Università
degli Studi
di Padova

Dipartimento
di Ingegneria
dell'Informazione

Master Thesis in ICT for Internet and Multimedia

Clustering Patients using Longitudinal Data

Master Candidate

Pargol Golmohammadi

Student ID 2005932

Supervisors

Prof. Saikat Chatterjee

KTH University, Sweden

Prof. Federica Battisti

University of Padova, Italy

Co-supervisor

Ashish Kumar

Karolinska Institute, Sweden

Academic Year
2023/2024

To my husband
and parents

Abstract

The management of longitudinal datasets in the context of clinical research, particularly in the presence of missing data, is a complex and diverse task that requires meticulous deliberation. Longitudinal datasets are very important in the context of evaluating disease development and treatment success due to their ability to record information over multiple time points. Nonetheless, the occurrence of missing data might be attributed to a range of factors, including patient dropping out, irregular follow-up, or technical errors. In order to tackle this problem, researchers often use advanced statistical methodologies such as imputation methods, which we have used in this work to handle missing data. In our case, we worked on longitudinal height and weight data of 3897 patients between 0 to 24 years old and the missing data ratio of our dataset was around 35%. As we wanted to get the BMIs of the patients and cluster them, at first we replaced these missing data with different imputation approaches, and according to the obtained results, we chose the Mean Expected Growth approach and then calculated the BMIs of the patients. Choosing the best clustering method depends on the nature and distribution of data and the problem definition and requirements raised in a project. In this research, the Gaussian Mixture Model (GMM) was selected as the clustering algorithm due to the Gaussian distribution of the data. The objective was to comprehend the dynamic changes in patient clusters using a novel forgetting factor approach in the context of longitudinal data to identify age-adjusted BMI growth trajectories. Forgetting factor is an approach used in time-series analysis and forecasting that involves assigning weights to previous data that decrease exponentially with time and analyzes previous observations' effect on future outcomes. Our dataset had a very high percentage of missing data, therefore we chose to cluster the data in two different ways. In the first scenario, we separated the data that did not have missing data, performed clustering on them, and considered it as a gold standard. Then, in the second scenario, we imputed the missing data and performed clustering on the entire dataset. By focusing on early life factors such as gestational smoking, lactation, and pre–gestational and gestational BMI control, our findings contribute additional evidence to the OECD guidance regarding high BMI risks and interventions (World Health Organization, 2016[20]).

Sommario

Contents

List of Figures

List of Tables

List of Acronyms

BMI Body Mass Index

GMM Gaussian Mixture Model

ILD Individual Learning Diagnostics

AI Artificial Intelligence

AR Adiposity Rebound

WHO World Health Organization

MCAR Missing Completely at Random

MAR Missing at Random

MNAR Missing Not at Random

LOCF Latest Observation Carried Forward

BOCF Baseline Observation Carried Forward

EM Expectation Maximization

MI Multiple Imputation

MICE Multivariate Imputation by Chained Equations

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

1 Introduction

In recent years, the widespread accessibility of longitudinal data has facilitated significant advancements in study across several disciplines, yielding useful perspectives on the progression and maturation of individuals. The analysis of longitudinal data has the potential to reveal significant patterns and trends, providing insights into the intricate relationship between physical characteristics and age-related changes. The analysis of longitudinal data holds significance in correlational studies that seek to establish connections between observations of identical variables over an extended duration. Examples of such studies include investigations into substance use or mental health in the field of psychology, recidivism behavior in sociology, and relapse or medication adherence in the realm of medicine. Longitudinal studies provide researchers the opportunity to evaluate and investigate the temporal fluctuations of the variables under investigation. Due to advancements in data collecting and storage, there has been a growing trend in the development of longitudinal studies that incorporate a substantial number of repeated measurements of a single variable per person over an extended period. When a substantial quantity of observations is included, the dataset is typically denoted as intense longitudinal data. The use of Individual Learning Diagnostics (ILD) offers the benefit of enabling a more detailed evaluation of changes occurring over a period, particularly when considering individual subjects. The use of appropriate models that account for the inherent structure of longitudinal data is essential for conducting effective analysis. The evaluation of variability is crucial since it is recognized that no two people exhibit similar characteristics. In addition to considering the existence of measurement variability within individual subjects, it is necessary for models to include variations that exist between individuals. In the investigation of medication adherence, it is possible to see significant variations in adherence levels among individuals during the duration of the study. An instance of such a modeling method is multilevel modeling [25]. An area of considerable interest is the examination of height and weight data obtained from individuals at developmental stages from infancy to early adulthood, which I worked on in this project. Our data set is very unique and has been collected by the Karolinska Institute in Sweden over 24 years and the height and weight of people have been measured during this period.

Artificial intelligence (AI) is a subfield within the discipline of computer science that aims to enhance the intelligence of computers. Learning is considered a fundamental prerequisite for the manifestation of intelligent behavior. The prevailing consensus among contemporary scholars is that the presence of learning is a necessary condition for the manifestation of intelligence. Machine learning is widely recognized as a prominent subject within the domain of artificial intelligence (AI). It is noteworthy that machine learning has seen substantial advancements, positioning it as one of the fastest-evolving subdisciplines of AI study. Medical dataset analysis began using machine learning methods. Machine learning offers various essential data analysis techniques nowadays. The digital revolution made data collection and storage affordable, especially in recent years. Modern hospitals have data-collecting equipment and big information systems. Machine learning is ideal for assessing medical data, especially minor, specialized diagnostic issues. Medical records at specialist hospitals or departments typically provide reliable diagnosis data. Simply enter patient data with proper diagnoses into computer software to execute a learning process. This is a simplification, but medical diagnostic information may be automatically extracted from prior instances. The generated classifier may then be used to help clinicians diagnose new patients faster, more accurately, and more reliably, or to teach trainees or non-specialist physicians to detect a specific diagnostic issue[15].

One of the most important methods in machine learning is clustering, which we have used in this project. Clustering is a widely used methodology in many domains such as data analysis, machine learning, and pattern recognition. It involves the grouping of a collection of objects or data points according to their shared similarities and differences. The objective of clustering is to generate clusters or groups in which the items belonging to the same cluster exhibit a higher degree of similarity among themselves compared to those in other clusters[3] [28]. Clustering is categorized as an unsupervised learning technique, indicating its independence from labeled data. It serves as a means of exploratory data analysis, facilitating the acquisition of insights and comprehension of the underlying structure within a given dataset[28]. There are different types of clustering methods such as Centroid-based Clustering, Density-based Clustering, Distribution-based Clustering, and Hierarchical Clustering[4]. In this work, I use the Gaussian Mixture Model (GMM) for Clustering which is a Distribution-based clustering method. GMM is predicated on the underlying premise that each cluster is derived from a combination of Gaussian distributions. GMM is a statistical model that is capable of capturing intricate patterns within a dataset and assigning each individual data point to a specific cluster with a matching probability score and it has the capability to effectively handle datasets that are not linearly separable or include overlapping clusters. This characteristic makes GMM a flexible and potent clustering technique[26].

1.1 Challenges

Within the framework of longitudinal studies, the matter of missing data acquires a central and crucial role, presenting complex issues that need careful and deliberate examination. In the context of these studies, every experimental or observational unit undergoes measurement at the initial stage and thereafter at various intervals during the study duration. Nevertheless, the prevalence of incomplete data is a frequent phenomenon, arising from several factors such as participant attrition or irregular involvement resulting in nonmonotone missing data patterns.

The field of longitudinal studies encompasses a wide range of disciplines, including clinical trials, quality-of-life studies, and environmental research, which together provide many instances of nonignorable missing data. In the context of health studies, it is possible for the occurrence of treatment side effects to influence the level of participation among participants. This may lead to a non-random pattern of missing data, which is directly associated with the outcome being studied. In the context of quality of life research, the level of adherence shown by participants may be subject to impact from their prognosis, hence introducing additional complexity to the structure of missing data[11, 21].

The presence of missing data within a longitudinal dataset has the potential to introduce bias and inaccuracies in clustering results. The presence of bias in cluster assignments, altered computations of similarity, decreased ability to detect patterns, and poor comprehension of temporal dynamics may arise as consequences. The process of imputing missing data presents difficulties and has the potential to impact the outcomes. These factors result in a reduction in statistical power, an increase in uncertainty, and a consequential impact on the interpretation of the results. It is vital to utilize specialized approaches and exercise great attention in order to successfully address the impacts of missing data and assure the correctness of insights while clustering longitudinal data. This can only be accomplished by paying close attention and using specific procedures.

1.2 Growth Trajectories of Patients

The age range of 4-11 years is of utmost importance in the context of child development due to its inclusion of significant developmental milestones. These milestones include the occurrence of the 'adiposity rebound' (AR), which refers to a dip in the growth curve between the peak of infancy and the peak of adolescence, typically observed between the ages of 4 and 6 years. Additionally, this age range encompasses the period of a 'mid-growth spurt', characterized by a rapid increase in growth velocity between the ages of 4 and 11 years. Furthermore, in certain individuals, this age range may also involve the onset of 'pre-puberty'. The form of the development trajectory throughout childhood has been shown to have a significant association with a considerable number of health issues associated with obesity. The temporal occurrence of the mid-growth spurt has been shown to have predictive value for the beginning of metabolic syndrome, non-alcoholic fatty liver disease, and type 1 diabetes. Similarly, the timing of adrenarche has been seen to be indicative of subsequent developmental milestones, such as the onset of puberty. Children who start AR after age 5.5 have comparatively fewer health concerns, but AR happening before age 5.5 increases the likelihood of adult obesity. The "significant gap" in the research surrounding overweight and obesity in children over 5 years old was noted by the World Health Organization (WHO) in 2016. This review synthesizes research published since 2010 that presents data on body mass index (BMI) values at the group or individual level within the age range of 4 to 11 years. The majority of treatments aimed at preventing childhood obesity typically focus on children between the ages of 6 and 12 years. This trend is expected to continue owing to the advantageous accessibility of school settings for reaching kids[22].

1.3 Contributions

My whole Master's thesis project's goal was to design and implement a machine-learning-based system to act as a medical assistant for physicians to predict diseases and prescribe preventive healthcare advices to their patients. In the following, you can find the contributions of this work:

- We worked on unique dataset of heights and weights provided by Karolinska Institute under the BAMSE project[2]. To the best of my knowledge, there is no previous work on this dataset for BMI clustering using longitudinal trajectory analysis through the implementation of unsupervised clustering.
- As we had around 35% missing data ratio in our dataset, I implemented the Mean Expected Growth imputation based on the mean of BMIs of patients at different time points and their growth patterns.
- In addition to the hard clustering which assigns each patient to only one cluster, we proposed the Forgetting Factor approach to analyze trajectories in a way that considers previous observations' effect on the current age's BMI cluster assignment.
- Our results adds additional evidence to the OECD guidance regarding high BMI risks and interventions (World Health Organization, 2016) by targeting early life factors such as gestational smoking, lactation, and pre–gestational and gestational BMI control.

1.4 Thesis Organization

In Chapter 2, the preliminary concepts of Clustering algorithms, Missing data, and, Imputation methods will be discussed. Then, in Chapter 3, Dataset description, Mean Expected Growth approach, Cluster selection, Experimental results, and Results discussion will be explained. Finally, in Chapter 4, you can find some content about possible future works and conclusions of this thesis.

2

Background

2.1 Chapter overview

The task of extracting significant structures from raw data is a fundamental aspect of analyzing information within this domain, clustering methods are of utmost importance. Clustering is a fundamental technique within the domain of unsupervised learning, whereby data points are organized into distinct groups based on their similar properties. This chapter will go into an examination of several clustering methodologies, each of which offers a distinct approach to the task of splitting data into coherent clusters. Through an extensive examination of many methodologies for data segmentation, including hierarchical and partitioning algorithms, density-based and model-based techniques, and others, our objective is to get a comprehensive understanding of the existing tools and strategies available for this task. You will learn about the important things researchers need to think about when picking the right clustering method for certain analysis tasks during this talk. It will also lay a foundation for the next study.

2.2 Clustering algorithms

2.2.1 K-means

The K-means clustering strategy is a widely used unsupervised learning method. This approach requires an integer k and n observations. The final result is a division of the n observations into k sets, where each observation is assigned to the cluster with the closest mean. The procedures of k-means are outlined in the stages that follow. Begin by initializing k cluster centers. In practical application, the task may be accomplished by using either a random selection method for k center selection. In the following and in Figure 2.1, you can find the algorithm of K-means steps:

1. The points obtained from the n observations or the randomly generated k center points.
2. Compute the distance between each individual observation and the respective cluster centers.
3. Assign every single point to the cluster with the shortest distance from its center among all cluster centers.
4. Calculate the cluster mean again using the locations of the k centers.
5. Recalculate the distance between each individual data point and the recently calculated centroids. Continuously iterate through steps 3 and 4 until all data points have been allocated to a single cluster and remain stationary.

In most cases, previous information about the nature of the data or the use of clustering validity metrics will have an impact on the selection of the value of k[17, 24].

The objective of k-means clustering is to divide a set of observations $(x_1, x_2, ..., x_n)$, where each observation is a d-dimensional real vector, into $k \leq n$ sets $S = \{S_1, S_2, ..., S_k\}$ in a way that minimizes the sum of squares in each cluster [14].

Objective:

$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg\min_S \sum_{i=1}^{k} \mid S_i \mid VarS_i$(2.1) where μ_i is the mean, commonly referred to as the centroid, of the points inside S_i:

$$\mu_i = \frac{1}{\mid S_i \mid} \sum_{x \in S_i} x \tag{2.2}$$

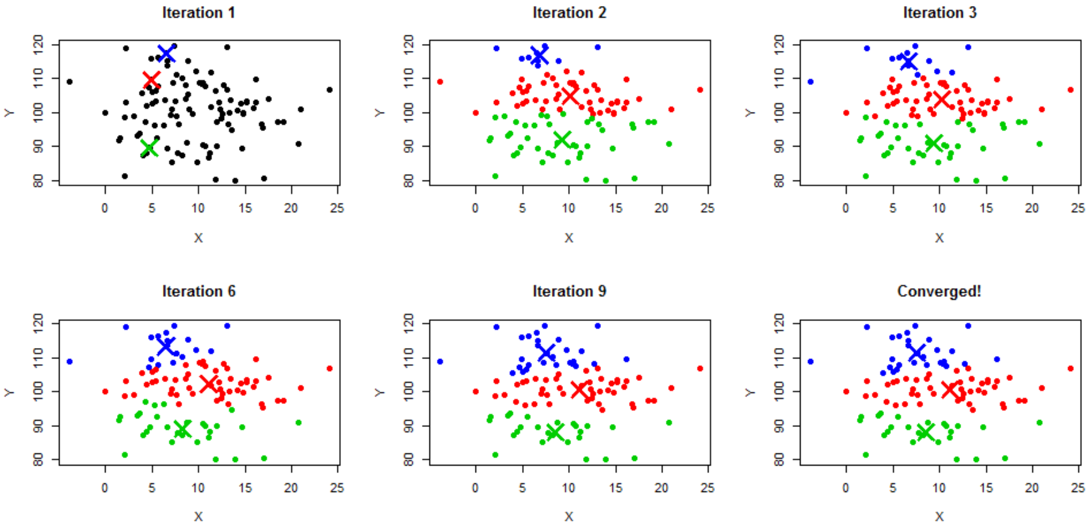


Figure 2.1: K-means Clustering [5]

- μ_i : Centroid or mean
- $\mid S_i \mid$: Size of S_i
- $\|.\|$: Norm

The idea of k-means clustering is very advantageous due to its simplicity of implementation and its adaptability in many application domains. However, it is important to acknowledge the existence of certain disadvantages and constraints. In the k-means algorithm, it is seen that all the clusters formed have a circular shape. This is due to the cluster centroids being updated repeatedly using the mean value. So, it couldn't be applied to a dataset in which the distribution of the points does not circular shape. Here a distribution-based model (Gaussian Mixture Model) will thus be used going forward in place of a distance-based approach[10].

2.2.2 Gaussian mixture model

Gaussian Mixture Models (GMMs) are a probabilistic framework used for the purpose of modeling real-world datasets. Gaussian Mixture Models (GMMs) are an extension of Gaussian distributions, enabling the representation of datasets that exhibit clustering patterns characterized by several Gaussian distributions in figure 2.2. The Gaussian mixture model is a probabilistic model that posits the generation of all data points from a combination of Gaussian distributions with parameters that are not known. The use of a Gaussian mixture model is applicable in the context of clustering, a computational activity that involves splitting a collection of data points into distinct groups.

The mentioned method may be used in finding clusters in datasets with uncertain or unclear cluster boundaries. Additionally, GMMs can be employed to estimate the probability that a new data point belongs to each specific cluster. Gaussian Mixture Models have a significant level of robustness against outliers, hence allowing them to generate accurate results even when faced with data points that depart from normal trends observed within the clusters. GMMs provide a significant level of adaptability and effectiveness when used for the task of data clustering [30, 1].

The GMM may be conceptualized as a probabilistic model in which Gaussian distributions are postulated for each group, characterized by their respective means and covariances determining their parameters. It is comprised of two fundamental components, namely the mean vectors (μ) and the covariance matrices ($\sum$). It is important to note that a Gaussian distribution is characterized as a continuous probability distribution that exhibits a bell-shaped curve. The Gaussian distribution is also often referred to as the normal distribution[26].

Here you can find different steps of the Gaussian Mixture Model(GMM) which you can find in figure2.3:

1. To determine the appropriate number of clusters for the given dataset, one may use many ways such as using domain expertise or employing statistical techniques like the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and Silhouette score.

2. Initializing mean, covariance, and weight parameter per cluster.

3. Utilize the Expectation Maximization method to accomplish the following objectives:

   - The Expectation Step (E step): involves the calculation of the probability of each data point being assigned to each distribution. Subsequently, the likelihood function is evaluated using the current estimation for the parameters.
   - The Maximization step (M step): involves updating the mean, covariance, and weight parameters from the previous iteration in order to maximize the anticipated likelihood obtained in the Expectation step (E step).
   - Continue iterating through these stages until the model reaches convergence.

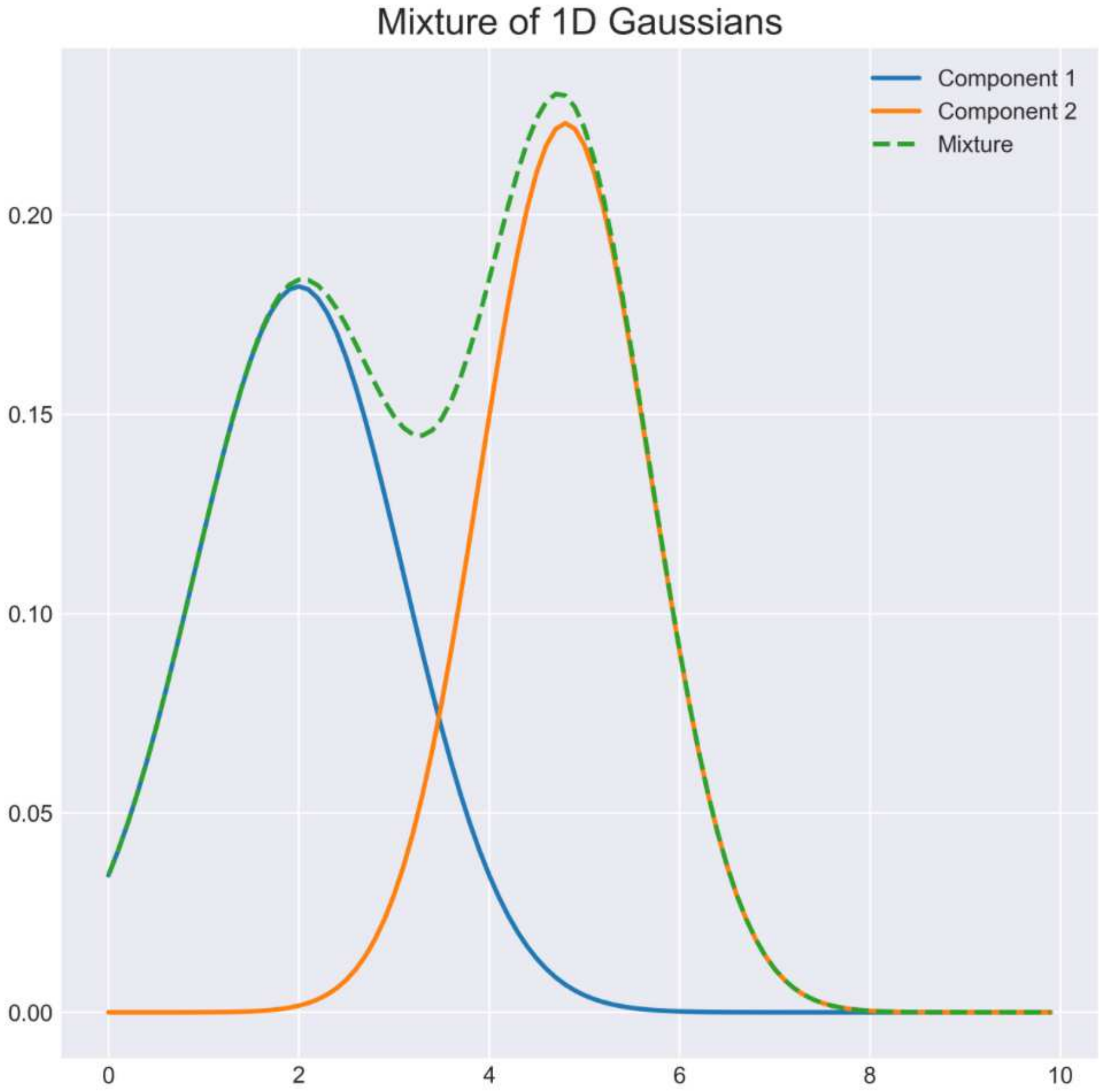


Figure 2.2: Mixture of 1D Gaussians [8]

Mathematics of Gaussian Mixture Model:

The Gaussian Distribution in one dimension is represented as:

$$G(X|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.3}$$

- μ: Mean
- σ^2: Variance of the distribution

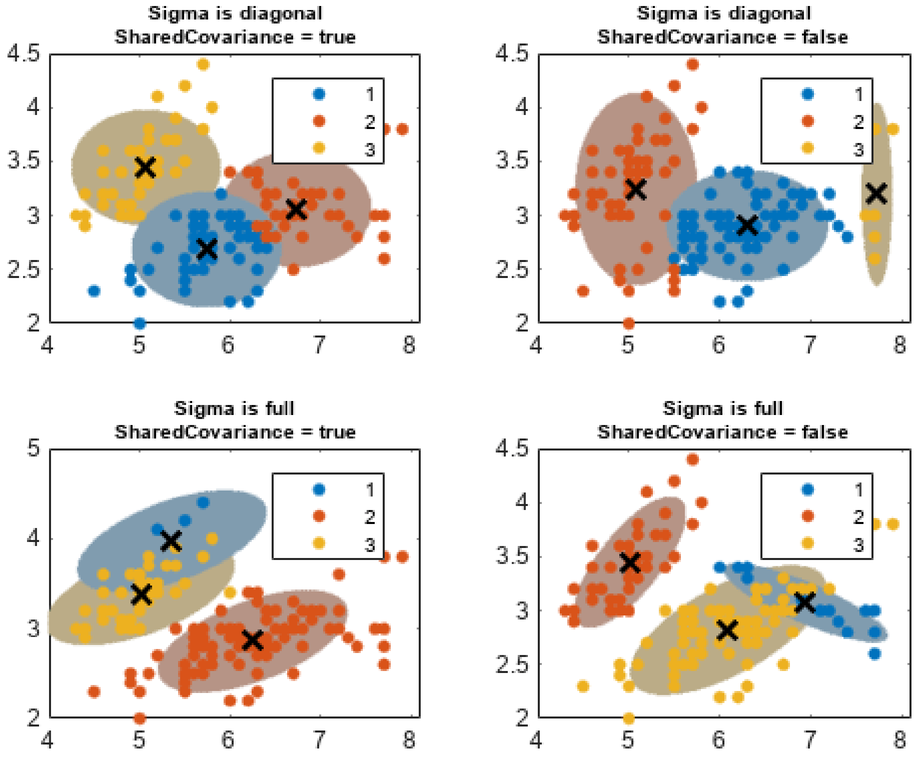


Figure 2.3: Clustering using Gaussian Mixture Model [8]

The probability density function for the Multivariate Gaussian Distribution is expressed as follows:

$$G(X|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(X-\mu)^T\mathbf{\Sigma}^{-1}(X-\mu)\right) \tag{2.4}$$

- μ: d-dimensional vector represents the mean of the distribution
- Σ: dxd covariance matrix

Assuming a predetermined number of clusters, denoted as K, let us consider the scenario. The parameters μ and Σ are also calculated for each value of k. If there had been just one distribution, the estimation would have been conducted using the maximum-likelihood technique. However, given the existence of K clusters, the probability density may be expressed as a linear function of the densities of all K distributions,i.e.

$$p(X) = \sum_{k=1}^{K} \pi_k G(X|\mu_k, \Sigma_k) \tag{2.5}$$

- π_k: The mixing coefficient for k^{th} distribution

To estimate the parameters using the maximum log-likelihood approach, it is necessary to calculate the following:

$$\ln p(X|\mu, \Sigma, \pi) = \sum_{i=1}^{N} p(X_i) = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \pi_k G(X_i|\mu_k, \Sigma_k) \tag{2.6}$$

Create a random variable $\gamma_k(X)$ such that $\gamma_k(X) = p(k|X)$.

Based on the Bayes theorem,

$$\gamma_k(X) = \frac{p(X|k)p(k)}{\sum_{k=1}^{K} p(k)p(X|k)} = \frac{p(X|k)\pi_k}{\sum_{k=1}^{K} \pi_k p(X|k)} \tag{2.7}$$

Now, in order to maximize the log-likelihood function, the derivative of $p(X|\mu, \Sigma, \pi)$ with respect to μ, $\sum$, and π should be equal to zero. Therefore, reorganizing the terms as follows:

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_k(x_n)(x_n - \mu_k)(x_n - \mu_k)^T}{\sum n = 1^N \gamma_k(x_n)} \tag{2.8}$$

And,

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x_n) \tag{2.9}$$

Therefore, it is rather obvious that the parameters cannot be evaluated in closed form. It is in situations like these when the Expectation-Maximization method comes in handy[9].

Expectation-Maximization (EM) Algorithm:

The Expectation Maximization (EM) algorithm is a widely applicable iterative computational technique for estimating maximum likelihood (ML) parameters. It is particularly effective in addressing incomplete-data situations, where more complex algorithms like the Newton-Raphson method may be less suitable. The EM method is used in a range of scenarios, including not just situations with clearly incomplete data, characterized by missing values, but also a diverse array of circumstances where the incompleteness of data is not inherently or immediately apparent. The EM approach has been used to address previously intricate maximum likelihood prediction problems in many contexts, either by directly resolving them or by simplifying the maximum likelihood prediction procedures that were previously sophisticated. The underlying principle of the Expectation-Maximization (EM) approach is to create a relationship between an incomplete-data problem and a complete-data problem. This link allows for the application of maximum likelihood estimation to the complete-data problem, making it computationally more viable and yielding closed-form estimates. To summarize, the EM algorithm comprises two independent procedures, including the E-step and the M-step. The E-step involves calculating the log-likelihood of the complete data issue by using the observed data set from the incomplete-data problem and the current parameter values. The M-step involves maximizing the log-likelihood that is created by the E-step. The aforementioned two procedures are iteratively executed until the condition of convergence is met[16].

2.2.3 K-medoid

One of the primary drawbacks of the k-Means approach is its susceptibility to outliers, since the presence of one item with an exceptionally high value may significantly affect the overall distribution of the data. Instead of using the mean value of the items inside a cluster as a point of reference, an alternative approach involves employing a medoid, which represents the object that is situated at the centermost position within the cluster, an example is shown in2.4. Therefore, the partitioning technique may still be executed by adhering to the idea of reducing the total dissimilarities between each item and its respective reference point. The K-Medoids approach is founded around this particular principle. The fundamental approach used by K-Medoids clustering algorithms is the identification of k clusters within a set of n items. This is achieved by first selecting a representative object, referred to as the medoid, for each cluster in an arbitrary manner. Every remaining item is grouped along with the medoid that it has the most similarity with. The K-Medoids technique uses representative items as reference points rather than computing the average value of the objects inside each cluster. The method accepts a parameter, denoted as k, which represents the desired number of clusters to be allocated among a given collection of n items [11]. The K-medoid algorithm is a well-established partitioning approach used for clustering, which aims to group a dataset of n items into k clusters. The value of k, which represents the number of clusters necessary, is to be provided by the user. The method operates based on the premise of reducing the total dissimilarities between each item and its respective reference point. The approach employs a random selection process to choose k items from dataset D as the first representative objects, which are referred to as medoids. A medoid may be characterized as the entity inside a cluster that exhibits the lowest average dissimilarity to all other entities in the cluster. In other words, it is the point that is most centrally positioned within the provided dataset. After each assignment of a data item to a certain cluster, the new medoid is determined for all medoids. The issue is that K-Medoids do not provide the same results on each run since the ensuing clusters are determined by the original random assignments. It is more resilient than K-medoids in dealing with the presence of noise and outliers, but it is more expensive to process than the K-medoid approach. Finally, since the ideal number of clusters k is difficult to anticipate, it is challenging for a user with no previous information to select the value of k[13].

Next, we will examine the internal workings of the k-medoids Algorithm, which may be described as follows:

1. The process of initializing k clusters is performed inside the provided data space D.
2. The process involves the random selection of k items from a set of n objects in a dataset, followed by the assignment of each selected object to a distinct cluster, ensuring that each object is allocated to just one cluster. Therefore, it is designated as the first medoid for every cluster.
3. Calculate the Cost(distance from all medoids as determined by Euclidean, Manhattan, or Chebyshev techniques) for all existing non-medoid objects.
4. Next, allocate each remaining non-medoid item to the cluster whose medoid has the smallest distance to that object, in comparison to the medoids of the other clusters.
5. Calculate the total cost, which refers to the sum of distances between each non-medoid item and its respective cluster medoid. This value will be assigned to dj.
6. Select a non-medoid item i in a random manner.
7. Next, do a temporary exchange of the object i with the medoid j. Subsequently, repeat the fifth step in order to recompute the overall cost and assign it to di.
8. If di is less than dj, the temporary swap made in step number 7 should be made permanent in order to create the new set of k medoids. Furthermore, it is necessary to reverse the temporary exchange that was executed in step number 7.

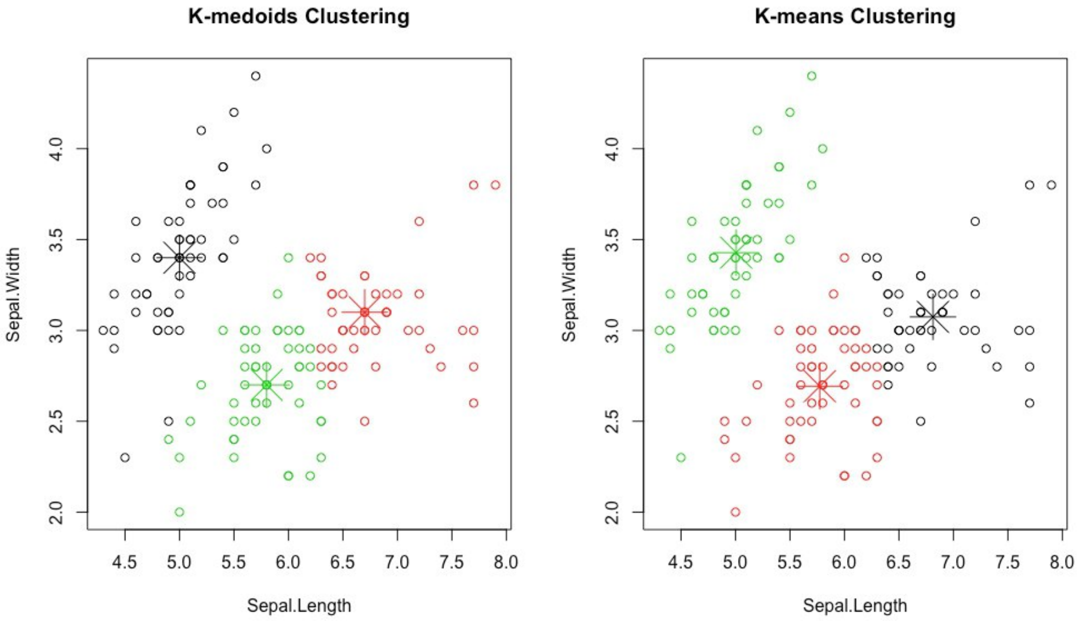


Figure 2.4: K-medoids vs K-means clustering [12]

9. Continuously iterate through the execution of steps 4, 5, 6, 7, and 8 until a state is reached where no more modifications occur.

2.3 Missing data

When there is no information recorded for some of the expected variables or individuals this is known as missing data. Incomplete data input, device failure, deleted files, and other accidents are only a few of the numerous possible causes of data loss. The presence of missing data is a frequent phenomenon that may have a substantial impact on the inferences that can be derived from the available data[19]. Research of all kinds, including that in the medical field, sometimes suffers from a lack of full data, particularly in longitudinal studies.In the context of a longitudinal study, it is customary to measure each observation at the initial stage and thereafter at regular intervals throughout the research duration. It is fairly uncommon to encounter incomplete data in studies using such designs, since some people may not be accessible for measurement at all time periods. Furthermore, it is possible for a subject to exhibit missing data at one follow-up time point and then have their measurements taken at one of the subsequent time points, leading to the emergence of nonmonotone missing data patterns. The statistical modeling of such data is a significant problem for the statistician.

In [11], Ibrahim et al. describe three different categories for missing data:

MCAR:

Missing data are said to be missing completely at random (MCAR) when the absence of a particular value is unrelated to any observable or unobserved values. For instance, consider a scenario where some elements of the variable y_i are absent, whereas the variable X_i is fully observed. The missing values of y_i may be classified as Missing Completely At Random (MCAR) if the probability of seeing y_i is not dependent on the values of Xi or the values of y_i that have been observed or would have been seen. According to the Missing Completely At Random (MCAR) assumption, the observed data may be regarded as a random sample drawn from the whole dataset. The use of a complete-case analysis may result in a decrease in efficiency, but, it does not add any kind of bias. According to the Missing Completely At Random (MCAR) assumption, the missing data mechanism may be represented as $f(r_i|X_i, \phi)$, where ϕ is a vector of unknown parameters. In other words, the values of the missing-data indicators $R_i = (R_{i_1}, ..., R_{in_i})'$ are not influenced by the outcomes y_{ij} in the model, where R_{ij} equals 1 if y_{ij} is observed and 0 otherwise.

MAR:

missing data are said to be missing at random when the reason for not seeing a particular value is unrelated to the unobserved values of y_i, given the seen values. However, the presence of missing data may be influenced by other variables that have been detected. For instance, let us consider the scenario when X_i is seen in its entirety, whereas some components of y_i may be absent. The missing values of y_i are considered to be missing at random (MAR) if the probability of witnessing y_i is not reliant on the specific values of y_i that would have been seen, but may still be dependent on the observed values of y_i and X_i. The assumption being discussed here is considered to be more realistic compared to the Missing Completely at Random (MCAR) assumption. However, due to the fact that the observed answers are no longer derived from a random sample, some changes need to be implemented. Performing a complete case analysis is likely to be both inefficient and biased. It is evident that in cases when data exhibits missing completely at random (MCAR) patterns, it may be inferred that the missingness is missing at random (MAR). In the context of a clinical study, if the occurrence of missing data is solely dependent on the treatment allocation, which is considered a covariate, then the missingness mechanism may be classified as Missing Completely At Random (MCAR). Consequently, it can also be inferred that the missingness mechanism is Missing At Random (MAR). In the context of Missing at Random (MAR), the missing data mechanism may be represented as $f(r_i|X_i, y_{obs,i,\phi})$, where $y_{obs,i}$ refers to the observed components of y_i.

MNAR:

The missing data mechanism is considered to be non-random if the non-observation of a certain value is dependent on the value that would have been noticed. For instance, consider a scenario where some elements of the variable y_i are absent, whereas the variable X_i is entirely seen. The missingness mechanism of the values of y_i may be classified as Missing Not at Random (MNAR) if the likelihood of y_i being missing is dependent on the missing values of y_i, even if it is dependent on the observed values of y_i or Xi. Missing Not At Random (MNAR) is a commonly seen phenomenon in longitudinal research including repeated assessments, representing a broad and prevalent scenario. In order to make valid conclusions, it is often necessary to either accurately determine the appropriate model for the missing data mechanism, or make assumptions about the distribution of the variable y_i, or both. The estimators and tests that arise from these assumptions often exhibit sensitivity. Hence, it is essential for the mechanism to assume a pivotal position in what are often referred to as sensitivity assessments. In the context of missing not at random (MNAR), the missing data mechanism may be represented as $f(r_i|X_i, y_{obs,i}, y_{mis,i}, \phi)$, where r_i denotes the missing data, Xi represents the observed data, $y_{obs,i}$ refers to the observed outcome, $y_{mis,i}$ represents the missing outcome, and ϕ represents the parameters of the mechanism.

The absence of some data raises a number of issues. First, there is a reduction in the statistical power of the test due to the lack of data. Statistical power is the chance that the experiment will ignore the null hypothesis when it is incorrect. Second, the missing data may introduce an element of bias into the parameter estimate process. Third, it may lessen the samples' ability to be representative of the whole. Fourth, it might make the interpretation of the research more difficult. Each of these issues poses a potential risk to the reliability of the trials and may result in wrong conclusions being drawn. In suggest some techniques for handling missing data as follows:

Listwise or case deletion:

Listwise deletion, in which instances with incomplete data are simply eliminated from analysis, is the method that is used the most often when dealing with missing data. This approach is commonly used, however, it has the potential to induce bias if the assumption of Missing Completely at Random (MCAR) is not satisfied. When this assumption isn't satisfied by the data, using listwise deletion might result in biased parameter estimations, despite the fact that MCAR considers it to be unbiased and cautious. When there is a big sample size and when power is not an issue, it is possible that listwise deletion is feasible. However, it is not the best method when used for smaller numbers of samples or when the MCAR threshold is not reached.

Pairwise deletion:

Pairwise deletion is a technique for managing missing data that keeps information if particular data points are necessary for testing assumptions. This approach may be used in situations when the information would otherwise be lost. It applies statistical testing to the data that is available, regardless of whether or not further data points are absent from the dataset. Pairwise deletion is superior to listwise deletion in terms of the amount of information it maintains. On the other hand, it does have a few drawbacks:

- It has the potential to result in the model parameters relying on distinct sets of data with changing statistics, including sample size and standard errors.
- It has the potential to yield an intercorrelation matrix that does not have a positive definite result, which may make future analysis more difficult.

When working with data that are Missing Completely at Random (MCAR) or Missing at Random (MAR), as well as when suitable mechanisms are included as variables, pairwise deletion is the method that produces the least amount of bias. However, if there are too many observations that aren't included in the analysis, it can turn out to be flawed.

Mean substitution:

The technique of mean substitution is used as a means of addressing missing data, whereby the average value of a given variable is utilized to substitute the missing values associated with that variable. This technology allows academics to make use of datasets that are not fully comprehensive. The justification for using mean substitution is based on the notion that the mean is a suitable approximation for randomly chosen observations from a distribution that follows a normal pattern. Nevertheless, in the case of non-random missing values, particularly when there is a notable disparity in the frequency of missing values across variables, using mean replacement may result in the introduction of inconsistent bias. Furthermore, this approach does not provide novel insights but rather serves to augment the sample size, perhaps leading to an underestimation of mistakes. Therefore, mean substitution is typically not considered a recommended approach for addressing missing data.

Regression imputation:

Imputation refers to the procedure of substituting missing data with approximated values. Instead of excluding any instance with missing values, this methodology retains all cases by imputing the missing data with estimated values derived from other available information. Once all missing values have been substituted using this methodology, the dataset is subjected to analysis using conventional procedures applicable to datasets with no missing values. Regression imputation involves using the available variables to generate a forecast, which is then used as a replacement for an actual observed value. The use of this method has many benefits, since it preserves a substantial amount of data compared to listwise or pairwise deletion, while also preventing substantial modifications to the standard deviation or distribution shape. In contrast to mean substitution, regression imputation involves replacing missing values with anticipated values based on other variables. This method does not provide any new information, but it does expand the sample size and decrease the standard error.

Last observation carried forward:

Within the realm of anesthesiology research, a considerable number of investigations are conducted using the longitudinal or time-series methodology, whereby the participants' measurements are collected periodically throughout a sequence of time intervals. The latest observation carried forward (LOCF) is a commonly used imputation approach in this scenario. The aforementioned approach involves substituting each missing value with the most recent observed value obtained from the corresponding subject. In instances when a value is absent, it is substituted with the most recent observed value. This approach offers notable benefits due to its inherent simplicity in facilitating comprehension and effective communication among statisticians, doctors, sponsors, and researchers. Despite its simplicity, this strategy makes a significant assumption that the missing data does not alter the value of the result, which seems improbable in several contexts, particularly in the context of anesthetic studies. The estimation generated by this method has a bias towards a particular treatment effect and tends to underestimate the level of variability associated with the projected outcome. The National Academy of Sciences has advised against the uncritical utilization of simple imputation techniques, such as last observation carried forward (LOCF) and baseline observation carried forward (BOCF). They assert that these single imputation methods should not be employed as the primary approach for handling missing data unless there is scientific justification for the assumptions upon which they are based.

Maximum likelihood:

Various solutions may be used within the framework of the greatest likelihood approach to address the issue of missing data. The assumption that the observed data represent a sample chosen from a multivariate normal distribution is comprehensible in these cases. The missing data are approximated based on the recently estimated parameters once the parameters have been calculated using the available data. In situations when there is a lack of full data, although the available data is rather comprehensive, the statistical analysis of the correlations between variables may be conducted via the use of the maximum likelihood technique. The estimation of missing data may be achieved through the use of the conditional distribution of the remaining variables.

Expectation-Maximization:

The Expectation-Maximization (EM) algorithm is a statistical technique often used for estimating missing values in a dataset based on maximum likelihood principles. The process consists of two distinct stages, namely expectation and maximization. During the expectation stage, the estimation of parameters including variances, covariances, and means is conducted, often commencing with the use of listwise deletion. The aforementioned estimations are used in the formulation of regression equations with the purpose of predicting data that is absent or missing. The maximizing stage then uses these equations to complete the missing data. The iterative procedure continues until stability is achieved, as shown by the presence of a covariance matrix that remains stable throughout.

It is worth mentioning that EM imputation incorporates random perturbations for each imputed value in order to address the uncertainty associated with the imputation process. But there are disadvantages to EM imputation:

- Convergence may be sluggish, particularly when there are a lot of missing data.
- Some analysts may find it hard to understand how complicated it is.
- The presence of bias in parameter estimations and the underestimation of standard errors may be seen as a consequence.
- Single imputation, which is often used as a consequence of the expectation-maximization (EM) algorithm, has a tendency to underestimate standard errors. This might possibly lead to an inflation of the perceived statistical power.

Hence, it is advisable to exercise caution while using EM imputation, and it is often advised to use several imputations in order to mitigate the uncertainty that arises from the imputation process.

Multiple imputation:

Multiple imputation is a robust and effective approach for addressing the issue of missing data. Instead of substituting missing data with a single imputed value, the approach involves replacing them with a collection of probable values that include the inherent variety and uncertainty associated with the actual values. The procedure starts by making predictions for absent data by the use of current data derived from other factors. Subsequently, the absence of data points is addressed by the substitution of estimated values, so generating a collection of imputed datasets. The iterative technique described herein produces numerous imputed datasets, each of which is then subjected to independent analysis using conventional statistical methodologies. The findings derived from these studies are aggregated to provide a unified overall analysis outcome.

Some major benefits of multiple imputation are:

- One potential approach to address the issue of missing data is to restore its natural variability via the incorporation of correlations with other variables.

- By incorporating uncertainty resulting from missing data, one may achieve reliable statistical inference.
- suited for low-volume data sets or those with significant missingness, as well as robustness to deviations from normality assumptions.

Although the concept of multiple imputation may seem intricate in principle, advancements in statistical software have greatly facilitated its implementation, making it more attainable for researchers. The aforementioned methodology has significant value in effectively tackling the challenges posed by missing data and generating dependable statistical outcomes.

Sensitivity analysis:

Sensitivity analysis is a research approach that examines the manner in which the uncertainty in the outcome of a model may be attributed to the many sources of uncertainty in its inputs. When doing an analysis of missing data, it is common to make additional assumptions on the causes behind the missing data. These assumptions are often relevant to the main analysis being performed. Nevertheless, the assumptions cannot be conclusively verified for their accuracy. Hence, the National Research Council has recommended the use of sensitivity analysis as a means to assess the resilience of the findings in relation to departures from the Missing at Random (MAR) assumption.

2.4 Imputation methods

2.4.1 Linear interpolation:

Linear interpolation is a method of imputation that posits a linear association between data points and uses neighboring non-missing values to approximate the missing values that you can find in figure2.5. It is a technique for creating new data points that fall within the bounds of a discrete group of existing data points.

The linear extrapolation method works like this:

- Locate the data point(s) that are missing in the dataset.
- Find the non-missing values that are located immediately next to the missing value(s) on each side.

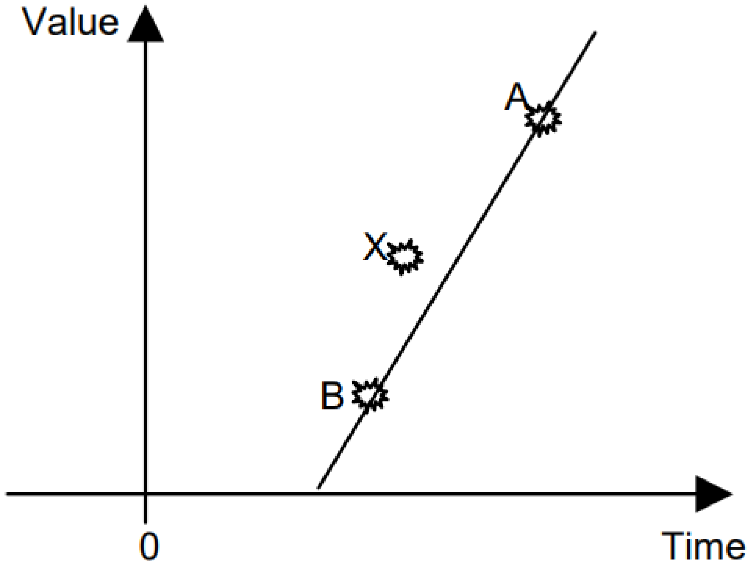


Figure 2.5: Linear Relationship Between Non-Missing Observations A and B, and Missing Observation [29]

- Find the slope of the line that joins the two non-missing data points.
- The slope of the data may be used to approximate missing value(s) while considering how close they are to the nearest existing values.

The mathematical representation for linear interpolation may be expressed using the following formula:

$$\frac{y_1 - y_2}{x_1 - x_2} \tag{2.10}$$

2.4.2 Forward-Fill(ffill) and Backward-Fill(bfill)

Forward-fill and backward-fill imputation are often used approaches for handling missing values in a dataset. Both methodologies are used in scenarios when a dataset contains missing values, necessitating the imputation of these values to facilitate calculations such as calculating the mean value.

The technique known as forward-fill, sometimes denoted as ffill, is used in data analysis to replace missing data in a dataset with the most recent non-missing value. The aforementioned strategy is used for extending the most recently verified observation progressively. The use of forward fill is beneficial in situations when the data has a series pattern, and it is expected that the missing values will resemble the previous values.

Backward fill, commonly referred to as bfill, is the antithesis of forward fill. The process involves substituting missing values with the subsequent non-missing value inside the dataset. The very first valid observation after a "NaN" value is propagated backward along the chosen axis using a technique called backward fill. The use of sequential or series data is advantageous in situations where the absence of values is anticipated to exhibit similarity with subsequent values.

Forward fill and backward fill may be used in combination with the group by an operation to execute filling inside certain groups. Furthermore, both approaches have the capability to restrict the propagation of valid observations to a certain number of rows utilizing the limit parameter.

The selection between forward fill (ffill) and backward fill (bfill) is contingent upon the particular contextual characteristics of the data. In a general context, the ffill method is considered more suitable for addressing missing values located at the beginning of a dataset, while the bfill method is deemed more acceptable for addressing missing values located at the conclusion of a dataset. Nevertheless, there are instances when the decision between using forward fill (ffill) or backward fill (bfill) may not be straightforward. In such situations, it becomes imperative to conduct empirical investigations employing both techniques in order to ascertain which one yields the most optimal outcomes[7].

3

Analysis

In this section, we review all the work done in this project in detail. At first, we have explained the dataset that we used, then we will talk about the imputation methods that have been used to impute missing data in our dataset. After that, we examined the different criteria by which the best number of clusters can be selected and finally analyzed the clustering results and their interpretation.

3.1 Dataset description

The research sample consisted of children who were part of the BAMSE cohort [2], a Swedish population-based birth cohort established in Stockholm between 1994 and 1996. In summary, a total of 3,897 children (each patient has 14 different values so there exist 54,572 values in the dataset) have been monitored throughout time by administering parental questionnaires to gather data on symptoms of allergic illness as well as environmental and lifestyle factors. During the clinical exams conducted at the ages of 4, 8, and 16 years, participants' weight was assessed using an electronic scale with a precision of 0.1 kg. The measurements were taken while the participants were wearing light indoor clothing. Additionally, height was measured with a wall-mounted stadiometer with a precision of 0.1 cm. Furthermore, within the original cohort, a total of 2,594 children (representing 63% of the cohort) had their weight and height measured at 10 specific ages ranging from 6 months (±2 weeks) to 12 years. These measurements were obtained from school and health-care records. The specific ages at which measurements were taken include 6 months (±2 weeks), 12 months (±4 weeks), 18 months (±4 weeks), 2 years, 3 years, 4 years, 5 years (±6 months), and 7 years, 10 years, and 12 years (with a range of −6 to ±11 months). Nevertheless, the precise age of the kid for each measurement was not provided. Data on weight and height, as reported by the individuals themselves, were gathered at two specific time points: at the ages of 12 and 16 years. The aforementioned data was used in cases when an individual had insufficient information available from either the health record (n = 455 at age 12 years) or the clinical evaluation (n = 458 at age 16 years). The validity of self-reported height and weight has been established throughout adolescence, namely at the age of 16, demonstrating a substantial level of concordance with objectively measured values.[6]. The information was obtained from the Swedish Medical Birth Register (the size of the dataset is 3,897). This dataset includes measurements of height and weight for children at different time points from birth up to 24 years of age. Specifically, data was taken at 14 distinct time points over this period. The time points include 0-month, 6-months, 18-months, 2-years(24 months), 3-years (36 months), 4-years (48 months), 5-years (60 months), 6-years (72 months), 7-years (84 months), 8-years (96 months), 10-years (120 months), 12-years (144 months), 16-years (192 months) and 24-years (288 months).

Unfortunately, our dataset had around 35% missing data ratio due to many reasons such as:

- Absence of participants during the measurement
- Data deletion due to negligence or device problem
- Participants might prefer not to disclose their medical information owing to worries about privacy and secrecy
- Use of several healthcare systems without adequate data integration
- Data entry errors refer to mistakes or inaccuracies that occur during the process of inputting data into a computer system or database. These errors may arise from several factors.

So, handling the missing data was one of our main challenges we figured it out by imputing these missing data using different imputation approaches. Before that, we visualized all the data to have a graphical representation of missing data which showed missing data in white color and non-missing data in black color using the missing data matrix Figure3.1. We found that the pattern of missingness in our dataset is MCAR(Missing Completely At Random).

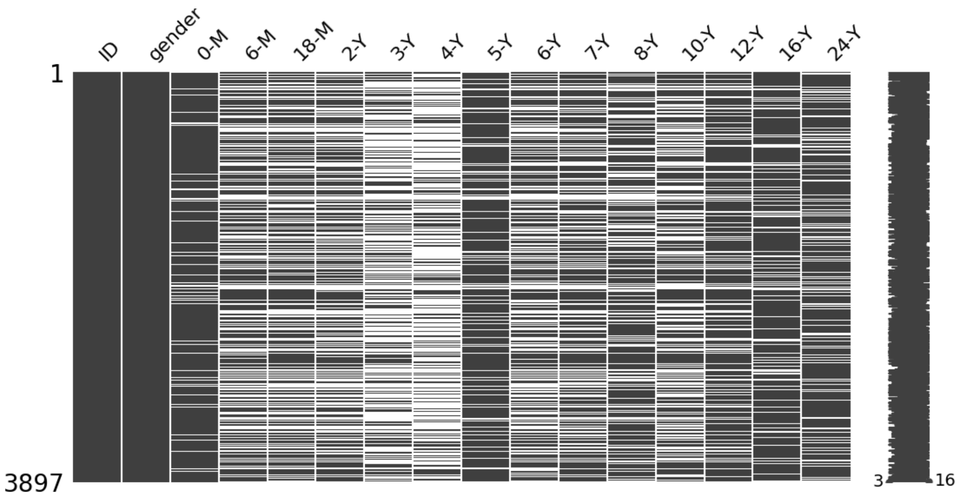


Figure 3.1: Graphical Representations of Missing data

There are some informative statistics that are shown in Tables 3.1 and 3.2 related to heights and weights respectively. As can be seen in Table3.1 the highest number of missing height data is related to the age of 4-years which is 2,630 and the lowest number of missing height data is related to the age of 0-month which is 365, in Table3.2 the highest number of missing weight data is related to the age of 4-years which is 2601 and the lowest number of missing weight data is related to the age of 0-month which is 39.

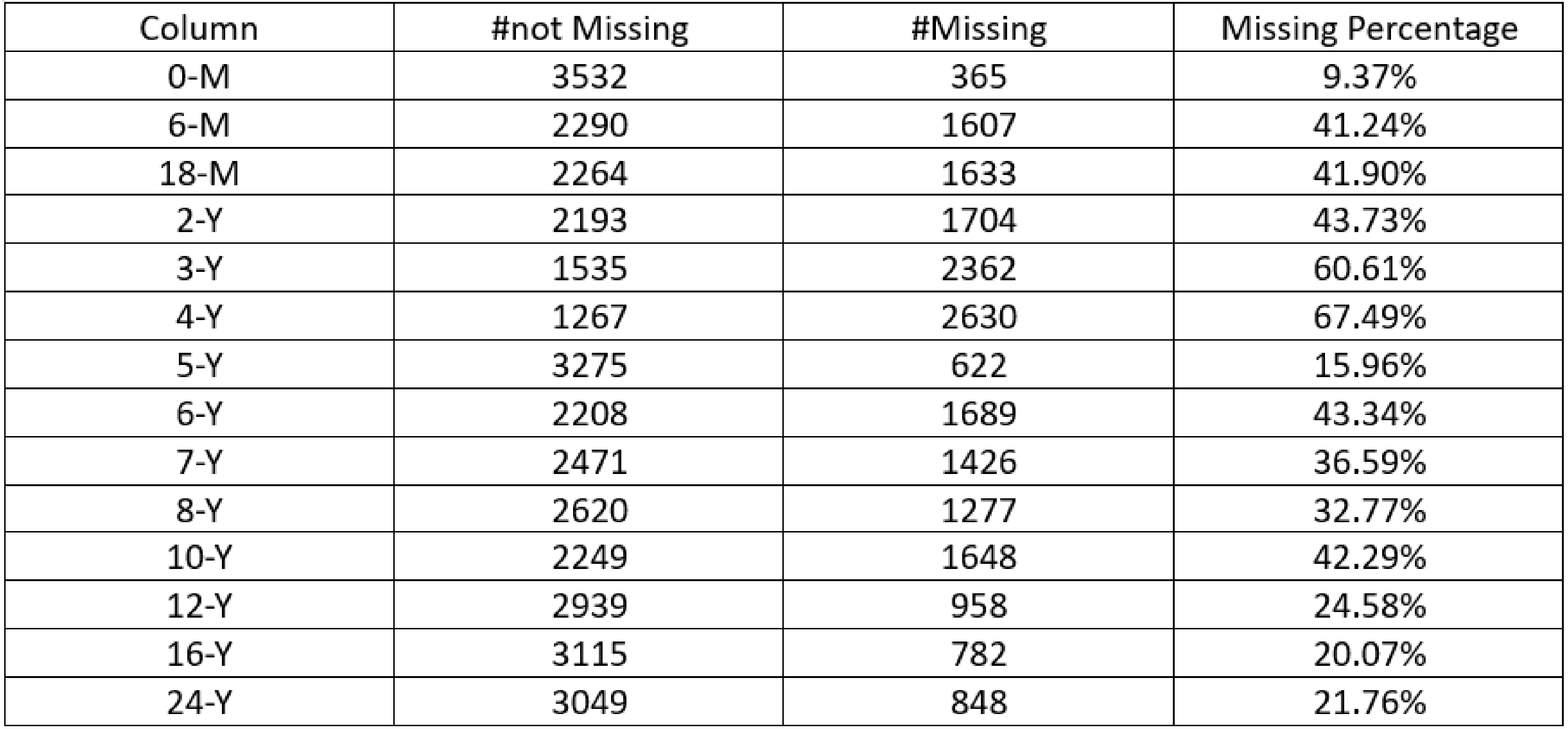
3.2 Mean expected growth

Mean Expected Growth is an approach that can be used to estimate missing values in a dataset In this way, by calculating the mean of the longitudinal data at each time point, we can obtain the growth rate of the data from each time point to another time point. In this project, we used Mean Expected Growth to impute missing data in the following steps:

- Computing the mean of data over each column which refers to time points including 0-month, 6-months, 18-months, 2-years(24 months), 3-years (36 months), 4-years (48 months), 5-years (60 months), 6-years (72 months), 7-years (84 months), 8-years (96 months), 10-years (120 months), 12-years (144 months), 16-years (192 months) and 24-years (288 months). So we have 14 different mean values.

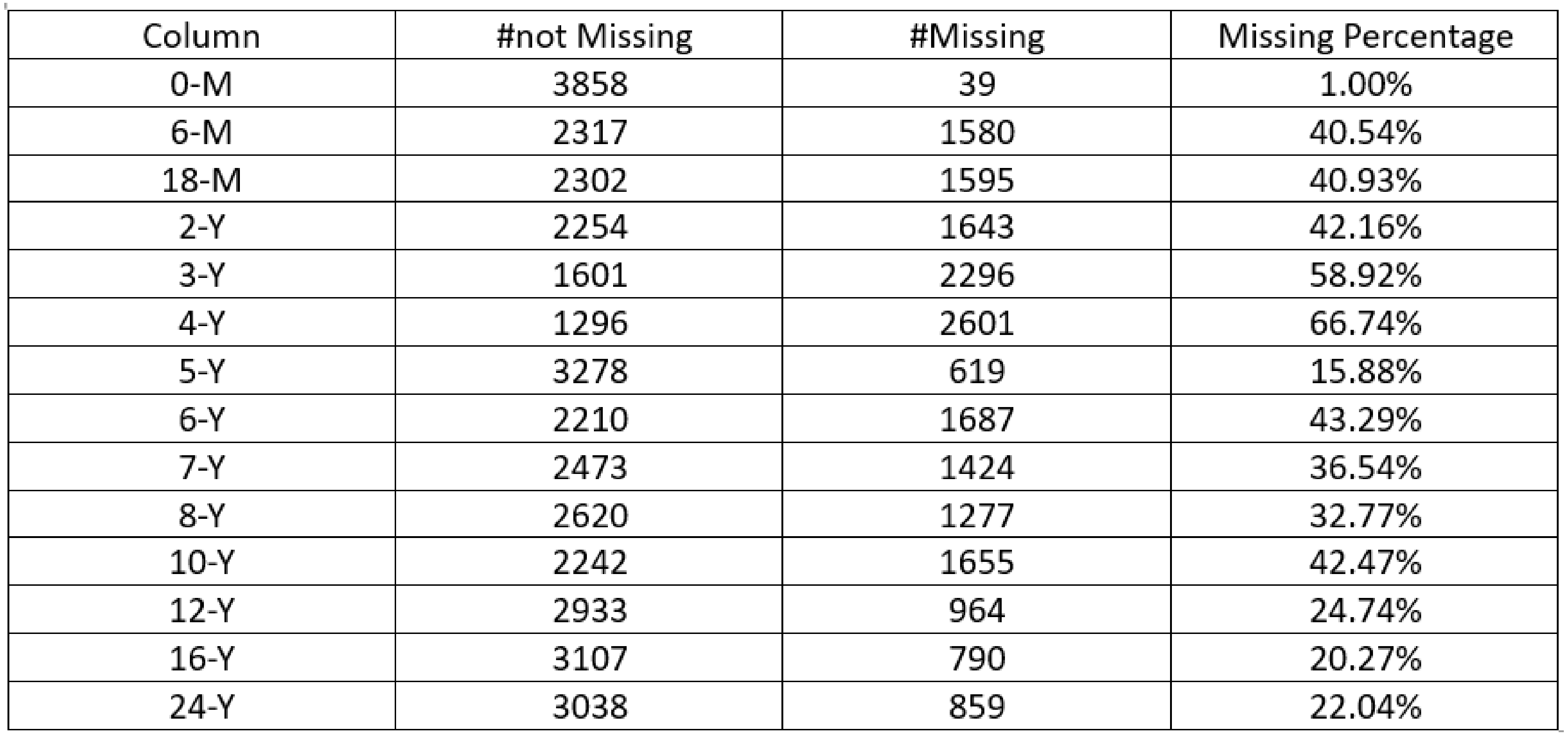
- Calculating the difference between $Mean_i$ and $Mean_{i-1}$

Table 3.1: Number of Missing and not Missing Values of Height Data



Column	#not Missing	#Missing	Missing Percentage
0-M	3532	365	9.37%
6-M	2290	1607	41.24%
18-M	2264	1633	41.90%
2-Y	2193	1704	43.73%
3-Y	1535	2362	60.61%
4-Y	1267	2630	67.49%
5-Y	3275	622	15.96%
6-Y	2208	1689	43.34%
7-Y	2471	1426	36.59%
8-Y	2620	1277	32.77%
10-Y	2249	1648	42.29%
12-Y	2939	958	24.58%
16-Y	3115	782	20.07%
24-Y	3049	848	21.76%

Table 3.2: Number of Missing and not Missing Values of Weight Data



Column	#not Missing	#Missing	Missing Percentage
0-M	3858	39	1.00%
6-M	2317	1580	40.54%
18-M	2302	1595	40.93%
2-Y	2254	1643	42.16%
3-Y	1601	2296	58.92%
4-Y	1296	2601	66.74%
5-Y	3278	619	15.88%
6-Y	2210	1687	43.29%
7-Y	2473	1424	36.54%
8-Y	2620	1277	32.77%
10-Y	2242	1655	42.47%
12-Y	2933	964	24.74%
16-Y	3107	790	20.27%
24-Y	3038	859	22.04%

- If we have missing data, considering the difference of means in two consecutive columns, we add the difference of means to the data before that missing data and replace the missing data with the obtained value.

$$x_i^c = x_{i-1}^{c-1} + (Mean^c - Mean^{c-1}) \tag{3.1}$$

- x_i^c: Missing data for a patient at time point $''i''$
- x_{i-1}^{c-1}: A data which refers to the previous time point $''i-1''$ for a patient
- $Mean^c$: Mean of data over a column at time point i
- $Mean^{c-1}$: Mean of data over a column at time point $i-1$

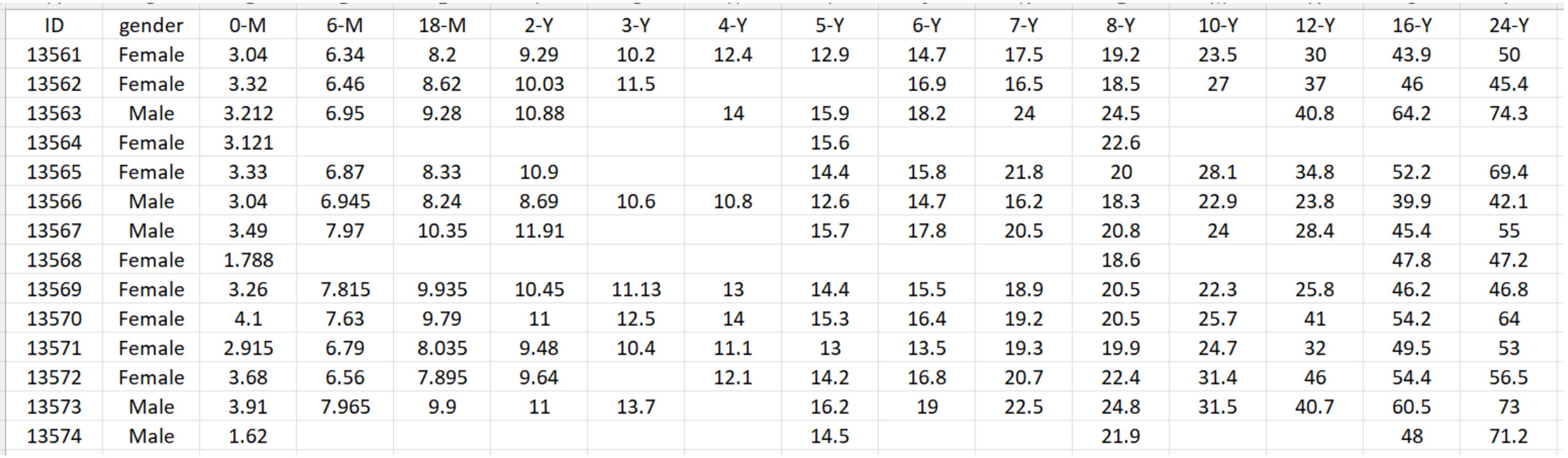
You can find how the missing data are imputed using Mean Expected Growth in Table3.3 refers to part of the dataset before imputation and Table3.4 after imputation.

3.3 Multivariate imputation by chained equation (MICE)

3.3.1 The core principles behind multiple imputation (MI)

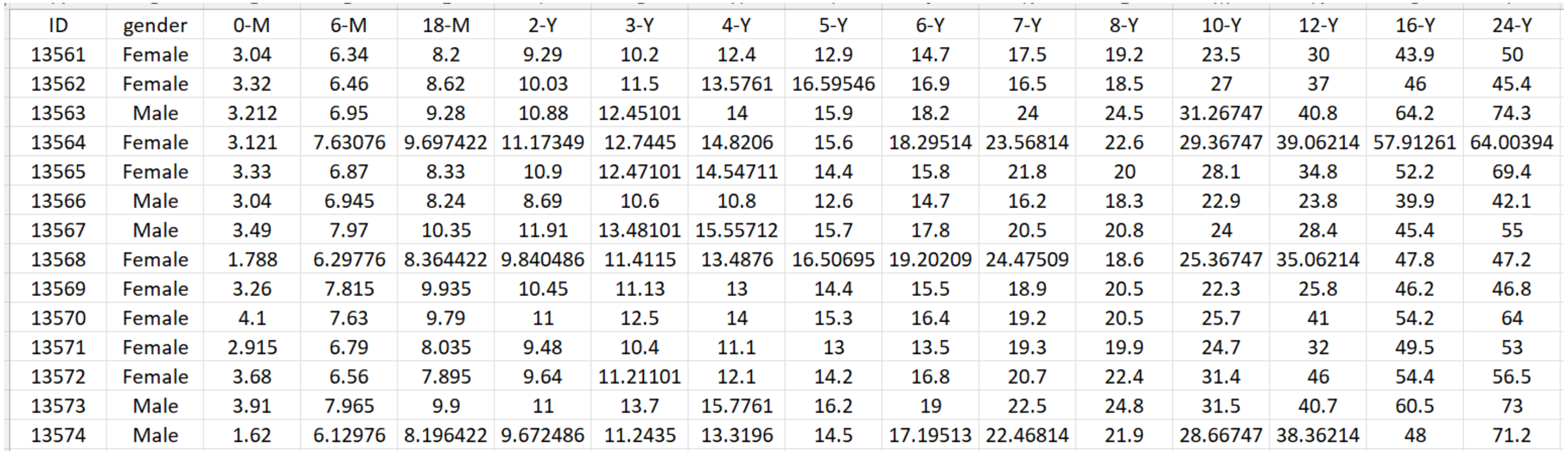
The multiple imputation (MI) process is used to replace each missing value with a range of potential values. In contrast to single imputation, this technique considers the uncertainty associated with estimating missing values. The methodology generates many datasets, from which estimations of relevant parameters may be derived. For instance, if one is interested in determining the coefficient for a covariate in a multivariable model, the coefficients will be predicted from each dataset, resulting in a total of m coefficients. Ultimately, these coefficients are aggregated to provide an approximation of the coefficient, while considering the inherent uncertainty associated with estimating missing values. The method of estimating the coefficient variance in this manner is less prone to underestimation in comparison to a single imputation. The imputation technique involves first constructing a predictive model for the target variable using the other variables that do not have missing values, Figure3.2. To put it another way, the variable that is being imputed is referred to be the response variable, while any other relevant variables are referred to as independent variables.

Table 3.3: The part of the original dataset for the weights (before imputation)



ID	gender	0-M	6-M	18-M	2-Y	3-Y	4-Y	5-Y	6-Y	7-Y	8-Y	10-Y	12-Y	16-Y	24-Y
13561	Female	3.04	6.34	8.2	9.29	10.2	12.4	12.9	14.7	17.5	19.2	23.5	30	43.9	50
13562	Female	3.32	6.46	8.62	10.03	11.5			16.9	16.5	18.5	27	37	46	45.4
13563	Male	3.212	6.95	9.28	10.88		14	15.9	18.2	24	24.5		40.8	64.2	74.3
13564	Female	3.121						15.6			22.6				
13565	Female	3.33	6.87	8.33	10.9			14.4	15.8	21.8	20	28.1	34.8	52.2	69.4
13566	Male	3.04	6.945	8.24	8.69	10.6	10.8	12.6	14.7	16.2	18.3	22.9	23.8	39.9	42.1
13567	Male	3.49	7.97	10.35	11.91			15.7	17.8	20.5	20.8	24	28.4	45.4	55
13568	Female	1.788									18.6			47.8	47.2
13569	Female	3.26	7.815	9.935	10.45	11.13	13	14.4	15.5	18.9	20.5	22.3	25.8	46.2	46.8
13570	Female	4.1	7.63	9.79	11	12.5	14	15.3	16.4	19.2	20.5	25.7	41	54.2	64
13571	Female	2.915	6.79	8.035	9.48	10.4	11.1	13	13.5	19.3	19.9	24.7	32	49.5	53
13572	Female	3.68	6.56	7.895	9.64		12.1	14.2	16.8	20.7	22.4	31.4	46	54.4	56.5
13573	Male	3.91	7.965	9.9	11	13.7		16.2	19	22.5	24.8	31.5	40.7	60.5	73
13574	Male	1.62						14.5			21.9			48	71.2

Table 3.4: The part of the dataset after imputation for the weights using Mean Expected Growth approach



ID	gender	0-M	6-M	18-M	2-Y	3-Y	4-Y	5-Y	6-Y	7-Y	8-Y	10-Y	12-Y	16-Y	24-Y
13561	Female	3.04	6.34	8.2	9.29	10.2	12.4	12.9	14.7	17.5	19.2	23.5	30	43.9	50
13562	Female	3.32	6.46	8.62	10.03	11.5	13.5761	16.59546	16.9	16.5	18.5	27	37	46	45.4
13563	Male	3.212	6.95	9.28	10.88	12.45101	14	15.9	18.2	24	24.5	31.26747	40.8	64.2	74.3
13564	Female	3.121	7.63076	9.697422	11.17349	12.7445	14.8206	15.6	18.29514	23.56814	22.6	29.36747	39.06214	57.91261	64.00394
13565	Female	3.33	6.87	8.33	10.9	12.47101	14.54711	14.4	15.8	21.8	20	28.1	34.8	52.2	69.4
13566	Male	3.04	6.945	8.24	8.69	10.6	10.8	12.6	14.7	16.2	18.3	22.9	23.8	39.9	42.1
13567	Male	3.49	7.97	10.35	11.91	13.48101	15.55712	15.7	17.8	20.5	20.8	24	28.4	45.4	55
13568	Female	1.788	6.29776	8.364422	9.840486	11.4115	13.4876	16.50695	19.20209	24.47509	18.6	25.36747	35.06214	47.8	47.2
13569	Female	3.26	7.815	9.935	10.45	11.13	13	14.4	15.5	18.9	20.5	22.3	25.8	46.2	46.8
13570	Female	4.1	7.63	9.79	11	12.5	14	15.3	16.4	19.2	20.5	25.7	41	54.2	64
13571	Female	2.915	6.79	8.035	9.48	10.4	11.1	13	13.5	19.3	19.9	24.7	32	49.5	53
13572	Female	3.68	6.56	7.895	9.64	11.21101	12.1	14.2	16.8	20.7	22.4	31.4	46	54.4	56.5
13573	Male	3.91	7.965	9.9	11	13.7	15.7761	16.2	19	22.5	24.8	31.5	40.7	60.5	73
13574	Male	1.62	6.12976	8.196422	9.672486	11.2435	13.3196	14.5	17.19513	22.46814	21.9	28.66747	38.36214	48	71.2

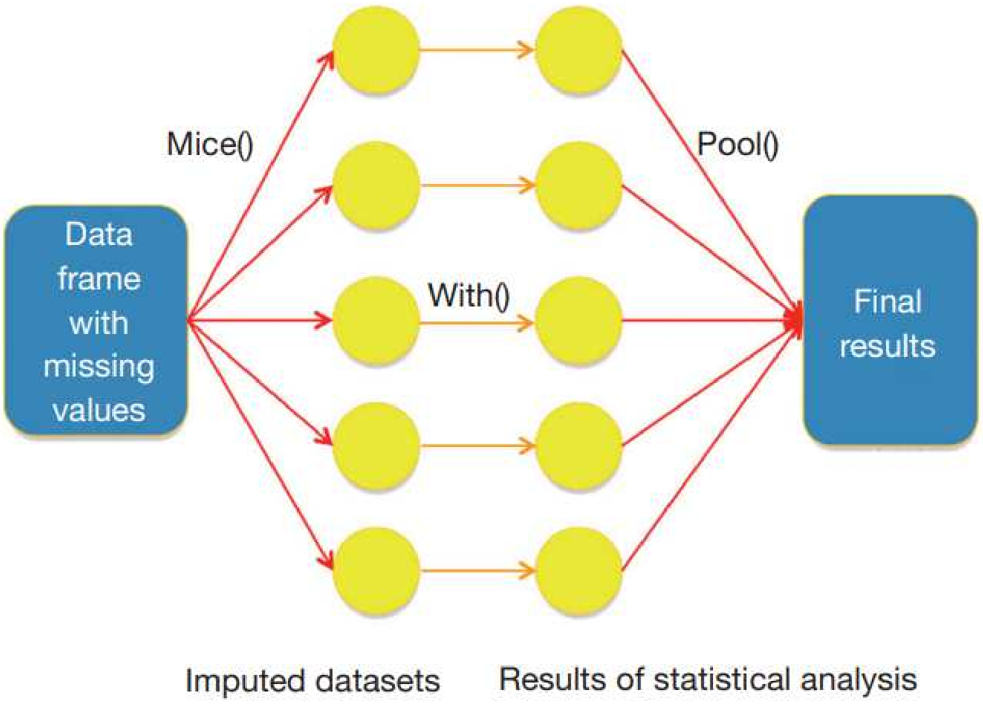

Mice()
Pool()
Data frame with missing values
With()
Final results
Imputed datasets
Results of statistical analysis

Figure 3.2: The schematic representation of the functionality of the MICE package in handling missing values inside a data frame. The sequential use of the mice(), with(), and pool() methods can be observed. [31]

3.3.2 MICE

The acronym MICE refers to Multivariate Imputation by Chained Equations. The statistical technique used to address missing data within a dataset is referred to as a missing data handling approach. The methodology employs multiple imputation methods to address missing data, followed by the integration of outcomes from numerous imputations to provide a conclusive imputed dataset [18]. The objective of Multiple Imputation by Chained Equations (MICE) is to maintain the association between variables in the original dataset while mitigating the potential bias that arises from imputing missing values. Multiple Imputation by Chained Equations (MICE) is a versatile and resilient approach for managing missing data.

The MICE (Multiple Imputation by Chained Equations) technique is an iterative imputation approach that uses a regression model to replace missing data with multiple imputations. The values that have been imputed are then utilized for estimating the values that are missing in the next iteration, continuing until the convergence conditions are satisfied.

The following are some justifications for why MICE is significant:

- Preserves relationships: The preservation of variable associations in the original data is a crucial aspect for ensuring the accuracy of findings in machine learning models, and the MICE approach fulfills this purpose well.
- Reduces bias: By iteratively imputing missing values and aggregating the outcomes, Multiple Imputation by Chained Equations (MICE) effectively decreases the potential bias that may arise from incomplete data.
- Flexibility: The MICE method is a flexible approach capable of effectively managing many forms of missing data, including both missing at-random and missing not-at-random data.
- Handles large amounts of missing data: MICE is especially beneficial for datasets with a significant number of missing values, when other imputation strategies may be ineffective.
- Comprehensive: MICE is an all-inclusive method for dealing with missing data that takes into consideration the uncertainty of imputation.

The procedure of the chained equation is divided into four fundamental parts that are iterated until optimum outcomes are attained. The first stage is substituting any missing data using the mean value of the observable variables, serving as a temporary substitute. The second step is resetting these imputed means to the position of 'missing'. The third stage involves doing a regression analysis where the observed values of a variable, denoted as x, are treated as the dependent variable while the other variables are considered as independent variables. The fourth step is substituting the missing data with the predictions obtained from the regression model. The imputed value would thereafter be included as one of the independent variables, alongside the observed values for other variables. The process of steps 2 to 4 is then repeated for every variable that has missing values, forming a single iteration. The regressed predictions based on the observed data are used to replace all missing values after one iteration. The imputed values are substituted after each iteration, and the number of iterations can change.

3.4 Cluster selection

Within the domain of cluster selection, the process of determining the most suitable number of clusters is a crucial and often complex endeavor. The procedure is a methodical assessment of several clustering methods in order to choose the configuration that most effectively encompasses the inherent patterns within the dataset. This selection procedure is greatly aided by a number of quantitative scoring techniques, including the elbow method, silhouette score, Davies-Bouldin index, gap statistics, Calinski-Harabasz index, and information criteria like AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion). The scores serve as objective measurements, providing perceptions of the quality and consistency of various clustering results. By using these scores, researchers may make well-informed judgments on the optimal number of clusters that best match the inherent structure of the data. This allows for the optimization of the usefulness and interpretability of the resulting clusters. Nevertheless, it is essential to recognize that relying just on one scoring technique may not always provide a conclusive outcome. Instead, using a comprehensive strategy that integrates quantitative measures, qualitative evaluations, and domain expertise often generates more reliable and significant clustering outcomes. Sometimes the choice of the best number of clusters depends on the form of the problem proposed in the project, and finally, we ourselves have to decide how many clusters are the best. However, at first, we just used AIC score, BIC score, and Silhouette score, which we have discussed in detail below:

3.4.1 AIC

The Akaike Information Criterion (AIC) is a statistical methodology used for the purpose of selecting the most suitable model from a set of potential models, with the aim of addressing a decision issue rather than a hypothesis testing problem. The AIC is widely recognized as a valuable tool for selecting variables in multivariable modeling. Additionally, it serves as a significant aid in determining the most appropriate representation of explanatory variables that have been gathered. The Akaike Information Criterion (AIC) is a statistical metric that quantifies the quality of fit of a model by taking into account both the empirical probability and the total number of parameters included in the model. The empirical likelihood, denoted as L serves as a metric that quantifies the predictive accuracy of a model in relation to the data it was built upon. Put simply, a stronger correlation between variables leads to a higher probability. The equation used to compute the Akaike Information Criterion (AIC) is as follows[27]:

$$AIC = -2\log_e(L) + 2p \tag{3.2}$$

- L = Empirical Likelihood
- $\log_e$ = Natural Log Function
- P = Number of Parameters in the Model

In the above formula, it can be seen that higher likelihoods, indicative of a better fit, correspond to greater values of $log_e(L)$, thereby leading to lower values of $-2\log_e(L)$. Increasing the number of parameters (denoted as p) in the model results in a proportional increase in the penalty. The first term, referred to as the goodness-of-fit term, will exhibit a drop in correspondence with enhancements in the fitting of the model. Conversely, the subsequent term, known as the penalty term, will manifest an increase in response to the inclusion of more complexities in the model. Models with lower AIC values are generally considered more favorable since they strive to strike an optimum balance between accurately fitting the data and maintaining simplicity. AIC plays a crucial role in identifying the essential variables required for accurate prediction while avoiding the problem of overfitting the model. AIC may be seen as a metric that quantifies the degree of distinction between the candidate model that has been fitted and the model that is assumed to have created the data. Consequently, models with lower AIC values are considered to be closer to the underlying "truth."

3.4.2 BIC

The Bayesian Information Criterion (BIC) is a statistical metric used in the process of model selection and comparison. In the domain of clustering, it aids in the identification of the most suitable number of clusters to be used in a clustering process. The important concept is to strike a balance between model complexity (the number of parameters employed in the model) and goodness of fit (how well the model describes the data). The fundamental principle behind the BIC is the assessment of the data's probability in relation to the model. The likelihood measure provides a quantification of the extent to which the clustering model effectively accounts for the observed data. The metric quantifies the probability of seeing the actual data points given the anticipated clustering model. In the context of clustering, the likelihood often encompasses the joint probability of data points being assigned to their respective clusters according to the underlying model. BIC implements a penalty for the level of complexity shown by a model. The penalty term exhibits a direct proportionality to the number of parameters included inside the model. inside the realm of clustering, parameters include many elements such as the number of clusters, the centers of these clusters, and maybe more characteristics that delineate the clusters or the distribution of data inside them.

The following is the formula for the BIC:

$$BIC = -2\log_e(L) + p\log_e(N) \tag{3.3}$$

- L = Empirical Likelihood
- $\log_e$ = Natural Log Function
- P = Number of Parameters in the Model
- N = The number of data points included inside the dataset.

In the context of cluster selection, it is customary to use the Bayesian Information Criterion (BIC) to evaluate and compare clustering solutions characterized by various numbers of clusters. The BIC value is computed for each clustering solution using the aforementioned algorithm. The objective is to determine the optimal number of clusters that results in the lowest BIC score. In essence, the optimal clustering solution is chosen based on achieving an optimal balance between the suitability of the model in explaining the data (model fit) and the level of complexity involved in the model (model complexity), including the number of clusters and associated parameters.

A decrease in BIC values suggests an improved balance between the ability to explain the observed data and the level of complexity in the model. Hence, the selection of the number of clusters that yield the minimum BIC is often regarded as the ideal decision. BIC is inclined towards simpler models that include fewer clusters, while simultaneously penalizing excessively complex models. Consequently, BIC serves as a valuable tool in the context of clustering analysis by mitigating the risk of overfitting.

3.4.3 Silhouette score

Various techniques exist for assessing the quality of clustering outcomes, including the Rand index, corrected Rand index, distortion score, and Silhouette index. Although the majority of performance assessment techniques often need a training set, the Silhouette index stands out as an exception, since it does not necessitate a training set for evaluating clustering outcomes. This modification makes it more suitable for clustering purposes. The Silhouette index is used in this study to assess the clustering performance. The following is the definition of the silhouette width $s(x_i)$ for the point x_i[23]:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{b(x_i), a(x_i)\}} \tag{3.4}$$

- x_i : A data in cluster π_k
- $a(x_i)$: The average distance of x_i to all other elements in the cluster π_k (in the context of dissimilarity)
- $b(x_i)$: $min\{d_1(x_1)\}$, among all clusters $l \neq k$
- $d_1(x_1)$: The average distance from x_i to all points in cluster π_l and $l \neq k$ (between dissimilarity)

The average distance, measured in terms of "between dissimilarity", between each point in cluster π_l and x_i. The Silhouette width, as shown by the formula, exhibits a range of values spanning from -1 to 1. A negative value is considered unfavorable due to its association with a scenario where $a(x_i)$ is greater than $b(x_i)$, resulting in a higher level of dissimilarity within a group compared to the dissimilarity between groups. A positive value is acquired when the value of $a(x_i)$ is less than $b(x_i)$, and the Silhouette width achieves its maximum value of $s(x_i)$=1 when $a(x_i) = 0$. A component has a better chance of being clustered in the right group if its (positive) s(xi) value is larger. Elements exhibiting negative $s(x_i)$ values are more prone to being concentrated inside incorrect clusters.

3.5 Experimental results

When we did some statistics on our dataset we found that there were 458 patients in the dataset that had no missing data, we called it the Full Data Patients dataset. So, we decided to analyze the Full Data Patients dataset and the original dataset with imputation separately.

3.5.1 Full Data Patients

As our dataset contains the heights and weights of patients we computed the BMIs of all patients using the formula that has been mentioned below:

$$BMI = \frac{Weight(in\ Kilograms)}{Height^2(in\ meters)} \tag{3.5}$$

Also, we created three different subsets of our dataset, namely, full (boys and girls together), only girls, and only boys. Then, we computed BMIs of each subset dataset and due to the presence of substantial changes in the range and distribution of features within a dataset we standardized BMIs using Min-Max Scaler in the preprocessing step to convert our dataset into a standard format so that they have a standard deviation of one and a mean of zero that is crucial and helps to improve the performance, stability, and interpretability of the clustering. The next step is clustering and we chose the Gaussian Mixture Model to cluster our data because the distribution of our dataset was a mixture of Gaussian. After that, we considered the number of clusters in the range of two to nine. Then we compared their results and finally found the best results for the numbers of clusters that were two and three according to different types of evaluation scores which are AIC, BIC, and Silhouette scores.

BIC scores have been shown in Figure 3.3 for full patients, Figure 3.4 for girls, and Figure 3.5 for boys. also, AIC scores can be seen in the following Figures which 3.6 stands for full patients, 3.7 for girls, and 3.8 for boys. In addition to BIC and AIC scores, there are Silhouette scores which can be seen in Figures 3.9 for full patients, 3.10 for girls, and 3.11 for boys. The analysis of these criteria is as follows: first, we checked the values of AIC and BIC for all the clusters together, and wherever both of these criteria have the lowest value, that value was set to the best number of clusters. In the subsequent stage, the evaluation of the silhouette score is conducted. Unlike AIC and BIC, a higher silhouette score indicates a more optimal number of clusters. In the final stage, the optimal number of clusters is determined by collectively evaluating these three criteria. This selection is based on the premise that the model achieves the most favorable number of clusters as indicated by these three criteria. We analyzed different subsets of our dataset separately. For the full patients, according to the plots, we understood that the three clusters case is the best model because BIC and AIC scores had the lowest values for three clusters, and the Silhouette score had the

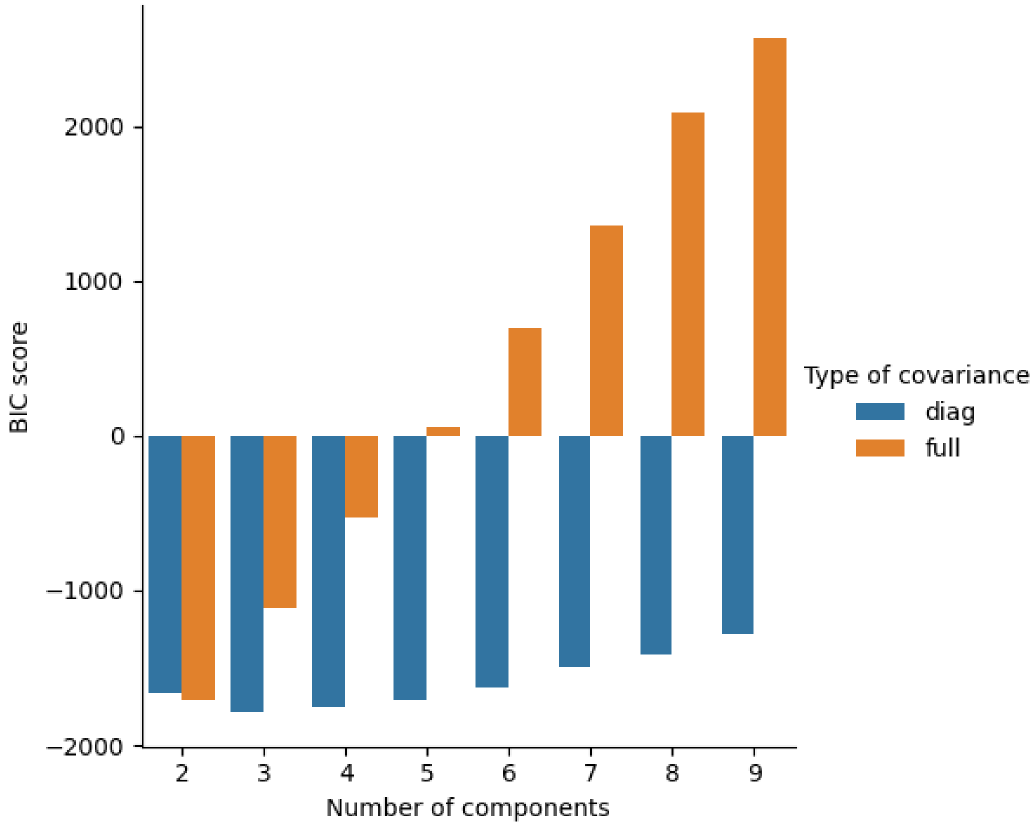


Figure 3.3: BIC scores of each number of clusters for Full dataset(Full Data Patients)

highest values for two and three clusters respectively, so we decided to choose three clusters based on these scores. For girls dataset, we selected three clusters, as the AIC and BIC scores had the lowest values for three clusters and Silhouette scores had the highest values for two and three clusters. But for boys, we chose two clusters because AIC and BIC both had the lowest values for two clusters and the Silhouette score had the highest value for two clusters.

After finding the best number of clusters of each subset of our dataset we visualized the data in each cluster, you can see the trajectories of our dataset in each cluster in the following Figures, 3.12 is related to full patients, 3.13 for girls, and 3.14 for boys.

Then to find how each patient can jump from one cluster to another at different time points, first we computed the mean of trajectories of each cluster and then chose five random patients. For each randomly selected patient at each time point, we compared the value of its BMI with the mean of different

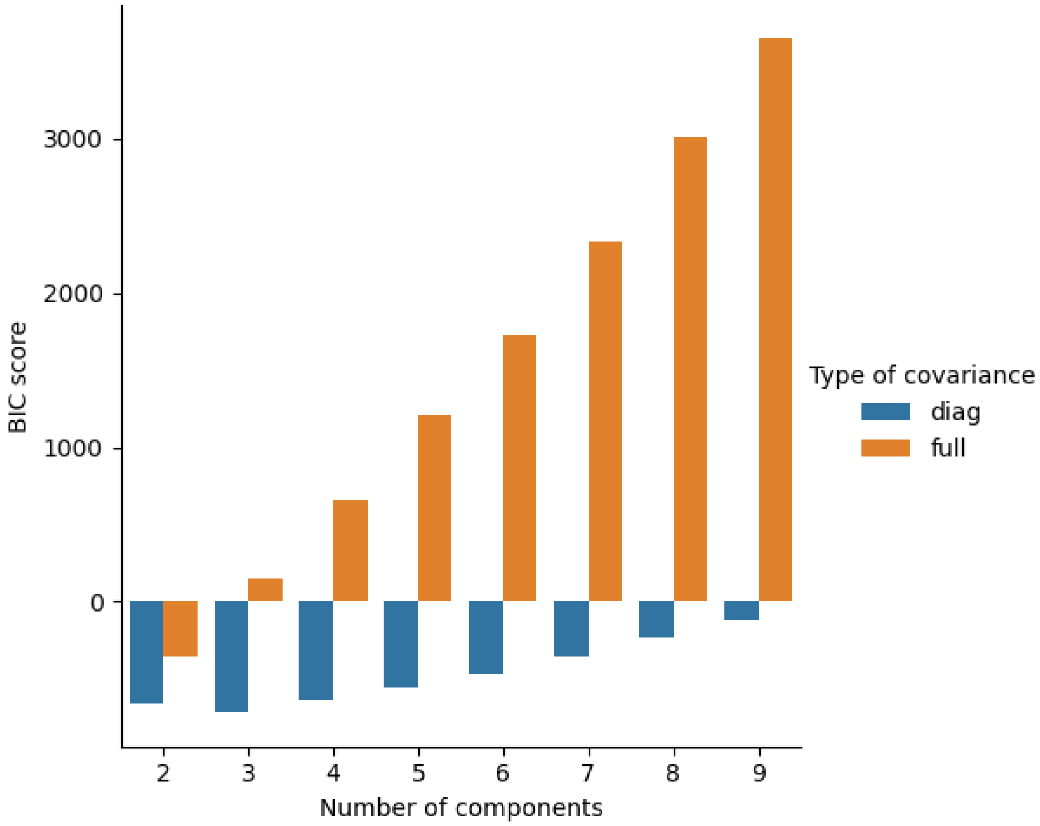


Figure 3.4: BIC scores of each number of clusters for girls dataset(Full Data Patients)

clusters' trajectory by computing Euclidean Distance and forgetting factor using the formula that has been mentioned below:

$$d = \sum_{i=1}^{n} w_i(x_i - y_i)^2 \qquad (3.6)$$

- x_i: BMI of a patient in time point i
- y_i: Mean of a trajectories in time point i
- $w_i = e^{-\lambda(t)}$: Forgetting factor in time point t

It is worth mentioning that the forgetting factor is a method used in time-series analysis and forecasting that involves assigning weights to previous data that decrease exponentially with time. This indicates that observations made more recently have a greater weight in shaping the present forecast or estimate, but observations made in the past have progressively less influence as time

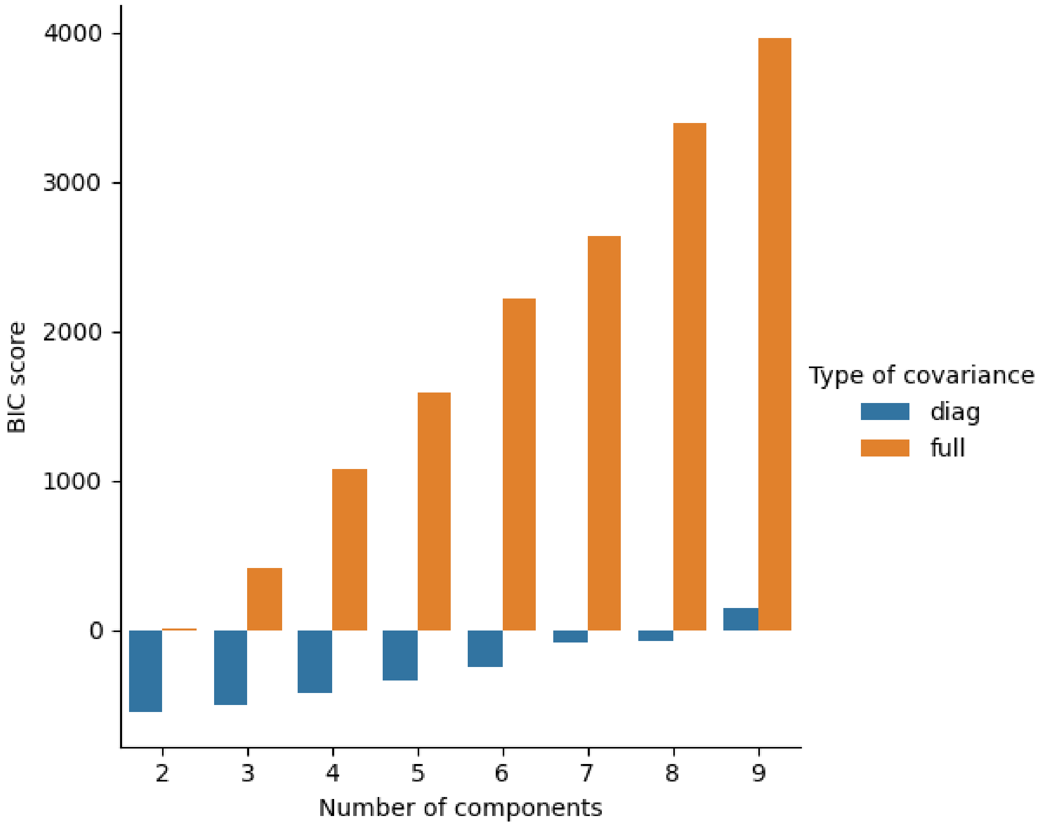


Figure 3.5: BIC scores of each number of clusters for boys dataset(Full Data Patients)

goes on. So, we used the forgetting factor to have more precise results. After calculating the Euclidean distance for different BMI values of each patient with the mean of trajectories at each time point, the data is assigned to a cluster that has a lower Euclidean distance.

We decided to observe two different values of λ including 0 and 0.5 to find out how it affects patients jumping from one cluster to another at different time points. We chose $\lambda = 0.5$ as a middle range and 0 to have equal weights to find Euclidean Distances. We selected five random patients and showed them with the different mean trajectories. Figure 3.15 refers to girls with $\lambda = 0$, Figure 3.16 for girls with $\lambda = 0.5$, Figure 3.17 for boys with $\lambda = 0$, Figure 3.18 for boys with $\lambda = 0.5$.

Also, an example of patient cluster changing for the five random patients has been shown in Table 3.19 for girls and Table 3.20 for boys, the first column shows the patient ID and the second and third columns show the changes of clusters

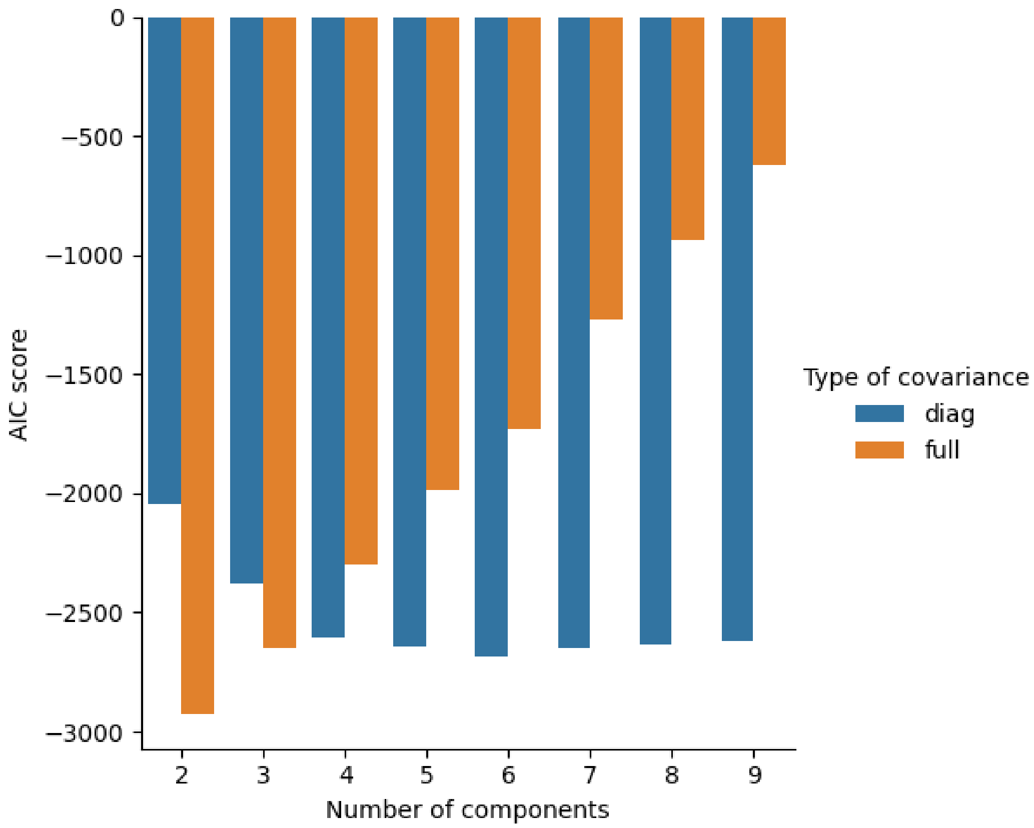


Figure 3.6: AIC scores of each number of clusters for full dataset(Full Data Patients)

for each random patient with $\lambda = 0.5$ and $\lambda = 0$.

3.5.2 Imputed Dataset

In the previous section, we clustered and analyzed the Full Data Patients dataset but in this section, we are analyzing the original dataset with missing data and imputed missing values. The total number of data was n = 3,897 including girls (n = 1931), and boys (n = 1966). We imputed the missing data using different imputation approaches such as Mean Expected Growth, Linear Interpolation, Forward-Fill, Backward-Fill, and MICE but found that the best approach was Mean Expected Growth which had better results, for instance in the case of MICE imputation When we compared the results of its clustering, which included the trajectories of the patients, with the gold standard, we realized that the behavioral pattern of the patients after imputation is very different from the gold standard, and the BMI of the patients often has many

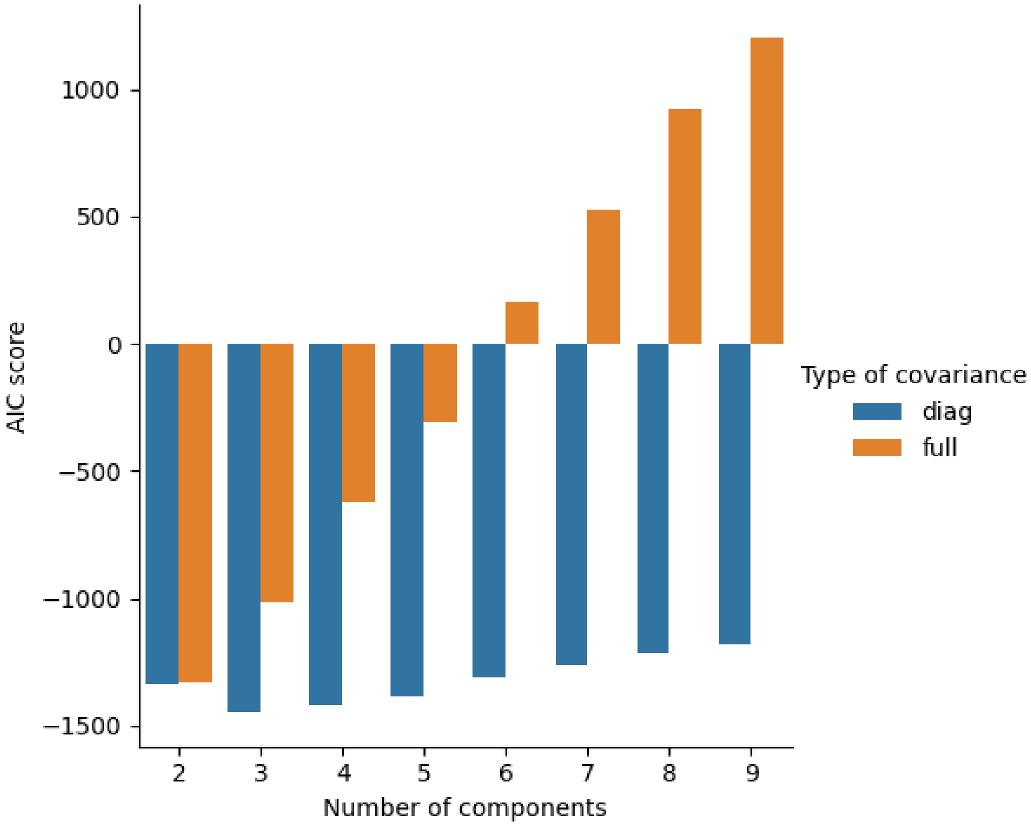


Figure 3.7: AIC scores of each number of clusters for girls dataset(Full Data Patients)

changes and ups and downs over time while we did not see such changes in the results of Mean Expected Growth and chose it as a best imputation approach in our project. Since our dataset contained heights and weights in 14 time points, at first we divided the dataset into two subsets of girls and boys then imputed them using Mean Expected Growth. The missing values in the dataset were approximated by computing the mean of the longitudinal data at each time point. This allowed us to determine the growth rate of the data between consecutive time points using the equation that has been mentioned before as follows:

$$x_i^c = x_{i-1}^{c-1} + (Mean^c - Mean^{c-1}) \tag{3.7}$$

- x_i^c: Missing data for a patient at time point $''i''$
- x_{i-1}^{c-1}: A data which refers to the previous time point $''i-1''$ for a patient
- $Mean^c$: Mean of data over a column at time point i

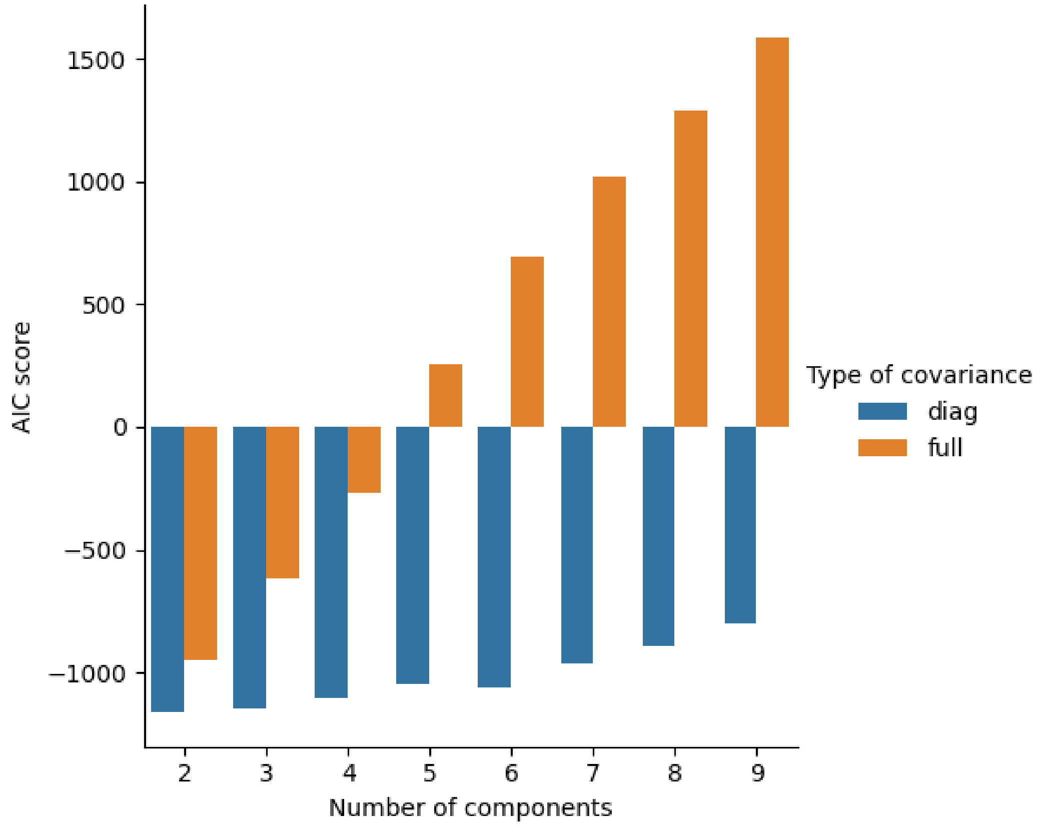


Figure 3.8: AIC scores of each number of clusters for boys dataset(Full Data Patients)

- $Mean^{c-1}$: Mean of data over a column at time point $i - 1$

After that, we calculated the BMIs of all patients using heights and weights using the formula that has been written below:

$$BMI = \frac{Weight(in\ Kilograms)}{Height^2(in\ meters)} \tag{3.8}$$

We considered three different subsets of our dataset, the same as for Full Data Patient in the previous section namely, full(girls and boys together, n = 3,897), only boys(n = 1966), and only girls(n = 1931). In the preprocessing step, we standardize the dataset to convert them into a standard form. After that, we clustered our dataset using Gaussian Mixture Models as a soft clustering method with the different number clusters between two and nine, then compared their results together and found the best number of clusters based on BIC, AIC, and Silhouette scores which was three for all the subsets of the dataset.

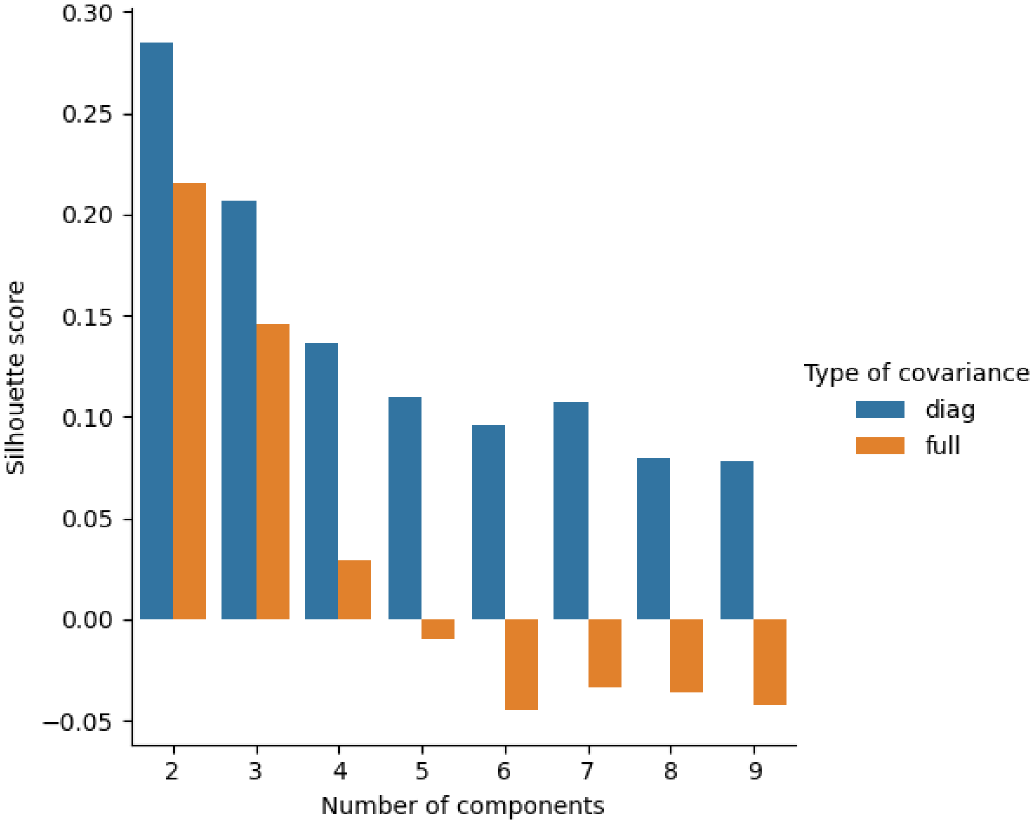


Figure 3.9: Silhouette scores of each number of clusters for full dataset(Full Data Patients)

BIC scores can be seen in figures 3.21 for full patients, 3.22 for girls, and 3.23 for boys, we have shown these scores for AIC in the same way in figures 3.24 for full patients, 3.25 for girls and, 3.26 for boys. Also, we got Silhouette scores for full patients in figure 3.27, girls in figure 3.28, and boys in figure 3.29. Same as the previous section we have, the smaller the BIC and AIC, the better the models, and the larger the Silhouette score, the better the model, so according to this and the problem definition and requirements we have chosen the best number of clusters for each subset of the dataset n this way, we considered the number of clusters for the full data to be three, three for girls, and three for boys, but as the result of the full dataset was not satisfying, we analyzed girls and boys separately.

After choosing the best number of clusters, here, like the previous procedure of Full Data Patient we visualized the trajectories of data in each cluster that can be seen in the following figures, 3.30 for full (girls and boys together), 3.31 for

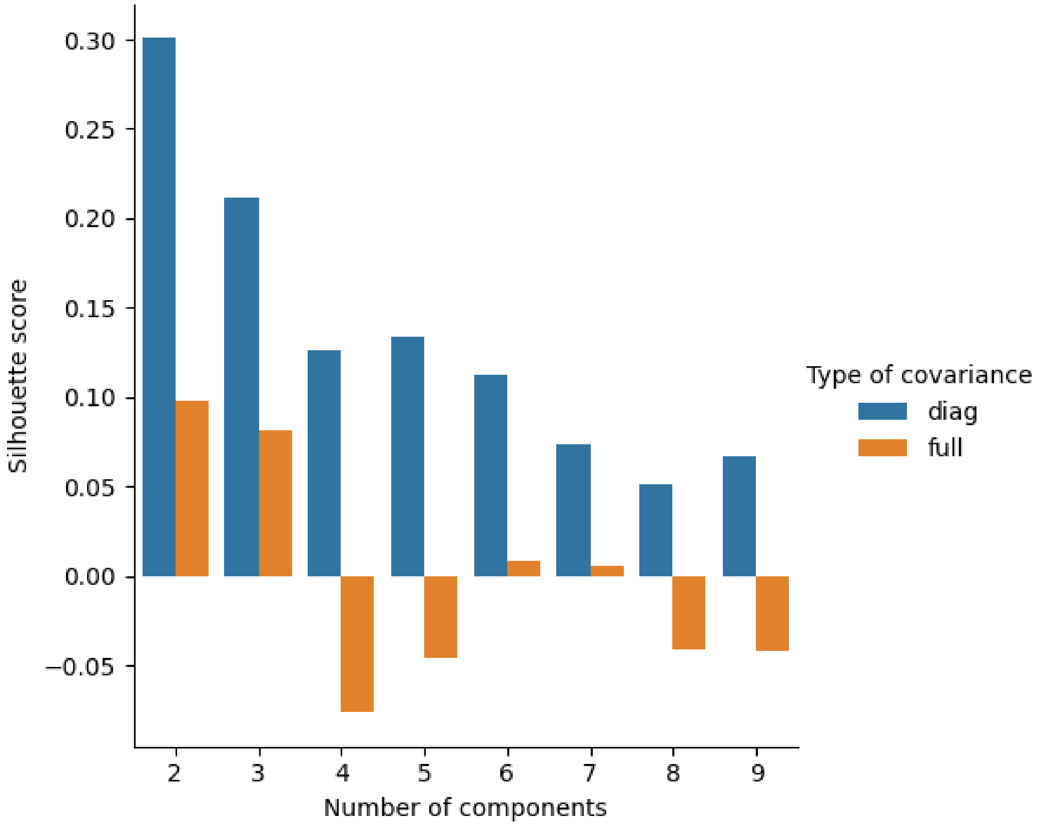


Figure 3.10: Silhouette scores of each number of clusters for girls dataset(Full Data Patients)

girls, and 3.32 for boys. Since we wanted to know how each patient changes their clusters from one time point to another one, first we computed the mean of trajectories of each cluster and then chose five random patients to show the mean of trajectories and random patients together to analyze it better. In this step, we wanted to compare the BMIs of each random patient with the BMIs of the mean of trajectories using Euclidean Distance and forgetting factor(weights) through the equation below:

$$d = \sum_{i=1}^{n} w_i (x_i - y_i)^2 \tag{3.9}$$

- x_i: BMI of a patient in time point i
- y_i: Mean of a trajectories in time point i
- $w_i = e^{-\lambda(t)}$: Forgetting factor in time point t

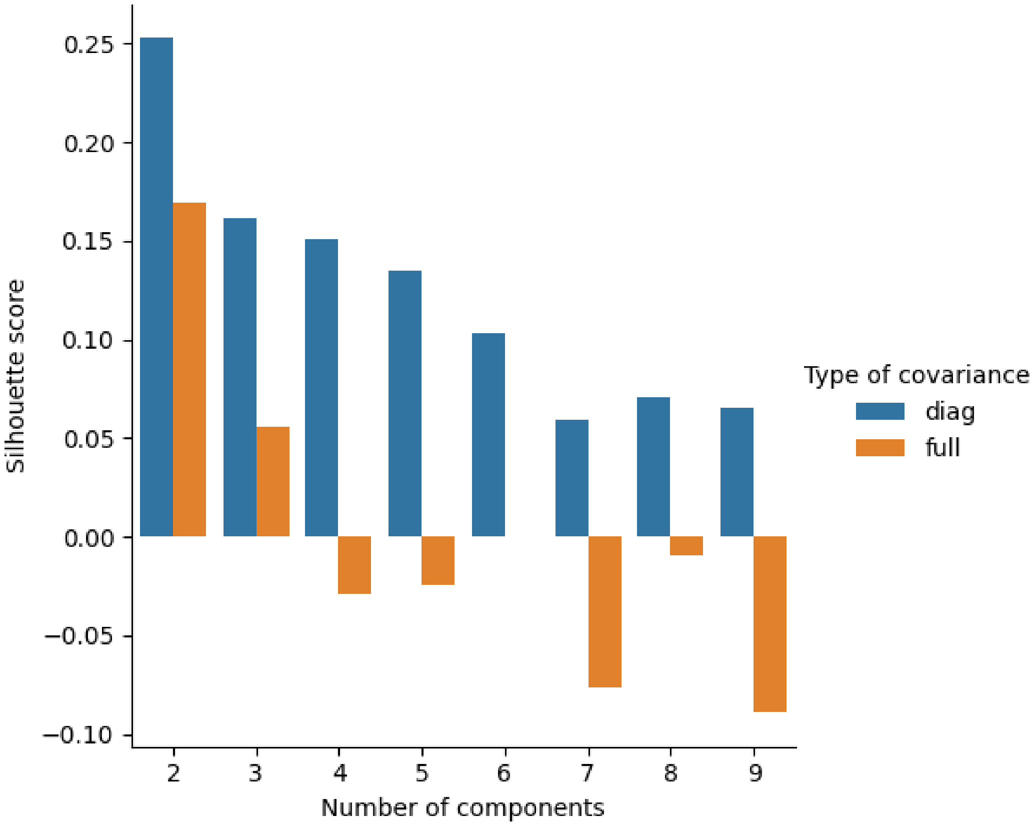


Figure 3.11: Silhouette scores of each number of clusters for boys dataset(Full Data Patients)

Everything about the forgetting factor method has been discussed in the previous section, so based on that, we found the Euclidean Distances for each BMI of random patients with respect to the BMIs of mean trajectories of different clusters in 14 time points and compared them together. Each patient is assigned to a cluster whose BMIs are less far apart(has the lowest Euclidean distance).

Choosing the λ was challenging for us. λ should be in the range of zero and one, so we decided to choose the middle range which is 0.5. When we had a $\lambda = 0$, equal weights were given to the earlier data so it was interesting to understand how random patients change their clusters with equal weights. You can find the figures of Five random patients with different λ (0 or 0.5) for each subset of the dataset, Figure3.34 refers to girls ($\lambda = 0$), and Figure 3.33 refers to girls ($\lambda = 0.5$), Figure 3.36 for boys ($\lambda = 0$), and Figure 3.35 for boys ($\lambda = 0.5$).

Furthermore, we presented an illustration of the alteration in patient clusters for the five randomly selected patients. Specifically, Figure 3.37 pertains to girls,

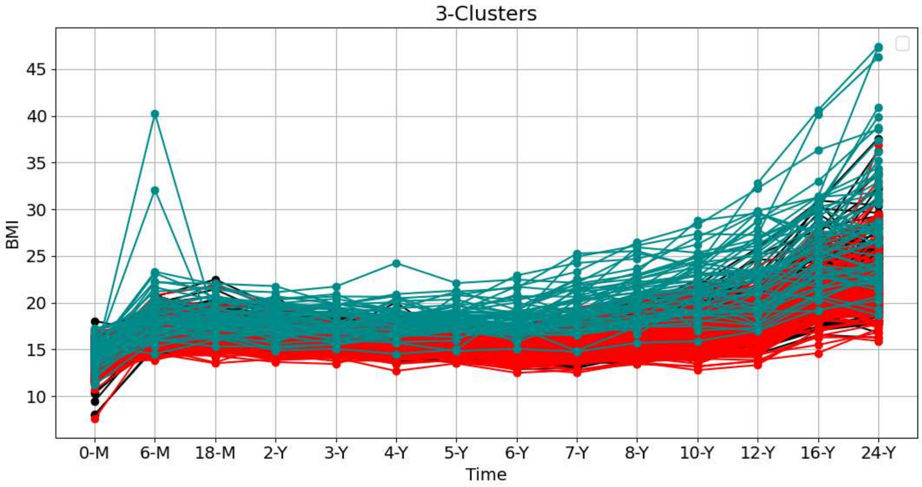


Figure 3.12: Trajectory of full dataset(girls and boys together) for Full Data Patients

while Figure 3.38 pertains to boys. The first column of these figures denotes patient ID, while the second and third columns represent the modifications in clusters for each random patient, with $\lambda = 0.5$ and $\lambda = 0$.

3.6 Results Discussion

This prospective cohort study utilized population-based longitudinal data from the ongoing Swedish birth cohort (born between 1994-96) called BAMSE[2]. Height and Weight measurements were collected from a combination of clinical follow-ups and merging school records after taking relevant consent from the study subjects. In the current study, three and two trajectories were identified in sex-stratified analysis in the non-missing dataset, Figure3.13 for girls and Figure3.14 for boys respectively, (considered the gold standard) where all fourteen time points were available.

The current study also provides benefits of imputing missing data as this issue has a big impact in longitudinal studies and needs to be addressed. With as much as 67% data at one time point, we evaluated linear interpolation, forward-fill (Ffill) and backward-fill (Bfill), Mean Expected Growth, and MICE. Some informative statistics are shown in Tables 3.1 and 3.2 related to heights and weights respectively. As can be seen in Table3.1 the highest number of missing height data is related to the age of 4-years which is 2,630 and the lowest

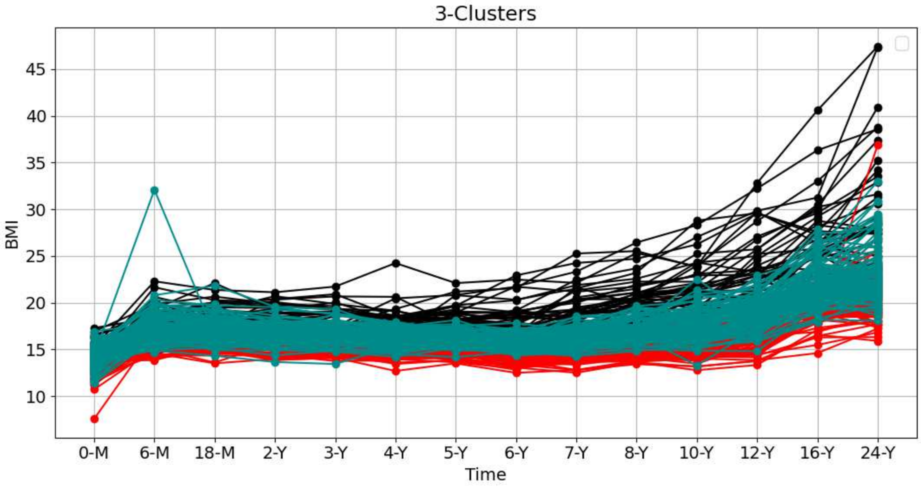


Figure 3.13: Trajectory of girls for Full Data Patients

number of missing height data is related to the age of 0-month which is 365, in Table3.2 the highest number of missing weight data is related to the age of 4-years which is 2601 and the lowest number of missing weight data is related to the age of 0-month which is 39, and we visualized all the data to have a graphical representation of missing data which showed missing data in white color and non-missing data in black color using the missing data matrix Figure3.1. We found that the pattern of missingness in our dataset is MCAR(Missing Completely At Random).

Given the size of our dataset, the compute time was similar for all methods. We compared these methods' performances and found the mean expected growth approach to give the best results, for instance in the case of MICE imputation When we compared the results of its clustering, which included the trajectories of the patients, with the gold standard, we realized that the behavioral pattern of the patients after imputation is very different from the gold standard, and the BMI of the patients often has many changes and ups and downs over time while we did not see such changes in the results of Mean Expected Growth and chose it as the best imputation approach in our project, although there are sensitivity analyses that could have been performed furthermore, but due to lack of time, we didn't manage to undertake them.

Another aspect of undertaking these longitudinal studies is the difference in the time points for data collection, their coverage of the growth spurt years and what impact this can have on the post-adolescence BMI trajectories. The widely

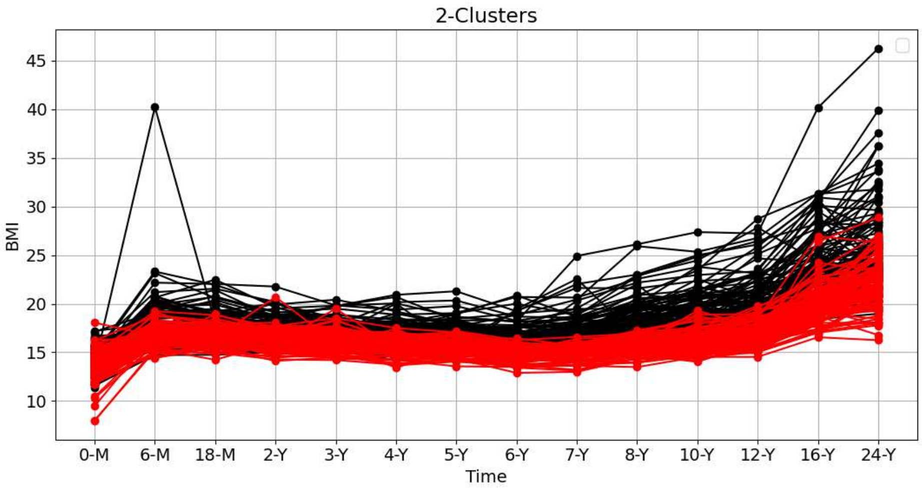


Figure 3.14: Trajectory of boys for Full Data Patients

accepted children growth spurt phases between 2-3 years and 6-7 years, show a sex-stratified effect in our gold standard and fully imputed dataset alike. Of the three clusters that we identify in the mean of girls clustering trajectories, the highest mean group (cluster 0 with 44% samples in fully imputed) doesn't see any significant dip after the 2-3 year phase Figure3.15 and goes on a steep increasing AR (adiposity round) after 6-years till 24 years time point in gold standard or fully imputed dataset, while cluster 1 and cluster 2 does see a dip after the 2-3 year phase, with nominal growth spurt at 6-year time point and a noticeable AR at 12-year timepoint. This significant finding adds further evidence to the OECD guidance about high BMI risks and interventions (World Health Organization, 2016 [20]) by targeting early life factors e.g. gestational smoking, breastfeeding, and control of pre–gestational and gestational BMI[22]. The same growth dip is again not noticeable in the highest mean group (cluster 0 with 26% samples in fully imputed) for boys, and their AR gains are noticeable at 6-year and 12-year time points. One further factor that could influence these trajectories is the significance of these growth spurts and any differences in measurement time points.

Most other studies don't account for these phenomena or give equal weight to every instance. While, we looked into the measure of forgetting factor, and how it can be used as a weighting technique to account for variation at the different measurement time points like growth spurts, different puberty impacts on boys and girls, impact of other linked known phenotypes like asthma, etc. In the

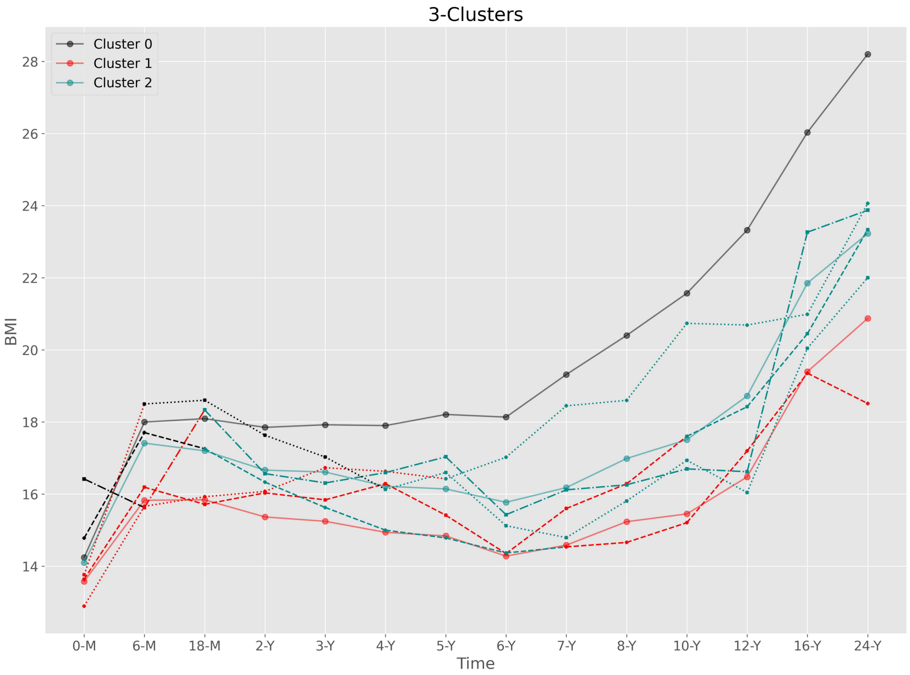


Figure 3.15: Mean of girls clustering trajectory with five random patients and λ=0 for Gold Standard (Full Data Patients)

current study, we looked at one alternative (λ=0.5) apart from the default (λ=0) and we observed changes in clustering patterns of some samples (Figure3.33), where they get allocated to a different cluster at λ=0.5, compared to λ=0 to jump. As a next step, we could look into how we can implement a biological link driven weighting technique for improving these sample clustering at each time point to create a more continuous growth curve. Since we wanted to know how each patient changes their clusters from one time point to another one, first we computed the mean of trajectories of each cluster and then chose five random patients to show the mean of trajectories and random patients together to analyze it better. In this step, we wanted to compare the BMIs of each random patient with the BMIs of the mean of trajectories using Euclidean Distance and forgetting factor(weights) through the equation below:

$$d = \sum_{i=1}^{n} w_i (x_i - y_i)^2 \tag{3.10}$$

- x_i: BMI of a patient in time point i

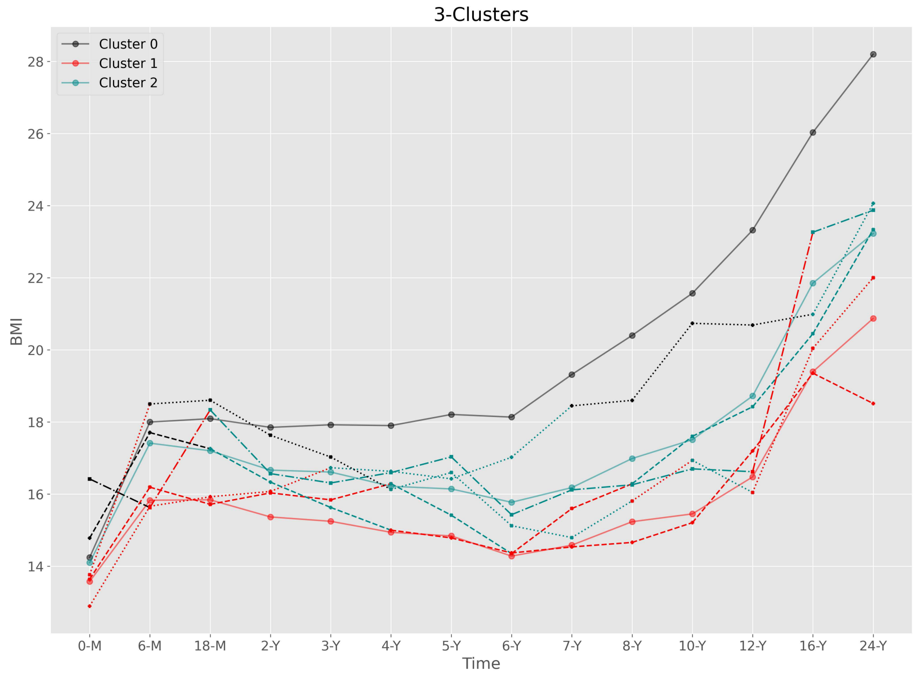


Figure 3.16: Mean of girls clustering trajectory with five random patients and λ = 0.5 for Gold Standard (Full Data Patients)

- y_i: Mean of a trajectories in time point i
- $w_i = e^{-\lambda(t)}$: Forgetting factor in time point t

In the equation that has been mentioned choosing the best values for λ was challenging for us, as the range of λ is between zero and one, we chose the middle range which is 0.5.

forgetting factor is a method used in time-series analysis and forecasting that involves assigning weights to previous data that decrease exponentially with time. This indicates that observations made more recently have a greater weight in shaping the present forecast or estimate, but observations made in the past have progressively less influence as time goes on. So, we used the forgetting factor to have more precise results. After calculating the Euclidean distance for different BMI values of each patient with the mean of trajectories at each time point, the data is assigned to a cluster that has a lower Euclidean distance.

To find the best number of clusters we evaluated different criteria including AIC, BIC, and Silhouette scores that have been explained and analyzed before in detail. The smaller the BIC and AIC, the better the models, and the larger the

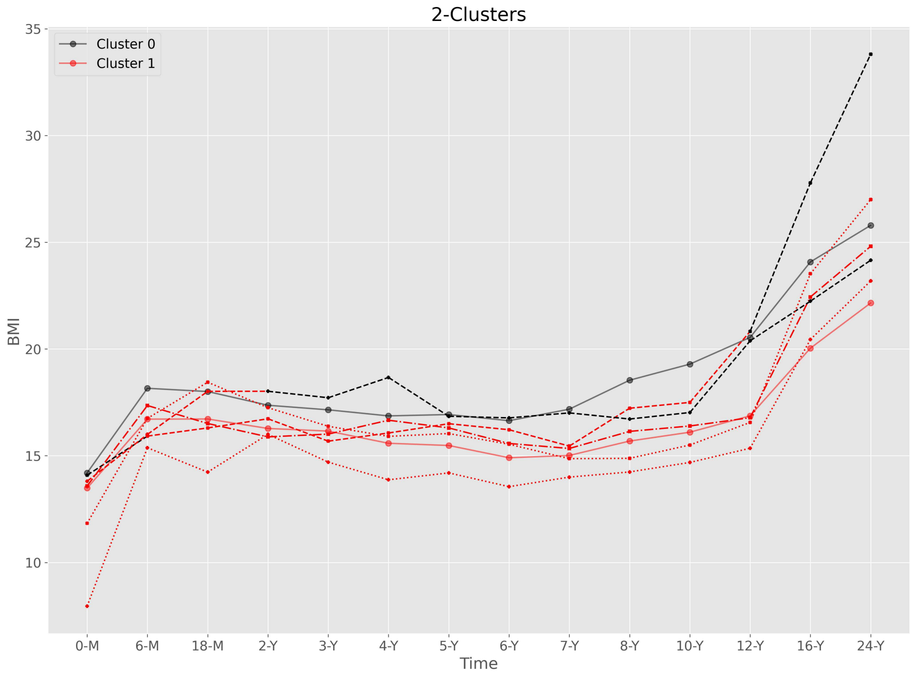


Figure 3.17: Mean of boys clustering trajectory with five random patients and λ = 0 for Gold Standard (Full Data Patients)

Silhouette score, the better the model. Based on these criteria we chose three clusters for the girls dataset and two clusters for the boys dataset in our gold standard and chose three clusters for the girls dataset and three clusters for the boys dataset for the Imputed dataset.

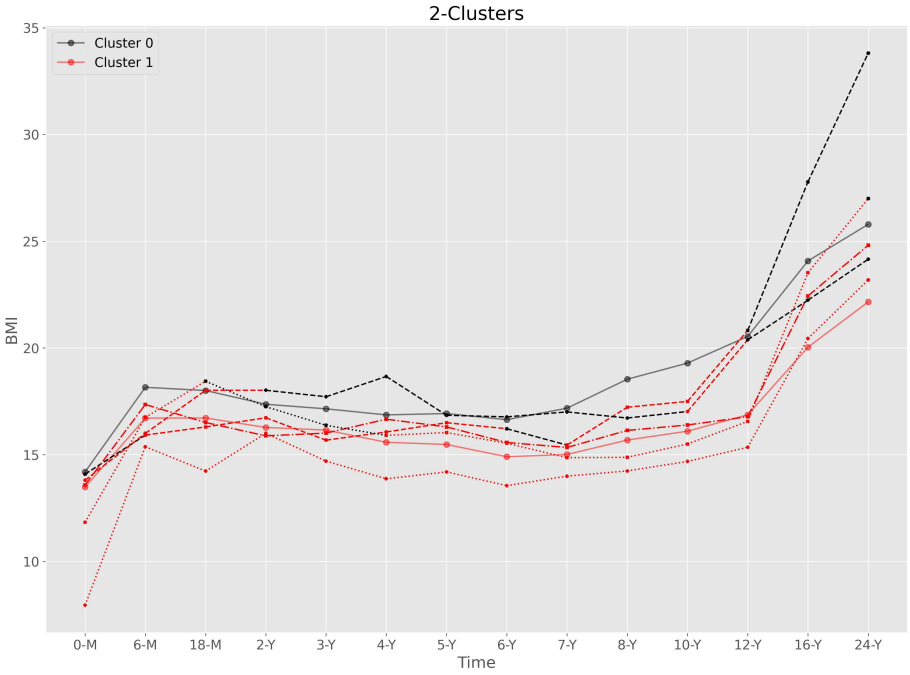
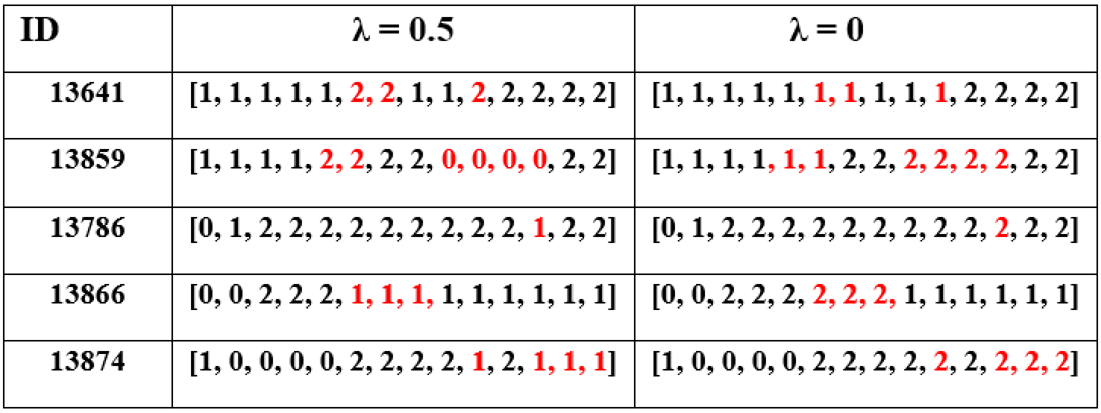
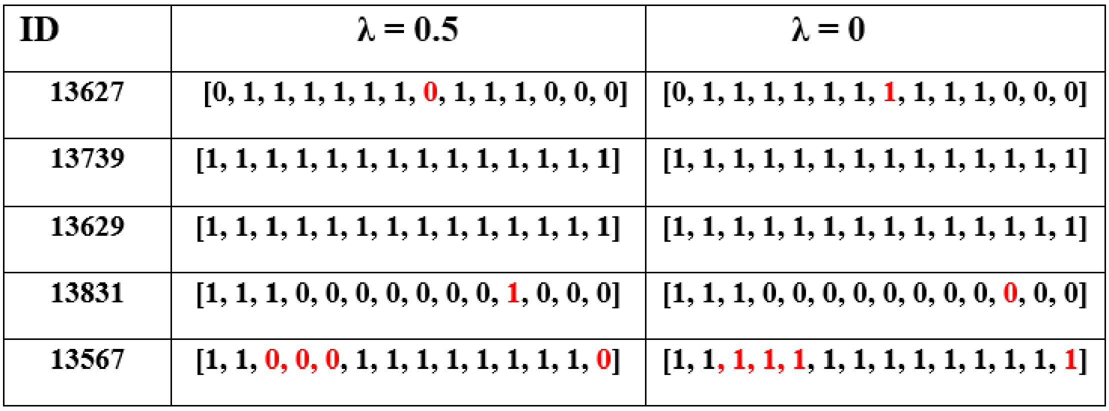


Figure 3.18: Mean of boys clustering trajectory with five random patients and λ = 0.5 for Gold Standard (Full Data Patients)



ID	λ = 0.5	λ = 0
13641	[1, 1, 1, 1, 1, 2, 2, 1, 1, 2, 2, 2, 2, 2]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2]
13859	[1, 1, 1, 1, 2, 2, 2, 2, 0, 0, 0, 0, 2, 2]	[1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2]
13786	[0, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2]	[0, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]
13866	[0, 0, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1]	[0, 0, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1]
13874	[1, 0, 0, 0, 0, 2, 2, 2, 2, 1, 2, 1, 1, 1]	[1, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2]

Figure 3.19: Patient cluster Changing of five random Patients for girls dataset(Full Data Patients)



ID	λ = 0.5	λ = 0
13627	[0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0]	[0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0]
13739	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
13629	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
13831	[1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]	[1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
13567	[1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Figure 3.20: Patient cluster Changing of five random Patients for boys dataset(Full Data Patients)

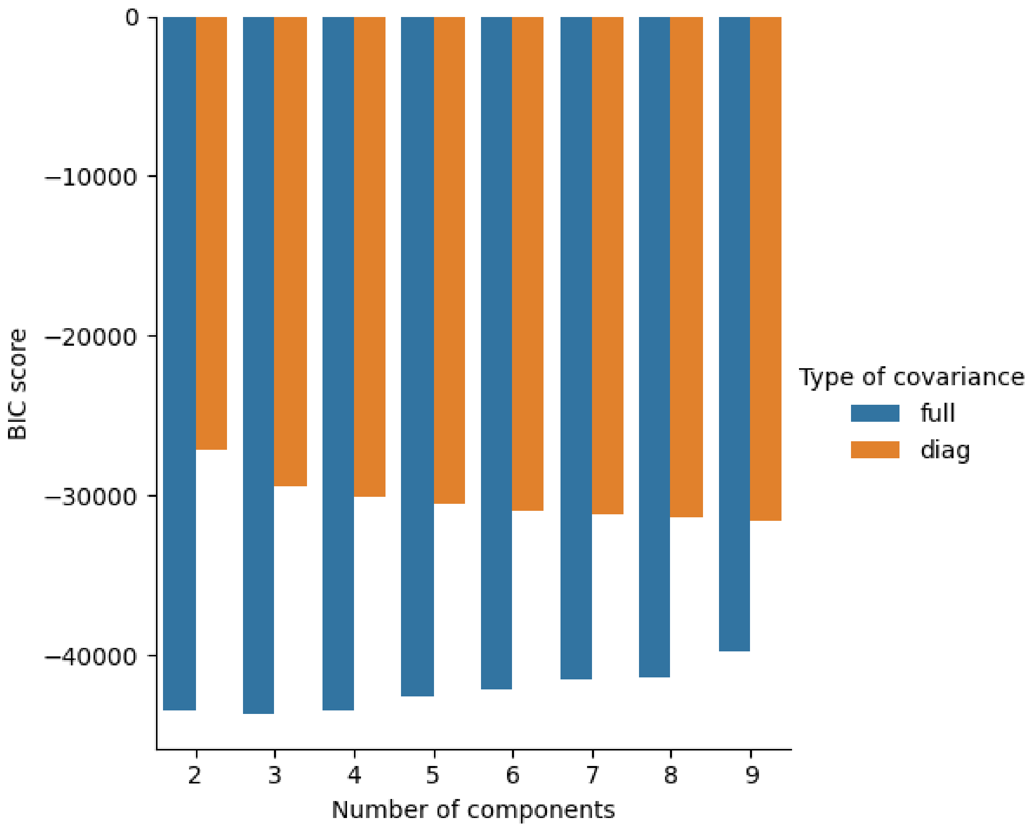


Figure 3.21: BIC scores of each number of clusters for Full dataset(Imputed Dataset)

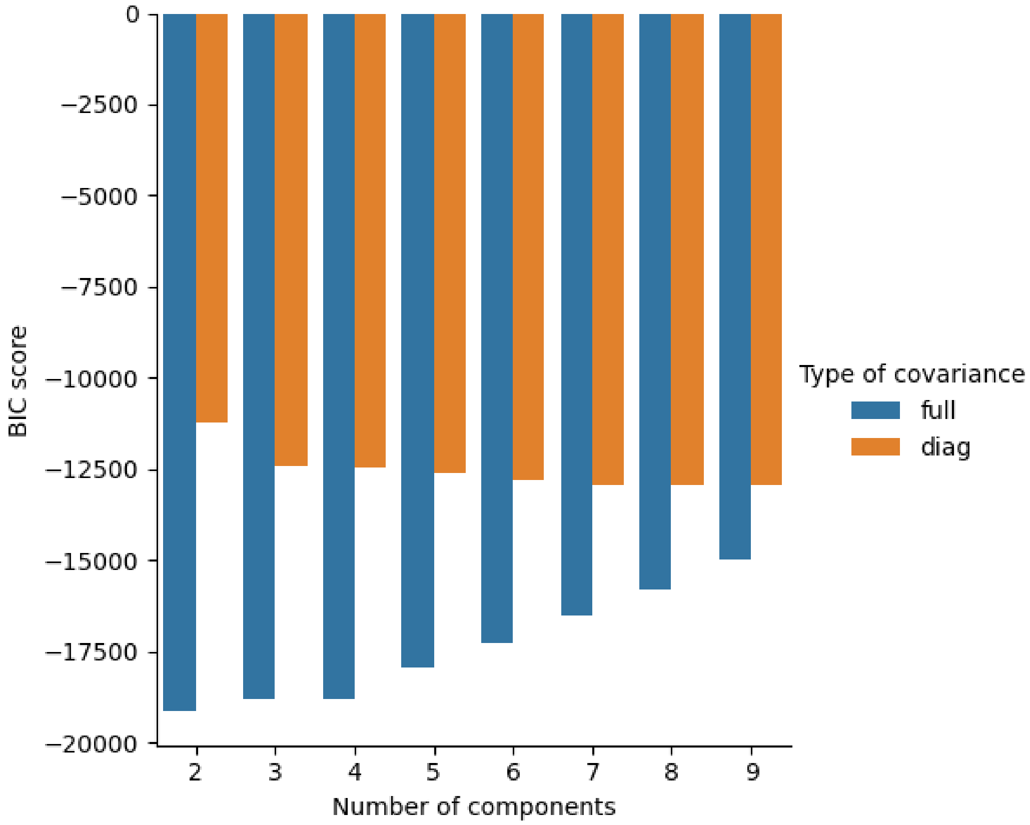


Figure 3.22: BIC scores of each number of clusters for girls dataset(Imputed Dataset)

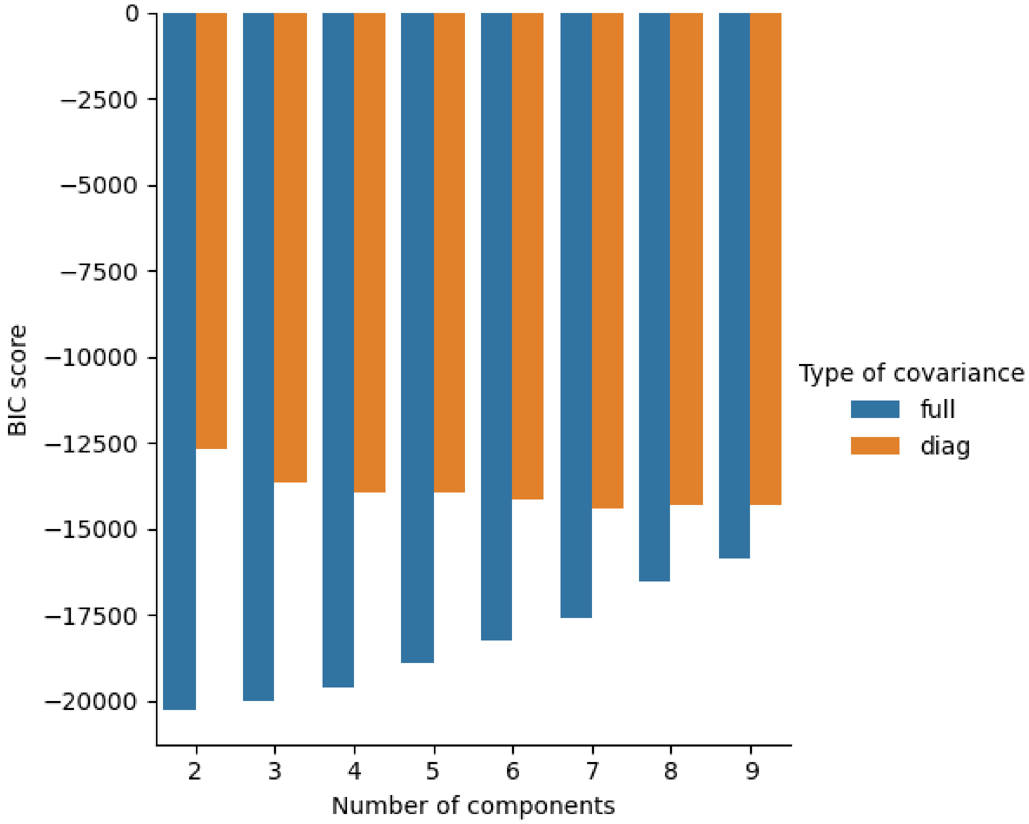


Figure 3.23: BIC scores of each number of clusters for boys dataset(Imputed Dataset)

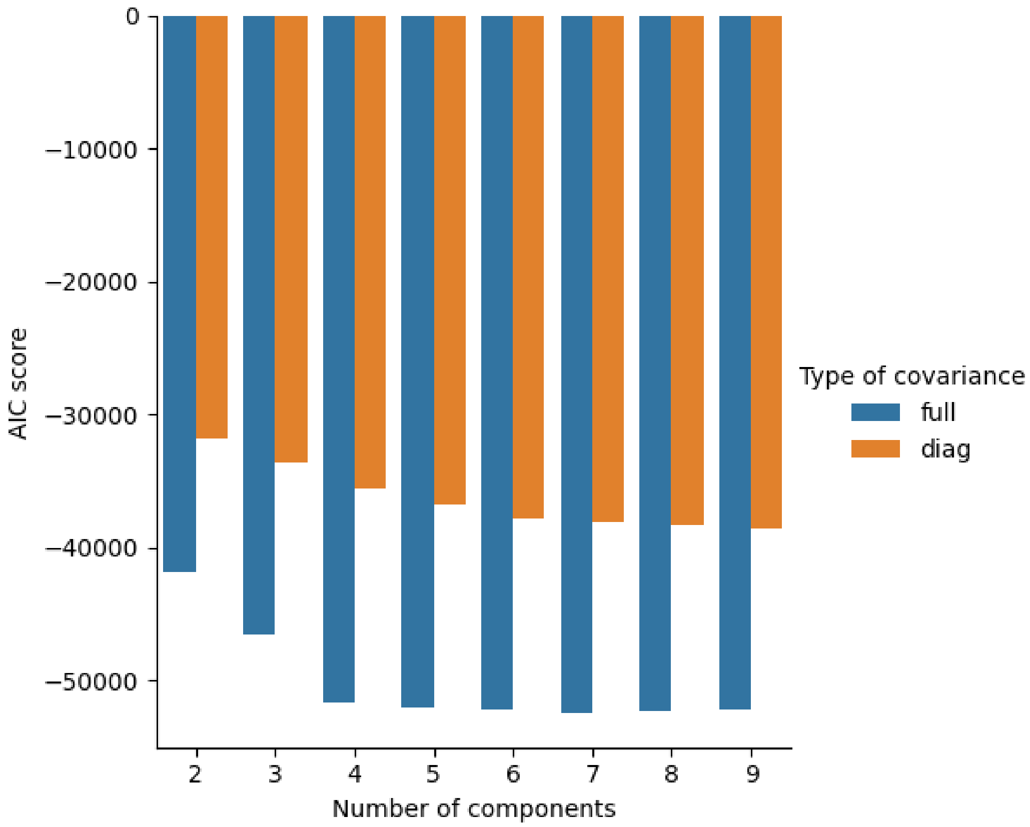


Figure 3.24: AIC scores of each number of clusters for Full dataset(Imputed Dataset)

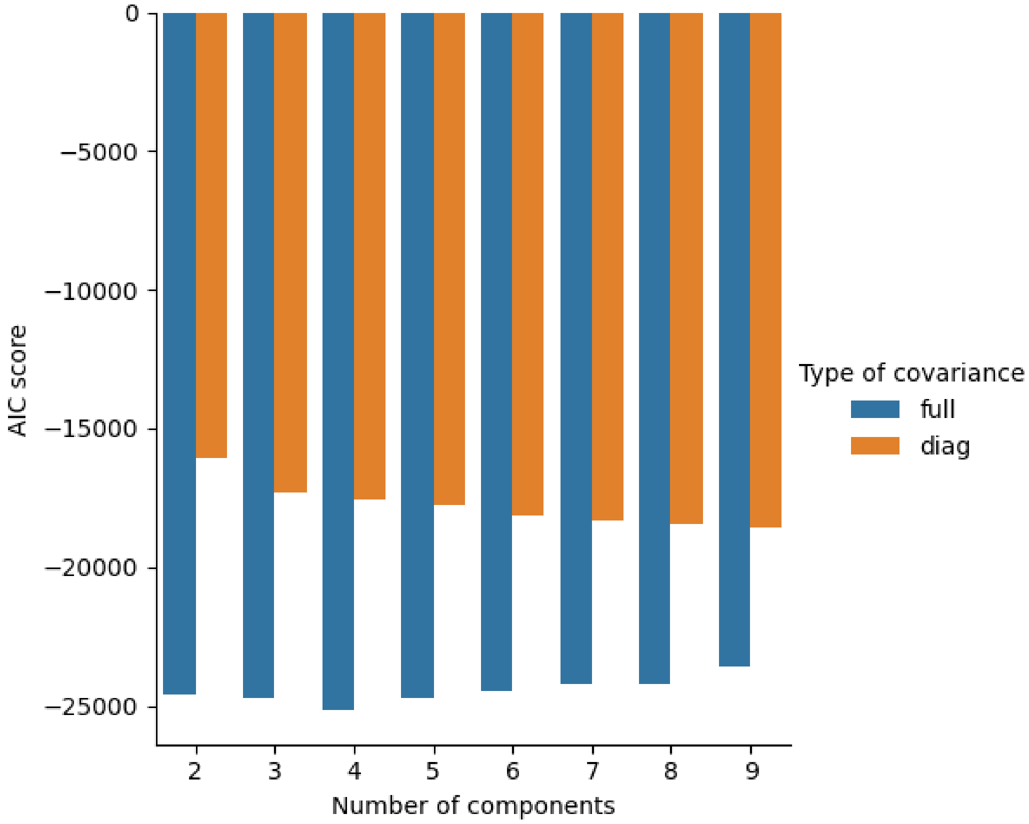


Figure 3.25: AIC scores of each number of clusters for girls dataset(Imputed Dataset)

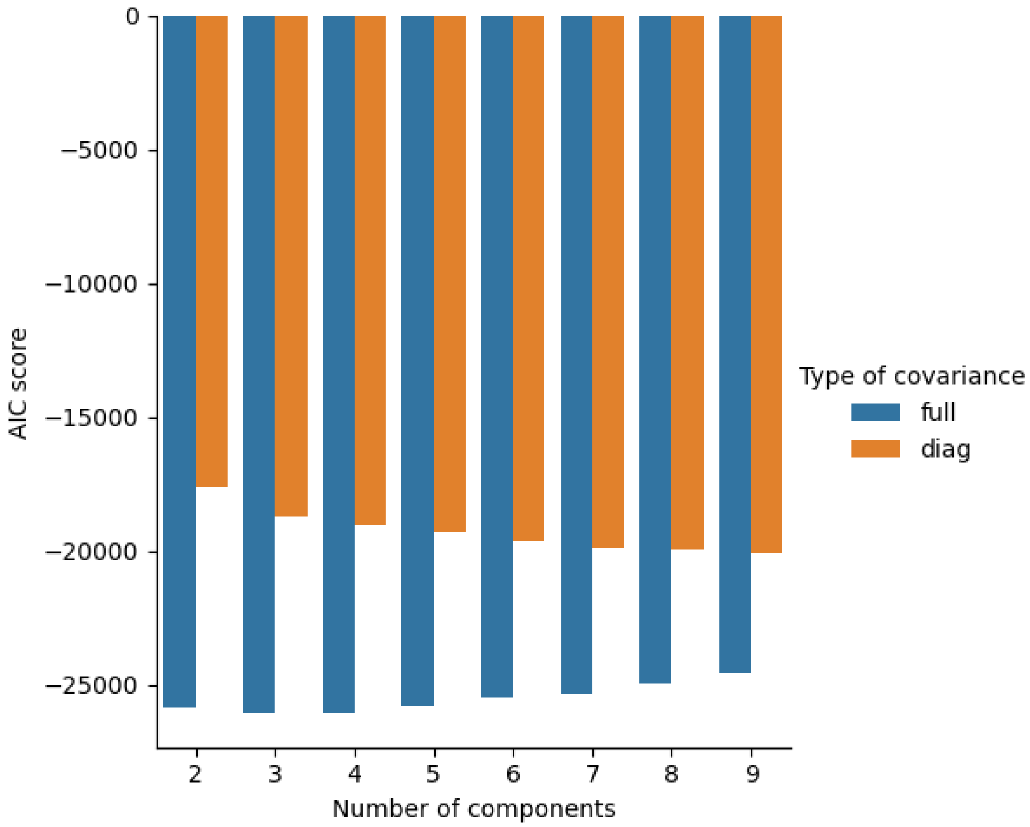


Figure 3.26: AIC scores of each number of clusters for boys dataset(Imputed Dataset)

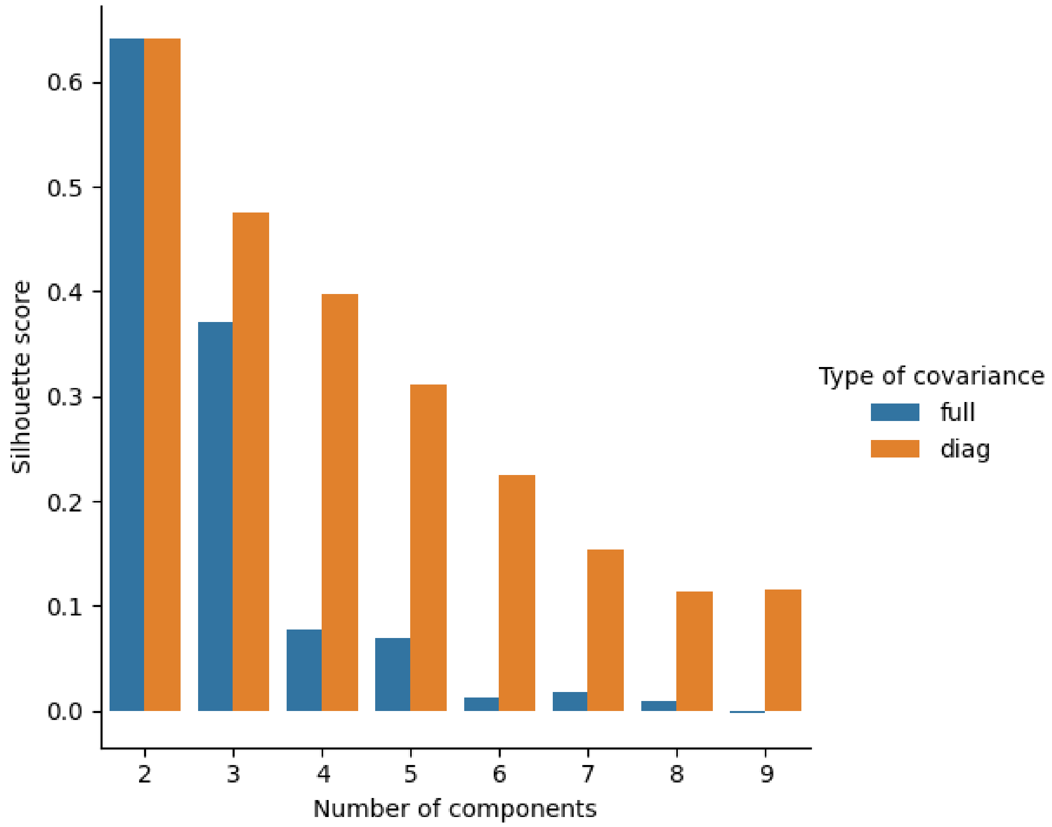


Figure 3.27: Silhouette scores of each number of clusters for Full dataset(Imputed Dataset)

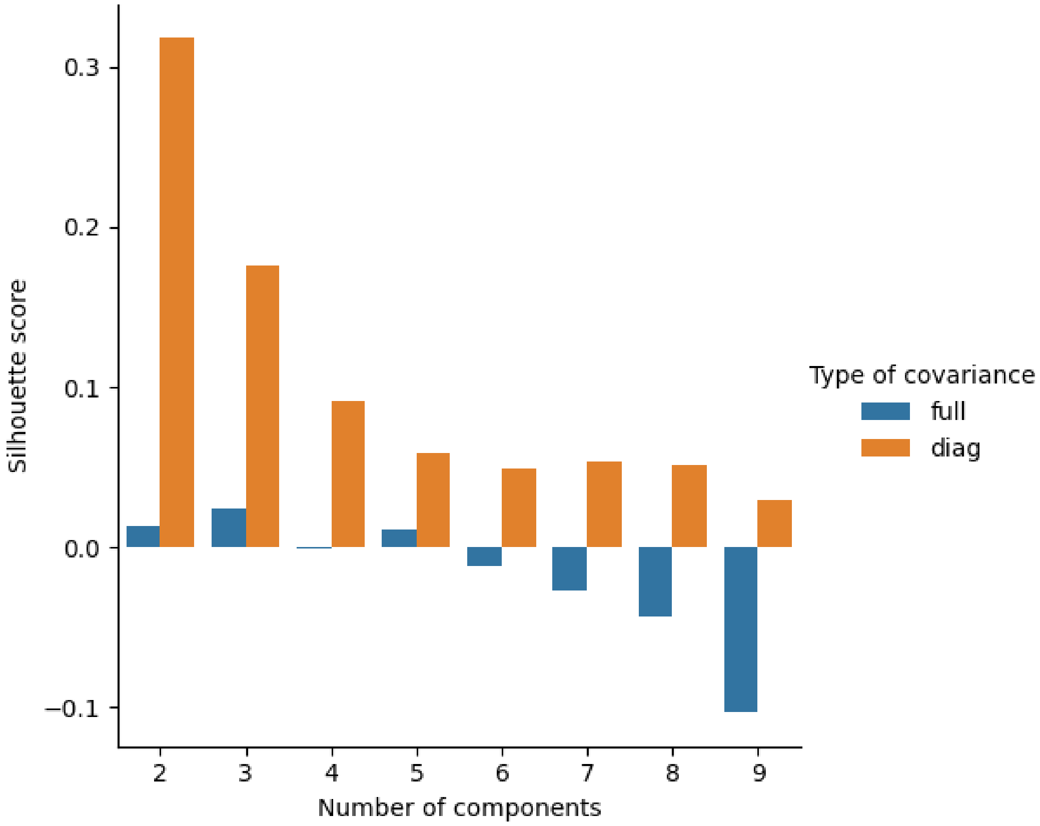


Figure 3.28: Silhouette scores of each number of clusters for girls dataset(Imputed Dataset)

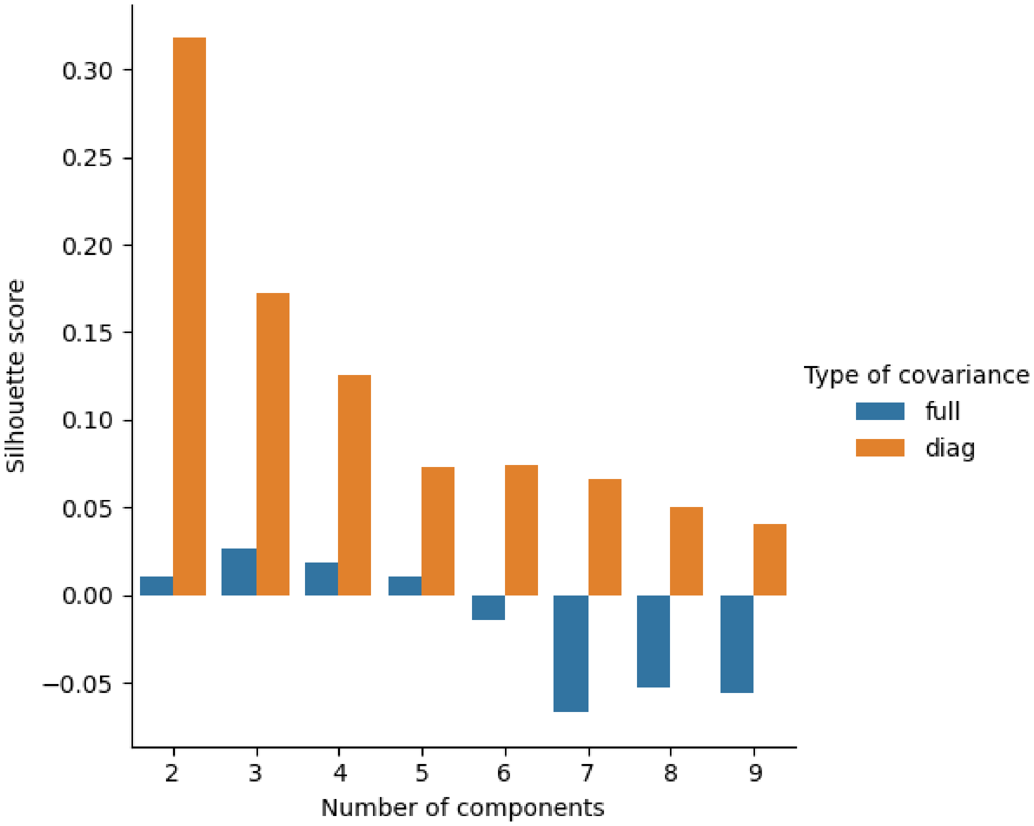


Figure 3.29: Silhouette scores of each number of clusters for boys dataset(Imputed Dataset)

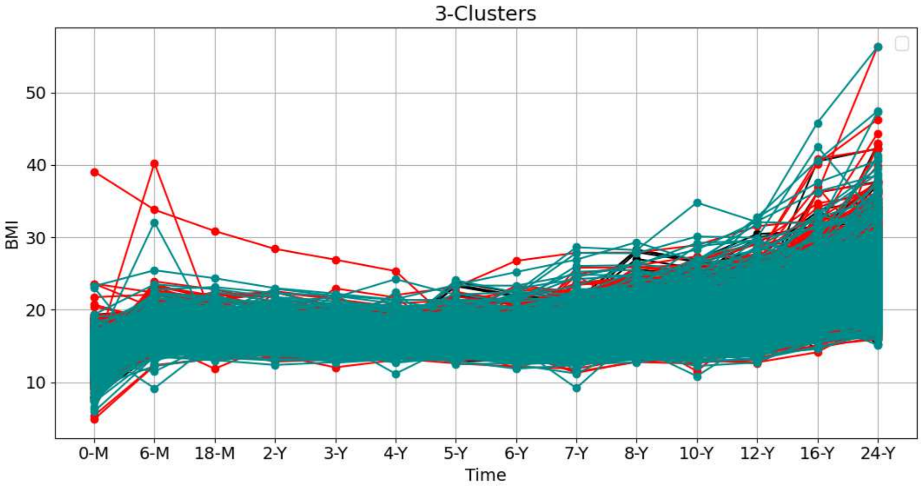


Figure 3.30: Trajectory of full patients for Imputed Dataset(girls and boys together)

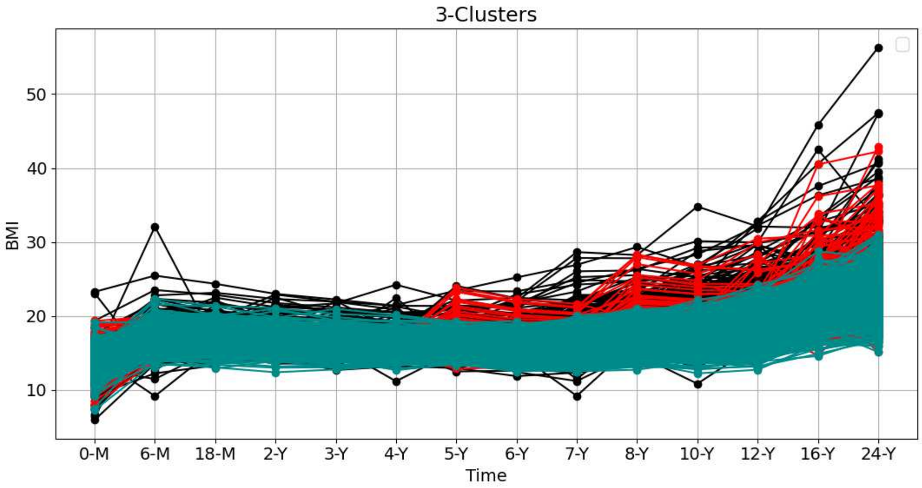


Figure 3.31: Trajectory of girls for Imputed Dataset

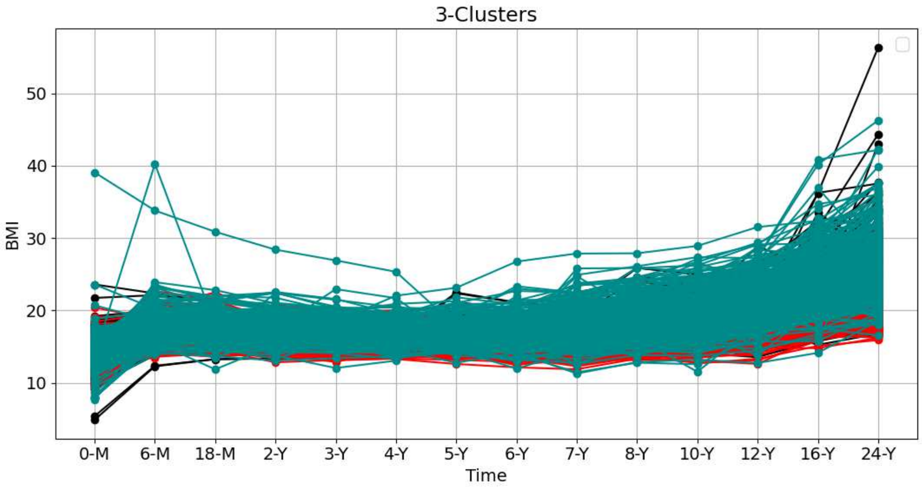


Figure 3.32: Trajectory of boys for Imputed Dataset

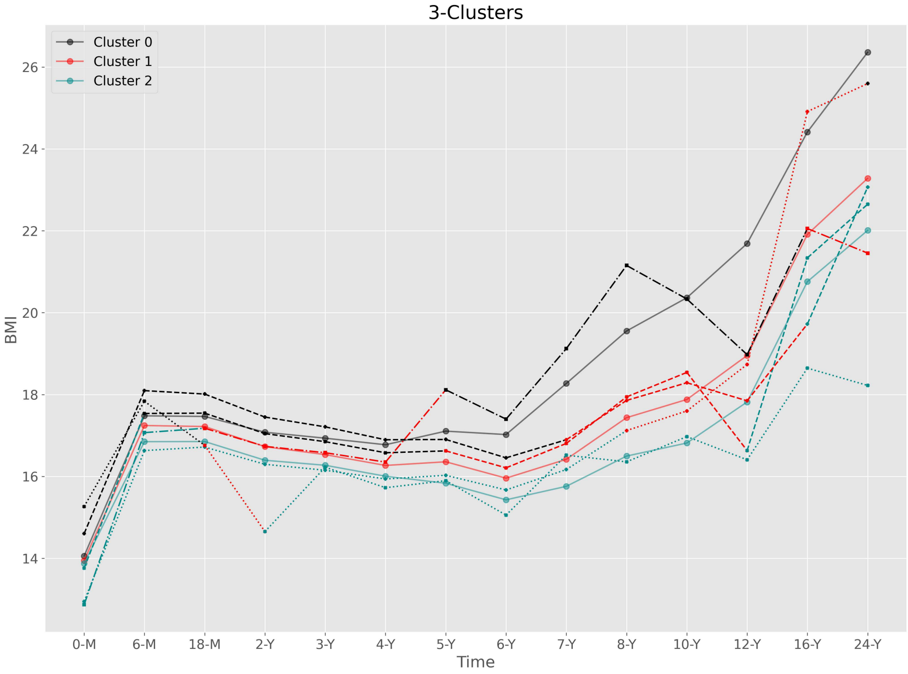


Figure 3.33: Mean of girls clustering trajectory with five random patients and λ = 0.5 for Imputed Dataset

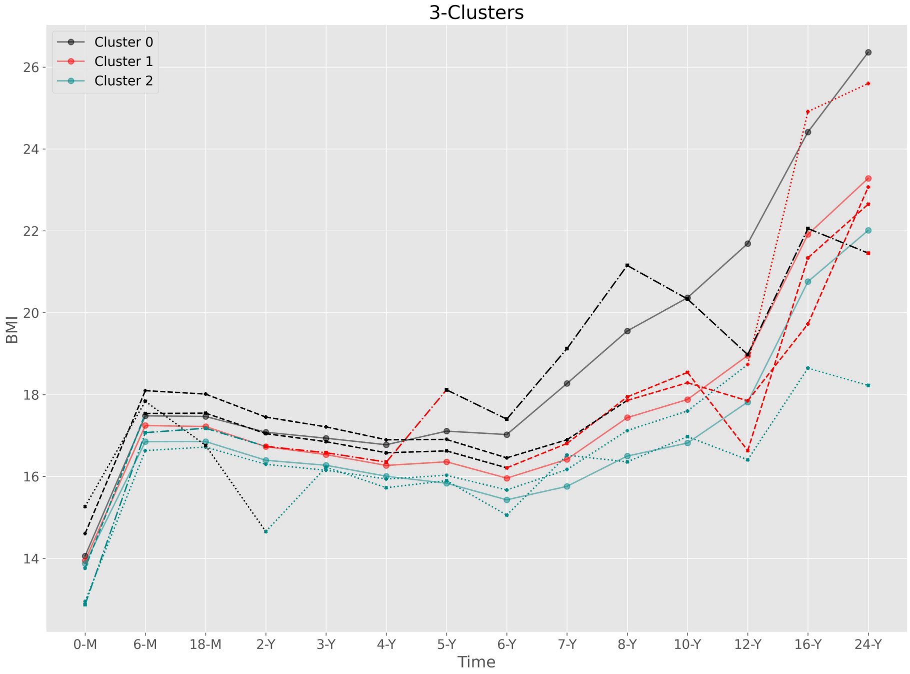


Figure 3.34: Mean of girls clustering trajectory with five random patients and λ = 0 for Imputed Dataset

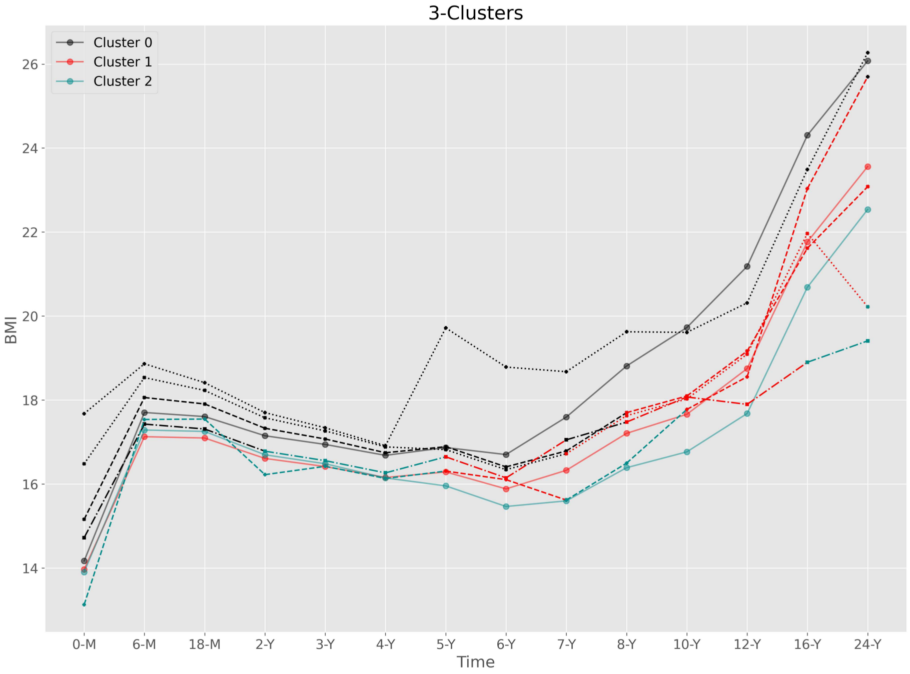


Figure 3.35: Mean of boys clustering trajectory with five random patients and λ = 0.5 for Imputed Dataset

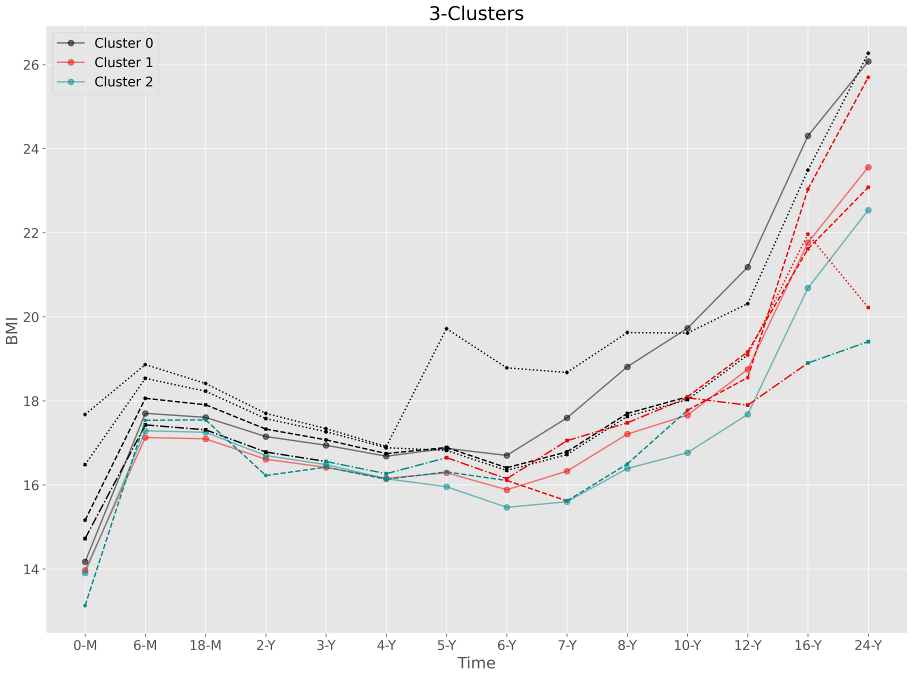
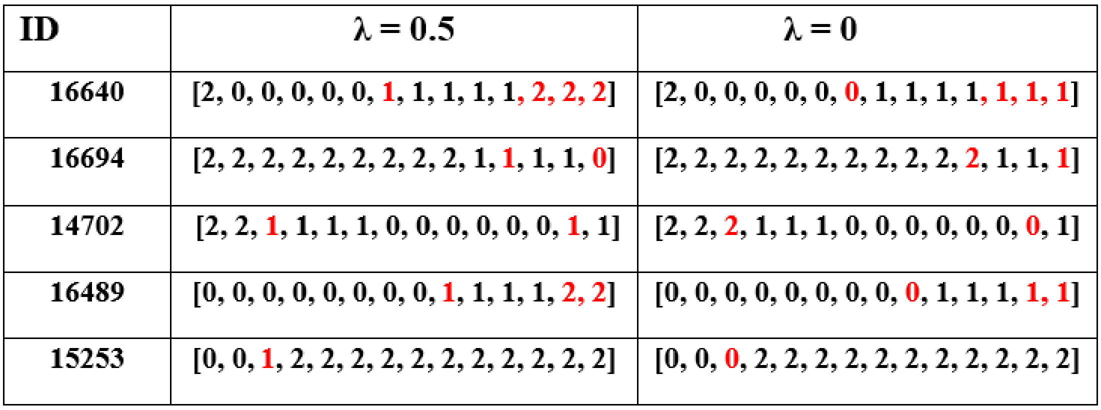
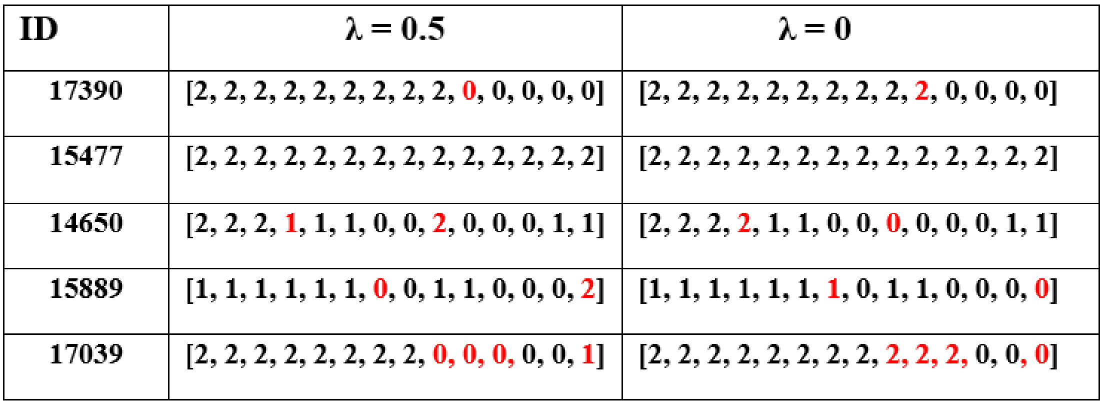


Figure 3.36: Mean of boys clustering trajectory with five random patients and λ = 0 for Imputed Dataset



ID	λ = 0.5	λ = 0
16640	[2, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2]	[2, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1]
16694	[2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 0]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1]
14702	[2, 2, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1]	[2, 2, 2, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1]
16489	[0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
15253	[0, 0, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]	[0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]

Figure 3.37: Patient cluster Changing of five random Patients for girls dataset(Imputed Dataset)



ID	λ = 0.5	λ = 0
17390	[2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0]
15477	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]
14650	[2, 2, 2, 1, 1, 1, 0, 0, 2, 0, 0, 0, 1, 1]	[2, 2, 2, 2, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1]
15889	[1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 2]	[1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0]
17039	[2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 1]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0]

Figure 3.38: Patient cluster Changing of five random Patients for boys(Imputed Dataset)

4

Conclusions and Future Works

This research explores the domain of longitudinal data analysis, with a concentration on the tracking of measurements of height and weight over a period of time. The dataset used for this analysis consisted of 3,897 data. The issues associated with missing data were thoroughly examined in this study. To ensure the integrity of our research, we applied many imputation techniques. After careful consideration, we ultimately selected the Mean Expected Growth strategy, which yielded superior findings compared to other imputation approaches. The proposed methodology involves estimating missing values within a dataset by computing the average of the longitudinal data for each specific time point. The growth rate of the data may be determined by comparing the values at different time points.

Additionally, the use of clustering techniques, notably Gaussian Mixture Models (GMM), has been emphasized in our study as a potent tool for revealing hidden patterns in longitudinal data. The usefulness of longitudinal data in comprehending patterns and trends through time, particularly in relation to anthropometric measures including heights and weights, has been effectively proven. The present investigation has yielded valuable insights pertaining to the patterns of development, levels of variability, and possible correlations with diverse elements. The forgetting factor was used to help with one of the goals, which was to get an understanding of the dynamic changes occurring within patients in each cluster via the use of a unique weighted technique to identify age-adjusted BMI growth trajectories. In time-series analysis and forecasting, the forgetting factor method includes giving increasingly smaller weights to older data.

The dataset used in our study exhibited a substantial number of missing data points. The consideration of missing data is a crucial aspect in the context of longitudinal research. We've evaluated how missing data may affect our results and used reliable imputation techniques to adjust for omitted observations. This has enabled us to reduce biases and enhance the precision of our findings.

The dataset was examined using two distinct approaches. Initially, patients with complete data were chosen (n=458) and subjected to clustering using Gaussian Mixture Models (GMM). The resulting clustering outcomes were treated as the Gold Standard. Subsequently, missing data in the entire dataset (n=3,897) were imputed using the Mean Expected Growth method, followed by clustering using GMM. The Gaussian Mixture Model (GMM) was used to cluster the data due to the presence of a mixture of Gaussian distributions within the data.

While the analysis of longitudinal data and the use of GMM have benefited greatly from our research, there are still a number of directions that warrant further research. One such direction is the use of advanced imputation techniques to investigate and develop more sophisticated imputation techniques that can handle missing data even more effectively. The enhancement of imputation accuracy may be achieved by using machine learning methodologies, including deep learning and probabilistic models. Additionally, it is noteworthy to remark that a comparative examination of several imputation approaches may be conducted using sensitivity analysis in order to choose the most optimal strategy.

Our findings support the OECD's recommendations on how to reduce the risks and improve outcomes associated with excessive body mass index (BMI) (World Health Organization, 2016 [20]) by focusing on preventative measures taken early in life, such as not smoking throughout pregnancy, breastfeeding, and maintaining a healthy weight before and during pregnancy.

Incorporating more patient parameters beyond heights, weights, and BMI may enhance the predictive power of our analysis. The inclusion of these new features expands the dimensionality of our dataset and offers a more complete perspective on the phenomena under investigation. This enhancement greatly enhances the predictive power of the models we use and facilitates a deeper comprehension and anticipation of events within our specific field of study. By expanding the dimensions of our data collection, we will introduce a broader spectrum of causes and variables that have the potential to influence our goal variable. The enhanced contextual information enables us to identify previously unnoticed connections, so yielding outcomes that are more precise and reliable. The recently implemented functionality, often obtained from an alternative domain, facilitates the acquisition of insights across other domains. The use of an interdisciplinary approach enhances the depth and breadth of our study by drawing from a wide range of knowledge sources and broadening the reach of our results. The incorporation of this particular function presents opportunities for further investigation in the future. The previously mentioned contribution not only enhances the current research but also establishes a basis for future inquiries, perhaps revealing unexplored facets of the phenomena and furthering our comprehension.

References

[1] Nikita Andriyanov, Alexander Tashlinsky, and Vitaly Dementiev. "Detailed clustering based on gaussian mixture models". In: *Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 2*. Springer. 2021, pp. 437–448.

[2] *BAMSE Project*. URL: `https://ki.se/en/imm/bamse-project`.

[3] *Cluster analysis*. URL: `https://en.wikipedia.org/wiki/Cluster_analysis`.

[4] *Clustering Algorithms*. URL: `https://developers.google.com/machine-learning/clustering/clustering-algorithms`.

[5] *Clustering in Machine Learning*. URL: `https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms`.

[6] Sandra Ekström et al. "Body mass index development and asthma throughout childhood". In: *American journal of epidemiology* 186.2 (2017), pp. 255–263.

[7] *Filling missing values using forward and backward fill in pandas dataframe ffill and bfill*. URL: `https://saturncloud.io/blog/filling-missing-values-using-forward-and-backward-fill-in-pandas-dataframe-ffill-and-bfill/`.

[8] *Gaussian Mixture Model*. URL: `https://medium.com/@joey.soehardinata/a-brief-introduction-to-gaussian-mixture-model-gmm-clustering-in-machine-learning-a00b9bb0fb17`.

[9] *Gaussian Mixture Model*. URL: `https://www.geeksforgeeks.org/gaussian-mixture-model/`.

[10] *gaussian mixture model*. URL: `file:///C:/Users/utente/Downloads/gaussian-mixture-models-clustering.pdf`.

[11] Joseph G Ibrahim and Geert Molenberghs. "Missing data methods in longitudinal studies: a review". In: *Test* 18.1 (2009), pp. 1–43.

[12] *Introduction to k-medoids Clustering*. URL: `https://subscription.packtpub.com/book/data/9781789956399/1/ch01lvl1sec08/introduction-to-k-medoids-clustering`.

[13] Noor Kamal Kaur, Usvir Kaur, and Dheerendra Singh. "K-Medoid clustering algorithm-a review". In: *Int. J. Comput. Appl. Technol* 1.1 (2014), pp. 42–45.

[14] *kmeans*. URL: `https://en.wikipedia.org/wiki/K-means_clustering`.

[15] Igor Kononenko. "Machine learning for medical diagnosis: history, state of the art and perspective". In: *Artificial Intelligence in Medicine* 23.1 (2001), pp. 89–109. ISSN: 0933-3657. DOI: `https://doi.org/10.1016/S0933-3657(01)00077-X`. URL: `https://www.sciencedirect.com/science/article/pii/S093336570100077X`.

[16] N. S. L. Phani Kumar, Sanjiv Satoor, and Ian Buck. "Fast Parallel Expectation Maximization for Gaussian Mixture Models on GPUs Using CUDA". In: *2009 11th IEEE International Conference on High Performance Computing and Communications*. 2009, pp. 103–109. DOI: `10.1109/HPCC.2009.45`.

[17] Youguo Li and Haiyan Wu. "A Clustering Method Based on K-Means Algorithm". In: *Physics Procedia* 25 (2012). International Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao, pp. 1104–1109. ISSN: 1875-3892. DOI: `https://doi.org/10.1016/j.phpro.2012.03.206`. URL: `https://www.sciencedirect.com/science/article/pii/S1875389212006220`.

[18] Yuan Luo. "Evaluating the state of the art in missing data imputation for clinical data". In: *Briefings in Bioinformatics* 23.1 (2022), bbab489.

[19] *Missing Data | Types, Explanation, Imputation*. URL: `https://www.scribbr.com/statistics/missing-data/`.

[20] *OECD*. URL: `https://www.who.int/docs/default-source/gho-documents/world-health-statistic-reports/world-heatlth-statistics-2016.pdf`.

[21] Kay I Penny and Ian Atkinson. "Approaches for dealing with missing data in health care studies". In: *Journal of clinical nursing* 21.19pt20 (2012), pp. 2722–2729.

[22] Heather A Robinson et al. "Post-2000 growth trajectories in children aged 4–11 years: A review and quantitative analysis". In: *Preventive Medicine Reports* 14 (2019), p. 100834.

[23] Meshal Shutaywi and Nezamoddin N Kachouie. "Silhouette analysis for performance evaluation in machine learning with applications to clustering". In: *Entropy* 23.6 (2021), p. 759.

[24] Kristina P Sinaga and Miin-Shen Yang. "Unsupervised K-means clustering algorithm". In: *IEEE access* 8 (2020), pp. 80716–80727.

[25] Niek Den Teuling, Steffen Pauws, and Edwin van den Heuvel. *Clustering of longitudinal data: A tutorial on a variety of approaches*. 2021. arXiv: `2111.05469 [stat.ME]`.

[26] *Unsupervised Learning: Clustering using Gaussian Mixture Model (GMM)*. URL: `https://behesht.medium.com/unsupervised-learning-clustering-using-gaussian-mixture-model-gmm-c788b280932b`.

[27] John VanBuren et al. "AIC identifies optimal representation of longitudinal dietary variables". In: *Journal of public health dentistry* 77.4 (2017), pp. 360–371.

[28] *What is Clustering?* URL: `https://developers.google.com/machine-learning/clustering/overview`.

[29] Julia Zhang and D Chen. "Interpolation calculation made EZ". In: *14th Annual Conference Proceedings, NorthEast SAS Users Group NESUG, Baltimore, MD*. 2001.

[30] Yi Zhang et al. "Gaussian mixture model clustering with incomplete data". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17.1s (2021), pp. 1–14.

[31] Zhongheng Zhang. "Multiple imputation with multivariate imputation by chained equation (MICE) package". In: *Annals of translational medicine* 4.2 (2016).

Acknowledgments

At this time, I would like to thank everyone who has assisted me with this research.

Prior to anything else, I would like to express my gratitude to my husband and my family for their unending support and love as I pursued my academic objectives. My accomplishments have been driven by the presence of my wife, my family, and my colleagues, and I will be eternally grateful to them.

My supervisors, Professor **Federica Battisti, Mr. Ashish Kumar, and notably Professor Saikat Chatterjee**, deserve my sincere appreciation for their outstanding advice, assistance, and expertise. Their assistance has been indispensable to my growth as a researcher and scholar. As I've labored to refine my craft, their understanding, encouragement, and incisive criticism have been invaluable.

In addition, I am thankful to the **University of Padua** for permitting me to pursue my academic interests. Academic resources and facilities have been indispensable to my success.

The KTH Royal Institute of Technology has provided me with the invaluable opportunity to pursue my thesis and collaborate with their distinguished students and faculty. KTH's assistance and resources were essential to the successful completion of this project.

I would like to express my deepest gratitude to **Erik Melén** for his indispensable role as the Principal Investigator (PI) of the BAMSE project, as well as to Gang Wang and the complete BAMSE working group for their unwavering commitment to scientific research. Their collective efforts in curating and distributing the dataset used in this study have been indispensable to its advancement.