclass-to-between-class ratio reached 0.91, highlights the superiority of the 2-cluster configuration. A higher interclass-to-between-class ratio for clusters with ten centers suggests that the clusters are less well-separated and may overlap to some extent. Conversely, the lower interclass-to-between-class ratio for clusters with two centers indicates better separation and distinctiveness among the clusters.

Analysis of Principal component

The table 4.7 presents the means and standard deviations for each principal component in two clusters.

Comp	Claster 1 mean	Std	Cluster 2 mean	Std	t-statistics	p-value
1	-0.86	0.85	2.52	1.50	-10.694	7.8×10^{-12}
2	-0.03	1.11	0.09	1.73	-0.322	0.750
3	0.02	1.02	-0.05	1.70	0.169	0.867
4	-0.08	1.03	0.23	1.38	-1.019	0.315
5	0.08	0.85	-0.23	1.57	0.945	0.352
6	0.09	0.94	-0.26	1.22	1.315	0.197
7	-0.05	0.91	0.15	1.08	-0.825	0.415
8	0.01	0.93	-0.04	0.95	0.243	0.810

Table 4.7: Means and Standard Deviations for Each PCA Component in Clusters

We used Welch's t-test to evaluate the differences, as the variances differ. The results indicate that the first component significantly differs for clusters and is the most distinguishing factor. It is also associated with common COVID-19 symptoms and emotional factors. The lower mean for cluster 1 suggests a weaker association than cluster 2, with a substantially higher mean value. The other components do not show significant mean differences between the two clusters.

Analysis of Unused Features

In order to analyze unused features, we performed calculations to determine the percentage of vaccinated individuals within different clusters and the percentage of males. We then used the chi-squared test to find out any differences. The chi-squared test is a statistical method used to evaluate the association between categorical variables. Similar analyses were conducted for other parameters, except that a t-test was used instead of the chi-squared test. We focused on characterizing the data for exploratory analysis, so we didn't perform corrections for multiple comparisons.

We can observe a significant difference based on the table 4.8. The second cluster contains more females and has a higher average severity of COVID-19. The subjective perception of the progression of the disease compared to

Characteristic	Cluster 1	Cluster 2	statistic	p-value
COVID Vaccinated (%)	91.53	100.00	3.07	0.0027
Mean BMI	25.16	26.62	-1.40	0.16
Average Age	44.05	42.95	0.28	0.78
Male (%)	52.63	19.23	7.46	0.0062
Average severity of COVID-19	5.22	7.17	-3.63	0.00045
Size	76	26		

Table 4.8: Cluster Characteristics

others, for instance, flu, describes this severity. Despite the cluster having a significantly higher vaccination rate, its members suffer from more symptoms explained by component 1 in PCA. These symptoms include worsening taste, headache, dizziness, memory problems, concentration problems, and anxiety. In conclusion, the clustering identifies a subset of female patients experiencing anxiety and perceiving COVID-19 more severely than the rest of the patients.

Chapter 5

Discussion

The analysis of simulated data provides an overview of the performance of various clustering methods across different dimensionalities. It is evident that these methods offer varying predictions for the number of clusters in the datasets. The results chapter presents the outcomes of various methods used to predict the number of clusters for simulated datasets with 2 true clusters, across different dimensionalities. While the Elbow Method consistently suggested 10 clusters, Silhouette Analysis, Calinski-Harabasz Index, and Davies-Bouldin Index consistently predicted 2 clusters, showing their effectiveness. Weighted Consensus Clustering, Hartigan's Rule of Thumb, and Cross-Validation showed poor performance. Gap Statistics performed well, except for the last dataset. Bayesian Information Criterion was effective but occasionally diverged. Akaike Information Criterion exhibited high variance.

For datasets with more clusters, Silhouette Analysis, Calinski-Harabasz Index, Davies-Bouldin Index, and Gap statistics consistently suggested the correct number of clusters, with Bayesian Information Criterion occasionally diverging, especially for datasets with 5 clusters.

These observations underscore the significance of selectively choosing clustering methods based on the dataset's characteristics. Methods such as Silhouette Analysis and Calinski-Harabasz Index demonstrate stability and reliability in identifying the optimal number of clusters, particularly in datasets with distinct cluster boundaries.

The COVID data analysis section provides a comprehensive examination of the dataset through exploratory data analysis (EDA), data preprocessing, principal component analysis, and cluster estimation. Initially, the EDA phase uses various visualization techniques such as boxplots and distribution plots to explore age distribution segmented by sex, marital status, and nationality, followed by correlation analysis of COVID-related variables. Subsequently, data preprocessing ensures data integrity by removing missing values, duplicates, and irrelevant columns, preparing the dataset for PCA. The PCA stage focuses on extracting meaningful features and reducing dimensionality, with emphasis on interpreting principal components and their associations with

COVID symptoms. The number of clusters is estimated using multiple metrics, including the Silhouette Score, Calinski-Harabasz Score, Davies-Bouldin Score, Bayesian Information Criterion, and Gap Statistics. Finally, cluster characterization examines the composition and characteristics of identified clusters, revealing significant differences in symptomatology, severity, and vaccination rates between clusters.

In summary, these findings highlight the importance of using robust evaluation methods and carefully considering data characteristics in clustering analysis.

Chapter 6

Conclusion

The assessment of clustering methods on simulated data emphasized the importance of choosing appropriate evaluation criteria based on dataset characteristics. While some methods such as the Elbow Method consistently showed limitations, others like Silhouette Analysis, Calinski-Harabasz Index, and Davies-Bouldin Index proved to be effective in identifying the correct number of clusters, with varying performance for complex datasets.

For the evaluation of real datasets, Silhouette Analysis, Calinski-Harabasz Index, Davies-Bouldin Index, Bayesian Information Criterion, and Gap Statistics emerged as suitable options based on their performance across different dimensionalities. Our study used exploratory data analysis, data preprocessing, principal component analysis, and number of clusters estimation. We could identify meaningful patterns and cluster structures within the dataset through these methodologies. The clustering analysis revealed significant differences in symptomatology, severity, and vaccination rates between identified clusters, providing valuable insights into patient subgroups' characteristics and progression. This study highlights the critical role of clustering analysis in uncovering hidden structures and patterns within complex datasets such as COVID-19 patient data. Researchers and healthcare professionals can gain deeper insights into disease progression, patient heterogeneity, and treatment outcomes using various techniques and evaluation metrics.

Overall, the study offers valuable insights into the strengths and limitations of various clustering evaluation methods, applied both for simulated and real-world datasets.

Bibliography

[1] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[2] SHAHLA Mousavi, F Zamani Boroujeni, and SAEED Aryanmehr. Improving customer clustering by optimal selection of cluster centroids in k-means and k-medoids algorithms. *Journal of Theoretical and Applied Information Technology*, 98(18):3807–3814, 2020.

[3] T. Soni Madhulatha. An overview on clustering methods, 2012.

[4] Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and S. Sarasvady. Dbscan: Past, present and future. In *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, pages 232–238, 2014.

[5] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[6] Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, and Muljono. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 533–538, 2018.

[7] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[8] Mojgan Mohajer, Karl-Hans Englmeier, and Volker J Schmid. A comparison of gap statistic definitions with and without logarithm function. *arXiv preprint arXiv:1103.4767*, 2011.

[9] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[10] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[11] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[12] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[13] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52:91–118, 2003.

[14] Boris Mirkin. Choosing the number of clusters. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):252–260, 2011.

[15] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[16] Daniel Berrar et al. Cross-validation., 2019.

[17] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

[18] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[19] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.